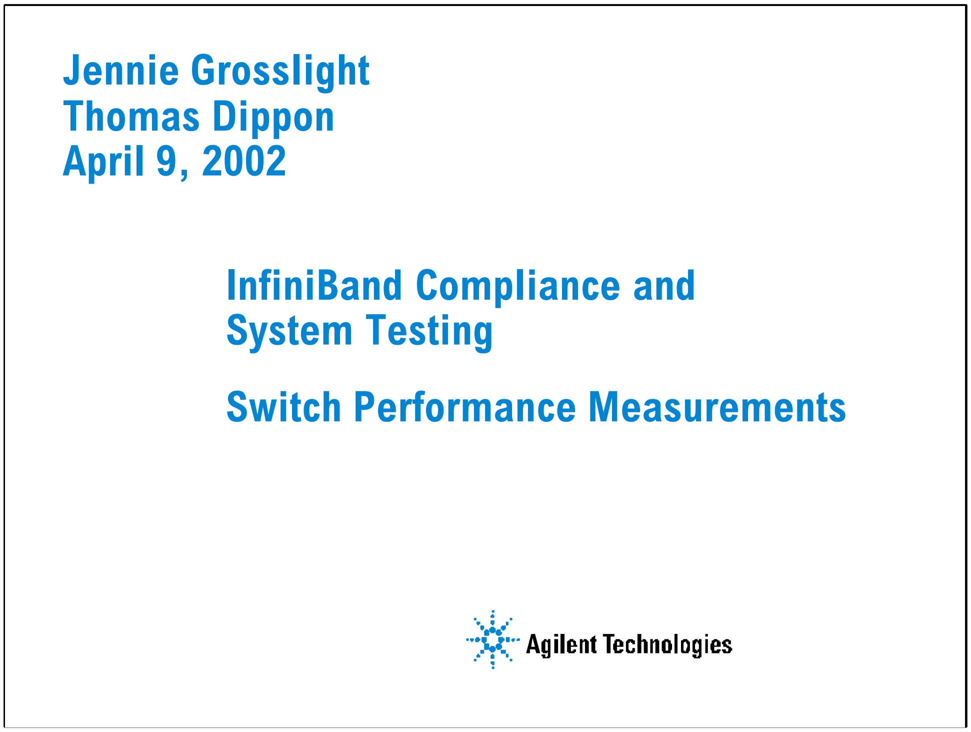


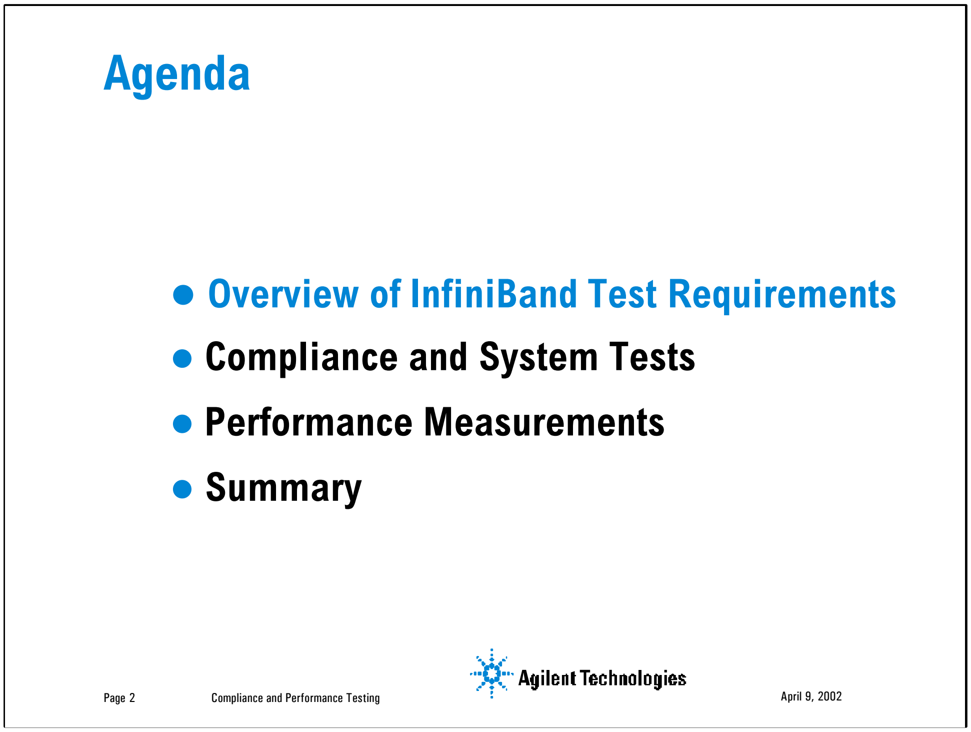
Jennie Grosslight
Thomas Dippon

April 9, 2002

InfiniBand Compliance and System Testing

Switch Performance Measurements

Agilent Technologies

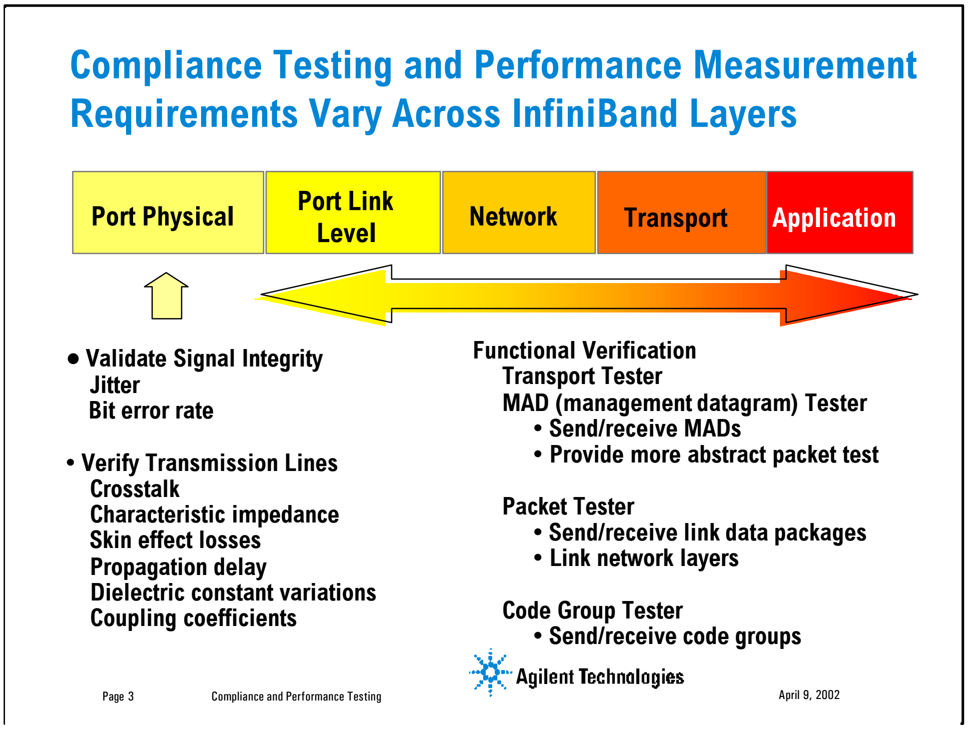


Agenda

- Overview of InfiniBand Test Requirements
- Compliance and System Tests
- Performance Measurements
- Summary

Page 2 Compliance and Performance Testing Agilent Technologies April 9, 2002

This presentation provides a brief overview of the test requirements for an InfiniBand (IB) system, examples of standards compliance and systems test, and new methods of measuring performance using Agilent test solutions.



Compliance Testing and Performance Measurement Requirements Vary Across InfiniBand Layers

Port Physical | Port Link Level | Network | Transport | Application

- Validate Signal Integrity
  Jitter
  Bit error rate

- Verify Transmission Lines
  Crosstalk
  Characteristic impedance
  Skin effect losses
  Propagation delay
  Dielectric constant variations
  Coupling coefficients

Functional Verification
  Transport Tester
  MAD (management datagram) Tester
  - Send/receive MADs
  - Provide more abstract packet test

  Packet Tester
  - Send/receive link data packages
  - Link network layers

  Code Group Tester
  - Send/receive code groups

Agilent Technologies

Page 3 Compliance and Performance Testing April 9, 2002

Supporting InfiniBand requires a huge, dynamic range of technologies--from microwave down to the lowest physical layers, all the way up through many layers of the protocol stack.

