digital interoffice memorandum

To: Wayne Rosing

cc: Distribution

Date: 16 April 1980
From: Rich Kalin
Dept: Office Systems
Advanced Development
Loc.: MK1-2/L2 DTN: 264-6145
Mail: VAX4::KALIN

SUBJECT: NI and Ethernet

The following position paper provides details and documentation supporting the statements I made last month about Ethernet. I hope this answers any questions you may have. As I said before, the issues are complex and some are highly technical, but if you follow what I have to say, I am sure that you will be led to the conclusion that:

1. NI must not be based on Ethernet. Ethernet is an laboratory toy. It lacks the RAMP features necessary for an NI product and an architecture capable of supporting RAMP enhancements.

Although not specifically discussed in the position paper, the research I undertook in its preparation led me to the following conclusions regarding our overall interconnect strategy.

1. We desperately need an interconnect architecture. We needed it a year ago. By not having it, there is little coherence to the proposed interconnect products. Complex software interfaces will be required and the construction of simple bus-to-bus adaptors has been made difficult, costly, and in some cases impossible.

2. The entire interconnect program is on its way to failure. The requirements of system and application designers have once again been ignored until after the major decisions have been made. The real costs of bottom-up design are visible in the estimates for producing supporting software. By not building the right functionality into the transport mechanism, we have massively increased the size and cost of the entire program.

3. Before going any further, we need to stop and take a closer look at the systems our customers hope to build. At least in the area for which I am responsible, the interconnect program fails to provide the basic products needed. Unless we adopt a more comprehensive interconnect strategy, Digital should forget about making future sales to the Office Systems market.

ETHERNET EVALUATION FOR OFFICE SYSTEMS INTERCONNECT

Richard Kalin

MK1-2/L2
16 April 1980

1.0 Summary and Introduction

This position paper discusses design requirements for local network interconnect (NI) in the Office Systems Marketplace and the suitability of Ethernet as a technological base for a NI product. Topics covered include: design methodology, requirements and frills, Ethernet pros and cons, and strategic alternatives. The main conclusion reached is that Ethernet fails to meet the needs of this market, and that to adopt it as an NI interconnect standard would eliminate major business opportunities. Acceptable alternatives do exist and recommendations are made as what should be done next.

Concepts under discussion:

- Local Area Network (LAN). The interconnection of computers, terminals, and other computer peripherals distributed around a campus or building. LAN's may span distances of up to several kilometers and may interconnect as many as 5,000 network drops.

- Local Network Interconnect (NI). A high speed, serial bus used in LAN communication subnets. A single NI bus may be as long as several kilometers and have several hundred network drops.

- NI Repeater. An adaptor for joining NI's together at the bus level, so that one NI becomes the logical extension of the other. NI's and Repeaters are the building blocks for LAN communication.

- Ethernet. An early (1972-) experiment at the Xerox Palo Alto Research Center (PARC) in Local Network Interconnect. The transmission media consisted of a one-kilometer long coaxial cable with pressure tap T-connectors allowing up to 256 network drops. Burst transmission rate was just under 3M bits per second. The Ethernet design proposed for NI would allow 2.5 kilometers of cable, 128 taps, and transmission rates up to 10M bits per second.

- Office Systems. The application of computers to the office environment. Today, this refers to the combination of Word Processing, Data Processing and Electronic Mail. Customer and forecaster expectations for the next ten years call for the integrated handling of voice, image, text and data.

The office has been called the next frontier of computer products. Over the last ten years office productivity has not increased, while rising office costs have consumed a larger and larger percentage of total business revenues (40-50% in many industries). Today's office is the least capitalized and most labor intensive area of any modern business. Only recently, have technological advances made it possible to computerize routine office functions.

The "Office of the Future" is envisioned by many to contain large numbers of small, distributed computer systems, tied together with a high speed, local area communications network. An intelligent terminal, with voice I/O and image display capabilities, will be found on every desk. Attached as other network drops will be intelligent copiers, specialized processors, data base controllers and archival storage devices, and telecommunications gateways to external networks.

