HEWLETT-PACKARD

JOURNAL

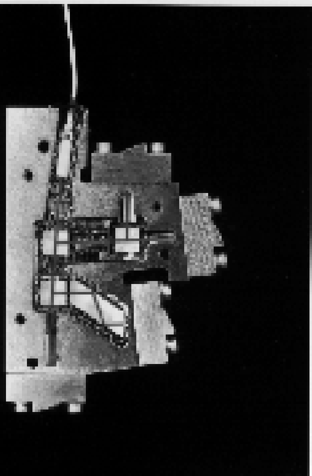


table of contents

February 1995,
Volume 46, Issue 1

Articles

1 **Broadband Frequency Characterization of Optical Receivers Using Intensity Noise**
by Douglas M. Baney, Wayne B. Soren

2 **Erbium-Doped Fiber Amplifier Test System**
by Edgar Leckel, Jurgen Sang, Rolf Muller, Clemens Ruck, and Christian Hentschel

3 **Multi-Quantum-Well Ridge Waveguide Lasers for Tunable External-Cavity Sources**
by Tirumala R. Ranganath, Michael J. Ludowise, Patricia A. Beck, Dennis J. Derickson, William H. Perez, Tim L. Bagwell, and David M. Braun

4 **Measurement of Polarization-Mode Dispersion,**
by Brian L. Heffner and Paul R. Hernday

5 **A New Design Approach for a Programmable Optical Attenuator**
by Siegmar Schmidt and Halmo Fischer

6 **Precision Reflectometer with Spurious-Free Enhanced Sensitivity,**
by David M. Braun, Dennis J. Derickson, Luis M. Fernandez, and Greg D. LeCheminant

7 **High-Power, Low-Internal-Reflection, Edge Emitting Light-Emitting Diodes**
by Dennis J. Derickson, Patricia A. Beck, Tim L. Bagwell, David M. Braun, Julie E. Fouquet, Forrest G. Kellert, Michael J. Ludowise, William H. Perez, Tirumala R. Ranganath, Gary R. Trott, and Susan R. Sloan

8 **Jitter Analysis of High-Speed Digital Systems**
by Christopher M. Miller and David J. McQuate

9 **Automation of Optical Time-Domain Reflectometry Measurements**
by Frank A. Maier and Harald Seeger

10 **Design and Performance of a Narrowband VCO at 282 THz**
by Peter R. Robrish, Christopher J. Madden, Rory L. VanTuyl, and William R. Trutna, Jr.

11 **Surface Emitting Laser for Multimode Data Link Applications**
by Michael R.T. Tan, Kenneth H. Hahn, Yu-Min D. Houng, and Shih-Yuan Wang

12 **Generating Short-Wavelength Light Using a Vertical-Cavity Laser Structure**
by Shigeru Nakagawa, Danny E. Mars, and Norihide Yamada

13 **A New, Flexible Sequencer Architecture for Testing Complex Serial Bit Streams**
by Robert E. McAuliffe, James L. Benson, and Christopher B. Cain

14 **Shortening the Time to Volume Production of High-Performance Standard Cell ASICs**
by Jay D. McDougal and William E. Young

15 **A Framework for Insight into the Impact of Interconnect on 0.35-mm VLSI Performance,**
by Prasad Raje

16 **Synthesis of 100% Delay Fault Testable Combinational Circuits by Cube Partitioning**
by William K. Lam

17 **Better Models or Better Algorithms? Techniques to Improve Fault Diagnosis,**
by Robert C. Aitken and Peter C. Maxwell

Broadband Frequency Characterization of Optical Receivers Using Intensity Noise

Methods for enhancing the dynamic range of the intensity noise technique for high-frequency photoreceiver calibration are proposed and experimentally demonstrated. These methods combine recently developed EDFA* technology with spectral filtering techniques. The intensity noise calibration technique is portable, easy to use, and field deployable.

by Douglas M. Baney and Wayne V. Sorin

Optical technology will play an important role in building the coming information superhighway through its capacity to provide high information throughput and low-loss transmission simultaneously. Optical sources such as semiconductor lasers provide an optical carrier whose intensity is modulated with information to be sent over fiber-optic cable. For high-data-rate communications, lasers typically operate near the 1.55-μm or 1.3-μm low-loss wavelengths in optical fiber. Optical receivers convert the information modulated onto the optical carrier to baseband electrical signals. As demands are made for more information throughput, the bandwidths of optical receivers are increased commensurately.

Accurate characterization of the frequency response of an optical receiver is important to ensure that the receiver is compatible with the data transmission rate. Currently, a number of techniques exist to characterize the frequency response of optical receivers. One method is to compute the Fourier transform from the time-domain impulse response.[1] Frequency domain techniques, such as that used in the HP 8703 lightwave component analyzer, employ frequency-swept sinusoidal modulation of the optical intensity using a high-frequency $LiNbO_3$ modulator. This allows commercially available response measurements from 130 MHz to 20 GHz. The optical heterodyne technique using two Nd-YAG lasers has demonstrated capability from ≈10 MHz to 50 GHz at a wavelength of 1.32 μm.[1,2] These techniques all involve specific trade-offs among frequency coverage, experimental complexity, and sensitivity, and none are completely satisfactory. It is desirable to have the capability to measure the broadband frequency response with a simple rugged optical instrument.

The intensity noise technique offers the possibility of measuring frequency response characteristics of photoreceivers across the entire frequency span of modern electrical spectrum analyzers (for example, 9 kHz to 50 GHz for the HP 8565E). This intensity noise method was first demonstrated using a semiconductor optical amplifier as a source.[3] This technique is of particular interest because the noise exists at all frequencies simultaneously, permitting very rapid optical receiver characterization. Additionally, an unpolarized short-coherence-length optical source is used. This is advantageous because it makes the measurements immune to polarization drifts and time-varying interference effects from multiple optical reflections, thereby allowing stable, repeatable measurements.

Intensity Noise Techniques

Intensity noise techniques take advantage of the beating between various optical spectral components of a broad-bandwidth spontaneous emission source. Any two spectral lines will beat, or mix, to create an intensity fluctuation with a frequency equal to the frequency difference between the two lines. This concept is illustrated in Fig. 1. Since the optical bandwidth of spontaneous emission sources can easily exceed thousands of gigahertz, the intensity beat noise will have a similar frequency content. The fluctuations in optical intensity are referred to as spontaneous-spontaneous, or sp-sp, beat noise.

There are many sources of broad-bandwidth spontaneous emission. Hot surfaces (such as tungsten light bulbs) can provide optical radiation ranging from the visible to the far infrared. Semiconductor sources such as edge emitting light-emitting diodes (EELEDs) provide increased power densities over a wavelength range of about 100 nm. Still higher power densities can be obtained from solid-state sources

* Erbium-doped fiber amplifier.

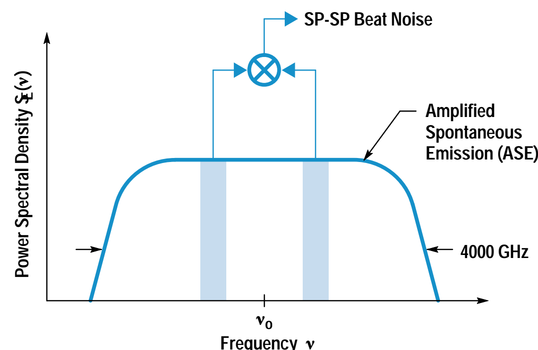


Fig. 1. Spontaneous-spontaneous (sp-sp) beat noise arising from mixing of the various spectral components from a thermal-like optical noise source.

© Hewlett-Packard Company 1995

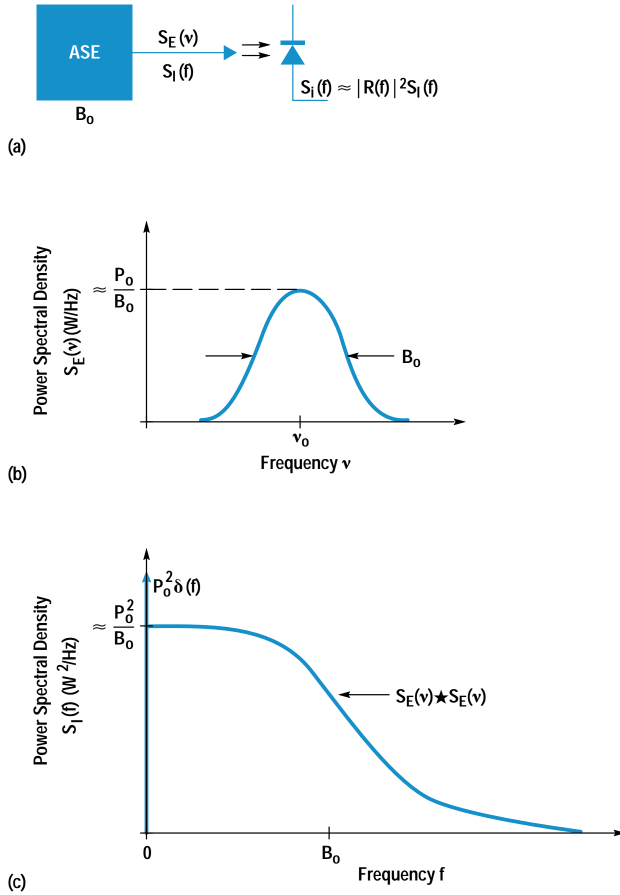


Fig. 2. (a) The intensity noise technique for optical receiver calibration. (b) The optical field spectrum as measured on an optical spectrum analyzer. (c) The optical intensity spectrum, which is proportional to the photocurrent spectrum, as measured on an electrical spectrum analyzer.

such as fiber-optic amplifiers (see page 9). The ability to couple these sources of broadband light efficiently into single-mode optical fiber is also important. Coupled power densities can range from about 500 pW/nm for a light bulb to greater than 1 mW/nm for amplified spontaneous emission (ASE) from a fiber-optic amplifier. The high power densities from fiber-optic amplifiers are particularly well-suited for intensity noise generation.

Bandpass-Filtered Intensity Noise Technique

Fig. 2 illustrates how the strength and line shape of the optical intensity beat noise are generated from a thermal-like source such as amplified spontaneous emission. Fig. 2a shows the optical power from an ASE source of spectral bandwidth B_o incident on the high-frequency optical receiver to be tested. The power spectral density of the optical field, $S_E(\nu)$, scaled so that its units are watts/Hz, is commonly used to characterize the output of optical sources. This quantity is typically measured using an optical spectrum analyzer. Fig. 2b shows a typical spectrum for $S_E(\nu)$ with full width at half maximum bandwidth B_o and center frequency ν_o. The average optical power P_o is found by integrating over the optical spectrum, that is,

$$P_o = \int_0^{\infty} S_E(\nu)d\nu,$$

which gives a peak value for the spectral density of approximately P_o/B_o. It should be noted that all power spectral densities discussed in this paper are single-sided, with energy only at positive frequencies.

Since photodiode current is proportional to optical intensity and not electric field, a more relevant quantity to consider is the power spectral density of the optical intensity, $S_I(f)$, which is scaled to have units of watts^2/Hz. This quantity is related to the optical field spectrum by the relatively simple expression:

$$S_I(f) = P_o^2\,\delta(f) + S_E(\nu) \star S_E(\nu), \qquad (1)$$

where $\delta(f)$ represents a delta function and $\star$ denotes the single-sided autocorrelation (integrated over positive frequencies). This result assumes unpolarized ASE and single-sided power spectral densities. This expression is valid for optical radiation with thermal-like noise statistics, such as amplified spontaneous emission. The effects of shot noise are not included in equation 1 because it is easier to account for this noise after the optical intensity is converted to a photocurrent. The result of equation 1 is graphically illustrated in Fig. 2c. Intensity beat noise exists up to frequencies determined by the spectral width of the ASE source. These frequencies are typically on the order of thousands of gigahertz. The magnitude of the beat noise at low frequencies is approximately P_o^2/B_o, the exact value depending on the line shape of the electric field spectrum. More accurate values for typically encountered line shapes are given later.

The power spectral density of the detected photocurrent, $S_i(f)$, in units of amperes^2/Hz, equals that of the incoming intensity spectrum except for the filtering effects of the photodiode. If the frequency dependent responsivity (i.e., transfer function) for the photodiode is given by R(f), which has units of amperes/watt, then the photocurrent spectrum can be expressed as:

$$S_i(f) \simeq |R(f)|^2 S_I(f) \qquad (2)$$

The receiver thermal and shot noises are not included in equation 2. The value for these two noise sources will determine the SNR (signal-to-noise ratio) for the measurement technique. Assuming that the optical beat noise signal is larger than the shot or thermal noise, this expression can be used to determine the magnitude $|R(f)|$ of the frequency response of the photodiode. This measurement is performed by observing the photocurrent spectrum $S_i(f)$ with an electrical spectrum analyzer. If the spectral width of the ASE source, B_o, is much larger than the frequency response of the photodiode, then $S_I(f)$ can be assumed constant, and the responsivity squared $|R(f)|^2$ of the photodiode is displayed directly on the electrical spectrum analyzer.

One previous difficulty in the practical use of this detector calibration technique is the small value for the intensity beat noise provided by typical ASE sources. This has presented a problem for high-frequency calibration because the thermal noise level associated with wideband receivers is typically quite large. In this paper we show that by combining the recent development of erbium-doped fiber amplifiers with spectral filtering techniques, we can overcome the previous limitation of small signal strength.

© Hewlett-Packard Company 1995

To understand how to optimize the ASE for detector calibration, the concept of relative intensity noise or RIN, which has units of Hz^{-1}, will be introduced. This parameter can be thought of as the fractional intensity noise associated with an optical source. The definition for RIN in terms of the optical intensity spectrum is:

$$RIN(f) = S_I(f)/P_o^2. \quad (3)$$

Thus RIN is the spectral density of the optical intensity at a given frequency divided by its integrated value at zero frequency. An equivalent definition equates RIN to the variance of the optical intensity $\langle \Delta I^2(f) \rangle$ in a one-hertz bandwidth divided by the average intensity squared. Since the optical bandwidth B_o is usually much larger than the electrical detection bandwidth, the RIN from an ASE source is usually considered to be constant, equal to its low-frequency value. From Fig. 2c, it can be seen that this value is approximately given by:

$$RIN \simeq 1/B_o. \quad (4)$$

This result is valid for unpolarized light with thermal-like noise statistics. A more accurate value for equation 4 requires knowledge of the line shape of the ASE spectrum.

Normally, for a given source, one would increase the average optical power incident onto a photoreceiver to increase the photocurrent beat noise and hence the measurement sensitivity. With the development of EDFAs (erbium-doped fiber amplifiers), the average optical powers attainable will easily saturate high-frequency photoreceivers. This means increasing optical power is no longer an issue, and for a given optical receiver, a power just below its saturation level should be used for calibration purposes. Since incident optical power can now be considered a constant, depending on the detector saturation level, the concept of RIN becomes useful because increasing its value increases the SNR of the intensity noise technique. According to equation 4, this means that decreasing the spectral width B_o of the ASE source will improve the SNR for detector calibration. The only limitation is that eventually there will be a roll-off in the high-frequency content of the beat noise.

Effects of Line Shape and Bandwidth

As the optical bandwidth is reduced, the maximum frequency separation of beating components also decreases. In Fig. 3, the RIN is shown as a function of frequency for four different optical bandwidths. The plotted curves correspond to the case of an unpolarized ASE source with a Gaussian optical spectral shape. The highest RIN is achieved at the lowest frequencies with the narrowest optical bandwidth. At frequencies comparable to the optical spectral bandwidth, roll-off in the RIN becomes apparent. The RIN for the 5-nm spectral width is relatively constant to several hundred gigahertz. Expressions for RIN as a function of frequency are shown in Table I for the case of Lorentzian, Gaussian, and rectangular optical field spectrums. The Lorentzian shape corresponds closely to the spectrum of commonly used Fabry-Perot optical filters. The Gaussian spectrum is sometimes used to describe the spectra of EELEDs or lasers operating below threshold. The rectangle function approximates some interference filters, grating-based monochromators, and chirped fiber gratings. The normalized (to unity) optical field spectrum is also included in Table I for reference. The line width B_o in each case is the FWHM (full width at half maximum) of the optical field spectrum.

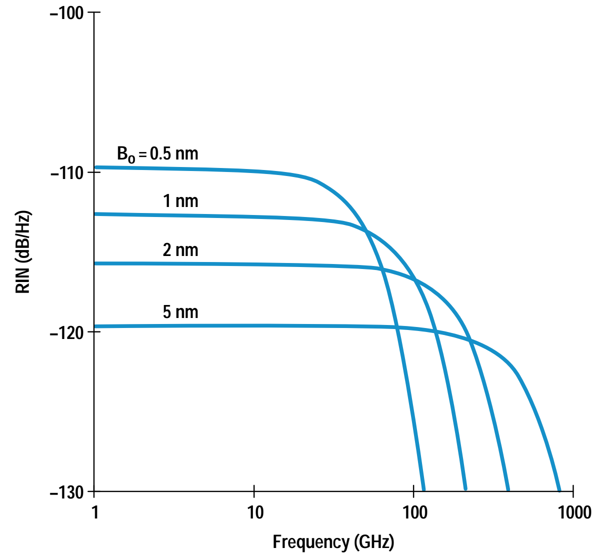


Fig. 3. Relative intensity noise (RIN) for a Gaussian-shaped optical field power spectrum centered at 1.55 μm.

Comparing the low-frequency RIN values for each of the different spectral shapes, the rectangle spectrum delivers the most RIN ($1/B_o$), followed by the Gaussian ($0.66/B_o$),

Table I
Relationship between Unpolarized Optical Field Spectrum and RIN

Optical Field Spectrum Shape	Normalized $S_E(\nu)$	RIN(f) (f > 0)
Rectangle*	$\Pi\left(\frac{\nu-\nu_o}{B_o}\right)$	$\frac{1}{B_o}\Lambda\left(\frac{f}{B_o}\right)$
Gaussian	$\exp\left\{-(4\ln2)\left(\frac{\nu-\nu_o}{B_o}\right)^2\right\}$	$\frac{1}{B_o}\frac{\sqrt{2\ln2}}{\sqrt{\pi}}\exp\left\{-(2\ln2)\left(\frac{f}{B_o}\right)^2\right\}$
Lorentzian	$\frac{1}{1+4\left(\frac{\nu-\nu_o}{B_o}\right)^2}$	$\frac{1}{B_o}\frac{1}{\pi}\frac{1}{1+(f/B_o)^2}$

* Π = rectangle function. Λ = triangle function.

© Hewlett-Packard Company 1995

1.55-μm Fiber-Optic Amplifier

The fiber-optic amplifier is a key element in modern high-data-rate communications. Its large optical bandwidth, ≈4000 gigahertz, makes it effectively transparent to data rate and format changes, allowing system upgrades without modifications to the amplifier itself. The most prevalent fiber-optic amplifier is called the erbium-doped fiber amplifier (EDFA). It provides amplification in the third telecommunications window centered at a wavelength of 1.55 μm. The EDFA has four essential components, as shown in Fig. 1. These are the laser diode pump, the wavelength division multiplexer (WDM), the erbium-doped optical fiber, and the optical isolators. To achieve optical amplification, it is necessary to excite the erbium ions situated in the fiber core from their ground state to a higher-energy metastable state. A diagram of the relevant erbium ion energy states is shown in Fig. 2. The erbium ions are excited by coupling pump light (≈20 mW or greater) through the WDM into the erbium-doped fiber. Commonly used pump wavelengths are 980 nm and 1480 nm. The ions absorb the pump light and are excited to their metastable state. Once the ions are in this state, they return to the ground state either by stimulated emission or, after about 10 ms, through spontaneous emission. Light to be amplified passes through the input isolator and WDM and arrives at the excited erbium ions distributed along the optical fiber core. Stimulated emission occurs, resulting in additional photons that are indistinguishable from the input photons. Thus, amplification is achieved. Optical isolators shield the amplifier from reflections that may cause lasing or the generation of excess amplified spontaneous emissions (ASE).

Good EDFA design typically requires reduction of optical losses at the amplifier input and minimizing optical reflections within the amplifier. EDFAs with greater than 30-dB optical gain, more than 10-mW output power, and less than 5-dB noise figure are readily achieved in practice.

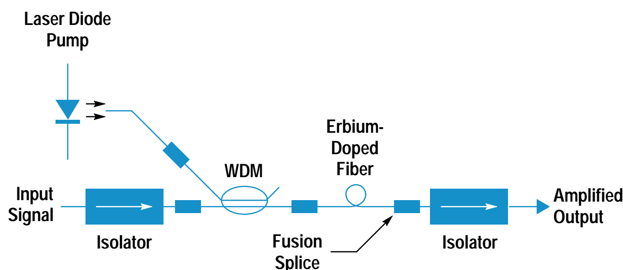


Fig. 1. Schematic of an erbium-doped fiber amplifier showing essential optical components. The WDM is a wavelength division multiplexer.

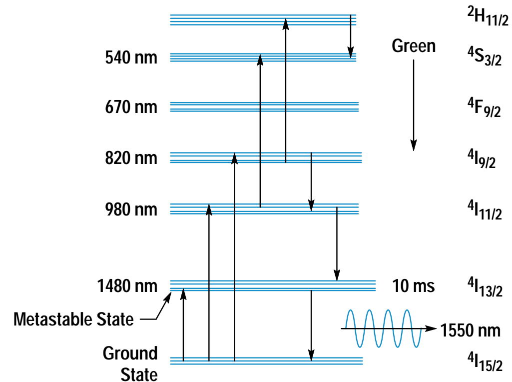


Fig. 2. Relevant erbium ion energy levels.

ASE is generated in optical amplifiers when excited ions spontaneously decay to the ground state. The spontaneously emitted photons, if guided by the optical fiber, will subsequently be amplified (by the excited ions) as they propagate along the fiber. This can result in substantial ASE powers (>10 mW) at the amplifier output. A typical spectrum of signal and ASE at the amplifier output is shown in Fig. 3. The ASE can extend over a broad spectral range, in this case in excess of 40 nm.

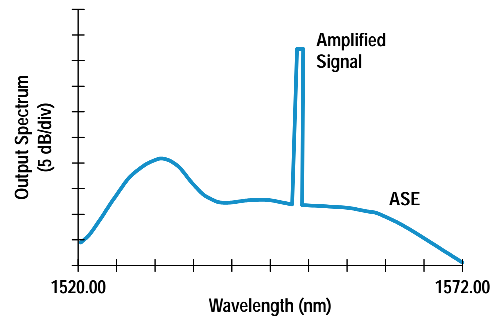


Fig. 3. Spectrum of amplified spontaneous emissions (ASE) and amplified signal at the amplifier output.

and next by the Lorentzian ($0.32/B_O$). If the ASE source is polarized, the RIN will be twice as large for all three cases.

If the optical receiver response measurements are performed in the flat RIN regime (see Fig. 3), no specific knowledge of the source line shape is required. However, if the spectral line shape for the ASE source is accurately characterized, additional measurement sensitivity can be obtained by using narrower optical filters and correcting for the roll-off in the response data.

Experiment: Intensity Noise Technique Using Optical Bandwidth Reduction

To demonstrate the filtered intensity noise technique, measurements were performed at a wavelength of 1.55 μm on a SONET receiver with 1-GHz electrical bandwidth. The measurement setup is shown in Fig. 4. The mirror at the end of the EDFA results in two-pass ASE generation. This doubly amplified ASE is then filtered by an optical bandpass filter with a 1-nm spectral width. All optical connections were fusion-spliced to eliminate optical reflections. The presence of multiple optical reflections could impart undesirable intensity ripple onto an otherwise constant RIN spectrum, which would be difficult to separate from the receiver frequency response. The bandpass filtering resulted in approximately a 15-dB improvement in SNR. An average optical power of 200 μW was incident onto the receiver; this was below its saturation level of 320 μW. Using a more complicated dual Nd-YAG optical heterodyne system, measurements at a wavelength of 1.32 μm were performed for comparison with the intensity noise technique. The intensity noise technique shows excellent agreement when compared with the standard heterodyne method as indicated in Fig. 5.

Periodically Filtered Intensity Noise Technique

As discussed earlier, for a fixed average optical power, the spontaneous-spontaneous beat noise in the photocurrent spectrum increases as the optical bandwidth B_O is reduced. This increase in signal strength is desirable, but if the optical bandwidth is reduced too much, the high-frequency content of the beat noise will start to roll off, making it unsuitable

© Hewlett-Packard Company 1995

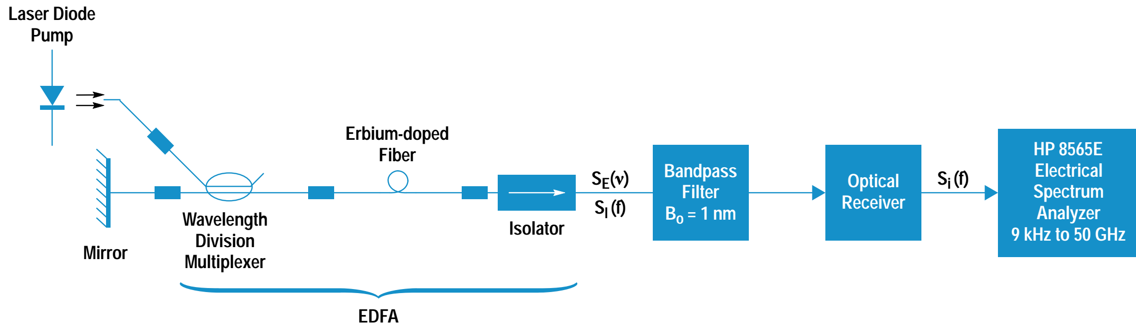


Fig. 4. Experimental arrangement for frequency response measurement of a SONET photoreceiver using the bandpass-filtered intensity noise technique.

for high-frequency detector calibration. In practice, this trade-off between signal strength and frequency content becomes a problem when trying to characterize high-frequency, low-gain optical receivers. Because of the large input noise figures (typically 30 dB) associated with high-frequency electrical spectrum analyzers, large values of photocurrent beat noise are required.

The periodically filtered intensity noise technique solves this trade-off problem by allowing the magnitude of the beat noise to be increased without loss of its high-frequency content.[4] This is accomplished by passing the amplified spontaneous emission through a Fabry-Perot filter and then on to the high-frequency optical receiver. This reduces the average optical power while maintaining the magnitude of the spontaneous-spontaneous beat noise at periodic frequencies spaced at intervals equal to the free spectral range of the filter. The result is an increase in RIN at periodically spaced beat frequencies. If needed, optical amplification can then be used to boost the average optical power to a value just under the saturation level for the receiver.

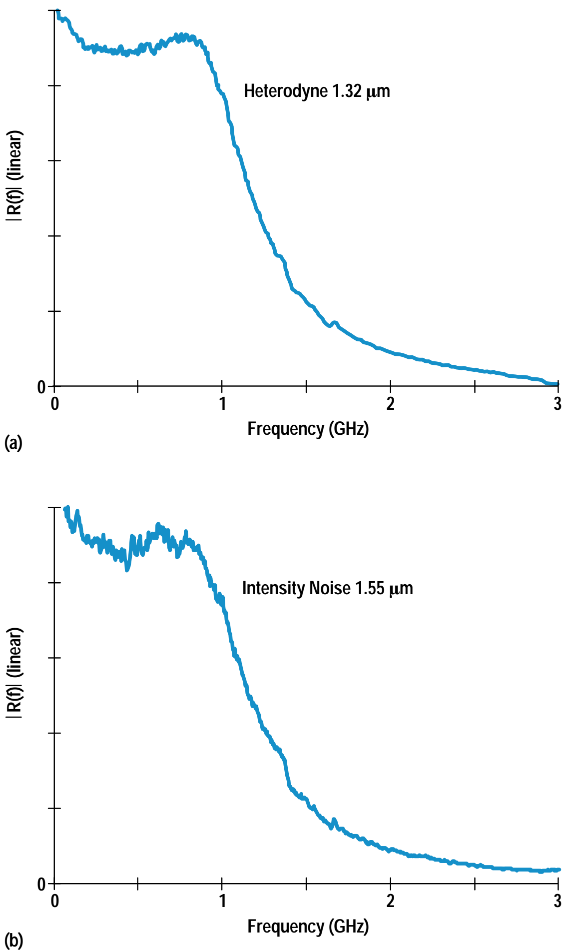


Fig. 5. Measurement of a SONET receiver frequency response using (a) a 1.32-μm Nd-YAG optical heterodyne system and (b) the 1.55-μm filtered intensity noise technique.

Details of the periodically filtered technique are shown in Fig. 6. Fig. 6a shows the ASE from a source such as an erbium-doped fiber amplifier passing through a Fabry-Perot filter and on to the test optical receiver. The transmission characteristics of the Fabry-Perot filter are determined by its free spectral range, FSR, and its finesse, F. FSR is the frequency separation of the transmission maxima and F is the ratio of the FSR to the width of the transmission maxima. From a signal processing point of view, the Fabry-Perot filter has a frequency-domain transfer function for the optical field given by $H(\nu)$. The squared magnitude of this transfer function is illustrated in Fig. 6b. For the ideal case, the transmission through the filter would be unity at frequencies separated by the FSR. After passing through the filter, the power spectral density of the input optical field $S_E(\nu)$ will be transformed to $|H(\nu)|^2 S_E(\nu)$. Fig. 6c shows the filtered spectrum, which is incident on the optical receiver. Strong beat signals occur in the intensity separated by frequency intervals equal to the FSR of the Fabry-Perot filter. As described earlier in equation 1, the power spectral density for the optical intensity $S_I(f)$ can be obtained from an autocorrelation of the input electric field spectrum:

$$S_I(f) \simeq P_o^2\delta(f) + |H(\nu)|^2 S_E(\nu) \star |H(\nu)|^2 S_E(\nu). \qquad (5)$$

This expression is valid for unpolarized amplified spontaneous emission (see equation 1). The result of equation 5 is illustrated in Fig. 6d. Intensity noise peaks are evident at frequency locations separated by the FSR of the Fabry-Perot filter. Relative to the dc signal, these peaks are a factor of F/π larger than the unfiltered case shown in Fig. 2c. For a typical finesse of F = 100, this corresponds to an increase in signal-to-noise ratio of about 15 dB for photodiode characterization. The penalty for the increased SNR is that beat signals only occur at specific frequencies. This constraint is not very severe because this frequency spacing can be set to any

© Hewlett-Packard Company 1995

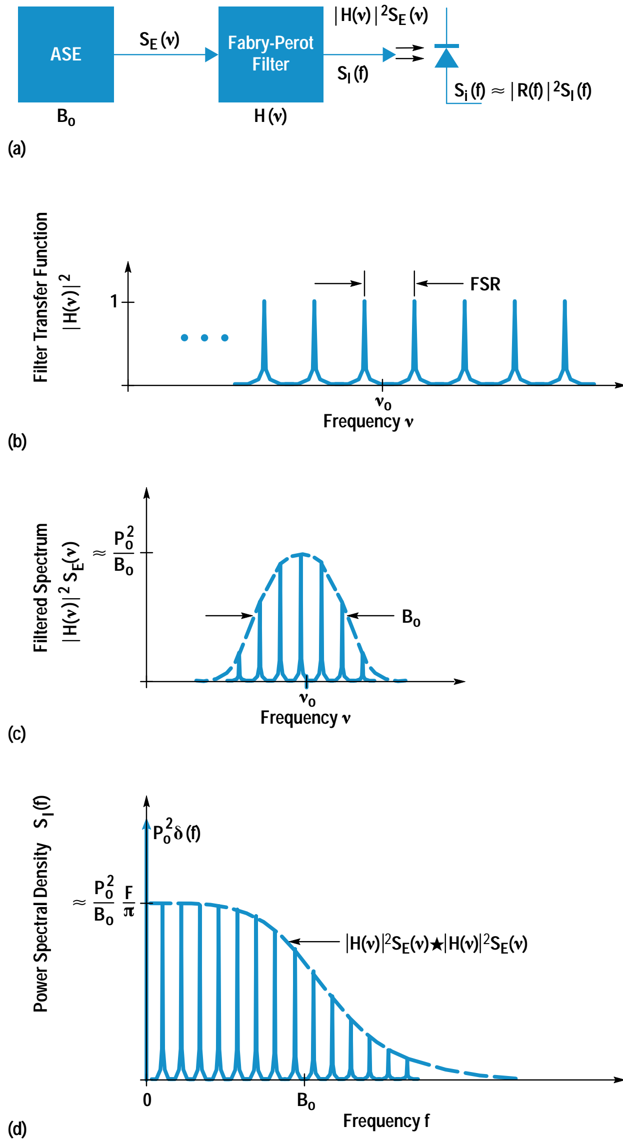


Fig. 6. (a) Block diagram illustrating the periodically filtered intensity noise calibration technique. (b) Transfer function of the Fabry-Perot filter. (c) Optical field spectrum incident at the photoreceiver. (d) Photocurrent spectrum as measured on an electrical spectrum analyzer.

desired value by proper choice of the filter FSR. As described earlier, the optical receiver frequency response can be obtained using equation 2, which relates the optical intensity spectrum to the photocurrent spectrum measured using an electrical spectrum analyzer. Assuming an optical bandwidth B_o much larger than the frequency response of the optical receiver, the photocurrent power spectrum is given by:

$$S_i(f) \simeq |R(f)|^2 S_I(f) = |R(f)|^2 \frac{P_o^2}{B_o} \cdot \frac{F}{\pi} \sum_k \frac{1}{1 + \left(\frac{f - kFSR}{\Delta\nu}\right)^2}$$

where $\Delta\nu$ is the spectral width of the Fabry-Perot filter and is given by $\Delta\nu$ = FSR/F. This result illustrates that the magnitude of the photocurrent beat signal can be increased relative to a fixed average current without sacrificing its high-frequency content.

Experiment: Intensity Noise Technique Using Periodic Bandwidth Reduction

To demonstrate the above result, the arrangement illustrated in Fig. 7 was used. The ASE obtained from the two-pass EDFA superfluorescent source generated an optical signal with a spectral width of about 40 nm centered at 1.55 µm. This ASE then passed through an optical isolator, which prevented the two-pass superfluorescent source from becoming a laser. The output of the isolator was then sent through a single-mode fiber Fabry-Perot filter with finesse F = 80 and free spectral range FSR = 680 MHz. To boost the average power, the filter output was optically amplified by a final EDFA postamplifier. Average output powers of several milliwatts can be obtained using this arrangement. Fusion splices between the superfluorescent source, the filter and the postamplifier were used to minimize reflections. This is important because multiple reflections will add amplitude ripple to the intensity power spectrum. The photoreceiver consisted of a high-speed 14-µm-diameter, InGaAs p-i-n photodiode followed by a traveling wave GaAs microwave amplifier. A 50-GHz electrical spectrum analyzer (HP 8565E) displayed the photocurrent power spectrum.

For comparison purposes, three different techniques were used to measure the frequency response of the photoreceiver. The results of these three measurements are shown in Fig. 8. The curve labeled A was obtained using unfiltered ASE. For

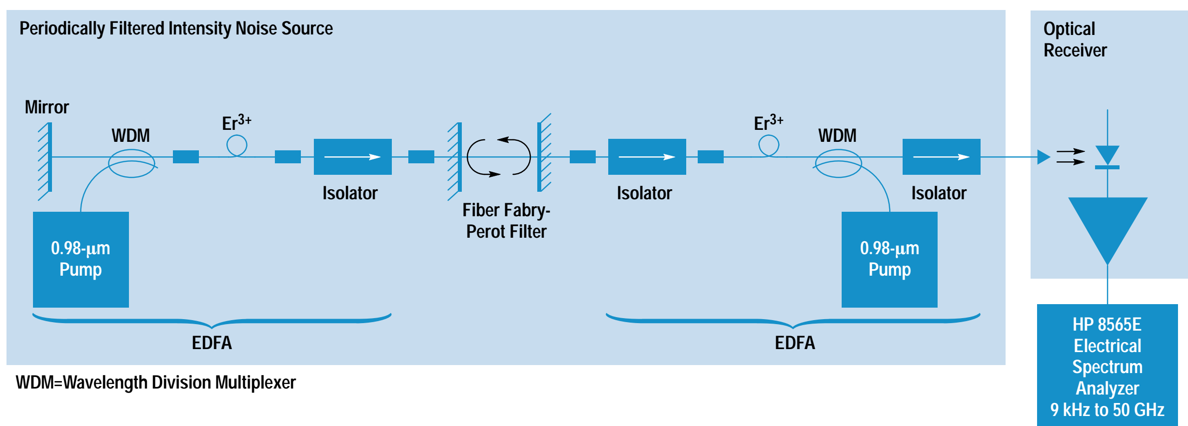


Fig. 7. Experimental setup used to demonstrate the periodically filtered intensity noise technique.

© Hewlett-Packard Company 1995

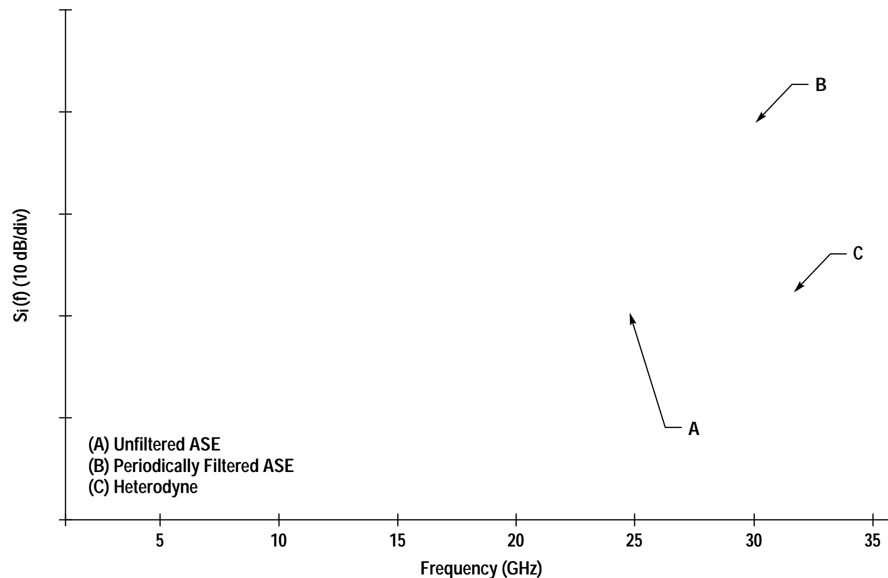


Fig. 8. Frequency response measurements of a high-frequency optical receiver using three techniques. The intensity noise curves were measured at 1.55 μm and the heterodyne measurement was made at 1.32 μm.

curve B, the periodically filtered ASE output was used. The average optical powers were held constant and equal for curves A and B. Curve C was generated using an optical heterodyne technique. Curves A and B experimentally demonstrate an SNR enhancement of approximately 17 dB between the filtered and unfiltered intensity noise techniques. Comparison between the heterodyne technique and the two intensity noise techniques shows very good agreement, illustrating the flat intensity noise spectrum obtained from the EDFA noise source.

Discussion

The use of a broadband intensity noise source for high-frequency detector calibration has several advantages over other frequency-domain techniques. One important advantage is that calibration of the frequency response of the optical source is not necessary since it can be made flat by choosing the ASE spectral width to be much larger than the frequency response of the photodetector. This is not the case for other frequency-domain techniques. For example, sinusoidal intensity modulation of an optical source using a $LiNbO_3$ modulator requires careful calibration of the frequency response of the modulator, which becomes increasingly difficult at high frequencies. Any errors in this calibration are passed on to the detector's frequency response.

Compared with heterodyne techniques, the intensity technique is rugged and field deployable and does not require stable polarization alignment. Since the intensity noise is present at all frequencies, it allows rapid measurements. Additionally, the long coherence length of laser sources in the presence of optical reflections makes the heterodyne measurement more susceptible to environmental effects.

The intensity noise method also has advantages compared to time-domain impulse measurement methods. For high-frequency measurements, the effects of the oscilloscope must be deconvolved from the measurement before an accurate calibration can be obtained. This can often be arduous since accurate impulse responses for high-frequency oscilloscopes are difficult to obtain. The advantage of not requiring careful source calibration for the intensity noise technique is an important consideration in developing a portable, low-cost, user-friendly calibration technique.

Summary

In this paper we have proposed and experimentally demonstrated methods for enhancing the dynamic range of the intensity noise technique for high-frequency photoreceiver calibration. By combining recently developed EDFA technology with spectral filtering techniques, we have shown that the magnitude of intensity beat noise can be increased to values practical for unamplified photodiode calibration. Using the periodically filtered technique, RIN values larger than −100 dB/Hz can be achieved over a frequency range in excess of 100 GHz. The intensity noise calibration technique has the potential for becoming a portable, easy to use, field deployable calibration method.

Acknowledgments

Mike McClendon and Chris Madden performed the optical heterodyne measurements and Mohammad Shakouri provided the 30-GHz photoreceiver. Steve Newton provided important support for this project.

References

1. D.J. McQuate, K.W. Chang, and C.J. Madden, "Calibration of Lightwave Detectors to 50 GHz," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 87-91.
2. S. Kawanishi, A. Takada, and M. Saruwatari, "Wideband Frequency-Response Measurement of Optical Receivers Using Optical Heterodyne Detection," *Journal of Lightwave Technology,* Vol. 7, no. 1, 1989, pp. 92-98.
3. E. Eichen, J. Schlafer, W. Rideout, and J. McCabe, "Wide-Bandwidth Receiver/Photodetector Frequency Response Measurements Using Amplified Spontaneous Emission from a Semiconductor Optical Amplifier," *Journal of Lightwave Technology,* Vol. 8, no. 6, 1990, pp. 912-916.
4. D.M. Baney, W.V. Sorin, and S.A. Newton, "High-Frequency Photodiode Characterization Using a Filtered Intensity Noise Technique," *IEEE Photonics Technology Letters,* Vol. 6, no. 10, October 1994, pp. 1258-1260.

© Hewlett-Packard Company 1995

Erbium-Doped Fiber Amplifier Test System

The HP 81600 Series 200 EDFA test system combines various instruments with powerful software to characterize erbium-doped fiber amplifiers. The system is a turnkey solution with fully specified uncertainty.

by Edgar Leckel, Jürgen Sang, Rolf Müller, Clemens Rück, and Christian Hentschel

Erbium-doped amplifiers (EDFAs) are the latest state-of-the-art solution for amplifying optical signals in lightwave transmission systems (see Fig. 1). They are used as booster amplifiers on the transmitter side to get as much power as possible into the link, as inline amplifiers to overcome the loss of the fiber, and as preamplifiers at the receiver end to boost signals to the necessary receiver levels. EDFAs can be used in single-wavelength transmission systems, in wavelength division multiplexed (WDM) systems, and in soliton transmission systems. To use EDFAs in the various applications it is necessary to characterize the single amplifier as a component.

Parameters and Measurement Techniques

The main parameters describing an EDFA are signal output power, total output power, gain, and noise figure.[1] All of these parameters are dependent on input power level and wavelength. To characterize these amplifiers fully it is necessary to measure their dependence on both input power and wavelength (see Fig. 2).

Signal output power is measured with an optical spectrum analyzer and the total output power is measured with a power meter. The gain is the ratio of the signal output power of the amplifier to the signal input power.

Noise figure is defined as the ratio of the signal-to-noise ratio at the input to the signal-to-noise ratio at the output of the amplifier under the following conditions: shot-noise-limited photodetector, shot-noise-limited input signal, and optical bandwidth approaching zero. The main problem in measuring the noise figure is that sources used for generating variable input power and wavelength also generate a broad LED-like spectrum called source spontaneous emission (SSE). The SSE is amplified and adds to the output power. The amplifier output consists of amplified signal and amplified spontaneous emission (ASE). To measure only the signal and ASE contribution from the amplifier we have to eliminate the contribution from the SSE.

There are two principal methods[2] for measuring the exact noise level of the amplifier. The first is called the amplified spontaneous emission interpolation subtraction method. In this method the SSE level of the laser source is determined during the calibration and stored in a calibration file. With this calibration and the measured gain, the SSE × gain

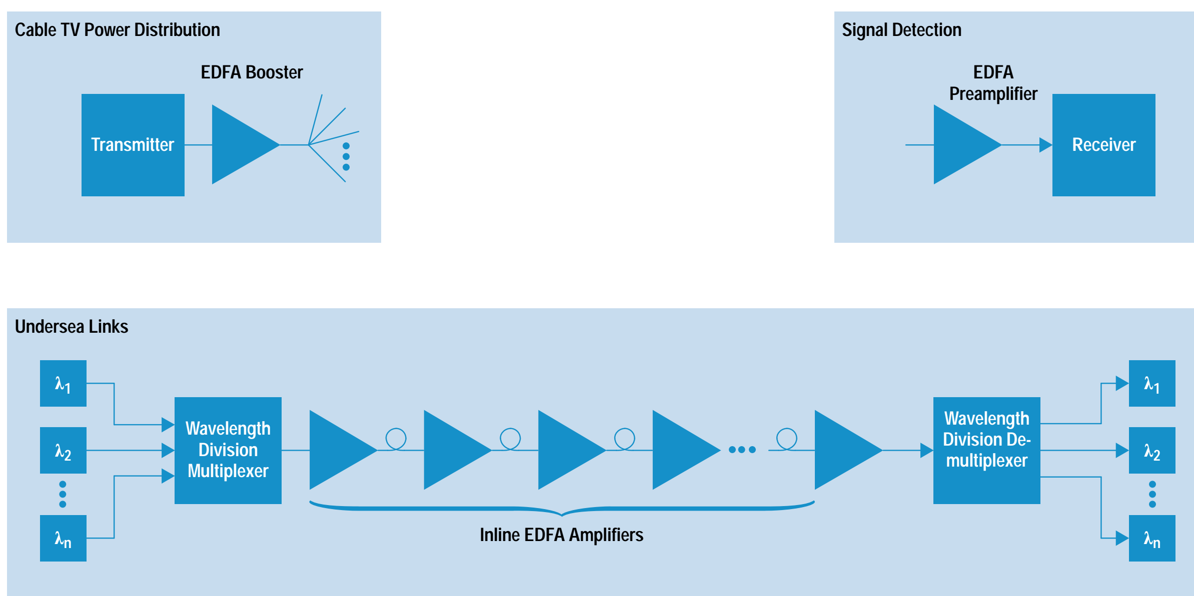


Fig. 1. Erbium-doped fiber amplifier (EDFA) applications in communication systems.

© Hewlett-Packard Company 1995

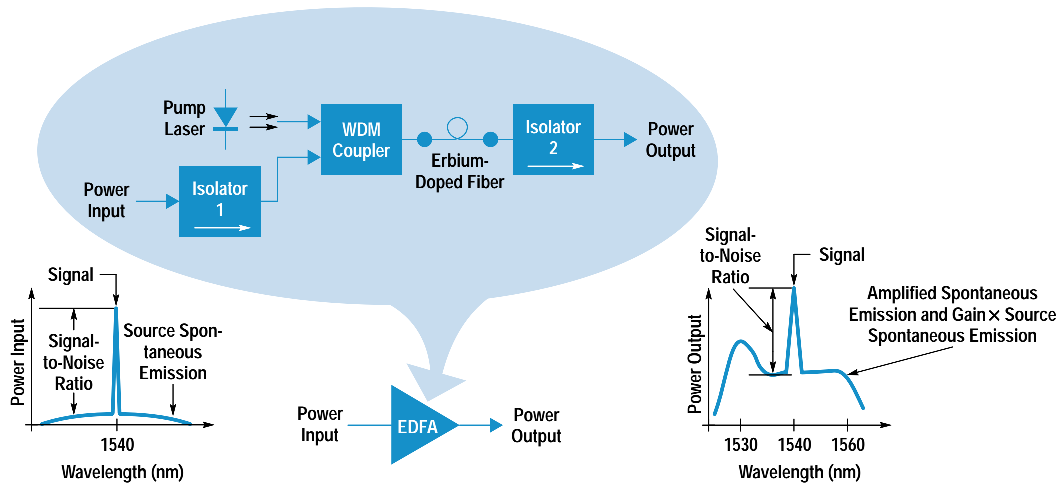


Fig. 2. Key components, optical signals, and measurement parameters for EDFA characterization.

contribution can be subtracted from the total spontaneous power level to obtain the ASE level of the amplifier itself.

The second method is called the polarization extinction method. It depends on the fact that the signal and the SSE of the laser source have the same state of polarization because there is a polarizer at the output of the tunable laser signal source. The amplifier's ASE is unpolarized. This makes it possible to extinguish the amplified SSE contribution by blocking the signal and therefore also the SSE contribution after the amplifier. The extinction is accomplished with a polarization controller/filter.

EDFA Test System

The HP 81600 Series 200 EDFA test system is shown in Fig. 3. Fig. 4 is its block diagram. The tunable laser source with built-in attenuator[3] provides the input power levels over the required wavelength range. To guarantee the absolute power level at the input of the EDFA, the power is monitored by means of a coupler and a calibrated power meter.[4] At the output of the EDFA the total output power is measured with a 5% tap and a power meter. Because of the high output power of the EDFA it is necessary to insert attenuators in front of the power meter heads. The coupler for the power

Fig. 3. HP 81600 Series 200 EDFA test system.

© Hewlett-Packard Company 1995

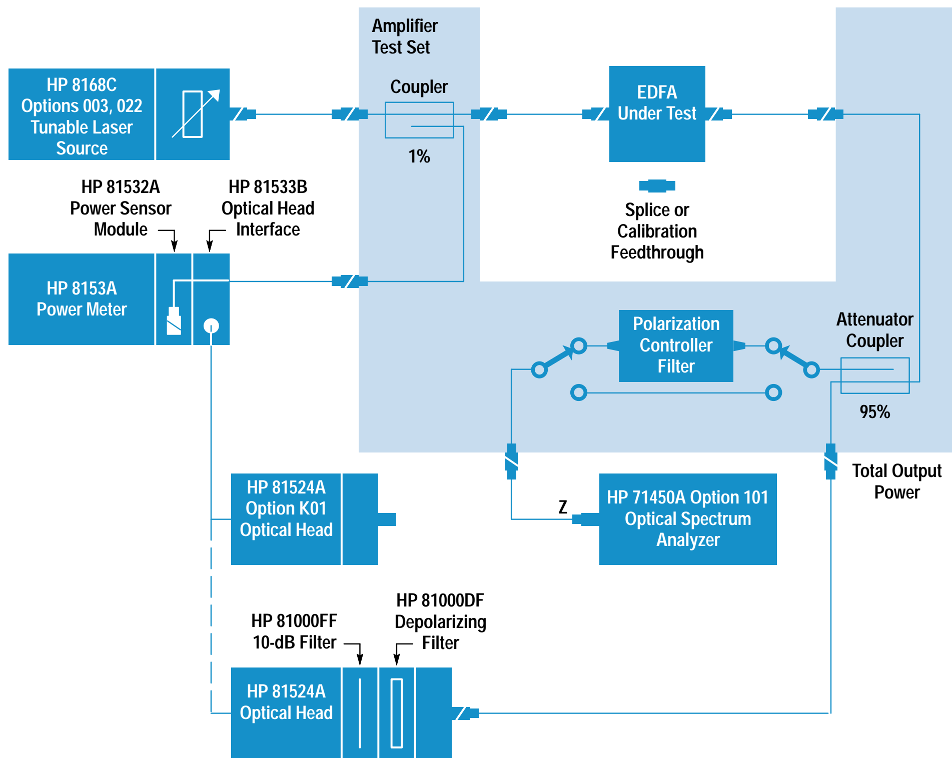


Fig. 4. Block diagram of the HP 81600 Series 200 EDFA test system.

meter acts as an attenuator for the optical spectrum analyzer. The coupler is followed by a switch and the polarization controller/filter arrangement. This makes it possible to measure the signal directly or via the polarization controller/filter with the polarization extinction method. The optical spectrum analyzer acts as a wavelength-selective power measurement device.

Amplifier Test Set

A specially developed instrument for this test system is the amplifier test set. The amplifier test set consists of couplers for monitoring the input and output power of the EDFA and a polarization controller/filter. The switches are used to select between gain measurement on the straight path and ASE measurement on the polarization controller/filter path.

The optical design of the polarization controller/filter is based on two retardation plates—one quarter-wave and one half-wave plate—and a linear dichroic polarizer (see Fig. 5). These parts are mounted on rotatable hollow shafts so that the collimated light beam can pass through the shafts and through the center of the optical codewheel. The whole assembly—optical parts and encoder—is driven by a dc motor coupled with a belt gear drive. With this design any incoming state of polarization can be transformed into any other state by rotating the retardation plates to defined angular positions. A polarizer is added at the output so that linear states of polarization can be extinguished.

In the polarization extinction method the polarization controller/filter is used as a polarization analyzer to determine the input state of polarization. Based on this measurement, the retardation plates are rotated to calculated angular positions, which change the signal and the amplified SSE to a linear state perpendicular to the pass direction of the polarizer. This causes the signal and the amplified SSE to be extinguished.

The amplifier test set contributes to the overall system uncertainty. Therefore, the couplers, switches, and polarization

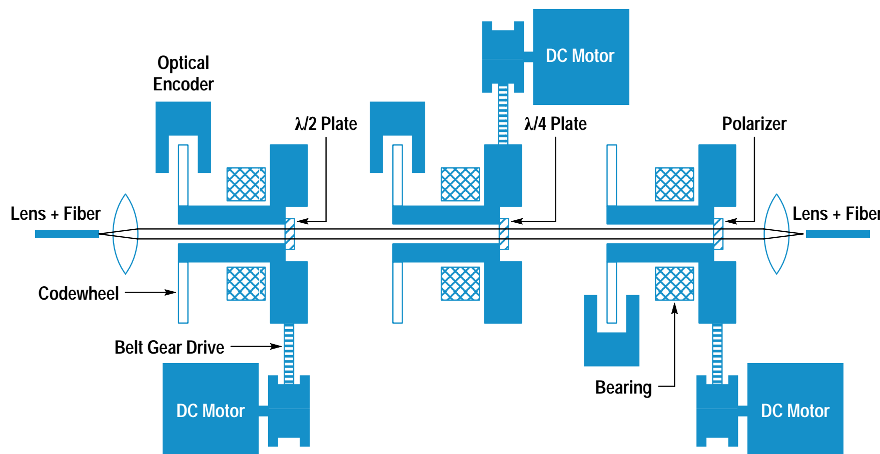


Fig. 5. Polarization controller/filter.

controller/filter were designed or selected for lowest polarization dependent loss. In addition, the switches were selected for very good repeatability and high return loss. The polarization controller/filter was also designed for low rotation dependent loss. Finally, an intelligent algorithm and optimized speed of the polarization controller/filter reduce the total measurement time.

Before starting a measurement, it is necessary to run a calibration. The first calibration step is to measure the coupling ratio of the coupler before the DUT with the help of the two power meter heads. The second step is to calibrate the optical spectrum analyzer and the power meter in conjunction with the attenuators and the losses of the paths. A feedthrough is used to connect the test system's input and output ports. The calibration is verified by measuring the feedthrough. In this case the gain is well-known (0 dB) and the output power must be the same as the input power.

EDFA Test System Uncertainty

Major efforts went into understanding, characterizing and, wherever possible, correcting the actual and potential sources of error in the HP 81600 Series 200 EDFA test system. This task is complex because the test system and measurement tasks are complex. All measurement tasks start with the calibration of the system, in which optical power traceability to PTB and NIST is ensured.[4] Optical power traceability is important in conjunction with the signal power, total power, and ASE measurements, the latter being used to calculated the noise figure. For the gain measurement, it is important that the power scale be linear. This parameter is also traceable to PTB through a chain of scale comparisons. Finally, wavelength traceability is ensured through comparisons to specific gas lamps and lasers, which can be considered natural physical constants.

A careful uncertainty analysis was carried out for each of four parameters: signal power, total power, gain, and noise figure. In each case, the entire process of calculation was analyzed. For example, the gain is defined as the ratio of signal output power to signal input power. Signal input power is measured by means of the input coupler and the HP 8153A optical power meter with the HP 81532A power sensor module. This measurement relies on the accuracy of the input calibration and the performance of the equipment involved. Signal output power is measured with the amplifier test set and an optical spectrum analyzer. This measurement relies on the accuracy of the output calibration and on the performance of the equipment.

As an example, the following represents a summary of the uncertainty analysis for the determination of noise figure. This analysis is the most complicated because the noise figure F is a function of the ASE power density ρ_{ase} and the gain G:

$$F = \frac{1}{G} + \frac{\rho_{ase}}{h\nu G}.$$

where h is Planck's constant and ν is the optical frequency. Analyzing this equation for the origin of the numeric values and the performance of the instruments involved leads to the following list of partial uncertainties for the noise figure:

- Polarization dependence of the amplifier test set and the optical spectrum analyzer
- Small errors in the optical spectrum analyzer scale fidelity
- Drift effects in the amplifier test set
- Uncertainty attributable to the input connector pair (the output connector pair cancels out because it influences the gain and the ASE in the same way)
- Calibration uncertainty of the HP 81524A optical head
- Finite accuracy of the cancellation of the source's spontaneous emission in determining the ASE level
- Loss uncertainty of the path through the polarization controller
- Uncertainty of the optical spectrum analyzer's resolution bandwidth
- Uncertainty of the input power measurement resulting from the polarization dependence of the input coupler and the HP 81532A power sensor module.

Table I shows a summary of the uncertainty analysis for the HP 81600 Series 200 EDFA test system in conjunction with the polarization extinction technique. Also shown are the total uncertainties, which are obtained by root-sum-squaring. They serve as the basis for the test system specifications.

Table I
Uncertainty Summary for HP 81600 Series 200 EDFA Test System

Uncertainty		Output Signal (dB)	Gain (dB)	Noise Figure (dB)
Gain Compression by SSE		±0.10	±0.10	
Optical Spectrum Analyzer Polarization Dependence		±0.10	±0.10	±0.05
Test Set Polarization Dependence		±0.07	±0.07	±0.07
Optical Spectrum Analyzer Scale Fidelity		±0.05	±0.10	±0.10
Drift of Test Set		±0.10	±0.10	±0.10
Output Connector Pair		±0.25	±0.25	
Input Connector Pair			±0.25	±0.25
Power Meter Scale		±0.11	±0.11	±0.10
Input Power			±0.05	±0.10
SSE Cancellation				±0.10
Polarization Controller/Filter Path Loss				±0.07
Optical Spectrum Analyzer Resolution Bandwidth				±0.10
Total Uncertainty:	Connectors	±0.39	±0.48	±0.39
	Splices	±0.27	±0.24	±0.25

Software

Testing an EDFA device in production requires software that is easy to use, since EDFA measurement methods are quite new and the software will be used by people who may not know all of the details of the test process. On the other hand, software flexibility was an important design goal to allow extensions for additional measurement methods and integration into existing customer processes and databases.

© Hewlett-Packard Company 1995

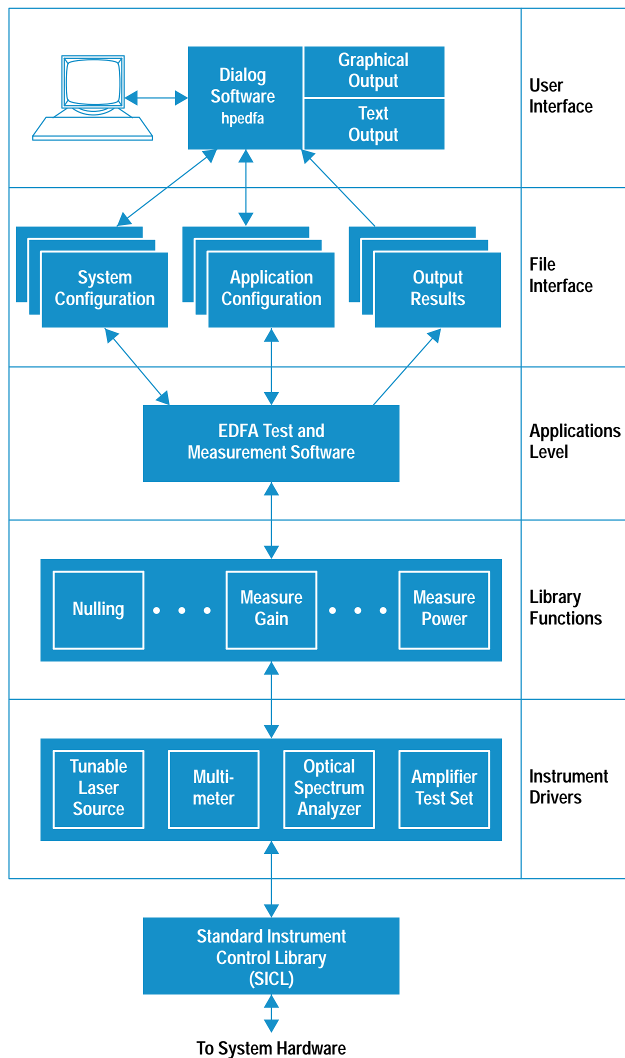


Fig. 6. EDFA test system software overview.

The EDFA test system software runs on an HP 745i workstation under the HP-UX* 9.01 operating system, and is implemented mostly in the C programming language. Fig. 6 shows an overview of the software structure. On the lower level, there is a set of drivers for the different instruments, along with some modules for accessing configuration information and help texts. They supply an easy and portable access to the instrument hardware shown in Fig. 2.

The communication layer for the instrument drivers is based on the SICL (Standard Instrument Control Library). Using the SICL, data can be written to and read from instruments with commands similar to those for reading and writing files in C. All calls to the SICL are handled through a control module called inst_drv (Fig. 7). Its purpose is to allow the logging of traffic going to and from the instruments and to support some basic functionality like detecting the presence of an instrument on the HP-IB (IEEE 488, IEC 625). It also supplies central error handling facilities.

For each instrument type there is a separate driver module. Each module can control several instruments of the same type at once, and keeps track of the internal instrument states to save execution time. The driver modules check parameters before passing them to the instruments, and support instrument error checking.

To speed up measurements where multiple instruments are involved, the drivers act asynchronously. For example, both channels of the HP 8153A optical power meter can fetch data while the HP 70950 optical spectrum analyzer is making a swept measurement. Additional commands are available to synchronize the instruments and make sure everything is settled before a measurement is taken.

The measurement application is built on top of this driver structure. It contains different algorithms for testing EDFAs, one of which is active at a time. Currently, the following measurement functions are integrated:

- Perform an input calibration of the test system
- Perform an output calibration of the test system
- Verify the calibration
- Measure an EDFA using the polarization extinction method
- Measure an EDFA using the amplified spontaneous emission method
- Perform a self-test.

While the measurement proceeds, intermediate results are checked against limits set up for the device by the operator, and against certain fixed system limits (to detect improper connections, etc.).

A C language interface allows customized extensions, such as controlling additional parameters (e.g., the EDFA pump current) or performing additional measurements (e.g., reading a voltmeter).

All results of the calibration and measurement steps are stored in a single file, along with information about the EDFA device, the test conditions, and the overall result (pass or fail). This file is organized in a structured ASCII format and allows easy extraction of all important test results for use in analysis programs and databases.

For user-friendly parameter input and measurement control, a sophisticated user interface was designed. Written with a

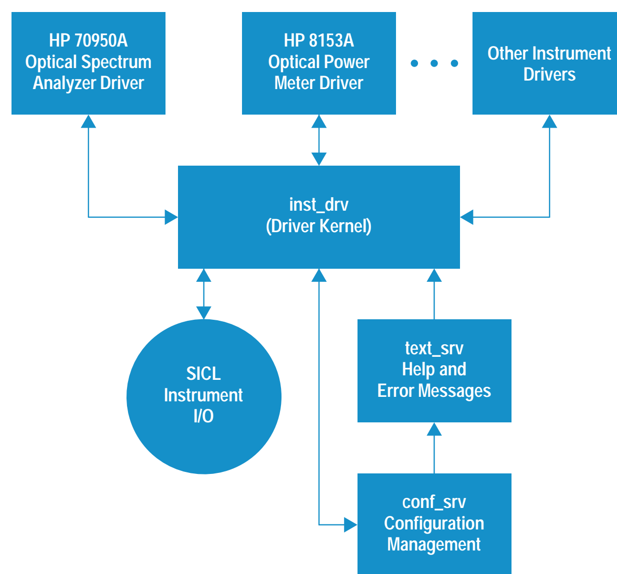


Fig. 7. Instrument driver overview.

© Hewlett-Packard Company 1995

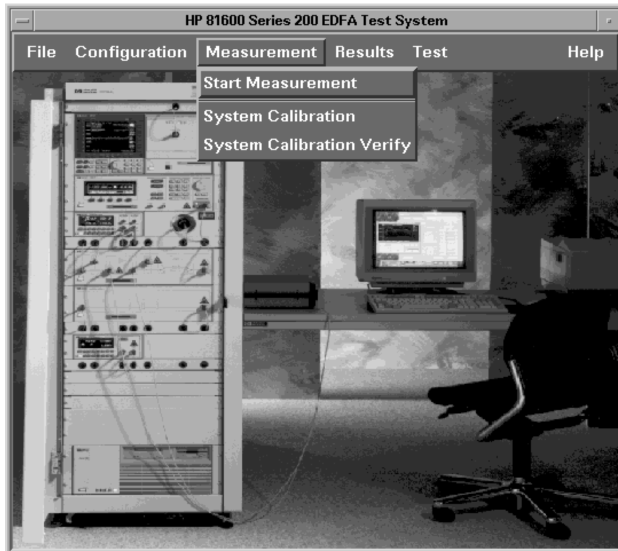


Fig. 8. Main window of the EDFA test system user interface.

user interface builder, along with some C code, it is based on X11/Motif (Fig. 8).