At the physical layer it is important to verify signal integrity with jitter and bit error rate measurements. Transmission line measurements are critical for signals moving at

2.5 Gb/s. Important tests as you move up the stack of InfiniBand layers include:

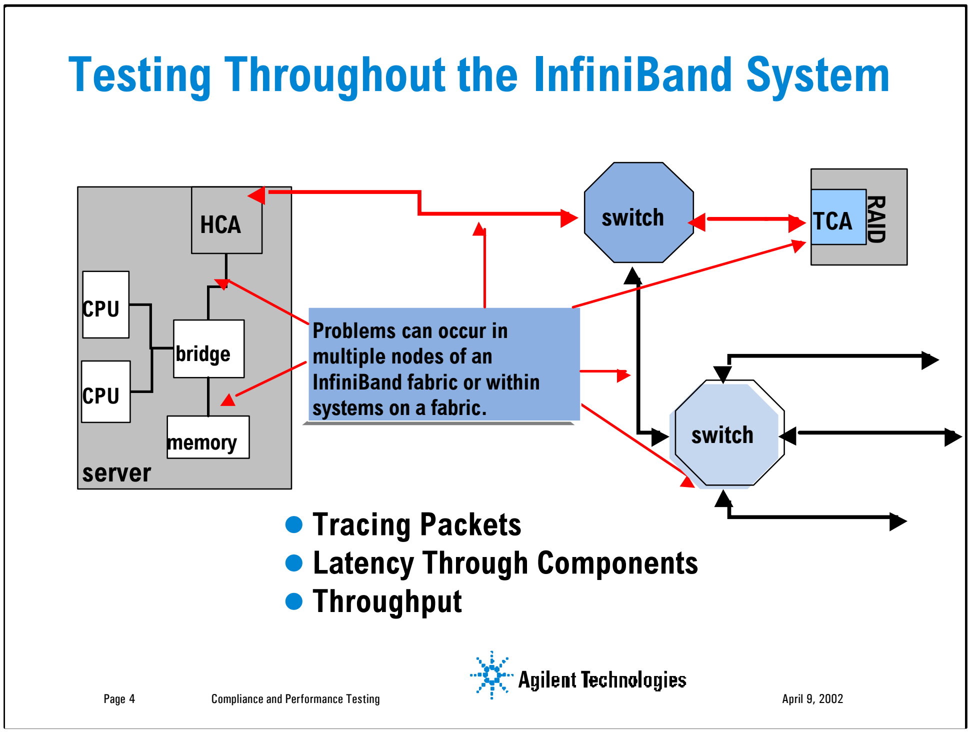
Functional verification

Transport tester

MAD (management datagram) tester

Packet tester

Code group tester



Testing Throughout the InfiniBand System

Problems can occur in multiple nodes of an InfiniBand fabric or within systems on a fabric.

- Tracing Packets
- Latency Through Components
- Throughput

Page 4 Compliance and Performance Testing Agilent Technologies April 9, 2002

Broad visibility, cross-correlation, and stimulus diversity are critical for effective system-level validation.

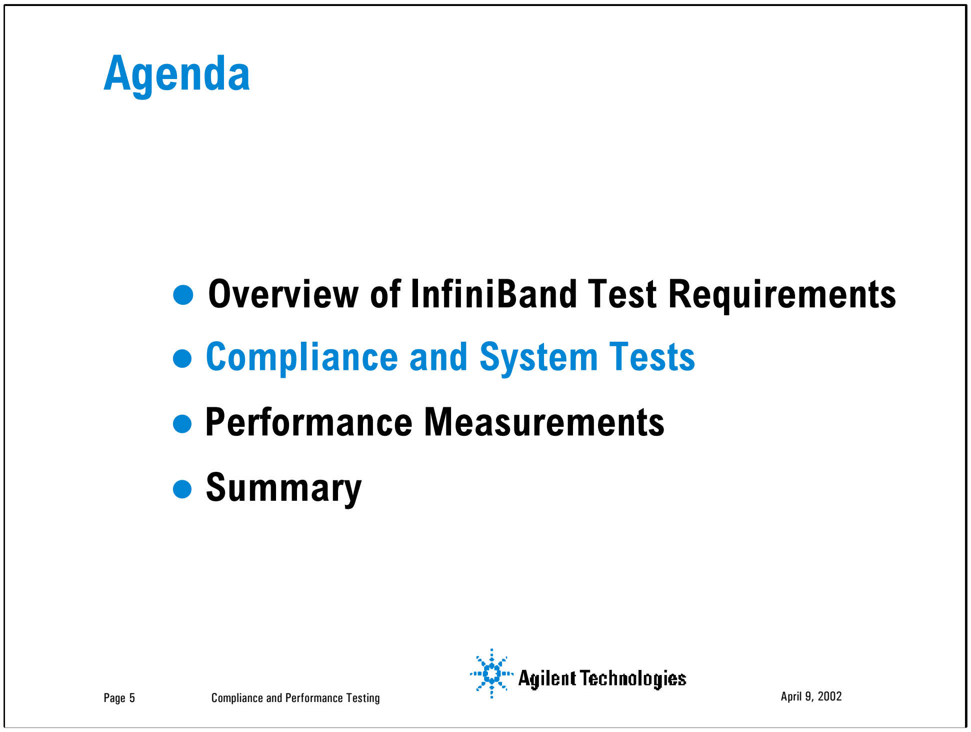
Broad visibility: Visibility across a range of business functions and links within the system is necessary for tracking down system-level problems. You need insight into the behavior of each link at the picosecond level all the way up to the highest packet level. Broad visibility is important for matching up request/response pairs in a protocol, for example.

Cross-correlation: You need the ability to correlate activity in different parts of an InfiniBand system. The symptom of a problem may appear in one part of the system, but the root cause may lie somewhere else. We'll examine two important kinds of correlation for system-level test.

Spatial correlation: You want to see how an event in one part of a system relates to an event in another part of the system and view both events simultaneously. Spatial correlation allows for latency measurements and tracing traffic through a system. Another form is temporal correlation--that is, how an event at one point in time relates to an event at another point in time. A good example is matching packets.

Stimulus diversity: You can't model 100% of the real world to cover all possible conditions and corner cases that your system will handle. You need to test your system with real-world stimulus, but you need a way of providing stimulus that is focused, directed, and controllable in order to reproduce and debug problems. Setting up regression tests that have very specific test cases, with very specific stimulus, that exercise the whole system, is a good practice.

In summary, you need to be able to work all across the system, to be able to correlate activities--from picoseconds to packets--and to be able to not only handle the billions of events that occur in the real world, but provide very focused, directed, controllable stimulus to your system.

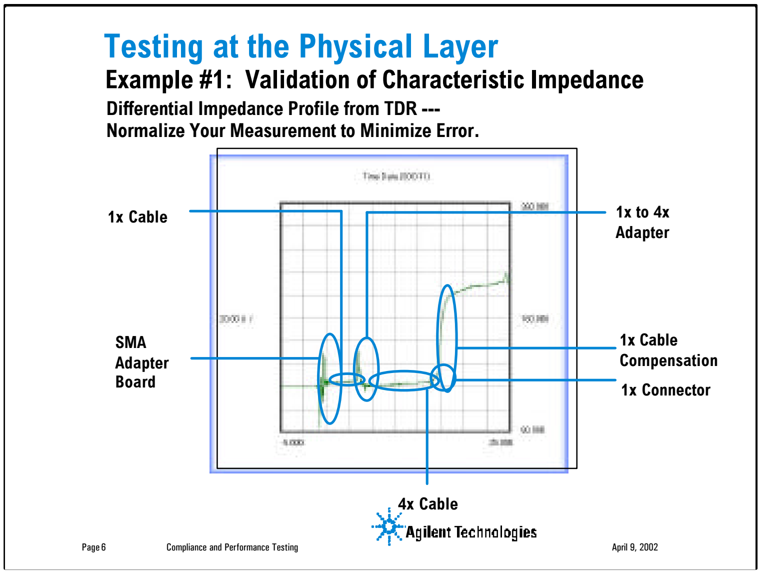


Agenda

- Overview of InfiniBand Test Requirements
- Compliance and System Tests
- Performance Measurements
- Summary

Page 5 Compliance and Performance Testing Agilent Technologies April 9, 2002

Now we will discuss several examples of standards compliance and system-level testing.



Testing at the Physical Layer

Example #1: Validation of Characteristic Impedance

Differential Impedance Profile from TDR ---
Normalize Your Measurement to Minimize Error.

Agilent Technologies

Page 6 Compliance and Performance Testing April 9, 2002

The example shown here is a surface mount adapter board connected to a length of 1x cable, a 1x to 4x adapter, followed by a length of 4x cable, and another connector.

The InfiniBand specification has rather tight tolerances on transmission line impedance. “Normalizing” or calibrating your time domain reflectometry (TDR) system provides the most accurate measurements. Systems that do not provide for complete calibration, or failure to calibrate, may result in measurement errors.

