TECHNOLOGICAL UNDERPINNINGS: HARDWARE

Scott E. Chapal

Joseph W. Jones Ecological Research Center, Route 2, Box 2324,
Newton, GA 31770

Abstract. Choosing appropriate computing hardware is challenging in this era of rapidly changing technologies. Hardware purchasing decisions affect long-term information management because of the large capital investment and the necessity to design a computing infrastructure which can withstand software upgrade cycles and provide operating system inter-operability. The increasing prominence of 'the network' in all aspects of information management has contributed to a re-alignment of hardware procurement budgets, with a larger proportion allocated to support 'bandwidth' requirements. Servers are re-establishing their positions of central importance in all computer networks while client computers are being standardized, simplified and increasingly required to perform terminal duties. The requirements of ecological information management are not extraordinary compared to other data-intensive endeavors, but the potentially contradictory demands of research, archival, analysis and collaboration can overwhelm an inappropriately designed computing infrastructure.

INTRODUCTION

The primary construct common to all modern collaborative computing is the Local Area Network (LAN). The dominant uses of LAN's are various forms of client/server computing (see Nottrott this volume, Schildhauer this volume). Client/server computing models allow flexibility in the allocation of data and processing resources and provide for the evolution of hardware use and network design. It is nonsensical to make hardware purchasing decisions without understanding operating system demands, application requirements, and the possible design alternatives available in various client/server models.

Perhaps more important than the technical considerations are site-specific needs that must be central to LAN design and, therefore, for hardware acquisition decisions. LAN design can be approached from several perspectives, but in the context of this chapter, three are paramount: 1) design to optimize the environment for research information management, 2) design to maximize system administration efficacy, and 3) design to maximize cost/benefit. Recommendations for specific hardware vendors, models or strategies are omitted herein because of the transient utility of that kind of information. The pace of change in these technologies renders specifics virtually obsolete by the time they can be drafted.

LAN design goals for research information management

A fundamental requirement of any computer environment is to provide access to the user population. Scientists, staff, students and temporary personnel must all have access to the computing resources of the site or project. A basic design goal that results from this need is to provide *interface consistency* and a *single log-on* paradigm for the user. The consequence of this goal for hardware acquisition is to minimize the number of supported platforms. The procurement ramification of this reasoning is to consolidate the number of vendors to a strategically selected minimum.

A related goal is to *accommodate mobility*, or to give users access to resources from all points on the network. This means that the design will achieve a many-to-many relationship of people to computers rather than a one-to-one relationship that was the hallmark of the stand-alone PC. Although there are researchers who do primarily use a single computer in an office, the flexibility and utility of the LAN is enhanced dramatically when this many-to-many relationship is established. Security implications of this arrangement are immediately apparent, but the details of security planning are best left to another forum (see Nottrott this volume).

It should be obvious that the facilitation of data collection and processing in the research environment is absolutely essential. Therefore, data entry protocols (see Briggs et al. this volume) must be available and integrated into the network design, and instruments must be interfaced to their respective computers and those computers to the network. Although details will vary from situation to situation, it is desirable to simplify and standardize, and this applies to hardware as well as to network protocols and applications.

In order to support a cogent data management framework, data storage, data organization and data security must be thoroughly considered and incorporated into fileserver design. Data can be centralized, secure, accessed by multiple client computers, adequately backed-up and redundantly configured on servers. In contrast, a completely de-centralized collection of client computers, all serving as data repositories in peer-to-peer relationships, is extremely difficult to manage and use. The utility of the client/server architecture is obvious given the lack of alternatives that can scale to accomplish increasing research demands.

Data management/analytical software tools represent an arena where standardization and consolidation should also be design goals. Planning for these tools should be on a time cycle that is longer than the upgrade cycle of operating systems and probably longer than the turnover frequency of hardware. The investment in these tools via programming and data structures can be quite high and should provide longevity and continuity to meet the long-term information and research demands. Much of the research agenda of individual scientists and institutions is now interdisciplinary and requires synthesis to address broad-scale and long-term questions. The simple fact that collaboration is necessary should be designed into computing infrastructure planning at all levels, including hardware specification and procurement.