The user is guided through the setup and calibration of the test system by easy-to-use menus and dialog boxes. Different calibration setups for specific wavelength ranges can be entered, as well as test conditions with limits appropriate for certain EDFAs (Fig. 9).

Where fiber connections have to be made in the calibration and measurement process, a graphic is brought up on the screen showing the necessary steps and connections. All

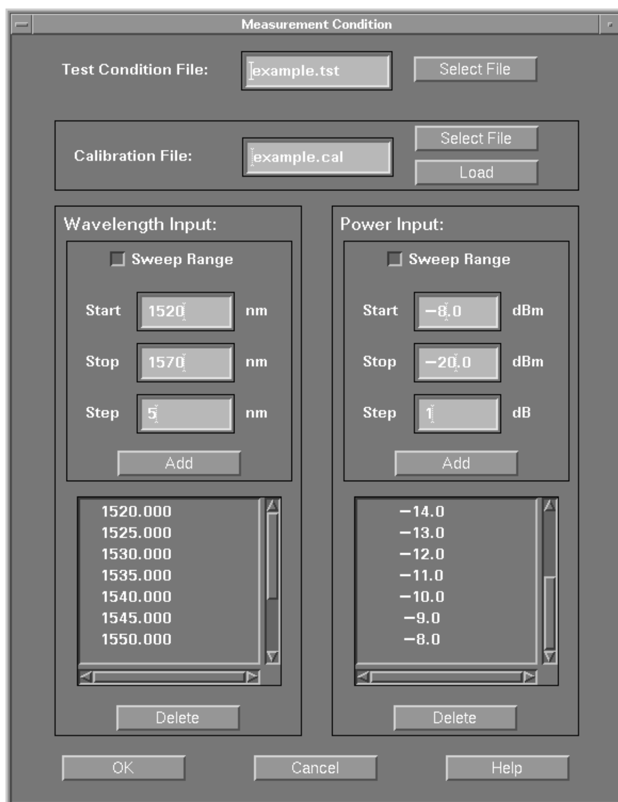


Fig. 9. Test condition input panel.

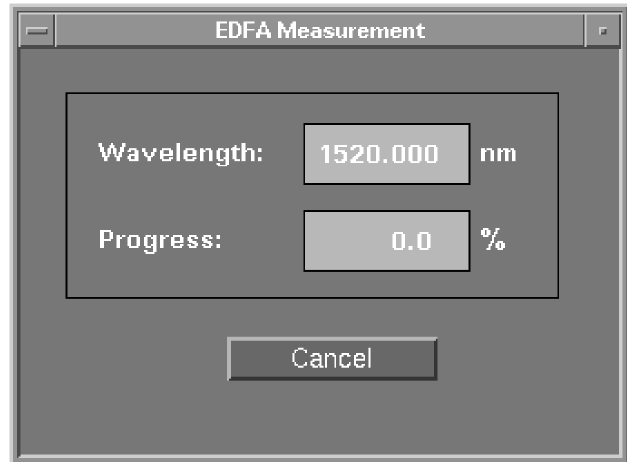


Fig. 10. Progress status window.

numeric data is checked for correctness immediately after it is entered. Before a measurement starts, the calibration is matched against the needs of the requested test.

The progress of a running calibration or test is shown with an information box on the screen (Fig. 10). Meanwhile, the program stays fully usable for entering new test setup data, or for reviewing the results of previous measurements.

Result data can be examined in textual form. A table showing input parameters and results can be viewed and printed. A graphical output system is included so that the user can review the data graphically and create graphs for EDFA device documentation.

A menu (Fig. 11) allows the user to choose an X-Y plot of any result parameter against any input parameter (including user-specified parameters). Fig. 12 shows such a plot. A single graph can have up to eight traces. The graph can be interactively scaled and zoomed, and markers can be placed either freely or bound to measurement points. The X-Y values of each measurement point can be read out with a single mouse click.

The graphics system also provides hard-copy functionality and allows the merging of different graphs on a specified number of pages, thus enabling the output of a user-defined data sheet which can be printed out automatically after each test is completed. Supported printers are the HP LaserJet series and the DeskJet series, the latter allowing color printouts.

Acknowledgments

We wish to thank Bernd Maisenbacher for project management and Emmerich Müller for help in the development of the polarization controller and measurement technique. We also want to thank Robert Jahn who was responsible for implementing the measurement routines and coordinating all software related issues. In addition we also thank Wolfgang Reichert for the mechanical design of the polarization controller, amplifier test set, and rack. Wolfgang was also responsible for coordinating the development of the amplifier test set. Also, we want to thank our production engineers Jürgen Mang and Reinhard Becker for writing test software. Last not least, Doug Baney of HP Laboratories provided the scientific support and background. Finally we want to thank

© Hewlett-Packard Company 1995

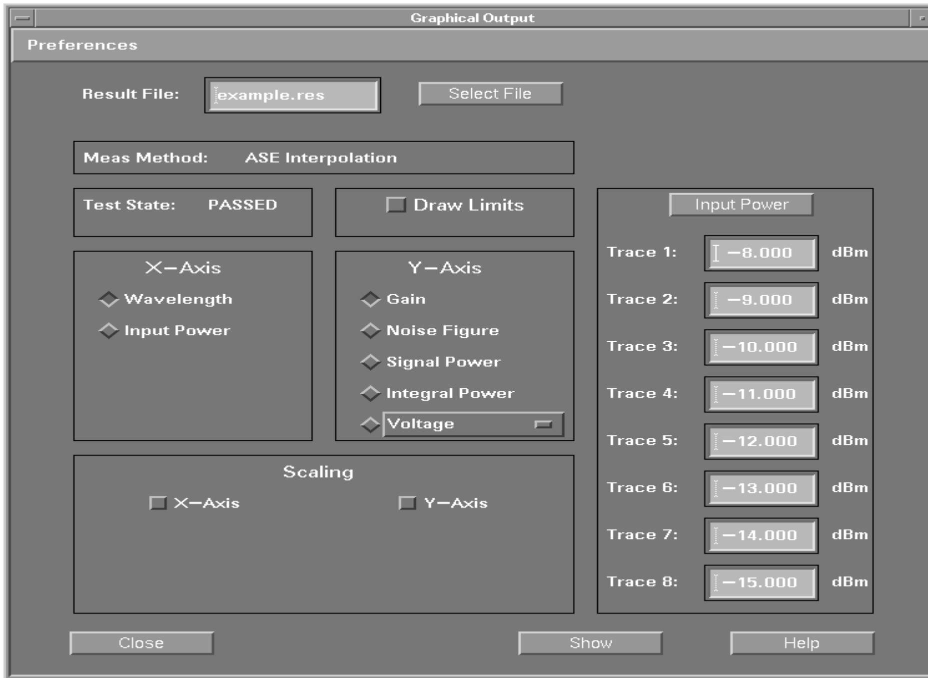


Fig. 11. Graphical output window.

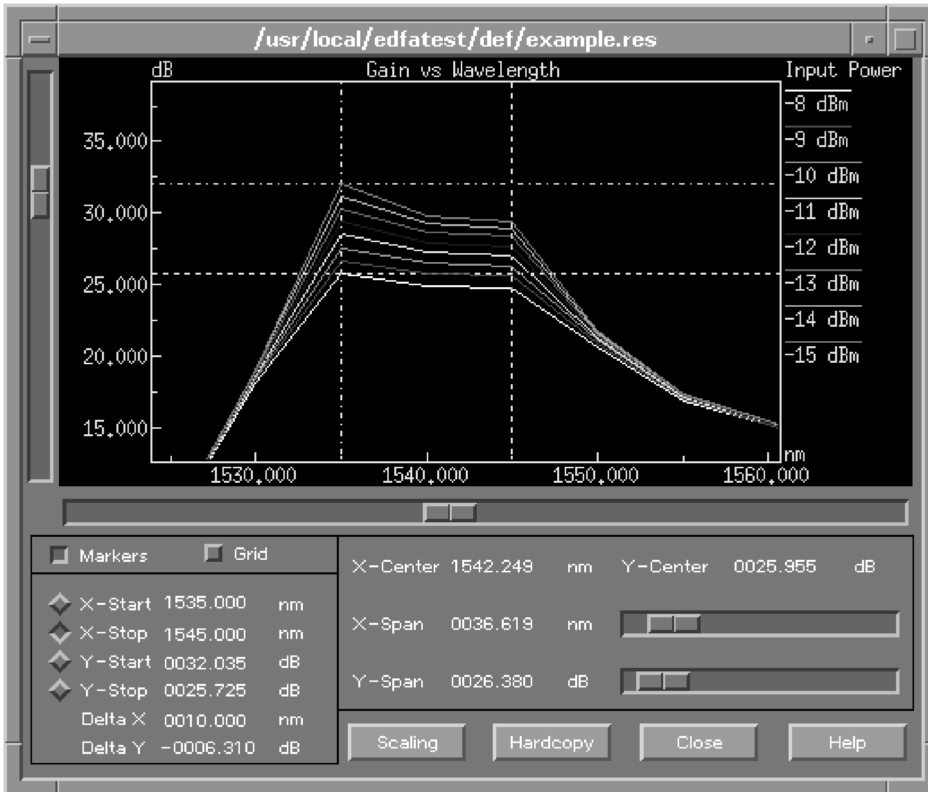


Fig. 12. X-Y parameter plot with markers.

Jack Dupre, Jim Stimple, Zoltan Azary, and Dave Baily at the Lightwave Operation for their support on optical spectrum analyzer related issues.

References

1. D. Baney, C. Hentschel, and J. Dupre, "Optical Fiber Amplifiers—Measurement of Gain and Noise Figure," *Hewlett-Packard Lightwave Symposium,* 1993.

2. C. Hentschel, E. Müller, and E. Leckel, "EDFA Noise Figure Measurements—Comparison between Optical and Electrical Techniques," *Hewlett-Packard Lightwave Symposium,* 1994.

3. *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 11-38.

4. *Hewlett-Packard Journal,* Vol. 42, no. 1, February 1991, pp. 58-83.

HP-UX is based on and is compatible with Novell's UNIX® operating system. It also complies with X/Open's* XPG4, POSIX 1003.1, 1003.2, FIPS 151-1, and SVID2 interface specifications.

UNIX is a registered trademark in the United States and other countries, licensed exclusively through X/Open Company Limited.

X/Open is a trademark of X/Open Company Limited in the UK and other countries.

Motif is a trademark of the Open Software Foundation in the U.S.A. and other countries.

© Hewlett-Packard Company 1995

Multi-Quantum-Well Ridge Waveguide Lasers for Tunable External-Cavity Sources

A new multi-quantum-well ridge waveguide laser enhanced for use in a grating-tuned external-cavity source has been developed. The device offers higher output power and wider tunability for improved performance in a new instrument. A core technology has been developed for use in a variety of light-emitting devices.

by Tirumala R. Ranganath, Michael J. Ludowise, Patricia A. Beck, Dennis J. Derickson, William H. Perez, Tim L. Bagwell, and David M. Braun

Tunable laser sources for testing of optical components and subsystems are an important part of a family of lightwave communication test and measurement instruments that HP has developed over the last decade. Currently, HP offers two tunable laser sources: the HP 8167A and HP 8168A.[1] These instruments function in the wavelength windows centered at 1300 nm and 1550 nm.

Custom requirements of test instruments cannot always be met by commercially available lasers. As a result, semiconductor laser development was begun to provide core material and device technologies that would produce suitable optical gain media chips for future tunable laser sources. The laser described in this article is the optical gain medium for the HP 8168C tunable laser source. The same technology forms the basis for other custom optical sources, such as the edge-emitting LEDs discussed in the article on page 43.

There are many different aspects to the development of a custom laser chip: device design, material growth, fabrication, testing, and reliability. Wide tunability and high output power are two very important requirements for the instrument application. The device must also be capable of operation in an external-cavity laser source like the HP 8168C.

A semiconductor laser is an optical oscillator, and like any oscillator, it consists of an optical amplifier (realized in a suitable semiconductor material system) together with feedback (provided by two atomically parallel cleaved facets). There are two classes of semiconductor lasers: gain guided and index guided. They are distinguished by the waveguiding mechanism for the optical field. Gain guided devices are relatively simple to fabricate but have high threshold current and astigmatism of the output beam. Index guided lasers can be more challenging to fabricate but offer low threshold currents, high linearity of output power, and good beam quality.

There are a number of variants of the index guided laser. The most sophisticated involve multiple regrowths of epitaxial material around the waveguide. A ridge waveguide laser,[2] while still index guided, has no regrowth steps, moderate threshold currents, excellent beam quality, and potentially high reliability. Ridge structures do not exhibit the best linearity of output power with current drive (as do the best buried structures) and therefore are not suitable for applications such as CATV. But for external-cavity sources, ridge waveguide lasers provide excellent performance.

The amplifying region in the ridge structure allows a single optical mode to propagate while a suitable electron-hole population distribution is maintained by an electrical current. Details such as material composition, layer thickness, and ridge dimensions impact the threshold current, differential quantum efficiency, and far-field divergence angles of the light emanating from the two device facets. The far-field divergence angles determine how efficiently the device can couple to a hybrid grating-tuned external cavity on one side and to a single-mode fiber output on the other. The device threshold current is the point at which the amplifier gain equals the total cavity losses. Useful optical power output is obtained only for currents greater than threshold. Differential quantum efficiency is a measure of the device's ability to convert electrical energy to coherent optical output power, once the injected current exceeds threshold.

Tunability is intimately tied to the electronic band structure of the semiconductor as well as our ability to grow a device's light-emitting region by means of a vapor phase technique known as organometallic vapor phase epitaxy (OMVPE). To operate in an external cavity, a device also needs an antireflection coating. Devices must be reliable, so time-consuming reliability studies are done. Groups of devices are stressed at high temperatures, high currents, or both for many thousands of hours. Also, devices must be tested in chip form and in subassembly module form.[3] To anticipate instrument operation a laboratory grating-tuned external cavity is used. This allows a quick look at tunability and output power for a given device.

Semiconductor Laser Device

In Fig. 1 we show a schematic cross section of a ridge waveguide laser. Essential features of our device are a multi-quantum-well active region surrounded by n-type and p-type

© Hewlett-Packard Company 1995

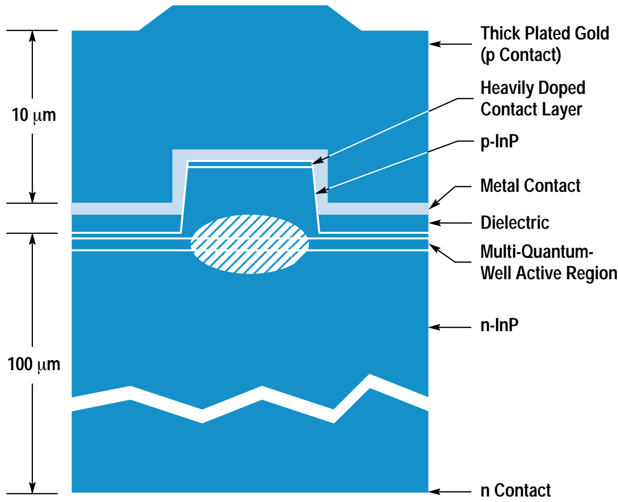


Fig. 1. Ridge waveguide laser cross section.

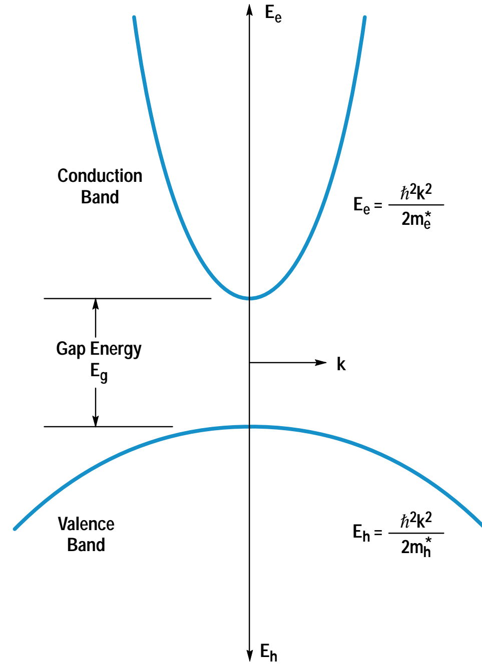


Fig. 2. Parabolic energy E versus crystal momentum k diagram (simplified) for a direct bandgap semiconductor having an energy gap E_g. The difference in curvatures reflects the difference in carrier effective masses for electrons (m_e^*) and holes (m_h^*).

InP cladding regions. The upper part of the device has an etched ridge whose function it is to confine light laterally, forming a single-mode optical waveguide. With a p ohmic metal contact on the ridge top and an insulator on either side, current is confined to a specific region. The completed device has front and back facets cleaved to form atomically flat and parallel end mirrors. The two end mirrors together with the optical waveguide form an optical cavity. The ridge is typically a few micrometers wide and a few micrometers high, while the completed chip is on the order of 100 micrometers thick and a few hundred micrometers on each side. The p contact has a thick gold pad plated to act as a heat spreader in addition to forming a bond pad. An n contact on the backside of the chip completes the device.

Active Region Design

Absorption, Stimulated Emission, and Optical Gain. In a direct bandgap semiconductor (say, the active region of a laser), an electron from a filled valence band state can absorb an incident photon (of energy $> E_g$, where E_g is the bandgap energy). In the process the electron gets transferred to a conduction band state, leaving behind a vacancy (hole). The absorption rate depends on the incident photon flux and the density of available conduction band states. In the reverse process, stimulated emission, an incoming photon of energy E_p stimulates an electron-hole pair to emit a second in-phase photon (of energy E_p). The stimulated emission rate depends on the incident photon flux, the density of electrons in the conduction band, and the density of holes (at the correct energy) in the valence band. A quantum-mechanical matrix element that measures the strength of the optical coupling between hole states (in the valence band) and electron states (in the conduction band) enters into the calculation for optical absorption and emission states.

It is this stimulated emission process, under conditions of electrical pumping (with a sufficient density of electrons and holes in the conduction and valence band states) that can provide optical gain. When the stimulated emission rate exceeds the (stimulated) absorption rate there is net optical gain. This is generally possible only under high electrical pumping conditions. In addition, an electron from the conduction band can spontaneously emit a photon by recombining with a hole from a valence band state. This process does not depend on the presence of an incident photon and gives rise to optical noise in laser devices.

Electron/Hole Density of States and Tunability. To gain an appreciation for how optical gain over a broad window arises, a brief examination of electron/hole states in a direct bandgap semiconductor is useful. In a bulk direct bandgap semiconductor, electron and hole states generally exhibit a parabolic relation between energy E and the crystal momentum k (see Fig. 2). In the conduction band,

$$E_e = \frac{\hbar^2 k^2}{2m_e^*},$$

where m_e^* is the electron effective mass, and in the valence band,

$$E_h = \frac{\hbar^2 k^2}{2m_h^*},$$

where m_h^* is the hole effective mass. The electron effective mass is about a factor of 10 smaller than the hole effective mass and this is reflected as a factor of 10 difference in the curvatures of the parabolas in the conduction and valence bands. The three-dimensional density of allowed electron and hole states per unit energy is proportional to $\sqrt{E}$ for a bulk semiconductor. For the case of a very thin ($L_z \approx 10$ nm) layer of the same bulk semiconductor sandwiched between two barrier regions of a higher bandgap material, a structure known as a quantum well, the electron and hole motions are restricted and the allowed energies for motion normal to the layer are quantized and occur at discrete energies (say E_e^1, E_e^2, ...) determined by the dimension L_z. Motion in the plane of the film is not restricted and allowed energy states form a continuum. The result is a constant density of states for both

© Hewlett-Packard Company 1995

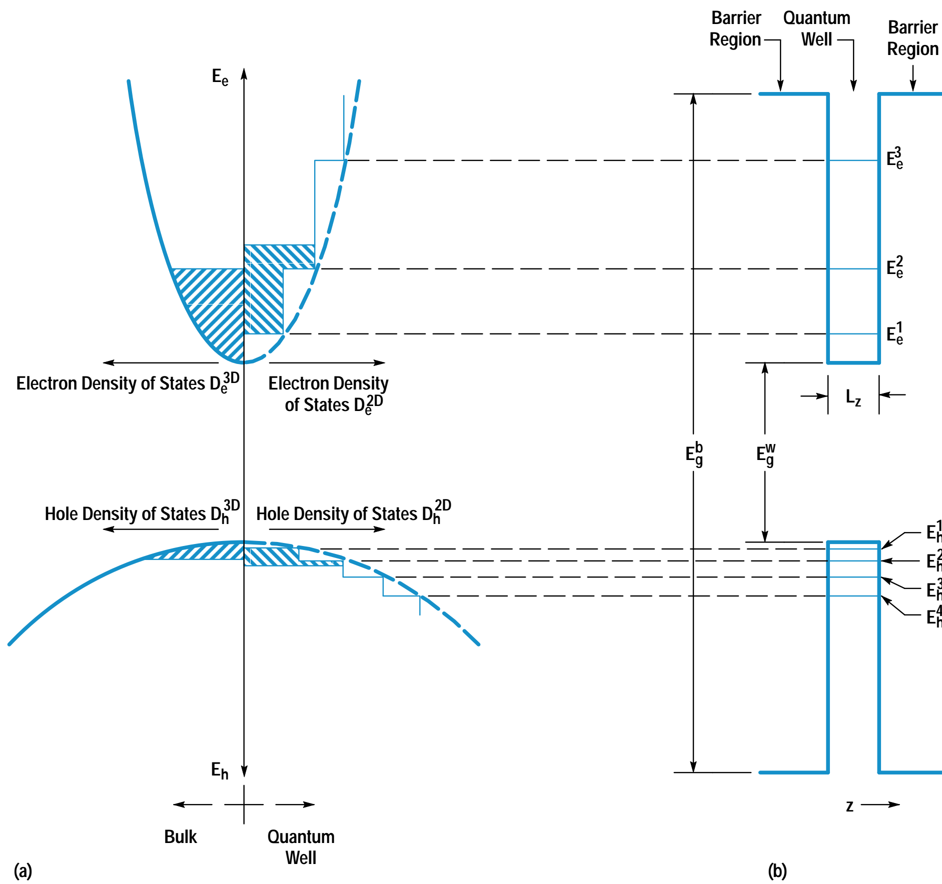


Fig. 3. (a) Parabolic and step density of states for bulk semiconductor (left of line) and quantum well (right of line). For a given volume density of carriers, electrons fill up to a greater depth for the quantum well case (right of line). (b) Discrete electron and hole energies from restricted z motion in a quantum well of thickness L_z. We have omitted the presence of a degenerate light hole band for this illustration.

electrons and holes.[4] Fig. 3 illustrates these observations for both bulk and quantum well semiconductor structures. Fig.3a shows (hatched regions) the same volume density of electrons (and holes) distributed in energy space for bulk (left of line) and quantum well (right of line) structures. The electrons fill up the available states and reach a higher energy for a quantum well than for a bulk active layer. It is this high energy state occupancy feature that allows a broad wavelength range of gain using quantum well active regions. Quantum well composition (which fixes the gap energy in the well region, E_g^w) and thickness (L_z), together with barrier region composition (which fixes the gap energy in the barrier region, E_g^b) determine the positions of the discrete levels E_e^1, E_e^2, ... (see Fig. 3b). Level E_e^1 determines the long-wavelength end of the gain spectrum. The short-wavelength (or higher-energy) end of the gain spectrum is determined by the loss of electron confinement in the quantum wells. This loss of electron confinement is primarily determined by the barrier energy gap E_g^b. The upshot is that one needs to pick carefully the barrier material (for E_g^b), the quantum well material (for E_g^w), and the thickness L_z to ensure a properly wavelength-centered and sufficiently broad gain spectrum to fit the instrument need. In practice, the modal gain provided by a single quantum well is quite modest, so multiple quantum wells are used in parallel to achieve the requisite gain.

OMVPE Growth

Organometallic vapor phase epitaxy (OMVPE) dominates contemporary production of the quaternary compound semiconductor GaInAsP because of its relative simplicity, high yields, and scalability to large-area production. The GaInAsP material system has been studied since the early seventies when it was realized that compositions lattice-matched to InP substrates would produce devices important for the then-incipient fiber-optic communication systems. OMVPE uses group III organometallic compounds such as triethylgallium (TEGa) or trimethylindium (TMIn, both metal alkyls) and hydrides of the group V elements, such as AsH_3 and PH_3, as source materials for the vapor deposition. Vapors of the metal alkyls are entrained in a carrier gas, usually hydrogen, and transported to the heated substrate. The organometallics readily crack into single metal atoms and ethane or methane chains. The metal deposits on the growing surface, and the hydrocarbon chains are swept away in the carrier stream. Similarly, the hydrides also crack thermally, depositing only the group V atoms. A nearly perfect 1:1 stoichiometry of group V to group III atoms is achieved in the solid because the crystal strongly rejects excess group V atoms, which are volatile enough to evaporate.

The deposition zone design must incorporate several critical requirements. The substrate temperature must be uniform and tightly controlled since the As/P solid ratio is highly temperature-sensitive. The fluid dynamics must be controlled to deliver uniform amounts of precursor chemical to all parts of the substrate, and to compensate for any downstream depletion. Finally, turbulent flow streams defeat abrupt gas composition changes injected by the precision gas switching manifold and must be designed out. The laser diodes for the tunable source require precise control of the composition, thickness, and interfaces of nearly 20 epitaxial layers. The ratios of the four main gas constituents control the epitaxial GaInAsP composition, and the total metal alkyl concentration determines the epitaxial growth rate. Abrupt changes in

© Hewlett-Packard Company 1995

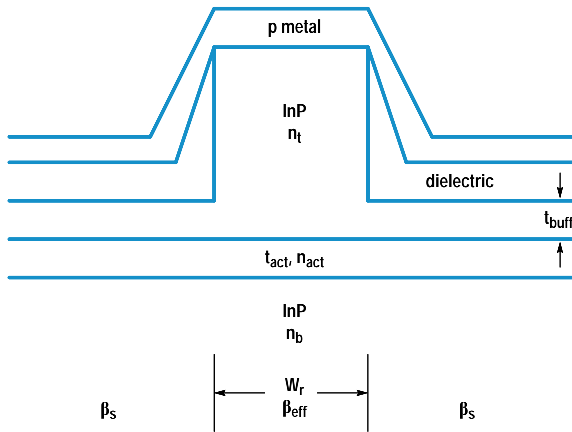


Fig. 4. Ridge waveguide laser design issues. β_{eff} and β_s are the optical propagation constants for equivalent slab waveguides whose layers are the same as in the ridge and side regions of the device, respectively.

composition are achieved by rapidly switching the gas composition. Slow variations in the metal alkyl vapor ratios and hydride gas flow produce gradually graded material. In all cases, the epitaxial layer must remain lattice-matched to the InP substrate to within 2 parts in 10^4 to suppress dislocation formation. This requires precise control over the alkyl temperature (≤ 0.1°C variation), carrier gas flow, and the system pressure at several points.

The use of TMIn introduces another control complication. Unlike other commonly used metal alkyls, which are liquid over practical temperature ranges, TMIn is a solid. The concentration of the TMIn vapor sublimed into the gas stream varies (often unpredictably) with the solid surface area and the source lifetime. The TMIn vapor concentration in the hydrogen carrier gas must be constantly measured and controlled throughout the deposition cycle. By timing the transit of an ultrasound pulse through a real-time sample of the TMIn-hydrogen mixture, TMIn concentration to within 1 part in 10^5 can be determined.

Finally, small interfacial composition transients generate massive dislocation networks that destroy device performance. By controlling the temperatures, pressures, flows, and gas switching timings, essentially perfect interfaces can be formed. This allows growth of quaternary quantum wells as thin as 25Å.

Ridge Laser Design Issues

The final device requires relatively small far-field divergence angles ($\Theta_\perp$ and $\Theta_{||}$), a small value of I_{th} and a high value for the differential quantum efficiency (η_d). Thickness (t_{act}) and effective index (n_{act}) of the active layer together with the refractive indexes of the top (n_t) and bottom (n_b) buffer regions determine β_{eff} and in turn control the angle $\Theta_\perp$ (see Fig. 4). (β_{eff} is the optical propagation constant for an equivalent slab waveguide whose layers are the same as the ridge part of the device.) The thickness (t_{buff}) of the buffer layer in the field and the width of the ridge (W_r) determine the angle $\Theta_{||}$. The thickness (t_{buff}) determines the size of the index discontinuity between the ridge and the field. A low threshold current is achieved by minimizing waste of carriers (to unnecessary lateral current spreading into the field regions) and by shielding the optical field from metal losses by a sufficiently thick dielectric layer. By judicious choice of dimensions these efficiency degrading current spreading effects can be minimized to obtain an extremely reliable device.

Once threshold (sufficient gain to account for all losses) is reached, any extra current produces lasing photons. These photons still face losses from various physical mechanisms: absorption in pumped and unpumped regions, scattering from defects, metal absorption losses, and useful light output from the end mirrors. The differential quantum efficiency can be expressed as:

$$\eta_d = \eta_i\left(\frac{\alpha_{mir}}{\alpha_{mir} + \alpha_{int}}\right),$$

where η_i is the internal quantum efficiency indicating the fraction of injected carriers converted to photons, α_{mir} is the loss resulting from light emitted from an end mirror, and α_{int} is the loss resulting from various internal factors. Both nonradiative recombination at surfaces and defects and Auger recombination† make η_i less than 100%. Since α_{mir} signifies useful output, one needs to keep α_{int} to a tolerable minimum to achieve a high η_d. For a semiconductor laser of length L_d and facet reflectivity R,

$$\alpha_{mir} = (l/L_d)\ln(1/R).$$

A value of L_d chosen to make $\alpha_{mir} = \alpha_{int}$ maximizes the output power. Factors dictated by instrument operation determine the actual device length.

Device Fabrication and Testing

Starting with a proper structure epitaxially grown on n-type InP substrate, a mask for the ridge is defined photolithographically. Using this mask, ridges of the chosen width and height are etched using a methane/hydrogen reactive ion etch.[6] A dielectric layer is deposited all over and selectively cleared from the ridge tops using a self-aligned photoresist process. Fig. 5 is an SEM cross-section showing photoresist on the sides, selectively protecting the dielectric layer that covers the entire top surface. Following this step, the dielectric layer is etched off the ridge top and the photoresist is removed from the field (see Fig. 6). A p contact/metal combination is deposited, covering the exposed ridge tops and the dielectric in the field. A thick gold pad is plated over the p metals to act as a heat spreader and as a bond pad. The wafer is thinned and an n contact metallization is deposited. Following a high-temperature contact alloying step the wafer

† Very briefly, Auger recombination[5] is a nonradiative and therefore lossy process that involves four particles: either three electrons and one hole or three holes and one electron. The Auger process is especially important in determining the temperature dependence of threshold current and η_i in long-wavelength semiconductor lasers such as ours.

© Hewlett-Packard Company 1995

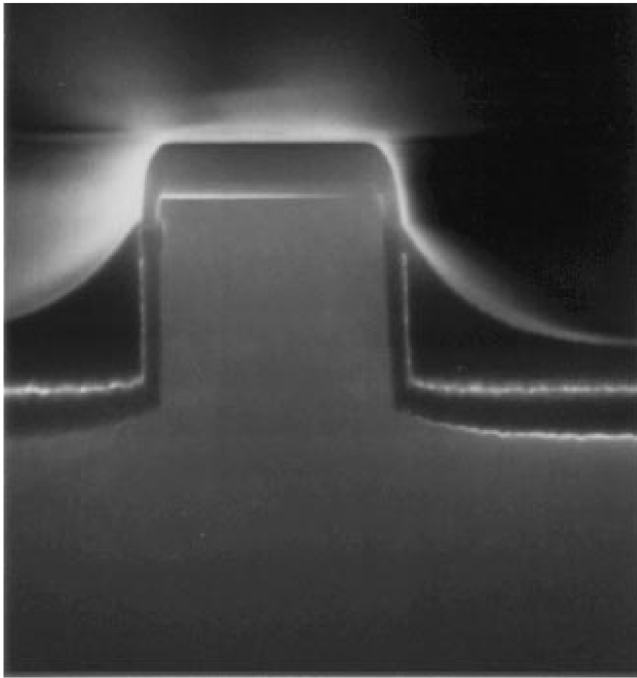
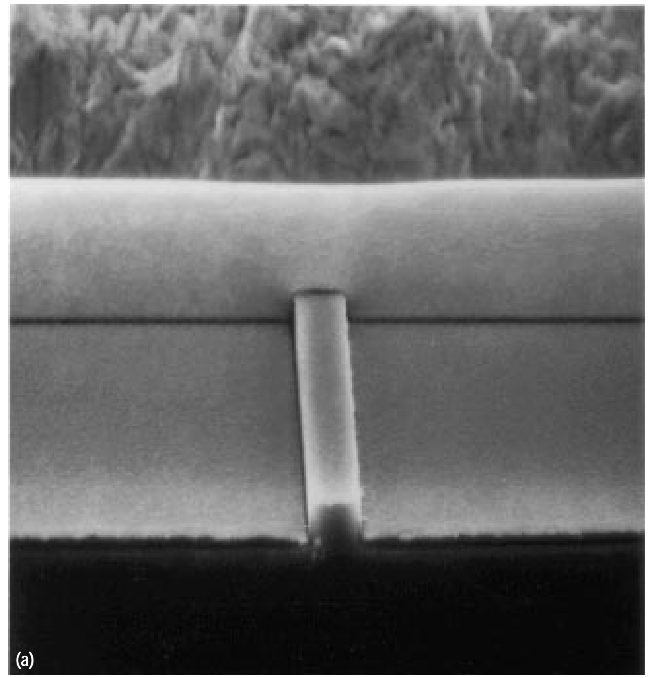


Fig. 5. Scanning electron microscope cross section of a ridge waveguide laser after exposure of ridge top with photoresist covering the oxide in the field.

is cleaved into bars for device testing. Fig. 7a is an SEM section of a completed device showing the ridge, a device facet, and the plated gold pad on top. Fig. 7b shows details of the device cross section with the active and contact layers clearly delineated.

Devices are first pulse tested in bar form to get an idea of their operation before die-attachment onto a heat sink. Devices 500 micrometers long have typical thresholds in the

Fig. 6. SEM cross section of a ridge waveguide laser after removal of the dielectric from the ridge top and photoresist from the field.



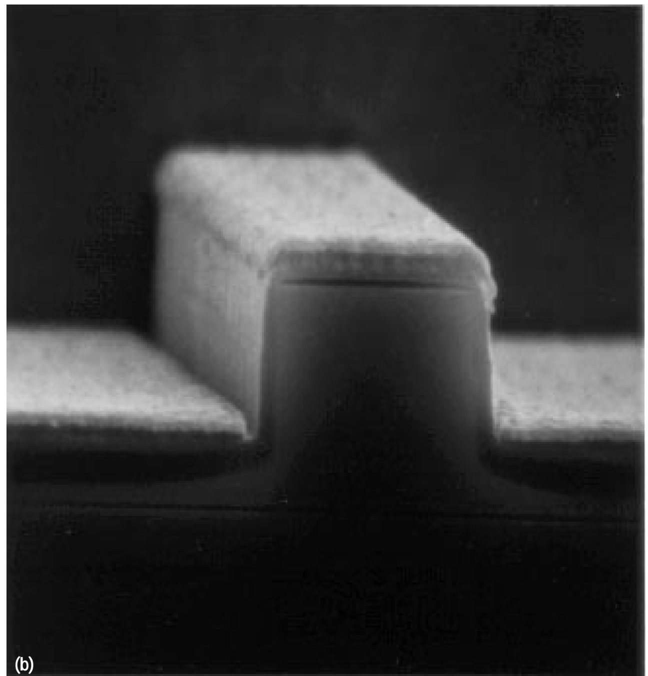


Fig. 7. (a) SEM section of a completed ridge laser showing the ridge and the plated gold. (b) SEM section of a ridge laser facet delineating the active and contact layers.

low 20-mA range with excellent slope efficiency and low divergence angles. Fig. 8 shows pulsed L-I and I-V characteristics, where L is light power output. Fig. 9 shows the far-field divergence angles perpendicular ($\Theta_{\perp}$) and parallel ($\Theta_{\parallel}$) to the plane of the device at different currents.

Device Reliability

A number of factors go into ensuring the reliability of our devices. To project device lifetimes under operational conditions, batches of devices die-attached to a suitable heat sink are subjected to elevated temperatures and a high bias current over many thousand hours. Periodically threshold

© Hewlett-Packard Company 1995

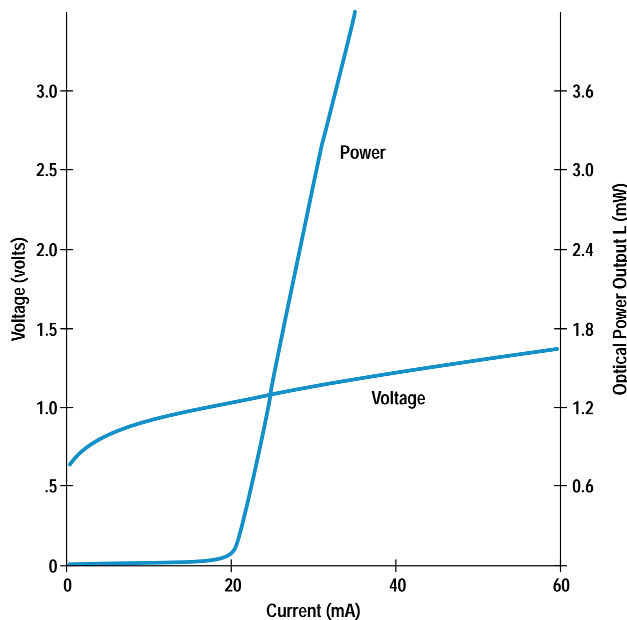
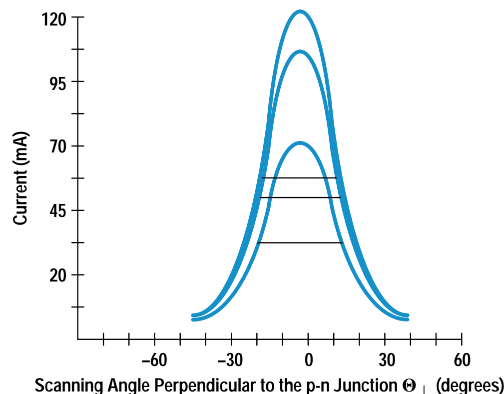
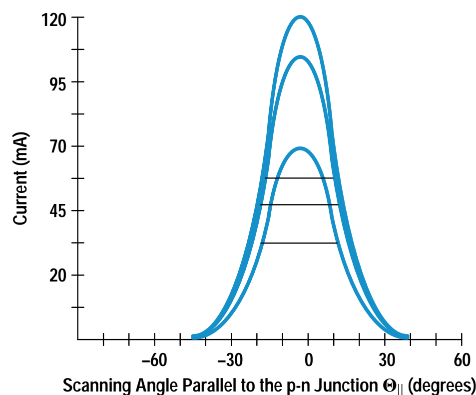


Fig. 8. Pulsed L-I (optical power output versus current) and I-V curves for a representative 1550-nm ridge waveguide laser.



Peak Current (mA)	Θ⊥ FWHM (Degrees)
70	32.6
100	32.7
120	32.7



Peak Current (mA)	Θ|| FWHM (Degrees)
70	23.4
100	23.5
120	23.9

Fig. 9. Far-field divergence angles perpendicular ($\Theta_\perp$) and parallel ($\Theta_{||}$) to the device junction plane.

current, power, and differential quantum efficiency are measured. The change in these parameters for devices stressed at different temperatures can be used to extract an activation energy for the principal failure mechanism. Device lifetime can then be predicted for normal operating conditions.

External-Cavity Operation

Completed individual devices are mounted on a heat sink. One facet is antireflection coated with a broadband multilayer coating.[3] The quality of this coating determines the suppression level of output power variations (from parasitic reflections), and ≤1 dB at the highest power levels of operation is achievable. With one of these devices properly aligned in an external cavity (see Fig. 10) we can generate a wavelength tuning curve at any operating current. Fig. 11 shows the power output as a function of wavelength for a representative device tested in an external cavity at an operating current of ≈120 mA.

Devices show >0 dBm output power (in a collimated beam) over the wavelength window 1465 to 1625 nm and put out >10 dBm over a smaller window, 1505 to 1615 nm. A single-mode fiber provides the output of the instrument (HP 8168C). With a reflection suppressing isolator included, these raw power output and tuning numbers are degraded somewhat. Nevertheless, these new custom ridge waveguide laser chips still provide significant increases in the power output and tuning window for the new instrument. The HP 8168C covers the range 1470 to 1590 nm with power increased by 8 dB and tuning window widened by a factor of two over the HP 8168A. Offering higher power (+2.5 dBm) in the range 1520

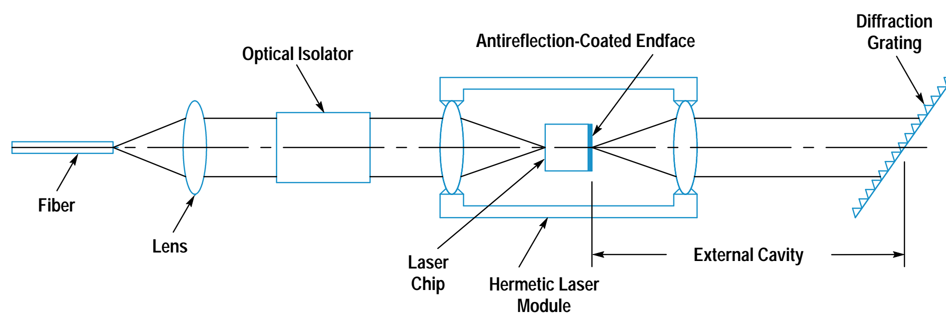


Fig. 10. Schematic of a laboratory grating-tuned external-cavity test setup.

© Hewlett-Packard Company 1995

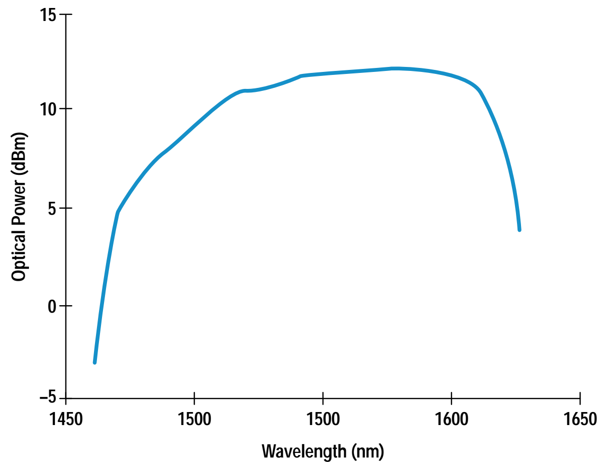


Fig. 11. Power output as a function of wavelength for a representative ridge laser in an external cavity.

to 1565 nm, this instrument is expected to be very useful for saturation measurements of erbium-doped fiber amplifier (EDFA) systems.

Summary

A multi-quantum-well ridge waveguide laser enhanced for use in a grating-tuned external-cavity source has been developed. The device offers higher output power and wider tunability in the new HP 8168C tunable laser source. This tunable optical source will allow simple testing of components for wavelength division multiplexing (WDM) as well as complete testing of EDFA systems. The core technology will allow the fabrication of other custom light sources in the wavelength window 1200 to 1700 nm.

Acknowledgments

The authors would specially like to thank Rose Twist for assistance in process development and device fabrication and Pamela Langhoff for assistance with epitaxial growth. Other people who have contributed to this project include: Jean Norman, Gary Trott, Joan Henderson, Mimi Planting, Nance Andring, Kari Salomaa, Nancy Wandrey, Deborah Sowers, and Roger Jungerman. We would like to express our appreciation for managerial support from Rick Trutna, Waguih Ishak, Kent Carey, Susan Sloan, Bob Bray, Jun Amano, and Ron Moon.

References

1. B. Maisenbacher, E. Leckel, R. Jahn, and M. Pott, "Tunable Laser Sources for Optical Amplifier Testing," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 11-19.

2. I.P. Kaminow, R.E. Nahory, M.A. Pollack, L.W. Stulz, and J.C. Dewinter, "Single-mode CW ridge-waveguide laser emitting at 1.55 µm," *Electronics Letters,* Vol. 15, 1979, pp. 763-765.

3. R.L. Jungerman, D.M. Braun, and K.K. Salomaa, "Dual-Output Laser Module for a Tunable Laser Source," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 32-34.

4. C. Weisbuch and B. Vinter, *Quantum Semiconductor Structures,* Academic Press, 1991, p. 21.

5. G.P. Agrawal and N.K. Dutta, *Long-Wavelength Semiconductor Lasers,* Van Nostrand Reinhold, 1986, Chapter 3.

6. U. Niggebrugge, M. Klug, and G. Garus, "A novel process for reactive ion etching on InP, using CH_4/H_2," *Institute of Physics Conference Series,* no. 79, 1986, pp. 367-372.

© Hewlett-Packard Company 1995

Measurement of Polarization-Mode Dispersion

Polarization-mode dispersion is defined and characterized, using Poincaré sphere and Jones matrix concepts. Interferometric, wavelength scanning, and Jones matrix eigenanalysis measurement methods are described. Instrumentation, especially the HP 8509B lightwave polarization analyzer, is discussed.

by Brian L. Heffner and Paul R. Hernday

New generations of high-speed undersea telecommunication systems and cable TV distribution systems feature an important new player: the erbium-doped fiber amplifier, or EDFA. Moving quickly from laboratory to mainline application, the EDFA will lower the cost and increase the reliability of long-haul telecommunications and greatly increase head-end distribution power for CATV.

In contrast to older systems in which propagation loss was compensated by detecting the optical signal and retransmitting it at higher power (regeneration), the EDFA-based system is a continuous glass pathway with amplification provided at intervals by short lengths of pumped, erbium-doped fiber. The absence of pulse regeneration must be offset by improvements in the dispersive characteristics of the pathway.

Historically, polarization-mode dispersion, or PMD is the third of a series of dispersive effects in optical fiber. The bandwidth of multimode fiber is limited because light separates into spatial modes of many different lengths. Single-mode fiber solves that problem but is limited by chromatic dispersion, in which the transmission medium allows adjacent wavelengths to travel at slightly different speeds. PMD, a more subtle effect, arises from slight physical asymmetry in the index of refraction, called birefringence. In fiber, it is caused by stresses induced by fiber manufacture, packaging, and deployment and is strongly influenced by environment.

When chromatic dispersion is sufficiently reduced, the pulse distortion and signal fading produced by PMD can be observed. In CATV systems, the combination of PMD in fiber and components, frequency chirp in the transmitter, and polarization dependent loss near the receiver produces composite second-order distortion. For high-speed, long-haul telecommunications, and high-channel-capacity CATV systems to realize their potential, PMD must be understood and controlled.

Polarization-mode dispersion is a fundamental property of single-mode optical fiber and components in which signal energy at a given wavelength is resolved into two orthogonal polarization modes of slightly different propagation velocity. The resulting difference in propagation time between polarization modes is called the differential group delay, commonly symbolized as $\Delta\tau_g$, or simply $\Delta\tau$. In most optical components, the polarization modes correspond to physical axes of the component and the differential group delay (and therefore the PMD) is nearly independent of wavelength. In practical lengths of optical fiber, differential group delay varies randomly with wavelength and the specification of PMD must be statistically based. Long-fiber PMD is commonly expressed as either the average value or the rms value of differential group delay over a wide wavelength range. For fibers that exhibit a large degree of coupling of energy between polarization modes, PMD scales with the square root of fiber length and is often specified in picoseconds per root kilometer.

How much PMD is too much? For modest impact, the instantaneous differential group delay of a telecommunication system must be kept below one tenth of a bit period, or 20 ps for a 5-Gbit/s NRZ pulse stream.

Characterizing PMD

PMD in many real systems varies over time and is best characterized by a statistical picture to account for its changing details. For the moment, however, let's consider how to characterize the PMD of a stable device or system that exhibits no time variation. Differential group delay, the most direct measure of the signal-distorting effects of PMD, does not tell the whole story. The PMD of a system is completely characterized by specifying any of the following three quantities as a function of wavelength or optical frequency:

- A pair of principal states of polarization and a differential group delay
- A three-dimensional polarization dispersion vector
- A Jones matrix (see page 28).

If a polarized, tunable optical wave is transmitted through a device, the polarization at the device output will in general trace out an irregular path on the Poincaré sphere[1] (see page 29) as the optical radian frequency ω is tuned, as shown in Fig. 1. Over a small range of frequency, any section of the irregular path can be approximated as an arc of a circle on the surface of the sphere. The center of such a circle, projected to the surface of the sphere, locates a principal state of polarization. A second, orthogonal principal state of polarization is located diametrically opposite on the sphere. The principal states of polarization are significant because they summarize how any output state of polarization evolves with frequency. As a function of frequency, all output states

© Hewlett-Packard Company 1995

Jones Calculus

Between 1941 and 1948, R. Clark Jones published a series of papers describing a new polarization calculus based upon optical fields rather than intensities. This approach, although more removed from direct observation than previous methods, allowed calculation of interference effects and in some cases provided a simpler description of optical physics. A completely polarized optical field can be represented by a two-element complex vector, each element specifying the magnitude and phase of the x and y components of the field at a particular point in space. The effect of transmission through an optical device is modeled by multiplying the input field vector by a complex two-by-two device matrix to obtain an output field vector.

The matrix representation of an unknown device can be found by measuring three output Jones vectors in response to three known stimulus polarizations. Calculation of the matrix is simplest when the stimuli are linear polarizations oriented at 0, 45, and 90 degrees (Fig. 1), but any three distinct stimuli may be used. The matrix calculated in this manner is related to the true Jones matrix by a multiplicative complex constant c. The magnitude of this constant can be calculated from intensities measured with the device removed from the optical path, but the phase is relatively difficult to calculate, requiring a stable interferometric measurement. Fortunately, measurements of many characteristics such as PMD do not require determination of this constant.

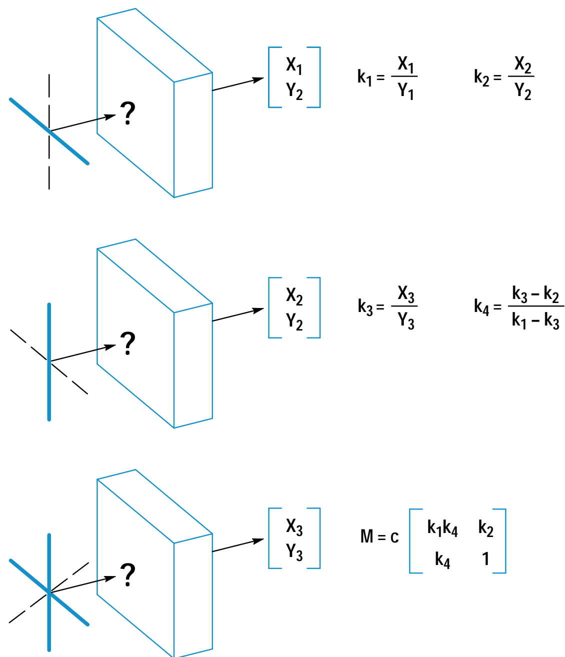


Fig. 1. Measurement of the Jones matrix requires application of three known states of polarization. Output electric field descriptions and ratios k_x are complex quantities. The Jones matrix **M** is found to within a complex constant c, whose phase represents the absolute propagation delay and is not required for PMD measurements.

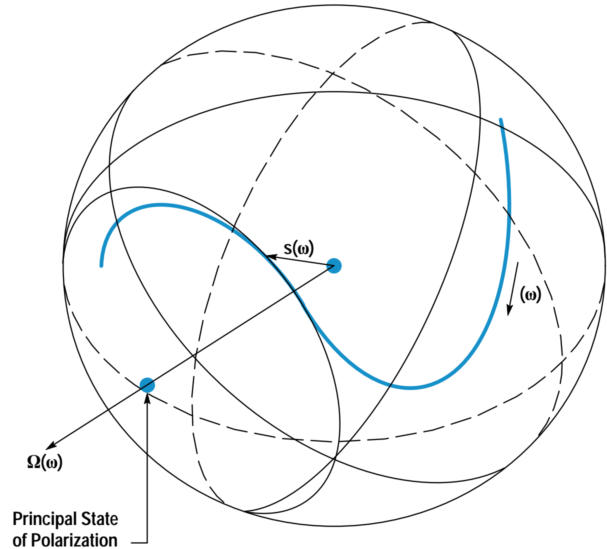


Fig. 1. Relationship between the polarization dispersion vector **Ω** and the output state of polarization **s** on the Poincaré sphere. The heavy line shows the path of the output polarization as the optical frequency changes. The output path is approximated over a small range of frequency as a circular arc generated by rotation about the polarization dispersion vector. The vector points in the direction of one principal state of polarization, and a second principal state of polarization is located diametrically opposite on the sphere.

rotate about a diameter connecting the two principal states of polarization. The rate of rotation is determined by the differential group delay.

The polarization dispersion vector **Ω** is probably the most intuitively meaningful representation of PMD because it is defined in the same real, three-dimensional space as the Poincaré sphere.[2] This vector originates at the center of the Poincaré sphere and points toward the principal state of polarization about which the output states of polarization rotate in a counterclockwise sense with increasing ω. |**Ω**|, the length of **Ω**, is the differential group delay. When the output state of polarization is expressed as a three-dimensional vector **s** composed of normalized Stokes parameters[1] locating the state of polarization on the sphere, the rotation about the principal states of polarization can be written as a cross product relation: $d\mathbf{s}/d\omega = \mathbf{\Omega} \times \mathbf{s}$. In the most general case **Ω** is a function of the optical frequency ω.

In some optical components, such as isolators and waveguide modulators, PMD originates in crystals or waveguides through which light propagates with different group delays for different polarizations. As the optical frequency transmitted through these components is varied, **Ω** remains fixed in orientation, while |**Ω**| might vary slightly as a result of chromatic dispersion in each polarization mode. Devices such as these, in which **Ω** is essentially independent of ω, are especially simple to describe and to measure. In particular, the differential group delay is constant over time and can be accurately characterized as a weak function of frequency. Such devices can be useful as standards because their characteristics can be expected to remain stable from one time and location to the next.

Modern single-mode fibers achieve a very low level of local birefringence in addition to low propagation loss, but the effects of small local birefringences along the fiber can accumulate to cause significant PMD through a fiber many kilometers long. Birefringence is a property of a dielectric describing the difference in the indexes of refraction for different polarizations. Local birefringence in a fiber can be caused by deviations from perfect circular symmetry of the fiber core, or by asymmetrical stress in the core region owing to the manufacturing process, bends in the fiber, or temperature gradients. Stresses can change with temperature and with time as the glass and the surrounding cable relax, leading to a birefringence along the fiber length that evolves

© Hewlett-Packard Company 1995

The Poincaré Sphere

The Poincaré sphere is a graphical tool in real, three-dimensional space that allows convenient description of polarized signals and of polarization transformations caused by propagation through devices. Any state of polarization can be uniquely represented by a point on or within a unit sphere centered on a rectangular (x,y,z) coordinate system. The coordinates of the point are the three normalized Stokes parameters describing the state of polarization. Partially polarized light can be considered a combination of purely polarized light of intensity I_P and unpolarized light of intensity I_U.

The degree of polarization $I_P/(I_P + I_U)$ corresponding to a point is the distance of that point from the coordinate origin, and can vary from zero at the origin (unpolarized light) to unity at the sphere surface (completely polarized light). Points close together on the sphere represent polarizations that are similar, in the sense that the interferometric contrast between two polarizations is related to the distance between the corresponding two points on the sphere.

Orthogonal polarizations with zero interferometric contrast are located diametrically opposite on the sphere. As shown in Fig. 1, linear polarizations are located on the equator. Circular states are located at the poles, with intermediate elliptical states continuously distributed between the equator and the poles. Right-hand and left-hand elliptical states occupy the northern and southern hemispheres, respectively.

Because a state of polarization is represented by a point, a continuous evolution of polarization can be represented as a continuous path on the Poincaré sphere. For example, the evolution of polarization for light traveling through a waveplate or birefringent crystal is represented by a circular arc about an axis drawn through the two points representing the eigenmodes of the medium. (Eigenmodes are polarizations that propagate unchanged through the medium.) A path can also record the polarization history of a signal, for example in response to changing strain applied to a birefringent fiber.

The real, three-dimensional space of the Poincaré sphere surface is closely linked to the complex, two-dimensional space of Jones vectors (see page 28). Most physical ideas can be expressed in either context, the mathematical links between the two spaces having previously been established for dealing with angular momentum. The graphical Poincaré description allows for a more intuitive approach to polarization mathematics.

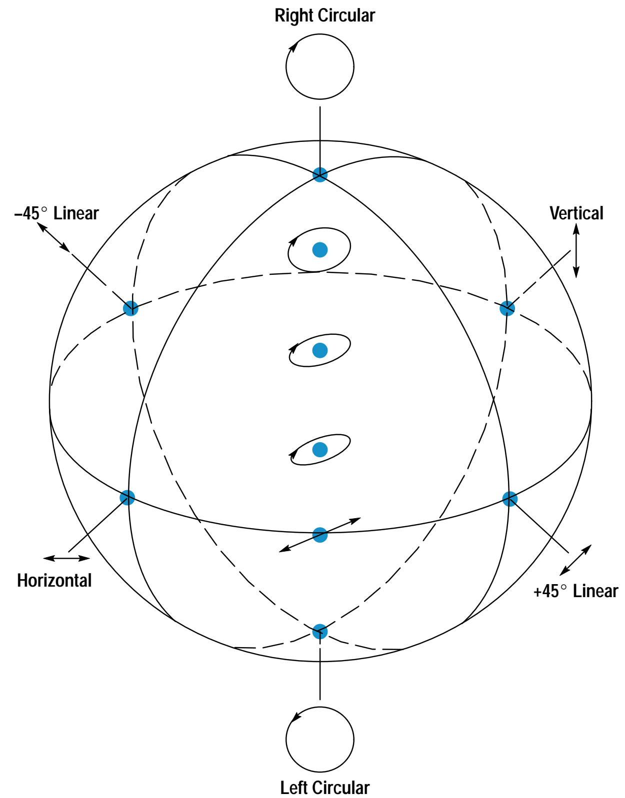


Fig 1. Points displayed on the surface of the Poincaré sphere represent the polarized portion of a lightwave. Linear polarizations are located on the equator. Circular states are located at the poles, with intermediate elliptical states continuously distributed between the equator and the poles. Right-hand and left-hand elliptical states occupy the northern and southern hemispheres, respectively. Example states are shown ascending the front of the sphere.

over time. This behavior leads to measurable PMD that is a function of both frequency and time, so that a statistical picture becomes the most appropriate view of PMD for the system designer.

A detailed statistical model of PMD has been developed from the basis of accumulated local birefringences, and has been experimentally confirmed.[3,4]

Each of the x, y, and z components of the polarization dispersion vector for a long fiber follows a normal distribution with zero mean. As a result, the orientation of $\mathbf{\Omega}$ is uniformly distributed, and the distribution of the differential group delay $\Delta\tau = |\mathbf{\Omega}|$ is proportional to $\Delta\tau^2\exp(-\Delta\tau^2/2\,\alpha^2)$. This is often called a Maxwell distribution because it is the same as the Maxwell distribution of molecular speed for a gas in thermal equilibrium. The distribution has an expected value $\langle\Delta\tau\rangle$ of $\alpha\sqrt{8/\pi}$.

The statistical theory predicts, and experiments confirm, that the differential group delay distribution measured at a particular frequency over a long period of time is identical to the distribution measured at one time over a large range of frequency. This fact allows statistics representing slow time variations to be measured very quickly by gathering data over a wide frequency range. As another result of the statistical model, when several long fiber sections are concatenated, the expected value of differential group delay for the concatenation is given by the root sum of squares of the expected values for the sections, that is,

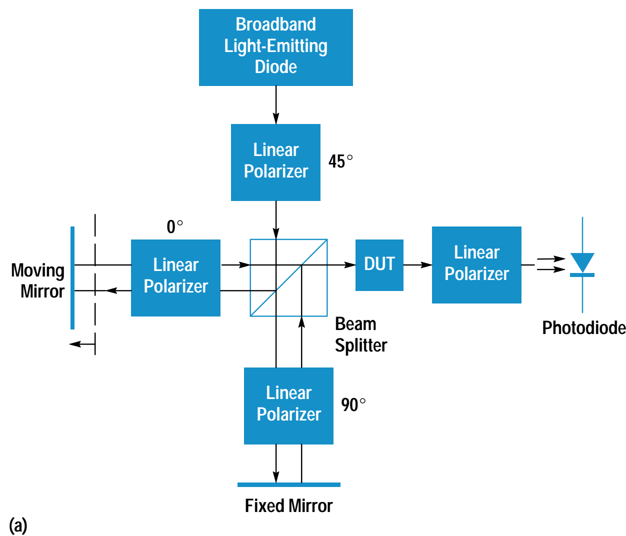
$$\langle\Delta\tau\rangle_{total} = \sqrt{\langle\Delta\tau\rangle_1^2 + \langle\Delta\tau\rangle_2^2 + \ldots}\,.$$

As a consequence, $\langle\Delta\tau\rangle_{total}$ grows proportionally to the square root of the fiber length, and the PMD of a long fiber is specified in units of $ps/\sqrt{km}$ with the understanding that the orientation of $\mathbf{\Omega}$ is uniformly distributed. In contrast, the PMD of a short section of fiber or of a fiber manufactured with a consistent birefringence over its length is specified in units of ps/km because it grows proportionally to the fiber length, and the orientation of $\mathbf{\Omega}$ is understood to be fixed relative to the physical orientation of the fiber. PMD in components is typically not statistical in origin, and is simply specified in ps.

Measuring PMD

Two polarization modes are transmitted through a device exhibiting significant PMD, each according to its own phase delay and group delay. Owing to the unequal group delays, propagation through such a device will change the mutual temporal coherence between the two polarization modes.

© Hewlett-Packard Company 1995



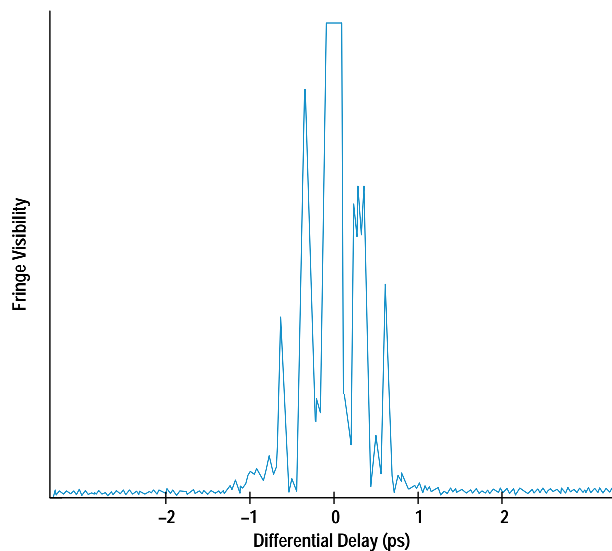
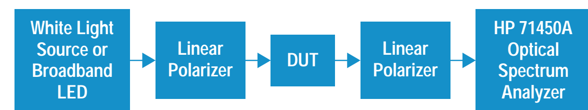


Fig. 2. Measurement of PMD using interferometry. (a) Interference fringes are detected only when the differential group delay between orthogonally polarized input states is compensated by the differential group delay of the DUT. (b) Measurement of a 44-km spooled fiber.

Likewise, the two unequal phase delays lead to a frequency dependent output state of polarization in response to a fixed input state of polarization. These physical characteristics make possible a variety of PMD measurement methods. An interferometric method[5,6] measures the effect of PMD on mutual coherence, and a wavelength scanning method[7] measures the effect of PMD, through variations of the output state of polarization, on transmission through a fixed analyzer. A method developed by Hewlett-Packard calculates the differential group delay and principal states of polarization as a function of frequency by analyzing Jones matrixes measured at a sequence of optical frequencies.[8,9,10] Most of the techniques currently used to measure PMD are similar in principle to one of these three methods.

A block diagram of the low-coherence interferometric method is shown in Fig. 2a. Collimated light from a broadband light-emitting diode is polarized and split into two



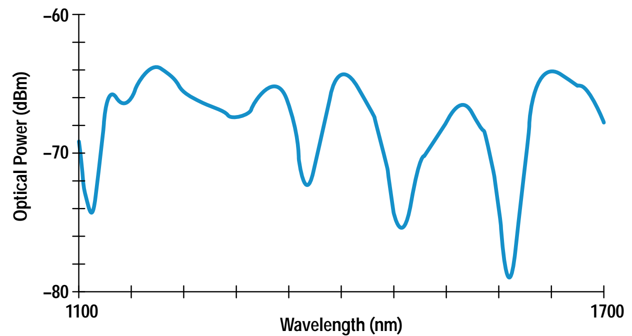


Fig. 3. Measurement of PMD using wavelength scanning. (a) System block diagram. (b) Measurement of a 4-km spool of single-mode fiber using a white light source.

mutually coherent beams. One mirror can be scanned in position, creating a differential delay between the two orthogonal polarizations, which are recombined and directed through the device under test (DUT). When photocurrent is measured as a function of the differential interferometer delay, coherent fringes can be observed only when this differential delay is compensated by the differential group delay of the DUT. Fig. 2b shows the envelope of the coherent fringes measured as a function of delay.