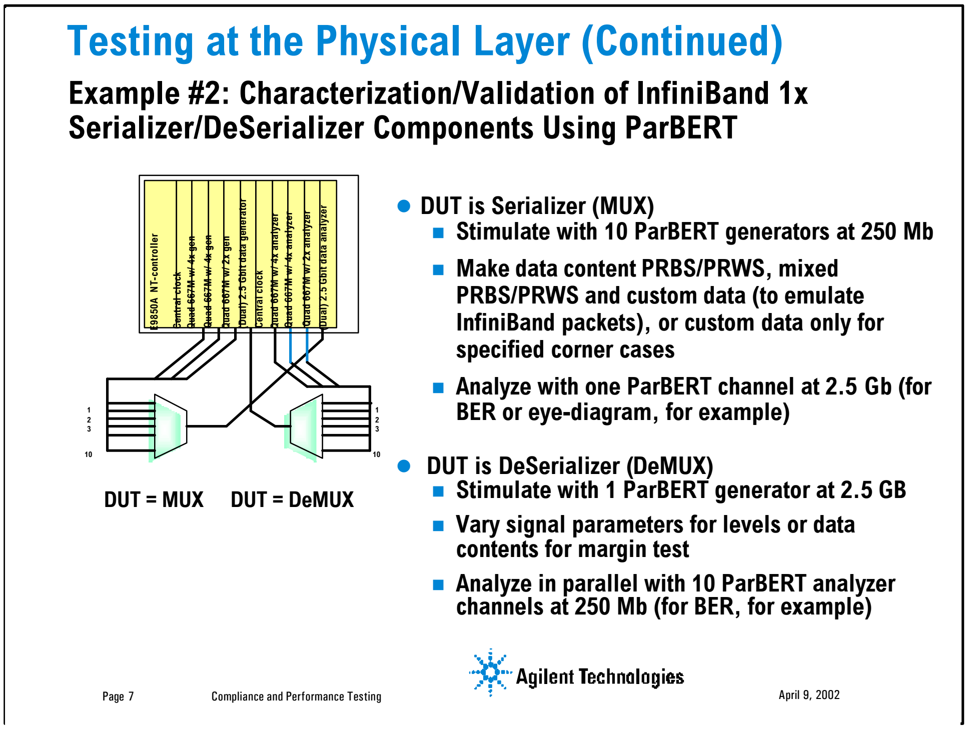
Vector network analyzers (VNAs) or TDR scopes make accurate and in-depth measurements useful for:

Finding elusive signal integrity problems

Extracting superior model parameters

Removing fixture effects

Measuring trace or cable length



Testing at the Physical Layer (Continued)

Example #2: Characterization/Validation of InfiniBand 1x Serializer/DeSerializer Components Using ParBERT

DUT = MUX   DUT = DeMUX

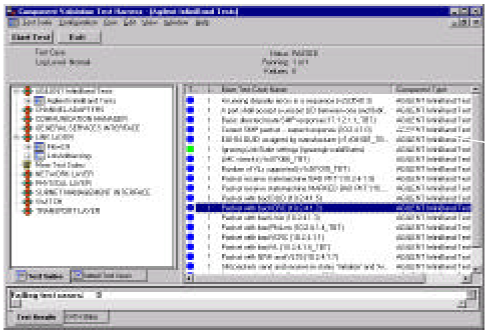
- DUT is Serializer (MUX)
  - Stimulate with 10 ParBERT generators at 250 Mb
  - Make data content PRBS/PRWS, mixed PRBS/PRWS and custom data (to emulate InfiniBand packets), or custom data only for specified corner cases
  - Analyze with one ParBERT channel at 2.5 Gb (for BER or eye-diagram, for example)
- DUT is DeSerializer (DeMUX)
  - Stimulate with 1 ParBERT generator at 2.5 GB
  - Vary signal parameters for levels or data contents for margin test
  - Analyze in parallel with 10 ParBERT analyzer channels at 250 Mb (for BER, for example)

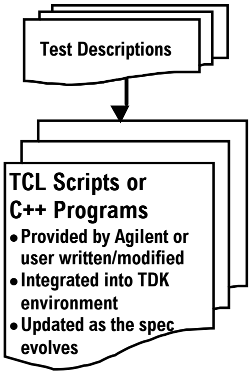
Agilent Technologies

We can characterize InfiniBand serializer and deserializer (SERDES) components by measuring bit error ratios with user-defined data, pseudo-random bit sequence/pseudo-random word sequence (PRBS/PRWS), or mixed data.

Automating Compliance Testing

Example #3: InfiniBand Compliance Test Suite Stimulus and Response





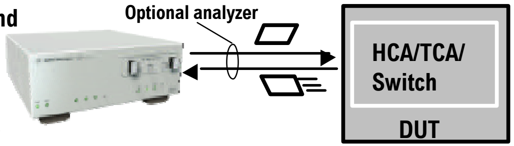
Test Descriptions

TCL Scripts or C++ Programs

- Provided by Agilent or user written/modified
- Integrated into TDK environment
- Updated as the spec evolves

Traffic Generator/Exerciser for InfiniBand

- Acts as a code group tester/packet tester/transport tester/MAD tester as needed
- Includes response and capture capabilities



The Compliance and Interoperability Working Group (CIWG) of the InfiniBand Trade Association (IBTA) is responsible for compliance testing specifications. Agilent has implemented CIWG test descriptions in its E2950 Series test software to provide a push button solution for compliance tests. Several test scripts are included with its exerciser software. Additionally, with the Agilent InfiniBand exerciser, users can customize their testing environment with the C++ or TCL scripting language. The compliance test software is available as InfiniBand exerciser option #100.

An easy-to-use graphical user interface (GUI) for the Agilent E2950 Series test system as well as a COM-Interface are specialized for InfiniBand real-time performance measurements. Other features include:

Functional validation

Compliance testing

Error injection (deterministic generation of error conditions)

Stress testing

Regression testing of corner cases

Application/vendor specific testing

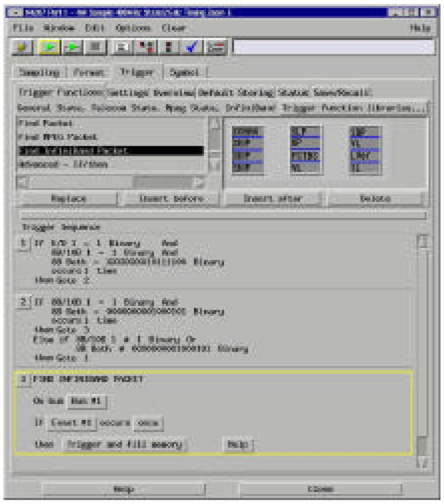
Emulation of host channel adapters (HCAs) / target channel adapters (TCAs) / switches with fully programmable characteristics

Agilent logic analyzers and protocol analyzers can help you examine areas of interest in detail. The logic analyzers are optimized for cross-bus measurements and provide a consistent way of looking at data, no matter what bus you're on.

The protocol analyzers offer a hierarchical view of protocols. You can click on specific packets and pieces of packets and expand them for more detail.

Example #3: Compliance and System Testing (Cont.)

- Set sequential triggering on logic analyzer
  TS1 followed by
  TS2 followed by
  Link packet
- Expand triggers or leave at high level
- Set exerciser to initiate TS



Agilent Technologies

This is an example of how you might set up an InfiniBand trigger on an Agilent 16700 Series logic analysis system. Sequential triggering helps to narrow in on specific events.

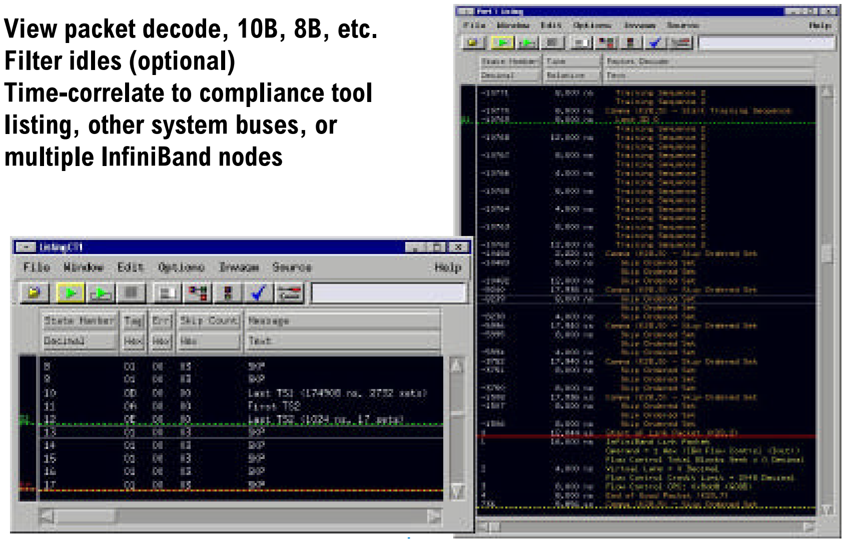
In this example, the trigger is set for a comma, followed by a training sequence, followed by a link packet. The exerciser initiates the training sequence. The logic analysis system triggers on the first link packet after the training sequence.

Trigger descriptions can be expanded to the bit level or viewed at a high level of user-definable names, such as “Event #1” or “Link packet”. Event # is the default value.

This way of setting up triggers and looking at results is very similar to the way that it traditionally has been done with microprocessors and other kinds of buses in a system. The logic analysis system is optimized for cross-bus measurements and provides a consistent way of looking at data across buses.

Example #3: Compliance and System Testing (Cont.)