LAN design goals for system administration

Another perspective from which to address infrastructure development, is from the system administrator. Given that resources for system administration are often limited and difficult to expand, it is prudent to make decisions that reduce administration workload. Obvious ways to simplify network operations are to centralize administration, and standardize the hardware, operating systems, and software supported. Applying conventions to all system administration functions (userID's, name service conventions, filesystem layout, computer names, IP address allocation, mail aliases), is essential to laying the groundwork for automation of tasks. Automation is a powerful way to accomplish repetitive tasks, thereby freeing the system administrator's time for problem-solving or project development. Simplifying installations of client operating systems and applications is especially important for organizations that have more than a dozen or so computers.

Cost/benefit perspective

Addressing hardware from a cost/benefit perspective is relatively straightforward -- maximize network functionality per dollar invested. This goal is conceptually simple, but its implementation is rather more complicated, and depends largely on the information management goals and system

administration constraints. A common strategy is to 'Right Size' which translates to: 'Don't buy what you don't need'. This is more challenging when decision-making for technology purchases is distributed. Individuals may not be aware of the broader organization's needs or priorities and often make decisions based only on a single project, investigator, or end-of-year surplus budget. Avoiding redundant purchases and budgeting across project boundaries can be difficult, but given the extreme cost of the technology, it can be well worth the effort. A structured approach to building consensus through a committee can be used to help with these decisions.

Another economic motivation is to attempt to *future proof* the investment in computer hardware. This can often be likened to forecasting the future with a crystal ball, but there are some basic assumptions that hold true. 1) Delay acquisitions to the extent possible because hardware gets cheaper, faster and better every day. 2) Try to extend the longevity of components by predicting their useful life-span and their potential to be re-deployed to secondary functions. The consequence may be to buy fewer, better components.

Hardware is the most persistent part of the infrastructure (if hardware components are purchased rather than leased), and therefore must be able not only to accomplish today's needs, but be sufficiently upgradeable or re-deployable to have enduring utility. Leasing may be a valid option for some organizations, projects, or individuals and must be analyzed on a case-by-case basis. Non-profit organizations, for example, may not have the tax incentives that make leasing attractive to some businesses.

RIGHT-SIZING THE COMPUTER INFRASTRUCTURE

As alluded to in the previous section, right-sizing the computer infrastructure means, fundamentally, to avoid investing in unnecessary hardware or technology that does not appropriately fill the need. Needs-based planning, to use both an ecological and utilitarian metaphor, should be both top-down and bottom-up. Staff may argue as to who or what's on top versus bottom, but regardless, planning needs to occur from both the vantages of: 1) budgetary and personnel constraints, and 2) projects, goals, and other aspirations. A technology decision-maker may have to compromise incongruous demands -- usually too much ambition for too little money. To consciously avoid this decision process, however, may undermine the utility of the network over the long-term by specifying too little, or may unnecessarily inflate the technology budget by specifying too much. An example of this short-sightedness would be if budgetary allocation were too heavily skewed toward client computers and not sufficiently devoted to network hardware (hubs, switches): bandwidth could become a bottleneck, ironically just when high powered CPUs could take advantage of it! It may be difficult to strike a balance, and that balance, once achieved, will definitely change quickly.

Most organizations, by necessity, have to incrementally improve their computer systems by building on legacy systems (the existing hardware and software that comprise the computing infrastructure). Usually, legacy systems are both an asset and a constraint, but the fact that they do accrue over time underscores the need to assess technology acquisitions (hardware purchases) for their entire life cycle and total cost of ownership. Evaluating cost of acquisition without regard to longer term personnel, budgetary, applications and research realities can result in more expensive solutions over time. A key is to implement for change. Hardware can be cycled to secondary functions as it ages, for example, but this perspective assumes institutional-level planning and coordination.

