==========
D I G I T A L
==========

I N T E R O F F I C E M E M O R A N D U M

TO: Distribution

DATE: 18 June 1982
FROM: Tony Sukiennik
DEPT: Cluster Task Force
EXT: 264-4727
L/MS: MK1-1G31
ENET: FREMEN::SMURF::Sukiennik

SUBJ: Analysis and Recommendations regarding the Clusters Program

Multi-computer systems, including those comprised of Personal Computers, are becoming more prevalent in the marketplace. Increased requirements for reliability are just one of the reasons for this trend. We can no longer concentrate solely on engineering, manufacturing, selling, and servicing single computer systems.

In the future, there will be less differentiation between the hardware and software of various computer vendors. Today's innovation will be a commodity tomorrow. Producing quality hardware and software will always be important, but we can no longer survive in the emerging markets simply by producing the best hardware and software.

The CI Cluster Program provides Digital with opportunity to gain valuable experience in the multi-computer space.

We must set up a structure which allows us to become a leader in the systems integration business. The vendor who is capable of properly characterizing, installing, and servicing their multi-computer systems will likely be the vendor of choice.

The CI Cluster Architecture provides significant technical innovation for Digital. We should not lose sales based on the functionality of the CI Cluster. Neither should we expect to win sales based solely on the functionality of the CI Cluster. Systems analysis and service capabilities tuned to the CI Cluster hardware and software will be of great importance.

The Challenge over the next year will be to ensure that we set up a structure to allow us fully leverage the Cluster Architecture. CI Clusters will be the first complete multi-computer systems Digital delivers to the market. This will allow us to gain valuable experience in the installation and servicing of multi-computer systems. This expertise will be necessary to survive in the newly emerging computing markets.

Clusters Analysis and Recommendations

The Market

'High Availability' is NOT the market for multi-computer systems. 'NonStop' will no longer be a point of differentiation as most vendors add fault tolerance to their products (the commodity effect!).

Enhanced availability is becoming increasingly important in the general purpose computer markets where we do most of our current business. In addition there is also a large demand for the ability to easily increase the capacity of a system. This is where the multi-computer cluster architecture is most applicable.

The newly emerging markets demanding 'systems' defined by personal computers and local area networks require a coherent multi-computer architecture. This market is commonly referred to as the 'Office'. CI Clusters closely match the needs of the traditional general purpose market. The technologies, such as shared data bases, developed for the CI Cluster program should be transportable to the emerging 'system' architecture needed for the office.

The Competition

Tandem Computers will rely less on their 'NonStop' architecture to sell systems in the future. They see the newly emerging market and will try to position themselves to be qualified in this space. They will use Distributed Data Base capabilities, Mail and Transaction Processing software, and Satellite Communications as their new levers into the account.

IBM will introduce products to compete in every market. They have stated their intentions to introduce fault tolerant extensions to one of their mainstream architectures. In most cases, IBM's hardware and software will not be the best available. The major threat from IBM is their potential (emphasize POTENTIAL) to be the best systems integrator (primarily of their own gear) in the industry. They have not always demonstrated aptitude for this, but they may view it as a strategic element in the near future.

Stratus, August Systems, and Intel are introducing multi-computer products into a variety of marketplaces. They are mentioned here to allow us to assess the impact of some of the new technologies. None of these vendors is an immediate threat to us, but their methods deserve watching.

Digital's Position

The CI Cluster architecture will make us competitive with any vendor now competing for general purpose applications with needs for enhanced availability or system capacity.

Clusters Analysis and Recommendations

WE SHOULD NOT LOSE SALES BASED ON THE FUNCTIONALITY OF THE CI CLUSTER. It is conceivable that we could lose on price for the lower end systems. We might also be vulnerable to price/performance attacks in the higher end configurations between now and the introduction of our next generation of VAX processors.

In order to properly leverage our efforts in the multi-computer space I recommend the following:

1. Agressively introduce the CI Cluster Program to the market with a program announcement (FALL 82) and product announcements (SPRING 83). The motivation behind the program announcement is to gain visibility in the market at the earliest possible time. We must not allow any further competition to gain footholds in application areas of importance to us.

2. Explore the possibility of providing a subset of CI Cluster functions on the NI.

3. Extend the 'systems oriented' programs recently intitiated in 32 bit engineering. We need to be able to fully characterize, install, maintain, and refine our multi-computer offerings. We must develop the tools and expertise to configure, test, monitor, and tune these multi-computer systems.

4. Provide services (both remedial and consulting) which complement the hardware and software being produced in the CI Cluster program. These services must be built upon the knowledge gathered in the systems programs initiated in engineering. Our service organization has a mass which can be used as a significant advantage over smaller vendors.

5. Start building a second generation of high speed computer interconnects. Serious thought should be given to merging the NI and CI programs. This effort would produce better price/performance and also avoid product confusion similar to that which exists today with the UNIBUS/QBUS.

6. Initiate programs to move the technology developed for the CI Clusters program (ie., shared data bases, etc.) to our other multi-computer programs (ie., Personal Computers, LANS, etc.). The hardware/software/service technology established by the CI Clusters program will make our other multi-computer architectures more viable.

7. Fully quantify the impact of the Ulysses communications switch. If it is perceived to be deficient in any way, we must be prepared to fill the holes with other available solutions.

8. The VAX Information Architecture provides Digital with powerful tools for both the general purpose market and newly emerging office markets. We must insure that VIA takes full advantage of new base VMS functionality produced as a result of the CI Cluster Program.

The attached DRAFT document goes into more detail for each of the points discussed above. Please get back to me with your comments.

Clusters

Analysis and Recommendations

Tony Sukiennik

18-June-1982

Outline

1. What is a Cluster?
2. The Market
    1. The High Availability Market
    2. The General Purpose Computing Market
    3. The Emerging Market
3. The Competition
    1. Tandem
    2. IBM
    3. The New Wave
4. CI Clusters
    1. Component Descriptions
    2. Possible Extensions
    3. The Future
5. Related Products
    1. Local Area Networks (NI)
    2. Personal Computers
    3. Data Base Management
    4. VAX Information Architecture
6. Impact of New Technology
    1. VLSI
    2. The Server Architecture
    3. Communications
7. The Issues
8. The Challenge

What is a Cluster?

The purpose of this paper is to clarify the concept of Clusters, accurately convey what we are building and the implied benefits, and to highlight the strategic importance of the technologies and methodologies associated with the Clustered System Program.

A CLUSTER is a group of cooperating COMPUTERS connected through a HIGH SPEED bus or link. CLUSTERS normally provide two important benefits as opposed to single computer systems. The first benefit is survivability. The second is modular expandability. Some of the identifying characteristics of a CLUSTER are:

1. The COMPUTERS are usually independent. Each COMPUTER has its own memory and its own copy of the operating system. Failure of one of the COMPUTERS in a CLUSTER should not affect the others. This topology is generally referred to as LOOSELY COUPLED MULTIPROCESSING.

2. The CLUSTER is utilized in much the same fashion as a single COMPUTER as far as the users are concerned. Rarely are the COMPUTERS in a CLUSTER expected to be secured or protected from one another. For this reason, COMPUTERS in a CLUSTER are usually located in close proximity to one another. Operational responsibility for the entire CLUSTER usually falls within a single organization.

3. For the short term, the other reason why a CLUSTER has a restricted radius is the need for a HIGH SPEED bus or link. Perhaps, in the future, communications technology will deliver HIGH SPEED, long distance links. HIGH SPEED can probably be defined as no less than 1 megabyte/second.

4. The COMPUTERS in a CLUSTER usually share a common file system or data base.

5. Because of the shared file system, users of a CLUSTER usually do not have a preference for which COMPUTER they connect to. Therefore, flexible communications switches or patches are usually used in conjunction with CLUSTERS.

6. The CLUSTER can be viewed as a single, larger, more dependable, more functional system than any of the component COMPUTERS. The shared resources and communications switching make the CLUSTER appear this way.

7. While we normally view clustering as being applied to larger computers (minicomputers are considered large these days), the same concepts can be applied to smaller computers, such as PERSONAL COMPUTERS. When PERSONAL COMPUTERS are clustered, no communications switches or patches are necessary, since the computer is also the terminal. Instead, Local Area Networks provide connectivity to all desired resources.

The Market

When we view the market for clustered systems, there is always a temptation to equate it with the so-called 'HIGH AVAILABILITY' market. Let us define the market and market size for clustered systems by breaking it into 3 segments. They are:

1. The High Availability Market
2. The General Purpose Computing Market
3. The Emerging Market

A sketch of each market and estimated market size follows.

1. The High Availability Market

   This is a small market segment whose main identifying characteristic is the need to maintain operation nearly 100% of the time and to recover from any failure within about a 1 second timeframe. Example applications in the High Availability Market are:

   1. Nuclear Power Monitoring and Control
   2. Air Traffic Control
   3. Space Flight
   4. Some Military Defense Applications

   These applications are characterized by the following:

   1. Need for totally redundant hardware, with little concern for cost. This hardware is required to have failover times of less than 1 second in most cases and 'milliseconds' in some cases.
   2. Need to support special process interfaces.
   3. High liability if the 'failsafe' system fails.

   Approximate market size: Less than 1% of all Data Processing Revenues.

   Digital has a policy about such applications. We do not bid for them.

2. The General Purpose Computing Market

   When most applications were automated for the first time, there was usually a manual backup system which could be invoked in the event of a failure. The main motivation for automating the application the first time was cost savings. Therefore, failures were a nuisance, but did not necessarily jeopardize the business. Payroll is an application with these characteristics.

Today, many applications make significant contributions to the profitability of a business. In some cases the computer applications are the reason the firm can compete. If the applications go down, the business is directly affected. A cash management/funds transfer system at a bank is an application with these characteristics.

As applications go through their second generation of automation, there is rarely a manual fallback system.

Reliable computing is becoming a prerequisite for the sale of computer systems. This does not imply that more applications are becoming 'HIGH AVAILABILITY' applications as defined in the previous section, but rather that all applications need some level of clean recovery.

We might call this facility 'PREDICTABLE RECOVERY'. For most general purpose applications, a small amount of downtime is not critical. What is critical is the ability to ensure that there be no loss of data, no corrupted data, and some facility to restore the computing resource in a timeframe selected by the user.

Just as important as reliable operation is the ability to easily expand the capacity of the system. This is one of the messages which we have always used for DDP, the Clustered System Architecture magnifies this message.

The requirements for reliable computing in the general purpose computing market are as follows:

1. Total Data Integrity. Protection against hardware destruction and software pollution of data.

2. Provisions for providing failover of hardware components. These facilities can be manual or automated. The customer must be given the choice of automated failover however.

3. Comprehensive services.

4. The ability to accomodate growth of the application without jeopardizing the user's current investment.

Approximate Market Size: 80% of all Data Processing Revenues.

The Clustered System Approach provides basic building blocks for both reliability and expansion. This manifestation of DDP could be considered a complete alternative to mainframe processing.

3. The Emerging Market

Over the last two years the computer industry has gone through an upheaval. Office Automation has become the newest and biggest buzzword in the industry. Personal computers have earned respect and have been 'blessed' by the two largest computer companies.

Office Automation and the increasing emphasis on personal computers have highlighted the need for modularity in computer systems. Local Area Network technology has given us hope that we can put together modular systems with personal computers in the Office environment and in other applicable environments.

The Emerging Market will attempt to utilize new technology to increase the overall productivity of their business entities, especially the office environment. The reason that the market must be labelled 'Emerging' is that there is no set definition for it yet. Every vendor entering this arena has their own definition concerning which technologies are key (usually emphasizing the technologies they have available at the time).

A few common threads run through the various definitions of the Emerging Market however. The technologies that seem to be required are:

1. Personal Computers

2. Local Area Networks

3. General Purpose Computers

4. Flexible Communications

5. Data Management (all data types, managed across networks)

Any vendor meeting all of the above requirements should fare well. However, the requirement for system integration is crucial in this market segment.

The successful vendor will supply the components listed above and also serve as a systems integrator. The biggest opportunities for startup firms in the computer industry today lie in the area of system integration. These 'Systems Houses' can pick and choose the best hardware and software available and add value by making it work together predictably.

The definition of computer system is changing. No longer can we measure only MIPS and I/O bandwidth to accurately project system performance. The new 'system' is not self contained. It uses personal computers, servers, various interconnects, and traditional processors (and clusters of processors).

The successful vendor in this market will be able to accurately define their new 'system' and also provide:

1. Hardware and software as listed above.

2. Accurate performance characterizations for their 'system' (integrated with the gear of others perhaps?). 'System' performance should be predictable. 'System' performance should be easy to monitor and tune.

3. Services oriented towards tuning their hardware and software 'system' (considering the gear of others perhaps?) towards the application goals of the user.

4. Ongoing support and maintenance of the new 'system'. Operation of the 'system' defined by personal computers and local area networks must not be perceived to be more complex than operating a simple time sharing system. Response times must be comparable with that of timesharing systems.

5. Consistent interconnects which allow new processing units and servers to be integrated into the 'system' while protecting the user's current investment.

Approximate Market Size: ?

Digital is going hard after this market.

The Competition

For this disussion we will limit the review of competition to the following:

1. Tandem
2. IBM
3. The New Wave (other companies with interesting potential, markets, or technology)

Further analysis will be made available at a later time.

1. Tandem Computers

   Tandem Computers was founded in 1974 to provide multi-computer systems (CLUSTERS) oriented towards transaction processing applications with critical uptime requirements. Tandem Computers will sell approximately $350 Million of such systems this year under the trademark 'NonStop'.

   Despite the name 'NonStop', Tandem is not marketing in the 'HIGH AVAILABILITY' market outlined earlier in this document. They have instead concentrated on the general purpose transaction processing market. Tandem has begun to do some repositioning into the newly emerging office market.