At current growth rates, total annual office costs are expected to reach $1.5 trillion by 1989. The market for "Office of the Future" products is then estimated to exceed $200 billion. This represents a prime opportunity for Digital's future growth. We have already declared our intention to become a major factor in this future market.

Unless otherwise indicated, quotes are from the Proceedings of the Local Area Communications Network Symposium, May 1979, sponsored by Mitre Corporation and the National Bureau of Standards. LAN design is a new and little understood area. Every network designer, it seems, can extoll the superiority of his own designs, while pointing out dangerous defects in all others. In my readings, I have found the LACN Proceedings to be most balanced statement of the ideas and issues available.

2.0 The Need for Top-Down Design

Excellence in design is possible only when the all the problems to be solved are fully understood. It is easiest, of course, when there is already a fully operational system to observe and work from. It is most difficult when attempting something new.

There is always the temptation when attempting a complex design problem to start with a small manageable component, to design it first, and then build upon it. There is nothing wrong with this as an intellectual exercise; it is a useful method of coming to grips with the overall problem. As an implementation strategy, however, it is extremely dangerous. To freeze the design of individual components before overall system requirements are understood invites disaster. Choices that look attractive early on may prove to be less than satisfactory as more details come to light. If the completed system ever works at all, it may turn out to be too big, too slow, or too expensive.

Computer networking is a classic case of where bottom-up design has failed miserably. Despite a history of disappointing efforts, we persist in defining low level interfaces and protocols first, before we understand the overall system requirements. The effect on higher system levels is complicated software, arcane user interfaces, and low performance operation. By not looking at the total problem, we postpone solution of its most difficult aspects until last. Responsibility for making everything work is left to the highest level software, often with inadequate tools provided for effective resolution.

By the time the difficult problems are recognized, it is usually too late to change the base components. Four years of network definition and development took place before the first distributed application ran on ARPAnet. Twelve million dollars had been invested in layered software, that was not about to be discarded.

It is now recognized that the many of the difficulties of writing ARPAnet applications are the result of design decisions made in 1968. In particular, the reliance on packet switching has caused major problems. Despite their conceptual simplicity, ARPAnet, Ethernet, and other packet networks have two major failings. They do not support simple interfaces and they complicate the design and delay the implementation of high level protocols.

To quote others on these points:

- John McQuillan, BB&N, "Networks have been designed backwards. The inside-out approach, starting from the lowest level functions and working towards the user, has not worked well. In the future, communication systems should be designed from the outside-in, starting with the functions and features that users need and moving towards the basic communications transport mechanisms which need to be built to support those functions.

  "Much of the work that is going on today in local network technology is misdirected. People should consider shifting their attention from the local network itself to the higher level protocols which will make using the network practical and cost-effective for a new generation of computing equipment."

  "It is unfortunate that we are not using the opportunity available to redesign packet technology to make it more generally useful. Local networks present the opportunity to design things right the first time and avoid costly re-design."

- Dave Farber, University of Delaware, "Too often we start with communication primitives and work up - an unhealthy exercise. One should look at the interprocess communication requirements and work down. We should deliberately put functionality into the communication system. (July 1979 R&D Seminar)"

- John H. Monahan, MITRE Corporation, "It goes almost without saying that since we are interested in human-related information distribution, our [LAN] networks must accomodate voice, video and data services. New combinations of data and voice -- such as voice cueing in conjunction with data transmission -- should also be anticipated."

- Robert Kent, Hanscom AFB, "The presumptive view that office automation will automatically increase office productivity, save money, eliminate paper, etc., has led to disaster in a number of instances. If you are going to change the world, you better find out what the world is all about first."

3.0 Functional Requirements for a Local Area Network (LAN)

An assumption of this section is that it is possible to derive the functional requirements for Local Area Communication Networks from an analysis of a major application area. This may not be true. Other application areas could impose additional requirements, not encountered in the Office Systems area. These would result in additional design constraints and possible design tradeoffs. But whether or not this is the case, no LAN design is acceptable that does not meet the mandatory requirements listed below.

3.1 Mandatory Requirements

- High Availability. John Monahan, Mitre, "As networks become more pervasive, more depended upon, reliability will become critical. Single points of network failure will become intolerable. A failure in one area must not affect the remainder of the network. Initial design of the network should include an on-line capability to monitor the health of the system, a capability to detect degradations that presage failures, and an ability to shut down failed units while maintaining service to the rest of the network."