It is important to understand the implications of standards on interoperability. Today's market leader can become tomorrow's albatross in the realm of proprietary network protocols and database technology. Therefore, it is a good idea to have at least a cursory understanding of the existence of standards in various areas of networking and database inter-operability. This

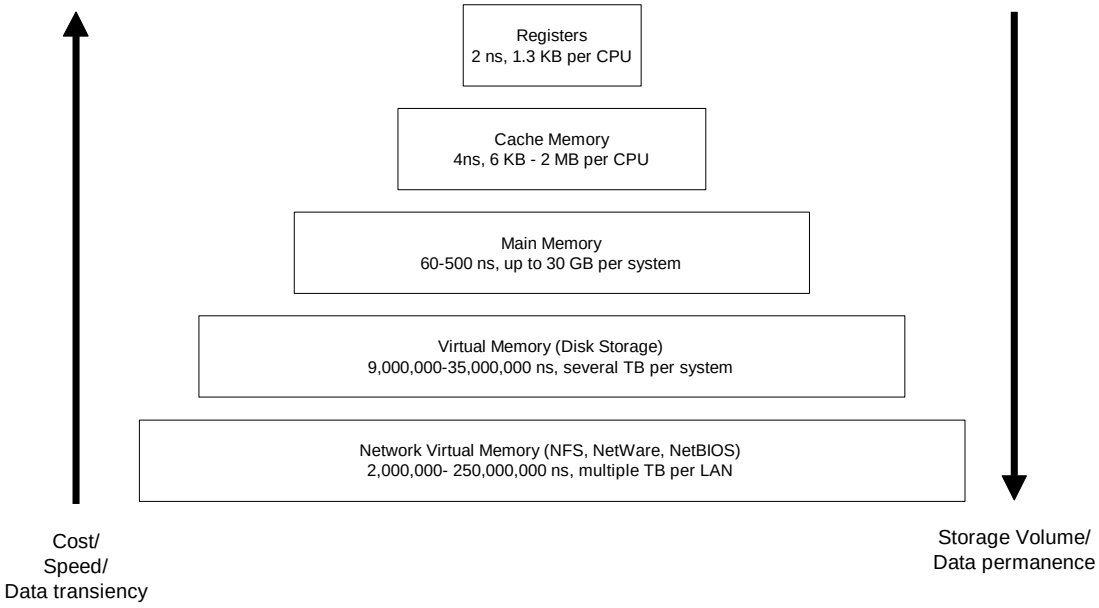
understanding will influence hardware purchases ultimately, since all component devices will need to communicate throughout their life-cycle. One only needs to talk to the systems administrator of a multi-protocol LAN to understand the complications that can result from a proliferation of proprietary network operating systems. Some of the standards that have evolved to address these issues are:

OSI - Open Systems Interconnection reference model
TCP/IP - Transmission Control Protocol/ Internet Protocol
POSIX - Portable Operating System Interface for Computing Environments
CORBA - Common Object Request Broker Architecture

The salient point is that standards can persist to a greater extent than proprietary implementations, and the universal motivation for them to do so is the need to interoperate. TCP/IP is the most recognized example of an accepted standard, and has become essentially ubiquitous in the Internet and in LANs. SNA and NetWare are testaments to proprietary architectures that functioned well in organizations, but their vendor-specific nature compromised their scalability and rendered them inelegant to incorporate into the Internet. For the Internet/WWW to continue to interoperate at ever greater levels of complexity, these *dejure* standards must continue to be respected and evolve, notwithstanding *defacto* standards that do achieve a level of interoperability such as Microsoft's operating systems dominance. It is worth noting that the address space afforded by the 32 bit Internet Protocol is quickly becoming saturated and the transition to IPv6 with it's128-bit address scheme is only achievable because of the acceptance of the standards process.

Right-sizing can also be viewed from the perspective of balancing the elements of the 'Virtual Storage Hierarchy' (Wong 1997). This is useful, because the utility of an individual computer or a LAN can be understood in terms of its ability to move data to where the data are needed: in the CPU, into RAM, cached on disk, on a server, or archived to some storage media. This illustrates the balance and distribution of processing and storage resources on a network and further emphasizes the advantage of designing functionality into the entirety of the LAN instead of into individual computers.