- View packet decode, 10B, 8B, etc.
- Filter idles (optional)
- Time-correlate to compliance tool listing, other system buses, or multiple InfiniBand nodes


Listing<1>
File Window Edit Options Invasm Source Help
Port 1 Listing
State Number Time Packet Decode
Decimal Relative Text

```
State Number  Tag  Err  Skip Count  Message
Decimal       Hex  Hex  Hex         Text
8             01   00   03          SKP
9             01   00   03          SKP
10            0D   00   00          Last TS1 (174908 ns, 2732 sets)
11            0A   00   00          First TS2
12            0E   00   00          Last TS2 (1024 ns, 17 sets)
13            01   00   03          SKP
14            01   00   03          SKP
15            01   00   03          SKP
16            01   00   03          SKP
17            01   00   03          SKP
```

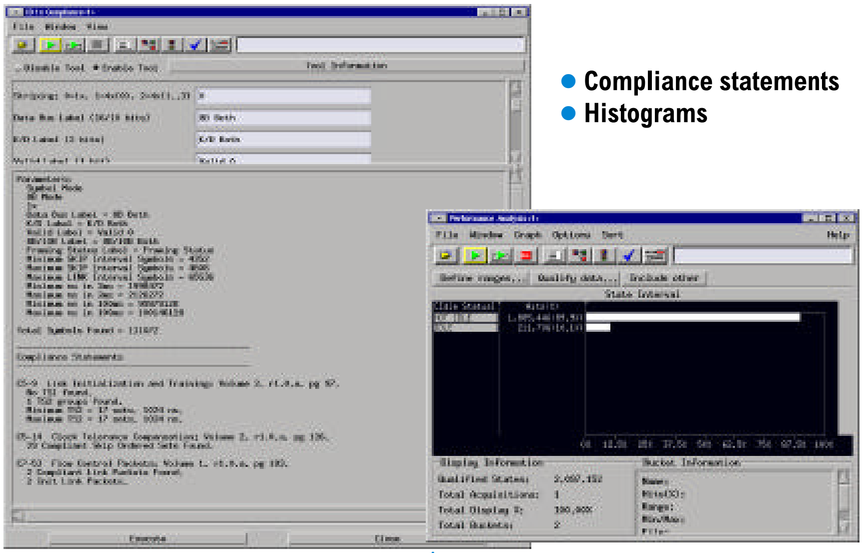
Agilent Technologies

The logic analysis system provides a view of the InfiniBand packet decode along with character codes for 10B, 8B, etc.

The logic analysis system has triggered on the first link packet after the training sequence. You can easily place global markers to move between a listing on a compliance tool and the InfiniBand packet decode, as shown. Global markers also track across all buses or InfiniBand nodes connected to the logic analysis system.

If the InfiniBand link had failed to come up and the logic analysis system had not triggered, we would reset the trigger to look for either the first TS2 or perhaps TS1. In this manner you could determine how far the system had progressed in the link power-up sequence. Traffic between the requestor and responder would point you to the defective node.

Example #3 Continued: Compliance and System Testing (Cont.)



```
Skipping: 0=1x, 1=4x(0), 2=4x(1..3)   0
Data Bus Label (8/10 bits)   8B Both
K/D Label (2 bits)   K/D Both
Valid Label (1 bit)   Valid 0

Parameters:
  Symbol Mode
  8B Mode
  1x
  Data Bus Label = 8B Both
  K/D Label = K/D Both
  Valid Label = Valid 0
  8B/10B Label = 8B/10B Both
  Framing Status Label = Framing Status
  Minimum SKIP Interval Symbols = 4352
  Maximum SKIP Interval Symbols = [illegible]
  Maximum LINK Interval Symbols = 65536
  Minimum ns in 2ms = [illegible]
  Maximum ns in 2ms = [illegible]
  Minimum ns in 100ms = [illegible]
  Maximum ns in 100ms = [illegible]

Total Symbols Found = 131072

Compliance Statements

C5-9  Link Initialization and Training: Volume 2, r1.0.a, pg 97.
  No TS1 found.
  1 TS2 groups found.
  Minimum TS2 = 17 sets, 1024 ns.
  Maximum TS2 = 17 sets, 1024 ns.

C5-14  Clock Tolerance Compensation: Volume 2, r1.0.a, pg 136.
  29 Compliant Skip Ordered Sets Found.

C7-53  Flow Control Packets: Volume 1, r1.0.a, pg 183.
  2 Compliant Link Packets Found.
  2 Init Link Packets.
```

```
Qualified States:      2,097,152
Total Acquisitions:    1
Total Display %:       100.00%
Total Buckets:         2

Name:
Hits(%):
Range:
Min/Max:
File:
```

- Compliance statements
- Histograms

Compliance tests on the logic analyzer include:

Link initialization and training (InfiniBand Specification V2, r1.0.a, page 97)

Clock tolerance compensation (InfiniBand Specification V2, r1.0.a, page 96)

Flow control packets (InfiniBand Specification V2, r1.0.a, page 183)

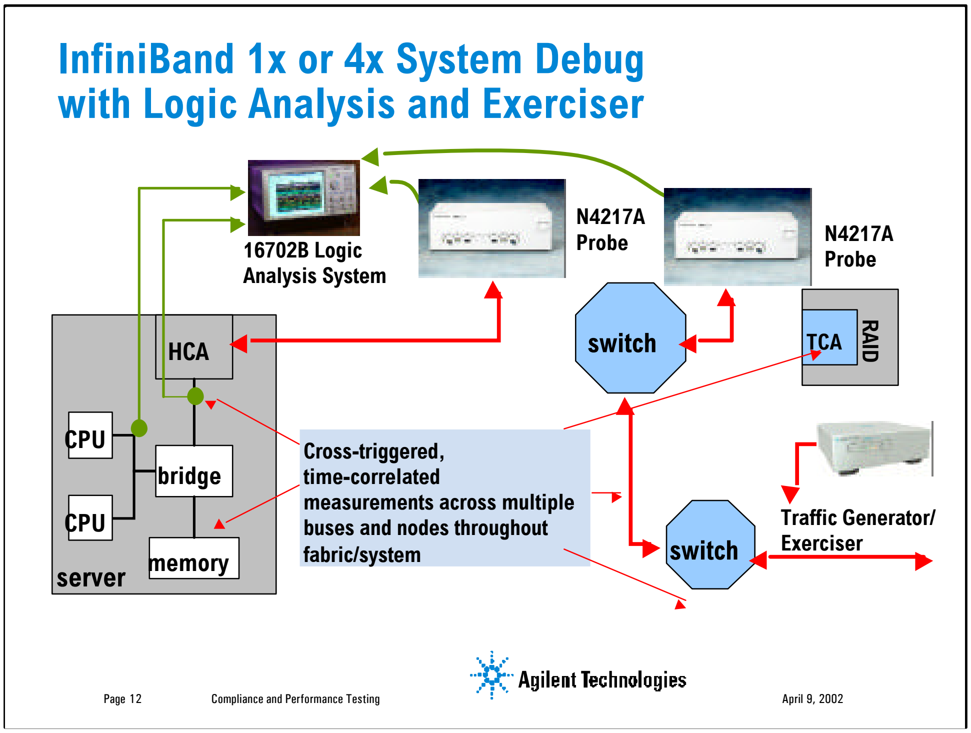
In this example:

No complete TS1 groups were found in the trace (rather shallow 32k trace)

One full TS2 group with17sets was found

There were 29 compliant skip ordered sets found

There were 2 compliant (Init) link packets found



InfiniBand 1x or 4x System Debug with Logic Analysis and Exerciser

16702B Logic Analysis System
N4217A Probe
N4217A Probe
switch
RAID
TCA
HCA
CPU
bridge
CPU
memory
server
Cross-triggered, time-correlated measurements across multiple buses and nodes throughout fabric/system
switch
Traffic Generator/ Exerciser

Page 12 Compliance and Performance Testing Agilent Technologies April 9, 2002

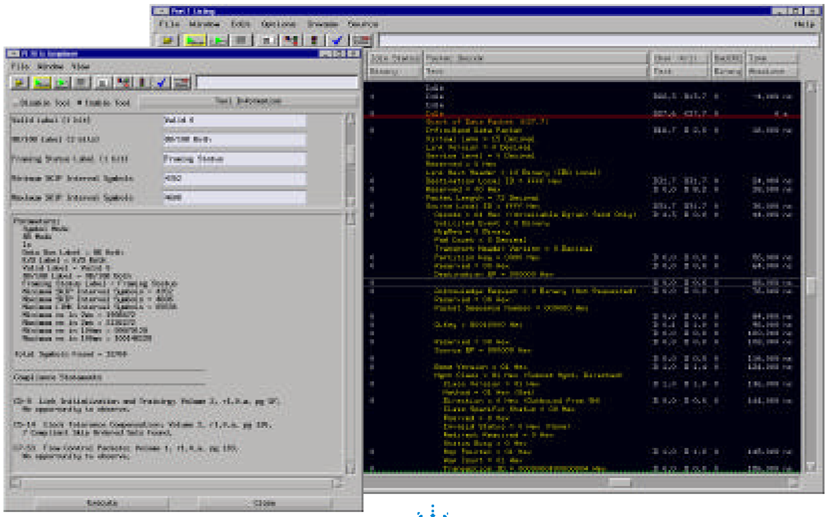
This diagram shows Agilent test instruments for complete system test and validation attached to multiple nodes in the InfiniBand fabric, a central processing unit (CPU) front-side bus, and a peripheral component interconnect (PCI) bus. We can probe different channels and buses to find problems and trace them to root causes, providing visibility into the different parts of the system.

Signal tracing of a faulty direct memory access (DMA) transfer is a straightforward example of system testing. The easiest way to debug a problem like this is to follow the data through the system by moving the scope probe from point to point through the circuit. We set up an analyzer to look at the InfiniBand traffic, trigger on the remote direct memory access (RDMA) operation, and follow the data as it goes through the PCI bus into the memory system. We also take a look at the CPU-to-chipset communication because that's probably how the memory management units (MMUs) are configured and programmed. Remember, InfiniBand DMA transfers are located in the virtual memory space, so MMUs are involved.