   They are positioning office automation as a natural extension of Transaction Processing (which has been their forte). By adding compatible support of new data types through their newly announced 'TRANSFER, TRANSFER/MAIL, and TRANSFER/FAX' software they are broadening their scope of applicability. Tandem announced intentions to pursue high speed/low cost transmission of data via satellite through a joint venture with American Satellite Corporation with a product called 'INFOSAT'.

   Tandem has created some new issues for the competition to address. They played this game with 'NonStop', positioning highly reliable operation as a primary requirement whether it was or not. Similarly, I expect them to make integration of all data types with satellite transmission, using distributed data base software, their new wedge into the account. Tandem believes that these new products will keep the competition on the defensive, while placing themselves in the Office Automation game.

   'NonStop' was (and will remain) a key buzzword in Tandem competitive situtations. 'Distributed Data Base', 'Multiple Data Types', and 'Satellite Transmission' will become the new buzzwords in Tandem competitive situations.

   Tandem believes that their Distributed Data Base is their biggest point of differentiation today.

Other key points derived from a presentation made by Tandem's president:

1. They view their competition as almost exclusively IBM (they explicitely stated this). Tandem states that IBM has a 'strike force' to compete with them. They also state that IBM will withdraw a bid rather than lose the decision.

2. They believe that IBM has traditionaly gained account control by controlling the central DP facility with centralized DP, centralized data bases, and hierarchical networks such as SNA.

3. Tandem wishes to garner account control through NonStop Distributed Data Processing, with Distributed Data Bases (they believe that their relational, distributed data base capability is the cornerstone of their entire system), and more flexible networking architecture.

4. Tandem perceives themselves as an 'End User' oriented company. They are committed to supplying very high levels of support (given that IBM is their competitive target, they wish to have a similar image). Tandem believes that they should be considered a 'mainframe vendor' which provides tools more in step with todays data processing needs.

5. They believe that their products have evolved as follows:

   NonStop Capabilities (including Data Integrity)

   extending to:

   Networks (including X.25, LANS, Gateways, Satellite)

   extending to:

   Distributed Data Base capabilities (including Data Integrity)

   extending to:

   Transaction Processing (layered on Distributed Data Base)

   extending data types to:

   Image (Facsimile, Xerox, Graphics, Video)

   Voice (Digitized)

   Text

   Binary

6. Tandem's target market segments are:

1. Large Banks and Financial institutions.
2. Large Manufacturing Companies
3. Travel
4. Transportation
5. Airlines
6. Telecommunications

2. IBM

IBM realizes the importance of the general purpose computing market. They realize that the CLUSTER approach to computing provides some very real benefits, particularly in the area of reliable operation. Because of these realizations, IBM has intimated that they are working on failsafe architectural extensions for one of their mainstream product families.

They perceive, and rightly so, that all vendors will have to improve the reliability of their general purpose computing products. IBM has not indicated any large interest in the 'HIGH AVAILABILITY' market outlined earlier in this document.

IBM has expressed great interest in the Emerging market defined by Office Automation and Personal Computers. Where IBM's hardware and software may not be up to standards at this time, they can be expected to improve.

The far more urgent threat from IBM is their potential to do very well in the systems integration part of the game. Having always been a service oriented company, providing a security blanket, IBM will invest heavily in being able to characterize the performance of their gear. They will also offer comprehensive services to help ensure that the expectations of the user are met (whether or not that means re-setting the user's expectations).

Digital is in for a major battle with IBM in the emerging markets. We have never been on more of a collision course with IBM in our history. For this reason, we must be prepared to invest heavily in the systems analysis, characterization, and service aspects.

3. The New Wave

1. Stratus

Stratus is a small startup company which has targeted the general purpose transaction processing market. They are using a different architectural approach than Tandem however. Stratus is relying almost exclusively on hardware redundancy to provide continuous processing.

In contrast to Tandem, who seem inclined to use custom logic for their processing engines, Stratus is using standard microprocessors in totally redundant configurations (they are using the Motorola 68000).

Because of the low cost of the microprocessors, Stratus sees fit to place two microprocessors in each processing unit with comparators to check for consistent results. In the event of an inconsistent result, failover takes place to a totally separate processing unit which also has dual microprocessors.

By using this approach, Stratus claims that invalid results will never get through the system. Thus they perceive no need for recovery software of any kind. Their's is a totally hardware oriented approach.

Contrast with Tandem or Digital's future offerings, which are combination hardware and software approaches.

This approach has great marketing appeal. It is easy for the customer to understand and easy to contrast with more complex approaches. However, the chip level redundancy should be viewed simply as an alternative way to implement error detection on a board. Our boards might be as reliable as theirs, but we have a more difficult time explaining how we do error detection. This is unfortunate, because our error detection is probably more comprehensive (since they only check on microprocessor failure).

It will be interesting to see how well Stratus' approach is received in the market place. Also how well the hardware only solution provides continuous processing. The approach warrants watching.

2. August Systems

August Systems is a small vendor targeting the 'HIGH AVAILABILITY' type of application outlined earlier in this document. The 'real time' nature of these applications differentiate them from the more 'data processing' oriented applications which Tandem, Digital, and IBM are targeting.

August Systems "Can't Fail" system uses triple-redundant microprocessor based logic, triplex process interfaces, and peripherals that can be triplicated depending on applications needs.

August Systems is mentioned here so that we can watch how well they perform in very high risk application segments.

3. Intel

Stratus is using the the M68000 microprocessor and is building a multiprocessor architecture around it.

Intel is in the process of introducing their 432 microprocessor into the marketplace. Its first point of differentiation is that has a very high level, object oriented instruction set.

Its second point of differentiation is that they have built the multiprocessor capabilities into the architecture.

Intel claims that the higher level, object oriented system will reduce the incidence of software failure. They also claim that the multiprocessor architecture will provide hardware fault tolerance.

They do have an Achilles Heel in the approach however. Their multiprocessor architecture uses shared memory with no provisions for memory subsystem failure. They do not provide automated methods to recover from component failures. They simply provide the right hooks to have many processors executing from a common bank of memory.

Where the high level, object oriented system might reduce the incidence of applications software failure, it will be interesting to see if implementing high level functions in logic and microcode proves less susceptable to system 'software' failures than implementing these functions in the operating system. Powerful tools are available for debugging operating system code today. Comparable tools are not yet available for debugging microcode and logic.

The Intel 432 architecture is worth watching. It does not pose an immediate threat in the reliable computing space. It has more potential to provide a wide performance range of high level, object oriented processing engines. This range of processing engines could be put together using Clustering techniques to provide highly reliable computing systems. This fits with Intel's strategy to market the 432 almost exclusively through OEM channels.

CI Clusters

In Q1FY84 Digital will offer a Cluster Architecture based on the CI. This section will summarize the key components of the architecture.

1. Component Descriptions

    1. Computer Interconnect (CI)

        The CI is a high speed (70 megabits/second), multidropped, short distance (90 meter radius) interconnect designed to pass data and control information among intelligent computers.

        The computers currently supporting the CI are as follows:

        1. VAX-11/780
        2. VAX-11/782
        3. VAX-11/750
        4. 2060
        5. 2080 (JUPITER)
        6. VENUS
        7. HSC-50 (I/O server)

        The CI port interfaces themselves are intelligent. The CI port interfaces have been designed to utilize the page tables and virtual addresses of the supported VAX systems, thus making bulk data transfers very efficient. Reliable transmission is guaranteed by protocols implemented in the port.

        A dual path facility has been built into the CI port architecture to provide for redundancy. Under normal operating conditions, the dual path facility can provide enhanced performance.

    2. I/O Server (HSC-50)

        The HSC-50 is an intelligent mass storage subsystem. When integrated into a CI Cluster, the HSC-50 is utilized as a common I/O Server for all host computers residing within the CI Cluster.

        Each HSC-50 is counted as a node in a CI Cluster. The HSC-50 is a computer, one which has been optimized towards managing the flow of information between large mass storage devices and one or more host computers.

        The HSC-50 relieves the host software of the burden of performance optimization, disk personality, and error recovery. The HSC-50 always presents 'logically perfect' volumes to the

host computers. For some applications, the I/O Server will also maintain shadowed copies of selected disk volumes. When both shadow volumes are online, a performance benefit can be expected since the system will access data from both volumes. An additional access arm to the data is available.

Plans are in place to provide utilities to perform volume backups from disk to tape without host intervention.

A bank of volumes can be dual ported between a pair of HSC-50s. The HSC-50s can share the I/O processing load (static dual porting only). If one HSC-50 in the pair should fail, the surviving unit can automatically restore service for volumes previously owned by the failed HSC-50. This failover takes place without loss of outstanding I/O requests.

The HSC-50 is a special purpose computer optimized for servicing I/O from large mass storage devices. General purpose computers with traditional mass storage interfaces are inherently less efficient at this task.

3. System Communications Architecture (SCA)

The SCA is a layer of software which implements the equivalent of network functionality between computers within a Cluster. To understand the difference between SCA and DNA we must first study the major differences between a Network and a Cluster.

1. Networks are usually geographically dispersed (although this is not necessary).

   The computers within a Cluster are usually co-located within the same facility. This is true since the primary motivations behind implementing Clusters are to provide larger capacity computer systems and to provide redundancy within a computer system.

2. Nodes within a Network are usually controlled and operated by several different organizations within a business entity.

   Computers within a Cluster are usually controlled and operated by the same organization within a business entity.

3. To access a Data Base on a remote node within a network, the requesting node must be given positive authorization by the serving node. It has been said that nodes within a network are 'mutually suspicious'.

   Data Bases and other resources are considered to be shared equally among all computers within a Cluster. When an additional computer is added to a Cluster, it is considered to be equal partner sharing all resources with the other computers in the Cluster. The computers in a Cluster are 'mutually benevolent'.

4. With current communications technology, routing techniques provide more flexible and lower cost network topologies.

   Communications technologies used to implement Clusters allow for full multidrop topologies. There is no need for routing functionality within Clusters.

5. Communications across networks primarily takes place between cooperating application processes. The system utilizes the network primarily to provide resource sharing functions.

   Communications within a Cluster takes place primarily between the member computer systems. Large data transfers between hosts and servers and resource contention control messages between cooperating hosts comprise the bulk of the traffic across the Cluster link. Of lesser magnitude are messages between cooperating processes on separate computers within the cluster.

The System Communications Architecture (SCA) was developed to provide for highly efficient data flow between computers within a Cluster. The SCA provides the backbone transport mechanism for all other cluster software. The efficiency provided by SCA is necessary to transform a group of independent computers into a cluster.

4. Mass Storage Control Protocol (MSCP)

A MSCP has been devised to allow for flexible connection of new mass storage devices to computer systems. With the advent of intelligent disk controllers such as the UDA and the HSC-50, it is now possible to implement disk drivers which can be insensitive to changes in the characteristics of the drives themselves, and also insensitive to changes in the transport mechanism from drive to computer memory.

These new drivers are called Class Drivers. These Class Drivers implement the 'master' side of the MSCP. The intelligent controllers implement the 'server' or 'slave' side of the MSCP.

Operating systems can support new disk technology in a more timely fashion by using this class driver scheme. In addition, new transport mechanisms, such as the CI architecture, can be more easily leveraged.

If a new interconnect is introduced it is now possible to support by simply writing a port driver interface to the new interconnect. If a new controller is introduced it is now supportable by simply writing the 'slave' side of the MSCP in the new controller.

Because every 'master' request must be positively acknowedged by a 'slave' when using MSCP, it is possible to cleanly implement device and controller failover in the system. I/O requests are never lost and can be retried in the reconfigured system in the event of a failure.

The 'slave' or 'server' side of the MSCP has also been implemented on VAX/VMS. This allows current VAX systems with local mass storage to be cleanly integrated into CI Clusters with no loss of user investment. Each VAX system with local mass storage can act as a server thus making its storage transparently available to other VAX systems within the Cluster.

5. Distributed Lock Manager

The Distributed Lock Manager allows VAX/VMS to implement a true shared file system across a Cluster. The Lock Manager is resident on each VAX system within the cluster.

An application process wishing to access a particular record within the shared data base of the Cluster makes a record lock request to the Distributed Lock Manager. Once the lock request has been granted, no other application process on that computer or any other computer within the cluster can secure a lock on that particular record.

The distributed implementation of the lock manager (and the disk ACPs) ensures that there is no single resource allocation bottleneck within the Cluster. Tradeoffs have been made in the lock manager to optimize for normal operations rather than failure recovery. The minimum amount of interprocessor information is passed during normal operations. Enough information is passed to allow surviving computers in a cluster to derive the lock information of a failed computer. The surviving computers can then release locks held by applications which were executing on the failed computer.

6. Common Journalling Facility (CJF)

The CJF provides a series of system services which allow any Data Base Management System to create and maintain journals of data base activity.

The CJF facilitates the creation of Before Image Journals which could allow data bases to be 'rolled back' to some known, consistent state.

The CJF facilitates the creation of After Image Journals which could be applied to Backup copies of the data base. This allows for 'roll forward' reconstruction of data bases destroyed by hardware failure or corrupted by software failure.

The CJF also allows for applications to maintain user defined audit trails of data base or other system activity.

The CJF allows for any number of Data Base Managers to share the same journal volumes. These journal volumes are normally magnetic tape, but can also be disk volumes.

The participating Data Base Management Systems are responsible for providing the utilities which apply journalled data in 'roll back' or 'roll forward' recovery situations. These utilities are currently being written for RMS and DBMS.

7. Recovery Units

The Recovery Units facility allow Data Base Management Systems to dynamicly maintain the consistency of their data bases in the face of transaction, or system failure.