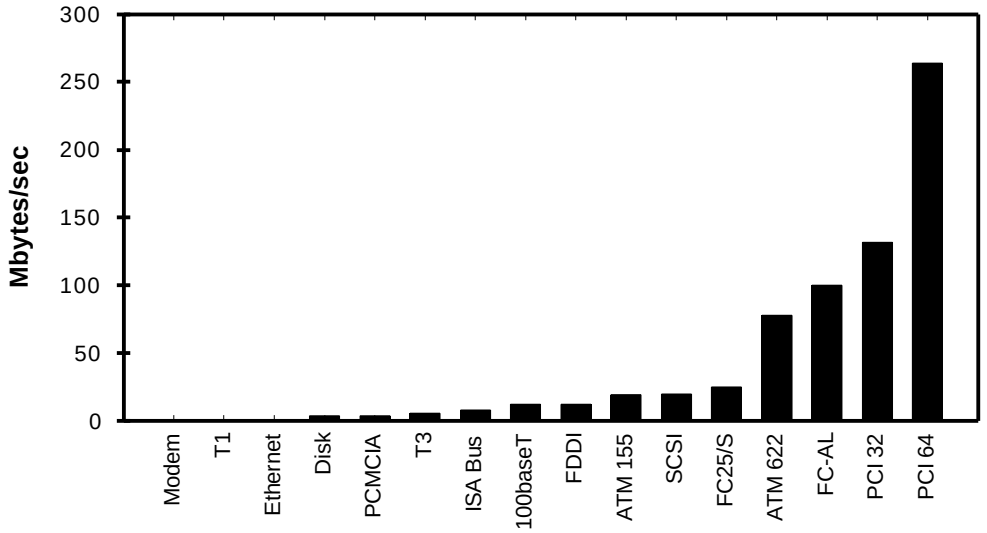
Figure 1. The virtual storage hierarchy (from Wong 1997).



A complete description of LAN topologies, media and types is beyond the scope of this chapter, but it is important to briefly describe the models that have become standard. Ethernet (10 Mbps) over twisted pair wiring is a dominant model in modern LANs. In fact, over 80 percent of all network connections were Ethernet by the end of 1996. Star topologies are commonly designed into buildings where the media can be easily reconfigured to centralized hardware. Token Ring persists in IBM environments, and FDDI and (increasingly) ATM provide backbone capabilities. Heightened demand for bandwidth has resulted in 100 Mbps ethernet to the desktop and even gigabit (1 Gbps) ethernet is slated for standards adoption [IEEE 802.3z] in early 1998 *(http://www.gigabit-ethernet.org).* Just as important is the transition to switched technology, which is rapidly replacing older shared ethernet segments, providing dedicated bandwidth improvements of an order of magnitude.

Although the network is increasingly critical to collaborative data access and processing, it is also the slowest part of client/server transactions (Figure 2). This is especially true in WANs, but even on LANs, data transfer is slow relative to the internal components of the computer. The situation is improving quickly, however, with gigabit ethernet, Fiber Channel Arbitrated Loop [FC-AL] *(http://www.fiberchannel.com)* peripheral interfaces (rapidly replacing SCSI), and PCI buses with dramatically increased throughput. If transfer rates (bandwidth) are represented graphically in common units of measure, the comparison is dramatic (Figure 2).

Figure 2. Bandwidth vs. component expressed in common units illustrating the relatively slow performance of the network.



Rationale for a client/server architecture

Given the constraints of network speeds and the costs involved in providing bandwidth and processing enhancements, the motivation to use client/server solutions may be obscure to many people. There are many reasons for the success of client/server as the dominant computing paradigm of our time, most of which derive from the economics of computers and the universal (corporate) requirement for integration and collaboration of many people into large projects. The business and technical incentives that favor client/server may be tangential to research information management concerns, but the application of client/server solutions to infrastructure needs in our domain is inevitable.

The benefits of client/server computing are simple to understand. Primary among these strengths is scalability: i.e., the ability to enhance or reconfigure components of the client/server architecture simply and in proportion to need. Scaleable designs are very important, both economically and for performance tuning and problem resolution. The fact that client computers have powerful processors is key to many client/server implementations and contributes to the scalability of the system by dedicating significant processing capabilities to the user. While the

basic role of the client computer is understood to mean providing the computer to appropriately address user requirements, the optimal size and description of the client is hotly debated.