Using a traffic generator, shown here, we can generate the RDMA request and tell the traffic generator to use a very specific data pattern in the DMA buffer to make it easy to follow the signature as it flows through the system.

We can trigger the 16700 Series logic analysis system on the DMA request and trigger the PCI and memory analyzers on the data pattern. If the request is making it out to the PCI bus or the memory bus, we'll find it and trigger. We can see where and when the request went. We also can look at the CPU bus for some insight into how the MMU is programmed.

InfiniBand 4x Compliance Tests and Packet Decode on the Logic Analyzer



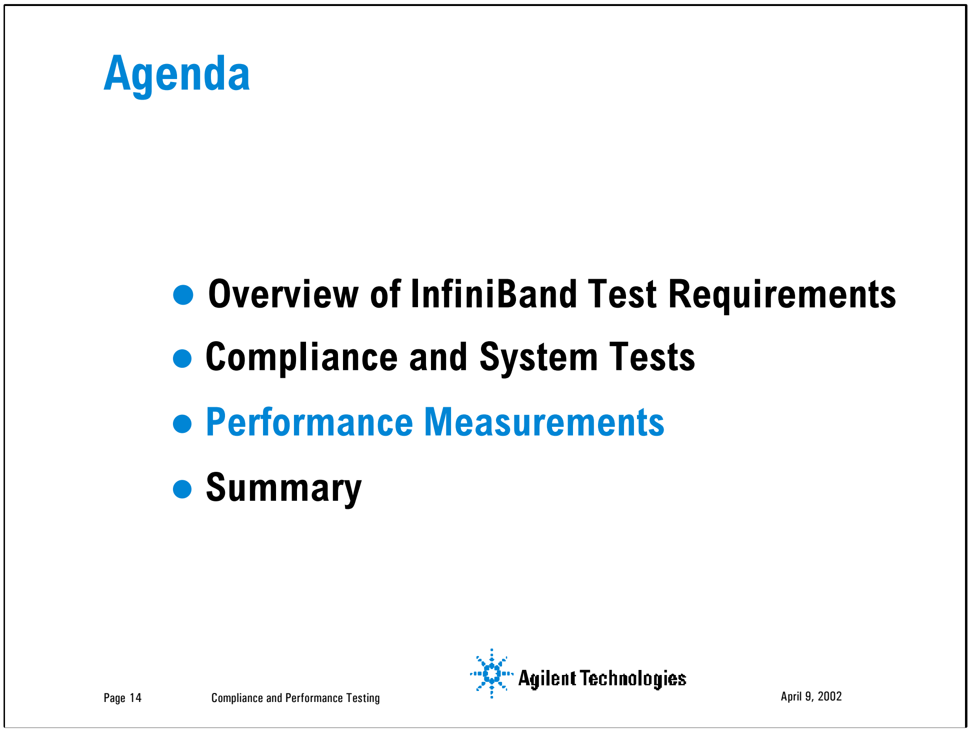
Examples of InfiniBand 4x triggering, viewing, packet decode and compliance testing are similar to 1x examples, except that Lane deskew is unique to 4x.

The InfiniBand specification provides systematic support for compliance testing. Distributed throughout the specification are statements that highlight key conditions all products must satisfy to comply. These statements often stand independently, but they also can refer to other parts of the specification that contain details supporting the compliance. The following is an example taken from Volume 1, Release 1.0.a of the InfiniBand specification:

"C7-53: Flow control packets shall be sent for each VL, except VL15, upon entering the LinkInitialize state. When in the PortStates LinkInitialize, LinkArm or LinkActive, a flow control packet for a given virtual lane shall be transmitted prior to the passing of 65,536 symbol times since the last time a flow control packet for the given virtual lane was transmitted." ©IBTA

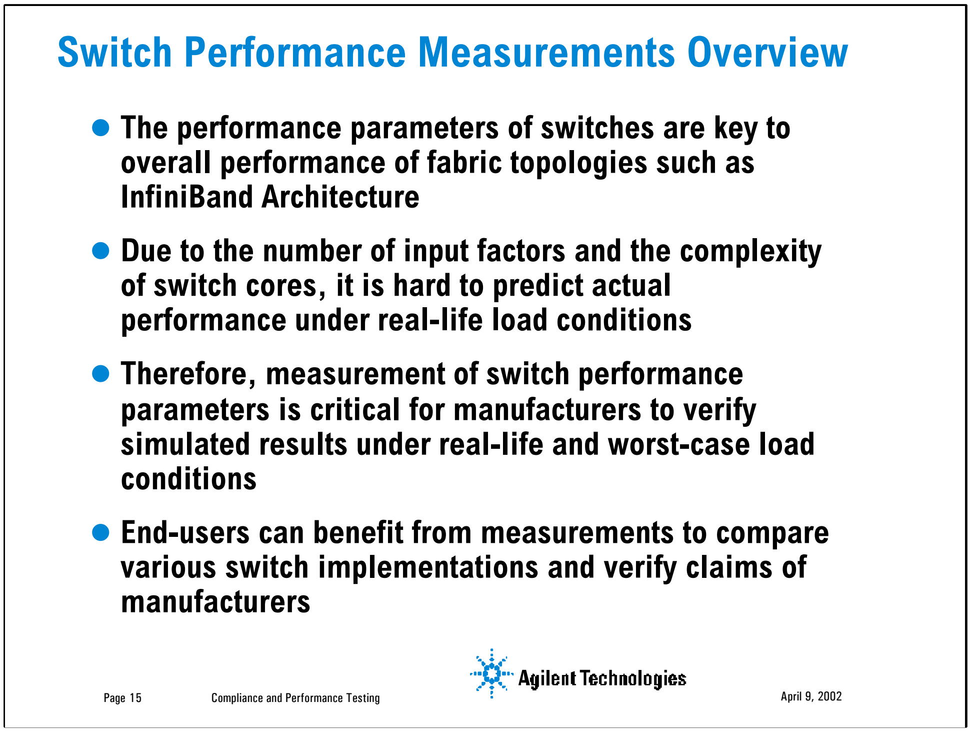
Each compliance statement is numbered for reference, and all statements are summarized into tables that appear at the end of each volume of the specification. Associated with each compliance statement are a set of test assertions, which are defined as atomic statements resulting from the decomposition of a compliance statement into individual components that can be shown to be either true or false by running a test or set of tests. For example, an assertion associated with statement C7-53 above could be stated as: "Flow Packet transmitted at least every 65536 symbol times."

From the assertions, a set of actual tests is defined by the InfiniBand specification. Each test may demonstrate the truth or falsehood of one or more assertions. A test identifies the assertion(s) it covers, the test equipment setup, and the test procedure. This information then can be used as a springboard for developing instrument control and analysis commands that implement a specific test.



Agenda

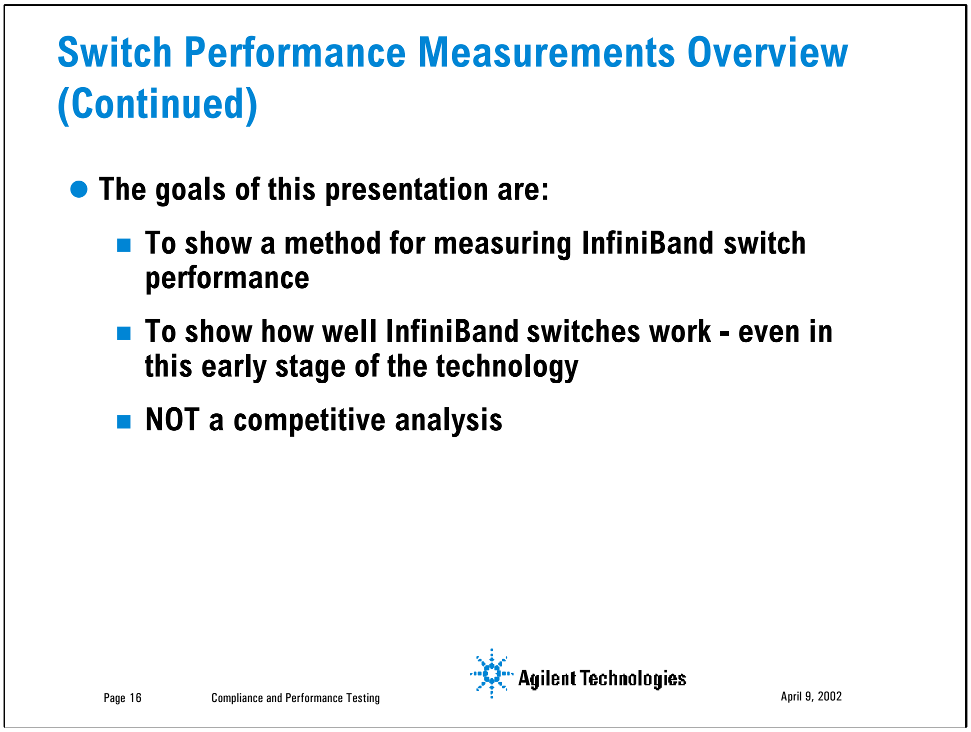
- Overview of InfiniBand Test Requirements
- Compliance and System Tests
- Performance Measurements
- Summary



Switch Performance Measurements Overview

- The performance parameters of switches are key to overall performance of fabric topologies such as InfiniBand Architecture
- Due to the number of input factors and the complexity of switch cores, it is hard to predict actual performance under real-life load conditions
- Therefore, measurement of switch performance parameters is critical for manufacturers to verify simulated results under real-life and worst-case load conditions
- End-users can benefit from measurements to compare various switch implementations and verify claims of manufacturers

This slide gives a couple of reasons why switch performance measurements are important for fabric topologies. Performance measurements can be interesting for both designers of switch chips, switch products and systems as well as end-users of the technology.

