

Storage Networking As I Understand It

OVERVIEW	4
SUMMARY.....	5
INTRODUCTION	6
SEMANTICS.....	6
HISTORY.....	6
<i>Networks: LANs and SANs.....</i>	6
<i>Sharing.....</i>	6
<i>Internal and External.....</i>	7
<i>Block-level Access vs File-level Access.....</i>	7
CHRONOLOGY OF BLOCK ACCESS	8
ESCON.....	8
SCSI	8
ATA	8
Review.....	9
Fibre Channel.....	9
Review.....	9
Naming Confusion Redux.....	9
URLs	10
CHRONOLOGY OF FILE ACCESS	10
NFS	10
NCP.....	10
PCLAN/SMB	10
Lantastic.....	10
AFP	10
CHRONOLOGY OF STORAGE MANAGEMENT	11
STORAGE ARCHITECTURES	12
DIRECT ATTACHED STORAGE.....	12
NETWORK ATTACHED STORAGE	12
STORAGE AREA NETWORK.....	12
NAS HEAD + SAN	13
WHERE'S THE BOTTLENECK?	13
PHILOSOPHY	15
ORGANIZATIONAL ISSUES.....	15
DIFFERENCES AND SIMILARITIES.....	16
<i>Ethernet and Fibre Channel.....</i>	16
<i>LANs and SANs.....</i>	17
<i>Operating Systems are Primitive.....</i>	17
THE DRIVER FOR STORAGE MANAGEMENT	17
THE KEY TO STORAGE MANAGEMENT SUCCESS	18
BUSINESS ISSUES	18
WHY FIBRE CHANNEL	18
WHY FIBRE CHANNEL SANs.....	19
WHY ETHERNET	19
WHY IP.....	19
WHY IP SANs	19
MARKET POSITIONING.....	20

DAS	20
Fibre Channel	20
NAS	20
iSCSI	21
Silos.....	21
TECHNOLOGIES AND MODELS.....	21
SHARING.....	21
HIGH-LEVEL VIEWS	22
I/O Trends	22
Physical Transport Networks.....	23
Storage Protocols.....	23
Protocol Details	23
A Few Words on InfiniBand.....	23
Connectivity Tinker Toys.....	23
Tiered Model.....	24
Data Management.....	24
BASIC DIAGRAMS	24
Direct-Attached Storage (DAS).....	24
Network-Attached Storage (NAS).....	24
Storage Area Network (SAN)	25
NAS Heads	25
Review	25
Technical Differences: FC and Ethernet/IP.....	26
LANGUAGE AND LEGENDS	29
BASICS.....	29
SAME WORDS, DIFFERENT MEANINGS ... DIFFERENT WORDS, SAME MEANINGS.....	29
Fibre Channel	29
Ethernet/IP.....	31
Common Semantics	31
Address Space Comparison.....	31
MYTHS	31
Fibre Channel is complicated.....	32
Fibre Channel NICs Cost More than Ethernet NICs	32
Fibre Channel is More Expensive than Ethernet	32
We Migrate to SANs to Increase Performance and Reliability.....	32
Fibre Channel is Faster than Ethernet	33
Fibre Channel Switches Cost More than Ethernet Switches.....	34
Glass is More Reliable than Copper.....	34
Fibre Channel Require Single-Mode Transceivers.....	35
Cut-Through vs Store-And-Forward.....	35
Hardware Forwarding vs Software Forwarding	35
Fibre Channel Has Half the Overhead of Ethernet/IP.....	36
Fibre Channel is Proprietary; Ethernet/IP is Open.....	36
Microseconds Matter	36
My Application Requires Fibre Channel	37
STORAGE APPLICATIONS.....	37
BACKUP & RECOVERY	37
CONTINUITY MANAGEMENT.....	37
HIGH AVAILABILITY	37
PERFORMANCE SCALABILITY	38
INFORMATION LIFECYCLE MANAGEMENT	38
THE ANATOMY OF A SAN	38

VIRTUAL SCSI CABLES	38
PARTITIONING	39
CHOOSING STORAGE SOLUTIONS.....	39
DESIGN PROCESS	39
<i>Guidelines</i>	39
<i>Outline</i>	39
TYPICAL STORAGE REQUIREMENTS.....	40
TECHNOLOGY CHOICES	40
APPENDIX	40
FUTURE TRENDS	40
<i>FC-Base T</i>	40
<i>FC over Ethernet</i>	41
<i>FC over Converged Enhanced Ethernet</i>	41
DISK ENGINEERING	41
REAL WORLD NUMBERS	41
LAB NUMBERS.....	42
THE CASE FOR FIBRE CHANNEL SANS	42
<i>Existing FC Networks</i>	42
<i>High-End Performance</i>	43
<i>Unreliable or Overloaded LAN</i>	43
THE CASE FOR iSCSI SANS	43
<i>Existing Ethernet/IP Networks</i>	43
DESIGNING iSCSI SANS.....	43
<i>Host-Based Routing</i>	44
<i>Microsoft iSCSI Initiator</i>	44
<i>The Case for a Private V-Net</i>	44
<i>The Case for Commodity Mingling</i>	47
<i>The Case for Server-Wide V-Net</i>	48
<i>The Case for Function-Specific Server Rooms</i>	48

OVERVIEW

This document outlines my understanding of storage networking. I acquired most of my original understanding from attending two of Howard Goldstein's Storage Networking Solutions workshops, presented at Interop/Las Vegas 2005.

- W915 Storage Networking Solutions: Leveraging NAS and SAN for Backup & Recovery, Data Sharing, Virtualization, and High Availability
- W911 Storage Networking Technologies, Design, And Performance: Making the right technology choices, Fibre Channel, IP Storage, SAS, and SATA

And since attending these seminars, I've been doing my own reading, attending additional seminars, and talking with vendors, to broaden and deepen my understanding.

Briefly, storage is a rich, complex field. Stepping up to this subject, I thought "how hard can this be? We're just talking about reading and writing files from/to a disk". In fact, though, the storage networking world has been growing in parallel for as long, if not longer, than the data networking world (though not as long as the voice and video networking worlds). Just as voice, video, and data have their richness and complexities, so too does storage. Just as voice, video, and data are gradually converging onto the single network, so too is storage. And, as each of these technologies merges toward the single network, that network changes to accommodate their unique requirements.

I've become a Howard Goldstein fan -- I learned a lot during these two days, Howard's speaking style meshes tightly with my learning style. Howard can go broad: business process, market analysis, human factors, and philosophical implications. And he can go deep: the protocol details of the latest SCSI transport scheme. Mixing broad and deep contributes heavily to my ability to grasp a new subject, and that's how I advanced my understanding of this subject so significantly during these workshops.

Caveats:

- I live in the open systems world; I know nothing about the z-Series (mainframe) environment.
- I work for a small company (less than 5000 employees) with low-end storage requirements (less than a hundred terabytes of backed up storage, no OLTP, minimal business-induced performance requirements). As a result, I know little about the high-end section of this space.
- I come from an Ethernet/IP and open source background – I am biased in favor of designs, technologies, and cultures which promote transparency.
- I have listened to more people with IEEE/IETF leanings than I have to people with ANSI leanings (meaning: I've heard more about iSCSI than I have about Fibre Channel).
- I know little about the large, complex, and rich fields of *storage management* and *server virtualization*. These are important, and difficult, subjects. But I don't know much about them. In this document, I focus on the smaller and simpler world of *storage networking*, i.e. transport for SCSI frames.

- I don't claim to be an expert on this subject (though I do claim to have listened to experts). If you detect inaccuracy or misinterpretation or oversight, please let me know.

SUMMARY

From a technologist's point of view, every form of host/storage connectivity can be described as a Storage Area Network: the terms DAS (Direct Attach Storage), NAS (Network Attached Storage), and SAN (Storage Area Network) are more a result of marketing strategies aimed at differentiating competitors than anything inherent in the technologies deployed. In general, beware of semantic distortion – IT suffers from this problem generally, and the storage arena is no exception. Having said that, I describe DAS and SAN as providing block-level access to storage, while NAS provides file-level access.

Networks specialize in the sharing of resources; Storage Area Networks are merely networks on which the customer intends to share storage. In other words, Storage Area Networking is about **asset utilization** ... allowing the customer to better utilize the space s/he owns. This precisely parallels what the customer is likely doing for other expensive assets, like printers ... attaching them to a network in order to push utilization toward 100%, making the very best use of an investment. In particular, Storage Area Networking is not about performance. If performance is your primary concern, stick to direct-attached SCSI Ultra 4: it is simpler and blows away Fibre Channel.

Storage attachment strategies are migrating from parallel to serial, resulting in the ability to move disk farther from its associated host.¹ Performance and reliability are improving. Differences between ATA and SCSI disks are shrinking (not vanishing, just shrinking). At one time, Fibre Channel was the primary game in town; these days, Serial Attached SCSI is challenging Fibre Channel and IP SANs at the low-end; while InfiniBand challenges from the high-end: imagine transferring a terabyte of data from one host to another while incurring a single interrupt.

The storage challenge with which most companies are currently wrestling is information life cycle management (ILM) – how do you keep track of your data, delete what you don't need, archive what you can, replicate the critical stuff, and what do you do about tapes anyway?

From a high level, the driver for SAN design are the *information flow requirements* of the applications hosted on the SAN.

¹ From parallel to serial to increase distance, and to parallelized serial streams, to increase throughput.

INTRODUCTION

Semantics

Human languages tend to employ inconsistency (different words meaning the same thing) and overloading (the same word meaning different things). Technical lingo copies this pattern. And the lingo around storage follows this trend, too. Sometimes, storage technology can seem confusing. When you feel confused, start looking for inconsistent or overloaded terminology.