  The need to diagnose bus failures on line is more important at the NI level than it is, say, at the UNIBUS level, because of the distances involved. UNIBUS devices are located within a few feet of one another. A field service person can quickly find most bus-related errors by swapping PC cards. LAN drops, in contrast, can be miles apart, locked in offices, or hidden in cableways. Finding and separately testing each one is both impractical and hopelessly time consuming.

  The LAN architecture should not preclude building networks with even more stringent availability requirements (e.g., self healing networks.)

- Modular Expandability. The network should be able to grow incrementally to meet the expanding needs of the organization. Service should be homogenous, new drops should offer the same protocols and access privileges as existing ones. Ideally, growth should be possible without limit, but a 5,000 drop limit

per LAN would be acceptable. It must be possible to add new drops and additional bandwidth while the network is running.

- Incremental Modernization. A large LAN will almost always be installed over a period of many years. During such time, many technological improvements will become available. For example, it should be possible to upgrade the interconnect media to a high-speed fiber optic link without affecting the NI interface or requiring either hardware or software changes to attached devices.

- Speed Independent Interfacing. Protocol definitions at the drop interface must be independent of network topology or NI data rates. Protocol timeouts should be provided by the NI adaptor and transparent to the tributary device. Tributary devices must be able to clock I/O at any data rate obtainable.

  There is no correct data rate for LAN tributaries. That is why data rate selection must be part of the interface definition. Every device is different and communication needs change with time. Today's terminal users would be happy with 9600 bps. Most voice is digitized at 64K bps. A Canon Lasar printer requires 700K bps. Full-motion video requires megabit rates. We can expect better data compression methods to reduce these requirements, but the need for supporting faster processors will drive them higher. All that can be certain is that future requirements will change and that the standard tributary interface must accomodate such changes.

- Real-Time Communication. A tributary must be able to request and receive service guarantees for bandwidth, uncorrected error rate and maximum message delays. This is required for real-time devices and highly desirable for other devices as well, as it greatly simplifies protocols, data base synchronization, buffering, and message flow control. The communication media must be treated as a resource of finite capacity that can be allocated dynamically to tributary connections that require virtual circuit capabilities.

- Standardized High Level Interface. A standard, high-level interface that is transparent to LAN technology or network configuration is essential.

- Low System Cost. The emphasis here is on the life-cycle cost of the entire LAN system, not just the cable, connectors, and interface electronics. Hardware costs can easily become insignificant when compared with costs of field maintenance operations, lost productivity on down time, protocol gateway conversion, system reconfiguration, and performance tuning.

3.2 Desirable Features

- Video data rates. The ability to send full-motion video through a LAN is a desirable, but unnecessary, feature in the initial product if it can become available through incremental upgrade. High availability and interconnectability are far more important than high data rates.

- Plug Compatibility with other Interconnects. It is axiomatic that NI's, BI's, CI's, etc. have to work together. Plug compatibility is ideal, but unnecessary if simple adaptors can be used. This desirable requirement underlines the need for a coherent architecture for designing interconnect products.

- Customer Installation. Cabling and installation rules should be simple enough for electricians to follow. Attachment rules that reference global requirements (e.g., maximum number of drops on a cable) need quick and simple, on-line test procedures associated with them.

  It is desirable to install all the cables in an area at one time and add drops only as they become needed. Attachment procedures should be simple and safe enough for end users to follow.

3.3 Frills

- Simple Line Protocol. This is the false god of bottom-up network design. While it important to keep the overall system as simple in design as possible, this does not mean starting with an elegantly simple line protocol. Many things, difficult to do in layered software, can more easily be done at the level of the physical link. In particular, I maintain that the DECnet implementations would be much simpler if flow control had been designed into DDCMP, instead being pushed into the NSP and application program layers.

- Contention Line Control. Much of the LAN literature makes the assumption that collision protocols are the only available alternative to using a centralized arbitration controller. (Central controllers are considered undesirable because their failure will bring down the entire network.) There are, in fact, other workable alternatives that have been used, such as, having each station control a separate, non-critical piece of the interconnect media, or allowing line control to migrate from station to station.