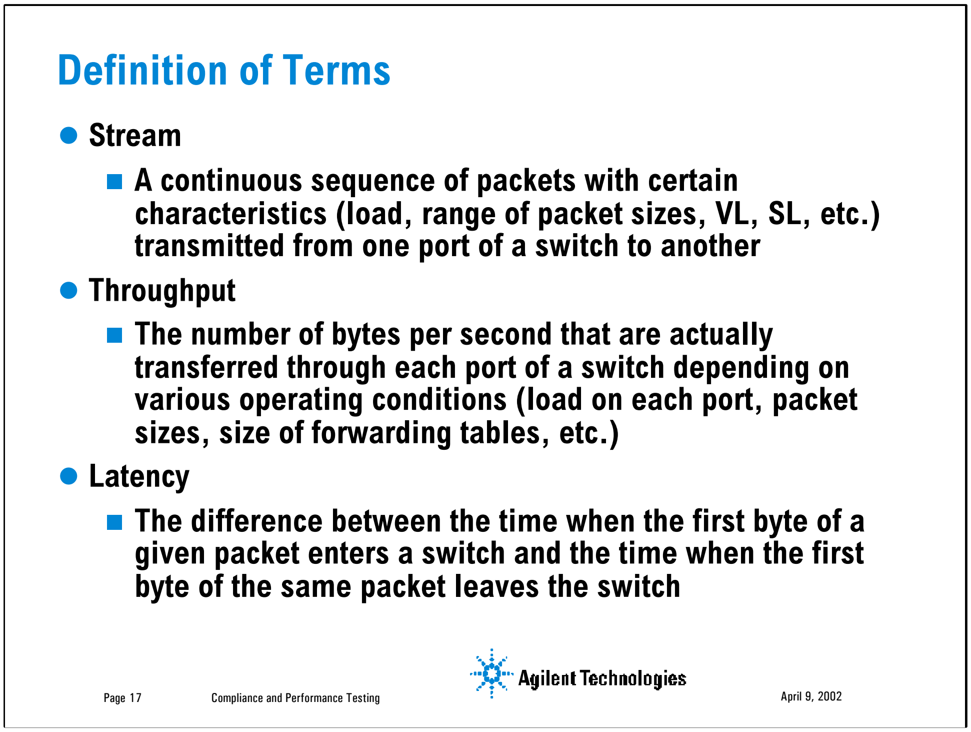


Switch Performance Measurements Overview (Continued)

- The goals of this presentation are:
  - To show a method for measuring InfiniBand switch performance
  - To show how well InfiniBand switches work - even in this early stage of the technology
  - NOT a competitive analysis

Page 16 Compliance and Performance Testing Agilent Technologies April 9, 2002

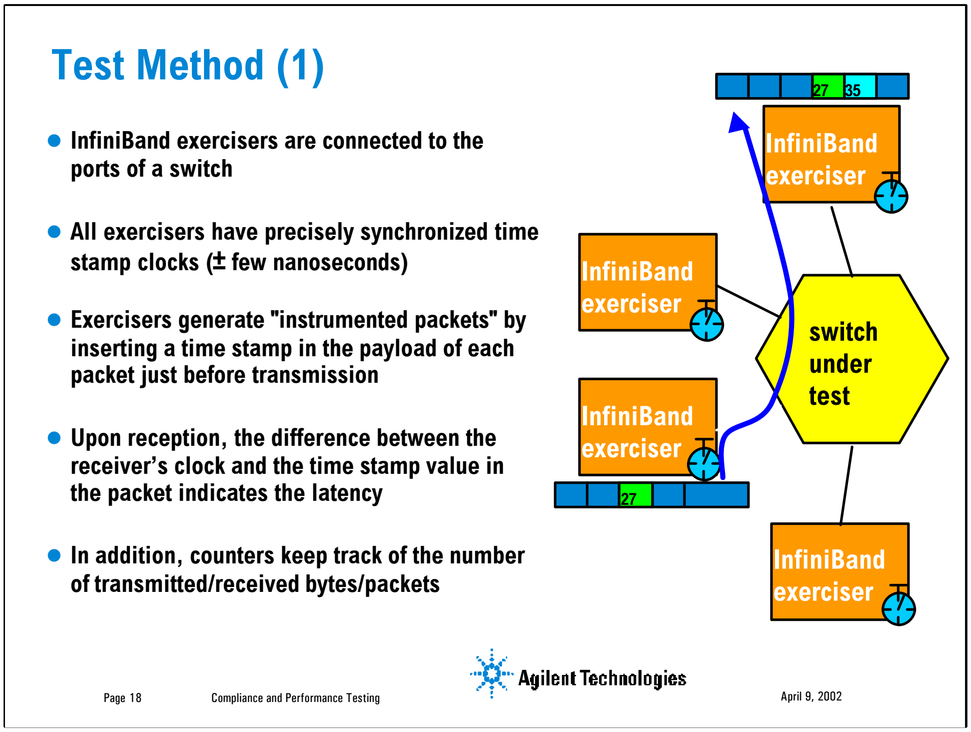
Summary of presentation goals.



Definition of Terms

- Stream
  - A continuous sequence of packets with certain characteristics (load, range of packet sizes, VL, SL, etc.) transmitted from one port of a switch to another
- Throughput
  - The number of bytes per second that are actually transferred through each port of a switch depending on various operating conditions (load on each port, packet sizes, size of forwarding tables, etc.)
- Latency
  - The difference between the time when the first byte of a given packet enters a switch and the time when the first byte of the same packet leaves the switch

A definition of terms that are used throughout the next couple of slides.



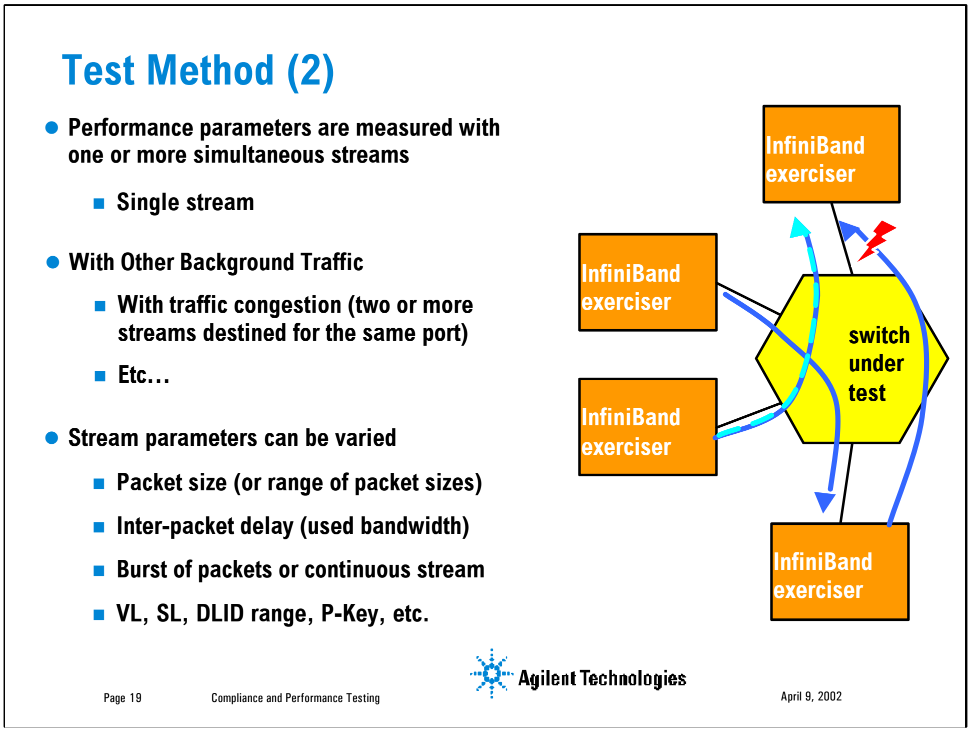
Test Method (1)

- InfiniBand exercisers are connected to the ports of a switch
- All exercisers have precisely synchronized time stamp clocks (± few nanoseconds)
- Exercisers generate "instrumented packets" by inserting a time stamp in the payload of each packet just before transmission
- Upon reception, the difference between the receiver's clock and the time stamp value in the packet indicates the latency
- In addition, counters keep track of the number of transmitted/received bytes/packets

Page 18 Compliance and Performance Testing Agilent Technologies April 9, 2002

With multiple InfiniBand exercisers, it is possible to conduct switch performance measurements under deterministic load conditions, using various packet sizes, VL, and so on.