Terminology confusion becomes a marketing strategy for vendors, particularly when they are trying to differentiate themselves from competitors or when they are trying to present their product as more closely meeting your needs than that of a competitor.

Remember that vendors' primary goal is to sell you what they have ... whereas your primary goal is to buy what you need ... these two goals are not always aligned ... and the strategies employed to meet these two goals sometimes conflict with one another.

Beware of Marchitecture ... Marchitecture is Architecture mangled by a sales teams' desire to sell you something ... typically looks like Architecture, only no matter how you slice it, the 'right' answer seems to be to purchase whatever product the vendor is selling.

History

NETWORKS: LANS AND SANS

Q: What was the original purpose of the network?

A: To share printers. Printers cost a lot, and sharing one printer amongst many people was and is a way to reduce costs. More generally, the original purpose of networks was to share resources, leveraging the value of a capital investment.

Q: What is the purpose of the network today?

A: Primarily sharing information, though for our purposes today, talking about storage, we will focus on the older purpose of sharing resources.

SHARING

Storage Area Networks are networks whose ostensible purpose is to share storage. As we will see, many implementations of this technologies don't do a good job of sharing² ... they tend to restrict access to a given chunk of storage to a single host. And most if not all implementations of Storage Area Networks permit the sharing of far more than just storage: in fact, Storage Area Network standards rather explicitly describe generalized networks designed to facilitate the sharing of a range of resources.

In some senses, the LAN/SAN dichotomy is a fraud: LANs and SANs are the same thing.

² There are excellent reasons why most Storage Area Networks don't do a good job of sharing! Read on for the explanation.

Both promote the sharing of resources ... the differences arise more from the desire of marketing people to differentiate their products from those of their competitors than from anything inherent in the technical choices involved. In the future, we may see marketing people grabbing onto something else as their lingo-of-choice for differentiation ... if that happens, we may then see one or the other term (probably 'SAN') fade away.

As a secondary benefit, SANs are sometimes used to off-load traffic from LANs ... since backups, and their sister, remote mirroring, tend to be the biggest hogs of LAN bandwidth, customers sometimes implement a parallel network to carry all this data and to reduce the chance of interfering with end-user activity ... typically, we call this second network a SAN. However, even when a customer re-uses the current LAN for storage traffic (something which we as an industry have been doing ever since Sun shipped NFS and Novell shipped Netware), we still call this a SAN ... a SAN overlaid on top of a LAN.

INTERNAL AND EXTERNAL

Technically speaking, there are two types of SANs: internal and external. Internal SANs typically have a range of, at most, a few meters and typically exist inside a single box. The original SCSI Bus is an example of an internal SAN. RAID controllers are another example of internal SANs (typically implemented as Fibre Channel arbitrated loops attached to SCSI disks ... perhaps 8-12 loops with each loop supporting 15 drives). 'External SANs' may extend throughout a cabinet, throughout a room, or in fact across the planet. For the most part, in this document, I skip over 'internal SANs' and focus on 'external SANs' – unless I specify otherwise, any mention of the word 'SAN' refers to an 'external SAN'.

External SANs come in a variety of flavors. Here are a few:

<u>Name</u>	<u>Description</u>
FC SAN	Fibre Channel network built in parallel with the LAN, carrying block traffic
IP SAN	Ethernet/IP networks built in parallel with the LAN, carrying file (NAS) and/or block (iSCSI) traffic
Overlaid IP SAN	Ethernet/IP file (NAS) and/or block (iSCSI) traffic overlaid on LAN, sometimes using existing NICs, sometimes using additional NICs

BLOCK-LEVEL ACCESS VS FILE-LEVEL ACCESS

Some applications want block-level access to disk. For example, high-end database management systems (Oracle, Sybase, SQL Server) prefer to read and write database records via SCSI (or ATA) block commands, rather than by going through the OS' file system.³

Some applications want file-level access to disk. For example, Microsoft Word doesn't want to think about SCSI ... it wants to issue a FILE OPEN command and just receive the entire document back.

³ However, some customers are exploring the use of NFS as a transport for database transactions – when NFS performance to a storage array meets or exceeds block-level performance, NFS becomes attractive: easier to manage and just as fast. And of course, for low-end requirements, customers employ even CIFS as a transport for database transactions: again, easier to manage than LUNs and 'good enough' in terms of performance.

Generalization: most applications want file-level access to disk. Operating systems and database managers are notable exceptions.

Chronology of Block Access

Aka 'A History of Confusing Naming Schemes'. See the *Block Level Protocol Stacks* diagram (<http://www.skendric.com/sans/block-level-protocol-stacks.pdf>).

ESCON

In the beginning, there was IBM. And life was good. Mainframes used a physical transport protocol called "byte and tag" to communicate with disks. From the primordial soup of "byte and tag" was born ESCON (Enterprise Systems Connection), an optical-only physical transport protocol employing the 8B/10B encoding scheme and interconnecting IBM mainframes with each other and with a variety of peripherals.⁴ Regrettably, the protocol implemented on top of ESCON was also called ESCON ... leading to a confusion of terminology which has continued to this day.⁵ ESCON (the upper layer protocol) contained exactly five commands.

SCSI

In the 1979, Larry Boucher (founder of Adaptec) lead the team which developed SCSI Bus (Small Computer System Interface)⁶, a physical transport protocol aimed at connecting hosts and peripherals ... an 'ESCON for the rest of us', if you like. Regrettably, the SCSI people called the protocol which rode on top of this physical transport 'SCSI' also ... thus perpetuating the naming problem which ESCON started. SCSI is way more complex than ESCON -- but then again, it works with all sorts of hosts and all sorts of peripherals from all sorts of vendors, not just with IBM big iron. The SCSI spec contains upwards of 2000 commands. ANSI ended up owning the SCSI development process.

ATA

In the 1980s, IBM developed ATA (Advanced Technology Attachment), a subset of SCSI intended to facilitate the production of cheap disk, for their new PC AT computer. Controllers implementing ATA tend to be cheaper than controllers implementing the full SCSI command set. Disks whose controllers speak ATA (later called "IDE disks", once the ATA controller was glued onto the disk itself) tend to be produced using manufacturing processes which are less rigorous than those used when making SCSI disks ... with the result being disk which costs less and fails sooner.⁷ Regrettably, the ATA people perpetuated the naming confusion: they call the physical transport network 'ATA' (a newer version of this transport network is called 'Serial

⁴ Kind of like a precursor to USB, a protocol which allows us to attach printers and disks to microcomputers.

⁵ The term 'ESCON' refers to the whole protocol stack supporting block transport ... this is kind of like calling both Ethernet and TCP/IP "Ethernet" ... or calling both John Doe and his car "Toyota Corolla".

⁶ "Small" because it addressed the needs of little machines ... machines which didn't require an entire room complete with tons of AC and filtered power in which to live.

⁷ An audience member claimed that vendors are moving toward producing all disks using the same manufacturing processes and only at the end gluing an ATA or SCSI controller chip onto them.

ATA') ... *and* they call the protocol riding on top of this (the subset of SCSI), 'ATA' also. ANSI ended up owning the ATA development process.

REVIEW

Notice that at this point we have three storage control protocols (aka three ways to read and write blocks from and to disks): ESCON, SCSI, and ATA. And we have three transport protocols: ESCON, SCSI Bus, and ATA.⁸ The storage control protocols are tightly tied to the transport protocols ... i.e. if you buy a SCSI disk, you gotta buy a SCSI adapter to go with it: you can't mix and match.

FIBRE CHANNEL

In 1994, ANSI ratified the Fibre Channel protocol. This is an optical-only physical transport protocol designed to interconnect hosts and peripherals. Its development was influenced heavily by ex-mainframe storage weenies from IBM. In some sense, it looks and feels like ESCON version 2. However, it disentangles the physical transport protocol from the upper layer protocol riding on top of it: one can run ESCON (the storage control protocol, not the physical transport protocol) on top of Fibre Channel (aka FICON); one can run SCSI (the storage control protocol, not the physical transport protocol) on top of Fibre Channel (aka SCSI/FC), and one can run ATA on top of Fibre Channel (aka SATA/FC). For that matter, one can run IP on top of Fibre Channel, too. Conceptually, Fibre Channel could have functioned as a competitor to Ethernet/IP, in the same way that ATM competed against Ethernet/IP.

REVIEW

Notice that at this point we have introduced a fourth transport protocol without introducing a fourth storage control protocol, i.e. one can reasonably call a disk an 'ESCON disk' or a 'SCSI disk' or an 'ATA disk' ... but there is no such thing as a 'Fibre Channel disk'. Fibre Channel is a transport protocol, not a storage control protocol; in the market place, it transports frames containing ESCON commands ... frames containing SCSI commands ... and, in some places, frames containing IP commands.

NAMING CONFUSION REDUX

At last, we have a chance to achieve a coherent naming scheme ... but did we take advantage of this opportunity? No, we did not. In common parlance, you will hear people talk about "Fibre Channel disks": vendors do this, customers do this, industry pundits do this. If you are in the mainframe world, you mean ESCON disks with Fibre Channel interfaces. And if you are in the open systems world, you mean SCSI disks with Fibre Channel interfaces. To be fair, what vendors are trying to express, by using this terminology, is how the manufacturing processes and drive engineering they use to build SCSI disks with Fibre Channel interfaces are superior than

⁸ Regrettably, the names of the disks (ESCON, SCSI, and ATA) have been re-used for the names of the transport protocols (ESCON, SCSI Bus, and ATA) ... in polite company, I call this *overloading*; amongst friends, I call this *confusing*.

those used to build, say, SATA disks with Fibre Channel interfaces ... aimed at a different market, with higher reliability and performance requirements and willing to bear a higher cost as a result.