- Minimal Hardware Cost. Minimal hardware means that there is less to break, with the expected result that MTBF is improved. However, it can also mean that important RAMP features are missing, with the result that downtime costs and MTTR can become completely unacceptable. This especially important in the LAN area, where the operation of the communication network becomes essential to business operations.

- Inexpensive Single Cable. Typical installation costs of $4.00/foot have been quoted for laying coax cable within an existing building. In comparison, the cost of the actual cable ($0.11/foot) is negligible. Pulling two cables costs little more than pulling one. Adding a second active cable is one option for achieving high availability.

- Optimal Line Utilization. The speed with which most LAN networks operate is limited not by the capabilities of the media, but by the speed of the processors that control it. The facination with line utilization is an artificial concern, useful if it can be achieved, but less important than the need to maintain acceptable tributary performance during periods of peak loading.

- Network Services. Some capabilities worth putting beneath the NI interface include: speed conversion, code conversion, flow control, echoing and data forwarding.

4.0 Ethernet Evaluation

Quotes in this section are either from the July 1976 CACM Ethernet paper by Metcalfe & Boggs or from John Shoch's draft NI-Ethernet Specification.

4.1 Good Points

- History of Use. Ethernet was one of the earliest LAN's to be built and has been the best modeled and studied. It has been in daily use at Xerox PARC for more than five years. An impressive amount of software has been written to support and make use of its capabilities.

- Simple design. Ethernet is the simplest of the LAN designs. MTBF has been high because there is little hardware to fail. Ethernet interface hardware could be manufactured at low cost.

- Packet Network Compatibility. The Ethernet protocol is similar enough to the packet network protocols that a "universal" gateway interface has been built.

- High Burst Transmission Speed. Even with several hundred drops, average line utilization of the PARC Ethernet averages less than 1%. A high proportion of all messages are sent without interference. Because tributary interfaces run at line speeds, downline loading and bulk data transfers take place at megabit rates.

- On-Line Reconfiguration. Pressure taps can be added or removed while the line is running. Tributaries are electrically isolated and can be separately powered on and off.

- Graceful Overload Degradation. Unlike the earlier contention networks the Ethernet protocol does not become unstable under overload conditions. The throughput apparent to individual tributaries drops, but channel utilization remains high.

4.2 Critical Defects

- Bottom up design. Ethernet was never designed to be a product. The original ARPA-funded experiments were directed at studying the channel utilization of Packet Radio protocols (ALOHA). The cable architecture dates to 1972, a time when the ARPAnet was still in its infancy and long before the problems of high-level packet protocols were recognized. Not surprisingly, a massive software effort at Xerox PARC was required to turn a laboratory experiment into a usable communication system.

- Lack of Alternate Message Paths. The notion of a single, half-duplex message channel is embedded into Ethernet. "There must exist only one path through the Ether between the source and destination; if more than one path were to exist, a transmission would interfere with itself, repeatedly arriving at its destination having traveled by paths of different lengths." Communication stops in the event of cable failure.

  Because any network failure that causes "pollution of the Ether" will render all communication impossible, transceiver design has been described as an "exercise in paranoia".

  One might think that the requirements for high availability and on-line failure diagnosis could be accomplished simply by running two Ethernets in parallel. This has not been shown, but it is expected that non-trivial transceiver and protocol redesigns would be required. As we have learned from Hydra, high availability must be put into, not layered on top of, base system products. Any requirement for non-stop operation changes the very character of the design.

- Data Reliability. Ethernet literature equates message reliability with CRC errors, but "packets may be lost due to interference with other packets, impulse noise on the Ether, an inactive receiver at the packet's intended destination, or purposeful discard." As Ethernet provides no mechanism for detecting or reporting that a packet has been lost, all such errors must be treated as failures of the transport mechanism.

  "Removing the responsibility for reliable communication from the transport mechanism allows us to tailor reliability to the application and to place error recovery where it does most good." It also complicates high level protocols considerably and upsets network layering by requiring upper protocol levels to contain details of the transport mechanism.