The wavelength scanning method, also called the fixed analyzer method, is shown schematically in Fig. 3a and can be assembled using equipment found in many optics laboratories without specialized equipment for PMD measurement. Polarized broadband light is directed through the DUT. PMD in the DUT causes the output state of polarization to trace out an irregular path on the Poincaré sphere when measured as a function of optical frequency, as was shown in Fig. 1.

By measuring the optical power transmitted through an output analyzer as a function of optical frequency (Fig. 3b), we effectively measure one dimension of the three-dimensional path on the sphere, leading to ripples in the spectral density measured by the optical spectrum analyzer. The average differential group delay over the measured frequency span is proportional to the number of spectral density extrema within the span.

Jones matrix eigenanalysis[8] is based upon Jones matrixes measured at a sequence of optical frequencies using the HP 8509B lightwave polarization analyzer and the HP 8167A and 8168A tunable laser sources,[11] as shown schematically in Fig. 4a. At each frequency a Jones matrix $\mathbf{T}_k$ is measured by stimulating the DUT with three accurately known states of polarization and measuring the response state of polarization at the DUT output.[12] The matrix product $\mathbf{T}_{k+1}\mathbf{T}_k^{-1}$ reveals the change in the polarization transformation caused by the change in frequency, which in this case is a rotation about $\mathbf{\Omega}$. The eigenvectors of $\mathbf{T}_{k+1}\mathbf{T}_k^{-1}$ yield the orientation of $\mathbf{\Omega}$ and

© Hewlett-Packard Company 1995

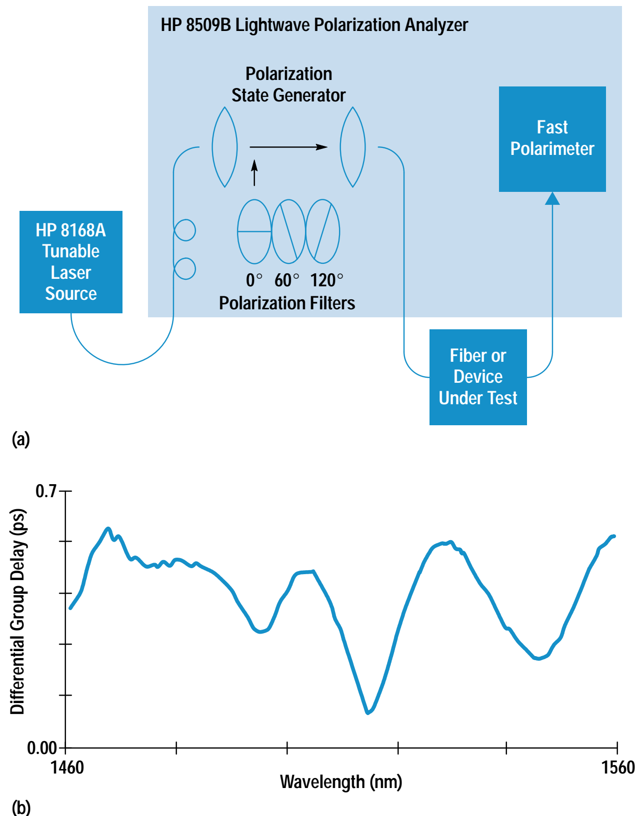


Fig. 4. Measurement of PMD using Jones matrix eigenanalysis. (a) System block diagram. (b) Measurement of a 8-km "high-PMD" single-mode fiber.

the eigenvalues ρ_1 and ρ_2 of $\mathbf{T}_{k+1}\mathbf{T}_k^{-1}$ yield the differential group delay through the relation:

$$\Delta\tau = |\mathbf{\Omega}| = \frac{\mathrm{Arg}(\rho_1/\rho_2)}{\Delta\omega},$$

where Arg is the argument function (the argument of a complex number is its polar angle, that is, Arg $\gamma e^{i\beta} = \beta$). Stepping pairwise through the sequence of matrixes $\mathbf{T}_k$, we obtain both principal states of polarization and the differential group delay as a function of optical frequency (Fig. 4b).

The ability to measure the frequency dependence of PMD can be helpful in diagnosing its physical origins, but in evaluating a highly mode-coupled device such as a long fiber only the average differential group delay is significant. This raises the question of how much data must be collected before a reliable estimate of the average differential group delay $\langle\Delta\tau\rangle$ is obtained. All three of these techniques gather data over a range of frequencies, and larger frequency or wavelength spans generally result in more reliable estimates of $\langle\Delta\tau\rangle$.

The interferometric and Jones matrix methods both can give an indication of when sufficient data has been collected by comparing measured data to curves theoretically predicted by the statistical model. Interferometry measures data we expect to be normally distributed, so the measured data is compared to a Gaussian curve to assess the estimate of $\langle\Delta\tau\rangle$. Fig. 5a shows a good fit. Eigenanalysis measures the magnitude $|\mathbf{\Omega}|$, which we expect to follow a Maxwell distribution. The curve of differential group delay versus wavelength is first converted into a histogram showing the number of measurements versus differential group delay, and then the histogram is compared to a Maxwell curve. Fig. 5b shows a good fit.

When the frequency span is not sufficient to produce a sufficiently good fit to the expected curve, new data can be collected by waiting for the physical properties of the fiber to change, assuming that the same statistical model remains valid. In the laboratory the characteristics of a spooled fiber are measured again after the fiber temperature is changed by a few degrees, while a deployed fiber must be measured again after several hours have elapsed. Multiple measurements over the same frequency span allow collection of a set of data that accurately predicts $\langle\Delta\tau\rangle$, as reflected by a good fit to the expected curve.

Instrumentation

The HP 8509B lightwave polarization analyzer, discussed on page 32, provides two automated methods for measurement of PMD: the Jones matrix eigenanalysis and three-Stokes-parameter wavelength scanning methods (Fig. 6). Both make use of the HP 8167/68A tunable laser sources.[11] The

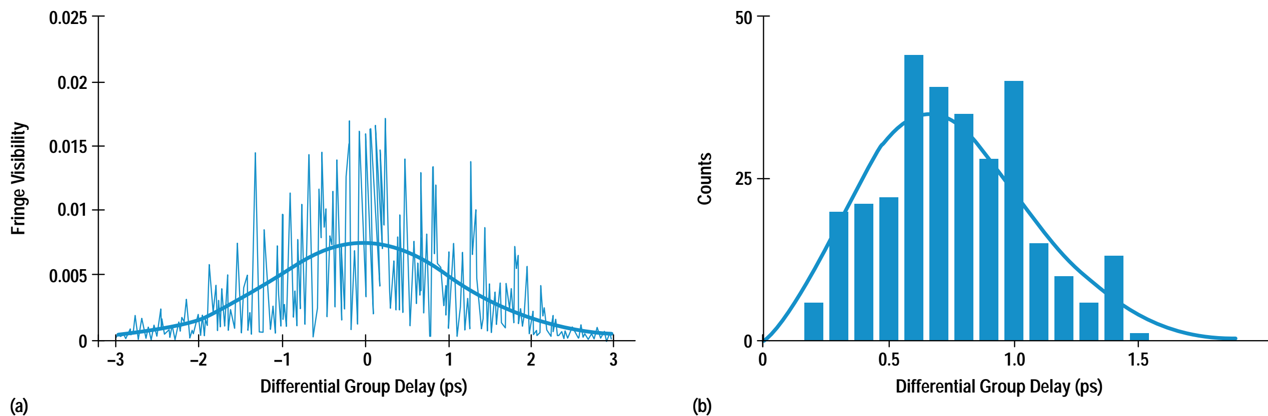


Fig. 5. (a) The smooth curve shows the expected normal distribution of the data measured using the interferometric method. In this case the measured data (jagged line) is a fairly good fit. (b) Eigenanalysis measures the magnitude $|\mathbf{\Omega}|$, which is expected to follow a Maxwell distribution (smooth curve). Here the measured data (histogram) is a good fit.

© Hewlett-Packard Company 1995

The HP 8509A/B Lightwave Polarization Analyzer

With the advent of the lightwave polarimeter, engineers in the fields of high-speed telecommunications, cable TV distribution, optical sensing, optical recording, and materials science can characterize polarization phenomena with the ease and graphical simplicity of the common oscilloscope. Supplemented by an optical source, a polarization state generator, and comprehensive measurement software, the polarimeter becomes a polarization analyzer, producing comprehensive measurements of both optical signals and two-port optical devices.

The HP 8509B lightwave polarization analyzer consists of an optical unit and a 66-MHz HP Vectra PC. The main display window, shown in Fig. 1, conveniently displays the polarization parameters of an optical signal and provides access to commonly used controls. The Measurements menu provides access to a variety of integrated measurement solutions addressing polarization-mode dispersion (PMD), polarization dependent loss, the Jones matrix, and optimization of optical launch into polarization maintaining fiber. The Display menu allows customization of the display window and the System menu enables the user to reconfigure system operating parameters, optimize performance at a particular wavelength, and automatically check the functional integrity of the instrument.

The heart of the HP 8509A/B is a high-speed polarimeter (see Fig. 4a in the accompanying article). A passive optical assembly (see cover) divides the optical signal into four beams and passes each beam through polarization filters to photodiode detectors. Autoranging amplifiers and 16-bit ADCs complete the circuitry. A series of calibration coefficients are determined at manufacture and stored in UV-PROM. The instrument interpolates among these coefficients to provide operation from 1200 to 1600 nm. Parallel filtering and detection combined with high-speed conversion and computation result in a measurement rate of 3000 polarization states per second.

A second optical assembly inserts three polarizing filters in the optical source path to allow measurement of the Jones matrix. The Jones matrix eigenanalysis PMD measurement method is based upon Jones matrixes (see page 28) measured at a series of wavelengths. Polarization dependent loss is also derived from the Jones matrix. In addition, the user can use external polarizers to define a physical reference frame, analytically removing the birefringence and polarization dependent loss of components between the polarizer and the polarimeter receiver. Once defined, the reference frame allows the measurement of absolute polarization state at a point far from the instrument itself.

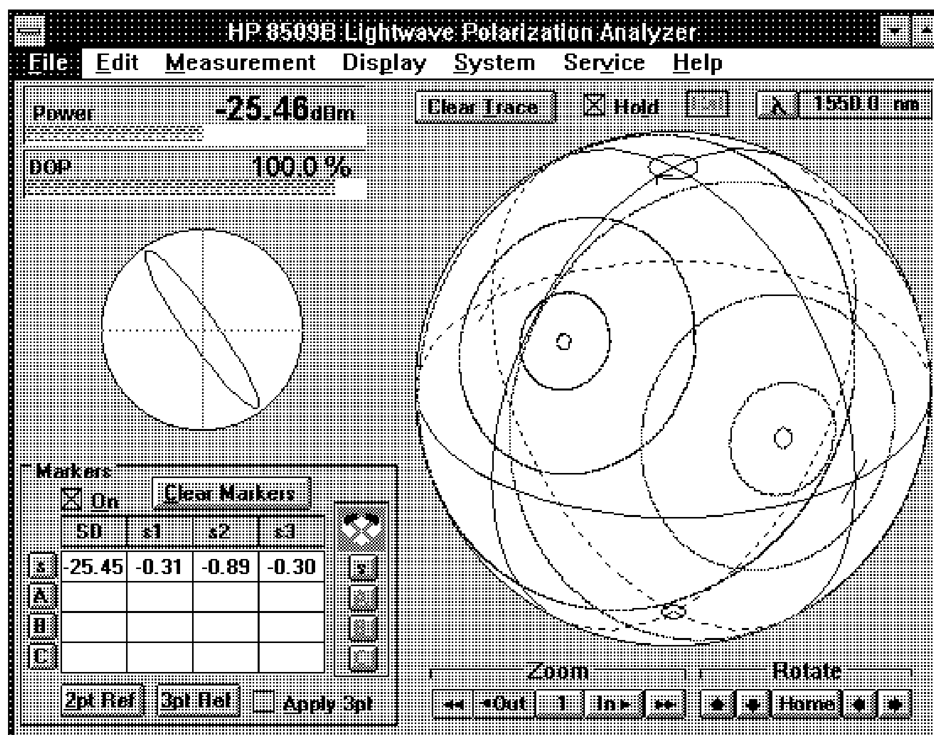


Fig. 1. Main measurement window of the HP 8509B lightwave polarization analyzer. Displays of average power and degree of polarization (DOP), along with elliptical and Poincaré sphere displays, fully characterize the polarization state of a lightwave. Shown on the sphere are the loci of output polarization states of a polarization maintaining fiber as the fiber is gently stretched. Red traces are on the front of the sphere, blue on the back. Different circles correspond to different states at the input of the fiber. The circles converge to points when polarized light is launched entirely on the fast or slow axes of the fiber.

wavelength scanning method determines three PMD values from changes in output polarization as observed along the three axes of the Poincaré sphere, then averages these results to provide a single value of PMD that is much less dependent on launch condition than conventional implementations of wavelength scanning. The measured normalized Stokes parameters are independent of signal power and are therefore immune to optical signal level changes, allowing a better measurement to be derived from a single wavelength sweep. Because the wavelength scanning method does not require the internal three-state polarizer, the method is also available on the HP 8509A.

The HP 71450A optical spectrum analyzer[13] with Option 002 (internal white light source) is a powerful foundation for traditional wavelength scanning PMD measurements. The interferometric method of PMD measurement is available for certain applications via the HP 8504A/B precision reflectometer.[14]

Acknowledgments

The authors wish to thank Greg Gibbons, Duncan Gurley, Richard Allen, Mike Hart, and Jeff Paul for the development of the original software of the HP 8509A/B lightwave polarization analyzer and Mike Fitzpatrick for recent enhancements, Roger Jungerman, Randy King, Don Cropper, and Jim Smith for the design of the polarimeter receiver, Fred Rawson and Matt Klein for circuit design, Luis Fernandez for overall product design, and Jim Yarnell and Ivan Hammer for contributions to the mechanical design. We are also grateful to Harry Chou for important enhancements in the

© Hewlett-Packard Company 1995

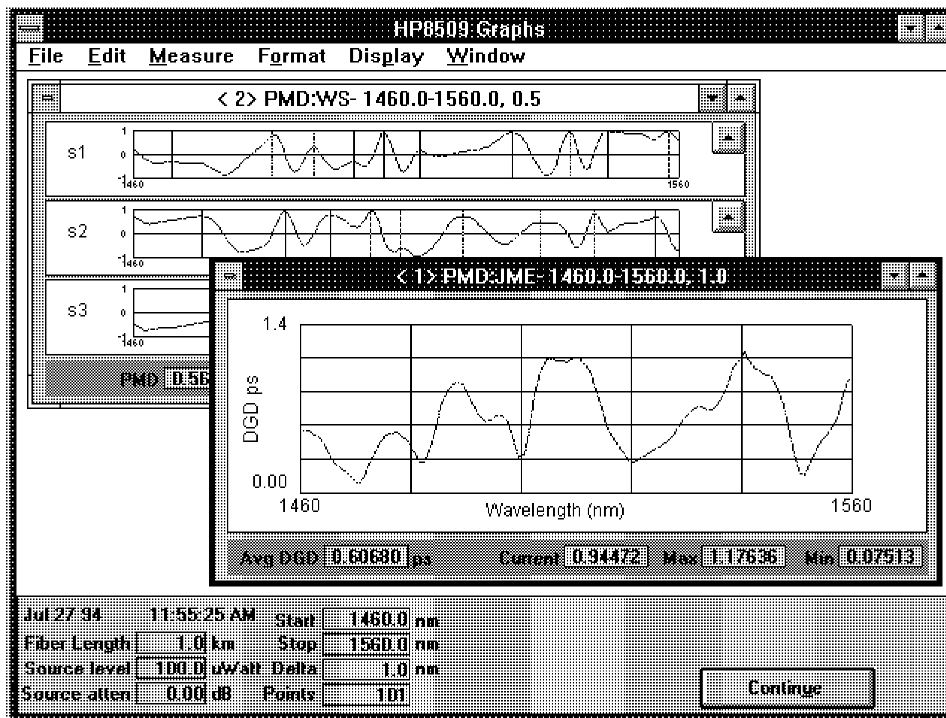


Fig. 6. HP 8509B lightwave polarization analyzer display of polarization-mode dispersion (PMD) measurement results. The Jones matrix eigenanalysis method (foreground) measures differential group delay (DGD) as a function of wavelength. The three-Stokes–parameter wavelength scanning method (background) is enhanced by analyzing three responses produced by a single wavelength sweep.

design and to Pamela Pitcher, Karl Shubert, Roger Wong, Jack Dupre, Hugo Vifian, and Steve Newton, for their strong support.

References

1. M. Born and E. Wolf, *Principles of Optics, Sixth Edition,* Pergamon, 1980.

2. N.S. Bergano, C.D. Poole and R.E. Wagner, "Investigation of polarization dispersion in long lengths of single-mode fiber using multi-longitudinal mode lasers," *Journal of Lightwave Technology,* Vol. LT-5, 1987, pp. 1618-1622.

3. C.D. Poole, J.H. Winters and J.A. Nagel. "Dynamical equation for polarization dispersion," *Optics Letters,* Vol. 16, 1991, pp. 372-374.

4. G.J. Foschini and C.D. Poole, "Statistical theory of polarization dispersion in single-mode fibers," *Journal of Lightwave Technology,* Vol. LT-9, 1991, pp. 1439-1456.

5. K. Mochizuki, Y. Namihira, and H. Wakabayashi, "Polarisation mode dispersion measurements in long single-mode fibers," *Electronics Letters,* Vol. 17, 1981, pp. 153-154.

6. N. Gisin, J-P Von der Weid, and J-P Pellaux, "Polarization mode dispersion of short and long single-mode fibers," *Journal of Lightwave Technology,* Vol. LT-9, 1991, pp. 821-827 and references therein.

7. C.D. Poole and D.L. Favin, "Polarization-mode dispersion measurements based on transmission spectra through a polarizer," *Journal of Lightwave Technology,* Vol. LT-12, 1994, pp. 917-929.

8. B.L. Heffner, "Automated measurement of polarization mode dispersion using Jones matrix eigenanalysis," *Photonics Technology Letters,* Vol. 4, 1992, pp. 1066-1069.

9. B.L. Heffner, "Accurate, automated measurement of differential group delay dispersion and principal state variation using Jones matrix eigenanalysis," *Photonics Technology Letters,* Vol. 5, 1993, pp. 814-817.

10. B.L. Heffner, "Attosecond-resolution measurement of polarization mode dispersion in short sections of optical fiber," *Optics Letters,* Vol. 18, 1993, pp. 2102-2104.

11. B. Maisenbacher, E. Leckel, R. Jahn, and M. Pott, "Tunable Laser Sources for Optical Amplifier Testing," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 11-19.

12. R.C. Jones, "A new calculus for the treatment of optical systems. VI: Experimental determination of the matrix," *Journal of the Optical Society of America,* Vol. 37, 1947, pp. 110-112.

13. D.A. Bailey and J.R. Stimple, "Optical Spectrum Analyzers with High Dynamic Range and Excellent Input Sensitivity," *Hewlett-Packard Journal,* Vol. 44, no. 6, December 1993, pp. 60-67.

14. D.H. Booster, H. Chou, M.G. Hart, S.J. Mifsud, and R.F. Rawson, "Design of a Precision Optical Low-Coherence Reflectometer," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 39-48.

© Hewlett-Packard Company 1995

A New Design Approach for a Programmable Optical Attenuator

The new HP 8156A optical attenuator offers improved performance, low polarization dependent loss and polarization-mode dispersion, and increased versatility. It uses a birefringence-free glass filter disk and a high-resolution, fast-settling filter drive system.

by Siegmar Schmidt and Halmo Fischer

For over eight years, HP programmable optical attenuators have offered high performance for many fiber-optic measurement tasks. Now, increasing data rates in digital transmission systems and the use of analog systems for cable TV require a new standard in test and measurement equipment. Optical attenuators with high return loss are essential for the measurement of bit error rates and noise performance in these systems. In addition, the longer links made possible by the use of erbium-doped fiber amplifiers as all-optical regenerators increase the importance of parameters that, until now, where considered less relevant. Typical examples are polarization dependent loss, polarization-mode dispersion, and high input power. In production, high cost pressures require high yields and throughput. The intense competitive situation demands reduced test margins to show the true performance of the devices tested, and this requires higher-performance test equipment that does not unduly influence the test results.

To meet these needs, the new HP 8156A optical attenuator, Fig. 1, has been developed with improved performance and with attention to parameters that have become more relevant than in the past. Compared with its predecessors, the HP 8156A shows improved performance with respect to linearity, accuracy, resolution, return loss, and settling time.

Fig. 1. The HP 8156A optical attenuator provides improved linearity, accuracy, resolution, return loss, and settling time, low polarization dependent loss and polarization-mode dispersion, and new features including a separate shutter and built-in software applications.

Its polarization dependent loss, polarization-mode dispersion, and input power level are specified, and it has several new features, including a separate shutter, built-in software applications, and options to tailor it to different fiber-optic applications. The options include standard-performance, high-performance, monitor, and high-return-loss options for single-mode fibers (fiber core diameter 8 μm), and a multimode fiber option (fiber core diameter 50 μm). The operating wavelength range covers the two fiber-optic wavelength regions around 1300 and 1550 nm.

Optical System

Commercially available fiber-optic attenuators, both variable and fixed, use a range of techniques for achieving optical attenuation. Some use techniques based on angular, lateral, or axial displacement of two optical fiber ends. Others use grayscale filters or polarizers to attenuate the light. The HP 8156A uses a circular grayscale filter and various bulk optic components.

As shown in Fig. 2, a first objective lens collimates the light of the input fiber to a parallel beam and a second objective lens refocuses it onto the output fiber. A mechanical shutter, a circular filter, a corner cube, and a prism are located between the two lenses. The circular attenuating filter consists of two disks of glass wedges glued together, one absorbing and the other transparent. The attenuation depends on the angular position of the attenuating filter, which is set by a motor that is controlled by a digital positioning system.

The attenuation of the wedged absorbing glass filter as a function of the angular position α is:

$$\mathrm{Att}(\alpha) = \frac{\mathrm{Att}_{max}}{2}(1 - \cos\alpha), \quad (1)$$

where Att_{max} = 60 dB. For angles α = 0 to α = 180 degrees the attenuation varies between 0 and Att_{max} in a strongly monotonic manner. No ranging effects and related overshoot and undershoot occur, that is, there are no dark spots. This is a very important feature for measuring thresholds in bit error rate measurements.

To linearize the attenuation characteristics, each attenuator is individually calibrated in production at wavelengths of 1300 nm and 1550 nm. An automatic calibration program characterizes each unit. The measured data is processed by the computer and transferred to an EEPROM in the attenuator.

 © Hewlett-Packard Company 1995

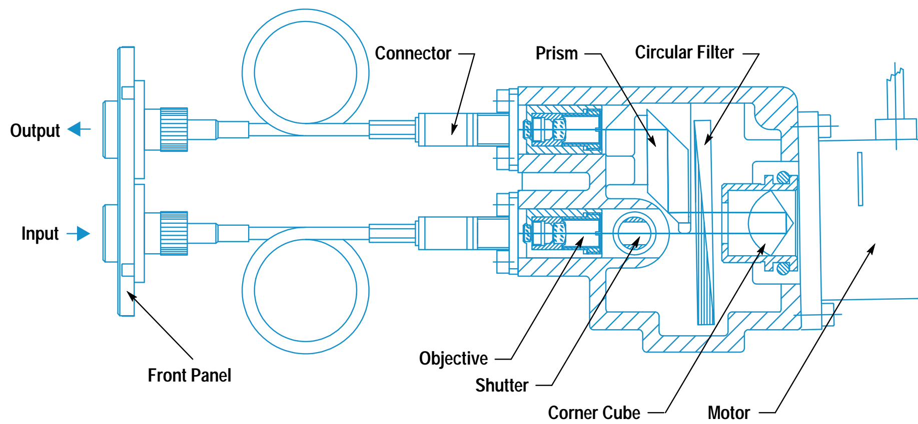


Fig. 2. Optical system of the HP 8156A optical attenuator.

Any errors caused by manufacturing tolerances of the attenuating filter are calibrated out. In addition, the wavelength characteristics of the filter are stored in the EEPROM. This permits the microprocessor to calculate correction factors for each combination of wavelength and attenuation. Fig. 3 shows the attenuation linearity of the HP 8156A over an arbitrarily selected 1-dB range.

To prevent beam steering effects, which cause unwanted loss changes, a double-pass design was chosen. The reason for the beam steering is that the filter disk is slightly wedge-shaped. This small wedge of 0.05 degree is necessary to prevent a resonant cavity in the filter which could cause unwanted attenuation changes when coherent laser light is used. Sending the optical beam twice through the wedged filter, once in the forward direction and once in the reverse direction, cancels the beam deviation to zero.

Polarization dependent loss effects are strongly reduced by using a glass absorbing filter instead of a filter with metallic attenuating coating. The birefringence of metallic coatings causes polarization dependent coupling loss in a fiber-to-fiber coupling of an attenuator. Glass filters exhibit no birefringence.

Performance

Typically, the polarization dependent loss (PDL) of the HP 8156A is on the order of 0.02 dB peak to peak, and the worst-case polarization dependent loss specification is 0.08 dB peak to peak (Option 101). The residual polarization dependent loss of the HP 8156A is independent of the filter position. It is caused by a very weak residual birefringence of all of the optical elements such as the lenses and the prisms. Fig. 4 shows the polarization dependent loss as a function of randomly changing input polarization states.

The HP 8156A has a low polarization-mode dispersion (PMD) of less than 4 fs, which also results from the weak birefringence of the optical components. The main contributions to the PMD come from the total reflections in the prism and the corner cube, which cause different phase shifts for the horizontal and vertical polarization vectors.

As a result of the design and the low polarization dependent loss of the HP 8156A, the attenuation is typically accurate within 0.05 dB, and the worst-case inaccuracy is 0.10 dB (Option 101). Fig. 5 shows the measured difference between the actual attenuation and the programmed attenuation over the entire attenuation range from 0 to 60 dB.

The return loss of the optical block of the attenuator is typically more than 70 dB. All optical surfaces are coated with antireflection coating. This helps reduce insertion loss and increase return loss. The optical surfaces are tilted to reduce the residual backreflections to less than 10^{-7}. The connectors at the optical block are angled and are also antireflection-coated.

Fig. 6 shows the spatially resolved return loss of the optoblock of the HP 8156A without the front-panel input and output connectors, as measured by the HP 8504A precision reflectometer. The main contribution to the return loss comes from the input and output connectors at the front panel of the HP 8156A. In the Option 101 version, straight contact connectors provide 40 dB of return loss. In the Option 201 version, angled contact connectors are used and a

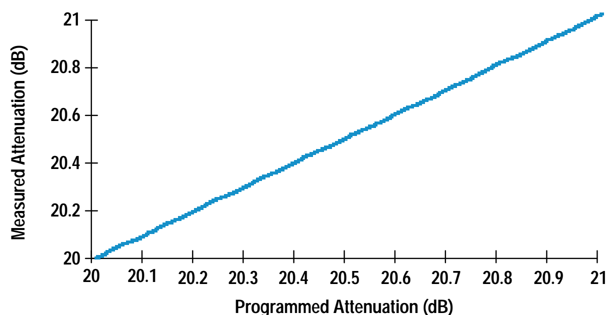


Fig. 3. Attenuation linearity of the HP 8156A optical attenuator over an arbitrarily selected 1-dB range.

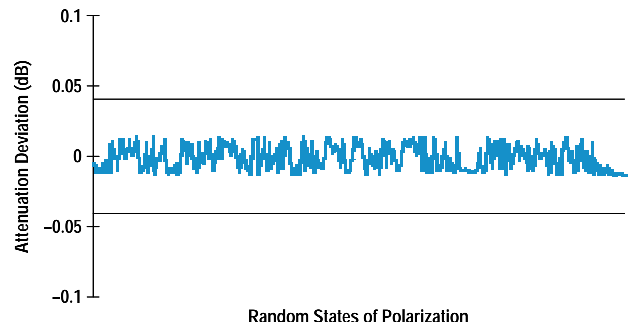


Fig. 4. Polarization sensitivity of the HP 8156A optical attenuator. The y axis indicates variations in attenuation resulting from changes in the polarization state of the light.

© Hewlett-Packard Company 1995

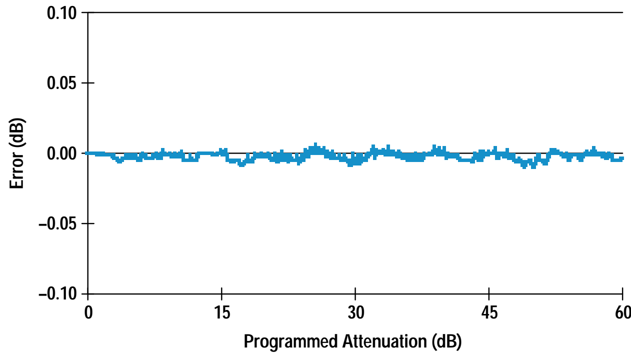


Fig. 5. Attenuation errors of the HP 8156A optical attenuator over the entire attenuation range of 0 to 60 dB (actual attenuation minus programmed attenuation).

return loss of more than 60 dB is achieved. Option 201 is the best choice for return-loss-sensitive noise measurements in cable TV applications or for bit error rate measurements in high-speed digital systems.

A separate mechanical shutter is used to interrupt the optical beam to disable the optical output. The shutter has an attenuation of more than 100 dB and works without changing the setting of the attenuating filter disk. The shutter protects power-sensitive devices under test from dangerous power levels.

The HP 8156A can be used as a calibrated and programmable backreflector to check the increase in bit error rate or noise performance of high-speed systems as a function of backreflection level. In this case the built-in backreflector mode can be used in combination with an external HP 81000BR backreflector connected to the output connector of the HP 8156A.

The HP 8156A enables the user to attenuate any optical signal up to 60 dB and up to power levels of +23 dBm in precise steps. A resolution of 0.001 dB with a typical repeatability better than 0.005 dB is achieved. Fig. 7 shows the attenuation repeatability over the entire attenuation range of 60 dB. This repeatability requires a precise filter positioning system, which is described in the following section.

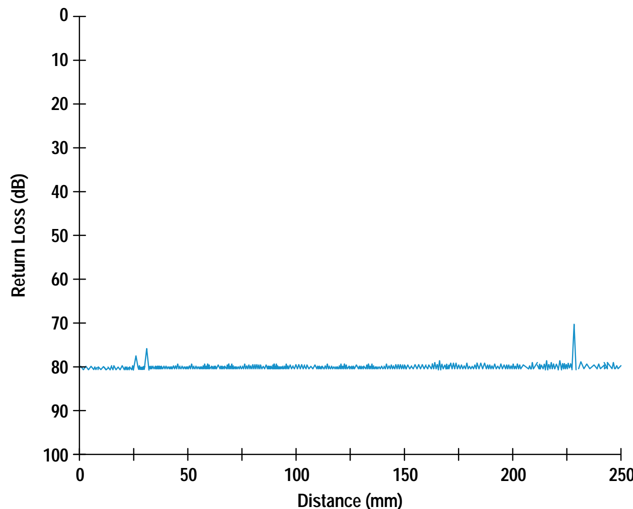


Fig. 6. Spatially resolved return loss of the HP 8156A optical attenuator measured by the HP 8504A precision reflectometer.

Filter Positioning System

For each optical attenuation setting, the filter positioning system of the HP 8156A optical attenuator rotates the circular filter to a corresponding angular position. The superior performance required of the HP 8156A in critical measurement applications leads to enhanced requirements for the filter drive regarding resolution, overshoot, and settling time.

Attenuation resolution is directly related to the angular resolution of the positioning system. The angular resolution requirements for this system can be directly determined from the filter characteristic described by equation 1. The goal of 0.01 dB attenuation resolution or a maximum resolution uncertainty of 0.005 dB results in an angular resolution of 0.009 degree, or about 40,000 data points per revolution of the filter disk.

For some applications, like bit error rate measurements, overshoot and spikes when changing attenuation settings are very disturbing. An overshoot of 0.5 dB is the upper tolerance limit for these applications. Transformed onto the circular wedge filter driven by a positioning system with an angular resolution of 0.009 degree, this results in a maximum overshoot of approximately 0.5% for an angular step of 180 degrees.

Optical attenuators are mainly used for device characterization in production areas, especially in automatic test systems. High throughput and yield are very important issues. Settling times between different attenuation settings have to be minimal, so a high-speed drive is obligatory.

Taken together, freedom from overshoot and fast settling mean a positioning system that has a strong aperiodic step response. To provide full performance under the conditions of environmental stress occurring in production and test areas, the positioning system also has to be highly insensitive to external vibration noise.

Filter Drive System

In the HP 8156A a direct-drive system is used. A motor-encoder assembly is directly attached to the filter shaft. This system has the advantages of compactness, ruggedness, and

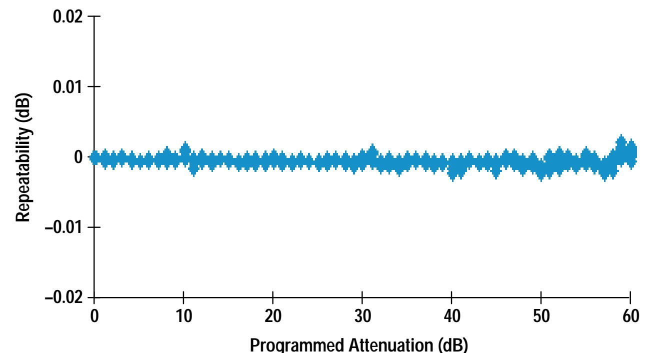


Fig. 7. Attenuation repeatability of the HP 8156A optical attenuator. The y axis indicates the variation in attenuation over ten measurements.

© Hewlett-Packard Company 1995

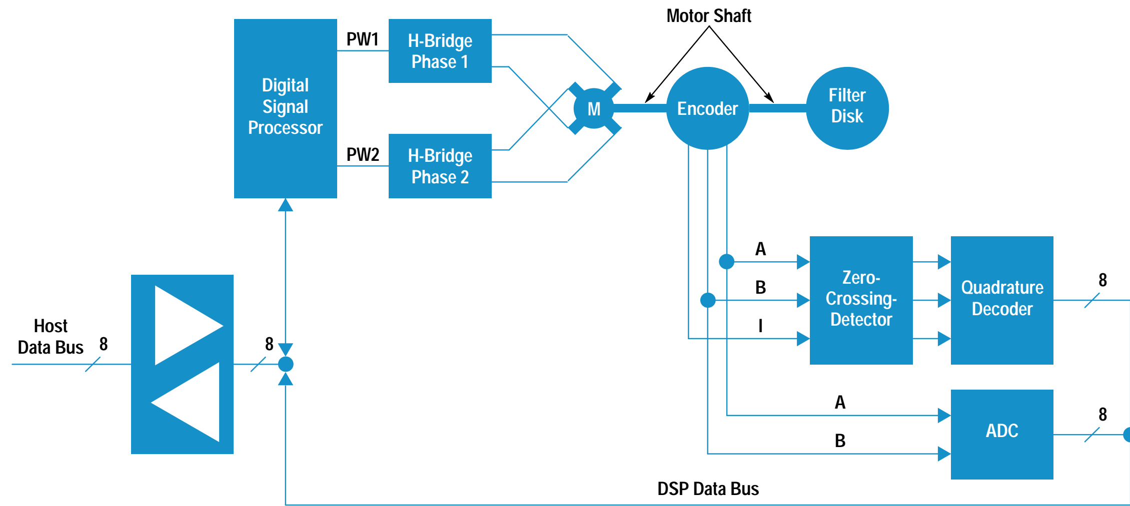


Fig. 8. Digital motor controller of the HP 8156A.

high speed compared with geared systems. There are no backlash errors and sensitivity to environmental changes is lower.

The motor is a brushless two-phase dc motor. Brushless motors combine the advantages of ruggedness, longer lifetime, and less friction, which is important in the control of high-resolution systems.

The drive currents for the two stator windings are generated by pulse width modulation of a 15V dc supply voltage. The pulse width modulation is done by two integrated full-bridge driver circuits. The MOSFET switches of these ICs have very short turn-on and turn-off times, allowing high pulse resolution, and they have very low on-resistance, which minimizes power dissipation. The pulse frequency is 25 kHz, so the acoustic noise caused by magnetostrictive effects in the motor is above the range of human hearing.

The encoder is an optical incremental rotary encoder with sine wave outputs. Its rotating disk has 1024 lines and is directly attached to the motor shaft. The main outputs, A and B, are in a quadrature relationship and their signal shape is very close to sinusoidal. A third output signal, I, provides an index pulse once per revolution for determining absolute position.

The zero crossings of the main outputs A and B are multiplied by four by a quadrature decoder circuit. The decoder output increments or decrements a coarse position counter. Both tasks are performed by a special decoder/counter IC. The resolution is further increased by an interpolation procedure that uses the sinusoidal shape of the main output signals.[1] These signals are fed directly into a multichannel 8-bit ADC, and the subsequent processing of the two digitized sinusoidal signals is done by a digital signal processor. The quotient of these signals is the tangent of the interpolated fractional position between the sine and cosine zero crossings. The filter rotation angle can be found by table look-up, using an arctangent table. Theoretically, with the eight bits of resolution in the analog-to-digital conversion process, an interpolation ratio of 256:1 can be achieved, but amplitude and phase distortion of the encoder outputs and nonlinearities of the converter limit the interpolation ratio to 64:1. This gives a total position resolution of $1024 \times 4 \times 64 = 262,144$ counts per revolution.

The servo loop is closed by a digital signal processor (DSP), a TMS320P14. This 16-bit signal processor has on-chip RAM and EPROM. It is able to work as a standalone single-chip controller. With its three high-resolution timers and a special event manager with six pulse width outputs it is ideally suited for servo control applications. The pulse width outputs feed the full-bridge drivers directly. They provide pulse width resolution down to 40 ns. This is important for achieving high-quality control because it enables the controller to measure its control effort very precisely. The DSP gets the position feedback from the decoder/counter IC and from the multichannel ADC over an external 8-bit-wide data bus. Position set values and all commands are given to the DSP by the main instrument microprocessor. Communication between the two processors takes place via an 8-bit-wide bidirectional interface.

Fig. 8 shows the block diagram of the filter positioning system.

Controller

To meet the demanding requirements of the drive, digital control is implemented, using an advanced control algorithm.

The first two considerations for the design of the controller are the control algorithm and the numerical range in which calculations take place. All digital control algorithms require one or more multiplications, and long execution times for a control step cause additional phase shift in the open control loop. Therefore, a control processor with a built-in multiplier is necessary to provide fast execution of the control algorithm.

The numerical ranges of the control variables and coefficients determine the required width of the internal processor data bus. Each filter position is described by a 22-bit number, composed of 16 bits delivered by the decoder/counter IC and 6 bits supplied by the interpolation process. For one revolution, 18 bits are sufficient. For an adequate control filter function, the filter coefficients of the control algorithm must be represented as at least 12-bit numbers. Therefore, at each control

step several 18-bit-by-12-bit multiplications have to be performed. Most of the low-cost processors offer only 8-bit-by-8-bit or 16-bit-by-16-bit multiplications in one instruction cycle. In our case, this means that it takes more than one instruction cycle for a multiplication. This means longer execution time, especially for advanced filter algorithms.

To overcome this problem the position error is divided into two number ranges: a coarse range that covers the upper 14 bits of the 16 bits delivered by the decoder/counter IC and a fine range including the 8 remaining bits (2 from the decoder/counter IC and 6 from the interpolation process). This allows fast, simple execution of the control algorithm in each range. Because of the nearly aperiodic step response of the whole system, a transition between the coarse and fine number ranges occurs only once during a position step movement.† At this transition all control variables have to be rescaled by simple shift operations. All calculations are performed in 16-bit fixed-point arithmetic with operands in twos complement representation.

The control filter function is implemented by a classical PID (proportional integral differential) algorithm enhanced with some special features. The integrator is switchable by software and works only when needed. This is a position servo, so the integrator is not used when the actual velocity magnitude is above a specified threshold. This technique strongly decreases overshoot because the integrated error signal doesn't accumulate over the duration of a move. The integrator output is limited to prevent windup effects.††

The differentiator input is not the error signal but the actual position signal, so any abrupt change in the target position signal is not differentiated. The stability of the system is not affected because the overall open-loop transfer function remains the same. Because of the high controller sampling rate, the differentiator input signal is decimated.

The PID coefficients can be switched on the fly depending on the error signal magnitude. This is necessary because the difference between static and moving friction changes the loop behavior when the system begins to move.[1] The angular resolution is so fine that the difference between the static and the dynamic cases becomes apparent. A special set of coefficients is set in the hold mode to increase robustness against external noise.

The controller function, which includes position interpolation, error ranging, PID filtering, output limiting, motor commutation, and pulse width modulation output, is implemented in the DSP, along with some special features. After executing the control routine, the DSP checks the state of settling and determines the overshoot and settling time.

In spite of the many tasks required for every control step, the execution time for one control step is a very short 150 μs. Therefore, a control rate of 4 kHz is possible. The remaining free time of the DSP is used for communication with the host processor.

† Nearly aperiodic means that the step response is generally monotonic. There is a small overshoot, but it does not cause a change in the upper 14 bits and therefore does not cause a transition out of the fine number range.

†† Windup effects would occur if the integrator output became greater than the controller output, which is limited by the pulse period. The result would be large amounts of overshoot, which is referred to as windup in rotating systems.

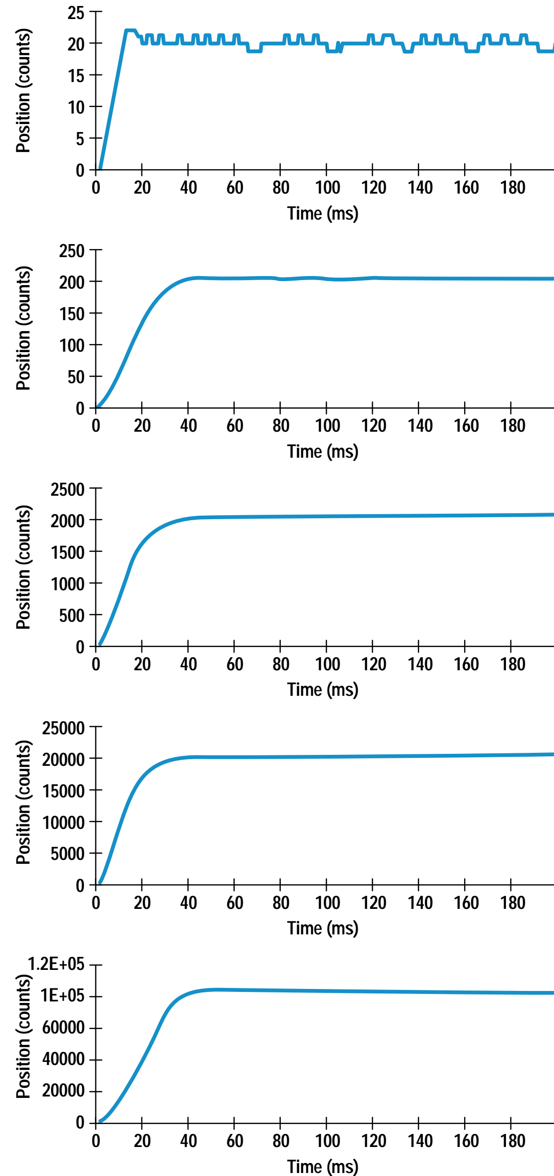


Fig. 9. Typical step response of the HP 8156A filter positioning system for various step sizes.

Well-selected PID filter coefficients are critical for the stability and step response of the servo loop. To see the step response directly, the system behavior was observed in the time domain, using a special sampling software tool developed for this purpose. A command from the host processor causes the DSP to send the variables describing the loop behavior, such as actual position and control effort, after each control step. The main instrument processor stores this data in a designated memory area. After the sampling procedure is complete, the stored data can be read by an external computer. The loop behavior is not influenced by this sampling. Using this tool, PID coefficients were found that provide a stable and strong aperiodic step response for either operating mode, ensuring that the system works well under all specified environmental conditions. Fig. 9 shows some typical step responses with different step heights.

© Hewlett-Packard Company 1995

Conclusion

Digital signal processors have been applied more and more in industrial motion control applications. Applying these processors together with a sophisticated control algorithm in an optical attenuator produces a system that provides overshoot-free, fast, and accurate positioning of the filter disk, even under noisy environmental conditions.

Acknowledgments

The authors would like to thank Wilfried Pless for project management, Michael Pott who was responsible for the design and implementation of the instrument firmware, and Rainer Eggert for the mechanical design of the optical attenuator. Special thanks to Joseph N. West at the Lightwave Operation of the HP Microwave Technology Division for his invaluable consulting and support regarding the filter drive system.

Reference

1. J.N. West, et al, "A High-Resolution Direct-Drive Diffraction Grating Rotation System," *Hewlett-Packard Journal,* Vol. 44, no. 6, December 1993, pp. 75-79.

Precision Reflectometer with Spurious-Free Enhanced Sensitivity

The HP 8504B precision reflectometer has an improved sensitivity of –80 dB at both 1300-nm and 1550-nm wavelengths. All spurious responses generated within the instrument itself have been significantly reduced. The instrument offers fiber-optic component designers and manufacturers the ability to pinpoint both large and small optical reflectances.

by David M. Braun, Dennis J. Derickson, Luis M. Fernandez, and Greg D. LeCheminant

A precision reflectometer is an effective tool for measuring the levels and locations of optical reflections in optical fiber systems.[1,2] The HP 8504B precision reflectometer (Fig. 1) uses an optical low-coherence reflectometry technique employing a Michelson interferometer with a low-coherence, broadband optical source to make spatially resolved measurements of optical reflections. One arm of the Michelson interferometer contains a translating mirror and in the other interferometer arm the device under test (DUT) is placed. When the optical path length to the mirror equals the optical path length to a reflecting surface in the DUT, the reflected signals add coherently, providing a calibrated reflectance response in the measurement trace. Measurements of reflecting surfaces within DUTs can be made in hundreds of milliseconds over distances as small as 1 mm or in tens of seconds over distances as wide as 400 mm, with a two-event spatial resolution of 25 μm at 1300 nm and 50 μm at 1550 nm.

Fig. 1. The HP 8504B precision reflectometer has enhanced –80-dB sensitivity with all spurious responses significantly reduced to greater than 65 dB below the largest reflection.

The HP 8504B precision reflectometer is an advancement of the HP 8504A, offering an improved sensitivity specification of –80 dB at both 1300-nm and 1550-nm wavelengths, with all spurious responses reduced to greater than 65 dB below the largest reflection. (A spurious response is a displayed signal that is generated within the instrument and not by the DUT.) Measurements of DUT reflections can now be made accurately across the entire measurement range without the need for interpretation of the measurement to eliminate instrument spurious responses.

A typical measurement application for a precision reflectometer is the analysis of low optical return loss in optical components and assemblies. Optical return loss is defined as the

© Hewlett-Packard Company 1995

ratio of the incident optical power to the reflected optical power in units of dB. Optical assemblies often have many internal reflections, all of which contribute to the total return loss. Precision reflectometer measurements identify which physical optical interfaces within the component are causing the greatest optical reflections and therefore are limiting the return loss. Another measurement application uses the time delay measurement capability of the reflectometer to make accurate measurements of the positions or thicknesses of elements within packaged fiber pigtailed components, the differential time delay through birefringent material, and the group index.[3]

Reflectometer Design

The key to the improved performance of the HP 8504B is a low-coherence source optimized for precision reflectometer measurements. HP has developed a family of powerful edge-emitting light-emitting diode (EELED) sources specifically designed to improve instrument sensitivity while simultaneously reducing spurious responses. Sensitivity is determined in large part by the source optical power level. The output power from the EELED was improved by the use of a long gain region. In standard EELEDs, high output power is often accompanied by large internal reflections that cause spurious responses in measurements. By careful control of these internal reflections, high output power and low spurious responses were achieved simultaneously. The article on page 43 provides a detailed description of the diode source and of how the power was increased and the internal reflections reduced.

Since all HP 8504B spurious responses were reduced, the need for the normal-sensitivity mode of the HP 8504A was eliminated. The instrument firmware was rewritten for one standard high-sensitivity mode, making the instrument easier to use and reducing factory calibration time and cost. This one mode of operation allows the simultaneous measurement of both large and small component reflections. Factory cost was an important consideration throughout the instrument redesign, resulting in a reduced manufacturing cost and a lower list price. All other important features of the HP 8504A have been retained in the HP 8504B.

Optical fiber communication systems continue to demand lower reflection levels. The improved performance of the HP 8504B addresses these increasing demands. The remainder of this article presents three measurement examples that illustrate different applications of the reflectometer.

Optical Connector Endface Characterization

Undesired reflections caused by contamination on connector endfaces are a serious concern for optical fiber systems. Contamination can decrease connector return loss and damage the endface.

Quantitative evaluation of endface cleanliness has been a difficult task. Visual inspection, surface interferometry, and surface profilometry are unable to give quantitative measurements of the small particles or films on connector endfaces. Because of its high sensitivity, the HP 8504B precision reflectometer can measure the small reflections caused by contaminants.

A setup for measuring endface cleanliness is shown in Fig. 2. The small reflections caused by contaminants were measured using a source wavelength of 1550 nm. The DUTs were custom ST connectors with endfaces polished to an 8° angle. When cleaned properly, these components can provide an optical return loss between 60 and 70 dB. Although there are multiple reflection sites in the test setup, the HP 8504B was set to examine only the connector endface of interest. The ability to distinguish between the reflection from the angled connector and the reflections from the other connectors and the ability to measure the reflection from the angled connector very accurately made the HP 8504B an excellent choice for this measurement application.

This measurement technique was used to measure the effectiveness of various cleaning processes on our manufacturing line. An optical test system began producing inconsistent results that seemed to be related to changes in the swab material used in the cleaning procedure. An evaluation was conducted examining the three materials commonly used in cleaning connectors: “lint-free” cotton, polyurethane foam, and polyester. Approximately twenty connectors were cleaned with propanol, rubbed with “lint-free” cotton swabs, and blown dry with filtered dry compressed nitrogen. Then the optical return loss was measured with the HP 8504B. The same group of connectors was cleaned again, this time using polyurethane foam swabs. The return loss was then measured again. This procedure gave a performance comparison between “lint-free” cotton swabs and polyurethane foam swabs. This process was repeated with a different batch of cables, using the cotton and polyester swabs. The measurement results are shown in Fig. 3.

The contamination clearly increased with the polyurethane foam swabs as indicated by a decrease in optical return loss. The average return loss degraded from 59.5 dB when cleaned with cotton swabs to 50.7 dB when cleaned with polyurethane foam swabs. The variability of cleanliness, measured by the standard deviation of the reflection, also worsened dramatically from 1.7 dB when cleaned with the cotton swabs to 8.9 dB with polyurethane foam swabs. Visual examination under a microscope indicated some evidence that polyurethane foam swabs may have left a film on the connector endfaces.

The polyester swabs performed much better. Here the average return loss for the connectors cleaned with cotton swabs was 60.3 dB and actually improved slightly to 61.7 dB with

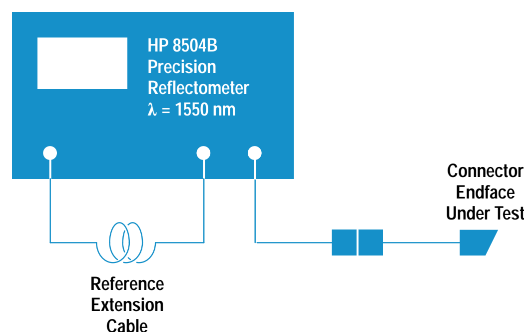
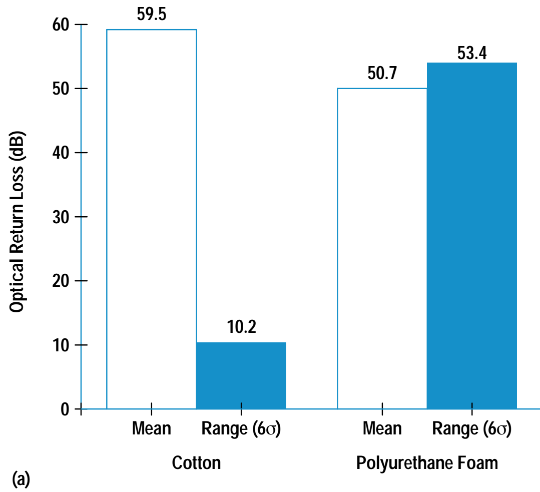


Fig. 2. Block diagram of a test setup for measurement of optical connector endface contamination. The system consists of an HP 8504B operating at 1550 nm with a single-mode fiber connected to the test port. The single-mode fiber is terminated with an ST connector that has been polished at an 8° bevel.

© Hewlett-Packard Company 1995



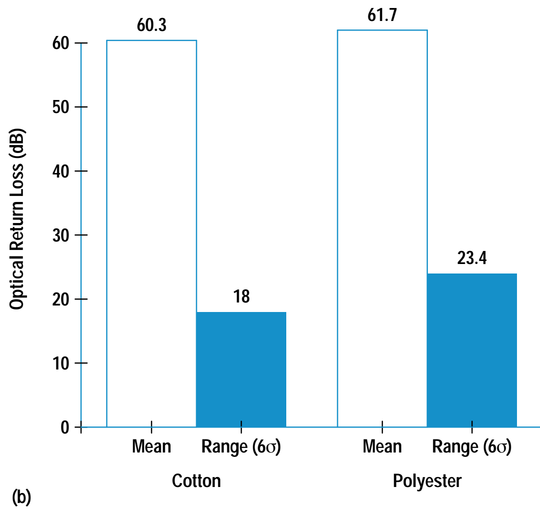


Fig. 3. Optical return loss measurement results presented as average reflectivity and range (6σ) values for cleaning procedures comparing (a) cotton swabs with polyurethane foam swabs and (b) cotton swabs with polyester swabs.

polyester swabs. The standard deviation for cleaning with cotton swabs was 3.0 dB but spread slightly to 3.9 dB with polyester swabs.

This test shows that there can be quite a variation in cleanliness just because of changes in the swab material. Using the HP 8504B, the measurement problems were linked to the use of polyurethane foam swabs. A switch to cotton swabs eliminated these problems.

This measurement technique can also measure connector optical return loss repeatability. As manufacturers improve the design of angled connectors, increasing the optical return loss, the HP 8504B can be used to measure their return loss up to 80 dB.

Optical Isolator Characterization

High-performance optical isolators are used to protect sources from reflected light by transmitting light in one direction and translating light away from the return optical path in the reverse direction.[4] Large levels of reflected light can cause linewidth and power output variations in distributed feedback lasers[5] and in Fabry-Perot lasers. In addition to attenuating light from any downstream reflections the

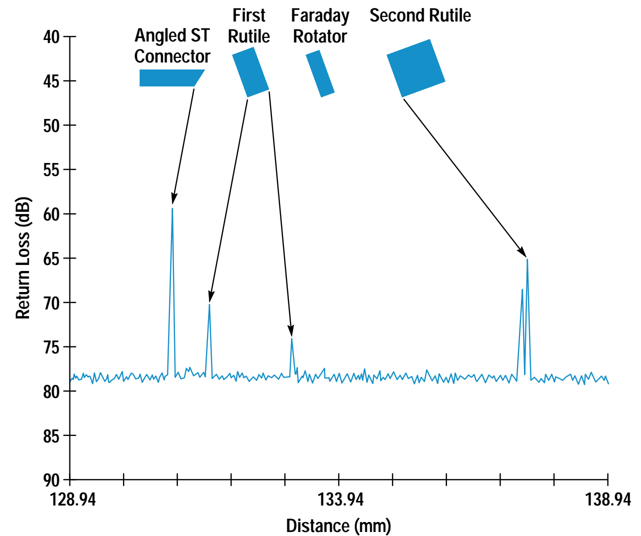


Fig. 4. A reflectometer measurement of an isolator component. Both large and small reflections can be measured with no spurious responses. Knowing the placement of the components within the isolator, the largest reflection is determined to be from the angled ST connector. A double reflection from the front face of the second rutile is observed because the first rutile is birefringent.

isolator itself must not cause reflections, or in other words, it must have high optical return loss. A precision reflectometer with high sensitivity is ideal for identifying which optical interface limits the optical return loss within the isolator.

Fig. 4 shows a measurement of an optical isolator. For this unit the highest level of reflected light occurs at the angled fiber tip with the next largest reflection occurring at the front face of the second rutile (TiO_2). With this information the designers can correctly focus their energy on the fiber as the subcomponent with the greatest potential for improving the optical return loss of the isolator. The large sensitivity and high spatial resolution of the HP 8504B enable the designer or production line to measure and identify these low reflections of the internal isolator components.

Polarization-Mode Dispersion Measurement

Polarization-mode dispersion (PMD) is a result of birefringence in both optical fibers and components. For an in-depth article on PMD see page 27.

Polarization-mode dispersion of lightwave components can be measured with a Michelson interferometer instrument such as the HP 8504B. Fig. 5 shows a block diagram of the measurement system. Placing the DUT in the reflectometer

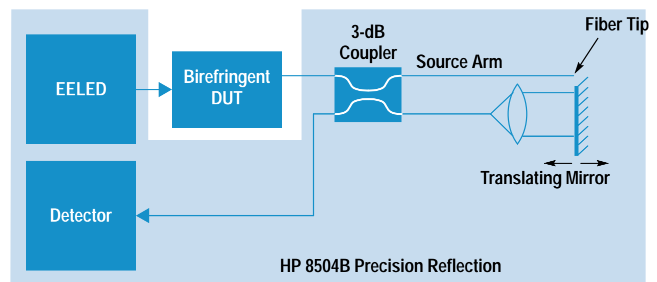


Fig. 5. Test setup used for measuring polarization-mode dispersion. The device under test is placed in the source arm of the HP 8504B.

© Hewlett-Packard Company 1995

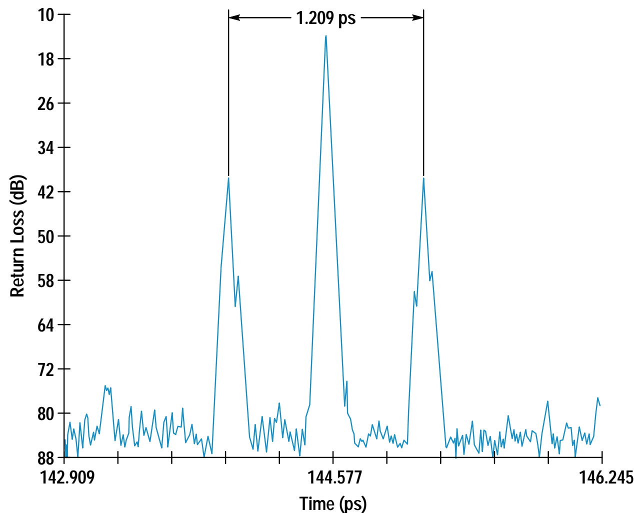


Fig. 6. Measured plot of a birefringent isolator. Three characteristic reflections are observed giving a polarization-mode dispersion measurement of 1.209 ps.

source arm allows partially polarized light from the HP 8504B EELED source to travel on both the slow and fast polarization axes of a birefringent device. The light on the slow polarization axis is delayed by an amount $\Delta\tau$ relative to the light on the fast axis. The HP 8504B produces a pair of responses when the moving mirror is positioned such that the optical path length of its arm is first longer and then shorter than the fixed-length interferometer arm by an amount corresponding to $\Delta\tau$. A response also occurs when the position of the moving mirror is such that the interferometer arms are of equal optical length, regardless of the PMD of the device.

Device PMD can be determined by configuring the HP 8504B horizontal axis in a time format and using the markers to measure the time difference between a symmetrical pair of DUT responses. Measuring between the pair of DUT responses corrects for the fact that as a reflection measurement instrument, the HP 8504B presents distance or time-of-flight measurement results in a "one-way" format. The minimum detectable dispersion is directly proportional to the HP 8504B's spatial two-event resolution and is 160 femtoseconds at 1300 nm and as low as 320 femtoseconds at 1550 nm.

This measurement technique lends itself very well to bulk optic devices such as isolators. Fig. 6 shows the measurement response at 1300 nm of an isolator placed in the source arm of the HP 8504B. The PMD value of the isolator is 1.209 ps.

The EELED source of the HP 8504B has a spectral width approaching 60 nm which means that the measured PMD is a composite over the bandwidth of the EELED. If there is significant mode coupling in the DUT, as can occur in single-mode fiber, the differential group delay (typically used to describe PMD at a single wavelength) can vary significantly as a function of wavelength. The HP 8504B measurement technique for such a DUT will not yield discrete interferometer responses. Jones matrix eigenanalysis and wavelength scanning methods, both available in the HP 8509A/B polarization analyzers, are preferred for characterizing single-mode fiber. For more on measuring PMD in single-mode fiber see the article on page 27.

Conclusion

An optical low-coherence reflectometer with large sensitivity, high spatial resolution, and an insignificant level of spurious responses is an effective tool for many measurement applications. The HP 8504B offers this measurement capability in a calibrated, easy-to-use instrument.

Acknowledgments

The development of the upgraded HP 8504B precision reflectometer benefited from the contributions of a large number of people. We acknowledge Howard Booster, Harry Chou, Mike Hart, Steve Mifsud, and Fred Rawson as members of the instrument design team and Susan Sloan, Tim Bagwell, Patricia Beck, Marilyn Planting, Joan Henderson, and Nance Andring for the edge-emitting light-emitting diode development. We thank Don Cropper, Steve Scheppelmann, and Tamas Varadi for assistance with the measurements of fiber endfaces. Finally, to Bob Bray and Jack Dupre go special thanks for their guidance and encouragement throughout this instrument development.

References

1. D.H. Booster, H. Chou, M.G. Hart, S.J. Mifsud, and R.F. Rawson, "Design of a Precision Optical Low-Coherence Reflectometer," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 39-48.
2. H. Chou and W.V. Sorin, "High-Resolution and High-Sensitivity Optical Reflection Measurements Using White-Light Interferometry," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 52-59.
3. W.V. Sorin, D.M. Baney, and S.A. Newton, "Characterization of Optical Components using High-Resolution Optical Reflectometry Techniques," *Proceedings of the Symposium on Optical Fiber Measurements,* Boulder, Colorado, September 1994.
4. K.W. Chang, S. Schmidt, W.V. Sorin, J.L. Yarnell, H. Chou, and S.A. Newton, "A High-Performance Optical Isolator for Lightwave Systems," *Hewlett-Packard Journal,* Vol. 42, no. 1, February 1991, pp. 45-50.
5. R.W. Tkach and A.R. Chraplyvy, "Linewidth Broadening and Mode Splitting Due to Weak Feedback in Single-Frequency 1.5-μm Lasers," *Electronics Letters,* 1985, pp. 1081-1083.

© Hewlett-Packard Company 1995

High-Power, Low-Internal-Reflection, Edge Emitting Light-Emitting Diodes

A new edge emitting LED has been developed for applications in optical low-coherence reflectometry. It offers improved sensitivity without introducing spurious responses.

by Dennis J. Derickson, Patricia A. Beck, Tim L. Bagwell, David M. Braun, Julie E. Fouquet, Forrest G. Kellert, Michael J. Ludowise, William H. Perez, Tirumala R. Ranganath, Gary R. Trott, and Susan R. Sloan

This article describes a new edge emitting LED (EELED) optimized for optical low-coherence reflectometry measurements.[1,2] Its use as a source for the HP 8504B precision reflectometer (see article, page 39) has resulted in improved measurement performance compared to the HP 8504A.

In optical low-coherence reflectometry, the output of a low-coherence source is coupled into an optical fiber and split in a 3-dB coupler as illustrated in Fig. 1. Half of the signal travels to a device under test (DUT), while the other half is launched into free space towards a mirror on a scanning translation stage. When the optical path length from the coupler to the mirror equals the optical path length from the coupler to a reflection in the DUT, the signals from the two arms add coherently to produce an interference pattern which is measured in the detector arm of the coupler. When the optical path length difference becomes larger than the coherence length of the source, this interference signal vanishes. The amplitude of the interference signal is proportional to the magnitude of the reflection from the DUT. Translating the mirror allows the reflectivity of the DUT to be mapped as a function of distance. With these new high-power EELEDs, the HP 8504B is able to measure return losses greater than 80 dB (reflections smaller than −80 dB) with a spatial resolution of 25 μm. In addition, reflections internal to the EELED have been reduced so that spurious signals are eliminated.

Source Characteristics

The important considerations in choosing a source for optical low-coherence reflectometry are spectral width, coherence length, spectral density, and output power. A broad spectral width implies a short coherence length and therefore a high spatial resolution capability. High output power and spectral density yield high sensitivity, that is, the ability to measure small reflections.

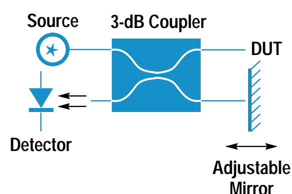


Fig. 1. Block diagram for optical low-coherence reflectometry measurements.