URLS

SCSI <http://www.t10.org>

Fibre Channel <http://www.t11.org>

ATA <http://www.t13.org>

Chronology of File Access

File-oriented protocols developed in the 1980s -- by 1986 (AFP), all these protocols were in place and widely deployed. See the *File Level Protocol Stacks* diagram (<http://www.skendric.com/sans/file-level-protocol-stacks.pdf>).

NFS

In the beginning of networked microcomputers, there was Sun. And life was good. Sun invented NFS (Network File System), and microcomputers could then share files with each other.

NCP

Novell developed NCP (Netware Core Protocol), and those geeks whose parents didn't let them play with Unix could now share files.

PCLAN/SMB

IBM and Microsoft developed PCLAN, which evolved into SMB (Server Message Block), so now mainframe geeks whose parents made them mess with these dumb little microcomputers could share files.

LANTASTIC

Artisoft promoted peer-to-peer file sharing (no centralized server -- every workstation shared its disk), so penny-pinchers could share files.

AFP

Apple followed with AFP (Apple Filing Protocol), because the rest of us wanted to share files, too.

Chronology of Storage Management

- In the beginning (of open systems), we had lots of microcomputers with their own disks ... connected to a network to share printers.
- We added a file server. The microcomputers still had their own disks ... but we tried to persuade people to save their files to the server's disk.
- We interconnected these departmental LANs into a big corporate LAN ... accelerating many-to-many connectivity.
- We centralized the administration of these departmental file servers ... and centralized IT began to stress about maintaining all these little servers.⁹
- So centralized IT started consolidating them, first into small clumps and later into bigger clumps (server farms, data centers). This step simplified management and administration and backups ... eliminated duplicate resources (keyboards, monitors) ... but storage was still dedicated to each server. Centralized IT spends more and more of its time taking servers down to add disk or shuffling groups of users from one server to another as they outgrow the storage located there.
- Data resides on enterprise servers, on departmental servers, and on clients. It may not be growing exponentially, but it is growing rapidly. The cost of managing this distributed storage starts to approach the cost of the devices themselves (some folks claim that the cost of managing the storage exceeds the cost of the storage itself ...)
- Then came Fibre Channel and the arrival of the era of the Storage Area Network. Multiple servers share the same storage device, which the administrator hacks into separate chunks and then dedicates to each server. The administrator can perform the following tasks, typically without taking down servers:
 - add or delete storage for a given server without touching physical disks
 - allow servers to automatically allocate additional space as their needs increase (from a pool of unallocated space)
 - share additional storage devices like tape and optical
 - add additional storage devices as space needs increase
- Storage networking has proved its worth, and now multiple technologies are tackling the problem, ranging from iSCSI and Serial Attached SCSI to InfiniBand, with players like 'Data Center Ethernet' on the horizon.

⁹ In grant-funded institutions, we tend to develop lots of centralized IT departments, reflecting the siloization of resources which our funding model creates. However, conceptually, the same pattern reoccurs -- each of these mini-centralized IT departments finds itself with more and more servers to manage.

STORAGE ARCHITECTURES

These are four of the major storage architectures, from a plumbing perspective. See the *DAS & NAS Architecture* (<http://www.skendric.com/san/das-nas-arch.pdf>) and *SAN Flavors* (<http://www.skendric.com/san/san-flavors.pdf>) diagrams.

Direct Attached Storage

This solution provides applications loaded on *one and only one* server *block-level* access to storage.

Definition

- One server running a general purpose operating system attached via a physical SCSI cable to one disk. Generally, adding disk or performing some disk/volume mappings requires down time.

Variations

- Multiple SCSI cables going to multiple disk arrays ... but still attached to a single server.
- Multiple (well ... a maximum of two!) servers attached to the storage devices via the magic of dual-ported SCSI -- popular in clustered or highly-available solutions.

Ignoring InfiniBand for the moment, this is the current performance leader in the storage world: dual-attached servers to a RAID 10 disk array equipped with 320MB/s disks can outperform even high-end Fibre Channel SAN solutions.

Network Attached Storage

This solution provides applications loaded on *multiple* servers *file-level* access to a *single* storage device.

Definition

- A server equipped with a special-purpose operating system providing NFS and/or SMB file services to other servers and directly to clients ... and nothing else. The internal disk is typically attached via SCSI or Fibre Channel. Generally, adding disk or performing some disk/volume mappings does not require downtime.

Variations

- Low-end solutions ship as appliances, often consisting of general purpose operating systems hacked to permit administrative access only via a custom Web interface and lacking 'zero downtime' features.
- High-end vendors bundle a SAN with their NAS ... creating a NAS Head (see below).

Storage Area Network

This solution provides applications loaded on *multiple* servers *block-level* access to *multiple* storage devices.

Definition

- Multiple servers and multiple storage devices attached to a dedicated network carrying block-level traffic (SCSI).

Variations

- The advent of iSCSI permits the previously dedicated network to be overlaid on top of the enterprise's commodity IP network.

In the 1990s, one built a parallel network (Fibre Channel) to carry SCSI traffic ... because no standards existed for carrying SCSI frames over existing Ethernet/IP networks, and because no one wanted to rip out their existing Ethernet/IP networks and replace them with Fibre Channel networks. And parallel storage networks remain a popular choice today, particularly amongst customers who have existing Fibre Channel networks ... these are a sunk cost ... so customers want to leverage this asset, adding additional Fibre Channel nodes as needed, or perhaps adding Ethernet/IP nodes to this parallel network.¹⁰

NAS Head + SAN

This solution provides applications loaded on *multiple* servers *block-level* and/or *file-level* access to storage.

Definition

- A NAS box plugged into a SAN back-end.

Typically, these solutions offer disk access via multiple transports, one or more (or all!) of NFS, CIFS, iSCSI, Fibre Channel.

Where's the Bottleneck?

Modern IT systems (clients, servers, and the networks which sit between them) are ever advancing in terms of capacity (CPU, memory, bus throughput, bandwidth, I/O processing, etc.) But these components of the entire system tend not to advance in lockstep with each other. At any moment in time, some are ahead of the pack, others trail.

Claim: At the moment, IT systems tend to have more CPU and bandwidth than they need. Vendors produce CPUs with ever higher clock speeds and motherboards with gigabit NICs in them (or switches with gigabit ports) because they can do so cheaply, not because customers need them.

¹⁰ Some vendors produces switches which sport both Fibre Channel and Ethernet (iSCSI) ports.

Claim: At the moment, disks performing random I/O go slower than memory and CPUs and networking pipes.¹¹

Claim: Analyzing storage requirements is hard. Most of us don't even know what questions to ask about storage, let alone what our particular answers would be.

Observation: Many customers, when they want to figure out how "fast" they ought to be going don't ask what their users and applications need; rather, they look at the labels on their disks (80MB/s, 160MB/s, 320MB/s) and NICs (100MB, 1000MB, 2000MB ...) and CPUs (2.0 GHz, 2.4, GHz, 2.8 GHz ...) and buy the ones with the biggest numbers. Kind of like going to Costco and buying the biggest containers of everything ... flour ... beans ... anchovies ... cayenne pepper ... toothbrushes.¹²

Analysis: Most IT systems can't even begin to touch the best-case throughput of their disks ... their data utilization patterns tend to involve random I/O, for example, instead of sequential I/O (disk throughput numbers like 80MB/s depend on sequential I/O) ... their system buses max out at 66MB/s (PCI)¹³, their SCSI controllers can't do better than 30MB/s anyway ... and even if they could optimize the channel to their disks, their CPUs would be overwhelmed by the interrupts required to handle the transfer.¹⁴

Moral: Beware of vendor claims that you "need" faster CPU, faster memory, faster disk, or more network bandwidth. Many IT systems today simply don't need more performance.

In well-architected environments, what limits IT systems these days tend to be:

- Cooling
- Power
- Host bus performance
- Storage management (disk is scattered amongst enterprise servers, departmental servers, and clients)
- System management (most organizations wait until a system blows up before fixing it ... rather than proactively detecting the signs of failure and scheduling downtime)

¹¹ When performing sequential reads, modern SCSI disks can, in theory, outpace networking pipes substantially. But sequential reads tend to occur in laboratories, not in the real-world, where the demands on storage tend to be bursty ... and random.

¹² If the cost of the analysis is higher than the cost of the gear, then this strategy is actually pretty smart ... plus, it optimizes both the customer and the vendor experience (customers save money/time on analysis; vendors receive big purchase orders). ☺

¹³ Theoretical maximum for PCI-X is 150MB/s, which outta outstrip a gigabit Ethernet port, which can push 100 MB/s.

¹⁴ This is where InfiniBand shines, promising to deliver a file with a single interrupt, rather than an interrupt for every single 2K (or 1.5K) frame.

PHILOSOPHY

Organizational Issues

Storage Area Networks are just data transport networks with a fancy name and do best when managed by people with networking expertise, as far as design and operations go – storage is merely another application running over the network. However, many organizations struggle with this decision because of human factors:

- SANs typically represent a significant investment, and whenever significant resources are involved, humans find it easy to slip into power struggles
- SANs offer the potential for reducing head count, because of the staff efficiencies which come from consolidating and sharing space. Loss of head count can lead to human power struggles.
- Some SAN implementations involve technologies which lie outside the comfort zone of the respective players (networks for system administrators, Fibre Channel for network types); again, when humans leave their comfort zones, they can slip into power struggles.

The speaker recommends:

- Identifying the underlying concerns of interested parties, in an effort to defuse the human dramas being played out.
- Focusing on defining the applications involved, along with their requirements – after all, these are what are most important to the business.
- Realizing that the underlying SAN technology choices are just plumbing and once installed and configured tend to carry water just like any other plumbing ... they become less of a human flash point some months after installation, when compared to the early decision-making period.