The Recovery Units facility provides two simple calls which allow applications to protect themselves from data corruption. The first call, normally invoked when the data base is in a known, consistent state, creates a recovery unit. Once a recovery unit has been created, the system is directed to secure (normally on disk) 'before image' copies of data base records affected by the transaction. The second call, normally invoked when the transaction has been completed and the data base is once again in a consistent state, purges the recovery unit.

If the transaction or system should fail while the recovery unit is open, the system will 'roll back' the effects of the transaction, thus bringing the data base to the consistent state which existed at the beginning of the recovery unit. At that point, the transaction can be retried.

In the case of an application failure or aborted transaction, the VAX system on which the transaction was running will do the 'roll back'. In the case of a system failure, the surviving systems within the Cluster will 'roll back' all active recovery units opened by the failed system.

8. Checkpointing Facility

A Checkpointing Facility is being provided to allow applications with a critical investment in processing to protect that investment. This is useful in two scenarios. The first is the application which runs a single monolithic job for long periods of time (usually hours). This type of application is typical in engineering and simulation applications. The second type is the application which requires that transactions be automatically retried (in the event of a failure) without additional operator interaction. In both cases the implementation is the same.

The application defines a checkpoint, usually at some consistent point in its execution. In the 'retry transaction' case, the checkpoint should be done immediately after all transaction inputs have been received. At the time of the checkpoint, the system secures all altered pages, in the virtual address space of the process, to a checkpoint file. The application can then resume processing for some amount of time. In the event of a failure, which causes the application to abort, the system can reinitiate the application from the point of the last defined checkpoint by reconstructing the state of the process from the checkpoint file. Because the checkpoint was secured to disk, the application could conceivably be brought up on a different processor within the Cluster (given that the new processor is the same type as the original processor - for example 780 -> 780).

The checkpoint facility, in many cases is combined with the Recovery Units facility. In this case, the recovery unit is 'rolled back' before the application is reinitiated from the checkpoint. The system will ensure that Recovery Units and Checkpoints are declared in a logical fashion when they are used in unison.

9. Ulysses Communications Switch

In previous sections the benefits of a Cluster were broken into to major categories. The first being the ability to add incremental processing capacity due, in most part, to the shared data base. The second benefit is the ability to provide 'spare' processing capacity with automatic failover to surviving units.

In order to fully realize the above benefits, there must be a facility to automatically switch terminals and communications lines from one computer within the cluster to another. To meet this requirement the Ulysses communications switch is being used.

There are two major points of differentiation for the Ulysses switch versus other similar switches. The first is the ability to concentrate lines and ports near their points of origin and use single high speed lines to the switch itself. The second point is that the switch is controlled primarily by software resident in the host computers within the cluster. With the flexibility of host control, some crude load leveling can be implemented.

The Ulysses switch can be configured in a fully redundant fashion.

Further detail on the Ulysses switch will be made available at a later time.

2. Possible Extensions

The Cluster Architecture could gain more flexibility by implementing Disk Volume Shadowing on disks directly connected to the host computers (through the UDA). This would allow for Clusters with a lower entry price. Currently Disk Volume Shadowing is available only on the HSC-50 I/O Server, thus requiring inclusion of an HSC-50 in order to provide the highest levels of data integrity.

The ability to perform Disk Volume Shadowing is also a capability useful outside the realm of Clustered Systems. Many single computer applications have stringent requirements for protection of data. In many cases, loss of the computing service is not critical, but loss of data can be a disaster. In these cases, providing Disk Volume Shadowing exclusive of the Cluster Architecture and the HSC-50 I/O Server would be desirable.

3. The Future

The CI Cluster Architecture has supplied Digital with some very significant benefits. They are:

1. The ability to provide survivable systems through extensions to the mainstream VAX family.

2. A hedge for high end systems. Except in the case where a very powerful compute engine is needed for monolithic compute jobs, the Cluster Architecture will allow us to cleanly increase the capacity of multi-user systems while protecting the user's current investment. This allows us to 'do the right thing' with technology at the high end by relieving the pressure to rush the next high end engine out the door.

3. The software technologies employed have solved some of the crucial problems associated with distributed data bases.

Because of the above general benefits, we should commit to providing similar capabilities on future members of the VAX family (and beyond?).

In the near future, this means initiating projects to provide 'CI like' capabilities on SCORPIO and NAUTILUS.

Related Products

The technologies explored and implemented in the CI Cluster Program are significant. The following programs have potential to enhance the Cluster Architecture. These programs can also leverage the experience we have gained in the design and implementation of CI Clusters.

1. Local Area Networks (NI)

   It is possible to provide lower cost (perhaps less functional) clusters by substituting CI with NI. We should be able to properly characterize the potential of this approach.

   As we move towards more modular systems in the future it will be difficult to protect our user's investment with two similar interconnects. Communications technology might allow us to produce a Local Area Network interconnect which approaches the speed of the CI. If this occurs, will there be a need for two separate interconnects with different sets of servers for both?

2. Personal Computers

   There is a requirement (if we intend to build 'systems' defined by personal computers and LANs) to provide for transparent data base access between Personal Computers, Servers, Networks, and general purpose computers.

   It would be desirable to extend the Mass Storage Control Protocol, Distributed Lock Manager, and related data base software, developed as part of the CI Cluster project to our Personal Computer Clusters. This may not be feasible until a 32 bit engine is available for our personal computers. Could we limit the scope enough to solve the problem with the 16 bit engine of today? Some of the research and prototype efforts for smart caches, distributed forms, and distributed editors may be applicable here.

3. Data Base Management

   The shared data base produced for CI Clusters has attacked many of the classic problems of the distributed data base. Is communications speed the only gating factor preventing us from having similar shared data bases across Networks? Perhaps the security issues are a major obstacle.

   True DATA BASE MACHINES (as constrasted with I/O and File Servers) should be considered carefully.

4. VAX Information Architecture

   The Common Journalling Facility, Distributed Lock Manager, Recovery Units, and Checkpointing Facility provide a much more solid base for the VAX Information Architecture than exists today. We must ensure that the higher level components of VIA take full advantage of these powerful new capabilities.

In addition, the Application Control Management System (ACMS) and the Transaction Processing Development System (TPDS), formally called TPSS, provide the highest layers of VIA. The high level Application Control facilities provided by ACMS should make Office applications much easier to conceive and implement. In addition there is potential for ACMS to make the Cluster System Architecture more powerful by providing load balancing or job partitioning functions. We should move agressively towards integrating these products cleanly into the Cluster System Architecture.

Impact of New Technology

1. VLSI

The obvious impact of VLSI is that it should provide better price performance in computers, servers, and communications ports. There will also be a trend towards putting higher level functions into the logic of the processors themselves. An example of this is the Intel 432 discussed briefly earlier in this document.

While striving for more reliable systems, the initial concentration has been on making the hardware more reliable. Very little practical work has been done to make software (or logic) more reliable. Hard failures are much easier to recover from than soft (or semi-soft) errors. There are some interesting theories in the area of software fault tolerance however.

It may be possible in the future to implement the higher level logic of a processor several different ways on a chip (since silicon area will not be at a premium). By applying success criterion to the operations, and providing facilities to back out non-successful operations, several different algorithms could be tried. Peter Lee of Digital's Advanced System's Research Group is our resident expert in this area of concern.

2. The Server Architecture

As we move closer to system architectures comprised solely of personal computers and a complement of servers, we must successfully deal with the transition from the traditional 'host computer' architecture of the past. There is a tremendous desire to produce and deliver these new 'systems' today, but our investments in new processors (the ones with active, funded projects today) are considered more as follow ons to our traditional lines of computers. Thus, the transition from traditional computing to the new Server Architecture becomes more difficult since we tend to lock our customers into the follow on traditional computers.

One way to deal with this problem is to ensure that the interconnects for our traditional system Clusters and the interconnects for our Server Architecture 'systems' converge. In this way, the two approaches can more closely complement one another. Additionally, we might start looking at future traditional processors as servers (even if they are considered high performance compute servers).

Servers will expected to be very highly reliable nodes within the new system architecture. Perhaps chip or modular level redundancy (ala the Stratus approach) should be studied for these critical components.

3. Communications

Communiations technology, particularly satellite, may allow us to cleanly migrate our Cluster Architecture functionality to more geographically dispersed topologies (Tandem is heading in this direction). What are the security implications? Encryption of satellite is probably a must.

Issues

The CI Clusters Architecture is a significant milestone in Digital's engineering history. Although the initial design center for the program was the 'HIGH AVIALABILITY' space, the program has provided contributions in other areas. The program has been in place for some time now, and products are due to be delivered in Q1FY84. The following is a list of issues which need to be addressed to ensure that the program is successful in the marketplace and to ensure that we are able to leverage technological advances produced by the CI Cluster Program in other strategic programs within Digital.

1. Product Introduction and Promotion

   It is well understood what products will be delivered in Q1FY84. The introduction and promotion of this program should have very high priority. We cannot miss the opportunity to leverage the technical innovation produced by the CI Cluster Program. The CI Cluster Program and product announcements scheduled for Q2FY83 and Q1FY84 should be treated in a fashion comparable to the Ethernet program announcement and other VAX family announcements of recent years.

   Suggested action: Firm budgets for program and product announcements.

2. System Characterization

   It is vitally important that we be able to properly characterize the performance of CI Clusters. The recent formation of a Systems oriented group within 32 bit engineering is a step in the right direction. The current level of funding for this group will ensure that the VAX-11/780 is thoroughly tested. Some additional maintainability tools will also be produced.

   More committment should be made to testing VAX-11/750 configurations.

   Committments should be made to include measurements of CI Clusters in all applicable performance studies within Digital.

   Suggested action: Additional emphasis on Systems testing within our engineering organization.

3. Services

   In order to ensure a smooth introduction of the CI Cluster Program into the marketplace, our Hardware and Software services organizations must have programs tailored to this new architecture.

   These programs should include:

   1. Remedial Software Support

2. Hardware field service support complementary to that provided on current single system offerings. It must be clarified how programs like the 'Guaranteed Uptime program' will relate to Clusters.

3. Consulting Service offerings to allow customers to take full advantage of both the performance and redundancy benefits of the Cluster Architecture. Clusters will be more difficult to tune than the single computer systems we are selling today. These consulting services must be in place to ensure that the initial customers for CI Clusters are successful.

Service offerings should be the delivery mechanism for the knowledge we acquire in our 'systems group' within engineering.

Suggested action: Continued interaction with the appropriate service organizations.

4. Interconnects

The Cluster Architecture is built around the CI today. This is a good match. It is possible however, that the NI could also provide a reduced level of functionality within the Cluster Architecture. This opportunity should be explored.

We must now start thinking of a second generation of interconnects to provide follow ons to both the CI and NI. If the programs converged it might be easier to provide servers for a wider range of application needs. It might also be easier to protect our user's investment in equipment over time. If the programs do not converge it is possible that we will have a problem similar to the UNIBUS/QBUS on PDP-11s.

Suggested action: Inclusion of NI into the Cluster Architecture. Research into a second generation of interconnects.

5. The new 'system'

It will be necessary to extend functionality now provided only within the context of the CI Cluster Architecture to the new 'system' defined by personal computers and local area networks. How much can we leverage experience gained in the development of the CI Cluster Architecture?

Suggested action: Formation of a new 'systems' group.

6. Communications switching

Although the Ulysses communications switch provides a flexible solution to most switching problems, it doesn't cover all the problems.

There are no plans to failover DECNET links for example.

Lower priced Clustered systems may need less generalized, lower cost communications switching.

Suggested action: Study alternatives to Ulysses for lower priced systems.

The Challenge

Multi-computer systems, including those comprised of Personal Computers, are becoming more prevalent in the marketplace. Increased requirements for reliability are just one of the reasons for this trend. We can no longer concentrate solely on engineering, manufacturing, selling, and servicing single computer systems.

In the future, there will be less differentiation between the hardware and software of various computer vendors. Today's innovation will be a commodity tomorrow. Producing quality hardware and software will always be important, but we can no longer survive in the emerging markets simply by producing the best hardware and software.

The CI Cluster Program provides Digital with opportunity to gain valuable experience in the multi-computer space.