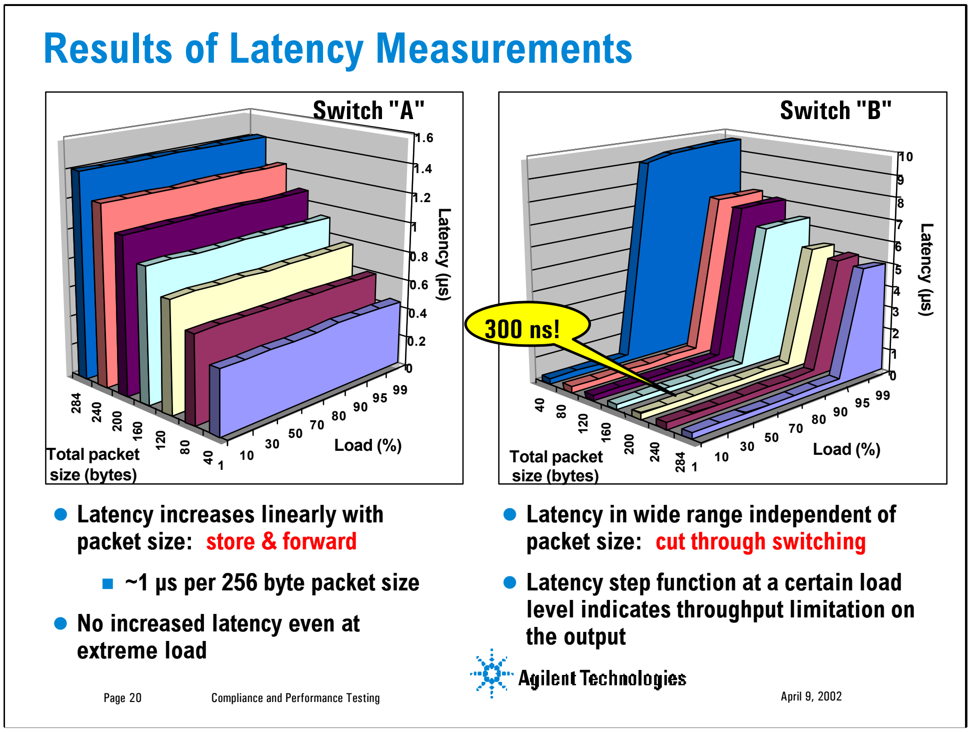
The basic idea behind measuring latency through a switch is to transmit so called "instrumented packets" from one port to another. Instrumented packets have a time stamp in their payload which is generated by the exerciser that transmits the packet and evaluated by the exerciser that receives a packet. With multiple exercisers generating traffic on multiple ports, latency can be measured under stress conditions.



Test Method (2)

- Performance parameters are measured with one or more simultaneous streams
  - Single stream
- With Other Background Traffic
  - With traffic congestion (two or more streams destined for the same port)
  - Etc…
- Stream parameters can be varied
  - Packet size (or range of packet sizes)
  - Inter-packet delay (used bandwidth)
  - Burst of packets or continuous stream
  - VL, SL, DLID range, P-Key, etc.

The concept of multiple "streams" allows performance measurements not only for a single packet traversing an otherwise idle switch, but to measure the switch's behavior in congestion scenarios, or when multiple ports are in use. Another important aspect is the dependency of performance and the type of "traffic" going through the switch. Are there many small packets or fewer large packets? What is the inter-packet delay? Does the switch have to perform P-Key checking. Is the linear or random forwarding table used?



Results of Latency Measurements

- Latency increases linearly with packet size: store & forward
  - ~1 µs per 256 byte packet size
- No increased latency even at extreme load

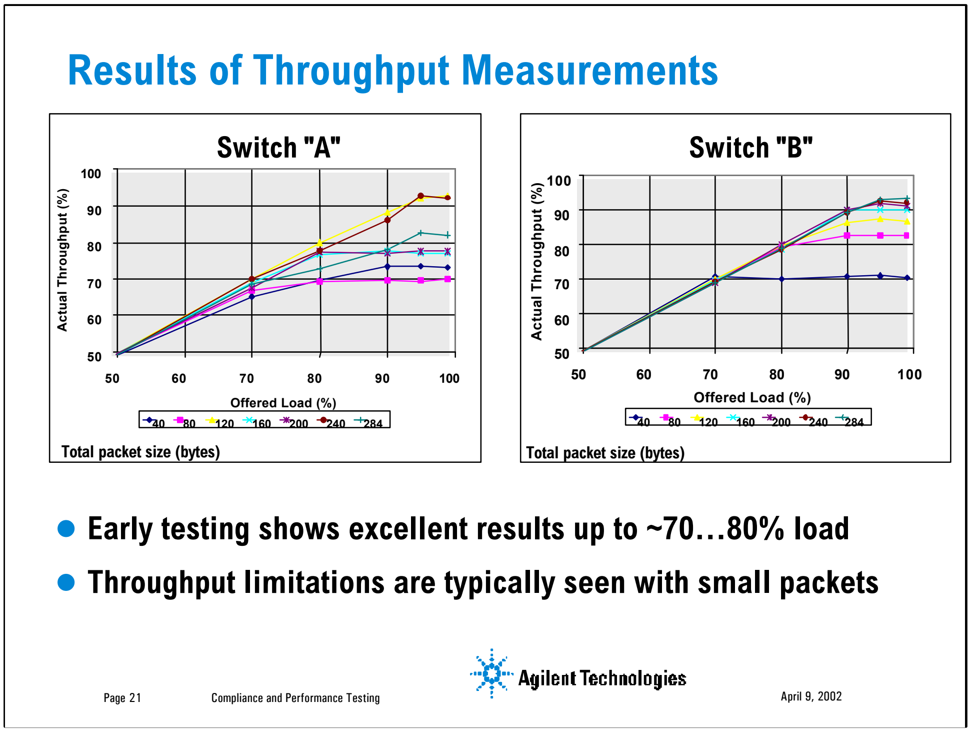
- Latency in wide range independent of packet size: cut through switching
- Latency step function at a certain load level indicates throughput limitation on the output

Agilent Technologies

Page 20 Compliance and Performance Testing April 9, 2002

This slide shows a series of latency measurements on two existing InfiniBand 1x switches. In this measurement, only two ports of the switch were connected to InfiniBand exercisers and traffic was sent from one exerciser to the other (single stream).

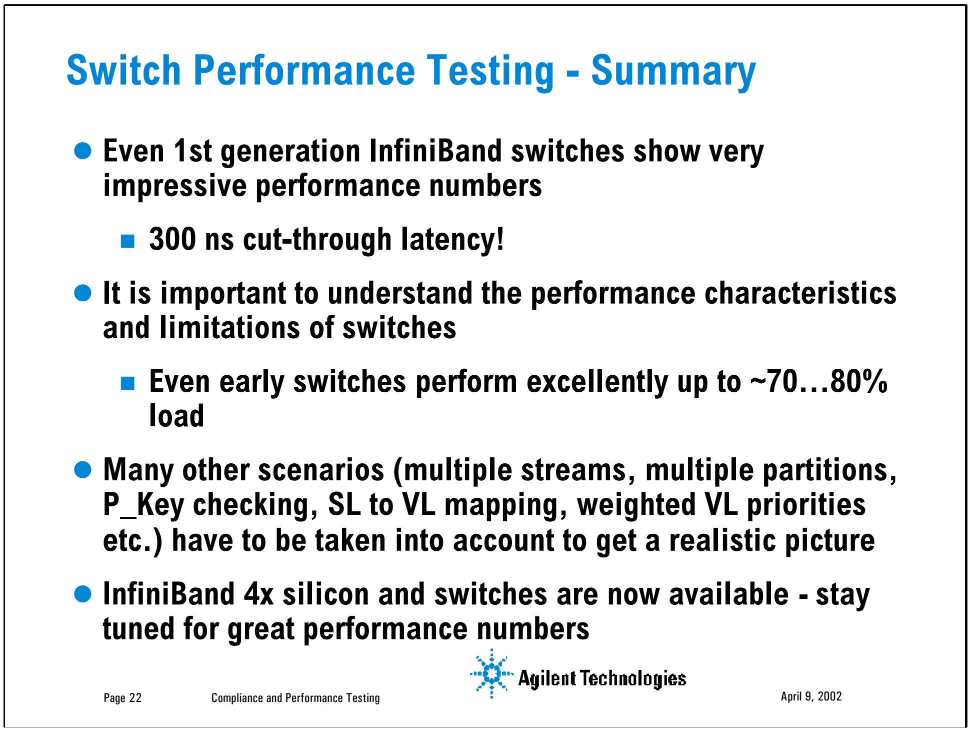
The results are quite different for the two switches in the test. Whereas switch "A" on the left shows a constant latency independent of the load offered on the input, switch "B" shows a significant increase in latency above ~ 90% load, especially with small packets. On the other hand, below 90% load the performance of switch "B" is extremely good-- around 300 ns independent of the packet size. Switch "A", on the other hand, shows an increased latency with larger packets. This is due to the store & forward behavior of switch "A". The first byte of the packet is only transmitted after the last byte has been received.