The interesting and complex part of SANs lies in the applications they support; this area lies firmly in the domain of the system administrators.

The speaker sees effective use of SANs where organizations land their design and operation in a networking group. The pit-fall here occurs when that networking group and the system administration group don't collaborate well -- like LAN deployments, SAN deployments require tight coordination between the two technology areas.

He sees ineffective SAN experiences where organizations land their design and operation in the system administration group. However, he notes that some organizations happen to have system administrators who are network-savvy; he has seen particularly effective SAN deployments in which the SAN sits in the hands of these hybrid geeks, typically re-orged into a 'storage' sub-group within the system administration group or within the data center group.

For Fibre Channel in particular, deployments go better when the folks designing and implementing the SAN have a close working relationship with the system administration or data center group ... the speaker sees this collaborative approach in roughly half the organizations he visits. This because, in Fibre Channel designs, the vendors build many of the plumbing-related services into the Fibre Channel switches themselves ... compare this to the Ethernet/IP world, where plumbing services (like name services, dynamic address allocation, authentication,

directory services) tend to be implemented in outboard servers. Conceptually, Fibre Channel switches resemble LinkSys or NetGear multi-function boxes (switch, router, NAT gateway, DHCP server, VPN tunnel terminator).

In summary, the speaker recommends employing a suitable mix of sys admin, storage, and networking skills when designing, installing, and operating storage area networks. This is not a radical claim: you're already doing this (we hope!) in your current environment.

Differences and Similarities

ETHERNET AND FIBRE CHANNEL

From a technical point of view, Ethernet and Fibre Channel are highly similar: they both facilitate the sharing of resources, they both create a flexible, many-to-many environment, they both value performance and reliability ... heck, they both share the same encoding technique on the wire (8B/10B). In theory, they could replace one another ... they both offer similar feature sets ... Ethernet can carry SCSI frames inside IP frames; Fibre Channel can carry IP frames inside Fibre Channel frames.¹⁵

From a sociological point of view, Ethernet/IP and Fibre Channel were born and raised in different cultures. Ethernet/IP belongs to international standards processes involving a wide range of players. The IEEE and the IETF promote interoperability and openness; their stakeholders make money from the widest possible dissemination and implementation of the standards these groups develop. Fibre Channel¹⁶, on the other hand, belongs to a much smaller collection of players who, in general, make money through differentiation, integration, and bundling and don't accrue significant financial benefit from interoperability.¹⁷ With the arrival of iSCSI (block transport over Ethernet/IP) and Cisco (a company which grew up in the IEEE/IETF world), the Fibre Channel world's evolution has accelerated, achieving unprecedented degrees of interoperability as well as expanding and embracing competing technologies.

¹⁵ In practice, of course, Fibre Channel/IP has lost the larger race: its deployment is too small, its costs are too high, for it ever to replace Ethernet/IP. And many customers have so much invested in Fibre Channel that they are reluctant to replace it with Ethernet/IP, no matter the appealing economics. But from a technical point of view, both environments offer similar services: that's the point in this paragraph.

¹⁶ It's all a matter of perspective, of course. From the big iron point of view, Fibre Channel is wildly open ... unlike its ESCON parent, it works on more than one vendor's hardware. However, from an IEEE/IETF point of view ... Fibre Channel looks mostly proprietary.

¹⁷ I personally find these cultural differences particularly difficult to manage. I grew up in the IEEE/IETF way of thinking, reinforced by my succession of jobs in academic research environments, where openness and information dissemination is a requirement for success. In the IEEE/IETF world, I can ask the question "How does this work?" and expect to acquire an answer which goes as deep as I care to go. In the Fibre Channel world, by contrast, I'm not confident that many of the people with whom I've interacted grasp the "How does this work?" question ... nor am I confident that they see value in either asking that question. They certainly don't answer it. A desire for secrecy? Ignorance? A lack of curiosity? I know that I don't yet understand why the Fibre Channel world operates as it does.

LANS AND SANS

When actually implemented, LANs and SANs tend to diverge from one another. LANs prize many-to-many connectivity, and while security pressures are eroding this, the value which a LAN delivers tends to be proportionate to how many destinations a given node can reach.

On the other hand, SANs, when implemented, tend to offer many-to-many connectivity during their first few minutes of life ... and then their administrators shut down all connectivity ... and hack them up into one-to-one mappings ... because most operating systems still believe that they own a given storage device ... and if you allow two hosts to write to the same chunk of disk, they will stomp on each other's data ... few operating systems contain mechanisms for sharing storage. To use hyperbole for the moment, the first thing SAN installers do, after installing their SAN ... is to rip it out ... i.e. to chop it up so that a given host sees only the storage allocated to it and not anything else ... back to the original one host – one disk model. All this is not to denigrate the benefits of SANs – they offer lots of benefits ... but to point out that because modern operating systems are still in their infancy, as far as shared storage goes, SANs end up looking a little different than do LANs, when actually implemented.

OPERATING SYSTEMS ARE PRIMITIVE

To expand on the notions above ... in many ways, operating systems are currently the most primitive objects on a network today.

- They have limited if any knowledge of remote disks ... they tend to believe that they are writing to a local disk ... the network protocol stack running inside them performs magic to sustain that illusion (typically called 'network redirectors', because they redirect what the OS believes is a local disk write to a remote disk on a file server).
- They have few interfaces for requesting network-specific services, like QoS
- They offer few if any interfaces to applications for communicating information flow requirements (lazy writes/reads, hyper writes/reads, etc.)
- They are generally deaf to any unsolicited input coming from their attached devices ... they only listen if they have requested information.¹⁸

As storage networking matures, expect to see operating systems mature as well

The Driver for Storage Management

Storage Management exists because customers want more tools to manage space ... to better utilize this asset. (Customers tend to put data into that space ... but at a fundamental level, storage amounts to space, and space is what customers end up managing, whether it contains data or not.)

¹⁸ For example, when a host boots, its SCSI controller scans the SCSI bus for devices, and the OS asks for the resulting list and responds accordingly. Thereafter, the SCSI controller will typically scan the bus periodically, to see if device status has changed. However, few Oses will listen to the controller when it announces that it has discovered a new device ... many Oses require a reboot to respond to a change in SCSI device presence.

The Key to Storage Management Success

The key to storage management success is *virtualization*, aka *indirection*. When one host is attached to one disk, our choices are limited. But once we abstract the connection between host and disk, all sorts of capabilities arise:

- Transparent growth and shrinkage of space
- Transparent increases and reductions in performance
- Transparent changes in reliability
- Transparent increase or reduction in number of clients (typically hosts)

This is where storage area management technologies spend their energy and their complexity: implementing one (or more!) layers of indirection between the application and the disk.

This indirection can be implemented in the disk, in the host, in the network out-of-band, in the network-in-band ... there are many ways to do this.

BUSINESS ISSUES

Why Fibre Channel

Question: Why does Fibre Channel exist?

Answer: Because ESCON was running out of gas and was limited to the big iron world -- the FC designers wanted a beefier solution and one which would support the open-systems world.

Question: Why didn't the Fibre Channel crowd learn from the mistakes of the Ethernet/IP crowd?

Answer: They didn't think they were inventing a networking architecture ... they thought they were inventing a faster bus ... they weren't networking weenies ... they were mainframe weenies ... and that's why they reinvented the wheel, making the usual mistakes along the way.

Question: Why do multiple competing solutions exist? (Fibre Channel, iSCSI, InfiniBand ...)

Answer: Because of a mixture of vendors' desire to leverage their investments and customers' desire to leverage their investments. Both sets of players want to maximize the value they are extracting from what they currently own.

Additionally, when deciding where to locate an application, customers want to leverage current resources. For instance, if a customer owns a stressed-out LAN (minimal spare capacity) and an over-architected SAN (plenty of spare capacity), s/he will tend to locate a new application on the SAN rather than on the LAN. Remember, the primary business driver for SANs is **asset utilization**.

Different solutions exist because they have evolved in different eco-systems. Fibre Channel and Ethernet/IP have done just this ... only now, their two eco-systems are merging, and we are seeing the resulting conflict, as each competes with the other. The storage world is full of

emerging technologies, scrabbling for their corner of the market and meeting customer needs in different ways.

Why Fibre Channel SANs

- For a decade, Fibre Channel-based SCSI transport was the only game in town, the only way to create multi-host storage area networks
- Designed, built, and supported by people who specialized in storage

Why Ethernet

Ethernet survived the primordial soup of micro-computer networking¹⁹ (1980s) because it best exploited two key insights:

- Generally, things work – assume that the transmitted packet will arrive at its destination and get on with the business of sending the next packet.
- There are lots of end-stations: make them simple. There aren't many transport-only devices (hubs, switches, routers) – push complexity into them.
- Transmission over unshielded twisted pair works.

After that, market momentum did the rest.

To date, only WiFi and Fibre Channel have successfully challenged Ethernet's dominance.

Why IP

IP survived the primordial soup of networking²⁰ (1970s-1990s) because it best exploited three key insights:

- Generally, things work – assume that the transmitted packet will arrive at its destination and get on with the business of sending the next packet. In other words, don't invest effort up front in protecting against lost packets – wait until it happens and then recover.
- Physical media changes; the need for ubiquitous transport does not. Divorce yourself from the underlying media and thus position yourself to run over anything, from barbed wire to hard vacuum.
- "Rough consensus and running code" – don't wait to get things perfect; deploy early and often.

To date, only Fibre Channel has successfully resisted the IP steamroller.