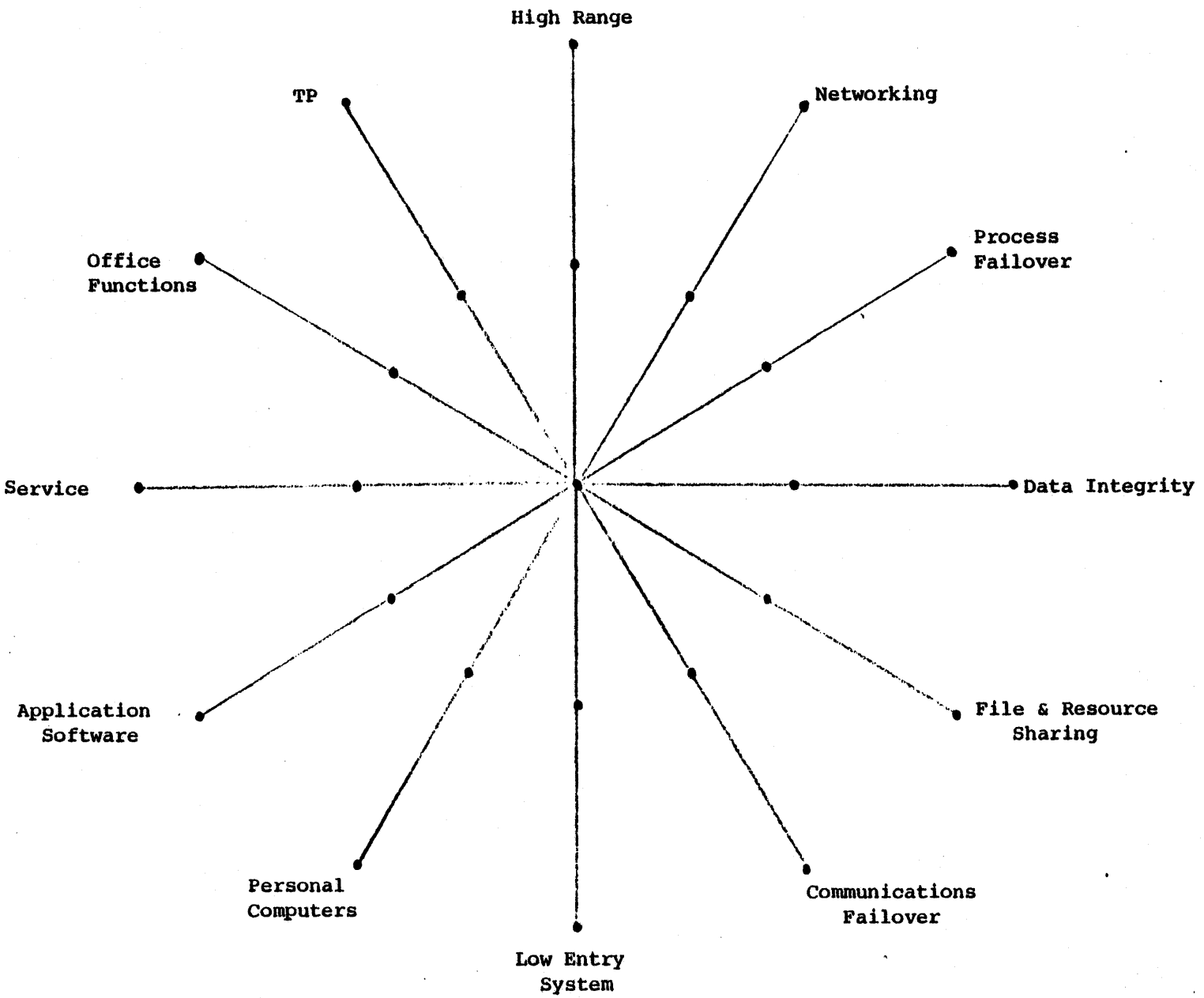
We must set up a structure which allows us to become a leader in the systems integration business. The vendor who is capable of properly characterizing, installing, and servicing their multi-computer systems will likely be the vendor of choice.

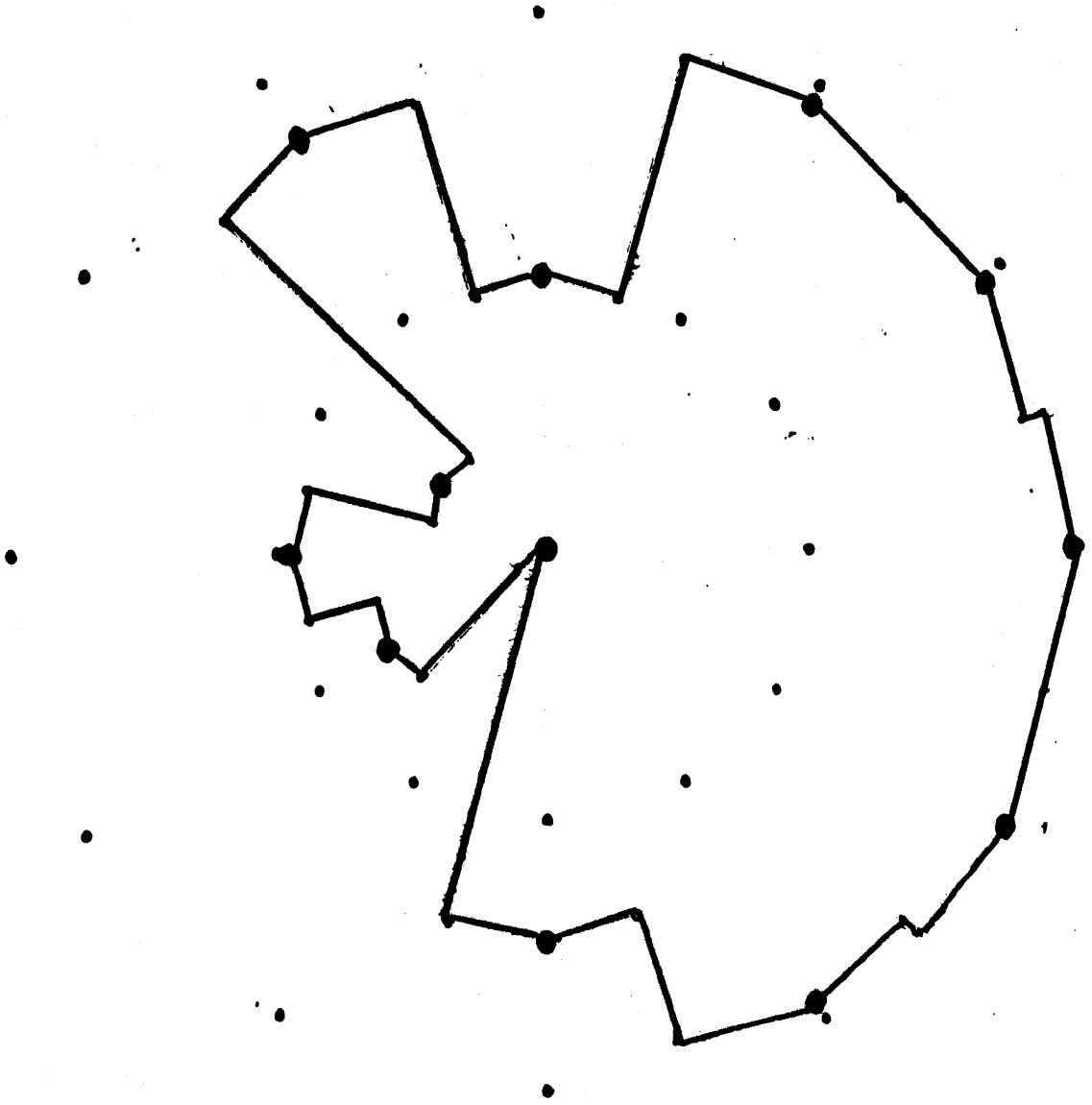
The CI Cluster Architecture provides significant technical innovation for Digital. We should not lose sales based on the functionality of the CI Cluster. Neither should we expect to win sales based solely on the functionality of the CI Cluster. Systems analysis and service capabilities tuned to the CI Cluster hardware and software will be of great importance.

The Challenge over the next year will be to ensure that we set up a structure to allow us fully leverage the Cluster Architecture. CI Clusters will be the first complete multi-computer systems Digital delivers to the market. This will allow us to gain valuable experience in the installation and servicing of multi-computer systems. This expertise will be necessary to survive in the newly emerging computing markets.

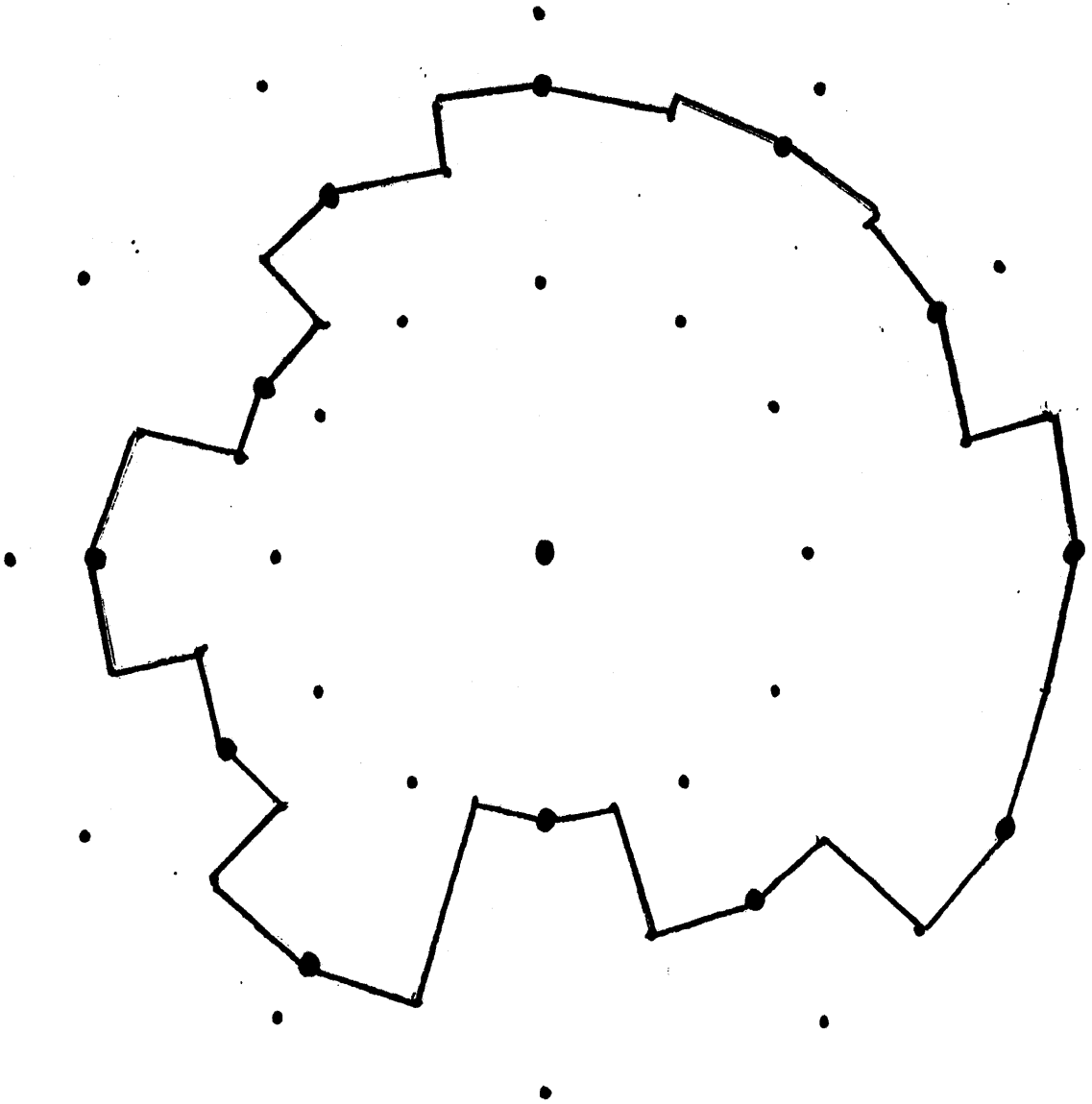
AGENDA

- Market Scenerios / Competitive Analysis
- Software Architecture
- Hardware Architecture
- Lower Price Systems
- Competitive Analysis
- Announcement Plans

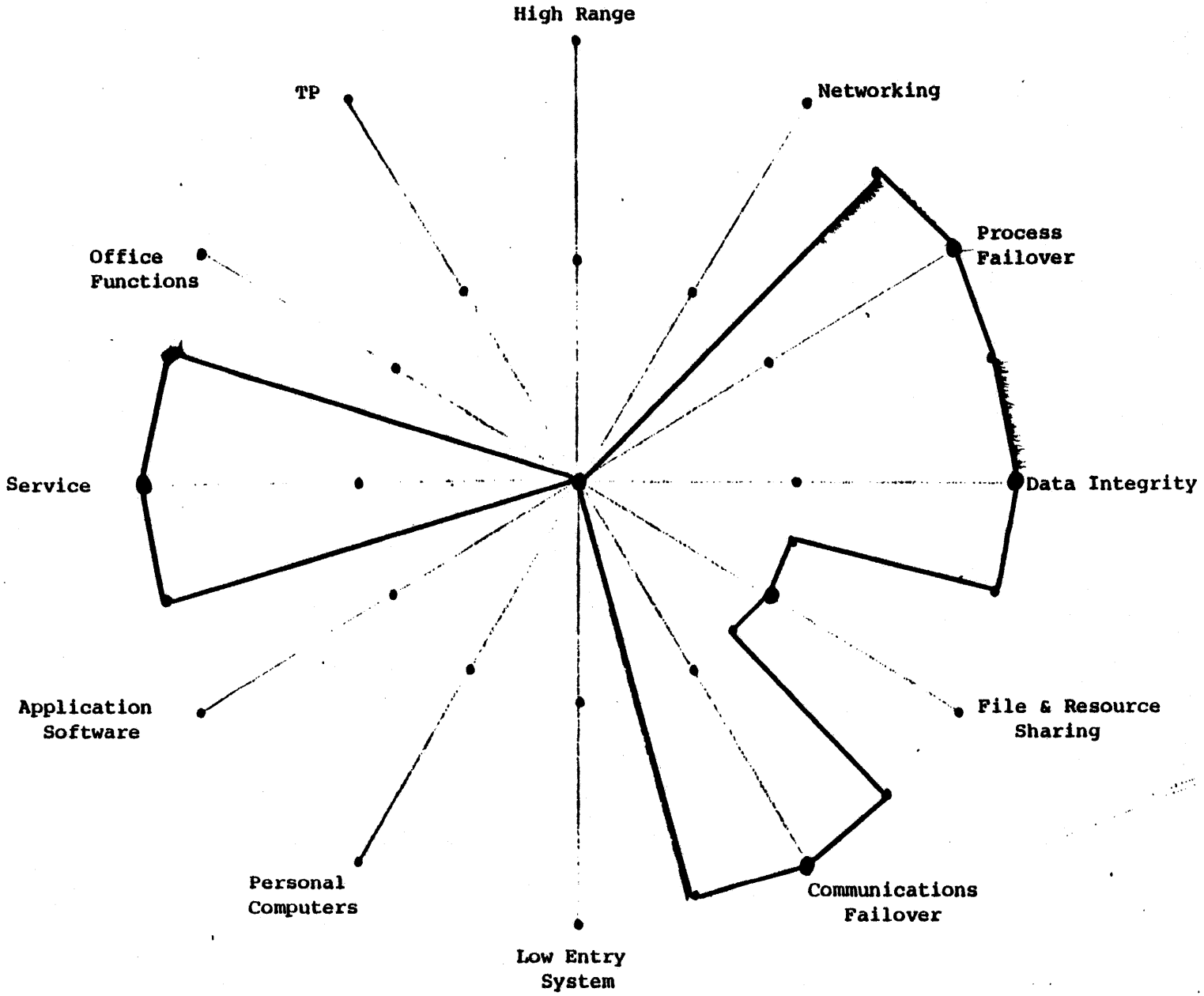

High Range
TP
Networking
Office Functions
Process Failover
Service
Data Integrity
Application Software
File & Resource Sharing
Personal Computers
Communications Failover
Low Entry System