Many optical sources exhibit either a broad spectral width or high spectral density, but achieving both characteristics at the same time is more difficult. A tungsten-filament lamp has a very wide spectral width, but low spectral density (approximately −63 dBm/nm) when coupled into a single-mode fiber.[3] The amplified spontaneous emission (ASE) from an erbium-doped fiber amplifier (EDFA) provides a high spectral density (approximately 0 dBm/nm), but a narrower spectral width (30 nm).[4] EDFAs are relatively expensive and restricted to wavelengths near 1550 nm. Surface emitting LEDs provide moderate spectral density (approximately −45 dBm/nm) and broad spectral width (100 nm).[5]

EELEDs provide a relatively high spectral density (approximately −27 dBm/nm) with a spectral width of 40 nm to 80 nm. We chose the EELED for optical low-coherence reflectometry measurements because they offer a good balance of cost, power, and spectral width (spatial resolution).

An EELED is biased at a high drive current to achieve high output power. At elevated drive currents, undesirable internal reflections in commercially available EELEDs produce large spurious responses in optical low-coherence reflectometry measurements. Along with real reflections from the DUT, several undesired signals will be displayed, potentially confusing the interpretation of the reflection structure of the DUT.

We have succeeded in producing an EELED with reduced internal reflections, allowing high-sensitivity optical low-coherence reflectometry measurements to be made without spurious responses. The following sections describe the design and fabrication of this high-power, low-internal-reflection EELED.

Two-Segment EELED

Fig. 2 is a scanning electron micrograph (SEM) of the EELED structure. This InGaAsP/InP-based device uses a ridge structure. The ridge serves two functions in the device. It forms an optical waveguide, confining light in the direction parallel to the semiconductor surface. It also confines the pump current to be near the ridge. Fig. 3 shows a SEM and a pictorial diagram of the EELED cross section. The continuous ridge waveguide has two electrical contacts: a gain contact and an absorber contact. Under normal operation, the gain contact

© Hewlett-Packard Company 1995

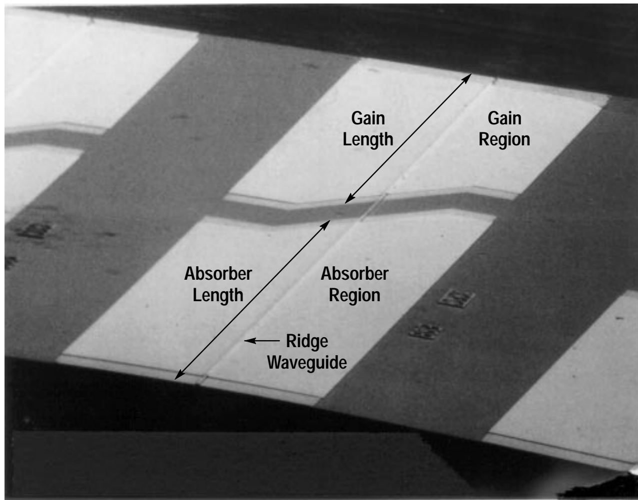


Fig. 2. Scanning electron micrograph (SEM) of the two-segment EELED.

is forward-biased and the absorber contact is reverse-biased. When a segment is forward-biased, an optical amplifier is formed. Under reverse bias, a highly absorbing optical detector is formed. The gain section generates amplified spontaneous emission (ASE) power at the device output facet from the forward traveling wave. The absorber section prevents the reverse traveling wave from reflecting off the back facet of the device. If the waves were allowed to reflect and circulate, the device could lase.

Organometallic vapor phase epitaxy (OMVPE) is used to grow the EELED epitaxial structure on InP (100) substrates. The first layer is an n-doped InP buffer. The light-emitting active region consists of an $In_{1-x}Ga_xAs_yP_{1-y}$ quaternary layer. The x and y fractions are chosen to provide the desired bandgap while maintaining lattice match to the InP substrate. Active region compositions for both 1300-nm and 1550-nm wavelengths have been developed. Zinc-doped p-InP is then grown above the active region. A heavily zinc-doped graded-bandgap layer is used to maintain low contact resistance to the top of the ridge.

The ridge location is first defined by a photolithographic process. The surrounding material is etched away in a CH_4/H_2 reactive ion etching chamber to a predetermined depth, stopping above the light-emitting active region. An electrically insulating silicon nitride layer is applied over the ridge structure. A window is opened in the silicon nitride on top of the ridge and a metal contact is deposited to form the topside gain and absorber electrodes. The semiconductor substrate is thinned to improve heat transfer and to facilitate cleaving. A backside metal is applied to form the substrate contact. The devices are finally cleaved to form the facets of the chip. More details on the processing steps are given in the article on page 20.

Operation and Spurious Responses

Fig. 4 shows the various paths that light can take to reach the EELED output facet. Two classes of signals can reach the device output. One is desired while the other is undesired. The undesired paths cause spurious responses during optical low-coherence reflectometry measurements.

Desired Path. When the gain region is forward-biased, light is spontaneously emitted isotropically from all points under the gain section. A fraction of the spontaneously emitted light is captured in the ridge waveguide. The spontaneous emission is amplified in both forward and reverse traveling waves. The desired output from the EELED is the forward traveling light (toward the output facet) with a single pass through the gain section. Most of the amplified spontaneous emission (ASE) leaving the output facet originates from the rear of the gain section (nearest the absorber) since these spontaneously emitted photons receive the largest amplification. An equal amount of ASE originates primarily at the forward end of the gain section and travels toward the absorber contact. An ideal absorber region prevents all of the reverse traveling ASE from reflecting off the back facet and becoming forward traveling energy. At the high pump currents used to drive the EELED, the absence of an absorber region could allow the device to lase. The source would then no longer have a broad, incoherent spectral width.

Undesired Paths. An undesired output occurs when photons reach the output along more than one path. These undesired paths start with a reflection from the output facet of the device. Although the output facet is antireflection-coated, such coatings are not perfect and some signal will be reflected. The signals reflected at the output facet travel back through the gain region toward the absorber section. An undesired output occurs if this signal is reflected a second time and travels back toward the device output. These twice-reflected signals will add to the desired output of the EELED and cause a spurious response. Possible sources of secondary reflections are distributed reflections along the length of the ridge waveguide, discrete reflections at the gain-to-absorber

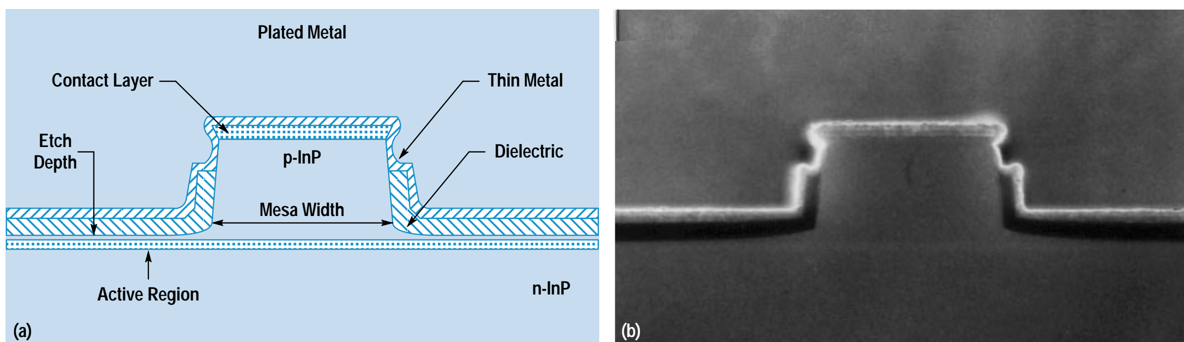


Fig. 3. (a) Cross section and (b) pictorial view of the ridge waveguide.

© Hewlett-Packard Company 1995

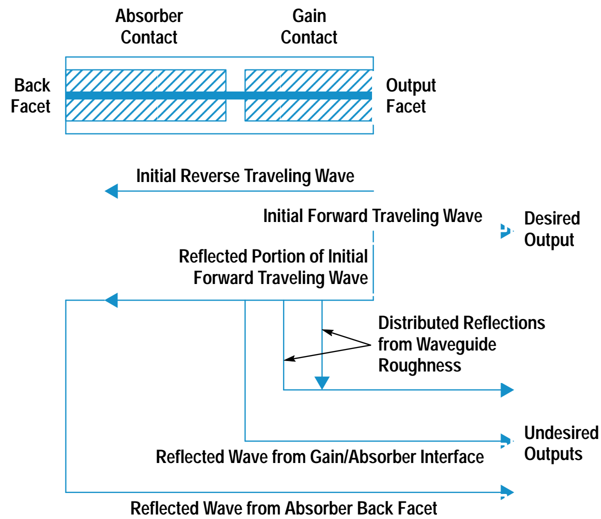


Fig. 4. Optical signal flow chart for the EELED.

interface, and discrete reflections at the back facet of the EELED.

If a source that has a large number of internal reflections is used to make an optical low-coherence reflectometry measurement, the resulting information can be confusing. Fig. 5 shows an optical low-coherence reflectometry measurement of a fiber-to-air reflection using an EELED source that has a significant level of internal reflections. This device is biased to deliver an output power of 40 μW. Ideally, the display would show only a single 15-dB return loss spike at the glass-to-air interface. The actual display shows several other spurious signals that arise from internal reflections in the device. The cause of each spurious response is identified in Fig. 5. The spurious responses appear symmetrically located about the true fiber tip reflection. The spacing between the spurious response and the fiber tip reflection is equal to the optical distance between the two reflection points in the undesired paths. Twice-reflected signals make up to three passes through the gain region, compared to one pass for the desired signal. Because of the three gain passes, the undesired signals rise very quickly as the gain and consequently

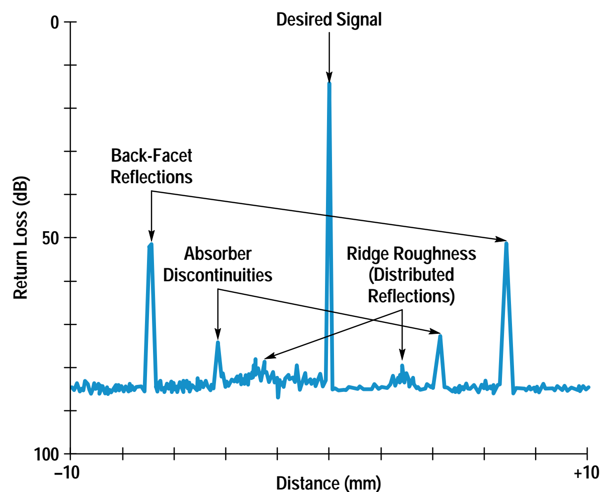


Fig. 5. Example of spurious signal generation in optical low-coherence reflectometry using a 1300-nm EELED source with a high level of internal reflections.

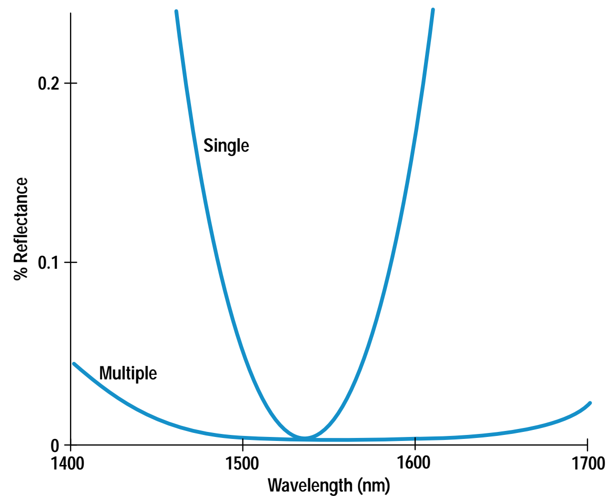


Fig. 6. Antireflection coating reflectivity versus wavelength for single-layer and multiple-layer designs.

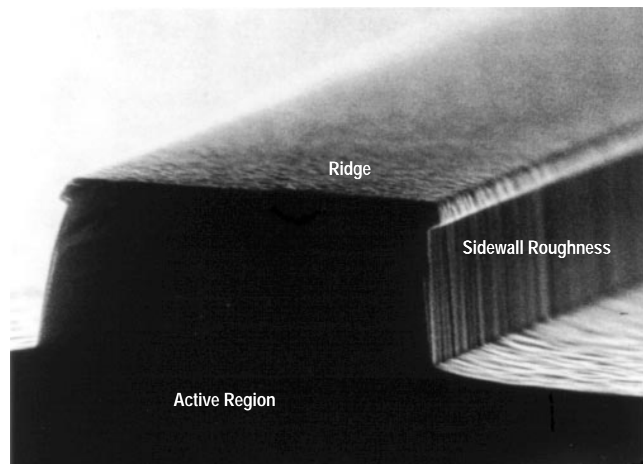


Fig. 7. SEM showing sidewall roughness in the ridge waveguide.

the power from the device are increased. For every 1 dB of increased output power, the spurious signals increase their level by 3 dB. Therefore, when attempting to use a high output power it is critically important to reduce all sources of internal reflection in the EELED. In the next section we will identify methods to minimize these internal reflections.

Reducing Reflections

It is very important to reduce the reflectivity of the output facet since all of the undesired twice-reflected signals depend directly on it. The wide spectral output of this device necessitates a broadband antireflection coating. Output facet antireflection coating reflectivity curves are shown in Fig. 6. The coating is made up of multiple layers, resulting in a broader low-reflectance window than a traditional single-layer, quarter-wavelength design.

The distributed reflections along the length of the optical waveguide are caused by variations in the ridge dimensions along the waveguide length. The conditions for the reactive ion etching of the ridge were chosen to minimize these waveguide variations. Fig. 7 shows the sidewalls and sloping floor of the ridge taken after the reactive ion etching process

© Hewlett-Packard Company 1995

step. There are small vertical striations in the sidewalls, but over a larger scale the surface is very smooth. The magnitude of the distributed reflections from these ridge striations is estimated to be at a level of less than 110 dB return loss. The distributed reflection level provides a lower limit to the multiple reflection characteristics of an EELED since a small amount of waveguide reflectivity is unavoidable.

Another source of secondary reflections exists at the gain/absorber interface. Since the gain section is heavily pumped and the absorber section is reverse-biased, the carrier density changes suddenly between these two sections. The index of refraction depends on the carrier density in semiconductors.[6] This abrupt change in carrier density can cause a reflection. An unbiased region between the gain and the absorber not only avoids conduction between the two sections, but also allows the carrier density (and thus the index of refraction) to taper off gradually, reducing the reflectivity. The gain contact is angled with respect to the waveguide to reduce the carrier density gradient and to direct any remaining reflection out of the waveguide.

The absorber characteristics are very important in obtaining low-internal-reflection EELEDs. The efficiency of the absorber at the long-wavelength end of the output spectrum is small. It is possible for deleterious levels of long-wavelength light to reflect off the back facet, thereby degrading optical low-coherence reflectometry performance. The characteristics of the absorber section have been measured experimentally and are presented in Fig. 8. The top curve shows the spectral density of the light from the EELED output facet as measured by an optical spectrum analyzer. The amount of light generated at long wavelengths is enhanced because of bandgap shrinkage and heating effects in the gain segment.[7] The output from the absorber section under several bias conditions is shown in the lower curves. The difference between the output curves and the absorber curves represents the absorption of the absorber segment as a function of wavelength. The absorber section effectively attenuates the light in the short-wavelength end of the output spectrum. The longer-wavelength light from the EELED travels through the absorber segment with relatively little absorption. Transmission through the absorber at long wavelengths is reduced by increasing the magnitude of the reverse bias on the absorber section. This increase in long-wavelength absorption is caused by the Franz-Keldysh effect[8] in bulk active region devices. We have also studied the absorber sections of devices using quantum well active regions. These devices use the quantum-confined Stark effect[9] to increase long-wavelength absorption. In our present bulk active region design, we have chosen a very long absorbing section. This reduces the back facet effective reflectivity to a level lower than that of the distributed reflection along the length of the gain section.

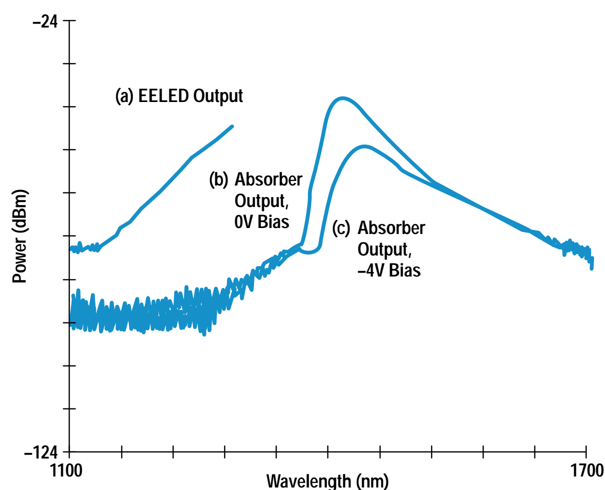


Fig. 8. (a) Gain segment output (forward traveling wave) as viewed in a 5-nm resolution bandwidth on an optical spectrum analyzer. (b) Output from the absorber section (reverse traveling wave) for a 0-volt bias and (c) a –4-volt bias.

Fig. 9 illustrates an optical low-coherence reflectometry measurement using the optimized EELED biased to produce an output power of about 40 μW, the same output power as the earlier device shown in Fig. 5. Fig. 9 shows a much lower level of internal reflections. The new EELED offers good sensitivity without the spurious responses that can be confused with real responses. If the output of the new EELED is increased to higher power (hundreds of microwatts), the spurious signals will eventually rise out of the noise floor, since spurious signals rise more quickly than the desired signals.

Power Characteristics

The EELED must produce high output power while maintaining low internal reflections. Power translates directly into increased measurement sensitivity for optical low-coherence reflectometry applications. In designing a device for high-power applications the following parameters must be considered: gain section length, ridge waveguide dimensions, active region designs, and device mounting techniques.

The output power of an EELED depends on the gain achievable in the optical amplifier (gain section). High net gain can be achieved with either a large amount of gain per unit length from a shorter segment or a smaller amount of gain per unit length from a longer segment. Fig. 10 compares the pulsed output power from EELEDs with 300-μm and 800-μm gain segment lengths. The absorber section lengths for both devices were identical. In each case, the light-versus-current

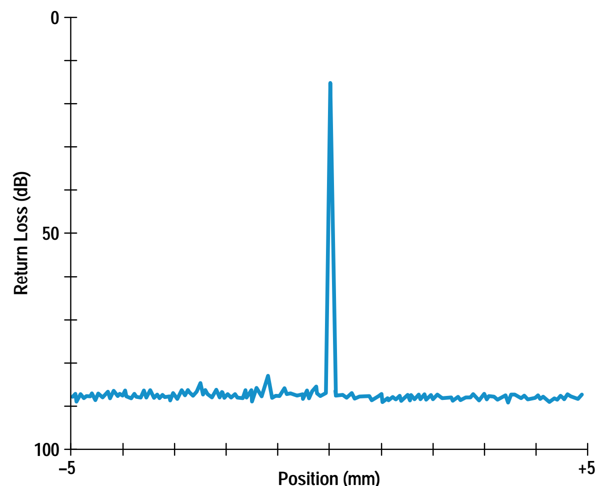


Fig. 9. Example of an optical low-coherence reflectometry measurement for an optimized 1300-nm EELED as described in this article.

© Hewlett-Packard Company 1995

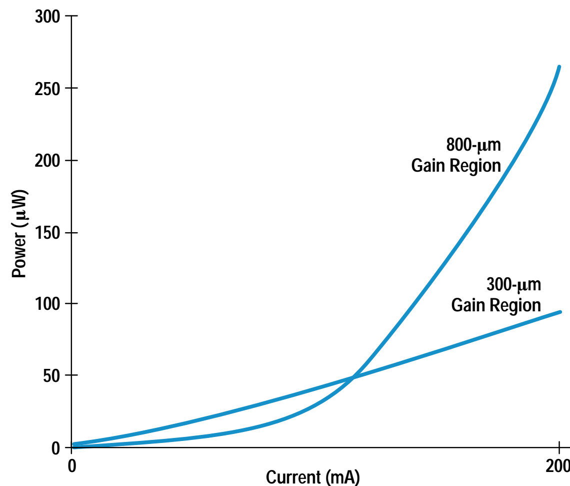


Fig. 10. Pulsed output power into a single-mode fiber versus current for 1550-nm EELEDs with gain section lengths of 300 µm and 800 µm.

curve is initially superlinear, meaning that the slope of the curve increases with drive current. At low current levels, the waveguide under the gain contact experiences a net loss per unit length, so that only the spontaneously emitted light near the output facet actually leaves the device. As drive current in the gain contact is increased, the device gain increases and the output power increases. The power output will eventually be limited by optical saturation of the gain region's output (by its own ASE), by nonradiative carrier recombination effects, by heating, or by a combination of these effects.

For lower drive currents, the device with the short gain region produces higher output power. For higher drive currents, the device with the longer gain region produces significantly higher output power. Under direct current bias conditions the output power difference between the short and long gain section lengths is even more dramatic. Several factors play a role in the increased output power from the longer device. The 300-µm device must obtain a large amount of gain in a short distance, requiring a higher current density and therefore a greater temperature rise than the longer device to achieve the same gain. The carrier density is also much higher for the shorter gain region length. As carrier density and temperature increase, a larger fraction of the pumping current goes into wasted nonradiative recombination compared to the desired radiative recombination. Auger recombination mechanisms are especially important at high carrier densities in the InGaAsP/InP material system and are more prominent in 1.55-µm devices than in 1.3-µm devices.[10]

The ridge design can also have a major effect on output power. We have investigated a variety of ridge widths ranging from a few to many micrometers. The narrower devices couple light more efficiently into single-mode fiber but the wider devices produce slightly lower levels of distributed reflections along the length of the gain section because the edge roughness lies farther away from the center of the waveguide. The ridge depth (distance from the ridge top to the material left above the active layer in the field as shown in Fig. 3) is also a key parameter in achieving a high output power. The depth controls both the transverse waveguiding of the optical mode and the current confinement in the device. Fig. 11 illustrates two extremes of etching. If the etch is not sufficiently deep, the current will spread so that a significant fraction of the current will be wasted pumping parasitic regions of the device in which the optical field is small. The optical mode will not be optimally confined. If the etch is too deep, the optical mode will intersect the lossy metal contacts, causing the waveguide loss to increase and output power to decrease. Etching into the active region can also introduce recombination losses. An optimal compromise between these two extremes was reached experimentally.

The output power capability of EELEDs with both quantum well and bulk active region structures has been studied.[2] For the same gain section lengths, bulk active region designs have produced higher power.

The spectral characteristics of the EELED depend on the amount of spontaneously emitted light captured in the waveguide at each wavelength and the bandwidth of the optical amplifier. Fig. 12 shows the spectral characteristics of a 1300-nm EELED and a 1550-nm EELED. The 1550-nm device exhibits a full width at half maximum (FWHM) spectral width of 56 nm at an output power of 350 µW. The 1310-nm device exhibits a FWHM of 51 nm at an output power of 80 µW.

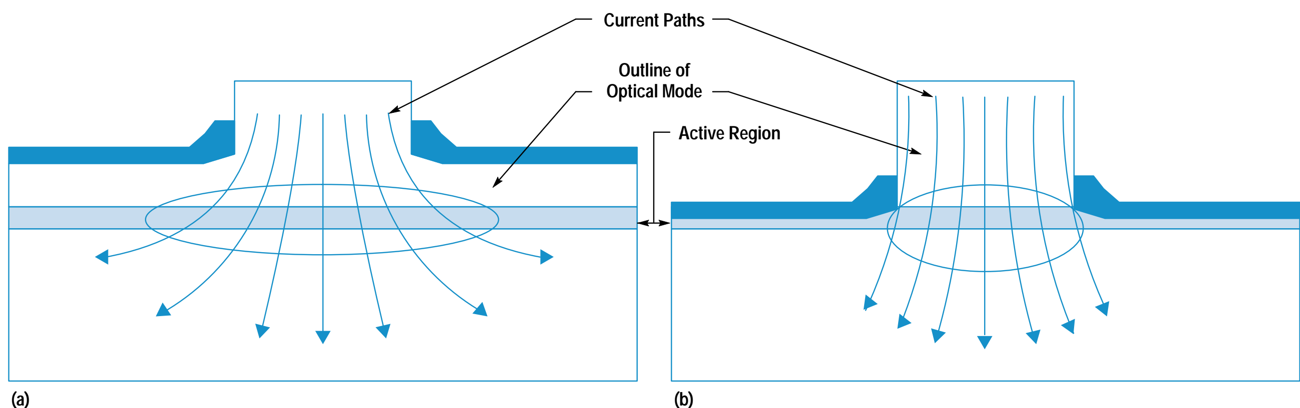


Fig. 11. Current and optical confinement for (a) shallow and (b) deep ridge etch.

© Hewlett-Packard Company 1995

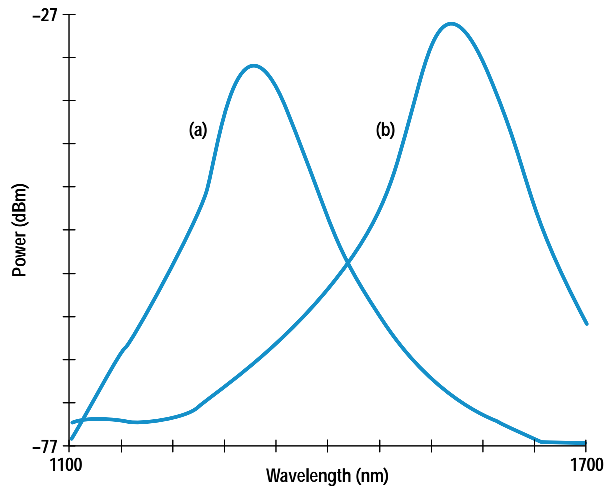


Fig. 12. Spectral characteristics of 800-μm-long gain section EELEDs at a current of 200 mA for (a) 1310-nm and (b) 1550-nm center wavelengths. The resolution bandwidth is 1 nm.

Reliability Results

Reliability of these ridge devices has been investigated, with over 6000 hours of accelerated life testing to date. Device failure is usually defined as a decrease of output power to one half of its initial value. Using this criterion, an extrapolation from measured data allows a lifetime prediction for the device. By measuring the lifetime at several temperatures we can determine the activation energy of the failure mode and infer what the actual lifetime will be at the operating temperature of the device. The lifetime of these devices is estimated to be over 800,000 hours at 25°C.

Summary and Conclusions

EELEDs optimized for optical low-coherence reflectometry applications have been developed. They are used in the HP 8504B precision reflectometer for increased measurement sensitivity and elimination of spurious signal responses in the display.

The devices have been optimized by reducing internal reflections and increasing output power. To reduce reflections we have: (1) produced smooth sidewalls on the optical waveguide during the ridge formation process, (2) extended the gain-to-absorber transition region and angled the metal contacts, (3) fabricated a long absorber section with the capability of applying a reverse bias to reduce back-facet reflections, and (4) applied a broadband antireflection coating to the output facet. The devices were also designed for increased power. The power was increased by: (1) employing a long gain section length, (2) optimizing the ridge etch depth, (3) choosing the ridge width as a compromise between efficient coupling to single-mode fiber and low distributed internal reflections, and (4) choosing a bulk active region design.

The characteristics that make these EELEDs valuable for optical low-coherence reflectometry also make them ideal sources for other applications requiring high power and low internal reflections. They could be used in fiber-optic gyroscopes, which require a very low-coherence source. In conjunction with an optical spectrum analyzer (such as the HP 71450/71451A), they can be used to measure the characteristics of optical components as a function of wavelength. Although 1300-nm and 1550-nm sources have been discussed because of their current importance to the telecommunications industry, EELEDs for other wavelengths between 1200 nm and 1700 nm can be produced by changing the epitaxy in the active region.

Acknowledgments

The authors would like to recognize the contributions of Nance Andring, Joan Henderson, Marilyn Planting, Mike Young, Henrietta Gamino, Johnny Ratcliff, and Shonna Close for processing the devices. Thanks go to Howard Booster, Bob Bray, George Patterson, Kent Carey, and Waguih Ishak for their strong support of this project.

References

1. H. Chou and W. Sorin, "High-Resolution and High-Sensitivity Optical Reflection Measurements Using White-Light Interferometry," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 52-58.
2. J.E. Fouquet, G.R. Trott, W.V. Sorin, M.J. Ludowise, and D.M. Braun, "High-Power Semiconductor Edge Emitting Light-Emitting Diodes for Optical Low-Coherence Reflectometry," *to be published in the IEEE Journal of Quantum Electronics.*
3. L.F. Stokes, "Coupling Light from Incoherent Sources to Optical Waveguides," *IEEE Circuits and Devices Magazine,* January 1994, pp. 46-47.
4. W.V. Sorin and D.M. Baney, "Measurement of Rayleigh Backscatter at 1.55 μm with 32 μm Spatial Resolution," *IEEE Photonics Technology Letters,* Vol. 4, no. 4, April 1992, pp. 374-376.
5. S.E. Miller and I.P. Kaminow, *Optical Fiber Communications II,* Academic Press, 1988, p. 467 ff.
6. J.I. Pankove, *Optical Processes in Semiconductors,* Dover, 1971, p. 89 ff and p. 392 ff.
7. G.P. Agrawal and N.K. Dutta, *Long-Wavelength Semiconductor Lasers,* Van Nostrand Reinhold Co., 1986, p. 28 and p. 86.
8. S. Wang, *Fundamentals of Semiconductor Theory and Device Physics,* Prentice Hall, 1989, p. 618 ff.
9. J.E. Fouquet, W.V. Sorin, G.R. Trott, M.J. Ludowise, and D.M. Braun, "Extremely Low Back Facet Feedback by Quantum Confined Stark Effect Absorption in an Edge Emitting Light-Emitting Diode," *IEEE Photonics Technology Letters,* Vol. 5. no. 5, May 1993.
10. G.P. Agrawal and N.K. Dutta, *op cit,* p. 23 ff.

© Hewlett-Packard Company 1995

Jitter Analysis of High-Speed Digital Systems

The HP 71501B jitter and eye diagram analyzer performs industry-standard jitter tolerance, jitter transfer, and jitter generation measurements on Gbit/s telecommunication system components. It can display both the jitter spectrum and the jitter waveform to help determine whether jitter is limiting the bit error ratio of a transmission system.

by Christopher M. Miller and David J. McQuate

Digital communication systems typically consist of a transmitter, some type of communications medium, and a receiver or line terminal unit. Typically, the digital pulses transmitted in these systems are attenuated and dispersed as they propagate through the transmission medium. To overcome signal attenuation along the transmission path, the signal may be reamplified. To overcome both attenuation and dispersion the signal may be regenerated. A regenerator receives the data stream of ones and zeros, extracts the clock frequency, then retimes, reshapes, and retransmits the digital data. Even if regenerators are not employed in the transmission system, the receiver always performs the process of extracting the clock signal to decode the data stream. Any fluctuation in the extracted clock frequency from a constant rate is referred to as jitter.

The International Telegraph and Telephone Consultative Committee (CCITT) has defined jitter as "short-term noncumulative variations of the significant instants of a digital signal from their ideal positions in time." A significant instant can be any convenient, easily identifiable point on the signal such as the rising or falling edge of a pulse or the sampling instant. The time variation in the significant instants of a digital signal is equivalent to variations in the signal's phase. A second parameter closely related to jitter is wander. Wander generally refers to long-term variations in the significant instants. There is no clear definition of the boundary between jitter and wander, but phase variation rates below 10 Hz are normally called wander.

Fig. 1 shows an ideal pulse train compared at successive instants T_n with a pulse train that has some timing jitter. The jitter time function is obtained by plotting the relative displacement in the instants versus time. Typically, the jitter time function is not sinusoidal. The jitter spectrum can be determined by taking a Fourier transform of the jitter time function.

Controlling jitter is important because jitter can degrade the performance of a transmission system by introducing bit errors in the digital signals. Jitter causes bit errors by preventing the correct sampling of the digital signal by a regenerator

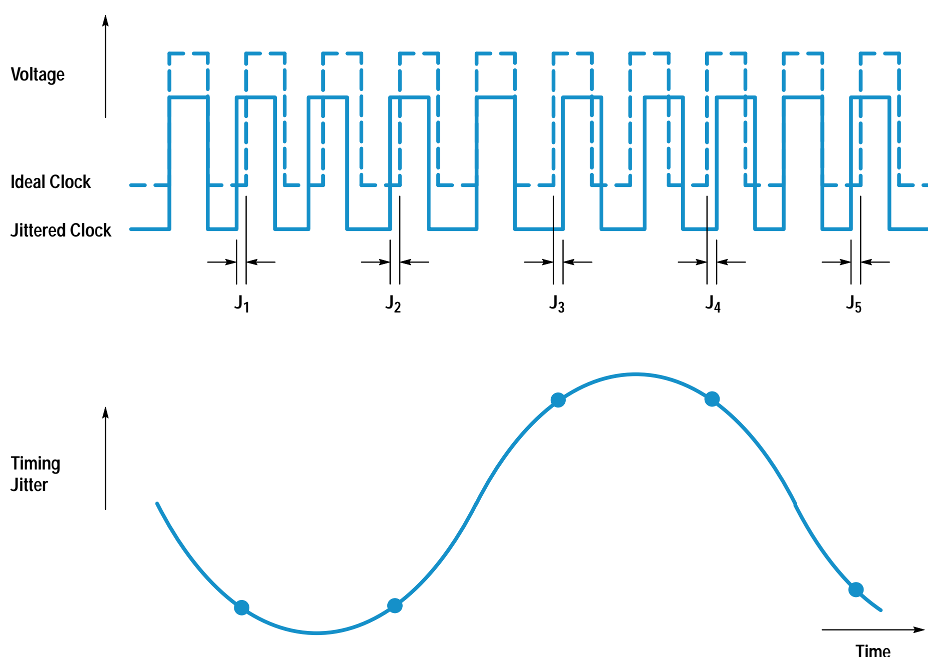


Fig. 1. Jitter time function derived from comparing a jittered clock with an ideal clock.

© Hewlett-Packard Company 1995

or a line terminal unit. Jitter can accumulate in a transmission network depending on the jitter generation and transfer characteristics of the interconnected equipment.

Jitter Measurement Categories

In an effort to standardize the high-speed telecommunication systems that are being developed and deployed, standards have been adopted for equipment manufacturers and service providers to use. Two such standards are the synchronous optical network (SONET), a North American standard, and the synchronous digital hierarchy (SDH), an international standard. Both standards are for high-capacity fiber-optic transmission and have similar physical layer definitions. These standards define the features and functionality of a transport system based on principles of synchronous multiplexing. The more popular transmission rates are 155.52 Mbits/s, 622.08 Mbits/s, and 2.48832 Gbits/s. The standards specify the jitter requirements for the optical interfaces with the intention of controlling the jitter accumulation within the transmission system.[1,2] The transmission equipment specifications are organized into the following categories: jitter tolerance, jitter transfer, and jitter generation.

Jitter tolerance is defined in terms of an applied sinusoidal jitter component whose amplitude, when applied to an equipment input, causes a designated degradation in error performance. Equipment jitter tolerance performance is specified with jitter tolerance templates. Each template defines the sinusoidal jitter amplitude-versus-frequency region over which the equipment must operate without suffering a designated degradation in error performance. The difference between the template and the tolerance curve of the actual equipment represents the operating jitter margin and determines the pass or fail status.

Each transmission rate typically has its own input jitter tolerance template. In some cases, there may be two templates for a given transmission rate to accommodate different regenerator types. Jitter amplitude is traditionally measured in *unit intervals* (UI), where 1 UI is the phase deviation of one clock period.

Jitter transfer is the ratio of the amplitude of an equipment's output jitter to an applied input sinusoidal jitter component. The jitter transfer function is also specified for each transmission rate and regenerator type. Jitter transfer requirements on clock recovery circuits specify a maximum amount of jitter gain versus frequency up to a given cutoff frequency, beyond which the jitter must be attenuated. The jitter transfer specification is intended to prevent the buildup of jitter in a network consisting of cascaded regenerators.

Jitter generation is a measure of the jitter at an equipment's output in the absence of an applied input jitter. Jitter generation is essentially an integrated phase noise measurement and for SONET/SDH equipment is specified not to exceed 10 mUI root mean square (rms) when measured using a high-pass filter with a 12-kHz cutoff frequency. A related jitter noise measurement is output jitter, which is a measure of the jitter at a network node or output port. Although similar to jitter generation, the output jitter of the network ports is specified in terms of peak-to-peak UI in two different bandwidths.

Jitter Measurement Techniques

Although these jitter measurements are made on digital waveforms, the tests themselves tend to be analog in nature. The most frequently encountered techniques to measure jitter usually employ either an oscilloscope or a phase detector. It is worth noting that there are additional jitter measurements that deal with asynchronous data being mapped into the SONET/SDH format. These tests examine the jitter introduced by payload mapping and pointer adjustments, and are performed by dedicated SONET/SDH testers.

Intrinsic data jitter, intrinsic clock jitter, and jitter transfer can be directly measured with a high-speed digital sampling oscilloscope. As shown in Fig. 2a, a jitter-free trigger signal for the oscilloscope is provided by clock source B, whose frequency reference is locked to that of clock source A. Clock source A, which is modulated by the jitter source, drives the pattern generator. The pattern generator supplies jittered data for the jitter transfer measurement to the device under test (DUT). The jittered input and output waveforms can be analyzed using the built-in oscilloscope histogram functions. The limitations of the oscilloscope measurement technique are the following. The maximum jitter amplitude that can be measured is limited to 1 UI peak to peak. Above this level, the eye diagram is totally closed. Secondly, this technique offers poor measurement sensitivity because of the inherently high noise level resulting from the large measurement bandwidth involved. Third, the technique does not provide any information about the jitter spectral characteristics or the jitter time function. Finally, the technique requires an extra clock source to provide the oscilloscope trigger signal.

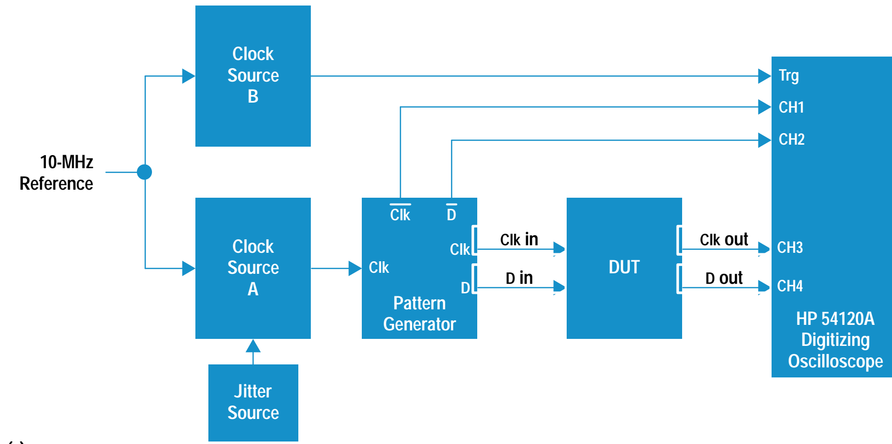
Many of the limitations of the sampling oscilloscope technique can be addressed by using a phase detector. The phase detector, in Fig. 2b, compares the phase of the recovered clock from the device or equipment under test with a jitter-free clock source. The output of the phase detector is a voltage that is proportional to the jitter on the recovered clock signal. The range of the phase detector can be extended beyond 1 UI by using a frequency divider. Intrinsic jitter is measured by connecting the output of the phase detector to an rms voltmeter with appropriate bandpass filters. A low-frequency network analyzer can be connected to the output of the phase detector to measure jitter transfer.

The phase detector method forms the basis for most dedicated jitter measurement systems. It is relatively easy to implement and provides fast intrinsic jitter measurements. However, there are several limitations. A jitter measurement system employing this technique usually consists of dedicated hardware, which only functions at specific transmission rates. In addition, the accuracy of the jitter transfer measurement with a network analyzer may be insufficient to guarantee that the specification in the standard is being met. Finally, the technique requires an additional clock source as a reference for the phase detector.

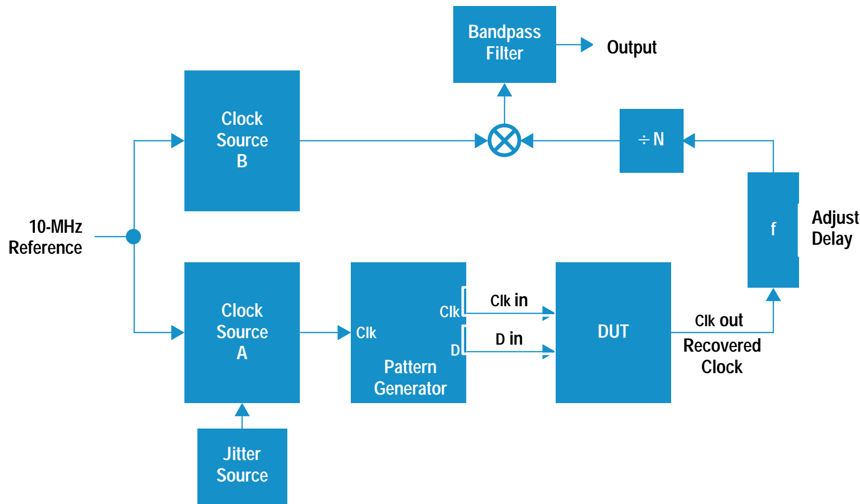
Jitter and Eye Diagram Analyzer

The HP 71501B jitter and eye diagram analyzer is a sampler-based instrument that offers a general-purpose solution to these jitter measurement requirements. To perform jitter

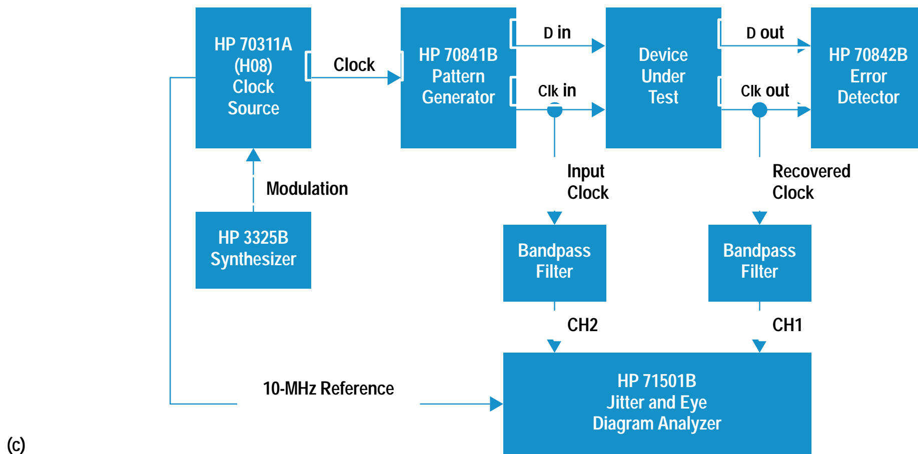
© Hewlett-Packard Company 1995



(a)



(b)



(c)

Fig. 2. (a) Oscilloscope-based jitter measurement system. (b) Phase-detector-based jitter measurement system. (c) HP 71501B jitter measurement system.

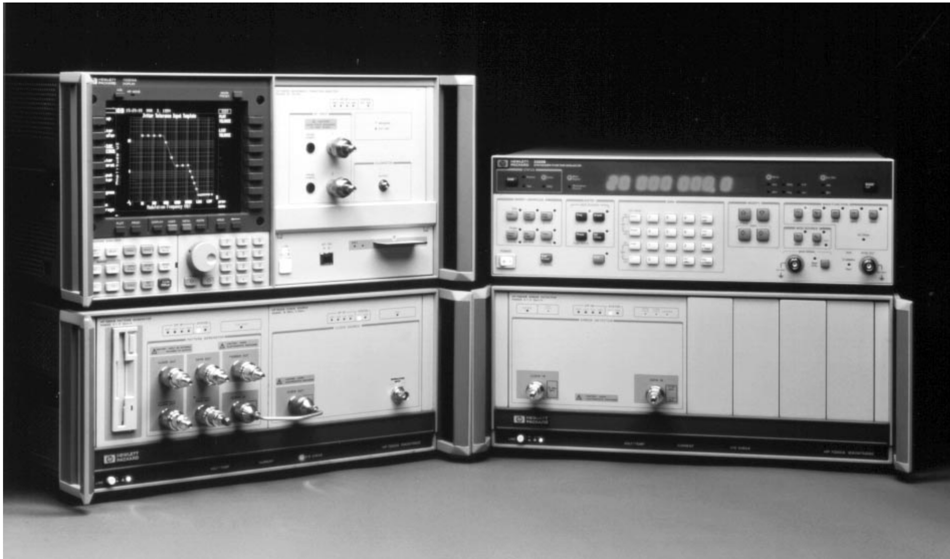


Fig. 3. HP 71501B jitter measurement system.

measurements, the HP 71501B combines the HP 70820A microwave transition analyzer module,[3] the HP 70004A color display and mainframe, and the HP 70874B jitter personality. The personality is stored on a 256K-byte ROM card and can be downloaded into the instrument. Shown in Fig. 3 is a photograph of the HP 71501B-based jitter measurement system. The system configuration, shown in Fig. 2c, includes an HP 70841B 3-Gbit/s pattern generator, an HP 70842B error detector, an HP 70311A Option H08 clock source, and an HP 3325B synthesizer, which serves as the jitter modulation source. The downloaded jitter personality allows the the HP 71501B to take control of all the other instruments in the jitter measurement system and to coordinate the measurements. This jitter measurement capability has been recently extended to 12 Gbits/s with the introduction of the HP 70843A error performance analyzer. Also, sophisticated eye diagram measurements can be made when the eye diagram personality is downloaded into the instrument.[4]

Sampler-based instruments like the HP 71501B typically operate by taking time samples of the data, then analyzing it using digital signal processing techniques. The HP 71501B has two input channels which allow it to analyze jitter transfer. Each signal processing channel can sample and digitize signals from dc up to 40 GHz, so the jitter measurement process is inherently frequency-agile. As shown in the block diagram in Fig. 4, input signals to each channel are sampled by a microwave sampler at a rate between 10 MHz and 20 MHz. The precise sample rate is set based on a determination of the incoming signal frequency and the type of measurement being made. The output of the samplers is fed into the dc-to-10-MHz intermediate frequency (IF) sections. The IF sections contain switchable low-pass filters and step-gain amplifiers. The dc components of the measured signal are tapped off ahead of the microwave sampler and summed into the IF signal separately. The output of the IF sections is sampled at the same rate as the input signal and then converted to a digital signal by the analog-to-digital converters (ADC).

Once the signals are digitized, they are fed into the buffer memories. These buffers hold the samples until the trigger

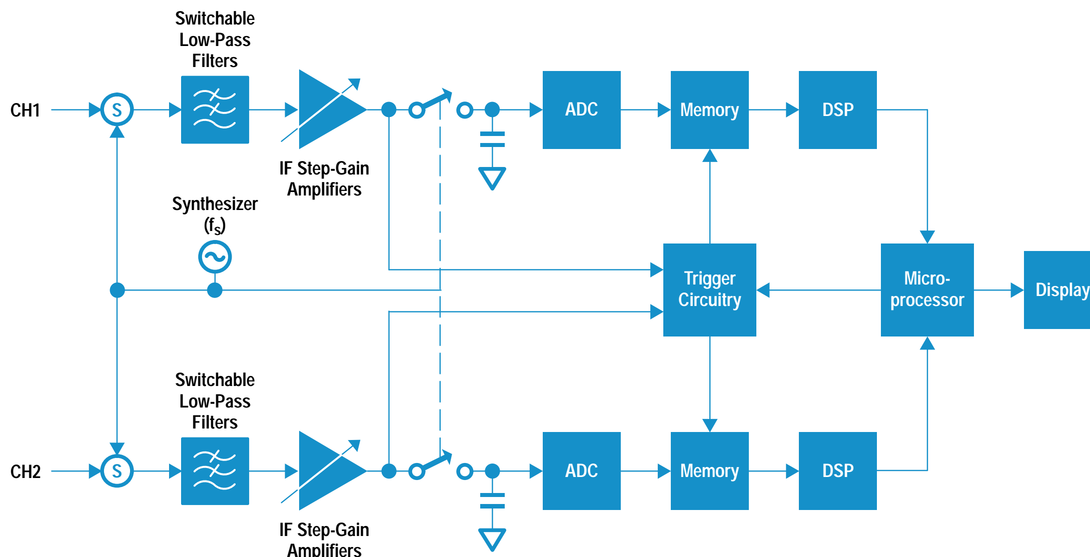


Fig. 4. Simplified block diagram of the HP 71501B jitter and eye diagram analyzer.

© Hewlett-Packard Company 1995

point is determined. By triggering on the IF signal, the HP 71501B is able to trigger internally on signals with fundamental frequencies as high as 40 GHz. Once the trigger point has been determined and all the necessary data has been acquired, the appropriate data is sent to the digital signal processing (DSP) chips. The time data in the trace memory buffer that is sent to the DSP chip has an FFT performed on it. With the time data now converted into the frequency domain, IF and RF corrections are applied to the data. The IF corrections compensate for nonidealities in the analog signal processing path. The RF corrections compensate for roll-off in the microwave sampler conversion efficiency as a function of incoming signal frequency. A Hilbert transform is then performed on the corrected frequency data to generate a quadrature set of data, and an inverse FFT is performed. This quadrature set of time-domain data is combined with the original sampled time data to form a complex-valued representation of the signal called the *analytic signal.* The analytic signal simplifies the manipulation and analysis of modulated waveforms.[5] Specifically, in this application, it is used to recover the jitter time function.

The HP 71501B can make and display measurements of the frequency spectrum or time-domain waveform of a jittered clock signal. It can also demodulate the jittered clock signal to display and perform measurements on the jitter spectrum or jitter time function. Fig. 5a shows a 2.48832-GHz clock signal displayed in the time domain. The jitter function in this example is a sinusoid at a 10-kHz rate with an amplitude that corresponds to a phase deviation of 0.25 UI peak-to-peak. This display is similar to what one would observe on a high-speed oscilloscope with the appropriate trigger signal. Fig. 5b shows the clock spectrum with jitter sidebands. This display is similar to what would be observed on an RF spectrum analyzer. Finally, shown in Fig. 5c are simultaneous frequency-domain and time-domain displays of the demodulated jitter function. As will be shown later, the measurement technique used by the HP 71501B depends on the spectral content and magnitude of the jitter time function, and the type of measurement being performed.

Measuring Sinusoidal Jitter

As previously stated, jitter is essentially phase modulation. For small amounts of sinusoidal phase modulation, a single pair of sidebands is observed, which are separated from the carrier (clock frequency) by the modulation frequency. For small values of modulation index, the magnitude of the sidebands is linearly proportional to the modulation index. As the modulation index increases additional sidebands appear and the relationship between modulation index and sideband magnitude becomes nonlinear. The amplitude of the nth sideband relative to the magnitude of the unmodulated carrier can be calculated using the nth ordinary Bessel function, with the modulation index β as an argument.

$$A_n = J_n(\beta),$$

$$\beta = \pi \times UI.$$

The modulation index is an indicator of the number of significant sidebands, significant being greater than −20 dB relative to the unmodulated carrier. The bandwidth BW of a phase modulated carrier, expressed as a function of the

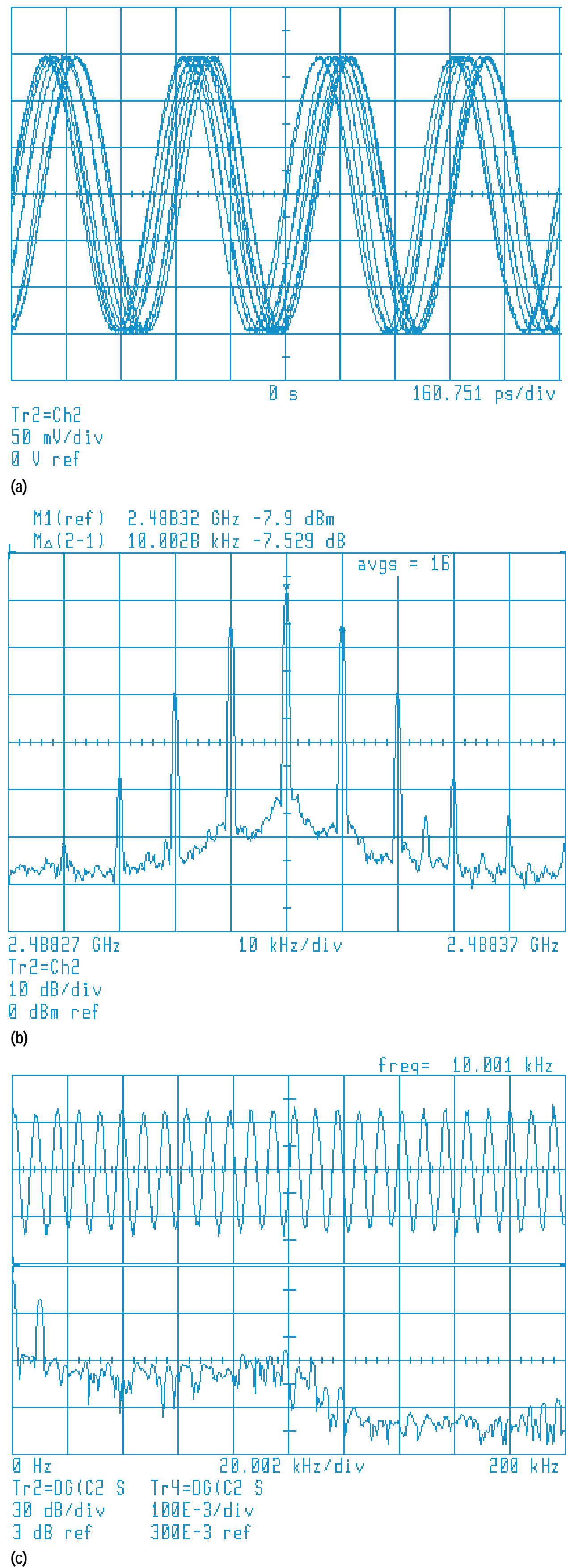


Fig. 5. (a) Jittered 2.48832-GHz clock signal. (b) Spectrum of the jittered clock signal. (c) Demodulated 10-kHz jitter waveform and spectrum.

© Hewlett-Packard Company 1995

modulation index and the modulation frequency f_{mod}, can be approximated using Carson's rule:

$$BW \approx 2f_{mod}(1 + \beta).$$

Theoretically, a bandwidth-limited signal can be accurately reconstructed if the sample rate is more than twice the signal's bandwidth. Since the HP 71501B's maximum sampling frequency is 20 MHz, accurate, unaliased sampled representations can be obtained of signals whose bandwidth is less than 10 MHz. The instrument's microwave sampler converts the jittered signal at the clock frequency, f_{signal}, down into the IF by mixing the signal frequency with a harmonic of the sample frequency, f_{sample}. This harmonic number is referred to as the comb number, N. The minimum comb number is:

$$N_{minimum} = \text{ceiling}\{f_{signal}/20\ \text{MHz}\},$$

where ceiling{x} is the smallest integer greater than x. This integer comb number is then used to compute the sample frequency for the given signal frequency with the following relationship:

$$f_{sample} = f_{signal}/[N + (\text{number of cycles}/\text{number of trace points})].$$

The jittered signal at the clock frequency would be mixed down to zero frequency in the IF without the term cycles/trace points, which is used to place the down-converted signal in the IF without having the sidebands fold over. In Fig. 6, an example is shown for a clock frequency of 2.48832 GHz, a minimum comb number of 125, trace points equal to 1024, and cycles equal to 256. The corresponding sample frequency is 19.866826 MHz. The effective IF bandwidth is

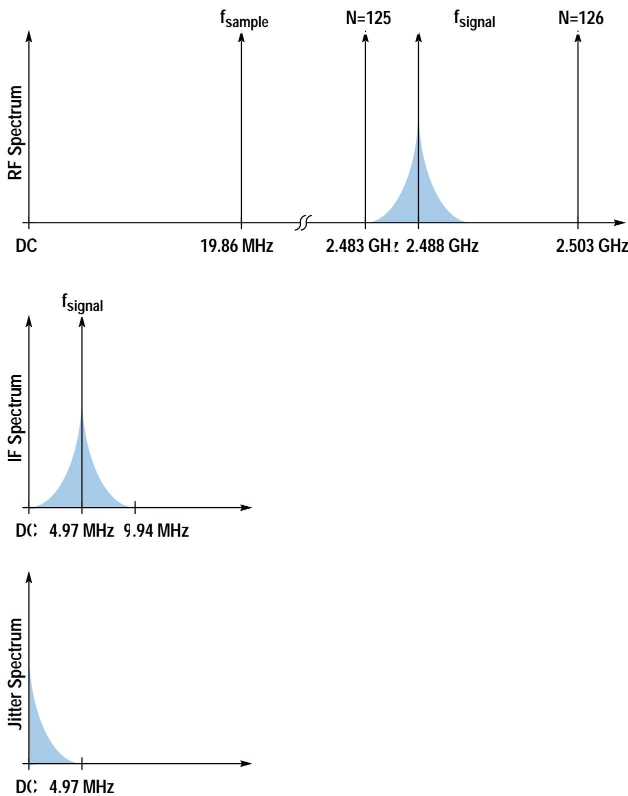


Fig. 6. Example RF, IF, and jitter spectra for a 2.48832-GHz jittered clock signal and an HP 71501B sample frequency of 19.866826

one-half the sample frequency or 9.939413 MHz in this case, and the down-converted signal is centered in the IF.

For these signals, the phase modulation waveform can be determined using the built-in DEG math function based on the analytic signal. The peak-to-peak jitter can be measured directly on this demodulated waveform. However, to improve the measurement signal-to-noise ratio, particularly for low levels of sinusoidal phase modulation, a discrete Fourier transform of the modulation waveform can be performed. The peak jitter can be determined from the magnitude of the spectral component corresponding to the modulation frequency. This component represents the energy in a small frequency band whose width is set by the window function used in calculating the transform. A flat-top window is used in the HP 71501B for single-frequency sinusoidal jitter measurements. This window function effectively serves as a resolution bandwidth filter that reduces the random noise. The resolution bandwidth RBW of the window function can be determined from:

$$RBW = \text{Window Factor} \times f_{sample}/\text{number of trace points},$$

where the window factor is 3.60 for a flat-top window. For a fixed sample rate, as the modulation frequency is reduced, the corresponding spectral component in the Fourier transform moves closer to that of the zero-frequency component. In the limit, the window functions of these two components overlap, and the magnitude of the modulation rate spectral component is contaminated. To counter this effect the comb number can be increased, reducing the sample rate and moving the modulation-frequency component away from the zero-frequency component. However, reducing the sample rate decreases the effective IF bandwidth, and thus the measurable signal bandwidth. Based on the specified jitter magnitude, the maximum comb number can be determined using the following relationship, with f_{sample} set equal to twice Carson's bandwidth:

$$N_{maximum} = \text{INT}\left\{\frac{f_{signal}}{4f_{mod}[1 + (\pi \times UI)]} - \frac{\text{number of cycles}}{\text{number of trace points}}\right\}.$$

As long as the sample frequency is greater than twice the down-converted signal's bandwidth, the transform measurement is performed. If not, the sample frequency is increased to meet the requirements of Carson's rule and the measurement is made directly on the modulation waveform. Typically, when this occurs, the specified jitter magnitude is quite large and measurement uncertainty resulting from random noise can be ignored. Often, the clock frequency, jitter modulation frequency, and magnitude are specified either by the requirements in the standard or by the customer, and in such cases the HP 71501B uses this information to determine the optimum measurement mode.

When the signal's bandwidth exceeds 10 MHz, the instrument's maximum IF bandwidth, the down-converted signal becomes aliased and the phase modulation waveform cannot be directly obtained. However, it is still possible to determine the modulation index of signals containing sinusoidal jitter. In this case, the instrument measures the signal's RF spectrum directly, setting the sample rate such that the carrier

© Hewlett-Packard Company 1995

and first-order sidebands fall at convenient places in the IF band. The magnitude of the carrier is measured, and the average of the magnitudes of the first order-sidebands is determined. Using the average of the upper and lower first-order sidebands significantly reduces any effect that incidental amplitude modulation (AM) may have on the phase modulation (PM) measurement. The modulation index β is calculated by numerically solving:

$$\frac{A_1}{A_0} = \frac{J_1(\beta)}{J_0(\beta)},$$

where A_1 is the average of the magnitudes of the first pair of sidebands, A_0 is carrier magnitude, and J_0 and J_1 are Bessel functions. Since only the carrier and first-order sidebands are measured, this technique can determine the modulation index for jitter levels up to that at which J_0 goes through its first null. This occurs at a jitter amplitude of 0.76 UI, which is about a factor of five larger than the jitter level of 0.15 UI specified in the standards for modulation frequencies that approach or exceed the instrument's maximum IF bandwidth of 10 MHz.

Jitter Tolerance and Jitter Transfer

Jitter tolerance and jitter transfer are both stimulus-response measurements. At each jitter modulation frequency, the amount of specified jitter is applied to the device under test (DUT), and its response is observed. As shown in Fig. 2c, The HP 71501B monitors the jitter level on the clock output of the pattern generator on Channel 2. The instrument uses its various sinusoidal jitter measuring techniques to adjust the amplitude of the HP 3325 synthesizer to calibrate the jitter level on the clock source to typically better than 1% accuracy. The jitter on the clock source is then transferred equally to both the data and clock outputs of the HP 70841B pattern generator. The jittered data output is applied to the DUT. The HP 71501B includes built-in input jitter amplitude-versus-frequency templates corresponding to OC-12, OC-48, STM-4, and STM-16 transmission requirements. In addition, the user can create and edit custom input templates, which can be saved to and retrieved from a RAM card. In Fig. 7, the maximum settable jitter amplitude is shown for the system relative to the requirement of the OC-48 input template. The maximum measurable jitter of the HP 71501B is 30 UI peak-to-peak for the unaliased measurements, and 0.7 UI peak-to-peak for the aliased measurements. Over most of

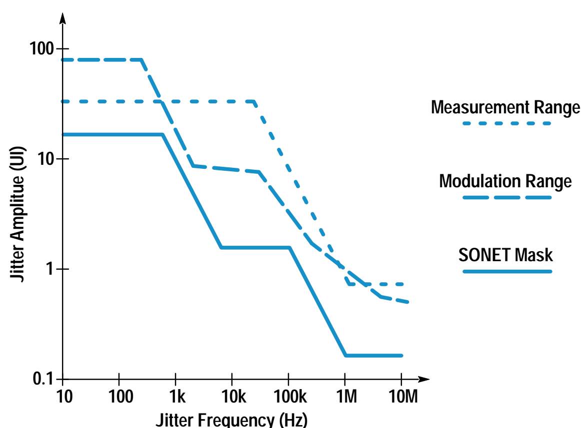


Fig. 7. Maximum sinusoidal jitter measurement range at 2.48832 GHz.

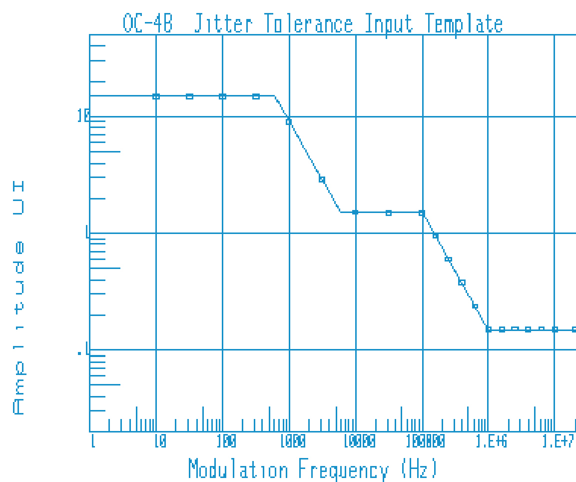


Fig. 8. OC-48 jitter tolerance measurement on a DUT.

the jitter frequency range, the maximum jitter amplitude is limited by the phase modulation capability of the HP 70311A clock source. In any case, these maximum limits are well in excess of the requirements of the standard template.

In the jitter tolerance test, the error performance of the DUT is monitored by the 70842B error performance analyzer. The aim of the test is to determine if the DUT's error performance is degraded by jitter at a specified frequency and amplitude level. Fig. 8 shows the result of a jitter tolerance measurement made on a clock and data recovery circuit. The input template was the standard template for OC-48, which corresponds to a clock rate of 2.48832 Gbits/s. Boxes correspond to measurement points that passed. Xs indicate measurement points that failed. Either bit errors or a particular error rate can be selected as the failure criterion.

In the jitter transfer test, the jitter on the DUT's recovered clock output is monitored on channel 1 of the HP 71501B and compared to the input jitter on channel 2. The ratio is then computed. This test is required to ensure that once installed in a system, these devices won't significantly increase jitter in any part of the spectrum. A cascade of similar devices, each with just a small increase in jitter, could result in an unmanageable jitter level. The standards specify a maximum value of jitter transfer of only 0.1 dB up to the specified bandwidth of the clock recovery circuit. This level has been difficult to measure accurately. The HP 71501B with its two matched input channels typically makes this measurement with an accuracy of hundredths of a dB. Shown in Fig. 9 is a jitter transfer measurement made on a DUT at OC-48. The solid line corresponds to the maximum specified jitter level at a given frequency. The boxes correspond to measurement points that passed. Xs correspond to measurement points that failed. Failures, if they occur, typically occur near the bandwidth limit of the clock recovery circuit.