Why IP SANs

- Ubiquitous transport
- Best of both worlds: greater connectivity while making fewer changes
- Extends distance to the WAN

¹⁹ RIP: OmniNet, ARCNet, LocalTalk/PhoneNET, Lantastic, Token Ring, StarLAN, ProNET, FDDI, ATM.

²⁰ RIP: DECNet, OSI, XNS, NCP/IPX, NetBEUI, AppleTalk, ATM.

- Customers have lots of IP expertise
- Customers feel intimidated by Fibre Channel
- Re-uses the currently installed packet transport infrastructure
- Supports a consistent management view

Market Positioning

- iSCSI will compete more with DAS (parallel SCSI)
- Fibre Channel will continue to grow, mostly in the high-end and with established customers
- iSCSI targeted at middle tier markets: small to medium sized SANs and anyone who hasn't taken the FC leap

DAS

Storage vendors like to portray DAS solutions as anachronistic, because they tend to be cheap, i.e. low-margin. Notice, however, that customers like cheap, and notice also that DAS solutions remain the highest performing solutions.

Positioning aside, DAS doesn't scale well and tends to be unsophisticated in its asset utilization abilities.

FIBRE CHANNEL

Fibre Channel vendors portray their products as enterprise solutions: high-performing, high-capacity, high-reliability. Their established customer base is the enterprise, and their products cost a lot – they don't have the volume advantage that IP SANs have.

Positioning aside, Fibre Channel solution providers tend to build high-end products, and if you need high-end features, you may well end up choosing a Fibre Channel solution, not because the technology uniquely offers them but because Ethernet/IP solutions tend to be constructed to meet the needs of less demanding applications.

NAS

NAS providers tend to aim their products at the small to medium business market, along with various niche markets. This is a particularly flexible field, given the ability to scale a NAS box from cheap to expensive, depending on functionality, and particularly given the new hybrid world of NAS Heads and multi-function NAS.

Positioning aside, NAS offers a strong ability to solve the problem of multiple hosts accessing the same file-based data. NAS Heads blend the best of NAS and SAN.

ISCSI

iSCSI solution providers tend to aim their products at the small to medium business market, because these are the growth markets, the markets where Fibre Channel has the least penetration. In addition, small and medium businesses tend to have fewer resources, and iSCSI solutions tend to be cheaper, building as they do on the volume advantages of Ethernet/IP.

SILOS

Psychologically, humans have trouble cooperating when they cross tribal boundaries (org chart boundaries). Functionally, if a technology straddles org chart boundaries, vendors must invest additional effort to figure out who pays for it and who fixes it. Vendors can find it easier to sell product if they can fit it neatly into the customer's org chart, selling to just one group, rather than having to pitch the product to multiple groups and then watch those groups struggle to collaborate. The converse is also true – customers may find it easier to purchase product which fits neatly into their current org chart.

Turning to storage specifically and the topic of convergence (in this context, Fibre Channel, iSCSI, InfiniBand), convergence proceeds more rapidly when:

1. IT silos merge²¹
2. Participants become convinced that the new technology has more to recommend it than the old technology (typically in terms of stability, but occasionally in terms of features).
3. The change is measurably small (this is a rephrasing of the "humans don't like change" adage ... the smaller the change, the more likely people are to accept it).

TECHNOLOGIES AND MODELS

Sharing

Storage is a resource like printers ... sharing it reduces costs. Attaching a disk to every server may be fast and easy ... but one ends up with lots of free space on one disk and not enough on the other ... and shuffling things around burns staff time and typically introduces service disruptions. The desire to reduce costs and improve efficiencies is what pushes customers to improve storage asset utilization ... i.e. to share storage space amongst servers. Remember, the name of the storage networking game is **asset utilization**.

Notice that we don't typically implement storage networking for performance reasons ... like printer sharing, storage sharing can actually slow things down (you have to wait in line at the printer for your job to come out; you have to wait in line at the network for your disk read to

²¹ For example, VoIP becomes successful when voice and data networking departments merge; in the storage arena, when server and networking groups merge, or storage and networking groups merge.

return)²². Rather, we implement storage networking in order to better utilize an asset (space on disks).

Sharing resources only gets you somewhere if you have resources to share ... one speaker told a story about a company which migrated their direct-attached data onto a SAN ... and then added a raft of new servers ... after all, the point of the SAN was to better utilize storage, and these servers needed access to space. But ... the company's direct-attached disk was almost full when they started ... and the networking pipes they installed to connect the storage to the hosts were almost full with frames ... they put roughly the same amount of disk into the SAN as they had had in their direct-attached configuration ... so they went from a place where they didn't have much spare space to another place where they didn't have much spare space. And then they added more servers to the mix, with their flow of I/O requests and demands for additional space ... processing came to a halt, and servers crashed, unable to read and write to their logical disks.

Notice that sharing disk is harder than sharing, say, a printer. Heterogeneous clients have different concepts of volume/directory/file formats, data formats, access control mechanisms, file attributes, naming conventions ... and even homogeneous clients run into trouble when two users attempt to read or write the same file/record simultaneously ... locking mechanisms are required in these situations. Caching introduces all sorts of issues, again, even with homogenous clients. Finally, virtually all clients tend to believe that they 'own' a disk ... operating systems are still primitive in this way ... and allowing two operating systems simultaneous access to the same block can result in a stomping issue. Similar issues occur with tape ... with the addition of the problem that one client can position a tape to read or write ... while another client issues an unload command ...

Typical solutions involve:

- Partitioning systems and devices ... i.e. effectively ripping out the many-to-many connectivity capabilities of the network and emulating the old direct-attached environment ('zoning' and 'v-sanning' are two techniques for doing this)
- Replicating data between two partitions (this tends to reduce functionality).
- Replicating the native file systems with a SAN-aware file system (this is an advanced feature and tends to involve proprietary choices)
- Using NAS (NAS devices excel at solving this set of problems)

High-Level Views

I/O TRENDS

- Increased throughput
- Greater distances
- More hosts wanting access
- High-availability

²² Although, given the fact that disks and networks today tend to be ahead of the overall IT system performance curve, a well-architected storage network will not be the performance bottleneck.

Typical I/O patterns are 80% random and 20% sequential. The exception comes with streaming media, where the I/O pattern is reversed. And in both worlds, typical I/O consists of 80% reads and 20% writes.

Notice that storage designs optimized for sequential reads stress throughput – jumbo frames in Gigabit Ethernet, Fibre Channel for high-end servers, caching. Storage designs optimized for random I/O don't particularly care about throughput; they care about iops (I/O operations per second) – mostly, looking for ways to reduce latency in the application, the operating system, the SCSI driver, the network, the disk – this driven by the command/response (e.g. ping-pong) nature of the SCSI command set.

PHYSICAL TRANSPORT NETWORKS

Name three components of a physical transport network:

- Media (or at least, the appearance of media ... copper wire ... airwaves ... glass)
- Protocol (Ethernet, Fibre Channel, frame relay, ATM)
- Transceiver (something which handles sending and receiving)

STORAGE PROTOCOLS

Name the three elements of a storage protocol:

- Commands (write this block, unmount this drive, read this block)
- Blocks (here's the block, {not relevant}, here's the block)
- Status (I wrote the block successfully, I unmounted the drive successfully, I read the block successfully)

All SCSI exchanges begin with a 'command' and end with 'status'.

PROTOCOL DETAILS

ESCON, SCSI Bus, and ATA are all half-duplex protocols, while Fibre Channel and Ethernet are full-duplex. They are all connectionless.

A FEW WORDS ON INFINIBAND

Currently the highest end choice ... where the PCI bus is just that, a (shared) bus; InfiniBand implements a switch inside a system and between systems ... along with the ability to copy data from one processor's memory space to another ... no platters, no seek time, no NICs, and often no interrupts involved.

CONNECTIVITY TINKER TOYS

iFCP

Most environments contain a widely-distributed Ethernet/IP network, and even organizations which also install parallel Fibre Channel networks often take advantage of their Ethernet/IP network to save dollars. All Fibre Channel NICs can simultaneously host a SCSI stack and an IP

stack, thus allowing them to communicate with a gateway which gives them access to a remote Fibre Channel island.

FCIP Gateways

Additionally, vendors produce FC/IP gateways which tunnel FC frames inside IP frames, for transit across an IP network. (In fact, many vendors integrate this functionality into their Fibre Channel switches.)

iSCSI / FC Gateways

Another option to tunneling FC over IP involves installing an iSCSI stack on a remote host and an iSCSI/FC gateway next to the Fibre Channel SAN.

TIERED MODEL

A popular storage model chunks space into four buckets: Tiers 1 – 4. Tier 1 is aimed at high-performance, mission-critical applications requiring a rich feature set and is often implemented using a SAN. Tier 4 typically means tape.

DATA MANAGEMENT

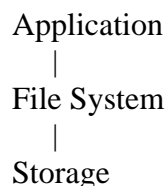
Information life cycle: how do you keep track of your data, delete what you don't need, archive what you don't touch, replicate the important stuff, and what do you do with all those tapes anyway? This is the leading challenge amongst customers today.

Basic Diagrams

See the *Storage Plumbing from a High Level* diagram (<http://www.skendric.com/sans/storage-plumbing-from-a-high-level.pdf>).

DIRECT-ATTACHED STORAGE (DAS)

One, or at most two, hosts share a common resource: storage.²³

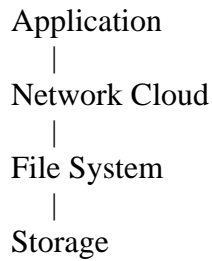


See pages 10-12 of the course book for different flavors of DAS

NETWORK-ATTACHED STORAGE (NAS)

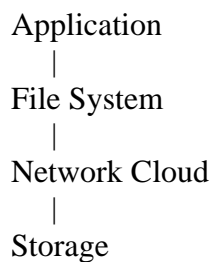
Numerous clients share a common resource: storage.²⁴

²³ Most computers sitting on most office desktops implement DAS ... i.e. they have local disk.