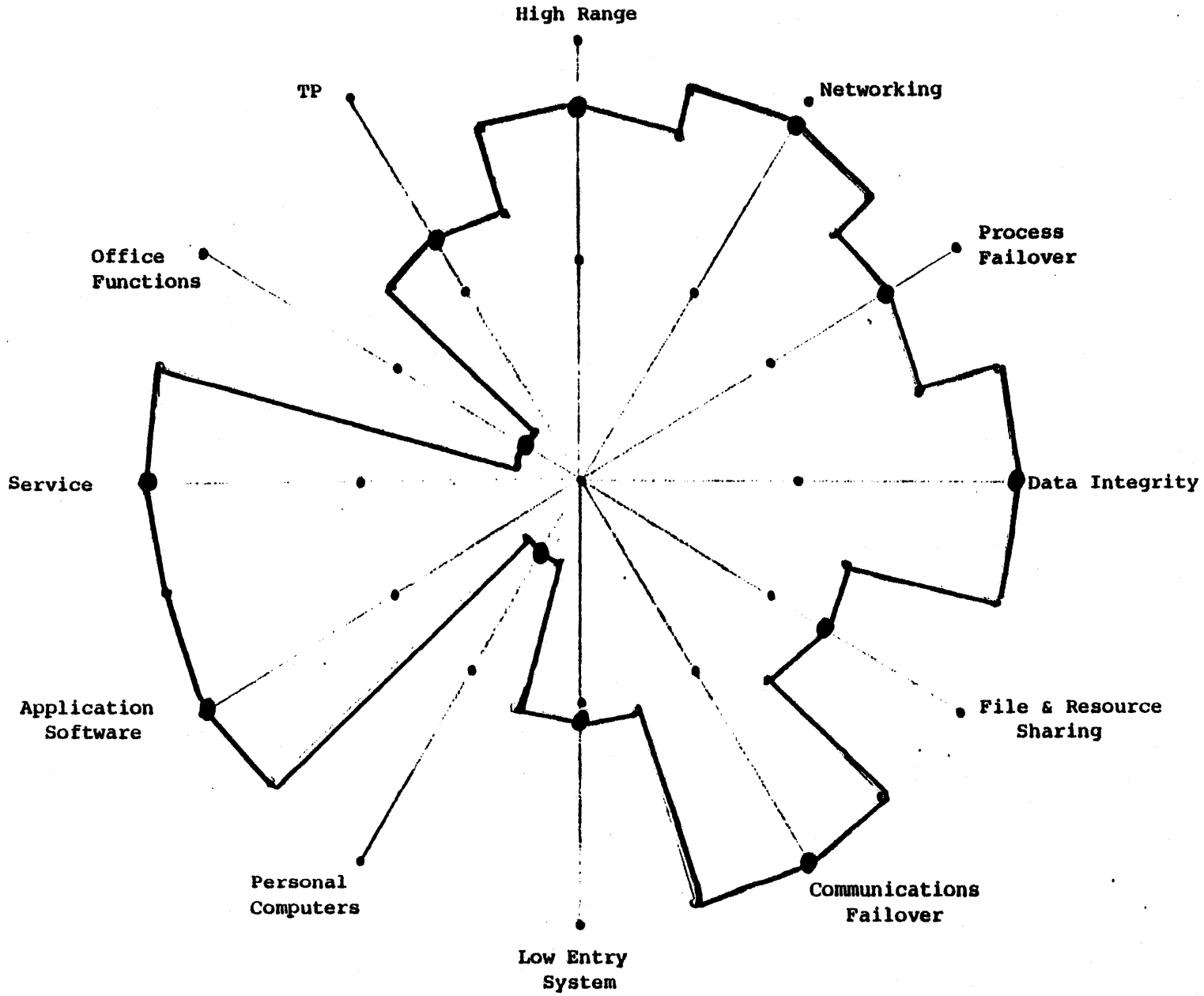
TANDEM CURRENT OFFERING



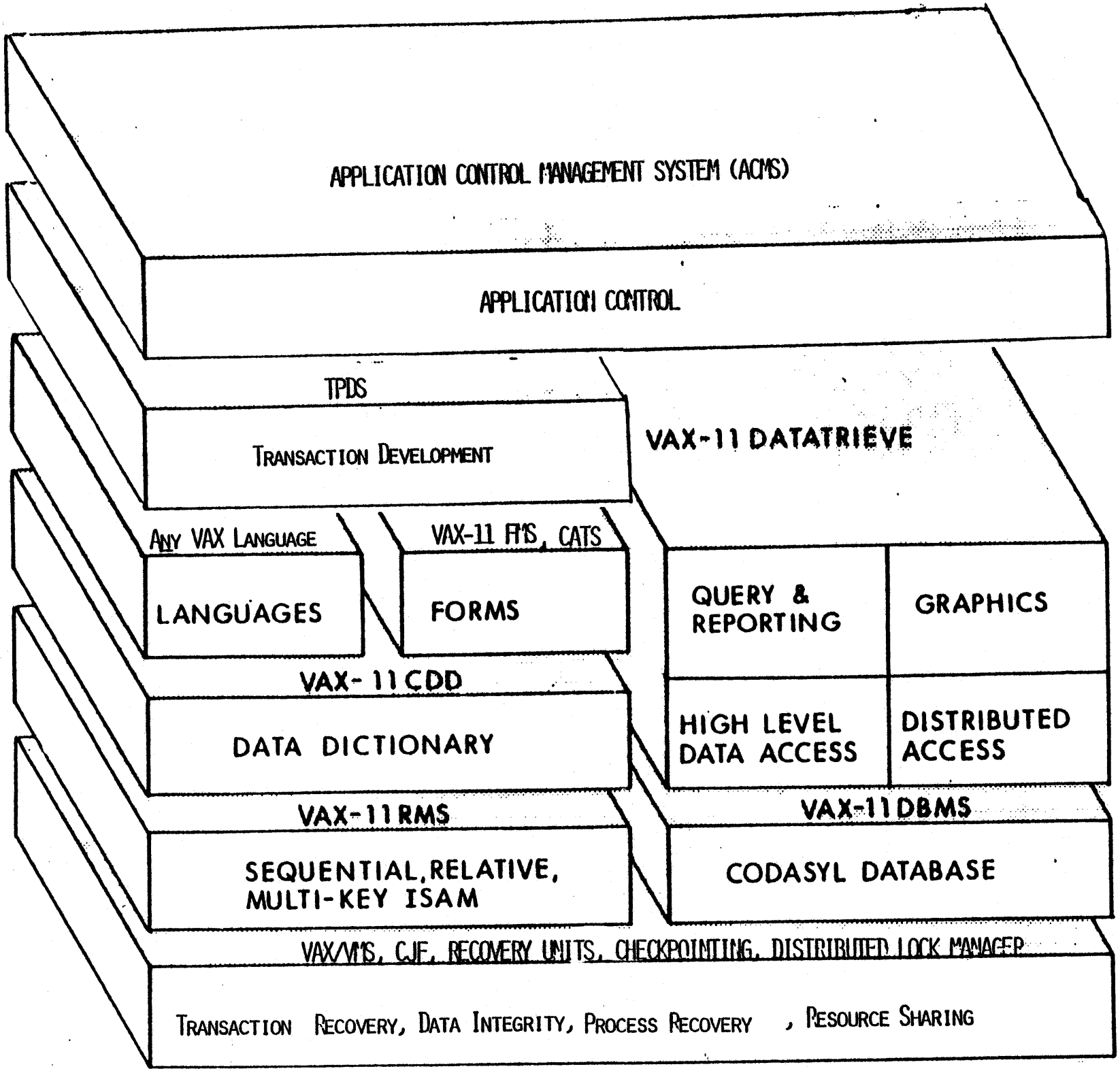
DIGITAL Q1FY84

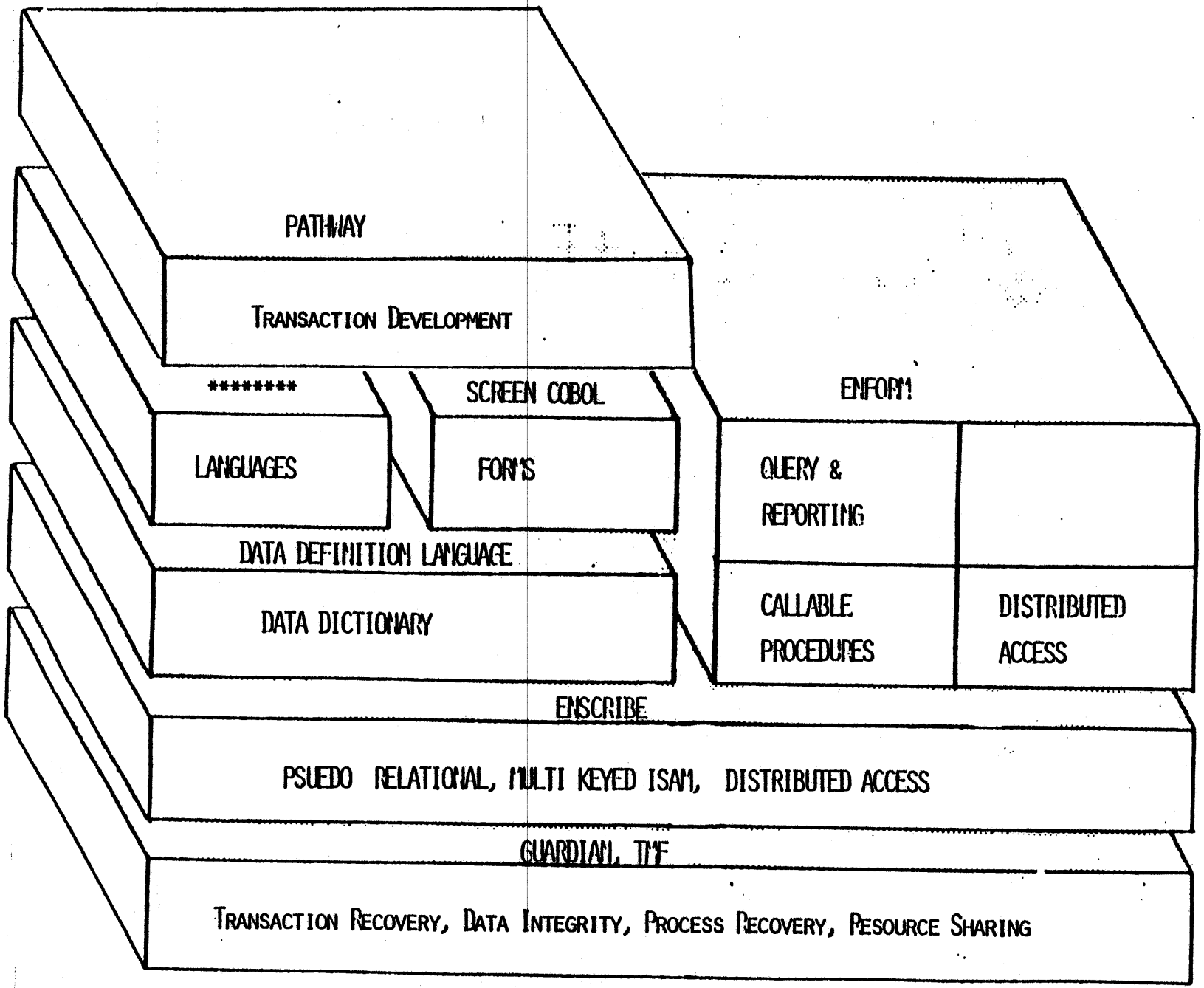


'NonStop' Application



BANKING FUNDS TRANSFER


APPLICATION CONTROL MANAGEMENT SYSTEM (ACMS)
APPLICATION CONTROL
TPDS
TRANSACTION DEVELOPMENT
VAX-11 DATATRIEVE
ANY VAX LANGUAGE
VAX-11 FMS, CATS
LANGUAGES
FORMS
QUERY & REPORTING
GRAPHICS
VAX-11 CDD
DATA DICTIONARY
HIGH LEVEL DATA ACCESS
DISTRIBUTED ACCESS
VAX-11 RMS
VAX-11 DBMS
SEQUENTIAL, RELATIVE, MULTI-KEY ISAM
CODASYL DATABASE
VAX/VMS, CJF, RECOVERY UNITS, CHECKPOINTING, DISTRIBUTED LOCK MANAGER
TRANSACTION RECOVERY, DATA INTEGRITY, PROCESS RECOVERY , RESOURCE SHARING