Results of Throughput Measurements

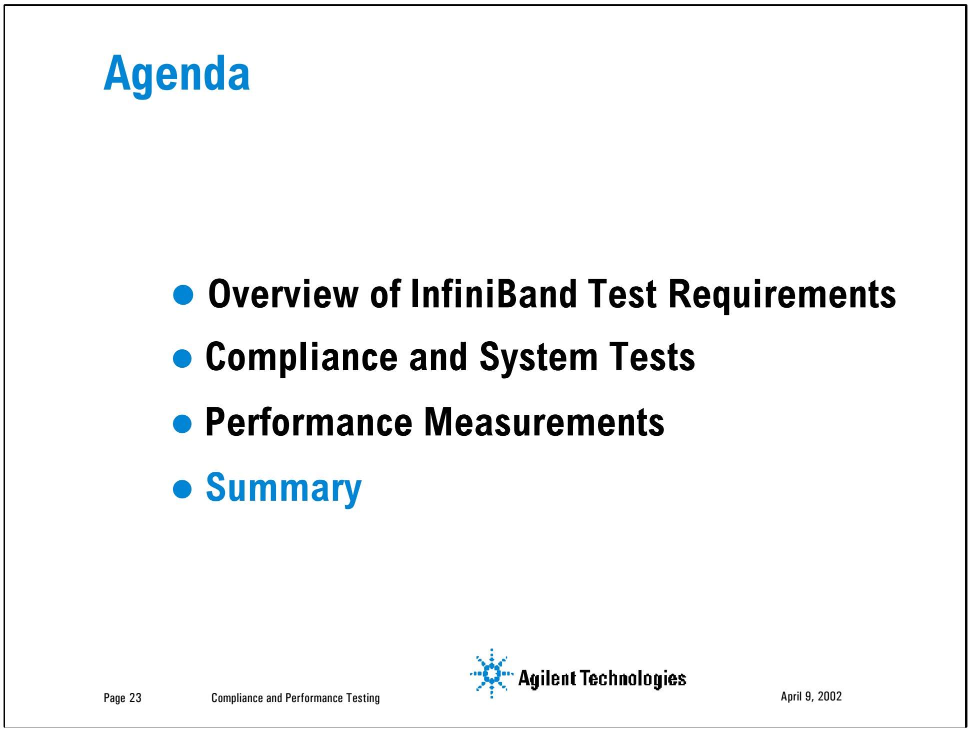
- **Early testing shows excellent results up to ~70…80% load**
- **Throughput limitations are typically seen with small packets**

The throughput measurements made with the same switches show that the switches can handle up to about 70% of the theoretical "wire" bandwidth without problems. Above 70...80%, they start to saturate - especially with small packets.



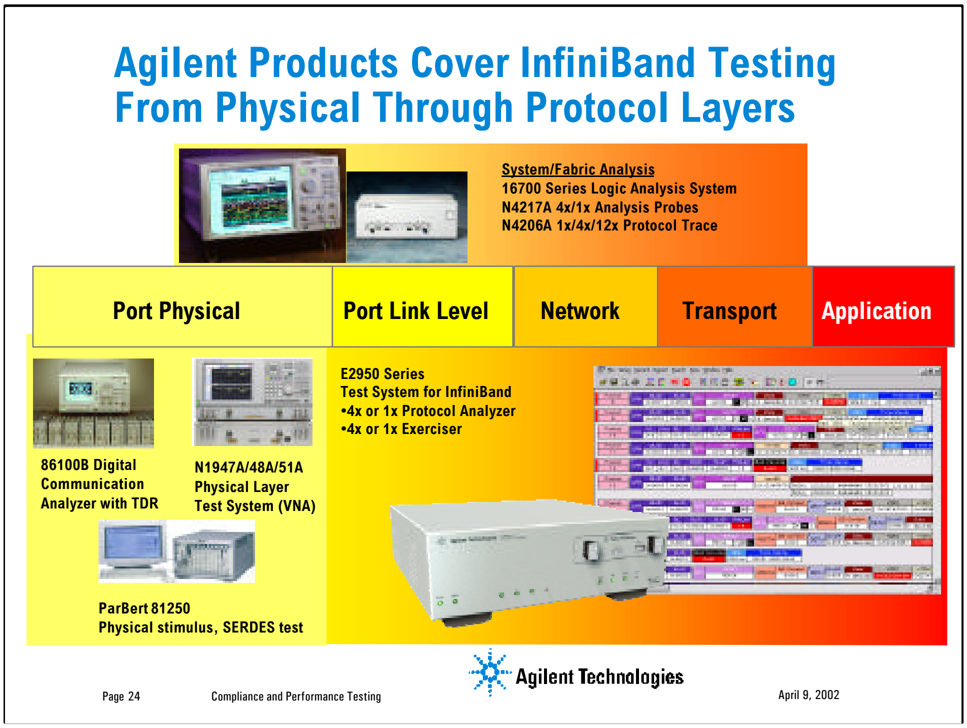
Switch Performance Testing - Summary

- Even 1st generation InfiniBand switches show very impressive performance numbers
  - 300 ns cut-through latency!
- It is important to understand the performance characteristics and limitations of switches
  - Even early switches perform excellently up to ~70…80% load
- Many other scenarios (multiple streams, multiple partitions, P_Key checking, SL to VL mapping, weighted VL priorities etc.) have to be taken into account to get a realistic picture
- InfiniBand 4x silicon and switches are now available - stay tuned for great performance numbers



Agenda

- Overview of InfiniBand Test Requirements
- Compliance and System Tests
- Performance Measurements
- Summary



Agilent Products Cover InfiniBand Testing From Physical Through Protocol Layers

System/Fabric Analysis
16700 Series Logic Analysis System
N4217A 4x/1x Analysis Probes
N4206A 1x/4x/12x Protocol Trace

Port Physical	Port Link Level	Network	Transport	Application

86100B Digital Communication Analyzer with TDR

N1947A/48A/51A Physical Layer Test System (VNA)

ParBert 81250
Physical stimulus, SERDES test

E2950 Series
Test System for InfiniBand
•4x or 1x Protocol Analyzer
•4x or 1x Exerciser

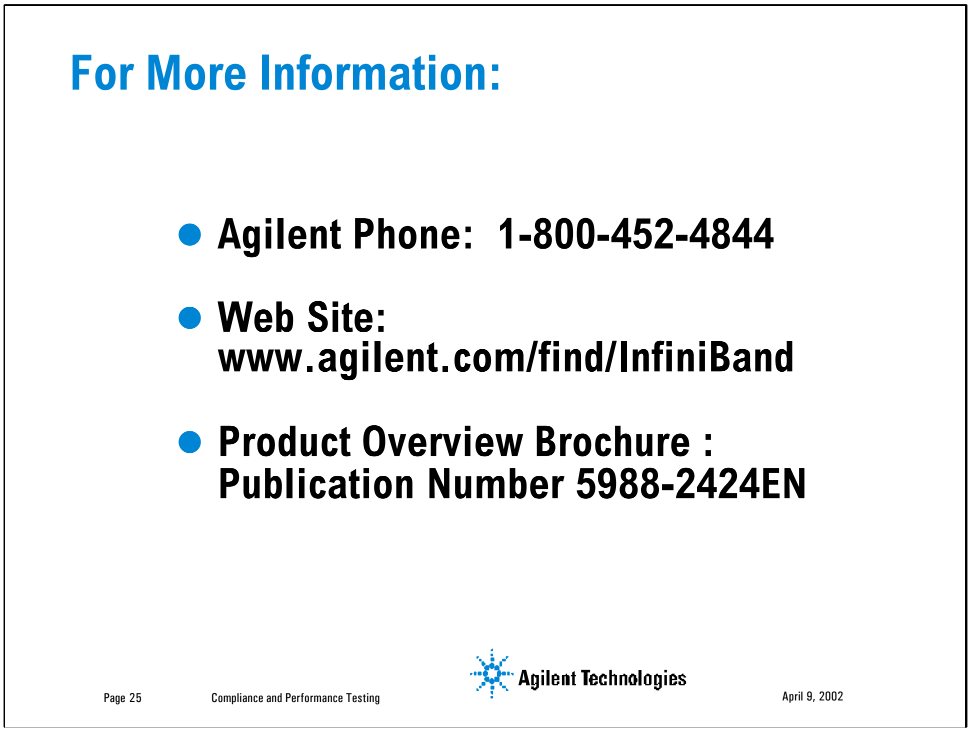
Agilent Technologies

Agilent provides the most complete set of solutions for testing InfiniBand devices and systems (fabrics). Testing can be done at the physical layer and at every layer of the InfiniBand protocol.

Here are examples of tools that can help you at various levels. At the physical layer TDR, scopes, bit error rate testers, VNAs, and spectrum analyzers provide insight.

Protocol analyzers work at the 10B level and go up through the networking, transport, and application levels.

Logic analysis systems straddle the middle. They start working at the 10B level and then move up into the networking and transport layers. The strength of the logic analysis system is in the ability to perform cross-bus analysis. As you move higher up you really want to transition to protocol analyzers as the tool of choice.



For More Information:

- Agilent Phone: 1-800-452-4844
- Web Site: www.agilent.com/find/InfiniBand
- Product Overview Brochure : Publication Number 5988-2424EN

Page 25 Compliance and Performance Testing Agilent Technologies April 9, 2002

Additional information on the Agilent test tools for InfiniBand is available.

The brochure "Test Tools for InfiniBand” provides an overview of the Agilent test solutions. For more detailed information, refer to the available data sheets.

The current E2950 software with the online manuals are available on our WebPages. The protocol analyzer software includes InfiniBand traffic samples to help you evaluate the product.