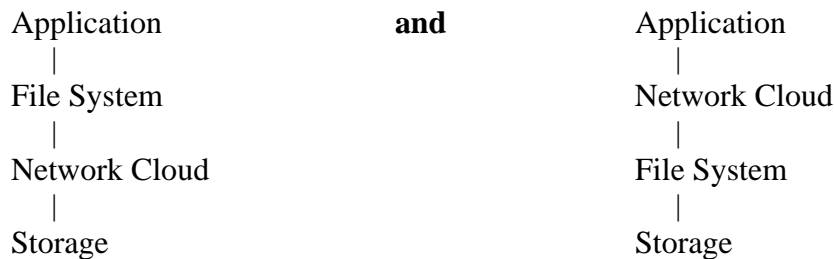
STORAGE AREA NETWORK (SAN)

Numerous hosts share a common resource: storage.



NAS HEADS

These devices combine a NAS front-end with a SAN back-end, delivering the best of both worlds. Most vendors these days include NAS Heads in their offerings.²⁵



REVIEW

See the *DAS & NAS Architecture* (<http://www.skendric.com/sans/das-nas-arch.pdf>) and *SAN Flavors* (<http://www.skendric.com/sans/san-flavors.pdf>) diagrams.

Strengths

Weaknesses

²⁴ Note that NAS is a misnomer ... more accurately, 'Network Attached File Services', or more simply, 'file server'. Most file servers are simply NAS devices (perhaps with extra functions piled on top ... or perhaps with extra functions stripped away).

²⁵ NetApp and BlueArc sell NAS Heads.

DAS

Simplicity
Flexibility
Ease of Use
IP Access

Supports only a few hosts

NAS

Flexibility
Ease of Use
IP Access

Doesn't support block access
Only supports one host

SAN

Block access
Scalability (many hosts)
Extended distance

Doesn't support file access

What's the difference between DAS, NAS, and SAN?

- DAS delivers plumbing: a physical SCSI cable linking a host to a disk.
- NAS delivers services: remote file services across heterogeneous systems.
- SAN delivers plumbing: virtual SCSI cable

Notice that vendors are blurring these distinctions with their NAS offerings ... NAS Heads and multi-access NAS devices combine both NAS and SAN, producing an as yet-unnamed fourth option. [Specifically, some NAS devices contain a SAN or can be attached to a SAN ... these are called NAS Heads ... and some NAS devices (and NAS Heads) support iSCSI, providing block-level access alongside file-level access.]

TECHNICAL DIFFERENCES: FC AND ETHERNET/IP

OSI Model

Ethernet sits at Layers 1 & 2 in the OSI model, while TCP/IP sits at layers 3 & 4. Fibre Channel occupies Layers 1-5 (FC0 – FC 4).

Frame Sizes

Fibre Channel frames max out at 2K; Ethernet frames max out at ~1.5K ... though Gigabit Ethernet w/jumbo support extends frame size to 9K.

Network-Aware Services

Fibre Channel includes support for various ATM-like modes, including bandwidth reservation, QoS, virtual circuits, and acknowledged delivery. In practice, everyone uses Class 3, which means connection-less, best-effort delivery (i.e. the only mode available in the Ethernet world).

Ethernet/IP contains limited support for CoS/QoS, and increasingly, Ethernet/IP networks are implementing these for latency-sensitive applications.

Repeated Bit Patterns

The SCSI protocol makes frequent use of repeated bit patterns (SCSI devices are constantly transmitting ... generally just IDLE frames). Fibre Channel has no provision for scrambling repetitive patterns – this leads to EMI problems when traversing copper media. This is the reason why Fibre Channel over copper is limited to short distances, perhaps 30m. Ethernet, along with most other networking technologies, uses scrambling to overcome this issue. The FCBaseT effort is aimed at remedying this, by borrowing the physical layer being developed by the 10GigE-over-UTP group. Hard to predict whether or not this effort will succeed in the market place.

Flow Control

TCP uses a sliding window mechanism to control flow, with each partner advertising to the other how many bytes it is willing to receive and incrementing or decrementing this as buffers empty or fill during the conversation. Fibre Channel can perform a roughly equivalent form of flow control, called End-to-End flow control, at Layer 4 as well (Layer 4 in the OSI and TCP/IP models; Layer 2 in the Fibre Channel model). However, this only occurs if the conversation is running in a Class of Service which includes acknowledgements. Generally, Fibre Channel conversations run in Class 3, which is an unacknowledged experience, and thus End-to-End flow control doesn't play a role.

However, at Layer 2 (in all three models), all Fibre Channel NICs and ports employ a Buffer-to-Buffer form of flow control based on credits: during login, each receiver hands its partner credits, typically expressed in number of Fibre Channel frames. A transmitter can spend credits by transmitting frames and must stop transmitting when it exhausts its credits. Receivers replenish credits by signaling Layer 1 to send the transmitter more credits ('R-RDY credits'). This is a deterministic form of flow control and guarantees that a port's buffer can never be overrun.

Compare this to Fast Ethernet and Gigabit Ethernet's flow control mechanism, which is non-deterministic. A transmitter can send as many frames as it likes ... until and unless the receiver emits a 'pause' frame, which tells the transmitter to shut up for a specified amount of time. (A transmitter can follow up with a 'pause' frame specifying '0' for the amount of time to pause; this functions as a "go ahead, I'm ready" message.)

Addresses

TCP-IP addressing includes the public and the private concept; Fibre Channel addressing has no parallel to the public concept: all addresses are private. Both types of addresses are composed of three parts: for TCP/IP, major.network.subnet.node; for Fibre Channel, area:switch:port.

Routing Protocols

A plethora of routing protocols exists within the IP world. Only one exists in the Fibre Channel world: Fibre Channel Shortest Path First (an OSPF clone). Fibre Channel vendors integrate the routing function, along with every other upper layer service, into their switches ... Fibre Channel switches are multi-function devices, whereas in the Ethernet/IP world, simple switches only handle Layer 2 packet forwarding, and even complex switches only integrate a few functions

(comparatively). In Fibre Channel nomenclature, switches ‘route’ frames to their destination ... one speaker prefers to use the word ‘steer’ ... Fibre Channel networks contain only a single broadcast domain, or in TCP/IP terms, a single subnet. [I don’t understand how Areas are used in Fibre Channel –sk.]

Hardware Offload

In the Ethernet/IP world, Ethernet NICs tend to just perform Ethernet functions. The host’s TCP/IP stack, built into the OS, performs TCP/IP functions.

In the Fibre Channel world, all NIC manufacturers use firmware or ASICs or both to off-load Fibre Channel FC0-4 processing from the host’s CPU. This is one reason why Fibre Channel NICs are expensive – they are doing a lot of work. This is called ‘hardware offload’; the goal is to shift work from the host CPU to the CPU on the NIC.²⁶

In the Ethernet environment, high-end adapters contain TCP/IP off-load environments (TOE), and sometimes even iSCSI off-load functions. These NICs cost more than the garden variety do. Generally, NICs equipped with hardware off-load allow the administrator to upgrade the on-board firmware, to fix bugs and add features.

Path Diversity

Both Ethernet/IP networks and Fibre Channel networks support multi-pathed environments. Multi-pathed environments mean that packets can arrive out of order. Both TCP and FC3 contain sequence numbers, allowing the receiver to correctly reconstruct such out-of-order experiences. However, in the Fibre Channel world, the NIC manufacturers haven’t implemented this²⁷; therefore, Fibre Channel NICs consider out-of-order packets to be an error and they discard them. Switch manufacturers, therefore, make sure that all packets belonging to a given conversation traverse only one path. This is not part of the Fibre Channel specification, merely the result of an informal arrangement between NIC manufacturers and switch manufacturers.

Timers

In the Ethernet/IP world, TCP maintains timers, tracking how long it has been since the partner acknowledged receipt of data. Fibre Channel environments contain numerous timers ... multiple timers within the various layers of the Fibre Channel stack and within the SCSI protocol riding on top of Fibre Channel. Both environments can require timer tuning, depending on latency and packet loss.²⁸

²⁶ I’ve heard varying opinions on how significant this effect is. Some folks say it is a big deal; others say that the most you’ll see is 3-5% drop in CPU utilization, if you replace a garden variety NIC with a hardware offload NIC.

²⁷ I’ve heard the rumor that Qlogic has, but that somehow their approach only works when talking with another Qlogic device ... I haven’t confirmed these rumors.

²⁸ My impression: TCP has a lot of sophistication in its timers – feedback mechanisms which allow for self-tuning – reflecting the decades of experience which the TCP community has with managing transport across widely varying networks, while Fibre Channel timers don’t have this kind of built-in tuning support and thus require manual fiddling when they encounter less-than-ideal conditions. I am expressing my opinion here, not supported by expert input.

Interoperability

The Ethernet world is heavily influenced by standards – customers take it for granted that an Ethernet NIC from one manufacturer will interoperate with an Ethernet switch from another manufacturer.²⁹

Originally, the Fibre Channel world had interoperability challenges – switches and NICs from different manufacturers didn't always interoperate. Even today, one can engage 'advanced features' on a NIC or a switch and run into compatibility issues. In general, however, industry players test their gear against one another and resolve problems, populating the 'compatibility matrix', a document which displays which hardware revisions loaded with which versions of software interoperate successfully with which competing products.³⁰

LANGUAGE AND LEGENDS

Basics

SCSI devices include adapter cards, disks and arrays of disks, tape, optical drives, printers, scanners

We talk about disks containing components like platters, actuator arms, cylinders, tracks, sectors, heads. Disk performance depends on revolutions per minute, seek time, latency, data transfer rate, caching mechanisms.