The ‘thin client’ as represented by the Net Computer Initiative, proposes to simplify client hardware and configuration to provide basic network access. This trend is premised on the increasingly central role of the server to Intranets via Web paradigms and the potential of computing platforms such as Java™. The criticism of traditional PCs is that they are difficult and expensive to manage and are inappropriate for many users who are mainly ‘data consumers’. The lack of early adoption of the Net Computer as a PC replacement is at least in part because of Microsoft’s antagonism to the model, until very recently. Even Microsoft, through its development of a multi-user version of NTServer - [Windows-Based Terminal Server - *Hydra*] is addressing the need to provide client/server computing to thin clients and aging, underpowered PCs, albeit in their characteristically proprietary manner. There are at least two end points on the spectrum of client computer evolution. The traditional PC is now characterised as a ‘Fat’ client while the Net Computer typifies the ‘thin’ client. Realistically, the entire spectrum will be represented for the foreseeable future.

Another consequence of the asymmetry of client/server design is that the client becomes increasingly generic and interchangeable while the server is managed for high availability. The partitioning of logic between the client and server is inherently flexible and further enhances the scalability of these designs. While the economics of computer system design dictate that it is much easier and less expensive to build 100 small computers than a single system that is 100 times as powerful, consolidation of processing on servers does occur to provide specialized functions. Specifically, fileserver, DBMS and computational server functions have been traditional to client/server, but increasingly, Internet/Intranet/Web functions are becoming important (e.g. http, email, ftp, usenet and firewall servers).

BANDWIDTH AND BOTTLENECK AVOIDANCE

Chief among issues which must be addressed in the evolution of the LAN are balancing data transfer demands on the network, in other words, how to size the ‘Network Plumbing’. Managing the growth of bandwidth demands is complicated by the fact that is quite difficult to predict future need based on historical data. The introduction of http has further exacerbated the situation by accelerating the rate of growth and introducing even more extremes in use patterns. It is undisputed that bandwidth demand will continue to grow at an accelerated rate so sizing solutions based on available and affordable technology must be balanced against predictions of availability of future cheaper technology. Sub-netting and segmentation are useful in managing traffic by isolating data to only those portions of the network where they are needed. Repeaters, bridges, routers and switches can be strategically implemented to achieve the necessary grouping of traffic. Bandwidth vs. latency (response time a requestor spends waiting for a result; Wang 1997) is an important distinction for planners to keep in mind because increases in bandwidth may not necessarily have the concomitant reduction in latency that users require.

A Redundant Array of Inexpensive Disks (RAID) is an important component of the data storage and transfer equation providing centralized, fast, fault tolerant disk space. RAID has been used to provide access to large amounts of data storage arranged on multiple physical devices. Large disks by themselves are not necessarily good solutions for providing access to large filesystems or databases. This is because the bandwidth available to an individual disk is not usually proportional to the size of the disk. Therefore, providing more, smaller component disks, with data striped across them and increased aggregate bandwidth is what RAID can accomplish. RAID devices have been largely responsible for the rapid adoption of the Fiber Channel Arbitrated Loop *(http://www.fiberchannel.com)* interconnect, which is replacing SCSI for data intensive

peripherals and is poised as a future network transport technology. Data redundancy and fault tolerance are also accomplished through RAID via mirroring [RAID level 1] or parity calculations [RAID level 5]. Software and hardware RAID solutions can be implemented alone or hybridized.

BACKUP AND ARCHIVAL

Data backup and archival are central to data management in the ecological research environment. The multitude of tape drive formats available presents a confusing array of choices including 8mm, DLT, DAT, and QIC. The highest volume/speed backup devices are now being built around DLT drives and 8mm drives to a lesser extent.

Table 1. Popular tape formats, 1997.

Format	Capacity	Transfer Rate
QIC/Travan	<1GB - 8GB	≤1MB/s
8mm	2GB – 125GB	1 – 6MB/s
DAT (4mm)	4GB – 24 GB	<1—3MB/s
DLT	10 – 100GB	1.5 – 10 MB/s

Future tape library migration path is at least as important a consideration as the technical attributes of an individual format. Robotic tape library devices are necessary for unattended backup of large data repositories. Care should be taken with the location of the tape library including plans for redundant storage, to ensure disaster recovery. Tape servers should be in secured locations and should be stable, competent computers.