Jitter Generation and Output Jitter

Both the jitter generation and output jitter tests are measurements of intrinsic random phase noise in a specific bandwidth with no external jitter applied. The HP 71501B uses the same measurement procedure for both jitter generation and output jitter, calculating both the peak-to-peak and rms

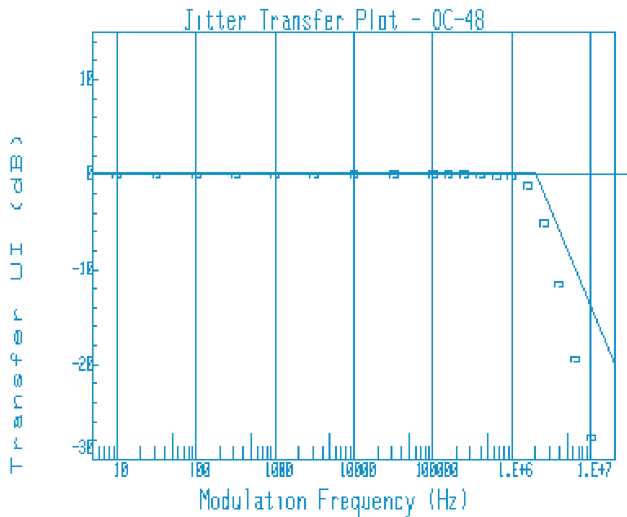


Fig. 9. OC-48 jitter transfer measurement on a DUT.

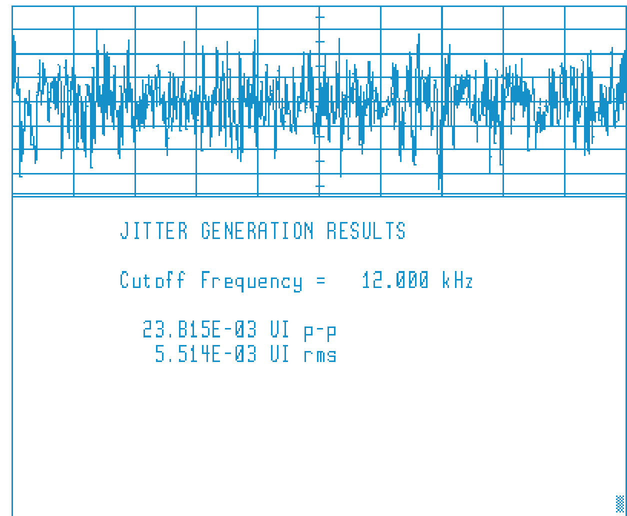


Fig. 10. OC-48 jitter generation measurement on a DUT.

jitter values. The upper end of the measurement frequency range is set by the hardware bandpass filter, shown in Fig. 2c, whose center frequency is equal to the clock rate. The low-frequency limit is implemented in the following manner. For each instrument sweep, the phase noise waveform is computed and an FFT is performed. A high-pass filter function is implemented by multiplying the appropriate elements of the Fourier transform by zero. The sample frequency is chosen so that an integer number of zeroed elements comes within the chosen accuracy of 5% for the filter cutoff frequency. This results in a sample frequency that is approximately 100 times the cutoff frequency. Since aliasing will occur if the signal bandwidth exceeds half the sample frequency, the amplitude of the jitter function is limited so that the resulting bandwidth is less than 50 times the cutoff frequency. By setting the bandwidth in Carson's rule to 50 times the cutoff frequency and working backwards the maximum measurable peak-to-peak UI as a function of frequency can be determined:

$$\mathrm{UI}_{max}(\text{peak-to-peak}) = \frac{1}{\pi} \times \left(\frac{25 f_{cutoff}}{f} - 1\right).$$

Up to 7.6 UI peak-to-peak can be measured at the high-pass cutoff frequency, with the measurable limit decreasing as 1/f at higher frequencies. The standards specify maximum limits of 0.15 UI peak-to-peak and 1.5 UI peak-to-peak for an entire bandwidth, which affords a comfortable amount of measurement headroom, as long as the intrinsic jitter spectrum falls off as 1/f or faster. Finally, the bandlimited result is transformed back into the time domain, where the peak positive and negative phase excursions are noted and the squares of all the samples are summed. When the requested number of sweeps has been completed, the rms value is calculated from the sum of the squares. Shown in Fig. 10 is a jitter generation measurement performed on an OC-48 clock recovery circuit. The specified cutoff frequency is 12 kHz and the measurement limit is 10 mUI rms.

Summary

Jitter measurements on components of high-speed telecommunication systems are necessary to ensure low-error-rate transmission and are required by the industry standards that define these systems. The HP 71501B jitter and eye diagram analyzer was designed in response to customer needs to make these measurements at the high transmission rates currently employed in optical systems. The HP 71501B can perform the industry-standard jitter tolerance, transfer, and generation measurements. In addition, its measurement technique is frequency-agile, allowing measurements to be made at proprietary transmission rates. Finally, its diverse measurement capability allows it to be used for diagnostics and jitter analysis.

Acknowledgments

The jitter analysis measurement capability was a collaboration between Hewlett-Packard's Lightwave Operation in Santa Rosa, California and Queensferry Telecommunications Operations in South Queensferry, Scotland. Initially, as a response to customer measurement needs, John Domokos wrote a preliminary application note with the help of John Wilson. Greg LeCheminant and Geoff Waters obtained additional market inputs that went into the product definition. Thanks go to Steve Peterson and John Wendler for their useful technical inputs. Finally, the jitter measurement personality was coded by Mike Manning, a software contractor from Hamilton Software.

References

1. *Digital line systems based on the synchronous digital hierarchy for use on optical fibre cables,* CCITT Recommendation G.958, 1990.
2. *Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria,* Bellcore TA-NWT-00253, 1990.
3. D.J. Ballo and J.A. Wendler, "The Microwave Transition Analyzer: A New Instrument Architecture for Component and Signal Analysis," *Hewlett-Packard Journal,* Vol. 43, no. 5, October 1992, pp. 48-62.
4. C. M. Miller, "High-Speed Digital Transmitter Characterization Using Eye Diagram Analysis," *Hewlett-Packard Journal,* Vol. 45, no. 4, August 1994, pp. 29-37.
5. M. Dethlefsen and J. A. Wendler, "Design Considerations in the Microwave Transition Analyzer," *Hewlett-Packard Journal,* Vol. 43, no. 5, October 1992, pp. 63-71.

© Hewlett-Packard Company 1995

Automation of Optical Time-Domain Reflectometry Measurements

The HP 81700 Series 100 remote fiber test system is a first-generation system consisting of a personal computer controlling one or more OTDRs and optical switches. It is well-suited for automated testing of small fiber networks such as company networks.

by Frank A. Maier and Harald Seeger

The world of telecommunications is changing very rapidly from a voice-based, nationwide network to one that is service-oriented and international. The network has to be capable of transporting voice, data, and video. New technologies like SONET, SDH, ATM, and SS#7 help meet the increasing demand for new services. However, the resulting complexity of the network calls for a higher degree of surveillance and management than the traditional network.

One of the driving forces of this multimedia age is the capability of optical fiber to support high transmission rates. Telecom operators already have large installed bases of fiber cables and are continuing to deploy fiber to meet the growing demands of their customers for new services. It is becoming increasingly important to have a more cost-effective maintenance strategy for the fiber-optic network than is in place today. The traditional method of using field-portable optical time-domain reflectometers (OTDRs) needs to be complemented with an automated testing solution. The HP 81700 Series 100 remote fiber test system (RFTS) is designed to meet this need. The RFTS helps improve overall network reliability and maintainability and reduces operating and maintenance costs.

The HP 81700 Series 100 RFTS is a first-generation system consisting of one or more OTDRs, an optical switch to share each OTDR between many fibers, and a personal computer to control these devices. The controller is capable of accessing several OTDRs through the normal telephone network by using modems.[1] The system is shown schematically in Fig. 1. It is based on the HP 8146A OTDR,[2] and the optical switch is supplied by an OEM partner. Fig. 2 shows an example of this system.

Systems of this kind provide fast, accurate fault location in case of an error and serve as a tool for preventive maintenance, allowing the analysis of long-term degradations through automatic periodic measurements. They are very effective if only a small system is required, for example for a small network of a private network operator or for a company private network. However, these systems are proprietary and do not provide full integration into the operations

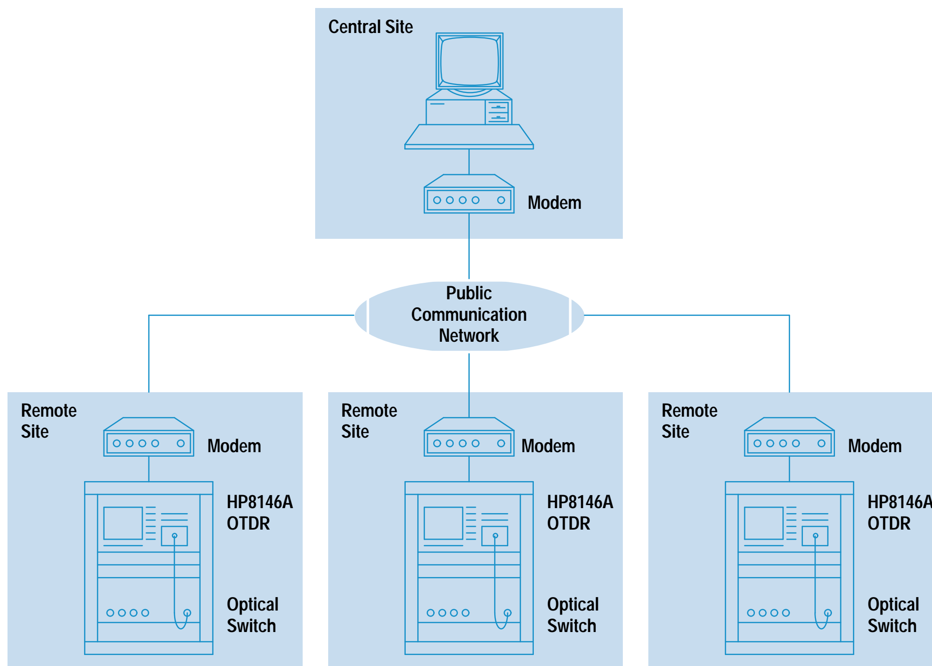


Fig. 1. The HP 81700 Series 100 remote fiber test system (RFTS) is based on the HP 8146A optical time-domain reflectometer (OTDR).

© Hewlett-Packard Company 1995

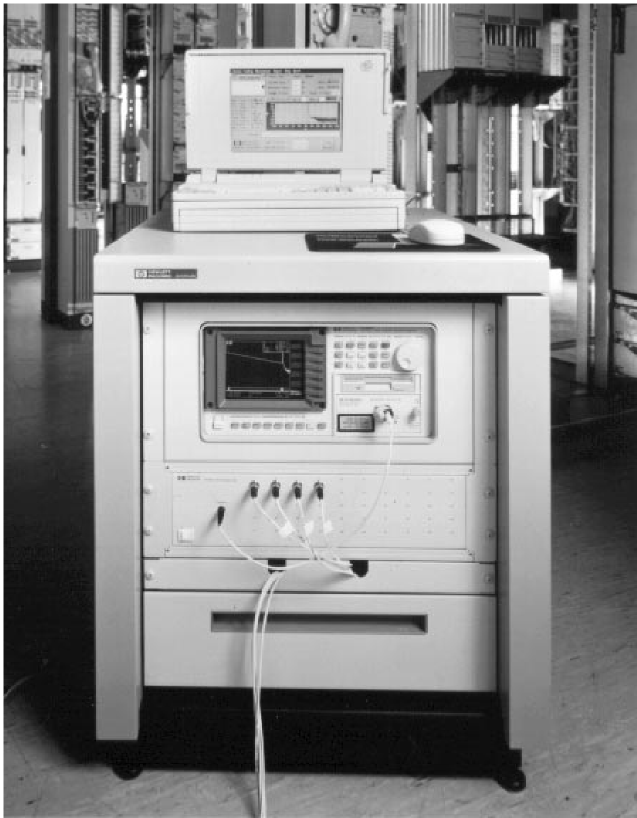


Fig. 2. A typical HP 81700 Series 100 remote fiber test system.

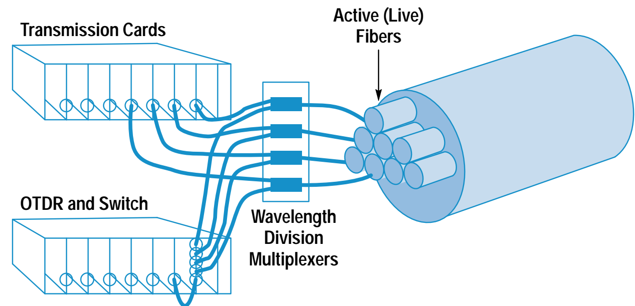


Fig. 3. Active (in-service) fiber testing using wavelength division multiplexing.

system of a telephone operating company. For this reason, a system of this kind can only be an intermediate step towards a solution of much higher complexity.[3]

Using the RFTS

There are three main methods of carrying out automated testing of a fiber link using the RFTS. The first is testing of a dark fiber, that is, a spare fiber within a cable that is never used for transmitting traffic. Bellcore states that 80% of all errors occur on the whole cable and not on an individual fiber, so with this method there is a high but not 100% certainty of catching the error.[4] With dark fiber testing there is no interference between the transmission signal and the test signal.

The second automated test method is testing of an active fiber in-service by using wavelength division multiplexing (WDM), as shown in Fig. 3. For traditional systems in which only a wavelength of 1310 nm is used for transmission, the measurement can be performed at 1550 nm. This is obviously not possible if either the transmission wavelength is 1550 nm or an upgrade to a 1550-nm system is being contemplated. In this case an out-of-transmission-band wavelength higher than 1550 nm must be used. Currently there is a trend towards monitoring at a center wavelength of 1625 nm. This wavelength is a good compromise: it is not too close to the transmission band, which reduces the requirements for the WDM wavelength separation, and it is not too high, which allows acceptable OTDR performance. (Increasing the wavelength decreases the dynamic range of an OTDR.) This method of testing the active fiber leads to 100% surveillance of the network. However, the need for WDM equipment means higher link losses and higher cost. Because the WDM equipment is nonideal, there is interference between the OTDR signal and the system signal, which must be analyzed carefully. In many applications no significant system degradation should occur. However, it may be necessary to incorporate additional filters into the link.

The third automated test method is testing of an active fiber out-of-service. This technique can be used if there is enough backup capacity so the signal traffic can be rerouted during the measurement. This method can be performed with the same wavelength as the transmission signal. It does not need WDM equipment or additional filters but only a device to combine the system signal and the measurement signal. This can be either a wavelength independent coupler, which is a low-cost device but adds a 3-dB loss to the link, or a 1-by-2 optical switch, which has less insertion loss but is extremely expensive. This method may be very suitable for a dual-ring structure because traffic can always be rerouted through the other ring. As shown in Fig. 4, the traffic that is normally routed directly from A to B can be rerouted via E, D, and C while the measurement is performed.

Overview Window

The RFTS tests the fibers of a network and informs the user if changes in the total link loss, or at specified points along the fiber, exceed the thresholds the user sets for acceptable performance. There are two levels of alarm. The first is the warning level, and the second is the failure level.

A summary of the status of the fibers under test is given in the overview window (see Fig. 5). If there are any fibers with failures these are listed at the top of the window and are marked with a red light. (When printed, the red light on the PC screen is shown as a box containing the word FAIL,

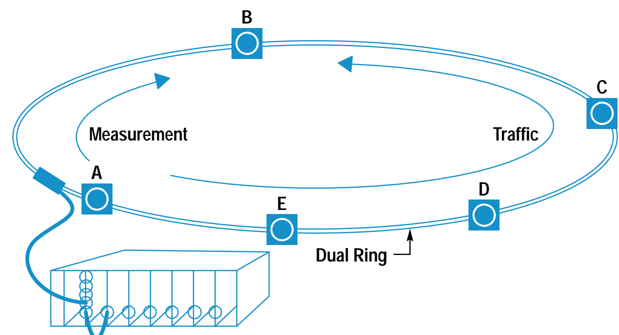


Fig. 4. In a dual-ring network, fibers can be taken out of service for testing while traffic is rerouted to the other ring (opposite direction).

© Hewlett-Packard Company 1995

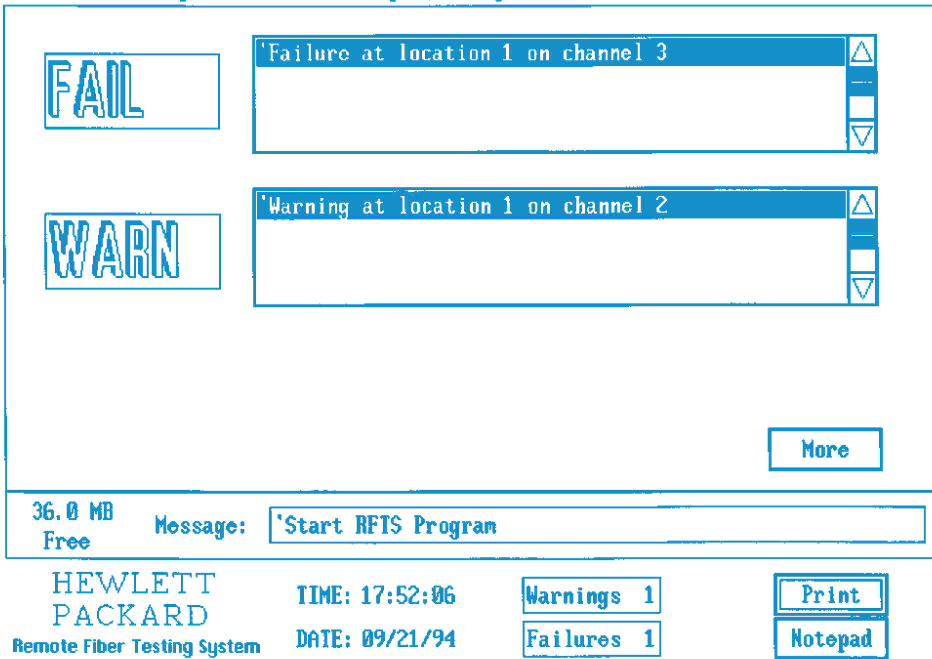


Fig. 5. RFTS overview window.

as shown in Fig. 5.) Fibers with warnings are listed in the middle of the overview window and are marked with a yellow light (printed as a box containing the word WARN). The red or yellow light starts flashing when the RFTS detects any new failure or warning. As long as no measurement result triggers any of the warning or failure criteria, a green light is displayed. Thus, the operator gets a quick picture of the overall status of the fiber network.

Report Window

Selecting More in the overview window shows the report window (Fig. 6), which lists all fibers of one specific location. The report window uses the same color scheme as the overview window: a red background for fibers with failures, a yellow background for fibers with warnings, and a green background for fibers without significant changes. The user can select any of 30 supported locations.

Trace Window

The user can examine fibers in greater detail by double-clicking on the list entry in the report window. The resulting trace window (Fig. 7) shows the current and reference measurements of the selected fiber, the fiber identification information, the overall loss of the fiber, and the measurement parameters that were used to take the OTDR measurements.

The user can set a movable marker and zoom the trace around this marker.

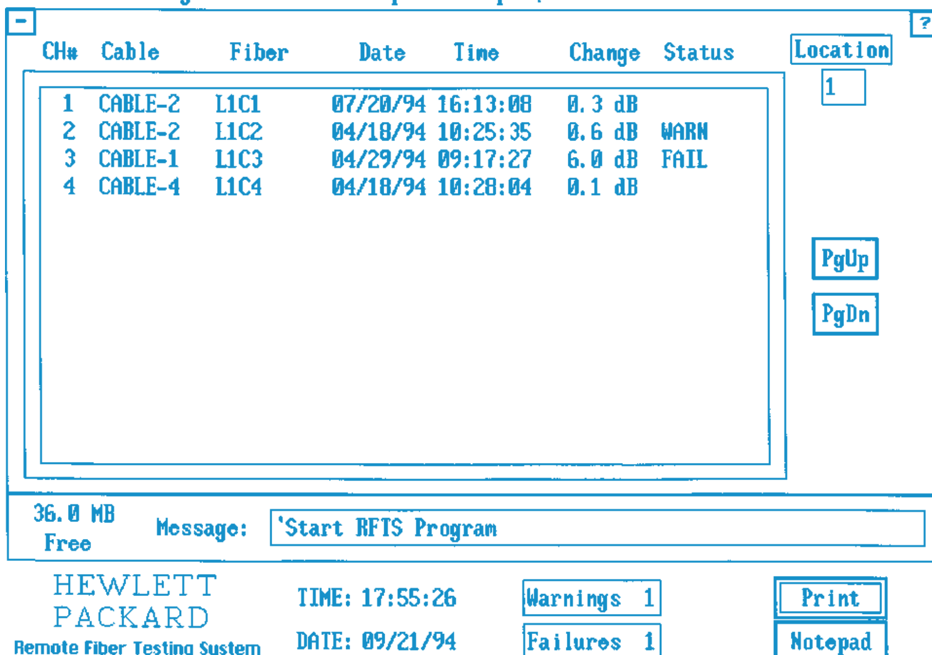


Fig. 6. RFTS report window.

© Hewlett-Packard Company 1995

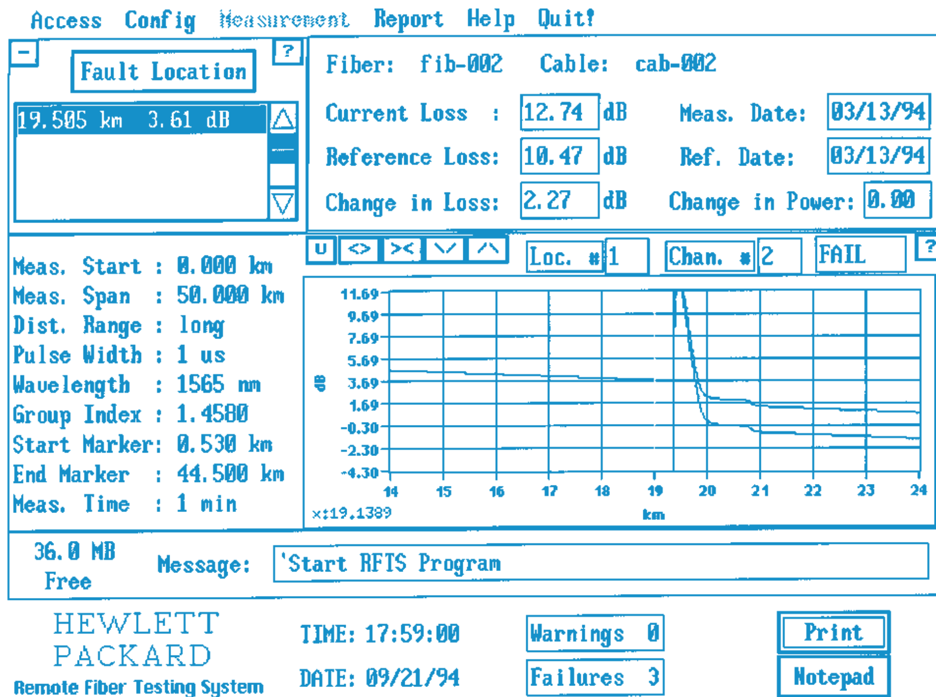


Fig. 7. RFTS trace window.

The current loss result, the reference loss result, and the change in the total link loss are shown above the traces. The loss is calculated as the difference between the power levels at two user-definable markers.

This screen also shows changes in launch power for the purpose of preventively maintaining the OTDR. This information can indicate when the laser power of the OTDR is degrading and repair is required. Without this information, the system might signal a warning or failure on a good fiber because of a problem in the OTDR.

Configuring a Fiber

The RFTS offers two user access levels: one standard user level and one password protected level. The standard user access level allows the user to monitor the system configuration and measurement results. The password protected user level allows an authorized user to configure and operate the system.

During configuration the user provides the information to link a fiber to the corresponding optical switch port at a specific location. The information for a fiber consists of the location number and the channel number of the switch, the ID of the cable to which the fiber belongs, the ID of the fiber, text fields to store names and addresses of responsible persons for installation and maintenance, text fields to store emergency procedures, the date when the channel was configured, the date when the reference measurement was taken, thresholds for warning and failure detection, OTDR parameters used to take measurements on this fiber, and information about cable access points that provides a map of the geographic positions of fiber locations or events. The user can deactivate a fiber, which means the fiber will not be measured in the future if the user starts a continuously cycling measurement or if the system starts a periodic measurement. This is helpful while a broken fiber is under repair.

Measuring Fibers

The RFTS makes two types of measurements: reference and actual. Reference measurements are those against which others are compared to determine the condition of the fiber. Typically, the user takes a reference measurement when the fiber is newly installed and is known to be in good condition. Actual measurements give the current condition of the fiber. The user must take a reference measurement before the fiber can be tested automatically.

The RFTS offers several ways to group measurements of fibers: measure a specific cable, measure all fibers at a specific location, measure all fibers at all locations, measure all fibers at all locations continuously, or measure all fibers at all locations periodically at a specific time.

Detecting Failures on Fibers

After a reference measurement has been taken, the OTDR scans the trace for anomalies. The algorithm builds an event table that lists all reflective events (connectors, mechanical splices), nonreflective events (splices), through loss and return loss values of the events, and the attenuation between two successive events. After each actual measurement, the data is compared with the reference measurement event table. Any problems identified are classified as warnings or failures.

The RFTS first calculates the total link loss of a measured fiber from the actual measurement data. The change in total link loss is tested against the channel-specific warning and failure thresholds. If the change exceeds a threshold, there are two possible reasons: degeneration of an existing event or one or more new events.

In this case the RFTS starts a new scan trace to find any new event, perhaps a new fiber end if the fiber is broken. These new events are merged into the actual measurement event

© Hewlett-Packard Company 1995

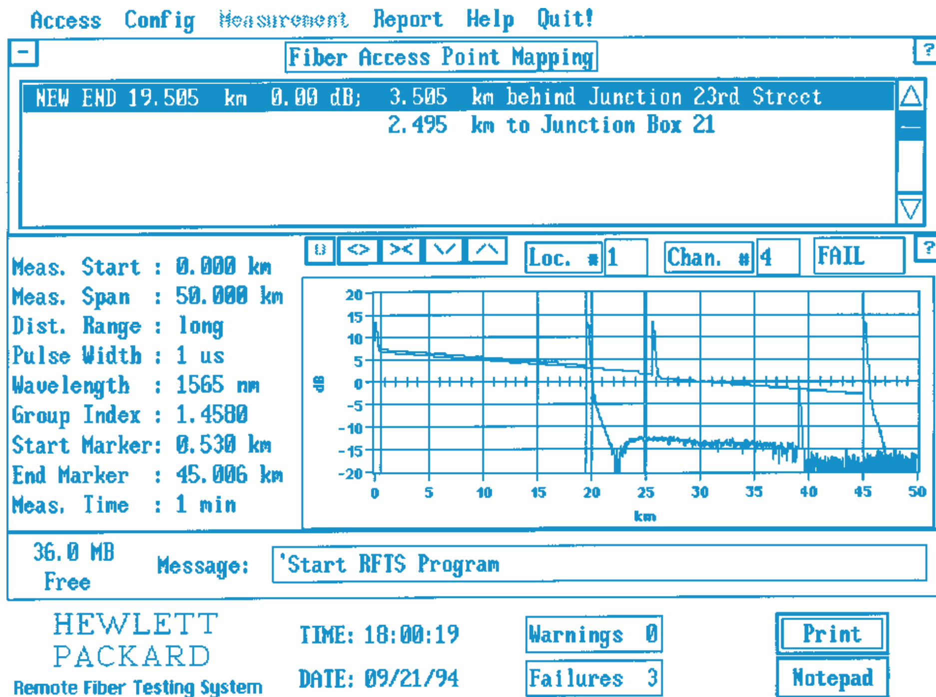


Fig. 8. Trace window with fault location box.

table. The actual event table is compared with the reference event table event by event. The loss values of all events are checked against the event loss thresholds. The user can define absolute thresholds for the through loss of nonreflective events, change thresholds for the through loss of nonreflective events, absolute thresholds for the through loss and return loss of reflective events, change thresholds for the through loss and return loss of reflective events, and change thresholds for attenuation between successive events. If any of these thresholds is exceeded, a warning or failure is signaled on this fiber.

Locating Failures on Fibers

After the RFTS has set any warning or failure on a fiber, the trace window shows a red or yellow line at the location of any event that failed one of the tests. The location and the type of a failure or warning is listed in a fault location box at the top of the trace window (Fig. 8). If the user sets up landmarks, the cable access points list box shows the fiber distance to the previous and the next landmark, if existing. The RFTS automatically prints out this information if the system is configured to do it.

History of Trace Results

Old measurement results are backed up to a separate directory before any new measurement is taken. To prevent the hard disk from being filled up with old data, the user can specify a time period after which the backup data will be deleted from the hard disk.

After a new reference measurement has been taken, the RFTS creates summary files for any measured fiber. A summary file is an ASCII file containing measurement parameters, date and time, total link loss, and event table results of the reference measurement. The total link loss and event table results of any ongoing actual measurements are appended to this summary file. Thus, the RFTS gathers historical information on the fiber network. Since the summary file is an ASCII file the user can load this information into charting programs or spreadsheets or store the information into databases.

Printing Reports

The user can print reports of the fiber status (Fig. 9). The report consists of a list of all active channels, showing for each channel the channel number, the fiber identification,

HP 81700 Series 100 Hewlett-Packard Remote Fiber Testing System

Location 1: Stuttgart

CH#	Cable ID	Fiber ID	Date	Time	Change	Status
1	CABLE-2	L1C1	07/20/94	16:13:08	0.0 dB	
2	CABLE-2	L1C2	04/18/94	10:25:35	0.7 dB	WARN
3	CABLE-1	L1C3	04/29/94	09:17:27	0.5 dB	WARN
4	CABLE-4	L1C4	04/18/94	10:28:04	0.1 dB	

Location 2: Freiburg

CH#	Cable ID	Fiber ID	Date	Time	Change	Status
1	CABLE-1	L2C1	04/18/94	10:29:32	0.8 dB	WARN
2	CABLE-2	L2C2	04/18/94	10:31:00	4.8 dB	FAIL
3	CABLE-1	L2C3	04/18/94	10:31:41	0.0 dB	
4	CABLE-1	L2C4	04/18/94	10:33:08	0.1 dB	

Location 3: Tuebingen

CH#	Cable ID	Fiber ID	Date	Time	Change	Status
1	CAB001	FIBER-1	07/20/94	18:53:36	0.0 dB	
2	CAB001	FIBER-2	07/20/94	18:55:47	0.1 dB	
3	CAB001	FIBER-3	07/20/94	18:47:52	0.0 dB	
4	CAB002	FIBER-1	07/20/94	18:50:11	0.0 dB	

Location 6: Heidelberg

CH#	Cable ID	Fiber ID	Date	Time	Change	Status
1	CABLE-1	L6F1	04/18/94	11:21:53	0.2 dB	
2	CABLE-2	L6F2	04/18/94	11:23:20	1.4 dB	WARN
3	CABLE-2	L6F3	04/18/94	11:24:49	0.4 dB	
4	CABLE-1	L6F4	04/18/94	11:26:16	0.1 dB	

Location 9: Munich

CH#	Cable ID	Fiber ID	Date	Time	Change	Status
1	CABLE-1	L9C1	04/18/94	17:52:47	0.4 dB	
2	CABLE-1	L9C2	04/18/94	17:54:15	0.0 dB	

Fig. 9. Fiber status report.

© Hewlett-Packard Company 1995

the cable identification, date and time of the last measurement, and the change in total link loss since the last reference was taken. If the measurement has crossed a warning or failure limit, the report also shows this information.

Reports can be manually printed for a single location or for all configured and activated locations. Reports can also be printed automatically once per day, once per week, or once per month. The user can configure the system to print out an alarmed trace automatically when the alarm occurs.

References

1. *HP 81700 Series 100 Remote Fiber Test System,* Hewlett-Packard publication no. 5091-8003E.

2. J. Beller and W. Pless, "A Modular All-Haul Optical Time-Domain Reflectometer for Characterizing Fiber Links," *Hewlett-Packard Journal,* Vol. 44, no. 1, February 1993, pp. 60-62.

3. F.A. Maier, "The Evolution of Fiber Analysis," *Proceedings of NFOEC 1993.*

4. J.W. Peters, "Integrated Approach to Remote Fiber Test Systems," *Proceedings of NFOEC 1992.*

© Hewlett-Packard Company 1995

Design and Performance of a Narrowband VCO at 282 THz

A single-mode optical signal source whose frequency can be voltage-controlled has been developed. We describe its design and performance.

by Peter R. Robrish, Christopher J. Madden, Rory L. VanTuyl, and William R. Trutna, Jr.

The development of extensive fiber-optic networks has increased the spectral range of communication carriers to frequencies in excess of 200 THz (wavelength = 1500 nm). Test instruments designed for use with such systems and their components often require signal sources with good frequency control and spectral purity.

A great deal of progress has been made in the development of broadly tunable optical sources with spectral linewidths less than 100 kHz. This has led to the development of commercial instruments such as the HP 8167A and HP 8168A tunable laser sources,[1] which use a semiconductor laser chip for amplification and the combination of a grating and an etalon for frequency control and tuning. In this article we describe an alternate approach to tunable laser sources that has high spectral purity and very rapid tuning.[2]

Noise characteristics of the electrically excited semiconductor amplifier place limits on the spectral purity of the signal source. The noise characteristics of the amplifying medium can be improved by using an optically pumped crystal and that is the approach taken in the design of the oscillator described in this paper. However, to achieve substantially improved spectral purity we sacrifice broad tunability. Therefore, the resulting source will be complementary to semiconductor laser sources, making it possible to address applications that require very high spectral purity within a narrow range of frequencies.

Laser Description

A laser is an oscillator operating at optical frequencies. Like all oscillators, it consists of an amplifier and a means for applying positive feedback to that amplifier. The tuning range of an oscillator is the set of frequencies for which the gain of the amplifier is large enough to compensate for losses in the system. To ensure single-frequency oscillation, one must design the feedback mechanism to allow only one frequency within the amplifier bandwidth.

In an optical oscillator, feedback is supplied by reflectors that form a resonant cavity containing the amplifier. The spacing of the reflectors determines the axial resonant modes of the cavity. Each cavity mode corresponds to a frequency for which an integral number of half wavelengths of oscillation will just fit within the cavity. Since typical laser cavities have lengths much greater than the optical wavelengths, a large number of optical modes can be defined by a particular cavity configuration. The frequency spacing $\Delta\nu$ of these modes is inversely proportional to the cavity reflector separation:

$$\Delta\nu = c/2nl,$$

where $\Delta\nu$ is the frequency spacing, c is the speed of light, n is the index of refraction of the material in the cavity, and l is the length of the cavity.

The amplifier in the system we have built is a crystal of yttrium orthovanadate, YVO_4, doped with 1.5% neodymium. The Nd atoms displace some of the Y atoms in the crystal structure and provide Nd^{3+} ions that have a set of energy levels that can be optically excited by the output of a semiconductor laser operating at 808 nm. The excited Nd ions can then emit radiation over a frequency range of about 240 GHz centered at 282 THz (1064 nm). This emission can occur spontaneously or can be stimulated by the presence of ambient 282-THz radiation, amplifying it.

Since the amplifier bandwidth is much narrower than the 9-THz bandwidth of the semiconductor chips used in the HP 8167/8A laser sources, we can use a relatively simple strategy to ensure that this laser will operate at a single frequency. If the cavity is made short enough so that its frequency spacing is greater than the emission frequency range of the Nd^{3+} ions then the condition for single-mode operation will certainly be satisfied. The index of refraction of the $Nd:YVO_4$ is about 2.1, so the cavity must be less than about 0.3 mm long to ensure single-frequency operation. For a cavity longer than 0.3 mm, one can still obtain single-frequency operation by controlling the length so that one of the cavity modes has a frequency near the peak frequency of the amplifier gain curve. However, as the cavity length is increased and the mode spacing decreases, it becomes more likely that a second mode will have enough gain to oscillate. This implies that the cavity must be as short as practical, but need be no shorter than 0.3 mm.

The requirement for a short cavity drove the choice of laser crystal. Of all Nd-doped crystals available in reasonable commercial quantities, $Nd:YVO_4$ is an ideal laser material for this application because its Nd ions exhibit a high probability for absorption of 808-nm light from commercially available diode lasers and very efficient reemission of light at 1064 nm. This means that a small amount of the material can have enough gain to reach the threshold for laser action at modest levels of optical pumping power. In addition, the emitted radiation is preferentially polarized along one of the crystal

© Hewlett-Packard Company 1995

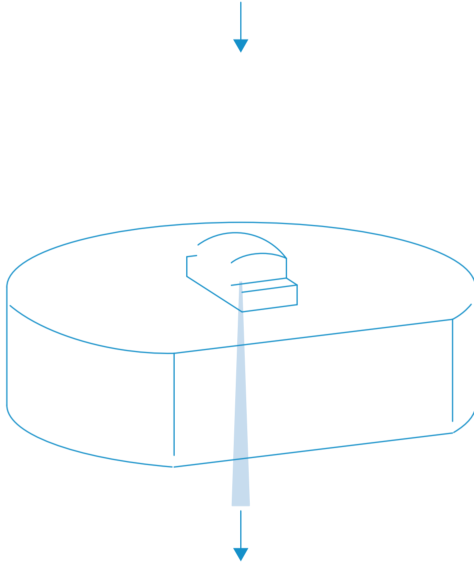


Fig. 1. Design of the narrowband laser operating at 282 THz (1064 nm).

© Hewlett-Packard Company 1995

Surface Emitting Laser for Multimode Data Link Applications

A surface emitting laser has been developed for use in a multimode optical fiber data link. The laser can operate in a high-order spatial mode, resulting in a spectral width as wide as one nanometer and a relative intensity noise (RIN) lower than –125 dB/Hz in a multimode fiber system. Electrical and optical characteristics of the surface emitting laser and the epitaxial growth methods are discussed.

by Michael R.T. Tan, Kenneth H. Hahn, Yu-Min D. Houng, and Shih-Yuan Wang

A platelet laser with light emitting perpendicular to the substrate was developed by Melngalis in 1965 at MIT Lincoln Laboratory.[1] By 1979, a pulsed double heterostructure InGaAsP surface emitting laser operating at cyrogenic temperatures was demonstrated by Professor Suematsu's group at Tokyo Institute of Technology.[2] Since the late 1980s many research groups have successfully demonstrated surface emitting lasers that were electrically pumped and operating CW at room temperature.

Why are surface emitting lasers the focus of so much work? The surface emitting laser structure is radically different from the conventional edge emitting semiconductor laser. The light emitted from the surface emitting laser is perpendicular to the substrate rather than in the plane of the substrate, as shown in Fig. 1. The optical cavity of a surface emitting laser is formed by distributed Bragg reflectors sandwiching an active layer.

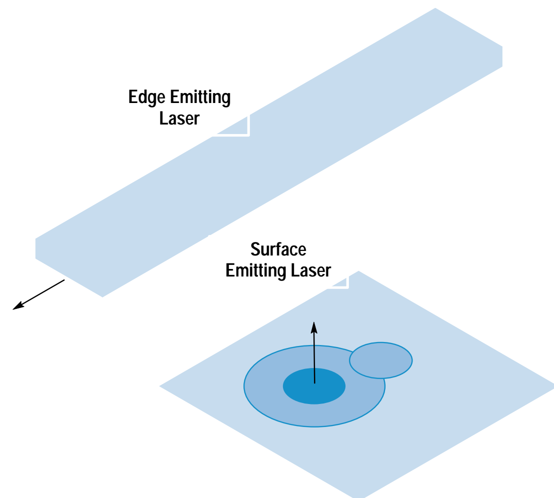


Fig. 1. Edge emitting lasers are cleaved into long bars typically 200 to 300 μm long and light is emitted from the cleaved facets. Generally, it produces an elliptical beam with a numerical aperture NA of 0.3 and 0.6. The surface emitting laser emits light in a direction perpendicular to the wafer with a circular beam and an NA as small as 0.05 for a single spatial mode. This small NA simplifies interfacing to optical fibers. In addition, the surface emitting laser can be manufactured like light-emitting diodes (LED) and is complete at the wafer level.

Fig. 2 shows the cross section of a bottom emitting laser (light emerging from the substrate) that has been developed in our laboratory. It has a hybrid Au-Bragg "back" reflector of 99.96% reflectivity (calculated) and an output mirror of 98.9% reflectivity (calculated). This configuration is amenable to high-volume manufacturing similar to light-emitting diode (LED) processing and therefore has the potential of very low cost along with high performance.

Some advantages of the surface emitting laser over the conventional edge emitting laser are: (1) the devices are completed at the wafer level and hence can be completely characterized, (2) the numerical aperture (NA) is smaller and symmetric and allows almost 100% coupling into optical fibers, resulting in simpler packaging, (3) operation is single-frequency, and (4) the structure can be integrated with monitor photodiodes or transistors, or in two-dimensional arrays as shown in Fig. 3.

Data Link Applications

High-speed optical data links for distances of under one kilometer for linking workstations, peripherals, and displays are becoming increasingly important. The optical source for such links has been the CD (compact disk) laser operating multimode or in the self-pulsating mode to broaden the spectrum to minimize the modal noise resulting from mode dependent loss in the multimode fiber system. Some limitations of the CD laser are that the laser has to be preselected for its self-pulsating characteristics and the modulation frequency is limited to approximately one third[3] of the self-pulsating frequency which is typically 1.5 to 2 GHz. A properly designed large-area surface emitting laser will not have these limitations and is an excellent light source for a multimode data link.[4,5]

Growth Method

The epitaxial layers of the laser shown in Fig. 2 were grown using a modified Varian Modular Gen II molecular beam epitaxy machine. In addition to the standard high-temperature effusion cells providing the group III sources of Al, Ga, and

© Hewlett-Packard Company 1995

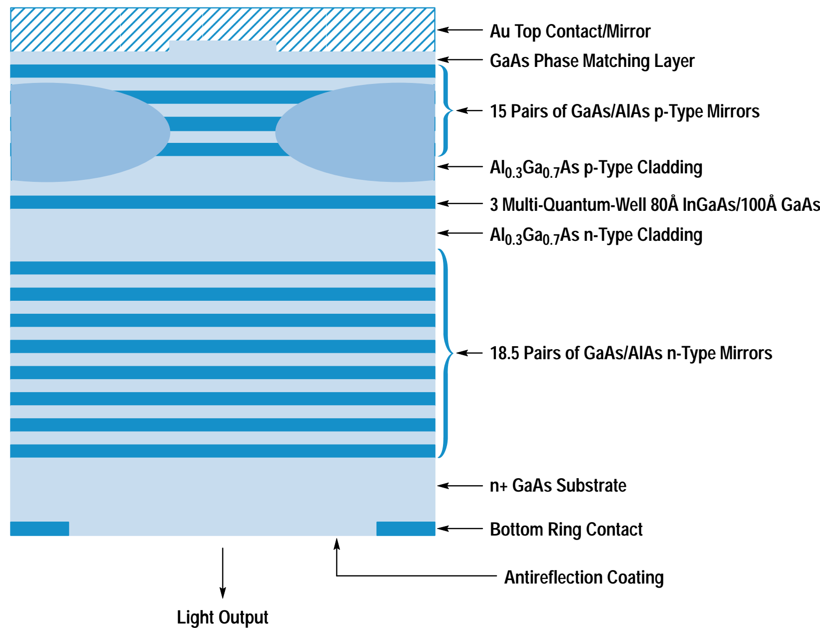


Fig. 2. This is a cross section of a bottom emitting laser with strained multiple quantum wells of InGaAs emitting at a wavelength of 980 nm. It has a totally reflective mirror consisting of hybrid Au/semiconductor distributed Bragg reflectors to minimize series resistivity and an output mirror consisting of semiconductor distributed Bragg reflectors. Proton ion implantation is used to confine the current.

In and the group V arsenic source As_4, the machine is also equipped with a high-temperature hydride cracker for introducing AsH_3 to provide arsenic and a low-temperature gas injector for introducing the p-type dopant of carbon tetrabromide (CBr_4). The n-type dopant used in this work was Si produced by elemental Si in a high-temperature effusion cell. The p-type dopant used is carbon. All growths were performed at 520°C on a 2-inch-diameter n+ substrate.

To maintain the alignment of the gain peak within 10 nm (blue shifted) of the Fabry-Perot wavelength, uniform control of the thickness and alloy composition must be maintained to better than 1% across the wafer. The total growth time for the bottom emitting laser structure is from 8 to 12 hours. To maintain stable growth over this time, an in-situ growth-monitoring technique using a pyrometer is used.[6,7,8] During the growth of the Bragg mirrors consisting of quarter-wavelength thicknesses of GaAs and AlAs, the emission intensity from the heated wafer is detected by a pyrometer. The signal is oscillatory in nature and is directly correlated with the growth of the alternating Bragg layers. Fig. 4 shows the run-to-run reproducibility using the in-situ monitoring technique, the Fabry-Perot wavelength can be achieved within ±1% for several different runs.

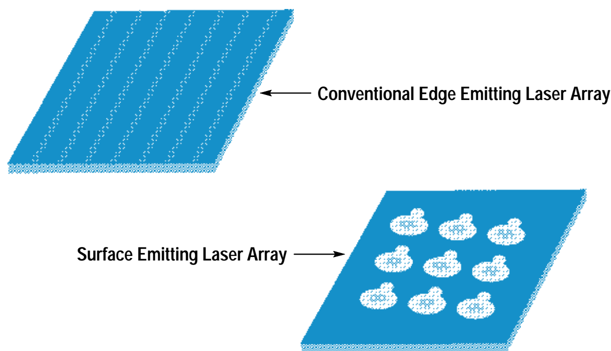


Fig. 3. Surface emitting lasers can be made into two-dimensional arrays and integrated with monitor photodiodes. It is much more difficult to accomplish these things in the edge emitting laser.

Device Design

The surface emitting laser is a bottom emitting structure with strained InGaAs quantum wells emitting at 980 nm. As shown in Fig. 2, it consists of 18.5 pairs of n-type GaAs and AlAs Bragg mirrors on the output face and 15 pairs of p-type GaAs and AlAs together with an Au mirror on the totally reflective face. The cavity is a single wavelength wide and consists of an active region of three 80-angstrom strained InGaAs quantum wells with 100-angstrom GaAs barriers and about 970 angstroms of $Al_{0.3}Ga_{0.7}As$ carrier-confining layers. The interface between GaAs and AlAs in the distributed Bragg reflector mirrors is digitally graded in eight steps using a chirped short-period superlattice. The final p-type GaAs phase-matching layer is doped to $3\times10^{19}/cm^3$ to provide a nonalloyed ohmic contact to the hybrid Au mirror, which also acts as a p contact. The GaAs and AlAs Bragg mirrors

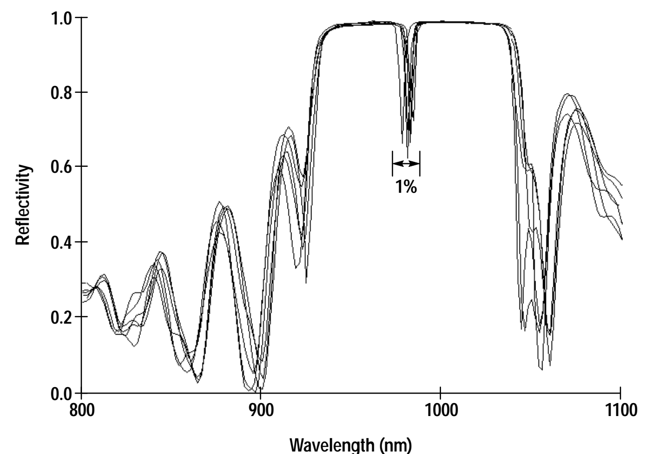
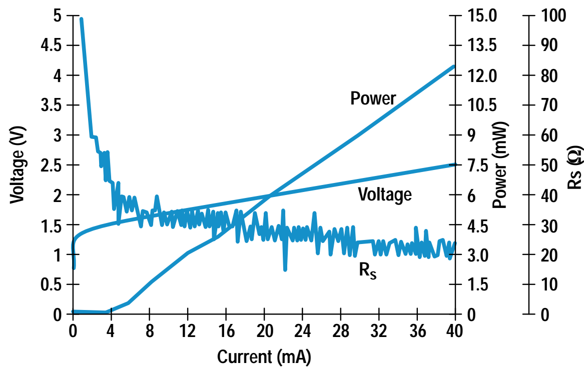


Fig. 4. The stop band characteristics (reflectivity versus wavelength plots) of six different epitaxial runs demonstrate the run-to-run reproducibility achieved with in-situ growth monitoring. The dip in the stop band is caused by the Fabry-Perot cavity formed by the two distributed Bragg reflector mirrors. Variation of the Fabry-Perot wavelength can be kept under 1%. The reflectometer is calibrated by the water vapor absorption line at 942 nm.

© Hewlett-Packard Company 1995



Typical Characteristics

R_S	~20Ω
V_{th}	1.40V
I_{th}	3.8 mA
P_o	>12 mW
η_{ext}	~27%
f_{3dB}	~6 GHz
Z_{th}	176°C/W

Fig. 5. Dc characteristics of a large-area 980-nm surface emitting laser. The curves show voltage, optical power output, and series resistance as a function of bias current. The table shows typical parameters obtained for such surface emitting lasers.

are uniformly doped to $1\times10^{18}/cm^3$ except for the digital grading region which is uniformly doped to $5\times10^{18}/cm^3$. The n dopant is Si and the p dopant is carbon which has been shown not to diffuse[6,7] out of the graded region.

Fabrication Steps and Device Characteristics

The basic fabrication steps for the bottom emitting laser are as follows. When the wafer is received from the grower of the epitaxial layers, its reflectivity is measured in a spectrophotometer to determine the stop band and the wavelength of the Fabry-Perot cavity. A small piece of the wafer is fabricated into a broad-area laser to determine the threshold current and the peak-gain wavelength. The Fabry-Perot wavelength and the peak-gain wavelength are important parameters for the surface emitting laser. Ideally, we would like the peak-gain wavelength to be blue-shifted by 10 nm with respect to the Fabry-Perot wavelength.

Next, the rest of the wafer is coated with gold film in an evaporator. The gold serves as a mirror in addition to the Bragg mirror, further boosting the reflectivity of the end mirror. A photoresist ion implant mask is then defined and the gold field is chemically removed. Protons of varying energy and dosage are implanted to confine the current. Photolithography is then used again to define a gold plating for die attachment. After gold plating, the wafer is lapped and polished to an accuracy of 0.005 inch. Finally, ohmic contacts and antireflection coatings are deposited, their areas defined by photolithography. This completes the surface emitting laser.

Surface emitting lasers with 24-μm active diameters have turn-on voltages as low as 1.40V and threshold current of 3.0 mA. Wallplug efficiencies† of 13% have been demonstrated. The I-V and L-I (light power output versus current) curves of the laser are shown in Fig. 5. The kinks in the L-I curve are from filamentation or higher-order spatial modes appearing in the laser cavity as the bias is increased. The 1.40V turn-on voltage is only 0.28V above the InGaAs bandgap energy. The series resistance of the device is 20 ohms.

Spectral Width

A wide spectral width is necessary to reduce the effect of modal noise resulting from mode-selective loss in multimode links. The surface emitting laser with an active width of 24 μm was found to give a spectral width of 0.3 to 0.7 nm. The wide active region is necessary to allow the accommodation of multiple filaments or higher-order modes whose simultaneous existence gives rise to the wide spectral width.

Fig. 6 shows the near-field pattern of the surface emitting laser and the associated spectrum as a function of the bias. As the bias is increased from 5 mA to 40 mA, the spectrum

† Wallplug efficiency is optical power out divided by electrical power in.

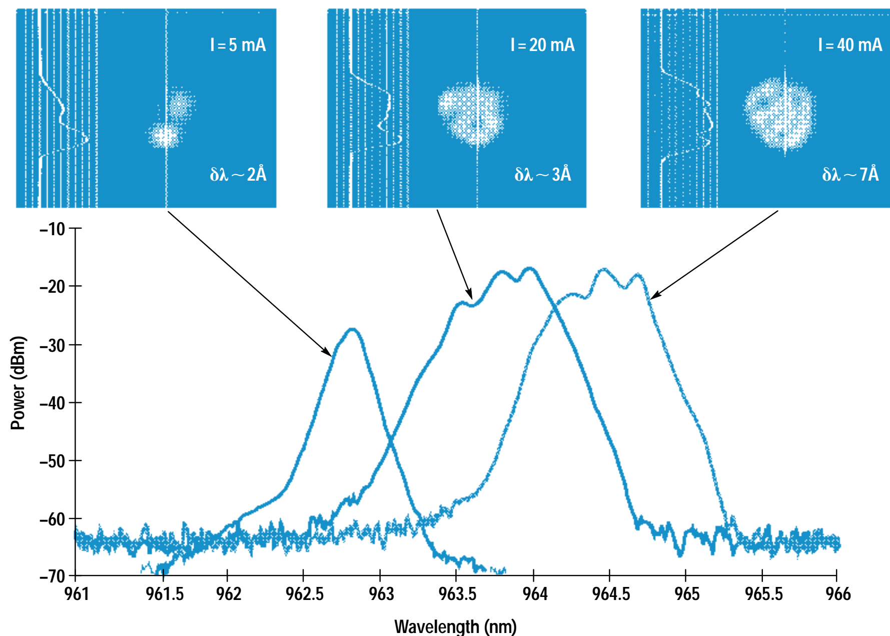


Fig. 6. A primary characteristic of the large-area surface emitting laser is its broad linewidth, which is important to minimize modal noise when using multimode optical fibers. This figure shows the near-field pattern and the associated spectrum as a function of bias current. The broad spectrum is a result of the increased number of spatial modes and the formation of multifilaments, each emitting at a slightly different frequency.

© Hewlett-Packard Company 1995

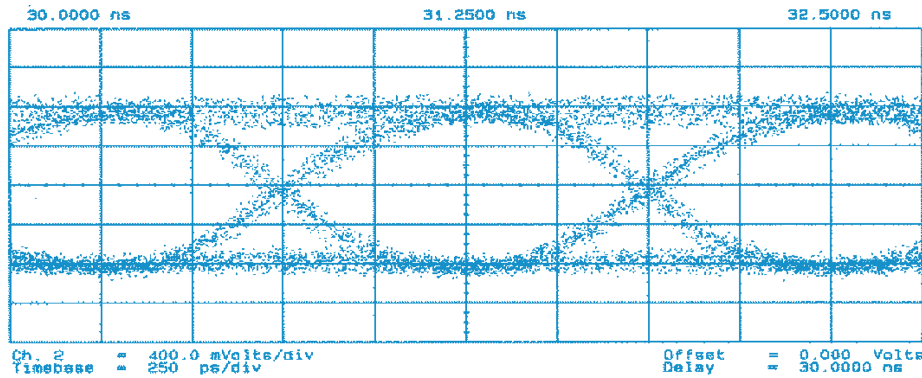


Fig. 7. The large-area surface emitting laser modulated at 1 Gbit/s using a 2^7–1 pseudorandom bit sequence exhibits a clean eye diagram. The laser is prebiased at 27.8 mA and a large modulation signal is applied. The modulated signal is detected after traversing a length of multimode fiber.

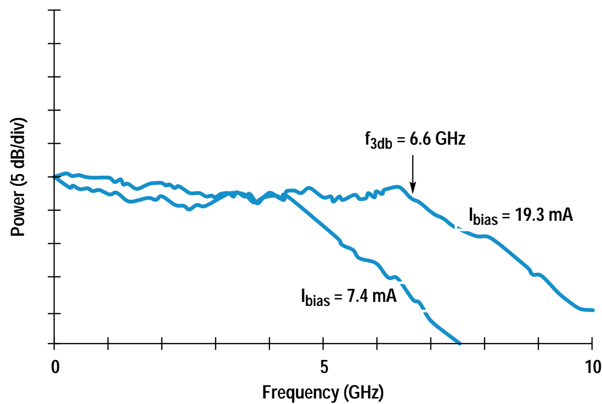


Fig. 8. This power-versus-modulation-frequency plot shows a 3-dB bandwidth of 6.6 GHz under small-signal modulation. The curve with the higher bandwidth corresponds to a bias current of 10.3 mA and the curve with the lower bandwidth corresponds to a bias current of 7.4 mA.

broadens from 0.2 nm to 0.7 nm. Fig. 7 shows the eye diagram at one Gbit/s modulation with the surface emitting laser biased at 27.8 mA using a 2^7–1 pseudorandom bit sequence. A bit error rate (BER) of better than 10^{-12} at 1 Gbit/s, good eye opening, and low modal and intensity noise have been obtained with these devices.

The small-signal frequency response of the surface emitting laser is shown in Fig. 8. The useful bandwidth is greater than 6 GHz.

Reliability

The major burn-in failure that we have observed is that of dark line defects and dark spot defects. The burn-in screening investigation showed that a short constant-current stress is effective in screening out early failures. The conditions used for the burn-in are 70°C and 10^4 A/cm^2 for 24 hours. The devices that pass the burn-in screening are stressed at 60°C at 1 mW and have lived over 4000 hours at the time of writing of this paper. Fig. 9 shows the light output power of 1 mW at 60°C as a function of time.

Conclusion

Large-area surface emitting lasers with wide linewidths are good candidates as sources for short-distance high-speed links using multimode fibers. These surface emitting lasers can replace self-pulsating CD lasers and offer higher bandwidths than the CD lasers.

Acknowledgments

A close collaboration was formed with the HP Optical Communication Division (OCD) to study the reliability of the surface emitting laser. The OCD team was led by Al Yuen and consisted of Tao Zhang, Chun Lei, Christa Tomasevich, Helen Crusco, and Jay Bhagat. The reliability results presented in this paper are the fruits of the joint effort of OCD and HP Laboratories. The authors also thank Long Yang for early contributions, Jean Norman for packaging, and Andreas Weber for discussions. We thank Lynette Martinez and Henrietta Gamino for device processing and Midori

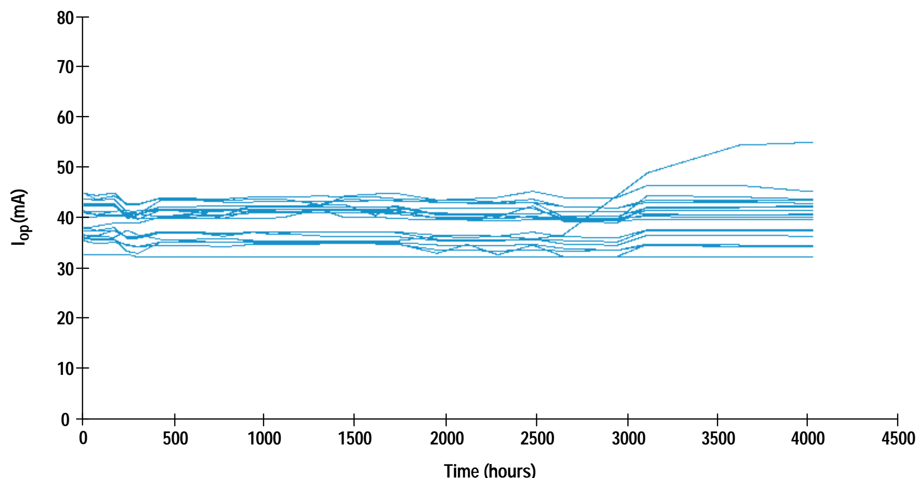


Fig. 9. The surface emitting laser is currently undergoing constant-power life testing. This figure is for 20 preselected lasers operating at 60°C ambient temperature at a light output power of 1 mW. The vertical axis is the current to maintain 1-mW output power and the horizontal axis is the test time in hours.

© Hewlett-Packard Company 1995

Kanemura for assistance in epitaxial growth. We are grateful to Waguih Ishak, Ron Moon, Bob Weismann, Kent Carey, Jeff Miller, David Dolfi, and Andy Liao for their support of this work. Discussions on systems applications with David Cunningham that helped launch this project are also gratefully acknowledged by S.Y. Wang.

References

1. I. Melngalis, "Longitudinal Injection Plasma Laser of InSb," *Applied Physics Letters,* Vol. 6, 1965, p. 59.

2. H. Soda, K. Iga, C Kitahara, and Y. Suematsu, "GaInAsP/InP Surface Emitting Injection Lasers," *Japanese Journal of Applied Physics,* Vol. 18, 1979, pp 23-29.

3. D. Sears, Hewlett-Packard Optoelectronics Division, *private communication.*

4. K.H. Hahn, M.R.T. Tan, Y.M. Houng, and S.Y. Wang, "Large-Area Multitransverse Mode VCSELs for Modal Noise Reduction in Multimode Fibre Systems," *Electronics Letters,* Vol. 29, 1993, p.1482.

5. K.H. Hahn, M.R.T. Tan, and S.Y. Wang, "Intensity Noise of Large-Area Vertical-Cavity Surface Emitting Lasers in Multimode Optical Fibre Links," *Electronics Letters,* Vol. 30, 1994, p. 139.

6. Y.M. Houng, S.D. Lester, D.E. Mars, and J.N. Miller, " Growth of High-Quality p-Type GaAs Epitaxial Layers Using Carbon Tetrabromide by Gas Source Molecular Beam Epitaxy and Molecular Beam Epitaxy," *Journal of Vacuum Science and Technology,* Vol. B11, no. 3, 1993, p. 915.

7. Y.M. Houng, B.J. Lee, T.S. Low, and J.N. Miller, "Growth of High-Quality AlGaAs/GaAs Heterostructures by Gas Source Molecular Beam Epitaxy," *Journal of Vacuum Science and Technology,* Vol. B8, no. 2, 1990, p. 355.

8. Y.M. Houng, M.R.T. Tan, B.W. Liang, S.Y. Wang, and D.E. Mars, "In-Situ Thickness Monitoring and Control for Highly Reproducible Growth of Distributed Bragg Reflectors," *Journal of Vacuum Science and Technology,* Vol. B12, no. 2, 1994, p. 1221.

© Hewlett-Packard Company 1995

Generating Short-Wavelength Light Using a Vertical-Cavity Laser Structure

Second-harmonic generation from a GaAs/AlAs vertical cavity fabricated on a (311)B GaAs substrate has been demonstrated. The experimental results and a theoretical analysis show that a GaAs/AlAs vertical cavity optimized both for efficient confinement of the fundamental power and for quasi-phase-matching can offer efficient second-harmonic generation.

by Shigeru Nakagawa, Danny E. Mars, and Norihide Yamada

There has been great interest in compact short-wavelength light sources, especially blue light sources, for a number of applications such as full-color displays and high-density optical storage. Full-color displays require compact light sources of three colors: red, green, and blue. Red light sources have been developed and are commercially available in the form of AlGaAs laser diodes and light-emitting diodes (LEDs). Green LEDs have recently become commercially available as well. Compact devices emitting blue light have been demonstrated, but their reliability is not good enough for commercial products. Optical storage uses laser diodes to read and write data. Shorter-wavelength laser diodes can give higher storage density. Blue lasers can store data at about three times the density achievable with the infrared lasers currently used for optical storage.

A laser diode made of a wide-bandgap semiconductor such as zinc cadmium sulfur selenium (ZnCdSSe) is one of the approaches being taken to realize compact blue light sources.[1,2] The advantages of this device include compact size and direct modulation of blue light. The feasibility of such a device has been demonstrated, but its reliability is not yet good enough for commercial applications.

Another approach being taken to make compact short-wavelength light sources is second-harmonic generation. In nonlinear optical materials, a fraction of the propagating fundamental optical wave is converted to an optical wave of double the frequency or half the wavelength. Using this second-harmonic generation technique, blue light at a wavelength of 430 to 490 nm has been generated from an infrared light of wavelength 860 to 980 nm. Reliable and compact short-wavelength blue light sources have been prototyped employing nonlinear dielectric materials such as lithium tantalate ($LiTaO_3$) and lithium niobate ($LiNbO_3$).[3-5] However, these second-harmonic generation blue light sources are hybrid and much larger in size than laser diodes.

Second-harmonic generation to realize compact short-wavelength light sources has been demonstrated in semiconductors such as GaAs and AlGaAs.[6-12] Second-harmonic generation in semiconductors can bring about monolithic short-wavelength light sources, since semiconductors like GaAs or AlGaAs can work as both laser materials and second-harmonic generation materials simultaneously. In some devices, a stack of GaAs and AlGaAs with a period of half a wavelength of the second harmonic has been incorporated to give quasi-phase-matched second-harmonic generation.[10-12] In most optical materials, the refractive index changes depending on the wavelength, causing a phase difference between the propagating fundamental light and the propagating second-harmonic light. The phase of the generated second-harmonic light is exactly twice that of the propagating fundamental light, so it is different from the phase of the propagating second-harmonic light, resulting in negative interference between the generated second-harmonic light and the propagating second-harmonic light. For some crystals, it is possible to cancel this phase difference and negative interference by choosing a certain axis for the light propagation direction. This is called phase matching. Another way to reduce the negative interference is by alternating the magnitude or sign of the generated second-harmonic field in phase with the phase difference. This phase-matching scheme does not eliminate the negative interference completely and so it is called quasi-phase-matching.

The second-harmonic power coming out of a GaAs/AlGaAs second-harmonic generator gets saturated in a limited distance because of the large absorption of second-harmonic power by GaAs or AlGaAs. To extract the second-harmonic power efficiently, most of the GaAs second-harmonic generators reported so far are surface emitters. The second-harmonic wave comes out normal to the surface after propagating through only a small distance in the absorbing semiconductor.[7,8,9,11,12] One way to increase the conversion efficiency in the limited GaAs/AlGaAs distance is to resonate the fundamental field with a Fabry-Perot cavity and to increase the intensity of the fundamental field inside the region, since second-harmonic power has a second-order dependence on fundamental power.