RAID typically improves performance and/or offers some level of fault tolerance. Cache can improve performance.³¹

Same Words, Different Meanings ... Different Words, Same Meanings

FIBRE CHANNEL

- NICs are called 'Host Bus Adapters'
- Switches are called 'switches'. But they typically have some routing functionality built into them.
- Switches with hardware redundancy (multiple power supplies, multiple brains cards) built into them are called 'Directors'.
- Switches often have multiple functions built into them, like routing, authentication, authorization, and name services.

²⁹ Having said that ... I can recall a certain manufacturer interpreting the auto-negotiation portion of the 802.3 spec differently from everyone else, such that their NICs and their switches interoperated (negotiating to full-duplex), but when mixed with NICs and switches from other vendors, the result would be duplex mismatches. We spent months tracking down these NICs and switches and upgrading them to fixed versions of their software loads.

³⁰ I've heard mixed reports on this – some folks recommend paying close attention to the compatibility matrix; others say they don't bother and just 'plug and play'.

³¹ The more predictable the I/O, the better the performance gain; the more random the I/O, the less important cache is. Generally can't hurt and is sometimes over-rated.

- Device addresses are called ‘identifiers’ or ‘IDs’: they are hierarchical: domain, area, device: e.g. switch 01, Port 02, Device EF = 0102EF
- ‘Loop switches’ are similar to hubs in Ethernet land.
- Expanders have no analog in Ethernet land, though one can sort of think of them as low-level hubs.
- Uses a credit-based system for controlling packet flow.
- Talks about segmentation, convergence, and reassembly (fragmenting blocks into chunks small enough to fit into Fibre Channel frames, ensuring that the chunks arrive in order at the other end, and reassembling them into the original block)
- Routing: steering a packet across one ‘appearance of media’,³²
- Fabric Services: the services which Fibre Channel switches offer:

Here are the Fabric Services:

Fabric Controller

Initialization, Configuration, Routing

Login Server

N-Port address assignment

Registry services

Name Server

Translates and look-up

N-Port IDs

World Wide Names

SCSI Target, LUN

Symbolic Names

Management Server

Fault

Configuration

Accounting

SNMP

Time Server

Expiration timers

Synchronization

Alias Server

Hunt groups

Multicast

³² Hearing that “switches perform routing” can be confusing to Ethernet/IP people ... but that’s the language which the Fibre Channel world uses. It kind of makes sense, once one realizes that Fibre Channel switches tend to have routing functions built into them.

ETHERNET/IP

- NICs are called NICs
- MAC ‘addresses’ aren’t really addresses: they are identifiers.³³
- IP addresses are hierarchical: major network, subnet, node (140.107., 74, 123)
- Additional functions like routing, authentication, authorization, and name services tend to be delivered in separate devices, rather than bundled together into switches and routers
- TCP uses a sliding window for controlling packet flow
- Talks about fragmentation and reassembly (fragments blocks into chunks small enough to fit into TCP frames, reorders and reassembles the fragments at the other end)
- Routing: forwarding a packet across many ‘appearances of media’

See page 30 of the course book for Name, Address, & Route discussion.

COMMON SEMANTICS

- Discovery (figuring out what is available on the network)
- Hardware offload for performance acceleration

ADDRESS SPACE COMPARISON

<u>Storage Network Scheme</u>	<u>Number of Available Addresses</u>
SATA	2
SATA II	15
SCSI Narrow	8
SCSI Wide	16
SAS ³⁴	16,384
Fibre Channel	16,777,214
IPv4	4,294,967.292
IPv6	Big

Myths

This is one of my favorite subjects. We humans love stories, we are highly social, and we are gullible. We sport brains which excel at seeing patterns ... we’re so good at seeing patterns, that we see them even when they don’t exist. Consider how cultures around the world looked at the night sky ... saw patterns in an effectively random distribution of points of light ... and then

³³ Addresses tell you how to get there ... postal addresses do this, as do IP addresses and even URLs, assuming you are geeky enough to interpret them. But MAC ‘addresses’ don’t do that ... they are just names ... i.e. identifiers.

³⁴ Serial Attached SCSI

constructed stories to explain those patterns. In the modern era, ‘urban legends’ take the place of deities battling amongst the stars.

What I enjoy about this subject is finding the kernel from which the story started ... the piece of accurate information which then ballooned into a whopper of a tale.

The IT world carries its own share of these myths. Here are few which are peculiar to the storage networking world.

FIBRE CHANNEL IS COMPLICATED

So is Ethernet/IP ... the two solve the same set of problems using similar strategies, and one only looks complicated when you aren’t familiar it.

FIBRE CHANNEL NICS COST MORE THAN ETHERNET NICS

Yes ... Fibre Channel NICs tend to cost more than Ethernet NICs ... but only when comparing apples with oranges. All Fibre Channel NICs come with hardware off-load for Fibre Channel processing, and this contributes to their cost. Ethernet NICs equipped with full hardware off-load for TCP/IP processing ... and particular those equipped with hardware off-load for iSCSI processing ... start getting pricey, too.³⁵ A more accurate way of summarizing this issue is to say that Fibre Channel doesn’t have a low-end play – there’s no equivalent of the simple, stripped-down \$19.95 NIC in the Fibre Channel world.

FIBRE CHANNEL IS MORE EXPENSIVE THAN ETHERNET

Sure. But Fibre Channel solutions, because of market positioning, tend to deliver higher capacities, more features, and greater MTBF than do Ethernet solutions, so you’re getting more for your money. Perhaps a more accurate way to express this would be that Fibre Channel doesn’t have a low-end play -- the NICs, switches, and support are expensive; licensing models add additional cost. By comparison, iSCSI can run over low-end NICs using freely available software stacks over commodity Ethernet/IP networks.

WE MIGRATE TO SANS TO INCREASE PERFORMANCE AND RELIABILITY

I think this one arose from vendor enthusiasm – no matter what a sales person is selling, s/he will describe it as “fast” and “reliable”. In fact, DAS remains the performance and reliability leader of the storage networking world. When we migrate to a SAN, whether implemented over Fibre Channel or iSCSI, we are replacing a short, simple, low-latency cable (latency measured in nanoseconds) with a mess of cables and switches and additional software (latency measured in microseconds if you’re lucky), full of multiple complex points of failure. In other words, when we build SANs, we build a maze of crummy virtual SCSI cables³⁶ ... we do this because we

³⁵ As of this writing, even vanilla Ethernet NICs ship with partial TCP/IP hardware off-load ... TCP checksum calculation, for example. In general, Ethernet NICs deliver varying amounts of hardware offload ... and the more off-load they include, the more they cost.

³⁶ Kudos to Howard Goldstein for this phrase.

want to better utilize our assets (**asset utilization**), and we take a performance and reliability hit in the process.³⁷

FIBRE CHANNEL IS FASTER THAN ETHERNET

Hmm, well, Ethernet counts in twos, whereas Fibre Channel counts in tens ... a gigabit in the Ethernet world is $1024 \times 1024 \times 1024 \times 8$... whereas a gigabit in the Fibre Channel world is $1000 \times 1000 \times 1000 \times 8$.³⁸

Quibbles aside, yes, Fibre Channel solutions tend to deliver more throughput than do Ethernet solutions. As of this writing, Fibre Channel comes in 1, 2, and 4 Gb flavors, whereas Ethernet tends top out at 1Gb (ignoring 10 GigE and 10GigFC for the moment). Fibre Channel NICs tend to come with hardware off-load features, which conserve host CPU, whereas Ethernet NICs tend to come with fewer off-load features, requiring more CPU involvement. High-end storage servers, equipped with Fibre Channel interfaces, tend to come with lots of cache, whereas low-end storage servers, equipped with Ethernet, tend to come with less cache. And the Fibre Channel culture tends to invest upfront analysis and back-end engineering (characterizing the bandwidth and latency requirements of each application and then designing a storage network to meet those requirements), whereas the Ethernet world tends to be less rigorous about such things.

In summary, Fibre Channel systems tend to outperform iSCSI systems because of several factors, including the market positioning of high-end products (they have FC interfaces on them, not iSCSI interfaces) and the culture which surrounds them. However, there is no technological reason stopping someone from building a high-performing iSCSI solution ... if s/he is willing to pay top dollar for high-end gear.

Higher level lessons:

- Fibre Channel has been the only player in this space for years ... iSCSI is a recent player. iSCSI is chewing away at the lower-end of the market, and thus Fibre Channel players are positioning their products at the higher-end of the game.
- All this talk about network performance generally misses the larger point: most I/O is random and bursty, which means that both the disks and the network are idling ... disk and network architectures are ahead of the performance curve these days -- it is system buses, or client applications, which are the bottlenecks. So really, who cares how many bytes are in a giga or how many giga-whatevers your network can carry? Don't get side-tracked by discussions of maximum theoretical performance.

In the future, expect iSCSI competition to push Fibre Channel costs down ... Fibre Channel to deliver 8GB, and both camps to deliver 10Gigabit solutions (i.e. iSCSI over 10Gig and 10Gig Fibre Channel ports/NICs).

³⁷ "Life is pain, Princess. Anyone who tells you otherwise is selling something" -- The Man in Black. Or, with a little less drama: All choices involve trade-offs.

³⁸The difference ends up being minor. One gigabit Fiber Channel delivers ~103MB/s throughput. One gigabit Ethernet/IP/TCP/iSCSI delivers ~114 MB/s.