Archival means different things in various environments. The exponential growth in data volume has further blurred the distinction between backup and archival. In particular, the shift from predominantly character data to object data, especially images, has largely been driving the massive increase in total volume. Images are large and in many cases are prime candidates for migration to archival media. The access to archive media can be provided via on-line, near-line or off-line solutions, using jukebox technologies and hierarchical storage software, or manual management.

OTHER CONSIDERATIONS AND REMOTE SITE SPECIFICS

Printers, plotters, film recorders, etc. should be shared among workgroups and the incentive to do this is both budgetary and practical, i.e. to simplify administration. Providing the printer with a network interface and queuing it from a capable print server will allow for relatively flawless printer access. High quality printing capabilities (high resolution, PostScript, etc.) can then be costed across many staff members and projects.

Fault tolerance and redundancy are qualities that should be prioritized relative to the institutional dependence on the equipment or service. For example, if a fileserver holds all research data, the availability of those data is critical to daily operation. Steps should be taken to ensure a level of fault tolerance that can be justified in the budget and in relation to all other priorities. These two attributes, fault tolerance and redundancy, are essentially two sides of the same coin from the hardware perspective. They can take the form of dual power supplies, mirrored disk systems, fail-over network paths, etc. Service contracts and spare-parts agreements can ameliorate the cost of redundancy. The relative merit of these approaches is highly site-specific and cost/benefit analyses are quite difficult. The need for UPS protection can not be overstated,

especially in sites prone to brown power. Surge suppression on network cabling should not be overlooked either, as lightning can wreak havoc on twisted-pair or any non-inert media.

Table 2. LAN reliability needs assessment (MTBF = mean time before failure).

RELIABILITY VALUE = COST OF DOWNTIME X SYSTEM MTBF X SITE RISK PROBABILITY
COST OF DOWNTIME = (SYSTEM TIME VALUE X MEAN TIME TO REPAIR) + COST TO REPAIR

In the realm of field computers, it can be categorically stated that traditional laptops are ill-suited to dirty field work. There are many varieties of Electronic Data Recorders or ruggedized handheld PC's which can be exposed to water, dirt, etc. The cost of these units may be higher than a standard laptop, but the life expectancy in extreme conditions can be significantly longer. Inexpensive palm top computers can serve some purposes, but are notably fragile and have cramped keyboards.

Because of the remote location of many field stations, other logistic and budgetary considerations contribute to hardware specification and network design. The cost of telecommunications connectivity, especially for leased digital services, are distance-sensitive. This means that there may be strong motivation to merge voice and data over the same line, a T1 for example, in order to leverage the investment for multiple functions. The investment in multiplexing/channel bank hardware to do this can often be recouped in a year or less. Expect radical changes in this arena over the next couple of years.

Budgetary limits are usually a major implementation constraint in research. Fractured budget sources and authority can further exacerbate the inability to execute information technology strategy and evolution. The cost of computing is often not fully integrated into planning, proposal development and budgeting in research projects or institutions. It is critical to bear in mind the total cost of ownership of technology, and not just the acquisition cost of hardware and software. Hardware and software management, system administration, user support, design planning, and alternatives analysis are all personnel costs which must be factored in the total cost of ownership equation. Balancing these demands for resources against the cost of research and operations is a challenge, but the necessity of information technology is now indisputable. Computer hardware provides technological underpinnings for virtually every aspect of research, communication and publishing, as we currently know it.

LITERATURE CITED

Internet Engineering Task Force. 1998. *http://www.ietf.org*

Microsoft, Inc. 1996. Microsoft Windows NT Server Resource Kit. Microsoft Press, Redmond, WA.

Ryan, H.W., Sargent, S.R., Boudreau, T.M., Arvantis, Y.S., Taylor, S.J., Mindrum, C. 1998. Practical guide to client/server computing. Auerbach, Boston, MA.

Stone, J.P. 1997. Handbook of local area networks. Auerbach, Boston, MA.

Wong, B.L. 1997. Configuration and capacity planning for Solaris servers. Sun Microsystems Press, Mountain View, CA.