We have experimentally demonstrated second-harmonic generation from a GaAs/AlAs vertical cavity fabricated on a (311)B GaAs substrate. A vertical cavity offers efficient confinement of the fundamental field because highly reflective mirrors can be fabricated on both sides of the cavity. In this paper, we will present experimental results and a theoretical analysis showing that a GaAs/AlAs vertical cavity optimized both for efficient confinement of fundamental power and for quasi-phase-matching can offer efficient second-harmonic generation.

© Hewlett-Packard Company 1995

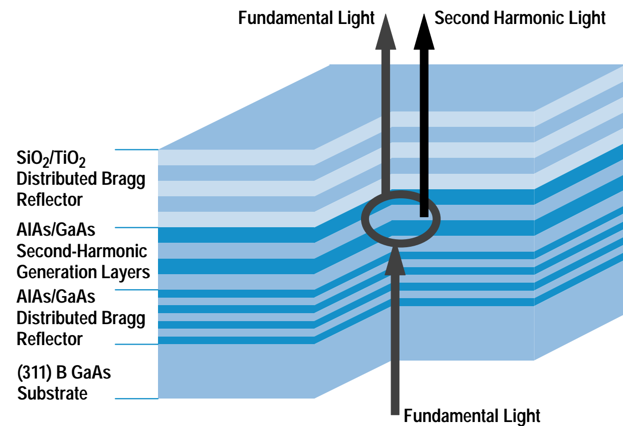
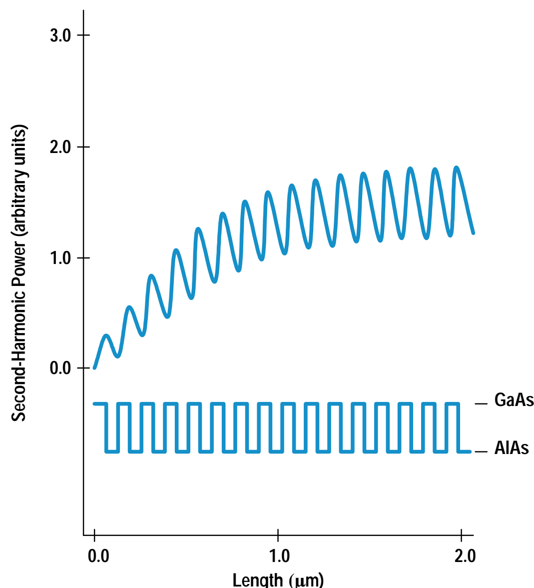


Fig. 1. Structure of the AlAs/GaAs vertical-cavity second-harmonic generator.

Structure of the Device

Fig. 1 shows the structure of the vertical-cavity second-harmonic generator, including a GaAs/AlAs distributed Bragg reflector and a TiO_2/SiO_2 distributed Bragg reflector as highly reflective mirrors. The GaAs/AlAs distributed Bragg reflector absorbs the second-harmonic power, while the TiO_2/SiO_2 distributed Bragg reflector has high transmission at the second-harmonic wavelength, so that generated light can pass through it. To generate second-harmonic power efficiently in the cavity, a periodic stack of AlAs/GaAs is incorporated for quasi-phase-matching. The conventional second-harmonic generation surface emitters demonstrated in the literature have also used a periodic structure for quasi-phase-matching whose period was equal to half a wavelength of the second harmonic.[10-12] However, our calculations indicate that the period should be a little shorter. This difference comes from the fact that while the absorption of second-harmonic power is assumed negligible in the conventional quasi-phase-matching scheme, second-harmonic power generated in the AlAs/GaAs layers is strongly absorbed, especially by the GaAs layers. From the calculated curves shown in Fig. 2, it can be seen that second-harmonic power in our quasi-phase-matched structure is generated much more efficiently than in a conventional half-wavelength quasi-phase-matched stack.

For the vertical-cavity second-harmonic generator, the second-harmonic field must be generated colinearly from the fundamental field. In zinc-blend crystals such as GaAs or AlAs, this is possible only when the substrate is oriented other than (100), such as (111), (110), or (311). A GaAs/AlAs distributed Bragg reflector (19.5 pairs) and a GaAs/AlAs stack of 19 layers were grown on a (311)B GaAs substrate by molecular beam epitaxy (MBE), resulting in good surface morphology. A TiO_2/SiO_2 distributed Bragg reflector (10 pairs) was fabricated after the MBE growth using an electron-beam evaporator and the substrate was polished to a thickness of 200 μm. We measured the total reflectivity of the device and observed a dip in the reflectivity at 984 nm, which indicated a resonance of the fundamental field in the Fabry-Perot cavity. The reflectivity of the GaAs/AlAs distributed Bragg reflector and the TiO_2/SiO_2 distributed Bragg reflector were measured to be 98.4% and 99.9% at 984 nm, respectively.



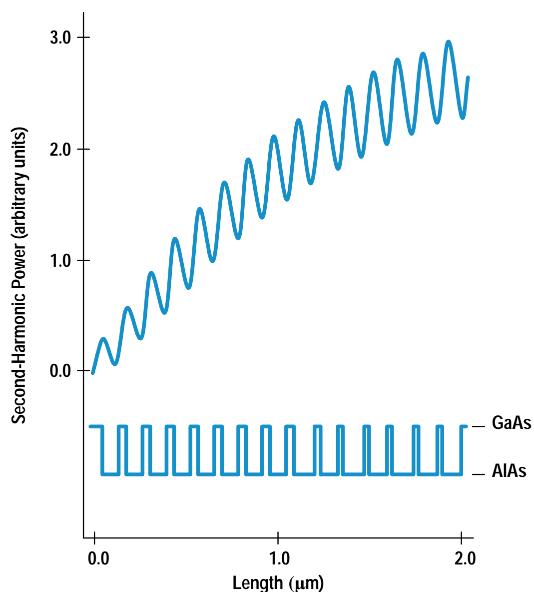


Fig. 2. Distribution of second-harmonic power in (a) a conventional half-wavelength quasi-phase-matched device and (b) a device that is quasi-phase-matched taking into account the absorption of second-harmonic power. The rectangular curves show the distribution of nonlinear coefficient in the devices. The higher and lower levels indicate GaAs and AlAs, respectively.

Results and Discussions

A frequency tunable Ti:Sapphire laser was used as a fundamental light source. The light was shot vertically through the polished GaAs substrate. We measured the second-harmonic power generated from the cavity and the fundamental power of the exiting beam. Fig. 3 shows how the second-harmonic power varies with the polarization angle of the fundamental field, in which the direction of the fundamental electric field is rotated toward the $\langle 01\bar{1} \rangle$ direction (90°) from the $\langle 2\bar{3}\bar{3} \rangle$ direction (0°). The points are measured data and the solid

© Hewlett-Packard Company 1995

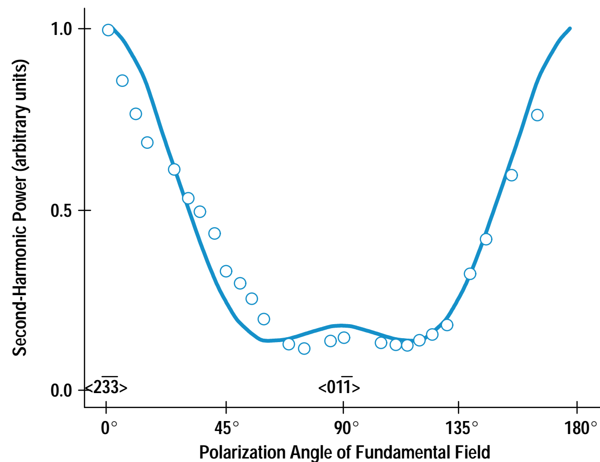


Fig. 3. Normalized second-harmonic power as a function of the polarization angle of the fundamental field. The points are measured data and the line shows the calculated values. 0° is identical to the $\langle 2\overline{3}\overline{3} \rangle$ direction.

line shows the calculated values. The details of the calculations are not given here because of their complexity. However, agreement of the measurement results with the calculations indicates that the observed power is purely second-harmonic power.

Fig. 4 shows the second-harmonic power at a wavelength of 492 nm coming out of the cavity as a function of the input fundamental power at a wavelength of 984 nm, indicating that the conversion efficiency of the device is 1.4×10^{-4} %/W. The conversion efficiency of second-harmonic generation is defined as the output second-harmonic power divided by the square of the input fundamental power (the generated

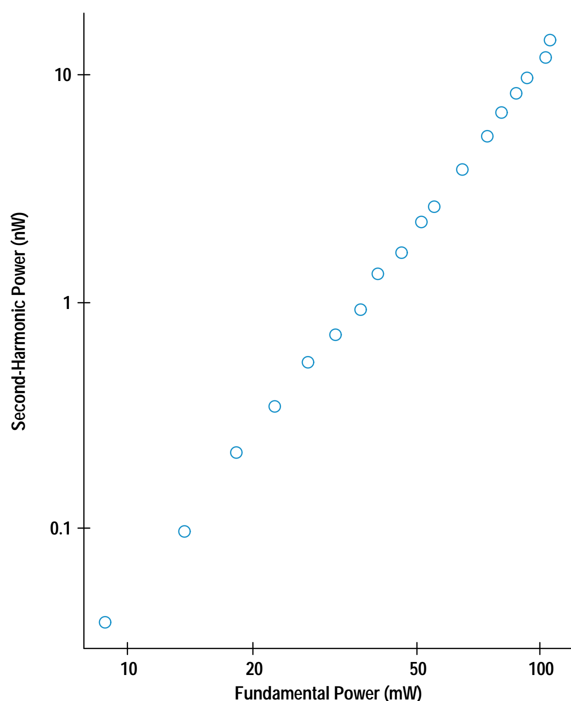


Fig. 4. Second-harmonic power coming out of the vertical cavity as a function of the input fundamental power. The wavelength of the second-harmonic power is 492 nm.

second-harmonic power is proportional to the square of the fundamental power). It is possible to increase the conversion efficiency in several ways. An efficient way is to improve the confinement of fundamental power inside the cavity. This will increase the second-harmonic power by the square of the factor by which the fundamental power in the cavity is increased, without any increase in the input fundamental power.

We found four ways to improve the conversion efficiency of our device. First, by comparing the resonant wavelength of the measurement with that of the design, we found that the actual layer thickness of the grown material was 4% smaller than the design value. As a result, our quasi-phase-matched condition was not completely satisfied. The second-harmonic power coming out of the device has been calculated both for the designed structure and for the actual structure. The calculations indicate that by making the device exactly as designed, the second-harmonic power will be increased 8.1 times compared to the device tested.

Second, the 4% reduction of the layer thickness also reduces the reflectivity of the AlAs/GaAs distributed Bragg reflector to 98.4% from the predicted value of 99.8%. This decreases the confinement of the fundamental field in the cavity. Third, the full width at half maximum (FWHM) of the Ti:Sapphire laser spectrum is 0.17 nm, which is much larger than that of the cavity, which is 0.03 nm. This causes a large part of the input fundamental light from the laser to be reflected, resulting in smaller confinement of the fundamental field in the cavity. Improvement of these second and third factors is expected to increase the fundamental power confined in the cavity to 90 times that of the present device, thereby increasing the second-harmonic power by a factor of 8.1×10^3.

Fourth, the fundamental beam from the Ti:Sapphire laser is focused on the device with a FWHM spot diameter of 18.6 μm. The conversion efficiency is proportional to the power density of the fundamental field, and would be increased 13.8 times with a fundamental beam diameter of 5 μm. Based on these considerations, we concluded that the conversion efficiency of the device would be 1.3×10^2 %/W if optimized.

Conclusions

We have demonstrated second-harmonic generation from an AlAs/GaAs vertical cavity fabricated on a (311)B GaAs substrate. We have observed a conversion efficiency of 1.4×10^{-4} %/W. We have theoretically shown that we can increase the conversion efficiency up to 1.3×10^2 %/W by optimizing both for quasi-phase-matching and for efficient confinement of the fundamental field.

Acknowledgments

Many people contributed to the second-harmonic generation from an AlAs/GaAs vertical cavity. We would like to express our special acknowledgment to Nobuo Mikoshiba, director of HP Laboratories Japan, for his support and for his technical discussions. We would like to thank Jeff Miller of HP Laboratories for his useful suggestions on the experiments and for his technical discussions. We also want to thank all of the members of the photonics group of HP Laboratories Japan for their support and help.

© Hewlett-Packard Company 1995

References

1. M.A. Hasse, J. Qiu, J.M. DePuydt, and H. Cheng, "Blue-Green Laser Diodes," *Applied Physics Letters*, Vol. 59, 1991, p. 1272.
2. H. Jeon, J. Ding, V. Nurmikko, W. Xie, D.C. Grillo, M. Kobayashi, R.L. Gunshor, G.C. Hua, and N. Otsuka, "Blue and Green Diode Lasers in ZnSe-Based Quantum Wells," *Applied Physics Letters*, Vol. 60, 1992, p. 2045.
3. E.J. Lim, M.M. Fejer, R.L. Byer, and W.J. Kozlobsky, "Blue Light Generation by Frequency Doubling in Periodically Poled Lithium Niobate Channel Waveguides," *Electronics Letters*, Vol. 25, 1989, p. 731.
4. S. Matsumoto, E.J. Lim, H.M. Hertz, and M.M. Fejer, "Quasi-Phase-Matched Second-Harmonic Generation of Blue Light in Electrically Periodically Poled Lithium Tantalate Waveguides," *Electronics Letters*, Vol. 27, 1991, p. 2040.
5. K. Yamamoto, K. Mizuuchi, and T. Taniuchi, "Quasi-Phase-Matched Second-Harmonic Generation in a $LiTaO_3$ Waveguide," *IEEE Journal of Quantum Electronics*, Vol. 28, 1992, p. 1909.
6. N. Ogasawara, R. Ito, H. Rokukawa, and W. Katsurashima, "Second-Harmonic Generation in an AlGaAs Double-Heterostructure Laser," *Japanese Journal of Applied Physics*, Vol. 26, 1987, p. 1386.
7. D. Vakhshoori, M.C. Wu, and S. Wang, "Surface Emitting Second-Harmonic Generator for Waveguide Study," *Applied Physics Letters*, Vol. 52, 1988, p. 422.
8. H. Dai, S. Janz, R. Normandin, J. Nielsen, F. Chatenoud, and R. Williams, "An InGaAs/GaAs SQW Laser Integrated with a Surface Emitting Harmonic Generator for DWDM Applications," *IEEE Photonics Technology Letters*, Vol. 4, 1992, p. 820.
9. R. Lodenkamper, M.L. Bortz, M.M. Fejer, K. Bacher, and J.S. Harris, Jr., "Surface Emitting Second-Harmonic Generation in a Semiconductor Vertical Resonator," *Optics Letters*, Vol. 18. 1993, p. 1798.
10. S. Janz, C. Fernando, H. Dai, F. Chatenoud, M. Dion, and R. Normandin, "Quasi-Phase-Matched Second-Harmonic Generation in Reflections from $Al_xGa_{1-x}As$ Heterostructures," *Optics Letters*, Vol. 18, 1993, p. 589.
11. D. Vakhshoori, R.J. Fischer, M. Hong, D.L. Sivxo, G.J. Zydzik, and G.N.S. Chu, "Blue-Green Surface Emitting Second-Harmonic Generators on (111)B GaAs," *Applied Physics Letters*, Vol. 59, 1991, p. 896.
12. D. Vakhshoori, "Analysis of Visible Surface Emitting Second-Harmonic Generators," *Journal of Applied Physics*, Vol. 70, 1991, p. 5205.

© Hewlett-Packard Company 1995

A New, Flexible Sequencer Architecture for Testing Complex Serial Bit Streams

Based on a generic model of serial communication systems, this architecture dramatically reduces the time needed to program functional and in-circuit tests for devices with serial interfaces. It is implemented in a new Serial Test Card and Serial Test Language for the HP 3070 family of board test systems.

by Robert E. McAuliffe, James L. Benson, and Christopher B. Cain

As serial bit streams become more prevalent in electronic products, the need for high-quality, thorough tests for those products increases dramatically.[1] Traditional automatic test equipment (ATE) architectures have limitations that make it extremely difficult (if not impossible) to write such tests. In this paper we discuss those limitations and describe a new sequencer architecture specifically designed to address the challenges of serial bit stream testing.

The architecture has been implemented as an enhancement to the HP 3070 family of board testers and has been used to emulate many challenging serial protocols and formats, including: ISDN S- and U-interfaces, I^2C, HDLC control channels, generic 64-kbit/s streams, IOM-2, ST-Bus, and various time division multiplexed (TDM) backplanes. Customers using the architecture have experienced dramatic reductions (up to 25×) in test development time, as well as significant increases (8× or more) in test throughput.

We assume the reader has at least some knowledge of manufacturing test procedures and equipment. A basic knowledge of telecommunications concepts may also prove useful, as many of the examples given are related to telecomm applications.

Throughout the paper we will refer to a *device under test* (DUT). In general the discussion will be in the context of functional board testing, in which case the DUT would be a complete printed circuit assembly. In this time of rapid technological changes, however, functionality that once required entire boards is now implemented as small clusters of components or as single integrated circuits. We will therefore use the term DUT to refer to whatever is being tested, be it an IC, a cluster of components, a board, or a complete system.

First we give a brief overview of traditional ATE sequencers and their shortcomings, and then discuss in more detail some of the special challenges of serial bit stream testing. We will show that many test development problems are caused by a fundamental mismatch between the ATE capabilities and the features of DUTs with serial interfaces.

Next we introduce a generic model of serial communication systems. This is essentially a definition of the general characteristics shared by all serial bit streams. Using this model, we were able to develop a test sequencer architecture more closely matched to the characteristics of serial bit streams and the DUTs that use them.

We then describe the architecture as implemented in the HP 3070 board test family. The architecture of the *Serial Test Card* (STC) and the *Serial Test Language* (STL) are described in these sections.

Finally, we present several case studies showing how the STC solves real-life testing problems.

Evolution of ATE Sequencers

Traditional ATE systems feature a test pattern sequencer capable of driving and receiving many simultaneous digital signals. Sequencers of this type first appeared over a decade ago. Early versions of these sequencers excelled at testing SSI/MSI components and simple circuit boards, but had difficulty with microprocessors and other VLSI components.[2]

In response to the test problems posed by microprocessors, ATE manufacturers enhanced their sequencers. New features like formattable pins, algorithmic pattern generation, memory emulation, and bus emulation were added to make the test sequencers a better match to these microprocessor-oriented DUTs.

Today, DUTs embody faster and more powerful microprocessors, concurrent processing technologies, serial communication channels, mixed signal functions, and a variety of custom circuitry (such as ASICs and FPGAs). Each of these characteristics brings with it challenges for the test engineer, but the widest gap between the DUT and traditional ATE seems to be in the area of serial communication testing and its associated concurrent processing technology. The single-sequencer, massively parallel architecture of traditional ATE is not suited to the special problems presented by these DUT characteristics.

 © Hewlett-Packard Company 1995

Our goal was to design a sequencer architecture better matched to the special test requirements of serial-oriented DUTs. As a first step toward that goal, we surveyed a large number of DUTs and serial protocols to identify specific characteristics that make them difficult or impossible to test with a traditional ATE system.

Characteristics of Serial DUTs

Serial-oriented DUTs have many characteristics in common with many other modern electronic devices. Conversely, they also have many characteristics that are fundamentally (or sometimes subtly) different. Some of the more common features of serial-oriented DUTs are:

- Complex physical (electrical) interfaces
- Multiple interfaces operating at unrelated bit rates
- Nondeterministic bit streams
- Bit streams with embedded clocks
- Hierarchical bit streams.

Complex Physical Interfaces

To maximize utilization of a transmission medium, many serial protocols abandon traditional binary digital formats in favor of three-level or four-level digital interfaces. The ISDN S-interface operates with three logic levels,[3] and the ISDN U-interface (2B1Q) operates with four. Other standard telecomm protocols are similar.

Many of these protocols require the transmitted waveforms to correspond to exactly specified shapes, usually to limit the high-frequency components of the signal to a reasonable level. Traditional sequencers have binary stimulus and response capabilities with programmable high and low logic levels. Some even provide rudimentary slew-rate control, but none are designed to interface directly with complex, nonbinary bit streams.

Multiple Interfaces Operating at Unrelated Bit Rates

Many DUTs contain more than one serial interface. In many cases, each interface is asynchronous with respect to the others (there is no specified alignment between bit centers from one interface to the next). One can sometimes force alignment by running each interface from a common clock, but this technique does not necessarily work with asynchronous protocols (see "Bit Streams with Embedded Clocks," below). Moreover, it is quite common to have interfaces running at entirely different bit rates.

The only viable testing approach for a traditional sequencer architecture is to apply vectors at a rate equal to the least common multiple of the clock rate of the two interfaces (assuming there are only two interfaces). The number of test vectors required for even simple tests can be formidable using this approach.

One of the authors has personally written a test for an ISDN S-interface device using a traditional sequencer and this "least common multiple" approach. The effort required three months and 13,000 lines of source code (roughly 90,000 compiled test vectors). Even with this huge effort, only a small fraction of device functionality was tested. Such long test development times are simply not acceptable in today's competitive markets.

Nondeterministic Bit Streams

The response of a DUT to an applied stimulus is not always deterministic, that is, many different "correct" responses to the stimulus are possible. An analog-to-digital converter, for example, will not generally produce exactly the same sequence of digital samples in response to different applications of the same analog stimulus. This does not mean that any one sequence is more correct than any of the others. It simply means a different measurement technique must be applied to the problem.

Some ISDN data link activation procedures include the transfer back and forth of HDLC-like packets of information.† According to these procedures, the packet address is, under some circumstances, generated by a pseudorandom generator on the DUT. This presents no problem if the address generated by the DUT is deterministic in response to a particular initialization sequence. On the other hand, if the address cannot be predicted, the test sequencer must be able to capture whatever address was generated and save it for use later in the test. The authors have not encountered a traditional sequencer with such on-the-fly storage capabilities.

Bit Streams with Embedded Clocks

Some serial interfaces are asynchronous in nature, that is, the clock specifying the bit boundaries is embedded in the bit stream. Traditional sequencers typically sample at predetermined times and are unable to interface properly to such a bit stream.

Embedded clocks can take many different forms. Some protocols guarantee data transitions at regular intervals. Other protocols provide no such guarantee but depend on the transmitter and receiver operating at a prearranged bit rate (asynchronous protocols such as those typically used with RS-232 connections work in this way).

In addition, *framing* information can also be embedded in the bit stream. Asynchronous protocols, for example, mark a frame boundary with a start bit. The HDLC and similar synchronous protocols indicate framing with a special flag pattern, usually eight bits in length. A traditional sequencer may be able to handle such cases if it has sophisticated branching capabilities, but it is very difficult to write a test that can synchronize to a complex framing pattern while simultaneously applying test patterns to other interfaces of the DUT.

Hierarchical Bit Streams

Many serial interfaces contain bit streams within bit streams. We call these *hierarchical* bit streams. A basic-rate ISDN interface is an example of a hierarchical bit stream. A basic frame of this interface consists of 16 bits of B channel data and 2 bits of D channel data.†† If the D channel bits from each frame are extracted and assembled one after the other into a separate serial bit stream, they form an HDLC-like bit stream.

† See references 3 and 4 for a complete description of ISDN and associated activation procedures. Details of HDLC and other bit-oriented protocols can be found in references 5 and 6.

†† This is a simplified description of an ISDN basic rate frame. Actual frame length and content vary depending on which reference point (S, U, etc.) is being considered.

© Hewlett-Packard Company 1995

This situation is very difficult to handle with a traditional sequencer because the *logical* channel the test engineer wants to communicate with is surrounded by a great many other bits of little interest. Complicated subroutines must be written to extract the data of interest from the large frame of bits or insert the data of interest into the large frame of bits. Furthermore, the test engineer must somehow attain bit and frame alignment for both the main frame and the embedded logical frame. This may make the test impossible to implement with a traditional sequencer.

Solutions

For many serial interfaces, the above problems can be solved using custom electronics, ranging from special circuits in the test system fixture to a "hot mockup" of the final system application of the DUT, or by simplifying and eliminating tests so that they can be more readily implemented with a traditional sequencer. We propose a third solution: use of a test sequencer designed specifically to address the special challenges of serial bit stream testing. By using the right tool for the job, the test programmer can develop a thorough, high-quality test quickly and without the need for complex fixture electronics.

Generic Model of Serial Communication Systems

In the last section we discussed the difficulties of serial testing using traditional sequencer architectures. For the most part, these difficulties are caused by an inherent mismatch between the DUT and the tester.[7] In each case, the problems presented could certainly be solved with custom circuitry provided by the customer or by the ATE vendor. However, this approach undermines one of the key advantages of commercial ATE systems: the advantage of being able to use the same tester to test many different DUTs (or many different pieces of the same DUT).

A better approach is to design an architecture that is specific enough to handle the peculiarities of serial testing but general enough to be usable for many different types of serial DUTs and serial protocols. To aid in the definition of such an architecture, we looked deeper into the test problems described in the last section in an effort to understand the fundamental nature of each. This led to the development of our generic model of serial communication systems, described below.

We designed our new sequencer architecture using this model as a guide, so essentially any bit stream compatible with the model is compatible with our architecture. Our particular *implementation* of the architecture was targeted at telecomm applications, so cost/performance trade-offs appropriate to that market have been made. The model (and thus the architecture) is more general and could theoretically be implemented in other ways for other serial test markets.

Definitions

The following terms are used throughout this section:

- *Communication System.* A means of transferring information from one place to another.
- *Serial Communication System.* A communication system that encodes information into digital signals and transfers this digital information from one place to another in a time-serial fashion, that is, the bits of digital information are transferred sequentially one after another according to a prearranged protocol. A serial communication system is typically composed of subparts made up of various bit processors (see below), which process and transform serial communication bit streams (see below).
- *Serial Communication Bit Stream.* A physical transmission path connecting two bit processors (see below) in a serial communication system. Bits are transmitted in a serial fashion, that is, the bits of a message are transmitted sequentially one after another on a common transmission medium. "Serial communication bit stream" will be abbreviated to simply "serial bit stream" or "bit stream" throughout the following discussions.
- *Bit Processor.* A hardware or software device that connects one bit stream to another. A bit processor usually transforms or filters the bit stream in some manner, but can also serve as a *source* or *sink*. A source generates the information to be transmitted over the communication system, and a sink receives and analyzes that information at the other end.

A serial communication system is composed of numerous pieces, each piece defined as either a bit stream or a bit processor. Bit processors serve to connect (transform) one bit stream to another. Or, equivalently, bit streams serve to connect one bit processor to another. Each of these elements—bit streams and bit processors—has certain well-defined properties. These properties are discussed below.

Properties of Bit Streams

Every serial bit stream possesses the following four properties:

- Physical specifications
- Symbol synchronization algorithms
- Framing algorithms
- Logical channel identification (multiplexing).

Physical Specifications. The physical specifications describe the electrical properties of the bit stream, the number of logic levels defined, and any other properties related to the physics of transferring the bit stream from one place to another.

Symbol Synchronization Algorithms. Since a serial bit stream is inherently composed of bits, there must be a way of demarcating the bit boundaries within the bit stream. A device designed to interpret the bit stream would use a symbol synchronization algorithm to locate the bit boundaries. The synchronization property of the bit stream is either *explicit* (a dedicated signal path for clocking is provided) or *implicit* (symbol synchronization information is encoded within the serial bit stream†).

Framing Algorithms. A raw serial bit stream is capable of transferring very little information. For example, a binary bit stream can represent only one of two states at any given moment (1 or 0, on or off). However, if a time reference is provided with the bit stream (an indication of when the bit stream "starts"), then bits can be assembled into larger units capable of carrying more information (bytes, words, messages, etc.). This is the purpose of framing: to provide a

† Asynchronous protocols are also considered implicitly clocked. In that case the symbol rate is defined as part of the bit stream symbol synchronization algorithm.

© Hewlett-Packard Company 1995

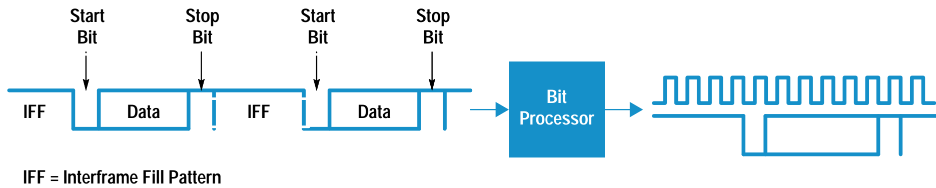


Fig. 1. Processing an RS-232 bit stream.

reference point so the bits of the serial stream can transfer more complex information.

There are two types of framing. Framing can be either *explicit* (a dedicated signal path for framing is provided), or *implicit* (framing information is encoded within the serial bit stream). An implicit framing algorithm may indicate framing using either a *bunched* framing pattern or a *distributed* framing pattern. A bunched pattern is made up of a group of contiguous bits. A distributed pattern is made up of a group of bits interspersed among the data bits.

The time interval between the end of one frame and the beginning of the next is called the *interframe gap*. This gap may be either zero-length (frames are back-to-back) or nonzero-length. If it is nonzero-length, the interframe gap is filled with an *interframe fill (IFF) pattern.*

Logical Channel Identification. Many serial bit streams simultaneously carry more than one independent logical channel of information. We say that these bit streams are *multiplexed* or *hierarchical*. If the independent logic channels are ever to be recovered from a multiplexed bit stream, there must be a means of uniquely identifying each individual logical channel. The multiplexing scheme specifies the method used to multiplex the logical channels and may be categorized as either *explicit* or *implicit*.

With explicit multiplexing, each frame contains a group of information from each multiplexed logical channel. The frame boundary provides a reference for the bit groupings. Time division multiplexed (TDM) highways, such as those found on telecomm line cards, and the ISDN S-bus are examples of explicitly multiplexed bit streams.

With implicit multiplexing, each frame contains a group of information from only one of the multiplexed logical channels. Information from other logical channels may follow in subsequent frames. In this case, logical channel identification is encoded in the bit stream (for example the address field of an HDLC frame or other packetized data.)

Properties of Bit Processors

The job of a bit processor is to convert one bit stream into another according to some specified algorithm. A bit processor therefore has two distinct properties: port definitions and transformation algorithms.

Port Definitions. A bit processor interfaces to bit streams through one or more *ports*. Most bit processors will have an *input port* and an *output port*, but some will have only one or the other.† Each port must have a *data interface* and must of course match the physical specifications of the bit stream. The directional sense of the data interface determines whether a port is an input port or an output port. Data always flows into an input port and out of an output port.

Other requirements are determined by the serial bit stream to which the port attaches. If the bit stream is explicitly clocked (uses an explicit symbol synchronization algorithm), the port must have a *clock interface*. Similarly, if the bit stream uses an explicit framing algorithm, the port must have a *frame synchronization interface*. These interfaces may consist of one or more physical signals that flow either into or out of the port. The implementation details can be determined by studying the bit stream specifications and the transformation algorithm.

Logical channel grouping algorithms do not directly affect the requirements of the port, although they may influence the design of the clock interface. For example, a bit processor designed to demultiplex a logical channel from a multiplexed bit stream could be implemented in one of several ways. One method might use gated clock signals: clock edges only occur when bits from the logical channel of interest are output from the data interface. Another method might use a continuous clock (*all* bits from the original bit stream are output at the data interface) and generate a separate "data valid" signal when bits from the logical channel of interest are present at the output data interface.

Transformation Algorithms. A transformation algorithm simply defines the function to be performed by the bit processor. The bit processor must manipulate the incoming bit stream in whatever manner is necessary to make it conform to the requirements of the outgoing bit stream.

A simple example will help to clarify the functions of bit processors and bit streams.

Applying the Generic Model

Fig. 1 shows an example of the generic model applied to a simple asynchronous communication system. The bit stream on the left is the incoming bit stream. Using conventional terminology, it would be described as an asynchronous bit stream using one start bit and one stop bit, no parity checking, and an eight-bit data field. The definition using the generic model would be something like this:

Physical Specifications:
- Logic levels: 2
- Voltage levels: RS-232-C levels; logic high = –3 volts; logic low = 3 volts

Symbol Synchronization Algorithm:
- Type: Implicit, known symbol rate
- Symbol Rate: 9600 per second

Framing Algorithm:
- Type: Implicit, bunched framing pattern
- Framing Pattern: "10" (at least one stop bit or IFF bit followed by the start bit)

† These are called *sources* or *sinks* and occur at the ends of a communication system. This is where the information sent back and forth in the serial bit stream is ultimately generated or analyzed.

© Hewlett-Packard Company 1995

Interframe Gap: >= 0 symbols
IFF Pattern: "1"

Logical Channel Identification:
Type: Explicit (none really, since there is a single eight-bit data field)

The bit processor receives this bit stream on the data interface of its input port. As shown in Fig. 1, the job of this particular bit processor is to generate a clocked, synchronous version of this bit stream on its output port. Here is the definition of the bit stream output from the bit processor (the bit stream on the right side of the figure):

Physical Specifications:
Logic levels: 2
Voltage levels: Standard TTL levels

Symbol Synchronization Algorithm:
Type: Explicit
Clock Rate: 9600 Hz

Framing Algorithm:
Type: Implicit, bunched framing pattern
Framing Pattern: "10" (at least one stop bit or IFF bit followed by the start bit)
Interframe Gap: >= 0 symbols
IFF Pattern: "1"

Logical Channel Identification:
Type: Explicit

Notice that the framing algorithm and logical channel identification specifications have not changed. The bit processor has merely generated a clock signal synchronized to the incoming bit stream. In this case another bit filter upstream of this one could further transform the bit stream and perhaps extract the data field. Since the output bit stream uses an explicit clock, it is not really necessary to specify the clock rate. Another bit processor that conforms to our generic model should be able to accept the bit stream knowing only that there is an explicit clock signal. Any practical implementation, however, will have a finite clock rate specification, so it is a good idea to specify all such performance requirements as part of the bit stream description.

The definition of the bit processor in our example looks like this:

Input Port:
Data Interface: Required. See bit stream definition for physical requirements.
Clock Interface: Not required, bit stream is implicitly clocked.
Frame Sync Interface: Not required, bit stream is implicitly framed.

Output Port:
Data Interface: Required. Standard TTL output specifications.
Clock Interface: Required, bit stream is explicitly clocked. The interface consists of a single clock output signal meeting standard TTL logic specifications.
Frame Sync Interface: Not required, bit stream is implicitly framed.

Transformation Algorithm:
1. Use a combination of the given baud rate, an internal high-speed clock, and data signal transitions to locate bit centers.
2. Sample the input data at the bit centers and retransmit on the output data interface. Transmit the sampling clock on the output clock interface.

The actual algorithm would probably need to be more sophisticated, but this simple example illustrates the process of mapping a real communication system onto the generic model. A more complex example is given in Fig. 2.

Modularity

The example just given was quite simple but can serve to illustrate an important feature of the model. As mentioned before, another bit processor upstream of the one described might further transform the bit stream, perhaps aligning itself to the framing markers and extracting the data field. The data field could then be passed to yet another bit processor, and so on. We call this feature of the model *modularity*. Using this idea, we can conceptually break up a serial communication system into smaller, simpler pieces consisting of a series of bit processors connected by a series of serial bit streams.

Modularity also aids the processing of hierarchical bit streams. The functions of demultiplexing and logical bit processing can be divided among more than one bit processor.

Hardware Architecture

As we have seen, real-world bit streams and communication systems can be described by carefully specifying each property of the generic model. In other words, the model can effectively mimic any serial communication system. A sequencer architecture that exactly implements the generic model would therefore be easily programmed to test *any* serial bit stream or protocol.

We knew of course that a variety of constraints (the laws of physics, for example) would limit our ability to implement the model exactly. We also knew an abstract model that could not be implemented well enough to solve real test problems would be useless, so we revisited the test problems described earlier. We examined each test problem, looking for issues that might affect the way in which we chose to implement the model.

Complex Physical (Electrical) Interfaces. The model handles this quite easily. One simply describes the characteristics on a sheet of paper. An actual implementation, however, is quite different. We decided early on that it was not feasible to implement hardware compatible with any possible electrical interface! Instead, we chose to implement hardware of several different classes, each class capable of handling a family of related physical bit streams.

Multiple Interfaces Operating at Unrelated Bit Rates. The generic model does not address this issue, although the model can be used to describe each individual interface. We decided that this implied a multichannel architecture in which each channel was capable of operating essentially independently of the others.

© Hewlett-Packard Company 1995

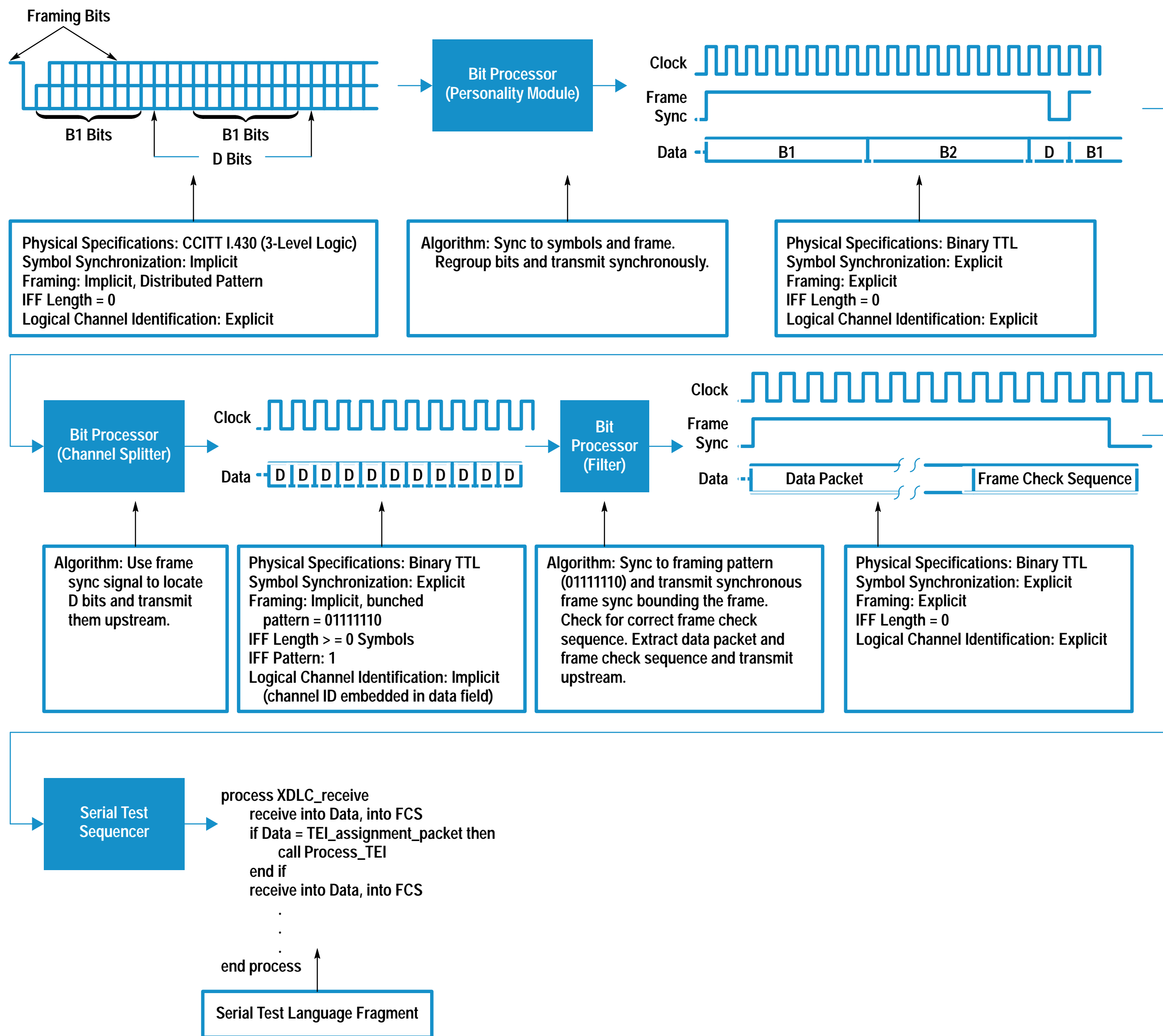


Fig. 2. Processing an ISDN S-bus D channel.

Although our original goal was to make test development faster and easier, we also realized that a multichannel, concurrent architecture could dramatically increase test throughput. For example, a long bit error rate test could be run simultaneously on all channels of a multichannel DUT.

Nondeterministic Bit Streams. In the model, any sort of calculation or adjustment of information in the bit stream is specified by the transformation algorithm. We thought there were two areas of implementation that could be affected by this issue. First, we thought it might sometimes be necessary for an algorithm to access information from both the transmit and receive bit streams simultaneously. This implied the need to process both bit streams within the same bit processor. Secondly, we knew that any actual implementation of a bit processor would have limits, so we wanted to be able to cascade or chain bit processors.

Bit Streams with Embedded Clocks. Our solution to the problem of complex physical interfaces applies to this problem as well: sets of different hardware each tuned to a class of embedded clock schemes.

Hierarchical Bit Streams. Hierarchical bit streams imply multiplexing, so we knew that our implementation needed to be good at handling multiplexed bit streams. This again implied the need for cascadable bit processors.

The architecture that eventually emerged from these considerations is shown in Fig. 3. The architecture is multichannel in nature. The figure shows a single channel of the architecture in the center with adjacent channels above and below. Each channel is designed to attach to a single bit stream of the DUT.

Each channel of the architecture contains two information sources/sinks called *serial test sequencers* (STS). In a typical application, each STS would process independent subchannels (logical channels) of the bit stream. In Serial Test Language, each logical channel is called a *substream* and is controlled by a *process* running on the STS. When required by the test program, STS resources from adjacent processing channels can also be attached to the bit stream, providing up to four substreams for each bit stream.

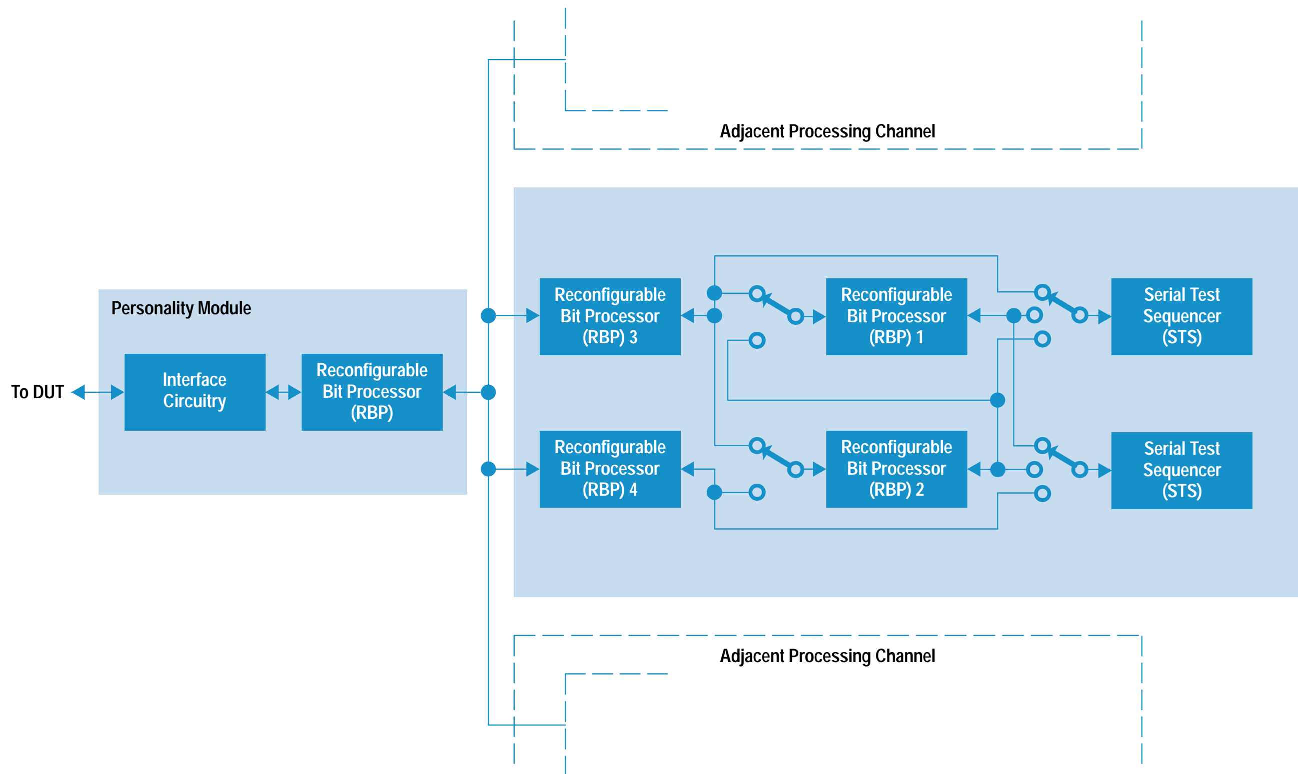


Fig. 3. A single processing channel of the Serial Test Card (STC) for the HP 3070 board test family.

Four channels of the architecture are implemented on the HP 3070 Serial Test Card (STC) and up to 12 STCs can be installed in a system.

Bit Processors

As shown in Fig. 3, each STC processing channel is composed of a series of bit processors connected by serial bit streams. The bit processors are implemented with the XC3000 series of field-programmable gate arrays (FPGAs) from XILINX Company. An important feature of the XC3000 series is their RAM-based configurability. The XC3000 can be programmed on-the-fly in the system, so no preprogrammed ROMs are needed. This feature allows the transformation algorithm of a bit processor to be changed on the fly and makes these devices ideal for this application. The transformation algorithms are implemented by circuits inside the XILINX devices. They are not really either hardware or software, so we have coined the word *circuitware* to describe this sort of reconfigurable circuitry.

Fig. 4 shows a more detailed view of our standard bit processor, called a *reconfigurable bit processor*, or RBP. In addition to the XILINX XC3042 FPGA, each RBP also includes a pair of 2K-by-8-bit RAMs. The RAMs are connected to I/O pins on the FPGA and are used as required by the circuitware designer. This RAM resource complements the architecture of the FPGA and provides a large, dense local storage element.† Each RBP has a downstream port (towards the DUT) and an upstream port (towards the STS), and each of these ports supports data transmission in both directions simultaneously.

† The XC3000 series of parts is structured as an array of D-type flip-flops fed by Boolean function generators. This structure makes them well-suited for state machine and random logic designs, but unsuitable for applications requiring a lot of storage elements.

The bit streams that interconnect the RBPs are defined according to the generic model. All internal bit streams are binary TTL logic signals and are explicitly clocked and explicitly framed. The RBP clock interface ports include *data valid* and *ready* signals to support multiplexed bit streams.

Personality Modules

We use *personality modules* to implement each interface class. There are currently personality modules available for TTL, ISDN S-bus and ISDN U-bus electrical formats. As a whole, the personality module serves as a bit processor and is responsible for converting the external serial bit stream into a format compatible with the internal STC serial bit stream definition.

Serial Test Sequencer

The serial test sequencer (STS) is the final bit processor in the chain. As such, it will often be an information source or

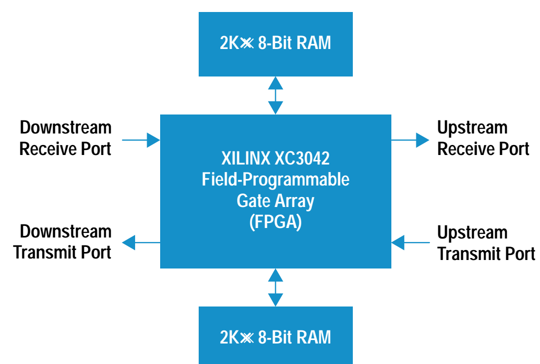


Fig. 4. Reconfigurable bit processor architecture.

© Hewlett-Packard Company 1995

sink. Recall that sources and sinks are where information is generated or analyzed by a communication system. In our case, sources and sinks are the means by which the test programmer communicates with the DUT (in traditional ATE terms, the sequencer or test pattern generator.)

The STS connects on one side to the internal STC bit stream format (as do all of our bit processors). The test programmer controls the transformation algorithm of the STS through a high-level programming language called the Serial Test Language (STL). The STS is implemented with a Motorola DSP56001 digital signal processor. The processor is well-suited to computationally-intensive transformation algorithms as well as general-purpose bit stream I/O.

Bit Processor Interconnect

When testing a multiplexed or hierarchical bit stream, the test engineer will typically choose to process (drive and receive) data for each logical channel independently of the others. At the same time, we have seen that it is often advantageous to be able to process a bit stream in a series of sequential steps. There is then a conflict, given a finite number of bit processors, between having a large multiplexing capability and having a large number of sequential processing stages. This is because multiplexed bit streams are best addressed by a wide, shallow array of bit processors, but multistep processing of a bit stream is best addressed by a narrower, deeper array of bit processors. Our RBP interconnect scheme and special resource assignment software minimize this conflict by allowing a wide variety of multiplexing and processing arrangements.

As shown in Fig. 3, there are several different interconnect paths between the RBPs and the STSs. Fig. 5 shows the topology of each possible interconnect combination. Multiplexing is supported by *channel splitter* circuitware loaded into two of the RBPs (the RBPs labeled 3 and 4 in Fig. 3). The channel splitters are labeled CS in Fig. 5. Resources can also be borrowed from adjacent processing channels, so additional branches can be attached to a personality module.

The system software manages this borrowing, minimizing the total resources required for a customer's test. If even wider multiplexing is needed, personality modules can be connected together at the same physical port of the DUT.

Performance

The STC serial processing pipeline was designed to accommodate maximum bit rates of up to 12.5 Mbits per second. There are obviously some applications that exceed this bit rate and cannot be addressed, but the performance of the architecture is a good match for proprietary telecomm PCM backplanes, automotive applications, LANs, asynchronous protocols at modem rates, ISDN basic rate interfaces, and many other similar applications.

Extensibility

The test capabilities of the STC are formidable, but there are still test cases that will require special capabilities or slightly different functionality. STC capabilities can be modified or increased by adding features to current circuitware, adding functionality to existing personality modules, developing new circuitware, or developing new personality modules. Since all of the circuitware is supplied with the system software and resides on disk, the first three of these can be achieved through simple software updates. The last requires a replacement of an existing personality module, but can easily be performed in the field.

Technologies

Several important technologies contributed to make the STC possible. The most important of these is the XILINX FPGA technology. Our architecture relies on the ability to load different circuitware (transformation algorithms) for each customer test.

To provide a great deal of functionality in as little space as possible, we wanted to use the smallest possible packages for the FPGAs and other parts. Surface mount technology was therefore a necessity, and fine-pitch surface mount was a very strong want. Unfortunately, because of mechanical registration limits, fine-pitch surface mount technology is extremely difficult to implement on large printed circuit assemblies like those used in the HP 3070 family of testers. To solve this problem, we decided to implement each channel of the STC architecture on smaller modules that would plug into the larger main printed circuit assembly. This provided two major benefits. First, we could use fine-pitch parts on these smaller modules, and second, trace routing on the main board was simplified considerably. Trace routing on the modules was still quite dense, but it only needed to be done once.

We also relied heavily on digital simulation technology. All of the circuitware developed for the STC was simulated and debugged before attempting bench turn-on. It was important to minimize debugging on the bench because of the highly

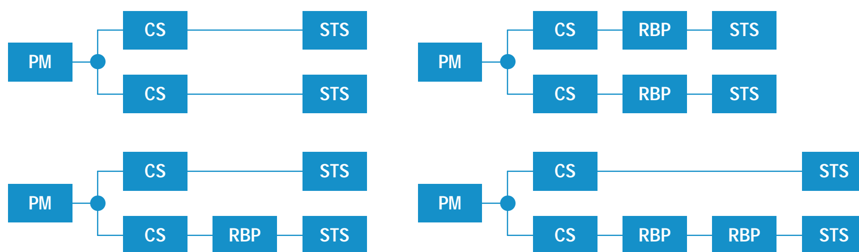


Fig. 5. Processing channel configurations.

© Hewlett-Packard Company 1995

integrated nature of the design. Although it is possible to probe internal FPGA nodes on the bench, it is much easier to probe during simulation.

Early in the project we also experimented with board-level simulation. The primary module, containing the STS and four RBPs, was simulated as a whole. These simulations were mostly intended to verify the circuitry surrounding the DSP56001 and to verify the RBP interconnect scheme. We decided not to simulate the entire STC card, mostly because of the lack of complete simulator models. In retrospect, this may have been a mistake because the vast majority of defects on the first STC prototype were in areas of the circuitry for which off-the-shelf models were available.†

Software Architecture

Up to now, we have discussed the generic model of serial communication systems and the hardware architecture based on that model. This section will provide a software overview of the Serial Test Language (STL).

Design Goals

To set the stage for the software discussion, we'll review the guiding design objectives for the STC and STL.

The process of studying serial bit streams, developing the generic model, and implementing our serial test architecture originally grew out of a desire to ease the test programming problems presented by serial-oriented DUTs. This led to our first design goal: to shorten the test development time by a factor of ten for DUTs with serial interfaces. As described earlier, existing test systems use a single parallel sequencer for controlling several serial bit streams. The resulting programs become very cumbersome and complex to create and maintain.

To address this problem, the STC uses a multiple-processor architecture to subdivide the overall programming task. With STL, we did not hide the hardware architecture from the programmer. Rather, we designed the STC user interface software to embody the generic serial protocol model. The software allows the programmer to segment a serial bit stream into logical pieces. Each logical piece can be programmed independently. This allows the programmer to concentrate on specific functions without keeping track of all the additional overhead found in the bit stream.

The hardware and software were designed concurrently. Several hardware changes were made to allow the software interface to better match the generic model. Many of these changes were facilitated by use of the FPGAs since they could be easily redesigned without a printed circuit board revision.

The concurrent nature of our architecture not only greatly eases the test programming burden, but also provides a clear test throughput advantage for multichannel DUTs. All channels of the DUT can be tested simultaneously, dramatically reducing the test time, especially for long bit error rate tests. We quantified this concurrent test strategy as our second design goal: to improve test throughput for multichannel DUTs by a factor of ten.†† We decided to address this problem through a parallel test strategy. The sequencer was carefully integrated into a complete ATE system which supports the use of many parallel sequencers. The software challenge was to provide an easy-to-program system capable of supporting many concurrently executing processors.

User Environment

Typical DUTs require system resources such as power supplies, nonserial digital drivers and receivers, and other signal sources and detectors to recreate the DUT's normal operating environment. The platform providing these system resources is the HP 3070 family of automatic test equipment.

The HP 3070 family supports both in-circuit and functional styles of testing.††† Both styles can be used together or separately on the HP 3070 system. The STC was designed primarily for board-level functional testing, but can be used for device-level functional test.

The software user interface of the HP 3070 supports the entire process of creating a suite of tests for a DUT. The in-circuit test development process involves describing the components and connections to the system. Using this information, the system generates the information required to build a mechanical interface between the DUT and the system. This information is also used to generate individual tests for each component.

The functional test process is similar. Only the edge connector need be described instead of all the components. The user then creates resource libraries describing the connection of HP 3070 resources to the DUT edge connector. The system uses the libraries and edge connector description to generate the mechanical interface description.

The highest-level user interface on the HP 3070 system uses the HP VUE environment running on the HP-UX* operating system. Specific portions of the test hardware are controlled by textual languages. These languages are similar in structure and syntax conventions and were designed for the specific purpose of testing DUTs. Overall test sequencing is controlled through a textual interface running an interpretive editor for the HP Board Test BASIC (BT-BASIC) programming language.

The standard digital sequencer is controlled by using the Vector Control Language (VCL). This sequencer provides parallel digital capability. The analog sources and detectors and external instrumentation access are controlled using the Analog Test Language (ATL). Both languages can be used together to test a particular DUT. A graphical Motif/X11

† This part of the design consisted mostly of discrete logic gates and flip-flops, so it was quite similar in character to the RBP circuitware designs. Experience has shown that these types of designs are more prone to design defects than higher-level, "cookbook" designs.

†† The choice of a factor of ten as a goal was somewhat arbitrary because the actual throughput improvement depends on the number of DUT channels. An eight-channel DUT might be tested only eight times faster, whereas a sixteen-channel DUT might be tested sixteen times faster.

††† In-circuit testing refers to methods for testing the various components of a DUT separately while they are in place on the board. This form of testing is based on the assumption that testing all components and connections between components ensures that the DUT as a whole has been properly tested. This form of testing produces excellent fault diagnosis to the failing component for repair. Functional testing refers to methods for testing a DUT by emulating the system environment into which it will later be integrated. Many DUTs have an edge connector interface, so this is also commonly called edge connector functional testing.

© Hewlett-Packard Company 1995

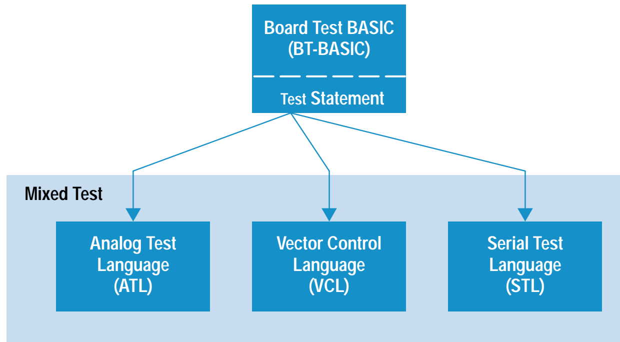


Fig. 6. Structure of HP 3070 testing languages.

environment supports development of VCL and ATL tests. Fig. 6 illustrates this software structure.

All textual sources are compiled before execution. The test is then executed from the BT-BASIC environment by use of the test statement. BT-BASIC is also used to control test resources like power supplies, test system operator interfaces, test results capture, and so on.

Serial Test Language Overview

STL was designed specifically for testing serial interfaces. Typically, these interfaces require the following capabilities:

- Simple, flexible input/output. All serial I/O can be segmented into a collection of bits called *frames*. The size of the frame varies depending on the application.
- Frame and bit manipulation features including comparison, modification, formatting, and concatenation.
- Complex conditional constructs that execute in real time. For many serial protocols, this means being able to perform 10 to 20 statements within 125 microseconds.
- Support for conversion of frame data to and from numeric data types. Many protocols require numeric processing of signals or numeric control.
- Support of multiple STCs running concurrently emulating a single serial bit stream or multiple serial bit streams.
- Triggering capabilities between concurrent processes (i.e., serial test sequencers) and the existing HP 3070 digital sequencer.
- Run-time control of STC personality modules and reconfigurable bit processors.
- Extremely flexible setup of STC personality modules and reconfigurable bit processors (RBPs). New RBP personalities and personality modules have been designed after initial product release. The user interface is designed to accept these with minimum software rework and without disturbing existing customer tests.
- Easy programming of the clock signal sources on the STC. These clock signal sources are used to simulate the bit and framing clocks found on many serial interfaces.

Each serial test for a particular DUT is contained within one source file. This source file controls all of the STCs required for that test. The number of tests depends on the level of diagnostics desired by the user. Generally, the user will write multiple tests to exercise the DUT fully.

All major units of the serial source code are organized as blocks. Each unit begins with a starting block statement and is terminated with an ending block statement. This greatly

```
serial;Count1,R1_frames

serial clock IOM_clock  is 2048 events
  events every 61.03515625n internal
  connect clk1 to "DCL"
  connect clk2 to "2048"
  at event 0 set clk1 to "1100"
  at event 0 set clk2 to "11110000"
end serial clock

stream IOM2_Master  type "synchronous"
  connect "tx_clock" to "2048"
  connect "tx_data" to "DD"
  transmit length unknown
  substream TX_All
    filter "xdlc CRC-CCITT"
    transmit bits all
  end substream
end stream

stream IOM2_Slave  type "synchronous"
  connect "rx_clock" to "2048"
  connect "rx_data" to "DD"
  set "rx_clock edge" to "falling"
  receive  length unknown
  substream RX_All
    receive bits all
    filter "xdlc CRC-CCITT"
  end substream
end stream

process TX_All
 loop
   transmit "\h5555"
   transmit "\haaaa"
 end loop
end process

process RX_All;Count1,R1_frames
 dim AAAA$[16]
 loop
    receive into AAAA$,"\hxxxx"
    exit if AAAA$ = "\hAAAA"
 end loop
 Count = 0
 loop
    receive "\h5555","\hxxxx"
    receive "\hAAAA","\hxxxx"
    Count = Count + 1
    exit if Count = Count1
 end loop
 R1_frames = Count
 initiate trigger digital
end process
```

Fig. 7. An example of a Serial Test Language (STL) program.

simplifies parsing and dramatically improves error diagnostic messages.

The overall structure for a serial test is illustrated in the example shown in Fig. 7.

The serial statement, serial;Count1,R1_frames, is used to define variables passed to or from the test execution environment (BT-BASIC). This allows flexibility in changing test execution parameters without needing to recompile. For instance, the user can pass in the number of seconds a bit error rate (BER) test is to be executed.

The optional serial clock block, serial clock IOM_clock ..., programs the STC clock sources to output specific clock signal patterns. Each clock section controls the four synchronous clock resources found on one STC. These clock generators can be synchronized to an internal or an external clock source. The clock signals can have variable lengths ranging from 2 to 65535 pattern changes (called events). The large number of events provides a very flexible format for defining

© Hewlett-Packard Company 1995

custom clock and frame sync signals. The width of each event is defined in nanoseconds or microseconds.

The stream block, stream IOM2_Master type "synchronous", is used to define the physical characteristics of the serial bit stream and protocol to the corresponding personality module. There are three personality module types: TTL, ISDN SBUS, and ISDN UBUS. The ISDN personality modules are used specifically for connection to ISDN interfaces. The TTL personality module is a flexible collection of programmable TTL-voltage level drivers and receivers used to interface to generic TTL-level serial bit streams. The stream type determines the personality module's mode of operation. In the above command, the keyword "synchronous" programs the personality module bit processor to expect an explicitly clocked bit stream. This type of serial bit stream requires a TTL personality module. An error would be signaled if this keyword were used and a TTL personality module didn't exist. The electrical levels of the personality modules are fixed, but the stream protocols are programmable. The example in Fig. 7 is an explicitly clocked and implicitly framed bit stream. Therefore, the TTL personality module uses two signal lines to receive data: rx_data and rx_clock. The explicit framing signal line, rx_frame_sync, isn't required by the bit processor because the HDLC format has the framing information embedded in the actual data transmitted. Different TTL modes are used to reprogram the bit processor to emulate a variety of serial protocols ranging from synchronous TDM interfaces to asynchronous interfaces like RS-232.

The mode of the ISDN personality modules can be changed as well. For example, the ISDN S-bus module can simulate terminal equipment or a network terminator. The automatic ISDN synchronization sequences are different depending on the ISDN mode selected. The mode type reprograms the personality module's circuitware to conform to different stream requirements.

The programmer connects the personality module to the DUT by using the connect statement, connect "tx_data" to "DD". This statement physically closes the correct STC relays to connect a personality module's resources to the DUT. The personality module's resources are multiplexed. During the test generation process, the system software will automatically determine the optimal connection method to the DUT. This connection method is then used to build the correct fixture interface to the DUT.

Each personality module mode has a set of programmable features. These are controlled by use of set statements: set "rx_clock edge" to "falling". The set statement is quite generic:

set "mode description" to <value>

The <value> parameter can be a numeric or string identifier like "enabled". Each mode has a specific set of features. If a feature is not explicitly set, then a default value is used.

By definition from our generic serial protocol model, each stream will have a structure of bits called a frame. A frame of bits is continuously (or on demand) transmitted or received. Depending on the protocol, we may or may not know the frame length. Certain serial protocols have the frame length embedded in the actual stream of logical bits; HDLC is an excellent example.

The transmit frame length and receive frame length statements are used to capture this information. This may be unknown (which is allowed as a keyword) or a number between 2 and 4096. This information is used by the channel splitter circuitware as it inserts or extracts bits from the stream.

This leads us to the next block structure, the *substream*. Each stream can have between one and four substreams. The substreams are defined as a block structure within each stream block. The substream programs the channel splitter to target specific bits within the frame. Each substream can receive and transmit bits from within each stream frame. The bits targeted by the channel splitter are therefore either extracted or inserted into the stream frame. Each substream has an associated serial test sequencer called a *process*. Associated substreams and processes have the same name. The substream defines which bits of the frame are passed to or received from the process. The process contains the actual program used to control the bit values.

If the stream frame length is known, then each substream is defined as a portion or all of the stream frame. Particular bits in the stream frame are enumerated from 1 to the frame length. The user tags certain stream bits for the substream to transmit or receive by use of the transmit or receive bits statements: transmit bits all, transmit bits 1 to 8, receive bits all. Multiple transmit or receive bits statements accumulate tagged bits. All these bits are concatenated (by the STC hardware) in order (bit 1 to bit n) and treated as a frame by the process associated with that substream.

If the stream frame length is unknown because it must be recovered from the bit stream, as is the case shown in Fig. 7, then the programmer must define a reconfigurable bit processor to transform the bit stream. This is identified in the substream block as a filter statement or block, filter "xdlc CRC-CCITT". Set statements can be used to control the filter in much the same way as set statements are used to control the personality module in the stream block.