TANDEM SOFTWARE ARCHITECTURE

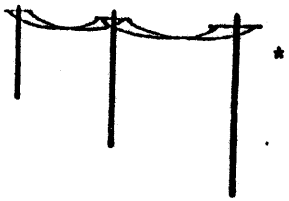
A. Multi-Processor Cluster

. Performance

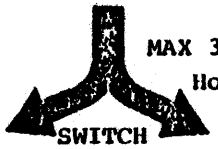
. Availability

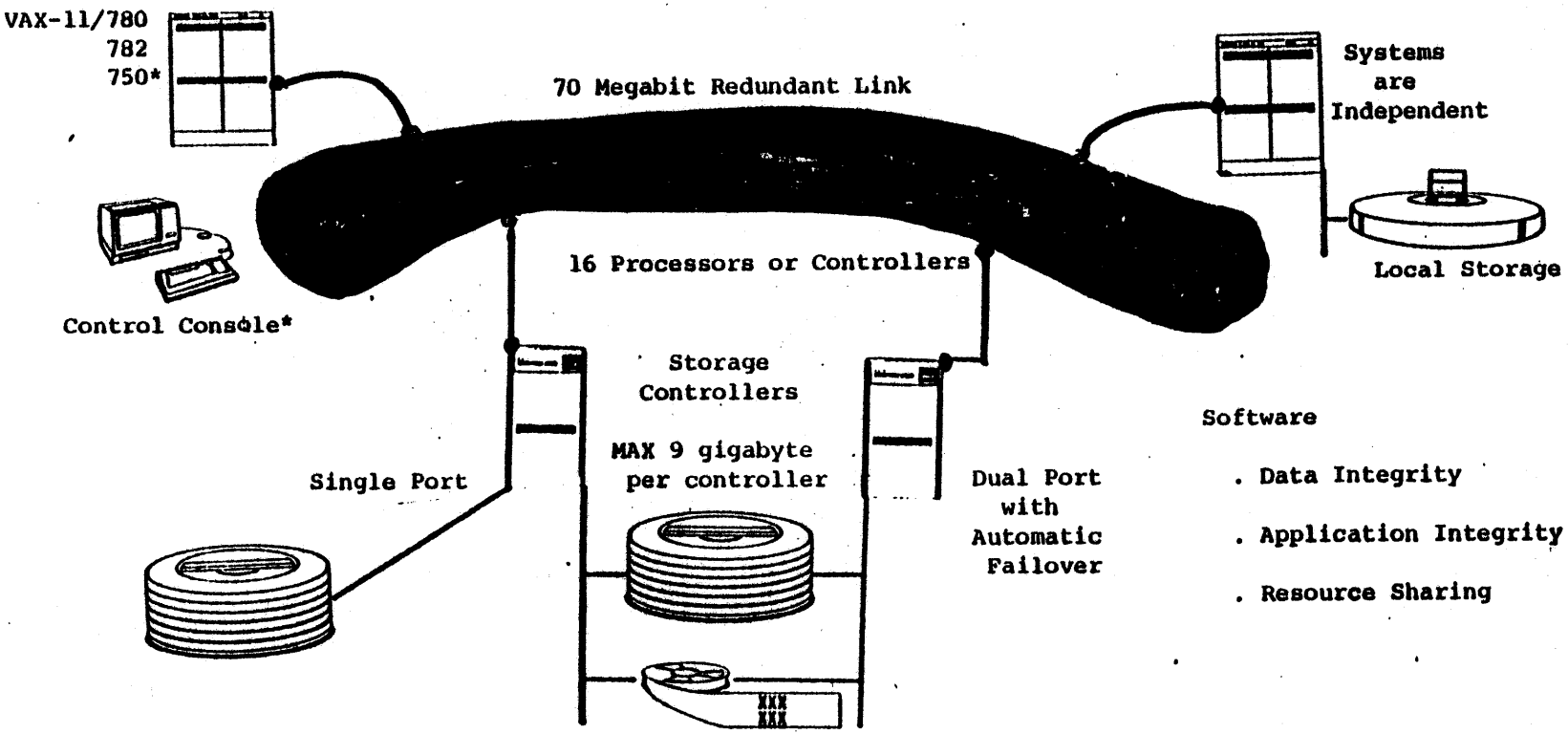
. Resource Sharing

OR



*





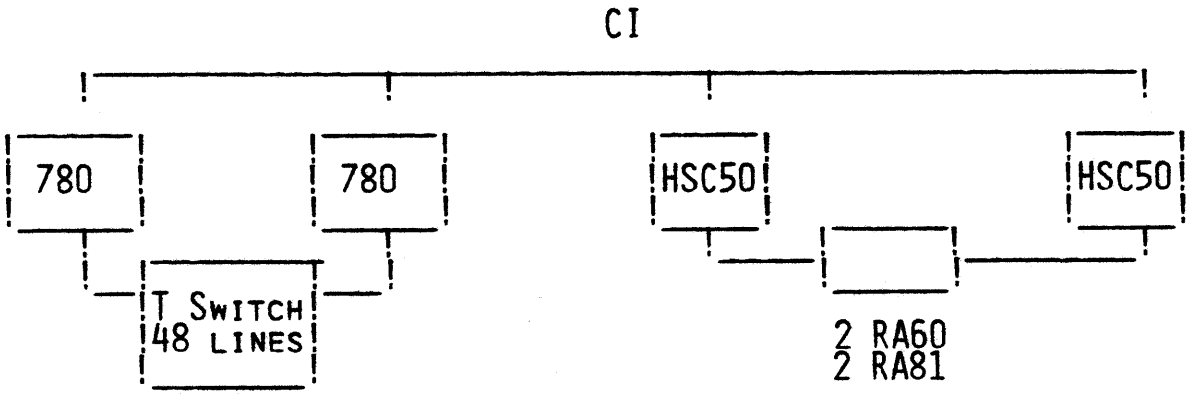
Software

. Data Integrity

. Application Integrity

. Resource Sharing

CI Clusters



"NO SINGLE POINT OF FAILURE"

MLP = $705K

FCS V3B

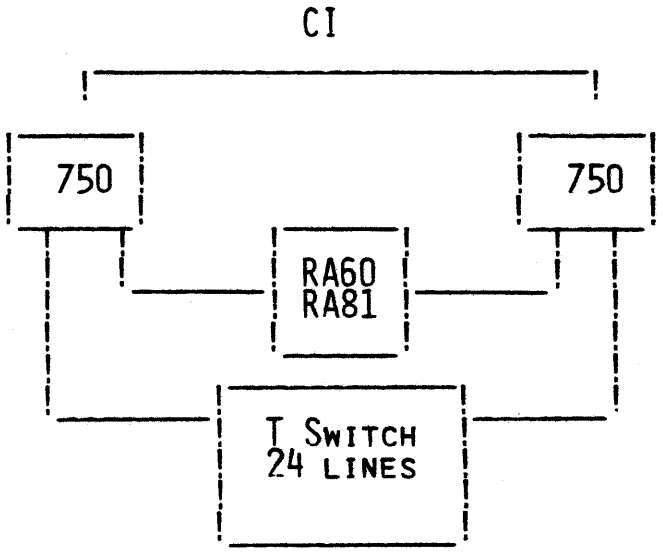
High Availability Systems Engineering

SCHEDULED PROJECTS

- 780 Cluster System Testing
- Maintainability Tools
- System Diagnostic Strategy
- Communications Switching
- "Hot Stage" Initial Ships

Budget: $1.2M

Low End Systems

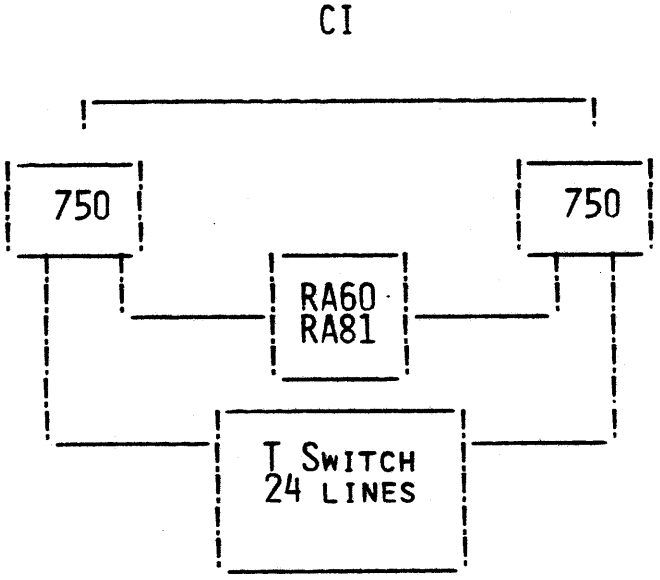


MLP $285K

FCS V3B

Note: Investigate Volume Shadowing with UDA

Low End Systems

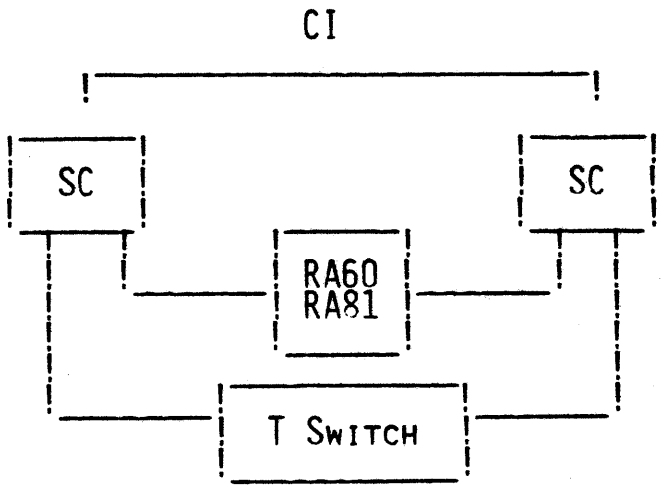


"Door Opener"

MLP $195K

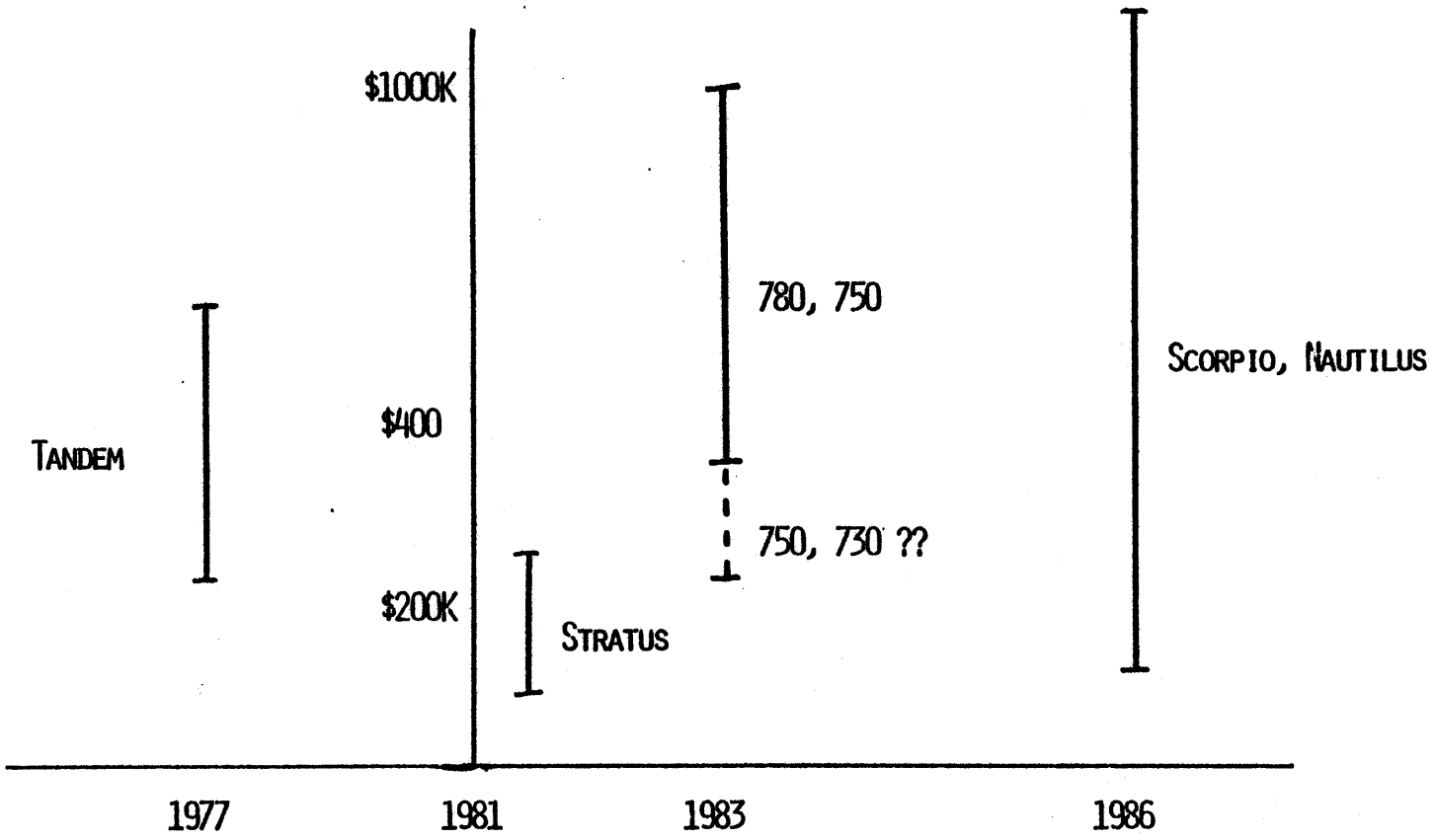
FCS V3B

Low End System Trends



FCS FY86

Incremental Budget (FY83-86) $4.5M


$1000K
$400
$200K
TANDEM
STRATUS
780, 750
750, 730 ??
SCORPIO, NAUTILUS
1977
1981
1983
1986

Revenue Projections

TVG	FY84	FY85	FY86
Low End $150K range	$94M	$123M	$144M
Mid Range $220K Range	$52M	$74M	$87M
High End	$16M	$12M	$12M
	$162M	$209M	$243M