- Susceptibility to Tributary Failures. Failure of a single tributary can render the entire Ethernet inoperable. Isolation of which tributary has brought down the network cannot be made on line. Physical isolation of the failure within a cable is difficult because only three cable splices are allowed and a single section can be 2500 meters long. It can be necessary to trace the length of the cable with test equipment, checking every drop along the way.

- Cabling restrictions. Limiting the topology of a single strand of cable makes it difficult to install a Ethernet without use of repeaters. These, as seen below, present problems in them selves. Once in place, an Ethernet cable cannot be spliced to accomodate changing requirements or office layouts.

  "The distribution of transceiver locations along the cable is critical, as improper placement can cause a severe increase in the magnitude of the apparent reflection caused by the shunt capacitance of the connection."

  Tap placement rules (e.g., "No tap cluster may contain more than 5 taps. No two non-overlapping clusters may be spaced less than 90 meters apart.") make it difficult to adapt cabling to common office layouts.

- Global Impact of Network Extensions. Adding repeaters to the net changes the timing of the protocol at the transceiver interface. Since there is no way for a station to determine that a repeater has been added, or how many repeaters stand between a transmitter and a receiver, high level protocols must be designed to keep track of network topology.

- Lack of Real-Time Throughput. Under overload conditions, the Ethernet algorithms for contention resolution have the effect of distributing channel capacity equally among all contenders. This concept of "fairness" is acceptable only if all stations can tolerate the degraded throughput and are considered of equal priority. (Imagine Digital attempting to allocate its manufacturing output the same way.) There is no way for higher protocol levels to circumvent this method of bandwidth allocation.

  This point may seem unimportant in light of the low utilization that Ethernet received at Xerox PARC, but the environment of the research laboratory says little about what will happen under field conditions. No matter what bandwidth NI supports, it will be used up. Overload conditions will occur and our systems must be designed to expect them. Ten megabit line speeds are no protection.

  Because the Ethernet contention protocol has the effect of allocating bandwidth on a message (not a byte) basis, the bandwidth available to a tributary can drop dramatically at

times of network overload. Under saturated conditions, a 10M bps Ethernet will transport as few as 1200 messages a second. If all 128 drops are competing for bandwidth, each can only expect to send less than 10 messages a second. For an application unfortunate enough to use single keystroke messages, this message rate hardly keeps up with typing speed. In contrast, the same application on an unloaded network could send almost 28,000 messages a second.

After studying Ethernet message delays, Ed Lazowska, University of Washington, reported, "Individual tributaries perceive Ethernet efficiency to be a linear function of network load. Response time becomes infinite and throughput drops to zero when the network becomes completely saturated."

Fouad Tobagi has commented, "As the number of users that have packets backlogged increase, the throughput decreases, thus these [Ethernet] systems exhibit two equilibrium conditions, one at high throughput/low delay, the other at low throughput/high delay."

The Ethernet contention resolution protocol has the unusual feature of favoring later messages. "Last come, first serve," is the apparent policy. The impact of network loading falls hardest on tributaries with the highest burst requirements.

The inability of Ethernet to allocate bandwidth where needed makes it an unsuitable media for real-time communication.

- Fiber optic cables. In many locations, building and electrical code regulations make it more attractive to use fiber optic cables the coax. Industry forecasters predict that declining costs will cause fiber to replace coax in most applications within ten years.

  Fiber optic cables are is unsuitable for Ethernet use for two reasons. The first is that lasar/diode transmitter/receiver technology does not lend itself to bidirectional communication over a single fiber. Secondly, there is the unsolved problem of how to make a low loss, low cost tap onto an fiber optic cable. (Fibernet is <u>not</u> the answer!) In contrast, ring and other LAN technologies can be easily adapted to use of fiber optic cabling.

  In addition, the Ethernet bus contention protocol degenerates at the gigabit data rates possible with fiber optic links. To get the full benefit of such high-bandwidth technologies, Ethernet would either have to be restricted to 25 meters of cable or messages would have to be 100,000 bytes long.