Our generic model allows infinite levels of bit stream hierarchy. The STC hardware supports two levels (stream and a single layer of substreams). Additional levels can be emulated in the serial language. This has not proved too restrictive since most protocols require no more than two levels. Refer to Fig. 2 for an example of how this works.

Multiple substreams can receive the same bits, but only one substream can transmit a particular bit. A substream may have no bits being transmitted or received. This capability is used to create a process that simply controls other processes.

As stated previously, each substream has an associated process, process RX_All;Count1,R1_frames. Each process represents an independent program that is executed concurrently with all other processes defined in a particular serial source program. Variables passed into the serial test are subsequently passed to individual processes. The STL process statements were modeled after BT-BASIC. Therefore, the statements and structures allowed in STL processes include:

- If-then conditionals
- Logical operators
- Assignment
- Loop/exit if/end loop construct

© Hewlett-Packard Company 1995

- Subroutines (with data scoping control)
- Bit error rate test functions
- Numeric functions
- Access to stream and filter features through control and status functions.

Bits to be transmitted or received are formatted as a data type called a *frame*. Frame data can be stored in a variable, A$, or as a constant, such as 10001. Each frame variable is simply a linear array of packed bits. From a programming perspective, the frame variable is treated like a string data type. Each frame variable has a maximum length declared by the dim statement. The default length is 16 bits.

Data is transmitted by the transmit statement, transmit "\h5555", A$, or received into the process by the receive statement, receive "\h5555", into A$.

All process frames are buffered internally within the serial test sequencer. These buffers allow the serial test sequencer to continue to execute the program and not be tied directly to the DUT transmit or receive rate.

All string operations like substring, concatenation, equality, insertion, and deletion are supported with the frame data type. Each bit of a frame works just like a character in a string. The user can use various notations (binary, octal, hexadecimal, decimal, or ASCII) to assign a value to a frame variable. Since string operations were modeled after those in BT-BASIC, programmers familiar with BASIC languages can quickly learn STL frame operations.

An STL process also supports integer, real, integer array and real array data types. Conversion between integers and frame data types is supported. These data types are used to support complex control algorithms and digital signal processing of data. One application requiring this capability is testing analog line cards. Line cards interface subscriber equipment, such as a phone, to the public data network. Typical tests require a significant amount of digital signal processing for testing analog-to-digital transmission transfer functions.

Time-order sequence control between processes and the digital sequencer is handled by use of level-sensitive, named triggers. The triggering mechanism has been implemented using a simple mailbox approach. The process sending a trigger places the appropriate trigger number in its assigned mailbox. Processes receiving triggers are told in which mailbox and for which number to look. This implementation allows any process to send or receive a trigger to or from any other process. It also allows multiple processes to receive the same trigger. Ten different triggers can be sent to any other process defined in the serial source. In addition to interprocess triggering, the system supports triggering to and from the parallel digital sequencer. For example, Fig. 8 shows two processes, A and B, sending and receiving triggers.

This simple trigger scheme allows a wide variety of sequence control among concurrently executing processes and the HP 3070 digital sequencer.

All statements in the process section are optimized for speed. Testing and use have shown that STL can perform a fairly complex set of operations in 125 μs† to a few milliseconds.

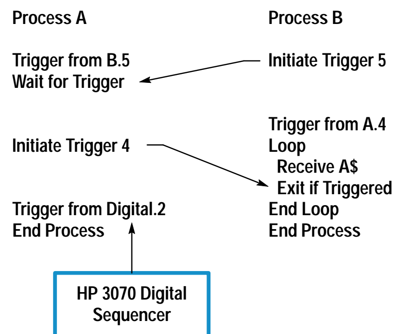


Fig. 8. Triggering in STL.

Debugging Serial Tests

The HP 3070 test debugging environment consists of a Motif/X11 interface that communicates with the HP 3070 system hardware through the BT-BASIC interactive editor.

The original debugging environment supports both the digital sequencer and the analog measurement subsystem. It provides an easy-to-use pull-down menu structure to control the analog and digital test functions, view digital test vectors graphically within a logic analyzer display, create measurement histograms, and so on. A serial mode was added to allow the user to view the status of variables, processes, connections, and so on. All modes support viewing of the textual source code and commands to execute the source program and view the data. If necessary, the source program can be modified, recompiled, and executed from the debug environment.

Fig. 9 shows the serial debug environment. The largest box contains the serial source program. It can be modified by the user in this pane. The compile-and-go button allows the user to quickly†† recompile just the serial source code and execute the test. The pane on the left side contains a list of STL processes. Clicking on one of these places the source program at the first line of that process.

The command pull-down shows the list of debugging commands available, such as viewing the current contents of variables, trigger log, current process status and line number, and status and contents of the transmit and receive buffers.

By various pull-down menu selections, specific variables or groups of variables of a particular type can be displayed. Frame variables can also be displayed in a variety of formats (binary, octal, hexadecimal, decimal).

The trigger log on each processor captures the last 20 trigger events. The events are displayed in chronological order. This helps resolve difficulties when triggering between processes or the digital sequencer. Each trigger event is displayed in the same syntax as the trigger statements in STL.

Each process keeps track of its own status and line number. The status debug command displays the current status and

† 125 μs is an important number in telecommunications applications because it corresponds to the basic 8-kHz frame rate used throughout the network.

†† Usually in under 15 seconds. The largest serial test to date takes 44 seconds to compile.

© Hewlett-Packard Company 1995

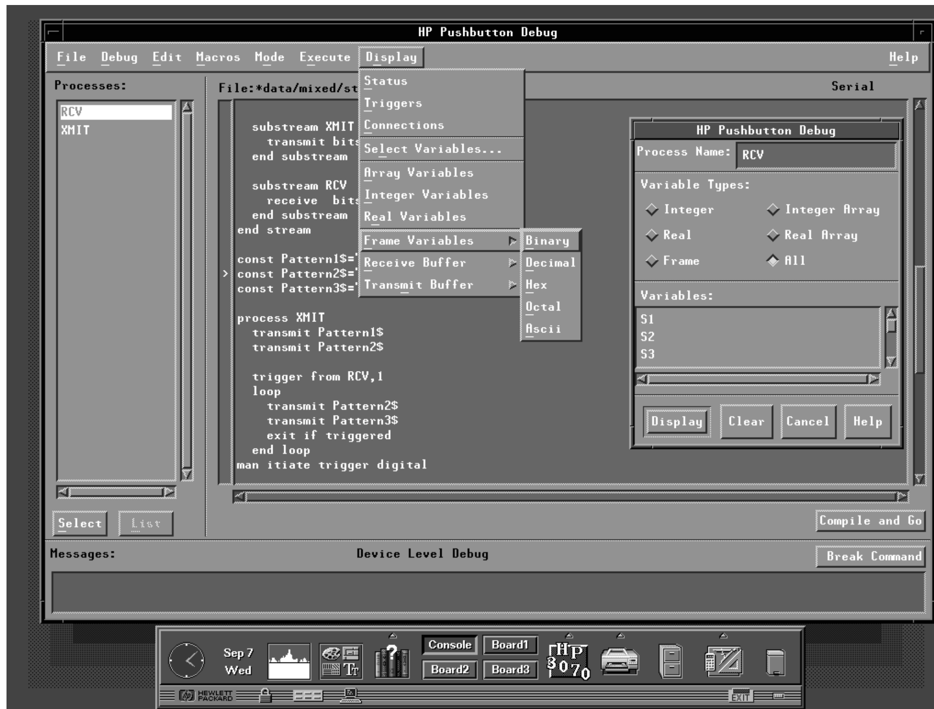


Fig. 9. Serial debug environment.

places the source pane at the line number the process was last executing.

Special commands are available to display the contents of the transmit and receive buffers. This allows the user to see what has been recently transmitted (in the transmit buffer), what was about to be received in STL, and what has been received by the STC hardware.

Breakpoints can be set by the user by means of the halt serial failing statement. Single-step execution is not possible because of the real-time nature of the STC. The serial test must execute until either the test ends normally or an exception occurs. If any exception occurs, all serial processes are immediately halted (this is implemented by a hardware signal) and the current status of all serial processes is available.

It is possible to switch to the digital (VCL) or analog (ATL) test modes by the mode pull-down menu. The interface for these modes is very similar to the STL interface.

Implementation

The STL compiler was written using C++ and object-oriented programming techniques. In an attempt to improve our software development process, the STC/STL group set several goals:

- Learn and use object-oriented programming development techniques (using C++ as a development language).
- Leverage libraries and tools as much as possible.
- Dramatically improve the quality of our products.

Object-Oriented Programming. When we started, object-oriented programming was a new technique in the realm of software development. After investigation and some training, we knew this technique was not going to be easy or immediately rewarding. In retrospect, the definition of classes is a time-consuming task, but is critical to the success of a project.

Use of an object-oriented design provides two primary advantages: better hardware abstraction and extensibility. An object-oriented abstraction models a given hardware construct like the STC quite naturally. Object-oriented programming allows a hierarchy of classes defining the various hardware levels to be created easily.

The abstraction of object-oriented programming made the STL compiler design much easier and more readable. Both the language constructs and the STC hardware are encapsulated in a hierarchy of classes. They meet at the code-generation phase of the compile, where language constructs (which are simply an instance of the statement class) invoke a message to the hardware code-generation classes.

An enhancement to the original software was the support of real and array data types. Typically, the addition of these data types would have taken approximately one engineering year. Because of the object-oriented programming benefits, the addition took only three engineering months.

Leverage. Leverage of existing tools and libraries saved a great amount of time and effort. We leveraged many areas of the STL/STC software development. We used the Codelibs library of C and C++ classes for data abstractions and algorithms. We used a third-party assembler for the STS sequencer programming. Debugging assembly code during development of the STL compiler was far easier than editing binary downloads. We eventually replaced the assembler with a direct binary output to speed up the compiler, but we left the assembly output mode as a debugging tool. We used wacco, a top-down recursive descent parsing tool, to simplify the parser and improve error handling, and we used and improved internal C++ classes for error reporting and file I/O.

The quality of the existing tools or libraries should be investigated carefully before leveraging them into the product. We

© Hewlett-Packard Company 1995

did not suffer any problems, but we were careful with our dependency on leveraged code.

Software Quality Assurance. There is no shortcut to quality. 30% of the product development time was spent in quality assurance.

There is no doubt that defects are far more easily fixed in earlier phases of software development, but a high-quality product is not ensured until the boundary conditions have been tested. The object-oriented programming design process and leverage of high-quality software components contributed greatly, but we also spent a great deal of effort in testing our software.

Our testing effort focused on two areas: early users (alpha sites) and automated regression testing. We developed tests on several customer DUTs and used two alpha sites to build our confidence in the ability of the STC to test serial DUTs and meet our project goals. We also created a set of tools that allowed our group to develop over 900 automatic regression tests. We used branch flow analysis† to refine these tests to get to 92% coverage within the STL compiler. The regression test suite is also used to verify that a particular change or defect fix has not introduced any other defects.

STL Summary

The key STL objectives were to shorten test development time and to increase test throughput for DUTs using serial devices. The ability to divide and conquer the serial bit stream using multiple processors has proved very successful in reducing the time to implement tests. In some cases, test development has been reduced from 4 to 6 months to 1 to 2 weeks. The ability to easily test multiple bit streams concurrently increases test throughput dramatically, especially in BER tests.

Customer Application Case Studies

The sequencer architecture we have described in this paper was derived from studies of our generic model of serial communication systems, which was itself developed from the study of a wide variety of serial protocols and DUTs. To test the effectiveness of the new architecture, we wrote functional tests for many serial protocols and customer boards. Two customer boards used in these case studies are described in more detail below.

Case Study I

The first customer board was a telecomm multiplexer. The board is used to multiplex and demultiplex four 2.048-Mbit/s bit streams to and from a single 8.448-Mbit/s bit stream. Converters in the fixture were used to translate the HDB3 signals in these streams to TTL-compatible levels. An additional serial bit stream is used to send and receive control information to and from the board. Each of the six serial bit streams was attached to an STC processing channel.

The customer had been manufacturing this particular board for five years, and had brought up their suite of functional tests for the board on two previous testers, both of which used a traditional pattern sequencer. Their test suite consisted of 15 functional tests, and they reported test development times of nine months and five months using the two previous board testers. Using the STC, they implemented their 15-test suite and 16 additional tests in four days—a 25-fold decrease from their best previous test development time. The tests implemented included a BER test on all four channels simultaneously and other tests they simply could not perform with the sequencer architecture available on the other testers.

Case Study II

The second customer board was an ISDN U-interface central office line card. This board has an ISDN U-interface for each of four subscribers and a serial backplane interface. Each of the five bit streams was attached to an STC processing channel. At the time this test was developed, we had not yet introduced our U-interface personality module, so commercial network terminators were used to translate the subscriber channels to ISDN S-interface format. The backplane of this board is a 2.048-Mbit/s serial bit stream conforming to the IOM-2 protocol.

Each subscriber port was split by the STC processing chain into two substreams, one to handle data transfer and the other to handle activation control of the ISDN interface. The backplane bit stream was also split into substreams, but in this case we chose to use a single process to handle all four control channels. This reduced the total number of STC channels that would otherwise have been required and did not overly complicate the test program. We also handled the four data channels with a single process, using a special, optimized BER function built into STL. This function can receive up to 32 independent BER data streams simultaneously with no intervention or special programming required of the test engineer.

Other Applications

During product development, we tested the architecture against many other serial protocols and formats, including ISDN S- and U-interfaces, RS-232-style asynchronous protocols, CEPT-30, T1, automotive serial interfaces, I^2C, HDLC control channels, generic 64-kbit/s bit streams, IOM-2, ST-Bus, and various TDM backplanes. In each case we were able to communicate with the bit stream with a minimum of programming time and effort.

Summary and Conclusions

Based on the case studies and other applications described above, we have found that when testing serial-oriented DUTs, the new architecture offers the following advantages over traditional sequencers:

- Much faster test development
- Much better test coverage (more functionality of the DUT can be tested more easily)
- Much better throughput (because of the ability to test multiple channels of a DUT simultaneously)
- A reduction in fixture electronics.

† This tool inserts probes into the source code that allow reporting of coverage of particular execution paths within a program.

© Hewlett-Packard Company 1995

The only potential disadvantage of the architecture is a slight increase in the capital cost of the test system. The relative weight of the advantages and disadvantages is determined by the type of DUT being tested and the type of test being run on that DUT. When testing boards with multiple identical channels, especially when the board tests include long conformance tests like BER, the n:1 increase in throughput one can achieve using n STC channels easily offsets the slight price premium of the hardware. Board tests that do not include long conformance tests and that involve significant overhead because of handling or heavy in-circuit testing will not see such a clear cost advantage, but even then the significant reduction in test development time may offset the cost of the hardware.

Acknowledgments

We would like to thank the following people for their work on the STC/STL project: John Siefers for the design and development of the STC main board and personality modules, John Algiere for his system design work, Greg Stander and Bud Cribar for the design and implementation of the serial test compiler, Eric Waldheim for the design and implementation of the DSP56001 assembler routines, Sunit Bhalla for writing the hardware confirmation and diagnostics test routines, Wilson Spence for his always enthusiastic suggestions and feedback, and Cullen Darnell and Lynn Schmidt for their management leadership.

References

1. R.E. McAuliffe, "Practical Production Testing of ISDN Circuit Boards," *Proceedings of the IEEE International Test Conference,* 1988, pp. 39-46.
2. J.T. Healy, *Automatic Testing and Evaluation of Digital Integrated Circuits*, Reston Publishing Company Inc., 1981.
3. CCITT, "Integrated Services Digital Network (ISDN), Overall Network Aspects and Functions, ISDN User-Network Interfaces," *CCITT Blue Book, Recommendations I.310-I.470, Volume III, Fascicle III.8*, 1988.
4. W. Stallings, *ISDN and Broadband ISDN*, *Second Edition*, Macmillan Publishing Company, 1992.
5. H.S. Stone, *Microcomputer Interfacing,* Addison-Wesley Publishing Company, Inc., 1982.
6. D.N. Chorafas, *The Handbook of Data Communication and Computer Networks,* Petrocelli Books, Inc., 1985.
7. R.E. McAuliffe, "Board Testing Modern DUTs: Solving the ISDN Test Challenge," *Hewlett-Packard internal communication*, 1988.

HP-UX is based on and is compatible with Novell's UNIX® operating system. It also complies with X/Open's* XPG4, POSIX 1003.1, 1003.2, FIPS 151-1, and SVID2 interface specifications.

UNIX is a registered trademark in the United States and other countries, licensed exclusively through X/Open Company Limited.

X/Open is a trademark of X/Open Company Limited in the UK and other countries.

Motif is a trademark of the Open Software Foundation in the U.S.A. and other countries.

© Hewlett-Packard Company 1995

Shortening the Time to Volume Production of High-Performance Standard Cell ASICs

Coding guidelines for behavioral modeling and a process for generating wire load models that satisfy most timing constraints early in the design cycle are some of the techniques used in the design process for standard cell ASICs.

by Jay D. McDougal and William E. Young

As time-to-market pressures continue to increase, the need for shorter design cycle times is more urgent than ever. At the same time, the demand for high performance in standard cell ASICs is also increasing. These trends are expected to continue as customers look to get the most return on investment from the latest IC process technologies. Many of the design considerations needed to achieve these high-performance goals compete directly against achieving quick time to market.

A typical standard cell design process includes several iterations in which individual steps must be repeated to adjust for data determined at later steps. A common example of this is the need to redesign portions of the circuit when physical results such as extracted parasitics are fed back into a simulator and performance goals have not been met.

This paper presents methods that help reduce or even eliminate the need for design iterations by increasing the chance of "first time perfect" at each design step. First, we discuss the methods developed for each of the steps in our ASIC design process (see Fig. 1) to get shortened throughput time and reduced design iterations while still producing high-performance components. Next, we present the results of applying these methods to the design of a CMOS ASIC that is used in HP's X-terminal products.

Behavioral Modeling

Since high-level behavioral modeling is done very early in the design flow, the way in which the code is written can have a dramatic effect on the downstream processes. We have developed a set of hardware description language (HDL) guidelines that, if followed, will reduce throughput time for the entire design flow. These guidelines were collected from several designers within HP and have been updated and modified as areas for improvement have been identified in each of the downstream processes.

The HDL coding guidelines include sections on clocking strategies, block hierarchy and structure, flip-flops and latches, state machines, design for test, techniques for ensuring consistent behavioral and structural simulation, and Synopsys-specific issues (Synopsys is an automatic design synthesis tool from Synopsys, Inc.).

Following these guidelines allows us to avoid many time-intensive steps later in the design process such as name remapping, race analysis, and iterative structural and behavioral simulation.

Synthesis

Our goal during synthesis is to be able to produce a design that, when routed, will meet performance goals with the smallest possible area. We also want to do this with as few iterations in the behavioral model and the synthesis scripts as possible.

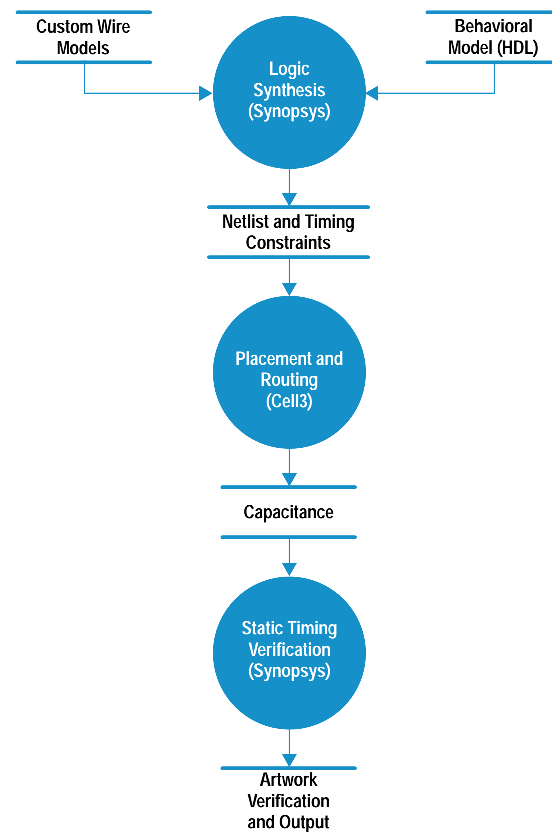


Fig. 1. Basic ASIC design flow.

© Hewlett-Packard Company 1995

Coding Guidelines. The HDL coding guidelines mentioned above enable the creation of behavioral models that synthesize predictably in a shorter amount of time with better performance than those created without guidelines.

Generic Synthesis Script. To streamline synthesis we create generic synthesis scripts for our design that contain all of the design-specific constraints such as flip-flop types, clock speed, behavioral model locations, and so on. These scripts also include input and output timing constraints, external loading, and drive capability. The scripts are used to synthesize all submodules in the design from the bottom up. Only those modules that do not meet timing requirements when incorporated into their parent module are resynthesized with scripts written specifically for them. This allows the majority of the blocks to be synthesized very quickly. Fig. 2 shows a portion of a generic synthesis script.

Custom Wire Load Models. One of our primary goals is to perform a single route without any iteration. To accomplish this, the wire loading (capacitance on the line) estimates that are used during synthesis have to be conservative. However, if they are too conservative then it is not possible to meet performance goals. To determine a wire loading model with the appropriate amount of conservative estimation for our library and tool methodology, we performed several wire loading experiments. These experiments consisted of synthesizing modules of various sizes and design types, routing them, and then verifying that their performance with actual wire loads satisfies the timing constraints defined for the module. Several passes were done for each module using a wire load model with varying levels of conservatism. Fig. 3 summarizes the process we used to generate the wire

```
/* These are fragments from the generic Synopsys dc_script */
/* the full script can be easily modified for a given model */
/*  -------------------------------------------------------------  */
/*  Read in Verilog HDL:                                 */
/*  -------------------------------------------------------------  */
read -f verilog mymodule.v
check_design

/*  ----------------------------------------  */
/*  Set the wire load model:                  */
/*  ----------------------------------------  */
set_wire_load block –library wire_loads

/*  ----------------------------------------  */
/*  Define the clocks:                        */
/*  ----------------------------------------  */
create_clock CLK –period 20 –waveform {0 10}
set_clock_skew –plus_uncertainty 0.5 –minus_uncertainty 0.5 –propagated CLK

/*  -------------------------------------------------------------  */
/*  Set block operating environment:                               */
/*  -------------------------------------------------------------  */
loader_pin = hp_cmos26g_table_slow/INVFF/A
driver_pin = hp_cmos26g_table_slow/NINVFF/Q
set_load 3 * load_of(loader_pin) all_outputs()
set_load load_of(loader_pin) all_inputs()
set_drive drive_of(loader_pin) all_inputs()
set_input_delay 10 –clock CLK all_inputs()
set_output_delay 10 –clock CLK all_outputs() –max

/*  -------------------------------------------------------------  */
/*  Compile the design
/*  -------------------------------------------------------------  */
compile

/*  ---------------------------------------------  */
/* Write out results                               */
/*  ---------------------------------------------  */
write –hierarchy –f verilog –output mymodule.vopt
write_constraints –format sdf –output mymodle. sdf –max_paths 1000
report_design >> mymodule.sn_rpt
report_hierarchy –full >> mymodule.sn_rpt
report_timing >> mymodule.sn_rpt
```

Fig. 2. A portion of a generic synthesis script.

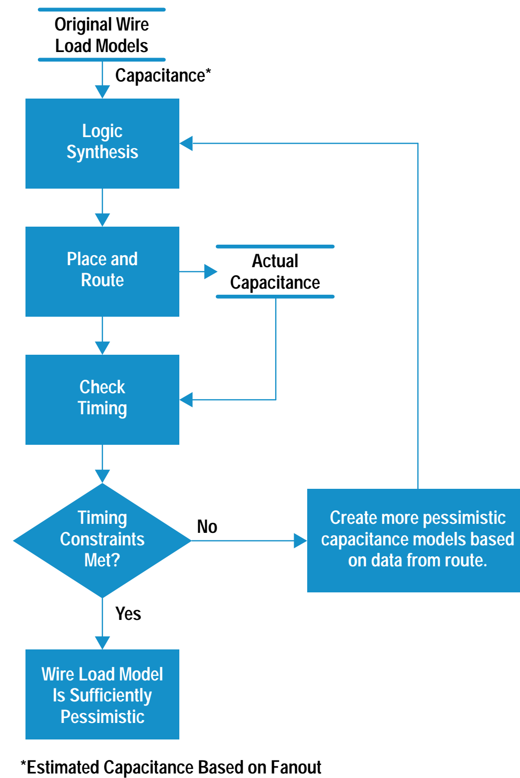


Fig. 3. The process for generating wire load models.

load model. This process was necessary because initial wire load estimates might be too optimistic. For example, the initial capacitance estimate for one connection between two inverters might be 0.04 pF, but after placement and routing the actual capacitance could be 0.1 pF, which might cause the chip's timing constraints not to be met.

During these experiments, each module was first synthesized using average wire load estimates. After routing, if the modules failed timing, additional experiments were performed using a progressively more pessimistic wire load model. The wire load model that was used was generated by selecting a point in the distribution of actual capacitances for each fanout that is greater than a given percentage of nets. Fig. 4 shows a sample distribution of interconnect capacitance for nets with a fanout of two in a typical module. In this example, we wanted to use a model that would predict a capacitance such that 90% of the wires would typically have actual capacitance less than the predicted value. To do this we simply used the capacitance value from the distribution that was greater than 90% of the other wire capacitances. This was done for each fanout to create a synthesis wire load model called a "90% model."

We used this method to test models that fell within the 50-to-95-percent range. We found that unless we used at least a 90% wire loading model we had some timing violations after back-annotation* that were not present during synthesis with estimated loads.

We also discovered that the magnitude and distribution of the routed capacitance were fairly consistent across a wide

* Back-annotation in this context refers to the process of taking the actual capacitance values extracted from routing and using them during static timing analysis.

© Hewlett-Packard Company 1995

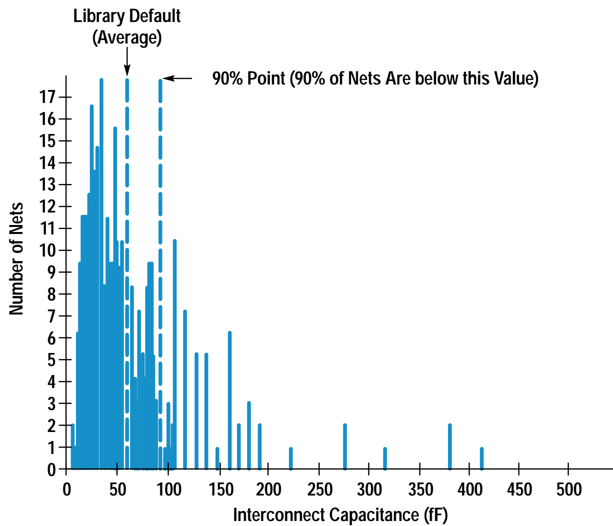


Fig. 4. A sample distribution of interconnect capacitance for a fanout of two in a typical module.

range of module sizes and design types. This was a surprise and it allows us to use the same wire load models for all designs done in the same library without having to repeat the experiment for each design.

We use the 90% wire load model derived from the above process to help guarantee single-pass routing for all modules using our library. In addition, we create another model for wires that are used for global* module interconnect. This is done using the same method described above except that the wires used to create the capacitance distributions are limited to global wires. This model more accurately predicts the higher capacitance associated with top-level interconnect. Synopsys, our logic synthesis tool, can automatically select the appropriate wire load model to use for each global wire.

Table-Driven Models. Another factor in our ability to achieve high performance and minimize cycle time is the accuracy of the Synopsys library timing models.

The biggest problem created by inaccurate synthesis timing models is the requirement for numerous iterations between synthesis and timing verification to fix timing violations. These iterations involve changing synthesis constraints, overconstraining, and replacing cells by hand in some cases. Timing inaccuracy also leads to poor optimization decisions, resulting in nonoptimal circuits in terms of performance and area.

These and other problems are essentially removed by the new nonlinear table-driven timing models in Synopsys. With these models, timing can be made to match the Spice characterization tables for each cell in the library. The transition time definition can also be made to match exactly so that it can be constrained properly.

The HP C26102SH library supports this new model and allows us to get accurate timing from our synthesis package so that there are no timing or operating condition violations in the timing verification step. Besides the obvious benefit of reducing iterations, this new library and timing model gives us performance and density improvements. The performance is improved because the synthesis tool is now working on the correct paths and is able to generate faster circuits. Density is also improved for the same reason. Since incorrect paths are no longer being optimized incorrectly and overconstraining is not necessary, the overall design size is smaller. In addition, the improvement in transition-time modeling and constraints avoids the need to apply a global transition-time constraint to the design which can lead to oversizing many cells.

Timing Constraints

As an additional method of ensuring that our performance goals are met without having to do multiple routing passes, we have added the process of driving the placement with constraints derived from synthesis.

Because Synopsys timing is very accurate with table-driven models, it can be used directly to create timing constraints imposed on the router. This is done using the Standard Delay Format timing output in Synopsys. Critical path timing is written in Standard Delay Format which can then be converted to Design Exchange Format and input to Cell3** with the netlist. These timing constraints become part of the overall constraint equation for the placer.

Several thousand constraints can be quickly and easily generated using this method. However, only those paths that are within a few percent of failing with estimated loads should be constrained. If an appropriate wire load model is used, the rest of the paths should meet their timing without being constrained. Having too many paths constrained will slow the placement process and may produce inferior results. However, we have successfully constrained up to a thousand paths.

Place and Route

Our goals during placement and routing are to implement the design in the smallest possible area, meet all specified performance goals, and minimize the number of iterations through the process.

Timing-Driven Placement. Timing-driven placement is the process of driving the placer with the timing constraints output from a timing analyzer (Synopsys, in this case). The following factors are important in successfully using this technique.

- Accurate timing models used for static timing analysis. As mentioned above, accurate timing models are critical to ensure that the synthesis program works on the right paths and that the timing constraints are accurate.
- Accurate cell delay information fed to the placer. Our placement program, Cell3, uses a two- or three-parameter delay equation. The two-parameter equation calculates delay with an intrinsic component and a load dependent component. The three-parameter equation modifies the intrinsic delay to reflect its dependency on the input slope. To drive the placer with accurate data, it is important that Cell3's delay calculation match that used by the timing analyzer as closely as possible. Table I shows the correlation obtained from each of these models compared to the data used by the timing analyzer. The data in the table is for 100 representative paths, varying in length from 4 to 71 cells.

* Global wires are the interconnecting wires between submodules on a chip.

** Cell3 is the placement and routing program we use, which comes from Cadence Design Systems.

© Hewlett-Packard Company 1995

Table I
Cell3 to Synopsys Correlation

	Timing Analyzer Data	Cell3 (two-parameter)	Cell3 (three-parameter)
Minimum Path	3.47 ns	3.56 ns	3.43 ns
Maximum Path	15.54 ns	18.51 ns	14.55 ns
Least Error versus Synopsys		+1.4%	0.0%
Greatest Error versus Synopsys		+26.2%	+7.4%
Error Range versus Synopsys		1.4% to 26.2%	−7.3% to 7.4%

As shown in Table I, Cell3's three-parameter model provides greatly improved delay modeling compared to the two-parameter model. Because of its improved accuracy, the three-parameter model is now used as the standard during timing-driven placement.

- Accurate estimates of interconnect. For synthesis, interconnect delay is specified by the 90% wire load models. For placement, it is important that the placer has a good idea of what it can expect in terms of average per-layer wiring capacitance for each signal. These numbers are determined by analyzing wiring distributions on previously routed chips.

Clock Tree Synthesis and Verification. Typical ASICS are driven by single or multiple high-speed clocks. These clocks drive thousands of flip-flops and must have low insertion delay and skew to meet performance requirements. To implement these balanced clocks, special placement and routing features must be used. Cell3's clock tree synthesis tools are used to insert a buffer tree for each clock (see Fig. 5). The clocks are then prerouted using various forms of Cell3's balanced routing techniques. This method has met with mixed success. Excellent results can be obtained by using the detailed balanced router, which is included as part of Cell3's clock tree synthesis option. However, since this tool has proven not to be as robust as we would like, we have found that an easier and much more robust method is to use Cell3's balanced global routing, which meets the needs of most high-performance standard cell ASICs.

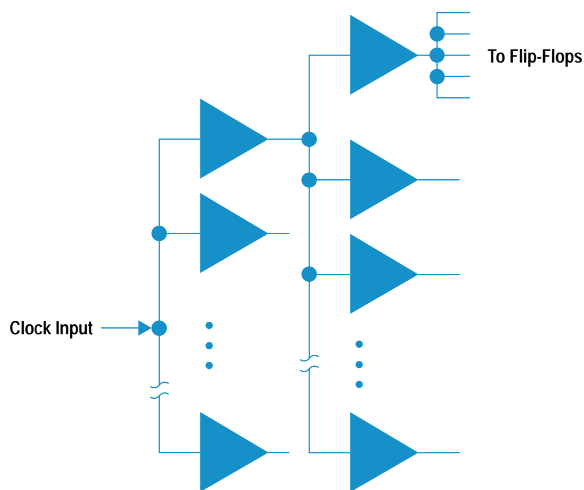


Fig. 5. A typical clock buffer tree. Insertion delay for this configuration is the average delay from the clock input to the flip-flops. Skew is the difference between the shortest and longest delays.

Actual clock delay and skew are verified using the RC extraction capability provided in CheckMate.* The general steps involved in using this approach include:

- Complete the physical implementation of the clocks.
- Extract (via CheckMate) RC information for all nets in the design.
- Create a Spice netlist containing only the cells and nets in the clock trees. (The desired nets and cells are specified to CheckMate's Spice writer.)
- Add custom circuit conditions (external loads, etc.).
- Run the Spice job.

Cell3 can also provide clock and skew information, but the steps listed allow us to obtain more accuracy and to build confidence in the Cell3 numbers.

Automatic Scan Insertion. Automatic scan insertion is done during the routing process. This makes it possible to include scan logic without having to design it in during behavioral coding. However, there are rules in the HDL coding guidelines that must be followed to make automatic insertion possible. Using automatic insertion reduces the complexity of the behavioral design and eliminates iteration because of scan clock skew, scan chain ordering, and timing performance issues. The use of automatic scan insertion is made possible by the use of an internal tool that performs insertion and optimization.[1]

Results

A chip in which the processes described above have been applied is a CMOS ASIC that is used in HP's X-terminals. Its main functions are memory and data path control. The chip contains 270,000 transistors and runs at multiple clock frequencies of 50, 100, and 33 MHz.

For this chip, the 1000 timing paths with the smallest timing margin were input as constraints fed to the Cell3 placer. Cell3 met all of the constraints on the first pass and subsequent timing analysis verified that all the paths were satisfied. More important, no other timing violations were created because of the constraints on the tightest paths. If that had occurred, we were prepared to exercise the Synopsys in-place optimization flow. This flow does in-place up and down sizing of cell drives as necessary to meet timing constraints. This step proved to be unnecessary for our X-terminal ASIC. The 90% wire load models were adequate.

The multiple clock domains on the ASIC are all derived from a 100-MHz input clock. The chip has three domains and each contains approximately 600 flip-flops. To meet performance goals, the skew across all of these domains had to be less than 0.5 ns and the insertion delays had to be matched within 1 ns. After placement and routing, clock delay and skew verification was done using the CheckMate RC extraction tool. Fig. 6 shows the Spice results obtained on one representative clock tree. All clock trees were found to have acceptable skew. This tight clock skew was a key factor in achieving the ASIC's aggressive performance goals.

* CheckMate is an artwork verification and parasitic extraction tool from Mentor Graphics Corp.

© Hewlett-Packard Company 1995

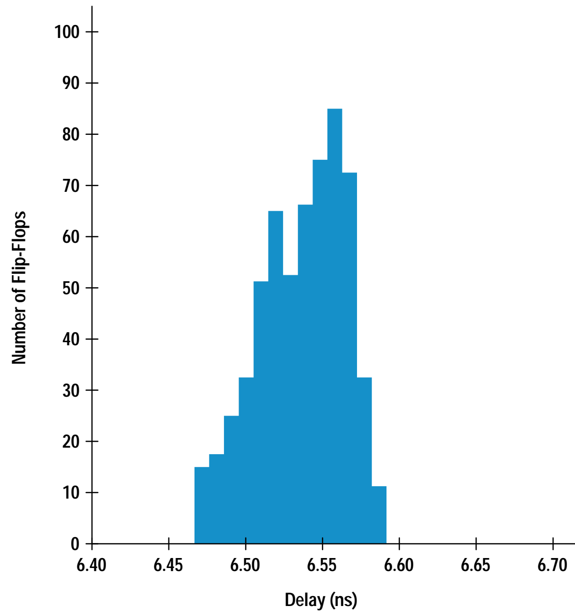


Fig. 6. Spice clock delay and skew results.

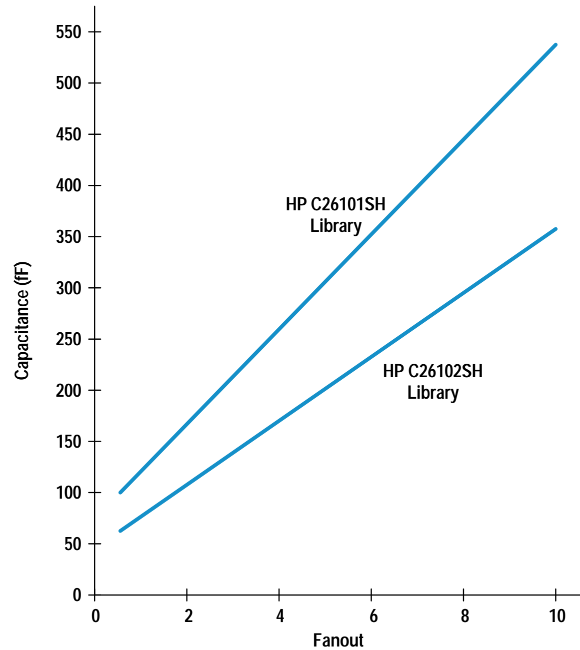


Fig. 7. Wire capacitance versus fanout in two cell libraries. The HP C26101SH library is the older CMOS26 library.

It is our goal to characterize Cell3's delay calculation to the point at which we can use its delay numbers for clock skew verification without the additional complexity of using Spice. Table II shows the correlation obtained between Cell3 and Spice for the clocks on the X-terminal ASIC.

Table II
Cell3 to Spice Correlation

	Cell3	Spice	Delta
Best Case	2.84 ns	2.56 ns	–9.9%
Worst Case	2.96 ns	2.57 ns	–13.2%

The results of this correlation are very encouraging. The offset is relatively constant, and the Cell3 numbers are always more conservative than the Spice numbers. This is because Cell3's per-layer capacitance constants are intentionally skewed toward the conservative end of the range.

Contributing Factors

The following factors played a part in producing the results mentioned above and providing a chip that met the specifications.

Library Design. One of the major contributing factors for meeting our density and performance goals is the use of high-quality libraries such as the new HP C26102SH (0.8 μm) standard cell library. This library is a scaled-up version of the HP C14104SH library developed for our CMOS14 (0.5 μm) process.[2]

One major improvement introduced with the HP C26102SH library is that it is tuned for Cell3 while maintaining compatibility with other routers. For a roughly square core, the library allows a very even distribution of horizontal and vertical routing resources. This gives the placer more freedom to find an optimal solution. The use of advanced layout techniques, along with the exploitation of new process design rules, allows smaller cells that provide the same functionality as previous library versions. Finally, advanced router model generation provides fully gridded routing and optimum pin locations, which improves both density and run time. These factors result in a reduction in average wire length for a given fanout. The improvement in wire capacitance as compared to the previous cell library is shown in Fig. 7.

Another area of improvement is in the drive sizes of the cells. The "stairstep design"[3] approach was used to determine appropriate drive increments. The stairstep design approach is a method of sizing gates to provide optimum area versus performance characteristics for each cell in the library. As a result of using this approach, the HP C26102SH library provides more drive increments for Synopsys to map to, avoiding the excessive area penalty imposed when a higher drive cell must be swapped into a path that is just missing timing.

Finally, this is the first library to use the more accurate table-driven Synopsys models. As discussed previously, this allows more accurate delay calculations and eliminates correlation issues with other timing verification tools.

Since the HP C26102SH library is a scaled-up version of the HP C14104SH library, all these advantages will continue into the next process generation.

Design Methodology. Our design methodology was carefully constructed to produce a chip that met aggressive timing goals in a reasonable amount of time. The 90% wire load models improved the chances of first time perfect through the routing cycle, albeit at some loss of performance and die size. This was an intentional choice designed to minimize the design cycle time, while still meeting performance targets.

The HDL coding guidelines enabled the creation of a design that was easily synthesized. They also enabled the use of automatic scan insertion and test vector generation.

© Hewlett-Packard Company 1995

Finally, by imposing a set of automated naming rules during synthesis, tool compatibility issues were minimized. These naming restrictions eliminated name remapping and cross-referencing throughout the design flow.

Point Tools. Several key point tools are necessary to support the design flow discussed in this paper. Synopsys with table-driven delay models is required to provide accurate static timing analysis and valid constraints to the placer. Cell3's timing-driven placement is required to implement the critical path timing output from Synopsys. An automatic test generation (ATG) tool is required to insert and optimize the scan chain automatically and produce appropriate test vectors.

Possible Improvements

Although we have been successful with the methods described in this paper, there are still some areas for improvement.

Timing Verification in Synopsys. Eliminate Verilog functional timing simulation by relying on Synopsys static timing analysis for verification.

More Automation of Clock Insertion. As mentioned, further characterization of Cell3's delay calculation is necessary to eliminate the need for Spice clock verification. Another area of improvement here is to work with Cadence to improve the balanced routing capability of Cell3.

Alternate Ways of Prelayout Capacitance Estimation. Other methods for accurate early capacitance estimation are available that don't require a preliminary route. Promising improvement areas here include using advanced floor planning earlier in the design cycle and netlist-based capacitance estimation that includes not just fanout but other factors such as estimated chip size and types of cells connected to each wire.

In-Place Optimization. Use less conservative wire loads for preroute estimation with expanded use of links-to-layout during the synthesis and placement loop to enable higher performance with minimal impact on cycle time.

References

1. B. Jung and J. McDougal, "An Optimal Scan Chain Auto-Connection Methodology and Scan Signal Insertion Scheme to Reduce Chip Area," *1993 HP Design Technology Conference Proceedings,* pp. 343-347.
2. S. Ratner, J. Eaton, A. Martinez, and H. Youn, "Development of a New Dense and Router Independent BiCMOS Compatible, Standard Cell Library Floorplan for (Bi)CMOS14," *1993 HP Design Technology Conference Proceedings,* pp. 477-484.
3. J. Eaton, "Stairstep Library Design: The Application of Optimization Techniques to the Design of the CMOS14 Standard Cell Library," *1993 HP Design Technology Conference Proceedings,* pp. 391-398.

© Hewlett-Packard Company 1995

A Framework for Insight into the Impact of Interconnect on 0.35-μm VLSI Performance

A design and learning tool called AIM (advanced interconnect modeling) provides VLSI circuit and technology designers with the capability to model, optimize, and scale total delay in the presence of interconnect.

by Prasad Raje

On-chip interconnect is having an increasing impact on the performance of VLSI chips. Previous work in this area from a technological perspective has concentrated mainly on the RC delay of the interconnect.[1] For cases in which the driving gate has been included in the analysis, there has not been an equal emphasis on accurate modeling of the resistance and capacitance of the interconnect and the interconnect's dependence on various dimensions.[2] The problem needs to be examined from the comprehensive perspective of including the gate in the delay analysis, using accurate models for the total delay, and including the dependence of delay on various parameters in the circuit and technological domains.

AIM (advanced interconnect modeling) is an efficient, accurate framework to analyze and optimize a fundamental building block of all VLSI critical paths, namely an arbitrary gate driving an arbitrary on-chip interconnect. AIM is a design and learning tool for both circuit and technology designers concerned about careful modeling, prediction, and scaling of total delay in the presence of interconnect.

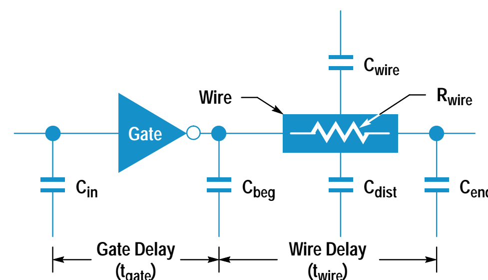


Fig. 1. Basic building block modeled by AIM.

Glossary

The following definitions explain terms as they are used in the context of the accompanying article.

C_{in} (input capacitance). C_{in} is proportional to the product of gate oxide capacitance per unit area, the gate length, and the gate width. The gate width is the total width of all the transistors tied to the input. C_{in} is often represented by the gate width in units of μm.

Circuit Domain. The circuit domain refers to the design realm of the circuit designer. Specifically, certain quantities are under the control of the circuit designer in the context of interconnect delay. These include the wire width, length, and space, or the gate width.

HIVE. HIVE is an internal HP software package that creates closed functions of wire capacitance components as a function of the relevant geometrical quantities. HIVE starts with the wire geometries, performs 2D numerical field simulations and arrives at closest-fit analytical functions.

Interconnect. Interconnect refers to the conducting wires on an integrated circuit chip that connect the components to each other and carry electrical signals.

Technological Domain. The technological domain refers to the design realm of the process technology designer. Certain quantities that affect gate delay in the presence of interconnect are under control of the technology designer. These quantities include wire thickness, interlayer spacing, transistor gate oxide thickness, and so on.

AIM includes circuit and process technology variables while providing a framework to manage a large design space. AIM is also computationally efficient while accounting for important effects like interline capacitance and distributed RCs. It also serves as a bridge between circuit and technology designers to allow for combined optimization of interconnects in both domains. This paper describes the delay model used in AIM, its implementation and verification, and some example analyses.

System Modeled by AIM

All critical paths of CMOS/BiCMOS VLSI chips can be divided into a sequence of basic blocks, each consisting of a switched active device driving a load and an interconnect. Fig. 1 shows a typical representation of a basic building block. The switched device (logic gate) can be represented without loss of generality by a simple inverter with input capacitance C_{in}. The load consists of all nonwire capacitances, typically gate oxide and source/drain junctions. These capacitances are located at three places: at the beginning of the wire (C_{beg}), at the end of the wire (C_{end}), and distributed along the wire (C_{dist}). Note that the distributed capacitance of the wire is distinct from C_{dist} and is discussed in detail below.

The interconnect wire presents a distributed resistance (R_{wire}) and a distributed capacitance (C_{wire}). R_{wire} and C_{wire} are functions of the wire geometry shown in Fig. 2. The

© Hewlett-Packard Company 1995

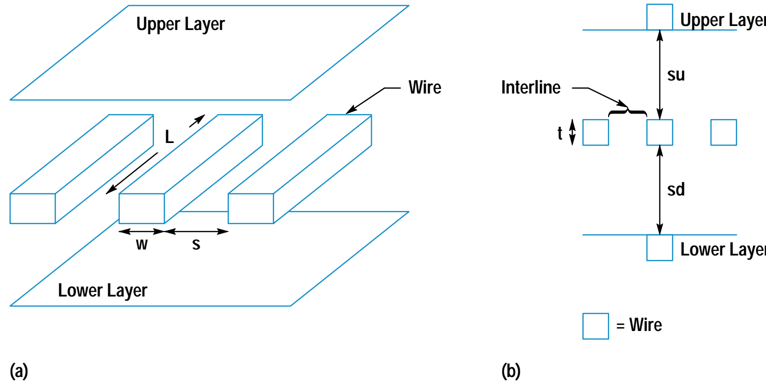


Fig. 2. Wire geometry. (a) Circuit domain variables. (b) Technology domain variables.

dimensions in Fig. 2 show the variables that represent the size (width w, length L, and thickness t) of a typical wire, and the separation of the wire from conductors that are above (su), below (sd), and adjacent (s) to it. L, w, and s are variables in the circuit domain, while t, su, and sd are variables in the technology domain.

A multilayer interconnect system, like the one shown in Fig. 2, provides different values of t, su, and sd depending on the wire layer* (polysilicon, M1, M2, M3, M4, etc.), upper conductor (M1, M2, M3, M4, or none), and lower conductor (substrate, polysilicon, M1, M2, M3, etc.). AIM allows all permutations of layer and upper and lower conductors. A layer variable in the circuit domain is defined that can take on any value selected from these permutations.

The upper and lower conductors, when present, are assumed to be continuous plates of conductors. To a first order, this is an acceptable approximation when the upper and lower layers have densely spaced wires. The adjacent wires are assumed to be equidistant on both sides. For simplicity, the R_{wire} and C_{wire} interconnect components are assumed to be uniformly distributed along the length of the wire. Any changes in layer or wire geometry that change R_{wire} or C_{wire} are important for capacitance extractions in actual layouts. There is no value to introducing this complication into AIM, where larger trends in delay on various parameters are of interest. Also, there would be no generality in choosing a particular change in R_{wire} or C_{wire} along the wire.

AIM Delay Model

A key feature of AIM is that the logic gate delay is included as a full participant in the interconnect delay analysis. Gate delay is defined as the delay from the gate input to the beginning of the wire (see Fig. 1). Wire delay is the delay from the beginning of the wire to the end, and includes the effect of C_{dist}. The total delay is the sum of gate and wire delays:

$$\text{delay} = t_{gate} + t_{wire}. \quad (1)$$

The gate delay is expressed in a delay-versus-fanout format instead of simplifying the gate as an equivalent resistance as described in reference 2. The fanout is the sum of all the capacitance seen by the gate as if it were all lumped at the output, divided by the input capacitance:

$$t_{gate} = t_0 + \text{slope}\,(C_{beg} + C_{wire} + C_{dist} + C_{end})/C_{in} - k \times t_{wire}, \quad (2)$$

where t_0 is the y-intercept of the of the delay-versus-fanout curve, slope is the slope of the delay-versus-fanout curve, and k is an empirical constant that represents a correction factor to account for "hiding" distant capacitance along a resistive wire ($0 < k < 1$ and is typically 0.5).

The wire delay is the usual RC delay including the distributed nature of both the wire and nonwire capacitance:

$$t_{wire} = R_{wire} \times (C_{wire}/2 + C_{dist}/2 + C_{end}) \quad (3)$$

C_{wire} consists of the interline capacitance of the adjacent wires on the same layer, and the interlayer capacitance of the upper and lower layers.

$$C_{wire} = C_{interlayer} + C_{interline} \quad (4)$$

The interlayer and interline components are expressed simply as the respective parallel plate capacitances. This formulation intentionally does not include fringing effects to make it easy to express the optimum width and optimum thickness formulas in the next section. Fringing effects are described in the section "Accurate Delay Modeling" on page 3.

The upper and lower layers are assumed to be quiescent but adjacent wires are allowed to have a signal switching in the opposite sense. A variable m accounts for the Miller effect and effectively doubles the value of the interline capacitance when the adjacent wires switch simultaneously in the opposite direction. If m =1 the adjacent wires are quiescent, and if m = 2 the wires are switched.

$$C_{wire} = \varepsilon wL/sd + \varepsilon wL/su + 2m \times \varepsilon tL/s \quad (5)$$

$$R_{wire} = \rho L/tw \quad (6)$$

where ε is the permittivity of the dielectric and ρ is the resistivity of the metal. The more general case of C_{wire} for upper and lower conductor switching can easily be constructed with extra Miller variables.

Optimum Wire Width and Thickness

The wire width w and thickness t appear in the numerator and denominator of the total delay expression. A larger width or thickness implies an inversely smaller R_{wire} but a larger C_{wire}. The net effect is a reduction of the wire delay (t_{wire}), but an increase in gate delay (t_{gate}). The total delay therefore is optimum at an intermediate value of w or t. The total delay is differentiated with respect to w and t to give the optimum values w_{opt} (optimal wire width) and t_{opt} (optimal wire thickness) at which the delay is a minimum.

* M1, M2, M3, and so on represent different types of metal layers.

© Hewlett-Packard Company 1995

$$w_{opt} = \sqrt{\frac{(1-k)}{2}\frac{r}{\varepsilon}\frac{C_{in}}{slope}\frac{sd \times su}{(sd+su)}\left(\frac{2m\varepsilon L}{s} + \frac{C_{dist} + 2C_{end}}{t}\right)} \quad (7)$$

$$t_{opt} = \sqrt{\frac{(1-k)}{2}\frac{r}{\varepsilon}\frac{C_{in}}{slope}\frac{s}{2m}\left(\frac{\varepsilon L}{sd} + \frac{\varepsilon L}{su} + \frac{C_{dist} + 2C_{end}}{w}\right)} \quad (8)$$

For the circuit designer, w_{opt} is an important quantity and one that can potentially be changed for each different net in a circuit to achieve the lowest delay. Wire widths must be increased:

- when driving longer wires (larger L)
- in the presence of an extra load along or at the end of the wire (greater C_{dist} and C_{end})
- when driving with stronger drivers (larger C_{in} or smaller slope)
- when adjacent wires are switching ($m > 1$).

For the technology designer, optimal wire thickness is the important quantity. The difficulty here is that it is not possible to change the wire thickness for each different net. Therefore, estimations must be made of parameters such as the expected range of wire length, driver size, wire spacing, and nonwire loads in the chips that are expected to be designed in the technology. Once these parameters are known, then one can state that the wire should be designed to be thicker when it is expected to be longer, driven by bigger drivers, or in the presence of a significant nonwire load. Equation 8 provides an analytical basis for the well-known interconnect design guideline that upper layers of metal that go over longer distances should be made thicker.

There is a subtle interaction between the optimum width and the optimum thickness of the wire. The variable w_{opt} depends on the thickness of the wire and vice versa. Thus, a wider wire may lead to a smaller optimum thickness according to equation 8. If the nonwire loads C_{dist} and C_{end} are small compared to both the interline and interlayer capacitances, then w_{opt} has no dependence on wire thickness, and t_{opt} has no dependence on wire width. In this case the technology designer can optimize the wire thickness from a delay standpoint without any consideration for the width of the wire.

Accurate Delay Modeling

The analytical delay model of the previous section provides important insight into the various parameters affecting wire delay. To make specific predictions about interconnect behavior in a technology, it is necessary to use accurate numerical values of the different components of the delay. These components are provided in AIM by HIVE[3] and Spice. The analytical expressions from HIVE are modified so that they can be used with Mathematica. Mathematica is an interactive, interpreted programming environment that allows one to do such things as express analytical equations, perform analysis, and create 2D and 3D plots.

HIVE for Wires and Spice for Gates

HIVE provides for some second-order effects not included in the C_{wire} expression (equation 5). The interlayer capacitances are still linearly proportional to width (w), but the second-order dependence on interline space (s) is included. This is the fringing effect which reduces the interlayer capacitances as interline space is reduced. The interlayer capacitance has the form:

$$C_{interlayer} = F_6(s) + \frac{w - w_{min}}{w_{max} - w_{min}}\left(G_6(s) - F_6(s)\right) \quad (9)$$

where $F_6(s)$ and $G_6(s)$ are sixth-order polynomial functions of s and $w_{min} < w < w_{max}$ is the range over which the fitting function applies.

The dependence of interline capacitance on s is modeled as a sixth-order polynomial rather than the simple linear s term in the denominator. The interline capacitance has the form:

$$C_{interline} = 1/H_6(s) + \frac{w - w_{min}}{w_{max} - w_{min}}\left(1/J_6(s) - 1/H_6(s)\right) \quad (10)$$

where $H_6(s)$ and $J_6(s)$ are sixth-order polynomials.

HIVE uses two-dimensional finite element simulation of actual geometries in an IC technology to obtain the coefficients of the sixth-order polynomials given above. Further, accurate values of these components are available for all values of the layer variable (M4 over substrate or M3 over M2 under M4, etc.).

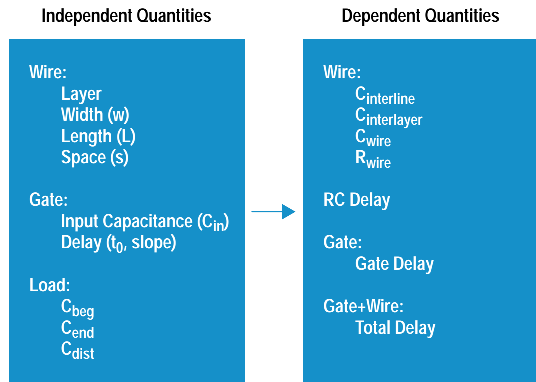
Spice simulations on various basic gates are performed to obtain accurate t_0 and slope values in the technology of interest. For the 0.35-μm CMOS technology (CMOSA) used in this paper, t_0 = 40 ps and slope = 23 ps/fanout. Empirical studies are also carried out to estimate the value of k in equation 2. The value of k lies between 0.4 and 0.6 in CMOSA technology. The dependence of gate delay on input slope could be included with the addition of one or two more fitting parameters. (For simplicity this is not introduced in the first implementation of the AIM model.) Also, the emphasis in the delay analysis is on the trends in delay as a function of various wire parameters. These dependencies are, to a first-order approximation, independent of the input waveform slope at the gate.

The delay predicted by the AIM delay model is compared to a Spice simulated delay of the same gate with the wire represented by a HIVE subcircuit. Delay calculations for 336 data points of various wire widths, lengths, and gate sizes are obtained and they show that the margin of error between the AIM and Spice results is $<3\%$ for 60% of the samples, $<5\%$ for 75% of the samples, and $<10\%$ for 93% of the samples. This provides confidence in the predictions made with the AIM delay model.

Implementation in Mathematica

With the more complicated wire capacitance expressions from HIVE, the delay model is no longer tractable by hand, but it is still in an analytical form that can be coded into Mathematica expressions. The basic delay expressions and subexpressions are in a single file. The technology dependent coefficients in the wire capacitance expressions are in a separate technology file. This allows different interconnect technologies to be analyzed by simply changing the technology file.

© Hewlett-Packard Company 1995



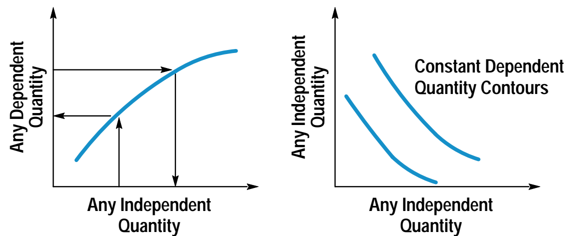


Fig. 3. AIM implementation showing the different types of analyses that are possible.

The analytical implementation in Mathematica allows powerful, fast, and accurate analyses of various dependent observable quantities as functions of various independent quantities (see Fig. 3). Independent quantities can be in numerical or symbolic form and include properties of the wire, gate, or load. These properties are typically specified by values in an input file. However, one or more of these properties may be left in symbolic form. The dependent quantities are expressions or formulas defined in terms of the independent quantities. Each piece of the wire capacitance, such as $C_{interline}$ and $C_{interlayer}$, is available separately. The most important quantity of interest is of course the total delay.

AIM provides standard routines to plot any dependent quantity as a function of any independent quantity. Similarly, given a dependent quantity and all but one independent quantity, the unknown independent quantity can be obtained. More complex analyses consist of plotting a dependent quantity in 3D versus two independent quantities, or plotting a contour plot of a constant dependent quantity with two independent quantities on the x- and y-axes. Examples of these analyses are discussed below.

Mathematica versus a Spreadsheet

The AIM model as it is implemented in Mathematica is highly customizable and many more types of analyses are possible. However, there is a barrier to using this implementation for designers not familiar with Mathematica. The model could be implemented in a spreadsheet, and although the graphical analyses would not be as easy, obtaining quick numerical results would be easier.

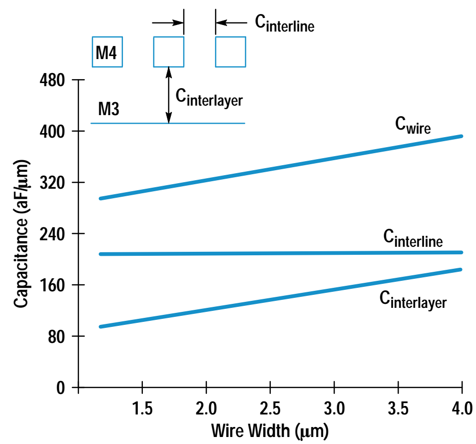


Fig. 4. Wire capacitance components for minimum-spaced M4 over M3.

Insights Provided by AIM

The following examples describe some special insights that are possible with the AIM model.

Interline Capacitance and Fringing

A simple but insightful analysis with AIM is the M4 wire capacitance versus the width of the wire. Fig. 4 shows the interline, interlayer, and total wire capacitance for minimum-spaced (1.6 µm) M4 wires over M3. Adjacent wires are assumed to be switching so that m = 2 (equation 5). The first observation is that the interline capacitance is larger than the interlayer capacitance. The interlayer capacitance increases linearly with width as expressed in equation 5. However, at zero width the extrapolated $C_{interlayer}$ line is not zero because of the fringing component. This behavior is included in the HIVE expressions.

Fig. 5 shows the same M4 wire with the only change being that it is over substrate instead of M3. There is a dramatic difference in the capacitance curves. The reduction in

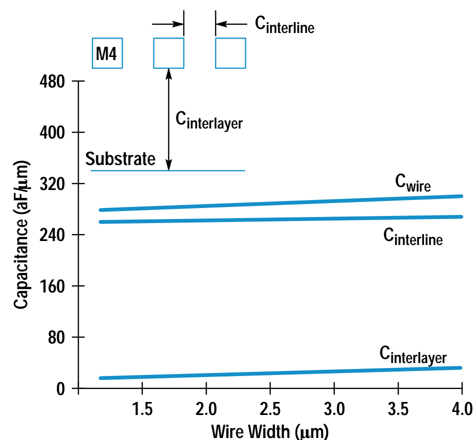


Fig. 5. Wire capacitance components for minimum-spaced M4 over substrate.

© Hewlett-Packard Company 1995

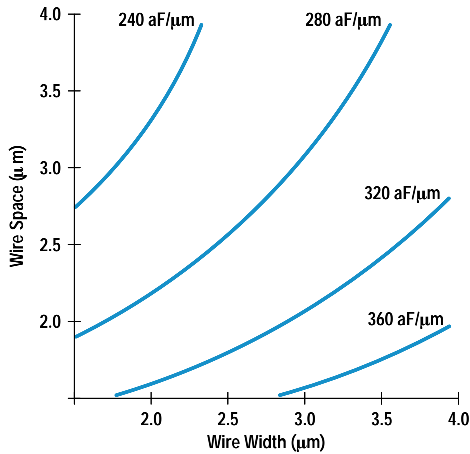


Fig. 6. Constant-capacitance contours for M4 over M3.

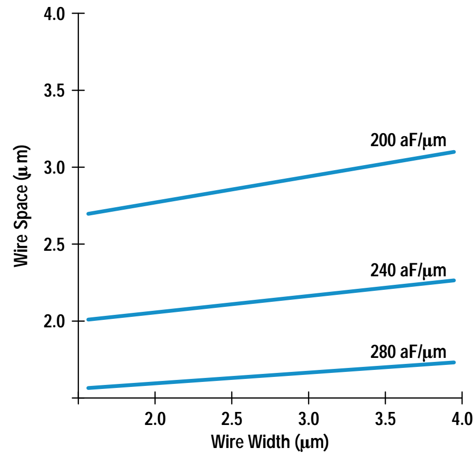


Fig. 7. Constant-capacitance contours for M4 over substrate.

$C_{interlayer}$ might be expected from the larger distance of the wire to the substrate, but the significant feature is that the total capacitance is not significantly reduced despite the larger distance from the substrate. This is because $C_{interline}$ has increased. This increase is because more lines of flux from the lower surface of the wire terminate on the adjacent wire instead of the lower conductor. In other words, the fringing component has increased. Another result of fringing is that the interlayer capacitance now has a very weak dependence on wire width.

Thus, the conventional wisdom that upper layers of metal enjoy much reduced capacitance because of their distance from the substrate does not hold. For one thing, the upper layers may run over wires in the immediate lower layer and even when they do not, the total capacitance is not much lower.

Capacitance versus Width and Space

A visual representation of the relative importance of width and space is obtained from a contour plot of wire capacitance in a 2D space of wire width and interline spacing. Fig. 6 shows constant-capacitance contours for M4 over M3 lines. The data in Fig. 6 is a superset of the information in Fig. 4. Along any horizontal line in Fig. 6 several contours are cut, indicating a rapid increase in C_{wire} (interline and interlayer capacitance) with width. The dependence on space is also significant. Fig. 7 shows the contours for M4 wire over substrate. The contours appear more horizontal indicating that there is a weak dependence on wire width and that a reduction in wire capacitance is easier to achieve with an increase in (interline) spacing. The contours have reduced in value but the reduction is substantial only when the wire spacing is large and the width is large. Such contours can be made for all the metal levels to provide a quick ready reference of wire capacitance for a range of geometries.

Optimizing the Width of Wires

RC delay is an important factor that causes a circuit designer to choose wider wires when they are long. However, it is important to realize that larger width comes with an increase in the total capacitance of the wire and therefore a possible increase in the total delay. There is an optimum width of the wire at which the total delay is a minimum. Equation 7 expresses the dependence of the optimum width on various parameters. Fig. 8 shows an example plot of the gate and wire delay components as a function of wire width. The optimum width is only 1.5 µm, which is relatively small for a 5000-µm long wire. There is a built-in function in AIM that provides the optimum width when the remaining variables in the system are specified.