Other PLs - No Detail Estimates between 15% and 40% of total Revenue

Program Announcement Phase I

September 30, 1982

- Announcement Brochure
- Slide Presentation (Architecture)
- Marketing Guide (Sales Update Format)
- Systems Summary (December 30, 1982)

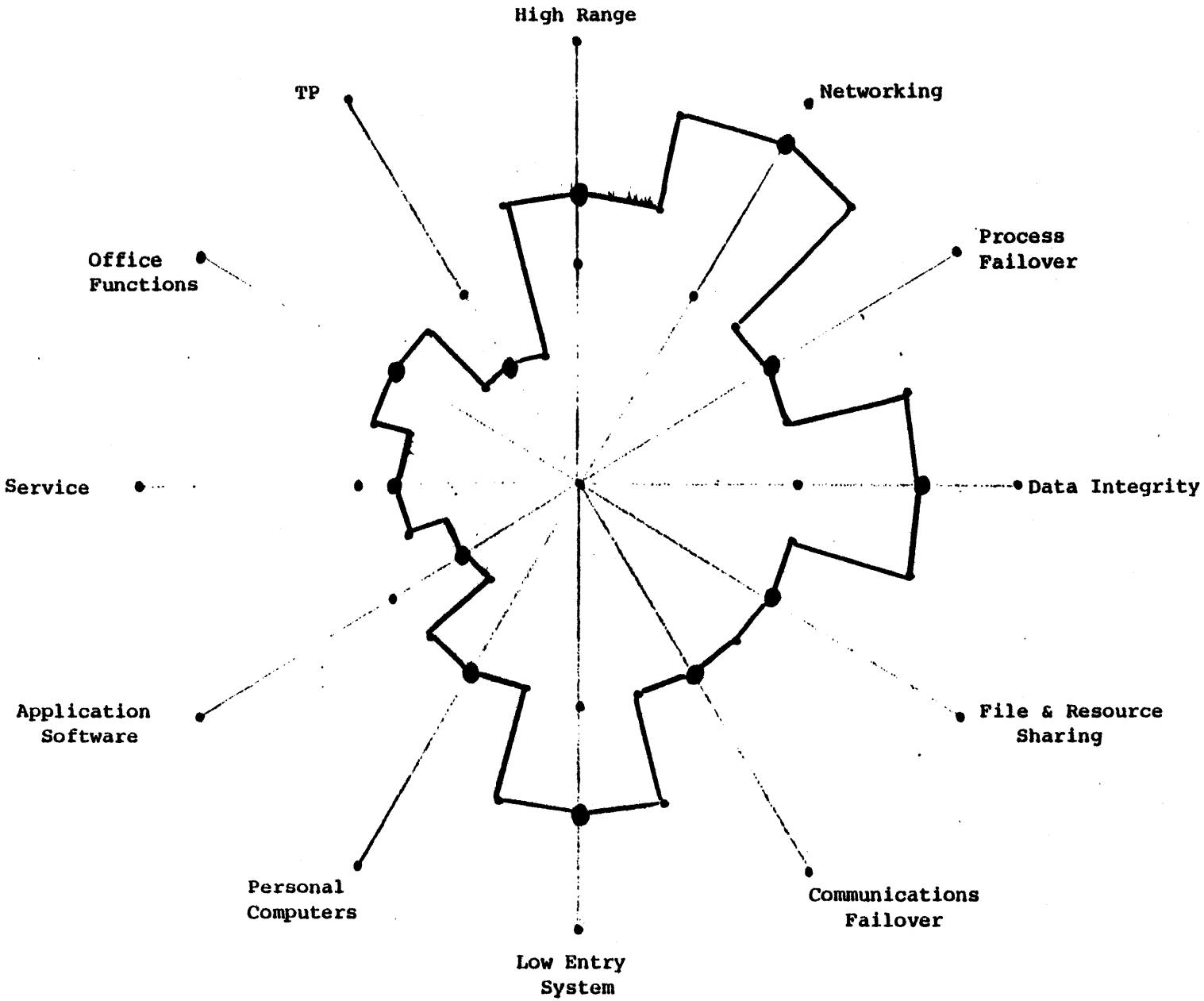
Cost/Effort = Ethernet/NI

Program Announcement Phase II

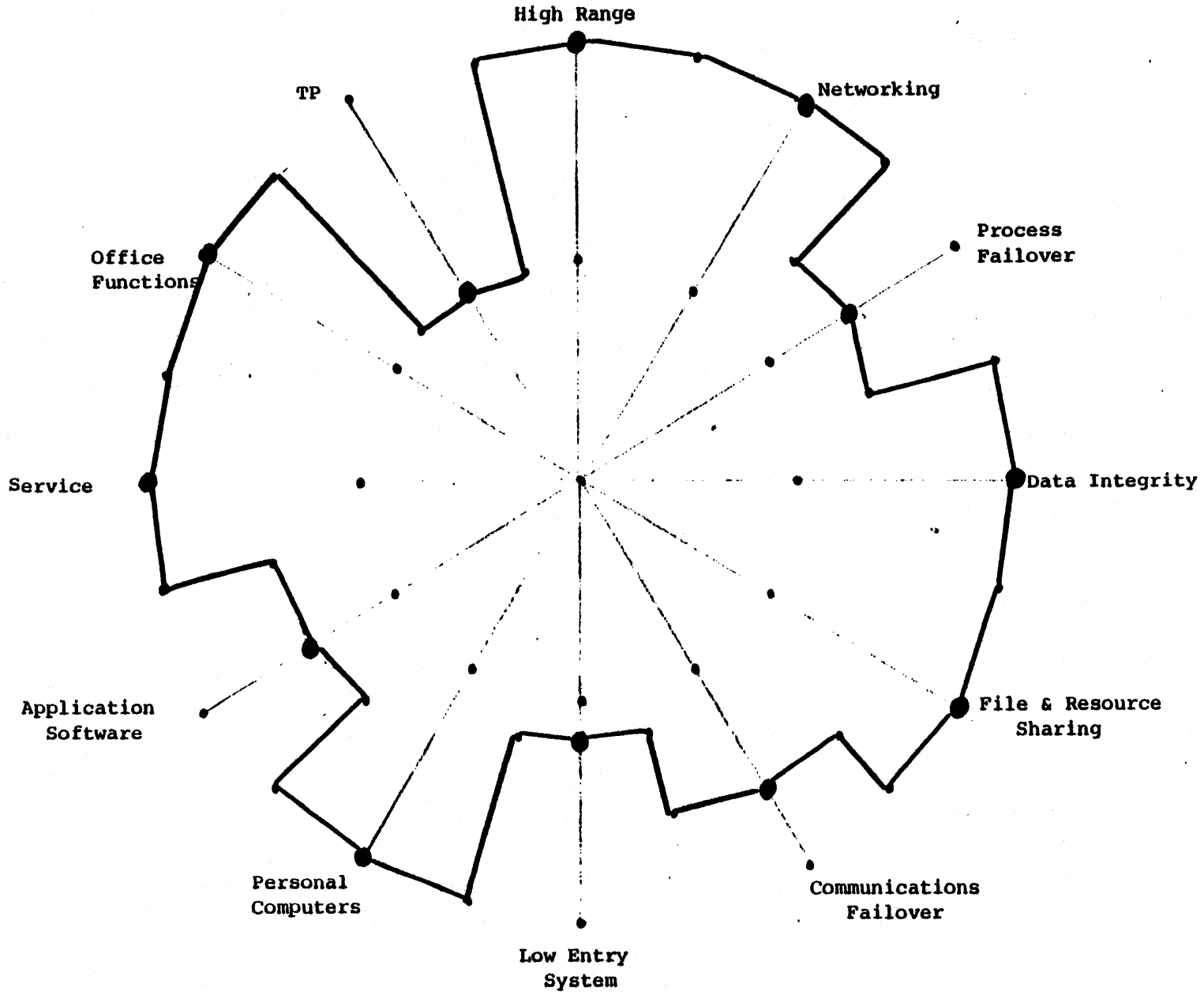
May-June 1983

- Update of ALL VAX/VMS Promotional Material
- Slide Presentation (Product Detail)
- Major Program/Product Announcement