- Extensibility. Repeaters are the Achille's heel of the Ethernet design. The simple "echo everything" repeater upsets the bus timing and efficiency of the backoff algorithm. The more complicated "filtering repeater" fails under conditions of

network overload and upsets the timing of higher level protocols. The Xerox PARC solution of tying Ethernets together using gateways and universal packet protocols (PUP's) is expensive both in hardware and protocol support.

- Incompatibility with Other Interconnect Products. The lack of a positive acknowledgement in the Ethernet line protocol makes it impossible to directly interface a Ethernet-NI to a CI bus. The higher level protocols that the adaptor would need to implement are outside the scope of the interconnect program.

4.3 Could be Improved

- Base-band signaling. (This is not a requirement for Ethernets, but is a feature of the design being discussed.) A single CATV cable provides about 300MHz of usable bandwidth. Ethernet uses only a small fraction of this bandwidth, but its base-band signaling scheme prevents the same cable from also being used for TV video or other signals. If transmissions were RF modulated, two logically separate Ethernets could share the same cable.

- Manchester encoding. Richard Hertzberg points out that the use of asynchronous character format, recovering the clock from the use of start/stop bits, is "more efficient, at 20% overhead, than traditional clock encoded data, e.g., Manchester encoding, which requires 50% overhead."

- Security. Transceiver taps can be added to a Ethernet cable in a completely passive way. Once on, they can monitor every message exchanged. Unless they themselves transmit, there is no way for other stations to detect their presence.

- Distributed contention control. Ethernet's contention protocol makes it difficult to study network activity or to prioritize message traffic across tributaries.

- Flow Control. If data arrives at a receiver unable to accept it into its buffers, that data must be thrown away. Bus band time is wasted. Higher level protocols are required to insure that the message is retransmitted. "There is no way for a receiver to quench the flow of such wasted transmissions or to expedite retransmission. (Metcalfe)"

- Contention algorithm. Ira Cotton, "As the data rate or the length of the transmission path (or some combination) becomes excessive (e.g., above 10M bits per second on a local network, or at any rate on a satellite system where distances are about 25,000 miles [sic]), these [Ethernet] contention rules begin to break down and cable utilization tends to regress to that of the slotted ALOHA system. For such systems, reservation schemes are used instead, where some packets are used to form a logical channel for control purposes (these being allocated by contention), the remaining (majority of the) packets being

reserved in advance for particular users through the use of the control channel." Note that the same problem occurs when the Ethernets are tied together by repeaters.

- Pressure tap T-connectors. The CATV industry has warned against the use of pressure taps because they are unreliable. Standard CATV T-connectors have excellent reliability, but cannot be used on Ethernet because of transceiver sensitivity to cable reflections. Since cable connectors are critical components of LAN systems, connector strategy should adopt, if possible, the proven technology of the CATV industry.

- Asynchronous clocking. The data rate at which each station transmits is determined by its own internal oscillator. Clock drift and phase differences can cause metastable "glitching" of receiver phase lock circuitry, introducing non statistical error patterns. Such errors become more prevalent with higher line speeds (10M bps vs. 3M bps of the original Ethernet).

- Inadequate Modeling Efforts. None of the performance measuring or modeling efforts undertaken by Xerox have considered the Ethernet performance as seen at the transceiver interface. Some still unpublished work has taken place, however, at the University of Washington. Real-time constraints do exist and models should be improved to include the delay distribution, not only the average.

5.0 Conclusions and Recommendations

The major conclusion of this analysis is that Ethernet does not and cannot meet the needs for local area networking as required by future Office System products. It lacks the RAMP features needed to become a DEC product and should under no circumstances become the NI standard.

In light of the above, I recommend that the following major actions be taken:

- Stop NI-Ethernet development until an acceptable alternative can be found.
- Define the LAN requirements we intend to fill, from the point of view of network functional capabilities, RAMP, extensibility, etc.
- Validate these requirements against current business plans and future strategies.
- Complement our own expertise with that provided by a variety of outside experts, each of whom can provide us with a different perspective on this area.
- Undertake a top-down design approach to the point where NI requirements are clearly understood in the framework of a solid LAN architecture.
- Then implement the bus(es) that evolve from this architecture and design methodology.