Minimum Metal Width Design Rules

The optimum wire width illustrated in the previous section depends on a number of parameters, the most important being length L, gate width C_{in}, wire spacing s, and wire layer. AIM can rapidly generate the optimum width for a large range of these parameters. Fig. 9 shows the optimum width versus length for a few example cases with a 200-fF fixed load at the end of the wire. The curves are not smooth because of the limited number of data points generated and the slow variation of delay with length. The approximately square root dependence on length predicted in equation 7 is illustrated in Fig. 9. If the gate width is increased to 200 µm, the curve moves to a higher w_{opt} (optimum wire width) as predicted in equation 7. A larger interline spacing reduces w_{opt} as illustrated by curve C, but this dependence is not strong. Curve D shows the optimum wire width for an M1 wire and is surprisingly close to A for the same conditions. This is because the interlayer spacings are similar when M1

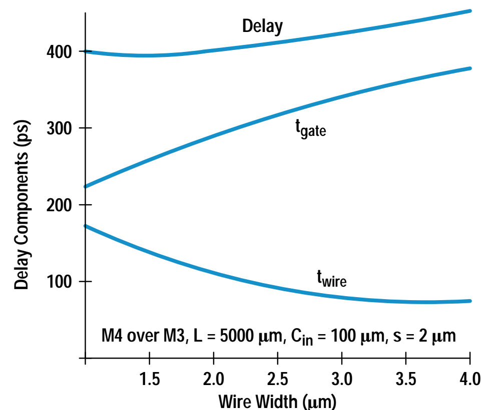
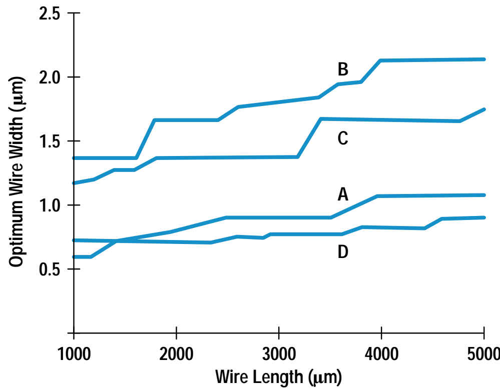


Fig. 8. Total delay versus width. RC delay is reduced but wire gate delay is increased with larger wire width.

© Hewlett-Packard Company 1995



Legend

Curve	Layer	Spacing	Gate Width
A	M4 over M3	1.6 µm	50 µm
B	M4 over M3	1.6 µm	200 µm
C	M4 over M3	10 µm	200 µm
D	M1 over Substrate	1.6 µm	50 µm

Fig. 9. Optimum wire width versus wire length under different layer, spacing, and gate width conditions.

and M4 are both surrounded by upper and lower conductors. Also, the dependence of w_{opt} on wire thickness is very weak.

If the minimum width design rule for M4 is 1.3 µm, then Fig. 9 shows that the optimum width can be smaller than the design rule width for many reasonable conditions. A full range of high and low values for the wire lengths, spaces, and inverter sizes can be simulated to determine the range of optimum widths of the wire. This can then provide guidelines for technology designers in setting minimum design rules for wires.

Delay versus Wire Length and Driver Size

An important analysis that encapsulates a lot of useful information for a circuit designer in a single figure is a plot of delay contours with gate width along the x axis and wire length along the y axis. Fig. 10 shows such a plot for M4 over M3 with 1.5-µm wire width, 2-µm spacing, and a 100-fF

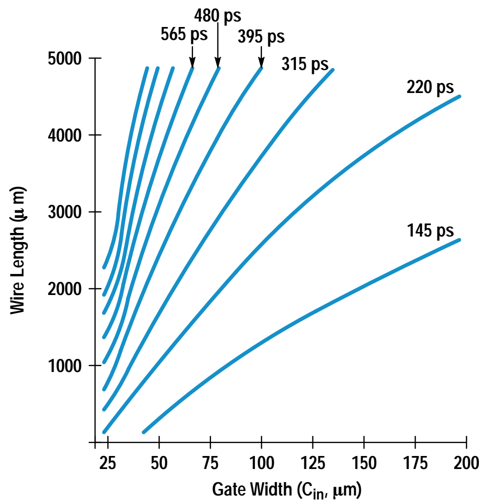


Fig. 10. Constant delay contours (ps) in a space of gate width (µm) and wire length (µm).

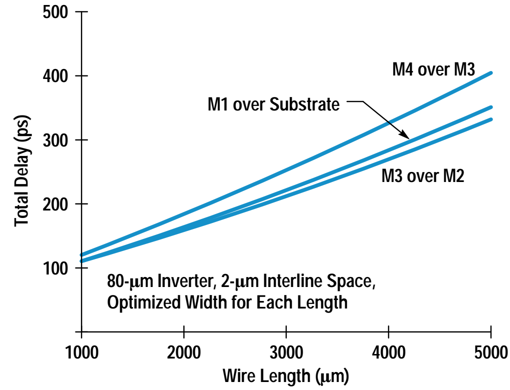


Fig. 11. Total delay versus wire length for various levels.

load at the end. This allows a ready reference for quickly looking up the gate width that would be needed to drive a certain wire length with a desired delay. Alternatively, the dependence of delay on wire length for a given gate width can be seen. For example, a 3000-µm long wire would require an 80-µm gate width to achieve less than 300-ps total delay. A 50-µm gate width would see a very rapid increase in its delay beyond a 2000-µm length as seen by the bunching of the contours. Similar plots for other wires or other conditions can easily be generated using built-in functions provided in AIM.

Delay versus Wire Layer

A common misconception is that an upper level of metal is always faster when driving long distances on the order of a few thousand micrometers. To analyze this, the built-in routines in AIM are used to plot total delay (a dependent quantity) versus wire length (an independent quantity). Fig. 11 shows this plot for M4, M3, and M1 wires, all with minimum interlayer spaces and a 2-µm interline space. The wire width for each level is optimized for each length as discussed earlier. A large 80-µm inverter size is chosen to emphasize the RC delay over the gate delay. The total delay is larger for the M4 wire than the M1 wire! Even though the wire delay of the M4 wire is lower, its higher capacitance leads to a larger gate delay. Also, the M4 wire suffers from higher interline capacitance than the M3 wire because of the passivating nitride over the M4 wire. While the M3 wire has the lowest delay, the M1 delay is remarkably close. The optimum width for a 5000-µm long M4 wire is 1.4 µm and that for an M1 wire with the same length is 1.2 µm. The fact that the M4 wire is slower than the M1 wire is not an artifact of AIM, but has been confirmed by Spice simulation with HIVE subcircuits.

Pitfalls in Algorithmic Shrinking

Many VLSI chips are shrunk from one generation of technology to the next by algorithmically scaling all the layers in the design to match the design rules of the new technology. The result on the wires in the circuit domain is a reduction in widths, spaces, and lengths. There may also be scaling of wire dimensions (thickness, interlayer spacings) in the technology domain. The result on the FETs is a reduction in gate width and length and also source/drain areas. It is relatively easy to predict the performance scaling of the delay for logic circuits that have a relatively small amount of interconnect. It is much harder to predict delay scaling in critical paths that have a large amount of interconnect. This is because the

© Hewlett-Packard Company 1995

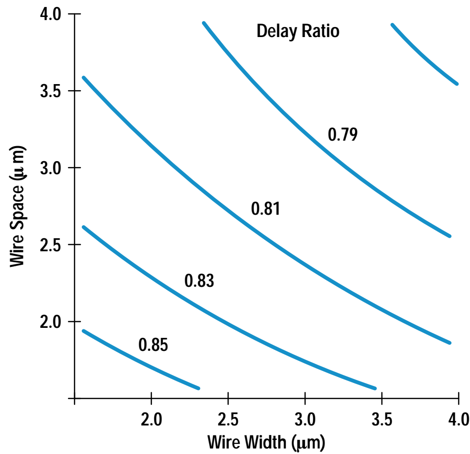


Fig. 12. Delay ratio CMOSA/CMOSB with 0.75× wire and FET shrink. Unscaled gate width = 100 µm, wire length = 5000 µm.

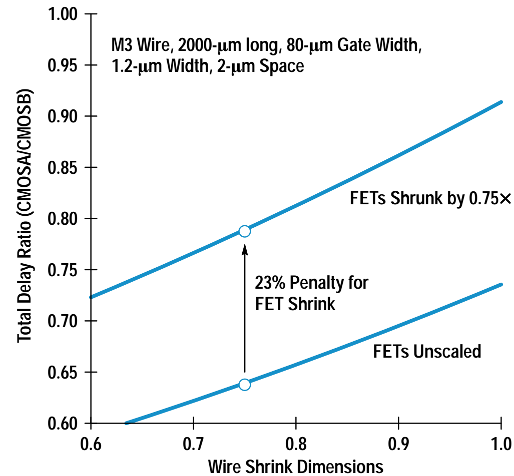


Fig. 14. Total delay scaling in going from CMOSB to CMOSA.

scaling factor depends on the interplay of a large number of parameters. AIM allows a rapid exploration of the design space and can pinpoint scenarios in which the delay improvement could be compromised.

To illustrate this capability, the scaling of delay from a 0.5-µm CMOS technology (CMOSB) to the 0.35-µm CMOSA technology is observed for a range of values of different variables. The gate width in each technology is characterized by the t_0 and slope values (see equation 2). The values for a CMOSA inverter are t_0 = 40 ps and slope = 23 ps/fanout. The values for a CMOSB inverter are 40% higher. This accounts for the change in the FETs in the technology domain. The change in the interconnects in the technology domain is accounted for by a new set of HIVE coefficients for capacitances, which get translated into a new AIM technology file. In the circuit domain, a shrink factor of 0.75 is applied to all wire dimensions (width, spacing, and length) and to the gate width and nonwire loads.

The resultant scaling of wire capacitance, RC delay, and so on is taken care of in AIM and only the circuit-domain scaling parameters are supplied as inputs. The data in Figs. 12 and 13 shows the delay ratio (CMOSA/CMOSB) as a function of wire space, wire width, wire length, and gate width. Fig. 12 shows that delay ratios are better for large wire space and wire width. Large wire space is often not available because of pitch limitations, while large wire widths can increase total delay. Large wire lengths are detrimental to delay scaling as are small FETs. AIM can generate these concise reference charts showing the dependence of delay scaling on various important parameters.

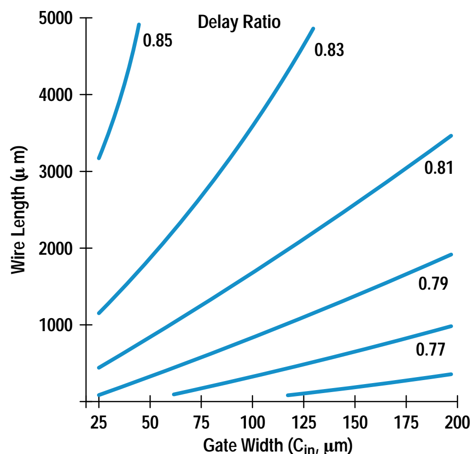


Fig. 13. Delay ratio in going from CMOSB to CMOSA with 0.75× wire and FET shrink, space = 2.4 µm, width = 2 µm.

The improvement in the delay is not significant for the wire dominated basic blocks considered in this paper. However, it is incorrect to assume that it is only the wire RC delay that is the cause of the problem. The reason is that when the wire dimensions scale down, the FET widths also scale down, reducing the drive to the wires. Further, even though the wire length reduction is beneficial to capacitance, wire spacing reduction increases interline capacitance. Also, fringing effects undermine the linear capacitance reduction expected from simplistic scaling of wire width. The resistance of the scaled wire is constant if the thickness stays the same. The net result is that both the gate delay and the wire delay do not scale well.

AIM allows one to examine the delay ratio of a typical basic block with independently varied scaling factors for FET and wire scaling. Fig. 14 shows the ratio of CMOSA to CMOSB delay as a function of wire shrink dimensions. If the FETs are kept unscaled and only the wires are shrunk, the delay ratio is 0.63. This substantial improvement would also be obtained for basic blocks that do not have significant wire loading. However, it is incorrect to expect this number when a whole chip is shrunk and the critical path consists of many wire-dominated basic blocks. Fig. 14 illustrates this scenario when the FETs are shrunk by 0.75×. The delay ratio is now only 0.78, a 23% increase over the previous case. This illustrates the importance of the capacitive load of the wire. AIM can be used to examine each net of a chip design to flag those nets that are susceptible to poor delay scaling if they are shrunk. These nets could either be redesigned or special cases made to keep the selected FET widths unscaled in a shrink. This can lead to guidelines for "design for shrinkability." In the meantime, the statement can be made that the ultimate technologically capable delay improvement is not possible in a pure shrink strategy.

© Hewlett-Packard Company 1995

Summary

AIM has been presented as a comprehensive framework to understand and optimize the performance of basic blocks in VLSI critical paths. The interconnect is modeled with highly accurate expressions that account for many second-order effects, and the gate driving the interconnect has been included as a full participant in the analyses. The design space is large because of the many variables in both the technology and circuit domains. This has been managed with a simple but accurate analytical delay model. The implementation in Mathematica provides quick and efficient analyses of many different types of technology and circuit variables.

The examples shown have illustrated only some of the capabilities of AIM. The myth of much lower capacitance for upper levels of metal has been shown to be unfounded. A visual insight into the relative influence of wire width and spacing on wire capacitance has been provided. The importance of the optimization of wire width has been demonstrated and its dependence on various parameters has been correlated with simple analytical equations. It has been shown that metal widths are often made larger than necessary and some minimum width rules may preclude optimal delay. A reference chart for circuit designers showing delay versus wire length and gate size has been demonstrated. It has been shown that upper levels of metal are not necessarily the best choice even for long wires. Algorithmic shrinking of chips from one technology to the next has been shown to suffer a substantial penalty in wire dominated basic blocks. The gate capacitive delay scales as poorly as does the wire RC delay.

Acknowledgments

The author would like to thank Gene Emerson for his encouragement, support, and guidance. Thanks also to K.J. Chang and Soo Young Oh for developing HIVE.

References

1. K.C. Saraswat and F. Mohammadi, "Effect of Scaling of Interconnects on the Time Delay of VLSI Circuits," *IEEE Transactions on Electron Devices,* Vol. ED-29, no. 4, April 1982, p. 645.
2. T. Sakurai, "Approximation of Wiring Delay in MOSFET LSI," *IEEE Journal of Solid-State Circuits,* Vol. SC-18, no. 4, August 1983, p. 418.
3. K.J. Chang, et al, "Parametrized Spice Subcircuits for Multilevel Interconnect Modeling and Simulation," *IEEE Transactions on Circuits and Systems,* Vol. 39, no. 11, November 1992.

© Hewlett-Packard Company 1995

Synthesis of 100% Delay Fault Testable Combinational Circuits by Cube Partitioning

High-performance systems require rigorous testing for path delay faults. A synthesis algorithm is proposed that produces a 100% path delay fault testable function with a minimal set of test pins.

by William K. Lam

To ensure that manufactured circuits meet specifications, the circuits must be subjected to static and dynamic testing. Static testing considers the steady-state behavior of a circuit (e.g., whether the output of a combinational circuit computes the required Boolean function). Dynamic testing examines the transient behavior of a circuit. In this paper, we focus on a specific kind of dynamic testing: delay testing, which is testing to determine how long it takes a circuit to settle to its steady state.

If we define a path in a circuit to be a sequence of gates from an input to an output of the circuit, then input signals propagate to outputs along paths in the circuit. Thus, the time for a circuit to settle to its steady state, called the delay of the circuit, is determined by the delays of the paths in the circuit. Hence, testing the delay of a circuit translates to testing the paths in the circuit. A common scheme for testing delays is shown in Fig. 1.

To test whether path a-b-c-f has a delay less than or greater than t seconds, a pulse is applied to input a and to input d where it is delayed t seconds to latch the output f. If the delay of path a-b-c-f is greater than t, we say the path has a delay fault. If the steady-state value of f is 0, then latching a 0 implies the delay of the path is less than t, provided the waveform at f has only one transition. If there is more than one transition at f, latching a 0 does not necessarily imply that f has settled to its steady-state value. The delay of the path cannot be inferred from latching the steady-state value if multiple transitions can occur at f. Therefore, a path's delay can be tested if there is an input vector such that a single transition at the input of the path propagates along the path and causes only one transition at the output, independent of the delays of the gates in the circuit. The significance of independence is illustrated in the following example.

In the circuit in Fig. 2, to test the delay of path a-c-f we set input b to 1 causing d to be 0. With this input value, a single transition at input a will propagate along path a-c-f and cause a single transition at f, independent of gate delays in the circuit. Therefore, path a-c-f can be tested for its delay. Similarly, the delay of path b-d-f can be tested by setting a to 0. However, for path b-c-f, a single transition at input b might cause a multiple transition at f, depending on the relative delays of the AND gate and the inverter. For instance, a rising transition at b produces a negative pulse (a falling transition followed by a rising transition) at f if the delay of the AND gate is longer than that of the inverter. On the other hand if the input pulse does not propagate to the output, f maintains a steady 1. Because we don't have prior knowledge about the relative delays of the AND gate and the inverter, we conclude that path b-d-f is not delay testable.

A path is called *robustly path delay fault testable* (RPDFT) if a single transition at the input of the path propagates along the path and produces a single transition at the output, independent of the gate delays in the circuit. Only RPDFT paths can be tested reliably for a delay fault. A necessary and sufficient condition for a path to be RPDFT is that there is an input vector such that during the course of a transition propagating along the path, for each gate on the path, all the side inputs of the gate take on noncontrolling values. A controlling value is an input to a gate that determines the gate's output regardless of the values at other inputs. For example, the controlling value for an AND gate is 0 and for an OR gate

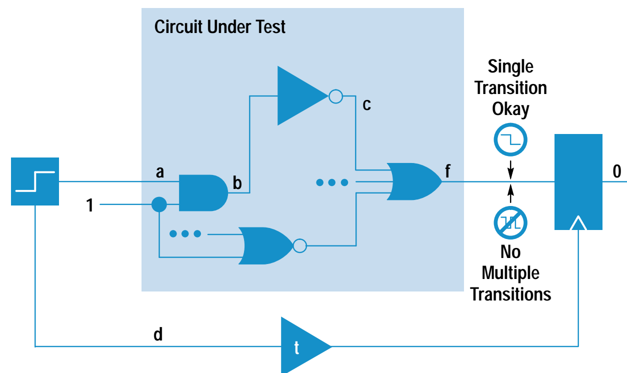


Fig. 1. Delay testing scheme.

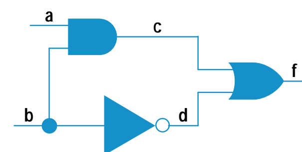


Fig. 2. Example circuit.

© Hewlett-Packard Company 1995

the value is 1. Since the delay of a circuit is determined by the delays of the paths in the circuit, to test the delay of the circuit, all paths in the circuit should be RPDFT. A circuit is said to be 100% RPDFT if all of its paths are RPDFT. Unfortunately, most practical circuits have very few RPDFT paths. This implies that most practical circuits cannot be fully and robustly tested for delay faults, even though many circuits are tested despite the presence of hazards.

In this paper, we propose an algorithm that always synthesizes 100% RPDFT circuits. First, we consider synthesis of 100% RPDFT two-level circuits from any given function. Then, we show how multilevel circuits can be derived from two-level circuits while preserving their delay fault testability.

Previous Work

Devadas and Keutzer[1] derived a necessary and sufficient condition for a path to be RPDFT and proposed an algorithm to synthesize a circuit to achieve a high percentage of RPDFT paths. However, their algorithm cannot always produce circuits with 100% RPDFT. It is known that there exist functions that do not have 100% RPDFT implementations. A natural question is: can any function be augmented so as to have a 100% RPDFT implementation? One way of augmenting a function is to add extra inputs. With this technique, Pomeranz and Reddy[2] demonstrated that many circuits can be made to be 100% RPDFT. However, it is not known whether any arbitrary function can be synthesized to be 100% RPDFT by using this technique or any other.

Synthesis of 100% RPDFT Two-Level Circuits

An example of a two-level circuit is an AND-OR implementation configuration (e.g., a programmable logic array) corresponding to a sum-of-products representation of a Boolean function. Any Boolean function can be represented as a sum of products. For example, $f = (a + b)(\overline{a} + \overline{b}) + bc$ has the sum-of-products representation $a\overline{b} + \overline{a}b + bc$, whose corresponding two-level implementation consists of three AND gates and one OR gate. Each AND gate implements a product term and the OR gate combines the outputs of the AND gates as shown in Fig. 3. The circuit in Fig. 3 is called a two-level implementation because the first level consists of AND gates and the second level an OR gate.

The path P in Fig. 3 from b through the AND gate for the term bc is not RPDFT because no matter what value input a is set to, a rising or falling transition at b through the path will produce multiple transitions at f or not propagate along P, depending on the relative delays of the AND gates. For instance, if input a is set to 1 and the delay of the left AND gate is shorter than the delay in the right AND gate, then a falling transition at b along P will be blocked from propagating to f (Fig. 4a). Thus, the delay of P cannot be reflected at

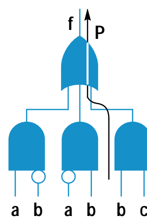


Fig. 3. A two-level circuit.

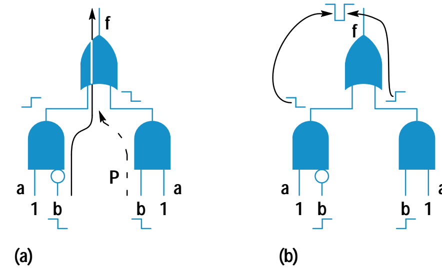


Fig. 4. Propagation of transitions. (a) Because the delay in the left AND gate is shorter than the delay in the right AND gate, a falling transition at b is blocked from propagating to f. (b) A rising transition at b causes a negative pulse at f.

f. Under the same setting, a rising transition at b will cause a negative pulse at f (Fig. 4b). If input a is set to 0, then a rising transition at b will be blocked from propagating to f because the output of the AND gate is forced to 0. Thus, path P is not RPDFT.

For more complicated functions, it would be difficult to perform the above analysis to determine whether paths are RPDFT. To make the task of identfying RPDFT paths easier, an algebraic method[3] is presented.

Definitions. Before stating the algebraic method, some terms need to be introduced.

- The *cofactor* of function f with respect to variable x (for positive phase), denoted by f_x, is derived from f by replacing the variable x in f with 1. Similarly for negative phase, $f_{\overline{x}}$ is derived by replacing x with 0.
- The *smooth operator* S on function f with respect to variable x, denoted by $S_x(f)$, is $f_x + f_{\overline{x}}$.
- Let C be a product term or *cube** in f. Then f − C is the function derived from f by eliminating C.

Cofactor f_x is the evaluation of f at x = 1. Smoothing f with respect to x gives the function independent of x.

Theorem 1: Let f be a function in a sum-of-products form of a two-level circuit, and π a path starting from primary input x and going through the AND gate of cube C. Then, path π is RPDFT if and only if there is an input vector ν = (…,x,…) such that:

$$S_x(C)\ \overline{S_x(f - C)}\ (\nu) = 1.$$

The vectors ν and ν′ = (…,$\overline{x}$,…) are a test vector pair for the delay fault on π.

For an arbitrary function in a sum-of-products form, $S_x(C)\ \overline{S_x(f - C)}\ (\nu)$ may be 0 for all vectors. This would mean that the path through C starting at input x is not RPDFT. To augment a given function so that it has a 100% RPDFT implementation, we add extra inputs called *test pins,* which equal 1 under normal operations and may be selected to be 0 in delay testing mode.

To construct a circuit with 100% RPDFT paths the set of cubes in a given function is partitioned into subsets such that each subset forms a 100% RPDFT function. Next, a pin is attached to each subset. To test a path in the subset, only the test pin of the subset is set to 1, while all the remaining test pins are set to 0. Since the subset is 100% RPDFT, the paths

* A cube is a product term. For example, $\overline{a}bc$ and $b\overline{c}$ are cubes, but a + c is not a cube.

© Hewlett-Packard Company 1995

are RPDFT under this setting of the test pins. Symbolically, let $f = \sum_i C_i$, where C_i is a cube which can be partitioned into subsets of cubes, S_i, such that each path in each S_i is RPDFT. The new augmented function is now $f = \sum_j T_j S_j$, where T_j is the test pin for cube subset S_j.

To test path π going through a cube in S_j, the test pins must be set such that $T_j = 1$ and $T_i = 0$ for $i \neq j$. So f becomes S_j, which is 100% RPDFT by construction, enabling π to be tested for a delay fault. In normal operation, all test pins are set to 1 allowing the augmented function $f = \sum_j T_j S_j$ to restore the original function $f = \sum_j S_j = \sum_i C_i$.

A natural question is: Can an arbitrary function be partitioned into such subsets? The answer is yes, because a partition in which S_i is a cube is such a partition. Further, the paths through the test pins do not need to be tested for delay faults because these pins are held constant during normal operation. Therefore, for any arbitrary function, a 100% RPDFT implementation is always possible with this cube-partitioning scheme. This fact is formally stated in the following theorem.

Theorem 2: Any Boolean function has a prime and irredundant two-level AND-OR implementation with 100% RPDFT and the possibility of adding new inputs. Further, if C is a two-level AND-OR implementation of f, then C can always be resynthesized to be 100% RPDFT.

To resynthesize a two-level circuit to be 100% RPDFT, the worst case is when a test pin is needed for each cube in the circuit. In this worst case, the additional area required is at most twice the original area, assuming each test pin is ANDed with the cube. This procedure allows designers to synthesize two-level circuits without considering delay fault testability because test pins can be added later to achieve the desired testability.

Because a test pin is provided for each subset, a minimum partition is desired. Of course, the designer does not have to make all paths RPDFT because test pins can be added only to the cubes in which the paths need to be tested. In this case, the number of test pins to add is bounded by the number of cubes that involve the paths to be tested. Nevertheless, we want an algorithm that produces a minimal partition.

The following algorithm produces a minimal cube partition by partitioning a set of cubes $f = \{ C_i \}$ into $\{S_i\}$ so that the sum of cubes in each S_i is a function with 100% RPDFT paths.

```
i=0;
while(f not empty) {
  i++;
  Si = {φ}
  for each cube C ∈ f {
    if(Si ∪ C is 100% RPDFT) {
      Si = Si ∪ C;
      remove C from f;
    }
  }/* end for loop */
} /* end while loop */
```

The test pins do not need to be connected directly to the outside world through pins on the package. A shift register, which can be an existing scan chain, can be used to shift in the test patterns. The extra pins needed are at most two, one for the shift register input and the other for its clock. A possible implementation is shown in Fig. 5. In the figure $T_1 \ldots T_n$ are the test pins whose values are set by the output of the decoder, which decodes the test patterns from the shift register. Each subset S_i needs one test pin.

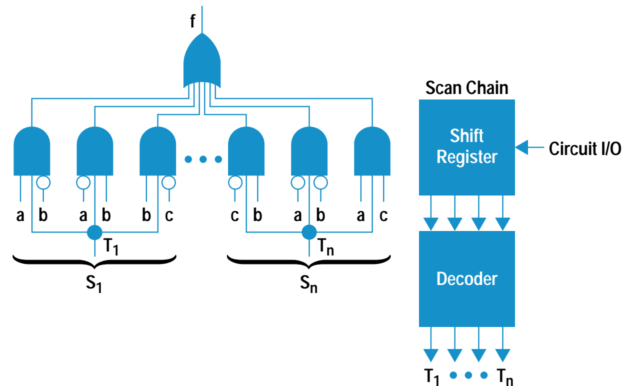


Fig. 5. 100% robustly path delay fault testable (RPDFT) implementation.

Multilevel Synthesis

A two-level implementation is a special case of a multilevel implementation and usually requires much more silicon area. This is because a multilevel implementation does more sharing of gates. For example, the multilevel circuit in Fig. 6a, which uses four two-input gates, would require eight two-input-equivalent gates if the same function were implemented using a two-level structure (Fig. 6b).

The multilevel implementation can be represented as $f = (a + b)(c + d) + e$, while the two-level representation is $f = ab + ad + bc + bd + e$. The multilevel implementation is simply a factored form of the two-level implementation. Thus, a two-level implementation can be transformed into an area-saving multilevel implementation by factoring out common terms. The question that comes up after these transformations is whether testability is preserved. That is, will a RPDFT path in the original two-level implementation remain RPDFT in the factored multilevel implementation and will a path newly created by these transformations be RPDFT? In the Boolean domain, factorizations like $ab = a(\bar{a} + ab)$ and $(a + b) = (a+ b)(\bar{a} + a)$ are valid. Factorizations involving the use of Boolean rules such as $a + \bar{a} = 1$, $a \cdot \bar{a} = 0$, and $a \cdot a = a$

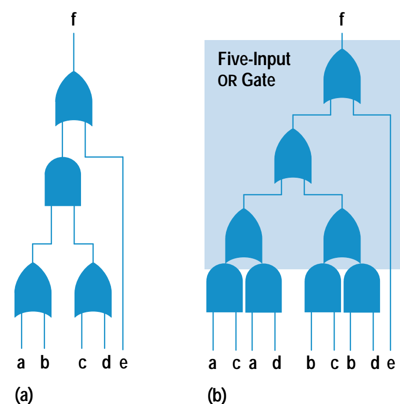


Fig. 6. (a) A multilevel circuit. (b) A two-level equivalent of the multilevel circuit in (a).

© Hewlett-Packard Company 1995

are called *Boolean factorizations,* and factorizations that don't use such rules are called *algebraic factorizations.* Hachtel, Jacoby, Keutzer, and Morrison[4] proved that a multilevel implementation derived from a two-level implementation using only algebraic factorizations preserves RPDFT of the paths in the original two-level implementation. This concept is summarized in the following theorem.

Theorem 3: If C, a two-level multiple-output circuit, is 100% RPDFT, and A is a multilevel circuit derived from C through the application of algebraic operations, then A is also 100% RPDFT.

Therefore, synthesis for a multilevel circuit with 100% RPDFT can be done in two steps. First, a two-level circuit with 100% RPDFT paths is synthesized using the cube partitioning method. Then, a multilevel circuit is derived from this two-level circuit by applying algebraic factorizations.

Selective Critical Path Testing

Because making all paths in a chip delay fault testable may not be area-efficient, only some representative paths are selected to be made delay fault testable. Theoretically, testing only a fraction of paths may not guarantee freedom from faults for the entire chip. However, because of the nature of delay tracking in IC processing, proper selective schemes can offer high confidence in testing.

Also, because gate delays within a chip track well, a long path is more likely to fail a timing specification than a short path, making longer paths good candidates to be selected for testing. If a selected path is not RPDFT, it can be made so by using one of the synthesis techniques discussed above. Specifically, to make a selected path RPDFT, find a maximal set of RPDFT cubes that contain the path and introduce a new test pin to isolate the set of cubes from the rest of the cubes using the minimal cube partitioning algorithm. This step is repeated until all selected paths are RPDFT. Finally, the number of test pins can be minimized by first repartitioning the cubes in the cube sets so that each cube belongs to only one new cube set, and then using one test pin for each new cube set.

Experimental Results

The cube partitioning algorithm was implemented on the Berkeley SIS (sequential interactive synthesis) platform and runs on an HP 9000 Model 735 workstation, which has about 150M bytes of RAM. Two benchmarks from the International Symposium on Circuits and Systems (ISCAS) and the Microelectronic Center of North Carolina were used on the algorithm. Table I shows the results of running these benchmarks through the cube partitioning algorithm.

The second, third, fourth, and fifth columns in Table I contain the total number of I/O pins, gates, paths, and non-RPDFT paths, respectively, in each circuit. The sixth column defines the original fault testability, that is, the fraction of paths that are RPDFT in a particular circuit. After these circuits were resynthesized, they became 100% testable (i.e., final testability = 1.0). The eighth column reflects the total number of test pins inserted to make the circuit fully delay fault testable. With the exception of circuit b12, after all the circuits were resynthesized they were made fully delay fault testable with six or fewer test pins. The ninth column is the area overhead, which is the ratio of the area increase over the original circuit area. Since any additional area adds some delay, the delay overhead for each circuit results from a layer of two-input AND gates for each test pin insertion. Finally, the last column is CPU execution time for each circuit. These times vary directly with the number of cubes in the circuit.

Circuits with very few non-RPDFT paths and circuits that did not finish within the 12-hour preset time limit are not listed in Table I.

Table I
Two-Level Synthesis of 100% RPDFT Circuits

Circuit	I/O Pins	Gates	Paths	Non-RPDFT Paths	Initial Testability*	Test Pins	Area Overhead (%)	CPU (Seconds)
table5	32	188	7259	495	0.93	4	3.25	141
table3	28	203	7381	687	0.90	4	3.38	282
rd84	24	263	3280	1456	0.55	2	0.48	821
apex1	90	296	9109	1515	0.83	6	6.16	25834
b12	24	449	1922	1845	0.04	23	5.53	423
ex1010	20	830	14710	2096	0.85	4	0.89	17434
z5xp1	17	148	4032	2558	0.36	4	5.26	336
z9sym	10	422	3780	3276	0.13	2	0.14	21694
ex4	156	676	4404	3632	0.17	3	1.14	175
alu4	22	1044	7875	3955	0.49	4	0.53	134
apex4	28	476	14958	4354	0.70	5	3.11	12368
misex3	28	1876	17971	5258	0.70	6	5.84	15936

* Final testability = 1.0.

© Hewlett-Packard Company 1995

Conclusion

In this paper, we studied the problem of synthesizing circuits with 100% RPDFT. We proved that for an arbitrary function, there exists a 100% RPDFT implementation, and we proposed a synthesis algorithm that always produces a 100% RPDFT implementation for any function and a minimal set of test pins. Further, we showed that a circuit synthesized using the proposed algorithm uses at most twice as much area as any two-level implementation of the circuit. For most practical circuits, the additional areas are small. Finally, we demonstrated how area-efficient multilevel circuits with 100% RPDFT can be constructed by applying algebraic factorizations to the synthesis algorithm.

Acknowledgments

The author would like to thank Barbara Fredrick for her enthusiastic support, Cheryl Harkness for a lesson on FrameMaker, Mark Heap for the many insightful discussions on the algorithms presented in this paper, and Robert Aitken and Peter Maxwell for reviewing the paper.

References

1. S. Devadas and K. Keutzer, "Synthesis of delay-fault-testable circuits: Theory," *IEEE Transactions on Computer-Aided Design,* Vol. 11, no. 1, January, 1992, pp. 87-101.

2. I. Pomeranz and S. Reddy, "Achieving Complete Delay Fault Testability by Extra Inputs," *International Test Conference '91,* Oct. 1991, pp. 273-282.

3. W. Lam and R. Brayton, *Timed Boolean Functions—A Unified Formalism for Exact Timing Analysis,* Kluwer Academic Publishers, 1994.

4. G. D. Hachtel, R. M. Jacoby, K. Keutzer, and C. R. Morrison, "On the Relationship Between Area Optimization and Multifault Testability of Multilevel Logic," *IEEE/ACM International Conference on Computer-Aided Design '89,* November 1989, pp. 422-425.

Bibliography

1. W. Lam, A. Saldanha, R. Brayton, and A. Sangiovanni-Vincentelli, "Delay Fault Coverage, Test Set Size, and Performance Tradeoffs," *IEEE/ACM Design Automation Conference '93,* June 1993, pp. 446-452.

© Hewlett-Packard Company 1995

Better Models or Better Algorithms? Techniques to Improve Fault Diagnosis

The simple stuck-at fault model paired with a complex fault diagnosis algorithm is compared against the bridging fault model paired with a simple fault diagnosis algorithm to determine which approach produces the best fault diagnosis in CMOS VLSI circuits.

by Robert C. Aitken and Peter C. Maxwell

Failure analysis is an important task for continuous improvement of both the quality of shipped ICs and the underlying fabrication process. Fault diagnosis (also called location) can aid the failure analysis process by producing a list of candidate faults given a set of observed tester failures. These faults are then mapped to potential defect sites, allowing a failure analysis engineer to target a specific and manageable portion of the chip.

Improvements to fault diagnosis have tended to be either improvements in fault modeling[1,2,3,4] or improvements in diagnostic heuristics and algorithms.[5,6,7,8] In this paper we attempt to analyze the relative contributions of models and algorithms by comparing the diagnostic ability of the simple stuck-at fault model paired with a complex location algorithm to the complex and realistic bridging fault model paired with a simple location algorithm. These models and algorithms are compared both on known bridging defects from actual chips, and since the available sample of known bridging faults is small, on a larger sample of simulated bridging faults.

Bridging and Stuck-At Faults

The most common approach for modeling IC defects is the stuck-at fault model.[1] This model states that defective lines in a circuit will be permanently shorted to either the power supply (stuck-at 1) or ground (stuck-at 0). The model has been popular for test generation and fault simulation because it is simple to use and because a complete stuck-at test set thoroughly exercises the device under test (it requires that both logic values be observed on all lines in a circuit).

With the advent of CMOS integrated circuit technology, the connection between the stuck-at fault model and actual defects has become somewhat tenuous. This is less important from a test generation perspective, since tests for stuck-at faults tend to be excellent tests for other types of defects as well.[2] For diagnosis, however, an accurate fault model might be more important. In the accompanying article, we consider bridging,[3] which extends the stuck-at model by allowing a defective line to be shorted to any other line in the circuit, not just the power and ground lines. Unlike the simpler stuck-at model, there are numerous variations of the bridging fault model, depending on which bridges are considered (all possible versus layout-based), and how they are presumed to behave (wired AND, wired OR, dominant signal, analog, etc.). Our model[4] considers possible bridges extracted from layout and models their behavior according to the relative signal strengths of the driving transistors.

References

1. R.D. Eldred, "Test Routines Based on Symbolic Logical Statements," *Journal of the ACM,* Vol. 6, 1959, pp. 33-36.
2. T.W. Williams and K.P. Parker, "Design for Testability—A Survey," *Proceedings of the IEEE,* Vol. 71, January 1983, pp. 98-112.
3. K.C.Y. Mei, "Bridging and Stuck-at Faults," *IEEE Transactions on Computers,* Vol. C-23, July 1974, pp. 720-727.
4. P.C. Maxwell and R.C. Aitken, "Biased Voting: A Method for Simulating CMOS Bridging Faults in the Presence of Variable Gate Logic Thresholds," *Proceedings of the International Test Conference,* 1993, pp. 63-72.

Only voltage testing is considered in this analysis. In addition, we employ a single fault model in all cases, both for simplicity and because in many of the cases of interest for diagnosis, single-site defects are more likely. A part that failed its functional package test, for example, probably contains only a single defect, or it would not have passed its numerous tests at the wafer level and its parametric tests at the package level.

We consider only full-scan circuits in this paper, since full-scan circuitry continues to be more amenable to diagnosis than nonscan or partial-scan circuitry. The main reason for this is that full scan does not require the fault models to predict future states accurately, since scan circuitry reloads the state after every test vector. This reduces the number of potential detections, and more significantly, reduces the dependency between potential detections, which can greatly complicate fault diagnosis. We find that about 10% of returned parts have failing scan chains and cannot be diagnosed, but for the remainder, the improvement in diagnostic accuracy is worth the costs of full-scan circuitry.

Fault Diagnosis Methodology

The fault diagnosis methodology is part of the overall failure analysis process. Not all failing chips are selected for failure analysis. Some typical candidates include chips that pass their component test but fail at board test, chips that fail component test in a similar fashion, and field returns. Fault diagnosis takes *failing test vectors** as input and returns a set of potential defect sites. Since these sites must be physically examined by a failure analysis engineer, it is important that there not be too many of them. Fig. 1 shows the tools we use and the files created during fault diagnosis.

* A given chip test consists of a set of hundreds or thousands of individual test vectors. On a given bad chip some of these vectors will produce outputs that fail the test, and it is these vectors that constitute the failing test vectors.

© Hewlett-Packard Company 1995

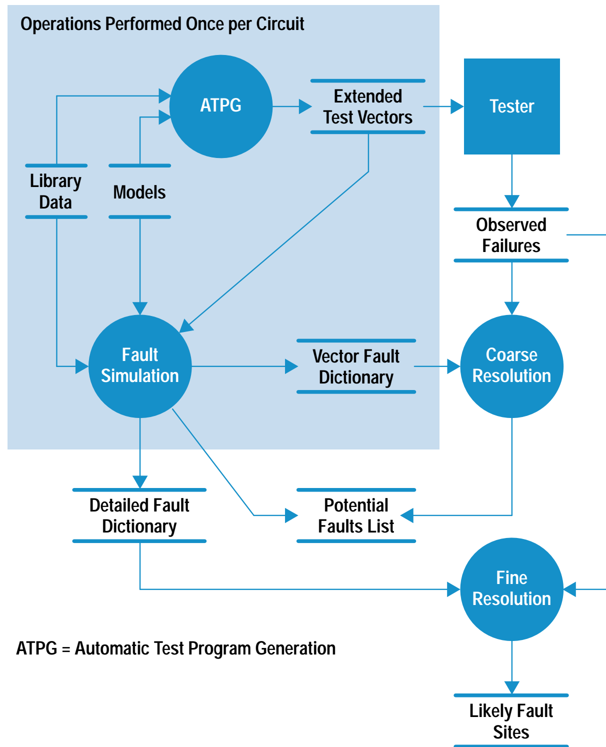


Fig. 1. Fault diagnosis tools and files.

Test Generation for Diagnosis. Scan vector sets for production test are usually optimized to achieve maximum single stuck-at fault coverage with a minimum number of vectors. This allows test costs to be reduced without compromising quality. A side effect of this process is that many faults tend to fail on the same set of vectors, making it difficult to distinguish between faults when using a vector fault dictionary,[9] which contains only failing test vectors. We improve the diagnostic resolution of production test vector sets by attempting to generate additional tests to distinguish between stuck-at faults which have identical failing behavior in the production set. This method seeks to maximize the diagnostic capability measures described in reference 10. The vectors produced from this effort are called an extended vector set.

The second shortcoming of production test sets is their reliance on the single stuck-at fault model. Stuck-at test sets can also be extended by vectors to target other faults such as bridging faults and transistor stuck-on/off faults. For the test sets used in our model, such faults were not explicitly targeted.

Fault Location Software. Once an extended vector set is available, it can be used in the diagnostic process. To avoid the large amounts of data associated with detailed fault dictionaries, we generate a vector fault dictionary, or fault coverage matrix (see Fig. 2). As with test vector generation, dictionary generation is performed only once for a given chip design.

The remaining steps are run on individual failing chips. Failing chips are run on the tester and their *observed failures* (test vectors and outputs where failures are observed) are logged. For chips with numerous failures, only the first few hundred failures are typically recorded. Diagnosis is a two-step process. Coarse diagnostic resolution is obtained by using the vector fault dictionary and the vectors from the observed failures to reduce the potential fault list from all faults to a manageable number.

The coarse resolution process eliminates the vast majority of faults from consideration. A more extensive fault simulation is then performed (using fast fault simulation techniques[11]) on the remaining faults to construct a detailed fault dictionary, showing not only which vectors are expected to fail, but also the scan elements and output pads where errors are expected to be seen. The output of this simulation is then compared with the tester data, and faults that match are sent on to failure analysis. This two-step process has been successful at reducing failure data and diagnosis time and predicting defect sites.

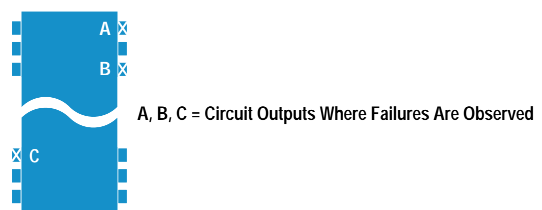
The success of the diagnostic software depends on the accuracy of the fault models and the ability of the algorithms to deal with unmodeled defects. The next section discusses the models and algorithms we selected for evaluation.

Fault Models and Algorithms

Improvements in the diagnosis process can be obtained by improving the fault models and the heuristics of the algorithms. Although previous work[5,6,7,8,12] has concentrated on the heuristics of algorithms, some results are available for diagnosis using more complex fault models,[3] often using I_{ddq}.[*,2,13] Waicukauski et al[12] reported that their diagnosis method would work on bridging faults. Pancholy et al[14] developed a simple test chip to examine the behavior of stuck-at and transition faults on silicon. Millman et al[15] examined the relationship between simulated bridging faults and stuck-at fault dictionaries, using an early version of the "voting model"[8] for bridging simulation. Our work builds on these results by examining the behavior of actual bridging defects on silicon and performing simulation using biased voting,[4] which is an extension of the voting model that takes logic gate thresholds into account.

We examine the relative effectiveness of two approaches to diagnosis: the simple stuck-at fault model with a complex

* I_{ddq} is the quiescent drain current.



Contents of Vector Fault Dictionary

Type of Fault	Failing Test Vectors
F1	1, 3, 5
F2	1, 2
F3	1, 3, 5

Contents of Detailed Fault Dictionary

Type of Fault	Failing Vectors + Circuit Outputs
F1	1-A 3-A, B 5-C
F2	1-A, B 2-C
F3	1-A, B, C 3-A 5-A, B

Fig. 2. The contents of a vector fault dictionary versus a detailed fault dictionary, which contains more data.

© Hewlett-Packard Company 1995

diagnostic algorithm and the more complex but realistic bridging fault model with a simple diagnostic algorithm. Using the stuck-at model and the simple algorithm together typically provides very limited success (although this is dependent on the manufacturing process). Clearly, using the complex bridging fault model with the complex algorithm would likely to be the most successful, but would give little insight into whether the algorithm or the model is the greatest contributor.

Stuck-at Fault Model. The stuck-at fault model continues to be the most commonly used model for test generation, fault simulation, and fault diagnosis. The model is simple and simulation using it is fast. This is important for dictionary-based diagnosis, especially if dynamic dictionary construction is used. Because of the model's ubiquitous nature, off-the-shelf CAD tools can often be used for much of the diagnostic process.

Bridging Fault Model. We use the biased voting model,[4] which is able to predict accurately the electrical behavior of a wide variety of bridging defects in standard cell CMOS circuits by considering the drive strength and logic thresholds of the circuit elements in the neighborhood of the bridge. Our implementation of the method runs approximately two to three times slower per fault than stuck-at simulation, but is still considerably faster than Spice, which it attempts to emulate. Use of a realistic bridging fault model requires the extraction of likely fault locations from the layout, which in turn requires an inductive fault analysis tool.[3] It is important that potential bridges be extracted between adjacent layers, as well as within layers.

Simple Diagnosis Method. The simple diagnosis method finds the best match between observed failure data and the predicted failure data. A failing output predicted by simulating a given fault for a given test vector matches the observed behavior when a failure that was observed on the tester appears at the same output for the same test vector. A fault is removed from consideration if it predicts a failure point at a vector (or vector/output pair for a detailed fault dictionary) where no failure was observed. This is equivalent to a somewhat relaxed fault list intersection algorithm.[8] In general, the best matches are chosen, and the actual number of matches depends on experience and the fault model being used. For our experiments, faults were only selected if they were able to predict all failures, which is the strictest selection rule.

Extended Diagnosis Method. The extended diagnosis method is similar to the bit-partial intersection method,[8] which extends the simple diagnosis method by not excluding faults whose simulation predicts failures that were not observed on the tester. Instead, the incorrect predictions for these faults are considered along with their correct predictions (i.e., simulated failures which were also observed on the tester). A final likelihood for each fault is determined by a weighted combination of the two measures, which is the extent to which the fault's predictions match the observed data. In general, a correct prediction is given a substantially higher weight than an incorrect prediction. For example, fault F3 in Fig. 3 (which has three correct predictions and one incorrect) has a higher likelihood than F2 (which has two correct predictions and zero incorrect).

It is possible to use an even more complex algorithm for diagnosis, such as the effect-cause method.[5] However, destructive scan (flip-flop outputs toggle during scan) on our example circuit precluded examining transition behavior, and no implementation of the algorithm was available for our experiments. Finally, it is easy to construct cases in which such algorithms can be misled by realistic bridging behavior, so the results are likely to be similar to those we obtained.

Experimental Results

Our experimental vehicle is a small, full-scan ASIC (nine thousand gates) implemented in a 1-μm process. Two experiments were conducted. In the first, parts with known bridging defects were diagnosed using the two approaches described above, and the results were compared with the known cause. In the second experiment, simulated bridging defects were diagnosed using the extended diagnosis method the stuck-at fault model.

Known Bridging Defects. The sample chips for this experiment came from two categories. Three chips had metal-to-metal bridging faults inserted with a focused ion beam (FIB). Two others were parts that failed at board test, and for which subsequent failure analysis revealed bridging defects as the root cause. The experiments were conducted so that the person running the diagnostic tools did not know the defect locations.

Both diagnosis methods described above are able to rank faults based on their ability to predict observed failures. For the example in Fig. 3, the simple diagnosis method would rank the faults F2, F1, with F3 being excluded (F2 and F1 have no wrong predictions and F2 has more correct predictions than F1), while the extended diagnosis method would rank them F3, F2, F1 (order is based on the number of correct predictions).

Since a failure analysis engineer usually does not have time to investigate a large number of defect sites, we declare a diagnosis to be misleading when at least nine simulated faults are assigned by the particular diagnosis method to have a higher likelihood than the actual fault of predicting observed failures because their simulated (or predicted) behavior closely matches observed behavior. We wanted to compare the two diagnosis methods by examining the number of misleading diagnoses.

Observed Failures* = A, B, C

Type of Fault	Predicted Failures**	Matches to Observed Failures	
F1	A	A	Correct
		0	Wrong
F2	A, B	A, B	Correct
		0	Wrong
F3	A, B, C, D	A, B, C	Correct
		D	Wrong

*Failures Produced by Applying Failing Test Vector Set 1 to a Failing Chip on the Tester (see Fig. 2)

**Failures Produced by Applying Failing Test Vector Set 1 to Fault Models

Fig. 3. The relationship between failing test vectors, observed failures, and the predicted failures produced by applying the failing vectors to a particular fault model.

© Hewlett-Packard Company 1995

The results of using these diagnostic methods on bridging faults are summarized in Table I. The entries under the fault model columns are the number of cells that have been predicted to have potential defects after performing fine resolution using the model in question. The extended diagnosis method which was used with the stuck-at fault model was able to identify the correct cell as the most likely defect for chips 2 and 3. For chip 4, the algorithm identified twelve other locations as being likely fault locations before it found the correct fault on the thirteenth try. For chip 1, on the other hand, faults on a total of 15 cells (including the actual defective one) predicted the observed failures correctly. However the stuck-at model on chip 1 also predicted failures on other outputs where no such failures were observed, which shows a danger in relying on a subset of failing outputs to create a dictionary of likely fault sites (see Fig. 1). Many of the faults on these 15 cells were equivalent,* so no stuck-at diagnosis could distinguish between them. Both chip 4 and chip 1 are thus misleading diagnoses because in both cases greater than nine faults match the observed failures better than faults at the actual defect site.

Table I
Results for Known Bridging Defects

Chip	Actual Defect	Type*	Fault Model	
			Bridging	Stuck-at
1	FIB bridge	Nonfeedback	1	15
2	FIB bridge	Feedback	1	1
3	FIB bridge	Feedback	1	1
4	Bridge	Feedback	1	13
5	Bridge	Feedback	2	5
6	FIB open	—	1	1
7	Open	—	2	2

FIB = Focused ion beam.

* See Fig. 4.

The bridging model was successful in each case using the simple diagnosis method. An unobserved failure was never predicted and all observed failures were predicted in each case. For chip 5, two bridges matched the observed behavior. The second possibility was also in the immediate vicinity of the defect.

We also analyzed two chips with known open failures (the final two chips in Table I). These failures can also be diagnosed to the correct location by both methods, showing that the bridging method did not produce misleading results in these cases. The actual rate of misleading diagnoses when the bridging method is used with nonbridging defects has not yet been determined primarily because of lack of data.

* Equivalent means that the same faulty behavior is caused by two different faults of the same kind.

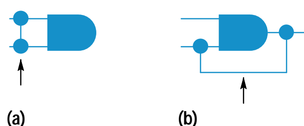


Fig. 4. (a) A nonfeedback fault. (b) A feedback fault.

Fault equivalence is much less common with realistic bridging faults than with stuck-at faults. Equivalent faults simplify test generation but complicate diagnosis because the nodes involved may be widely dispersed on the actual chip. The relative occurrence of equivalent faults is shown in Table II for all faults modeled in the circuit. The unique row shows the number of faults that behaved differently from all other faults in the test set. The vast majority of bridging faults behaved this way, but only 28% of stuck-at faults did, even though the test set was designed to maximize this behavior. The misleading row shows the number of faults with inherently misleading behavior, in that they were identical to at least nine other faults. Almost one sixth of the stuck-at faults belonged to this category, compared with less than 2% of bridging faults. This implies a substantial inherent misleading diagnosis rate even for stuck-at defects when the stuck-at model is used. The row labeled other represents faults that had between two and nine fault equivalencies. In total, the 25345 bridging faults exhibited 21540 failing behaviors, for an average of 1.18 faults per behavior, while the corresponding stuck-at figure was 2.04 for 21010 actual faults.

Table II
Equivalent Fault Behavior

Modeled Behavior	Bridging		Stuck-at	
	Number	Percentage*	Number	Percentage*
Unique	19404	76.6	5967	28.4
Misleading	412	1.6	3261	15.5
Other	1724	6.8	1086	5.1
Total	21540	85.0	10314	49.1

* As a percentage of actual faults (bridging faults = 25345 and stuck-at faults = 21010).

Simulated Bridging Defects. The data in the previous section shows that misleading diagnoses of bridging defects can occur with the stuck-at model, but are insufficient to allow a rate to be calculated. Rate is a measure of the probability of a misleading diagnosis. To get a better idea of the rate, the following simulation experiment was performed.

We selected 200 bridging faults at random from the set extracted for the circuit. Of these, 78 were feedback faults and 122 were nonfeedback. These faults were then simulated and a detailed fault dictionary obtained for each. The faults could not be used as observed failure files directly because they contained potential detection information.** Failure files were generated by assuming that 0%, 10%, 50%, 90%, and 100% randomly selected potential detections would result in errors. Our experience suggests that in practice this number is around 10% for our simulation method.

Coarse diagnosis (vectors only) was then performed on these files. Four stuck-at faults were considered for correctly describing the fault (stuck at 0 or 1 on either of the bridged nodes). One of these faults, which predicted the most observed failures, was then selected as the best match. All other faults that predicted at least as many failures as the best match were noted. Of these, the faults with no more than the number of failing predictions of the best match were included as candidates. This is meant to represent a

** In a simulation, a potential detection is a situation in which it is hard to prove that the fault was detected (see "Potential Detection" on page 115).

© Hewlett-Packard Company 1995

Table III
Results after Coarse Resolution

Category	Potential Detection (%)	Faults Remaining (%) 1-4	5-9	10-19	20-49	50+	Misleading Diagnoses (%)
Nonfeedback	0	46.7	15.6	5.7	11.5	20.5	37.7
	10	45.1	13.1	6.6	10.7	24.6	41.8
	50	41.8	13.1	8.2	8.2	28.7	45.1
	90	43.4	13.9	4.9	11.5	26.2	42.6
	100	42.6	13.9	5.7	11.5	26.2	43.4
Feedback	0	41.0	11.5	10.3	11.5	25.6	47.4
	10	35.9	14.1	15.4	5.1	29.5	50.0
	50	35.4	12.5	14.6	8.3	29.2	52.1
	90	33.3	14.1	14.1	7.7	30.8	52.6
	100	35.9	12.8	14.1	6.4	30.8	51.3
Bridge (self)	10	85.9	6.4	5.1	0.0	2.6	7.7

kind of optimal selection process, which is able to stop when it reaches the correct defect. In reality, the selection process would likely include other faults to guarantee that the correct fault is selected.

The coarse resolution results are summarized in Table III. As an example of interpreting the entries in the table consider the case in which 10% of predicted potential detections were considered to result in hard detections. For nonfeedback faults, 45% of the faults were resolved to fewer than five sites after coarse resolution, and the rate of misleading diagnoses was almost 42%. The average CPU time required for coarse resolution on an HP 9000 Model 735 workstation was 24 seconds for each feedback fault and 22 seconds for each nonfeedback fault. It seems evident that feedback faults result in a higher rate of misleading diagnoses than nonfeedback faults. This is not unexpected, since the behavior of feedback faults is more complex and their connection to stuck-at behavior is more limited.

As a reference point, one experiment (the last entry in Table III) was performed using the bridging model to attempt to diagnose the predicted bridging faults. In this case, the data for the feedback faults (which are more difficult to diagnose than nonfeedback faults) with a 10% rate of potential detections was diagnosed using a detailed bridging fault dictionary. The misleading diagnostic rate after coarse resolution was 7.7%. Most of these were cases where very few failures were observed. The average CPU time for coarse resolution was 38 seconds for this experiment.

Coarse resolution is only the first part of the diagnostic process, but the number of faults remaining after it occurs determines the time required to process the remaining faults. Fine resolution was performed on the faults remaining after coarse resolution. The results are shown in Table IV, which is identical in structure to Table III. The average CPU time required for fine resolution was 146 seconds for each feedback defect and 147 seconds for each nonfeedback fault.

It is interesting to note that in some cases the rate of misleading diagnoses actually increased after fine resolution. This may be because of a particular situation that occurs for some bridging faults. Sometimes vectors that fail closely match one of the stuck-at faults at one node of the bridge, while the actual failing signals propagate from another fault site. A typical example is shown in Fig. 5, in which the buffer is able to overdrive the NAND gate, so failing vectors can occur whenever the two nodes have opposing values. In this example, the buffer drives zero most of the time, and a stuck-at one failure on the buffer output matches the failing vectors extremely well. Since the buffer is always dominant, failures never propagate from that site and fine resolution would disregard that defect. In these cases, the coarse resolution process was modified to pick at least one fault from each site. This typically increased the number of faults remaining after coarse resolution by a factor of three or four. Note that the number of such defects is nontrivial, since global signals and buses tend to be driven with large buffers, and these signals pass or cross many others.

The faults that are difficult to diagnose using the extended algorithm and stuck-at fault model tend to share one of two characteristics: large fanout and/or similar drive strengths. In the first case, the so-called "Byzantine general"[16] behavior causes problems, where faults propagate along some fanout branches but not others. With reconvergence, this can cause failures where none would happen with a stuck-at fault on the stem and the reverse. In the second case, when neither gate in the bridge is dominant, the fault effects are dispersed and less closely tied to either of the bridged nodes.

The final row of Table IV again refers to using the bridging model to diagnose itself. As expected, misleading diagnoses are virtually nonexistent at this point, although this result is itself somewhat misleading, since the modeled defects are the same as those being diagnosed. Fine resolution took an average of 90 seconds per defect for this experiment. There were typically many fewer faults to resolve than with the

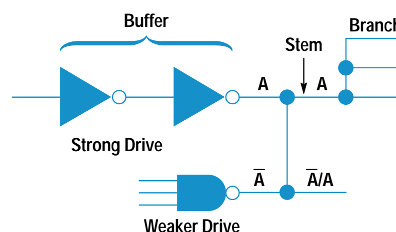


Fig. 5. Dominant bridging fault.

© Hewlett-Packard Company 1995

Table IV
Results after Fine Resolution

Category	Potential Detection (%)	Faults Remaining (%) 1-4	5-9	10-19	20-49	50+	Misleading Diagnoses (%)
Nonfeedback	0	43.4	14.8	8.2	13.1	20.5	41.8
	10	45.1	13.9	5.7	10.7	24.6	41.0
	50	42.6	14.8	11.5	12.3	18.9	42.6
	90	37.7	11.5	15.6	13.9	21.3	50.8
	100	36.9	14.8	13.1	12.3	23.0	48.4
Feedback	0	32.1	19.2	11.5	15.4	21.8	48.7
	10	29.5	16.7	14.1	24.4	15.4	53.8
	50	29.5	19.2	14.1	17.9	19.2	51.3
	90	35.9	19.2	11.5	15.4	17.9	44.9
	100	29.4	16.7	14.1	14.1	25.6	47.4
Bridge (self)	10	92.3	6.4	1.3	0.0	0.0	1.3

stuck-at model, which more than compensated for longer simulation time per fault.

Conclusion

We examined the occurrence of misleading diagnoses for an extended diagnosis algorithm with a stuck-at fault model and a simple algorithm with a realistic bridging fault model. A diagnosis was declared to be misleading when at least nine simulated faults were assigned a higher likelihood than the actual fault of predicting observed failures because their predicted behavior closely matched observed behavior.

In the results from the actual chips and the simulation experiments, it appears that the extended diagnosis algorithm, when used with the stuck-at fault model, results in a rate of about 40% of misleading diagnoses for realistic bridging faults. This is much higher than the rate obtained when using a simple diagnosis algorithm with a bridging fault model. This suggests that fault model improvement may be more beneficial than algorithm improvement in producing diagnostic success. Similar rates are seen for known bridging defects and for those simulated using the biased voting model,[4] although in the former case the sample size is too small for any conclusions. This work is continuing as a joint project between HP's Design Technology Center and HP's Integrated Circuit Business Division's quality group.

A second and somewhat counterintuitive result has also emerged from this work. In at least some cases, additional information can impede diagnosis. It was observed that vector dictionaries are sometimes better at diagnosing bridging faults than detailed fault dictionaries, particularly when the faults are of the one-gate-dominates variety. Additional work is underway to better characterize this behavior.

Potential Detection

A logic simulation will produce one of three values for a driven circuit output: 0, 1, and X, where X is unknown (the simulator cannot predict the value). The situation in which a circuit produces a known value, say 0, in a fault-free simulation but an unknown value, X, in the simulation of a fault, is called a potential detection. It is called potential detection because it will be detected if the unknown value is 1 on an actual faulty chip, but it will not be detected if the unknown value is 0. The following table summarizes these detectability conditions.

Good Value	Faulty Value 0	1	X
0	N	D	P
1	D	N	P
X	N	N	N

D = Detected
N = Not detected
P = Potentially detected

Acknowledgments

The authors gratefully acknowledge assistance from Jeff Schoper, Bob Shreeve, and others in the collection of chip data for this work.

References

1. J.M. Acken, *Deriving Accurate Fault Models,* Technical Report CSL-TR-88-365, Stanford University, Computer Systems Laboratory, October 1988.
2. R.C. Aitken, "A Comparison of Defect Models for Fault Location with I_{ddq} Measurements," *Proceedings of the International Test Conference,* September 1992, pp. 778-787.
3. A. Jee and F.J. Ferguson, "Carafe: A Software Tool for Failure Analysis," *Proceedings of the International Symposium for Testing and Failure Analysis,* November 1993, pp. 143-149.
4. P.C. Maxwell and R.C. Aitken "Biased Voting: A Method for Simulating CMOS Bridging Faults in the Presence of Variable Gate Logic Thresholds," *Proceedings of the International Test Conference,* October 1993, pp. 63-72.
5. J. Rajski and H. Cox, "A Method of Test Generation and Fault Diagnosis in Very Large Combinational Circuits," *Proceedings of the International Test Conference,* September 1987, pp. 932-943.
6. M. Abramovici, M.A. Breuer, and A.D. Friedman, *Digital Systems Testing and Testable Design,* W.H. Freeman and Co., New York, 1990.
7. P.G. Ryan, K. Davis, and S. Rawat, "A Case Study of Two-Stage Fault Location," *Proceedings of the International Reliability Physics Symposium,* March 1992, pp. 332-337.
8. P. Kunda, "Fault Location in Full-Scan Designs," *Proceedings of the International Symposium for Testing and Failure Analysis,* November 1993, pp. 121-127.
9. G. Ryan, W.K. Fuchs, and I. Pomeranz, "Fault Dictionary Compression and Equivalence Class Computation for Sequential Circuits," *Proceedings of the International Conference on Computer-Aided Design,* November 1993, pp. 508-511.

© Hewlett-Packard Company 1995

10. E.M. Rudnick, W.K. Fuchs, and J.H. Patel, "Diagnostic Fault Simulation of Sequential Circuits," *Proceedings of the International Test Conference,* October 1992, pp. 178-186.

11. J.A. Waicukauski, E.B. Eichelberger, D.O. Forlenza, E. Lindbloom, and T. McCarthy, "A Statistical Calculation of Fault Detection Probabilities by Fast Fault Simulation," *Proceedings of the International Test Conference,* November 1985, pp. 779-784.

12. J.A. Waicukauski and E. Lindbloom, "Failure Diagnosis of Structured VLSI," *IEEE Design and Test,* Vol. 6, no. 4, August 1989, pp. 49-60.

13. S. Chakravarty and M. Liu, "Algorithms for Current Monitor-Based Diagnosis of Bridging and Leakage Faults," *DAC-92 Proceedings,* June 1992, pp. 353-356.

14. A. Pancholy, J. Rajski and L. McNaughton, "Empirical Failure Analysis and Validation of Fault Models in CMOS VLSI," *Proceedings of the International Test Conference,* September 1990.

15. S. Millman, E.J. McCluskey and J. Acken, "Diagnosing CMOS Bridging Faults with Stuck-At Fault Dictionaries," *Proceedings of the International Test Conference,* September 1990, pp. 860-870.

16. L. Lamport, R. Shostak, and M. Pease, *The Byzantine Generals Problem,* Technical Report 54, Computer Science Laboratory, SRI International, March 1980.

© Hewlett-Packard Company 1995