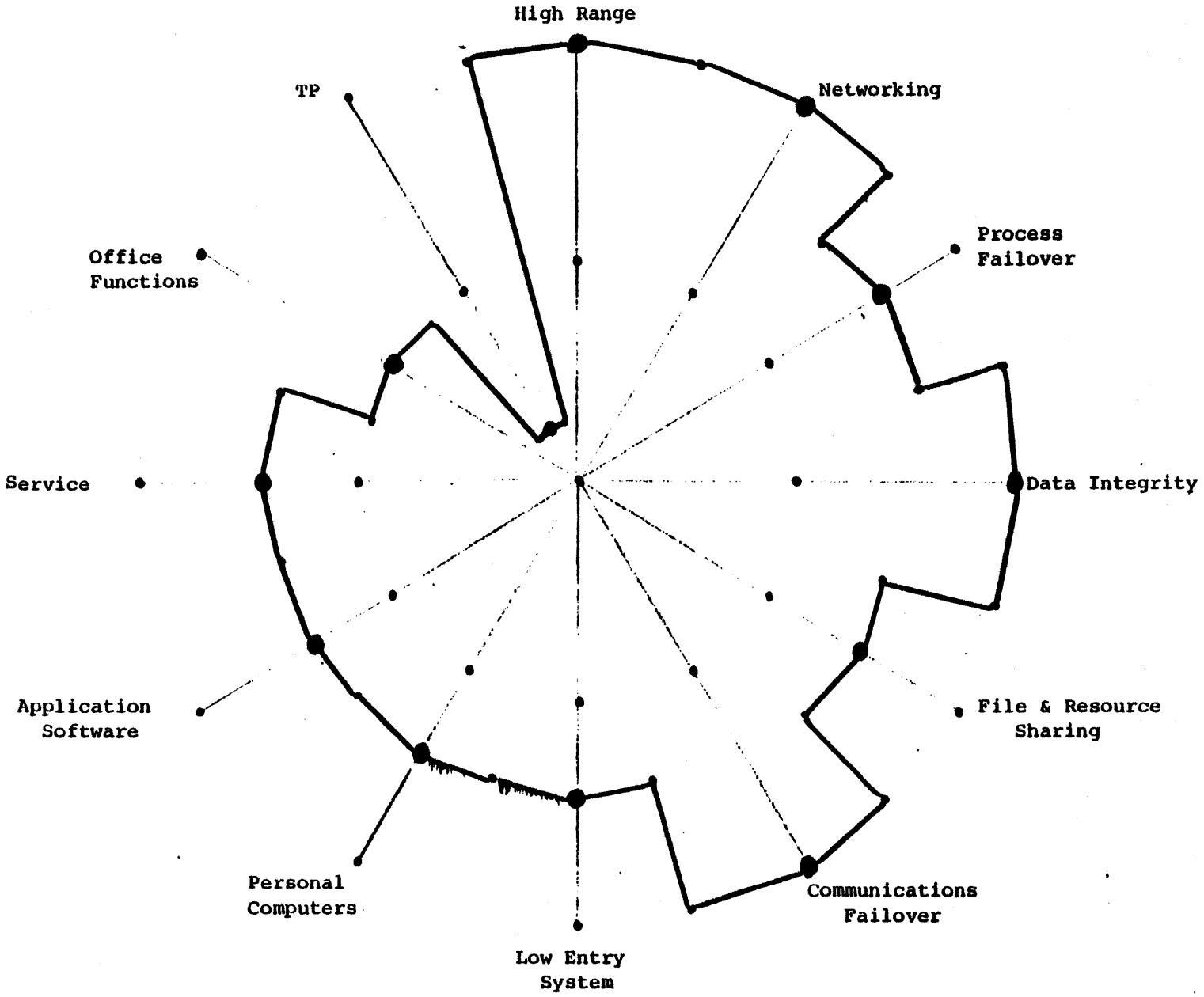
Cost/Effort GT VAX-11/730



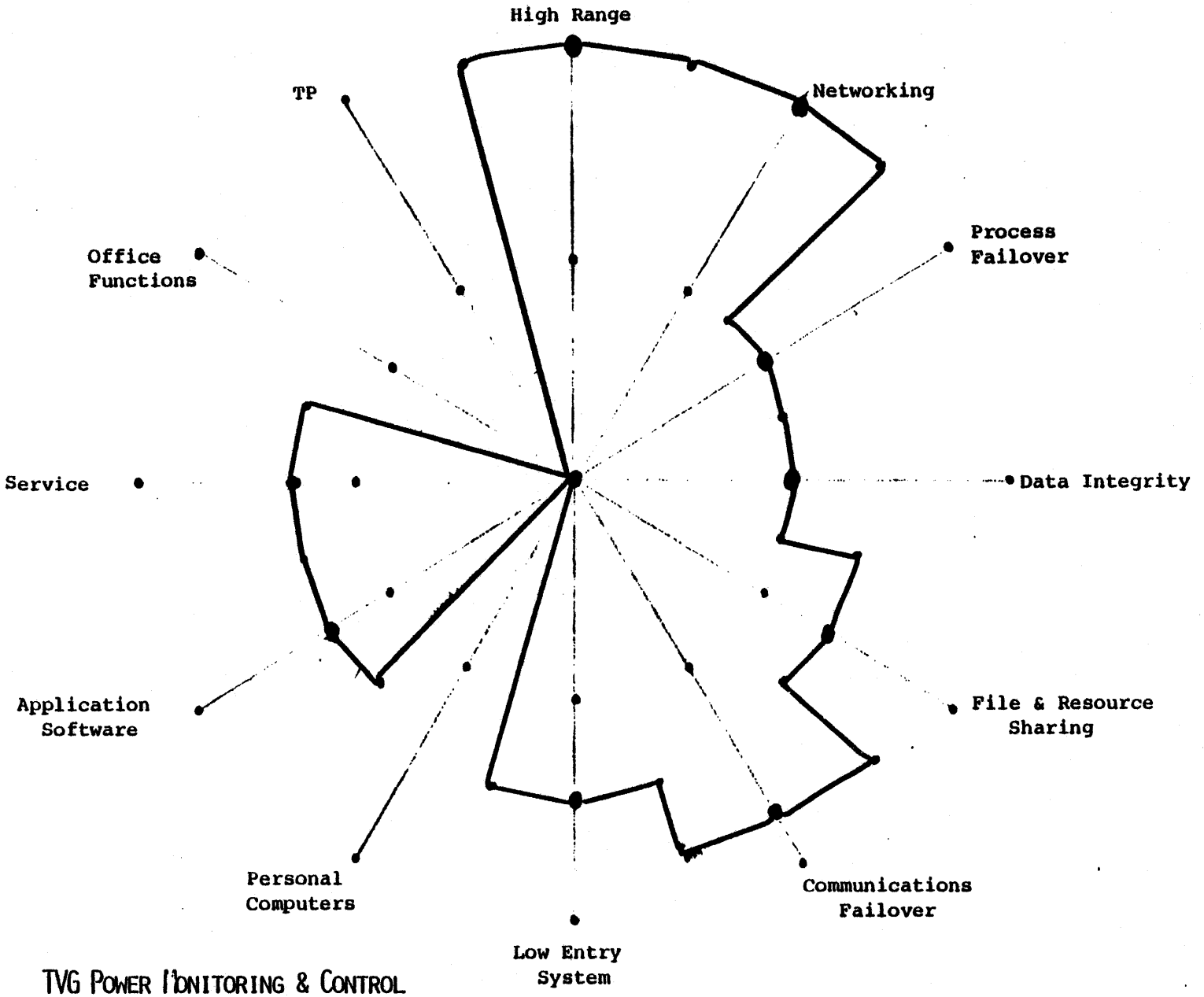
LDP Survey, Digital Today



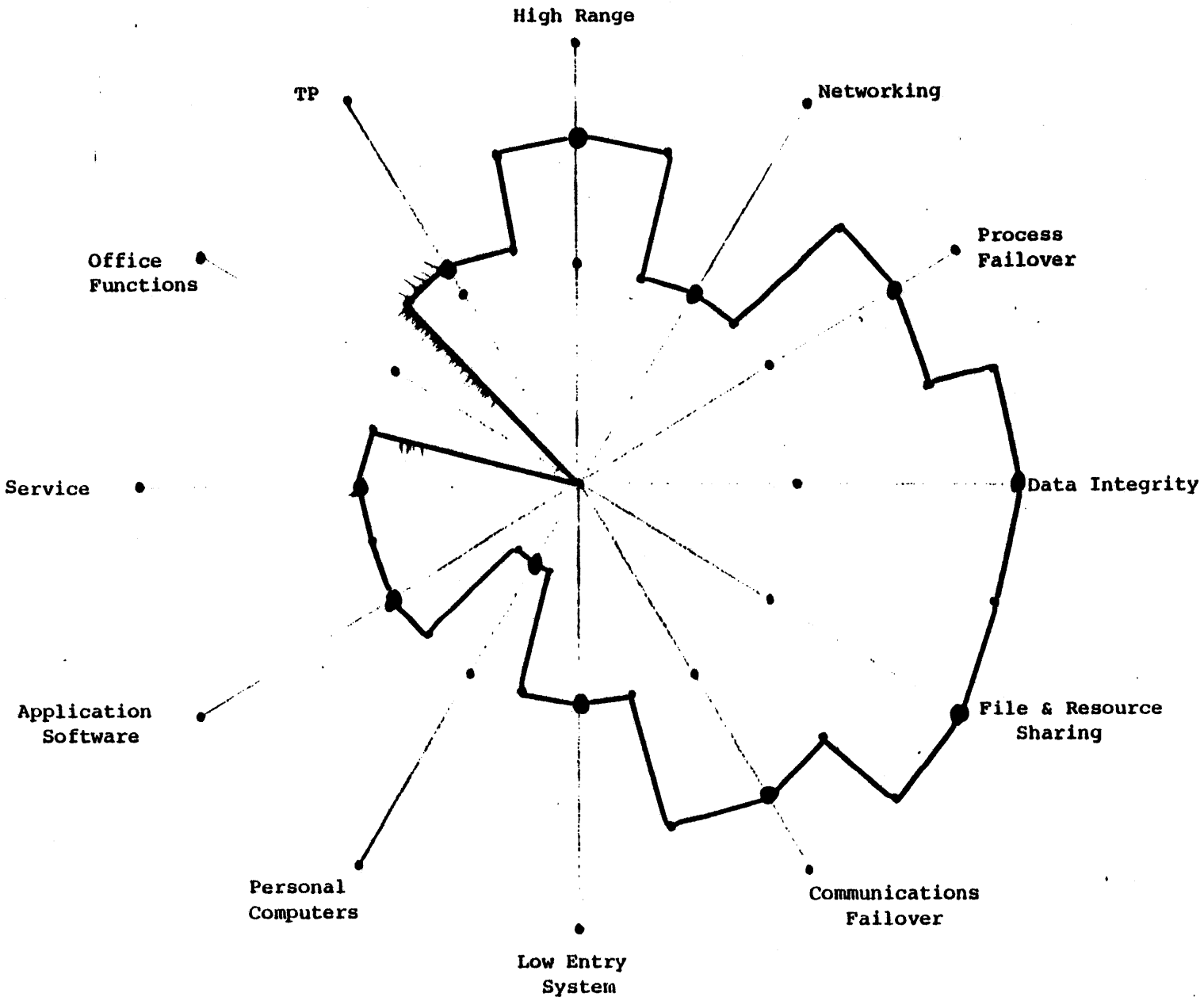
LDP Research Market Needs



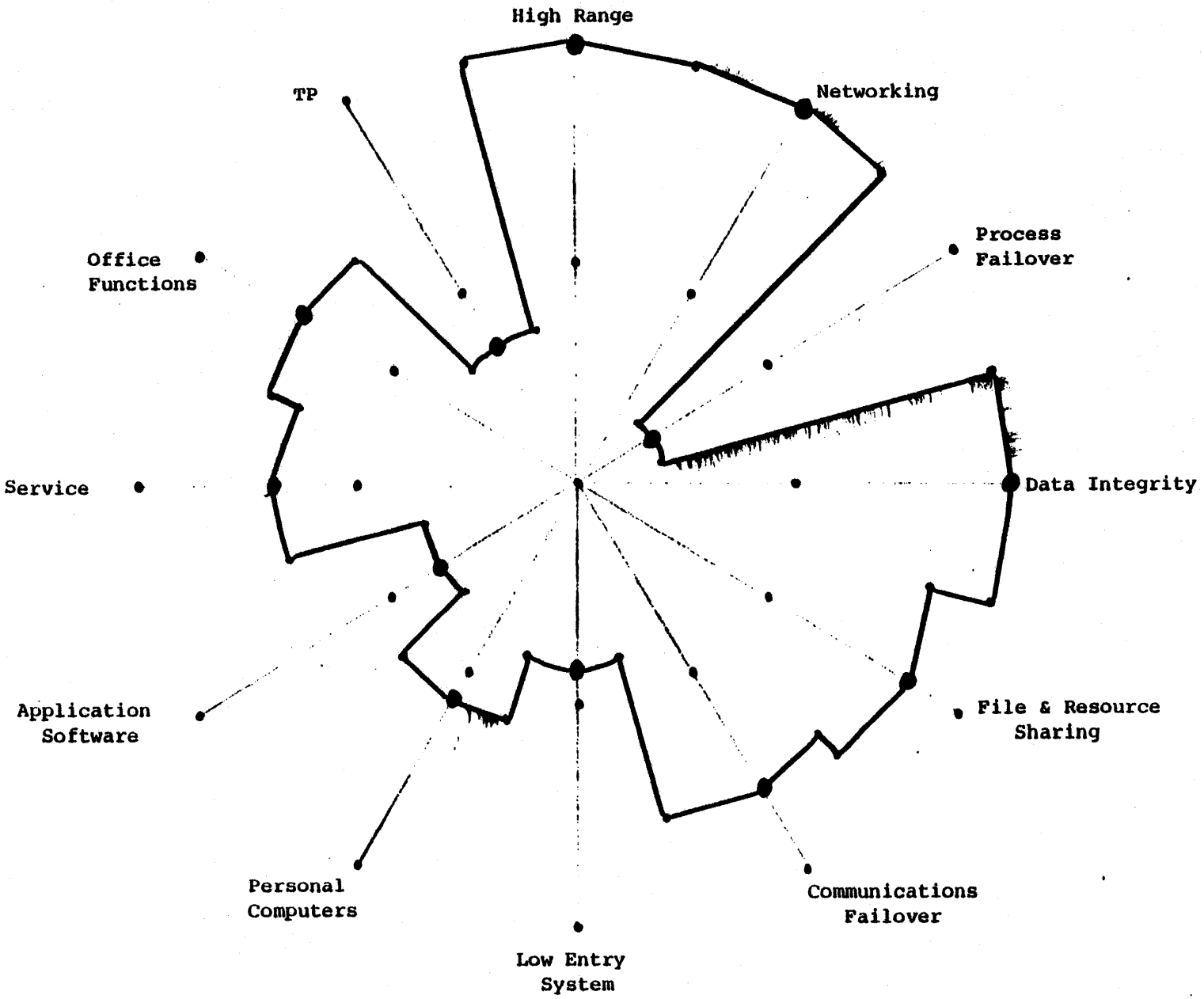
GSG Cluster System Req.



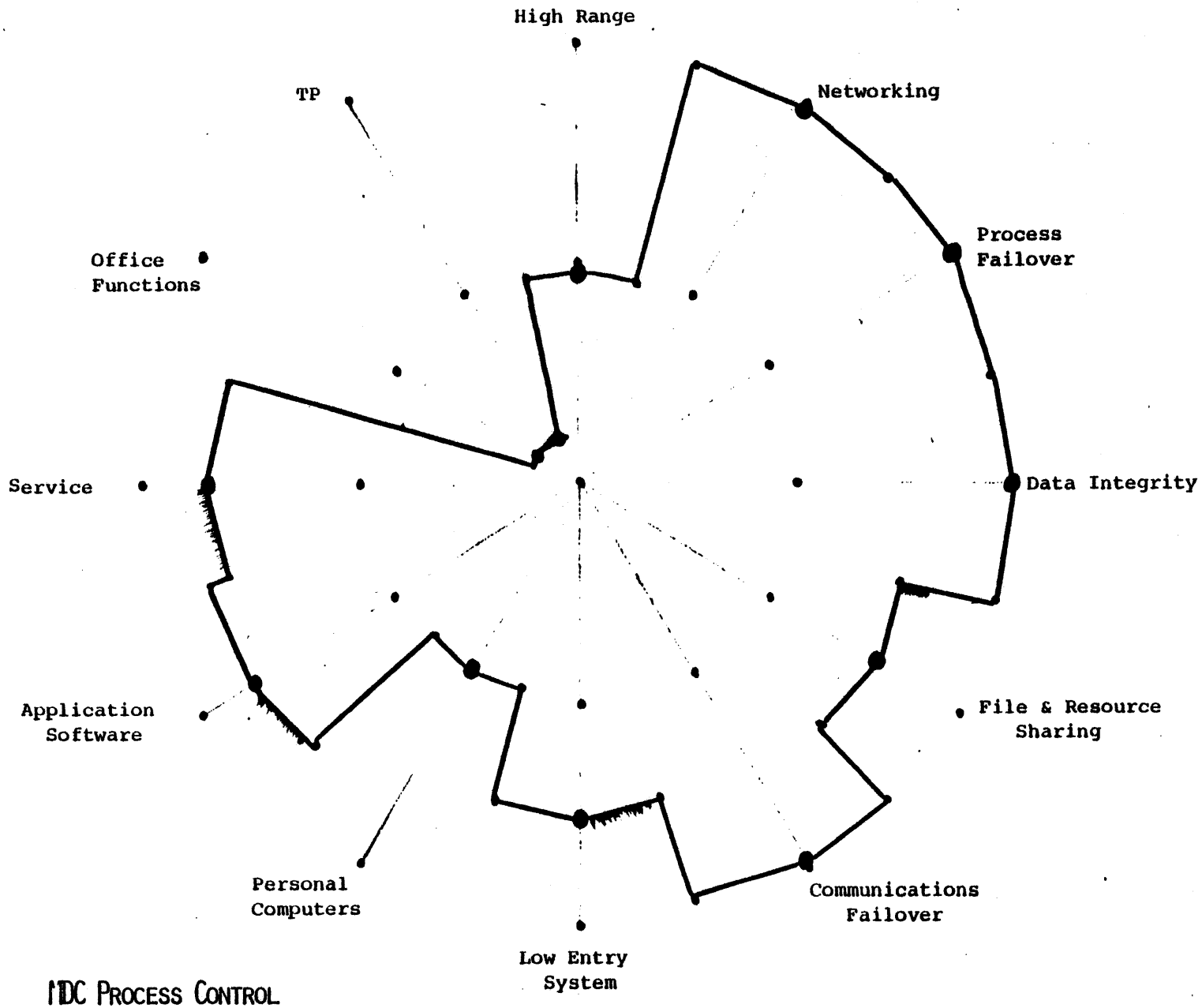
TVG Power Monitoring & Control



TVG PRINTING & NEWSPAPER



MDC Office Application



MDC PROCESS CONTROL