FIBRE CHANNEL SWITCHES COST MORE THAN ETHERNET SWITCHES

Yes, they do. However, it is worth noting that Fibre Channel switches are multi-function devices which include services typically provided by other devices in an Ethernet environment (devices like DNS, DHCP, directory, and authentication servers, as well as routers). Comparing the cost of the two infrastructures is hard, because of this bundling.

That being said, yes, Fibre Channel switches cost more per port than do Ethernet switches ... remember that Ethernet has a volume advantage over Fibre Channel.

GLASS IS MORE RELIABLE THAN COPPER

This is a wonderful myth. I first heard it in the late-80s, while supporting a campus-wide roll-out of high-speed networking over unshielded twisted pair. These days, storage networking weenies who believe this myth will earnestly explain how your precious data must not be entrusted to Ethernet; rather, it must be carried on Fibre Channel, because copper cabling isn't reliable.³⁹

Well ... let's dig into this one a bit. Both copper⁴⁰ and glass cabling can introduce physical layer errors, if improperly installed; both run error-free when correctly installed. Have a look at the error counters on your most heavily used Ethernet ports ... solid at zero, right? And if they aren't, then I bet you're going to investigate, looking for a bad NIC, a bad switch port, or bad cabling.

Consider your favorite critical application ... say, a stock trader punching buttons to sell ... a medical application monitoring patient vital signs ... or a military application. All three may present their human interface on a portable wireless device. The device emits a packet over WiFi ... which travels across air to a nearby wireless access point, where the packet is then transmitted onto the wired network, typically Fast Ethernet over Category 5 copper cabling. Switches or routers may convert the packet to Gigabit Ethernet over glass ... and then back again to Gigabit Ethernet over copper, arriving at a front-end server. The front-end server processes the request and sends a request to a middleware server (again over Ethernet-running-on-copper) ... which spits out a request to a backend database server (again over copper) ... which finally converts the requests to a SCSI write, sends it across a Fibre Channel network to a disk, where a bits inside a block change, implementing the end-user's request. Now, why doesn't the modern IT world fall apart, under the load of massive data degradation, with all this copper (and air!) munging up the bits? Because all these transmission schemes, from WiFi to Ethernet to, yes, Fibre Channel, contain error-checking and retransmission schemes. And because most cabling is installed correctly, so that it doesn't introduce bit errors.

Now, how did this one arise? Here are my guesses:

- Glass can carry signals much farther than copper can.
- Signals on glass resist EMP pulses from nuclear explosions, whereas signals-on-copper do not.

³⁹ Of course, Ethernet runs over glass as well as copper ... but folks caught up in this myth ignore this point.

⁴⁰ When I write 'copper', I mean unshielded twisted pair cabling meeting the Category 3 or higher specification.

- ESCON was an optical-only protocol; and the Fibre Channel developers copied that choice ... they did not develop a scheme to randomize repeated bit patterns, the way the Ethernet folks did.

Looking to the future, a T11 working group is tackling FCBaseT, i.e. Fibre Channel over copper. They are planning to re-use the phy layer which the IEEE 802.3 10GigE over copper working group is developing.

In the meantime, notice the pricing consequences of optical transport:

- Component costs go up, because lasers cost more than copper transceivers and glass is harder to manipulate than is copper.
- Fibre Channel customers have to go back to their cabling infrastructures and install a glass cabling plant, in parallel with their copper plant.

FIBRE CHANNEL REQUIRE SINGLE-MODE TRANSCEIVERS

No, it doesn't – it can employ multi-mode transmission. Just as in the Ethernet world, one makes the choice of single-mode vs multi-mode based on the distances involved and the available cable plant. I don't have any ideas on how this one arose.

CUT-THROUGH VS STORE-AND-FORWARD

I can vouch for the Ethernet side of this ... I know of no Ethernet switches which implement cut-through today, the last ones which did reaching end-of-life back in the 1990s; I've verified this using a hardware-based packet analyzer. Fibre Channel switch manufacturers tout the cut-through capabilities of their gear ... and where the myth appears is when sales people assure me that their switches implement cut-through switching even when pushing packets from 2Gb ingress ports to 4Gb egress ports and vice versa ... and I find this hard to believe.⁴¹

Nevertheless, perhaps Fibre Channel switches perform cut-through when they are forwarding frames from one port to another without a speed transition. I hope to borrow a Fibre Channel analyzer some day to verify this.

HARDWARE FORWARDING VS SOFTWARE FORWARDING

In this myth, folks say that Fibre Channel switches make their forwarding decision in hardware whereas Ethernet switches use software ... and hardware is faster than software. Nope. All switches forward using hardware ... and all switches use software to populate those hardware-based forwarding tables. I'm guessing this myth arose during the early days of Fibre Channel, when Ethernet-based routers (routers, not switches) performed forwarding decisions in software.

⁴¹ If you don't understand why, drop me a note ... the visualization is brings a chuckle to most folks, once they realize how absurd it would be to actually implement this.

FIBRE CHANNEL HAS HALF THE OVERHEAD OF ETHERNET/IP

Technically accurate ... but practically speaking, irrelevant. According to my calculations, Fibre Channel's frame overhead is ~1.5%, while Ethernet/IP's frame overhead is ~3% (for maximum sized packets) ... hardly something to write home about.

Fibre Channel:	32 bytes / 2048 bytes = 1.6%
Ethernet/IP:	52 bytes / 1514 bytes = 3.4%

FIBRE CHANNEL IS PROPRIETARY; ETHERNET/IP IS OPEN

This is one of those perspective things. Fibre Channel grew from the Big Blue world of mainframes ... Fibre Channel was born, in some ways, as ESCON version 2. From the Fibre Channel point of view, when compared to the closed architecture of mainframes, Fibre Channel is wildly open (you don't have to buy all IBM gear to get it to work, right?). Furthermore, no one vendor owns the Fibre Channel specification; rather, the specification grows within the ANSI structure. And while interoperability used to be a bear in the Fibre Channel world, it becomes more and more reasonable, particularly as of this writing, for Fibre Channel gear to interoperate, not only within a single vendor's offerings but also between vendors. In Fibre Channel-speak, the "compatibility matrix keeps expanding".

On the other hand, from an Ethernet/IP perspective, living within the world of IEEE and IETF standards ... sure, the Fibre Channel world feels proprietary ... fragmented ... full of competing, non-interoperable 'standards' ... or standards which are 'optional' and are only implemented by one vendor ... and this "compatibility matrix" concept makes IEEE/IETF weenies shudder. When's the last time you had to upgrade your Ethernet NIC driver in order to get it to work with some other vendor's Ethernet switch? When's the last time your Web browser refused to even connect with some company's Web server ... because of an incompatibility between the TCP/IP stack on your client and the TCP/IP stack on the server? At the end of the day, this one is a matter of perspective.

MICROSECONDS MATTER

Fibre Channel vendors tout their 2us forwarding decision latency ... and perhaps that is accurate.⁴² I've observed informal tests of the Layer 2 and Layer 3 Ethernet switches and routers in my environment, using Finisar THG analyzers (equipped with clocks accurate to 20ns), and I've recorded forwarding latencies ranging from .3us to 9us, depending on make and model.⁴³

The core of this myth is: it doesn't matter. In a world in which application latencies, OS latencies, and disk latencies are measured in milliseconds ... installing a switch which shaves a few microseconds of forwarding decision time off the whole experience just doesn't make a difference.⁴⁴

⁴² If you're willing to loan me your Finisar Xgig analyzer, I would love to verify these claims!

⁴³ Many thanks to Mike Pennachi and Chris Greer of Network Protocol Specialists for performing this work.

⁴⁴ Perhaps there exist high-end environments – big storage arrays with lots of cache, applications spitting out tens of thousands of IOPS, business requirements pushing for sub-second UI response – in which every microsecond does

My guess: switch vendors can't control application, OS, and disk latencies ... so they focus on what they can control ... and then tout it as a differentiator, when they are competing against other switch manufacturers ... and thus the legend is born.⁴⁵

MY APPLICATION REQUIRES FIBRE CHANNEL

Sometimes, this one is true, which perhaps explains why this story also appears as a myth. Some applications, typically found in vertical markets, were hard-wired to mess directly with Fibre Channel messages, and their designers don't see a business incentive for generalizing this. In other cases, again typically in vertical market situations, the application isn't hardwired for Fibre Channel ... but the operating system on which it runs doesn't have iSCSI support. (Sun has been slow to adopt iSCSI, so this is a major effect in some environments.) Until just recently, VMWare didn't support iSCSI.

However, in general, applications don't know or care about the transport involved: they reach for a LUN, and the operating system works its magic to deliver that LUN, whether over a Parallel SCSI cable, a Fibre Channel network, or an iSCSI network. In particular, database managers don't care (aka Oracle, SQL Server..) and e-mail servers (Exchange) don't care.

STORAGE APPLICATIONS

Backup & Recovery

Safeguard data against system failure or user error.

- Network, LAN-free, Snapshot, Server-free, NDMP, backup to disk, disk to disk, tape virtualization, frozen images (split mirror, copy-on-write aka snapshots, redirect on write), full, reference, incremental (differential, cumulative)
- Performance factors: client load, network load, media server load, disk drive transfer rate, tape drive transfer rate

Continuity Management

Replicate your environment to a distant location, as part of a larger plan to ensure continued operation in the event of major disruptions, typically called 'disasters'.

High availability

Hardened and/or redundant disk supports the design of highly available systems.

indeed matter. Remember, I live in a low-end world, vis-à-vis storage, and thus am not familiar with high-end requirements.

⁴⁵ And of course, latency matters under all sorts of circumstances ... try using a ping-pong application over a WAN with millisecond latency ... what I'm claiming here is that when the dominant factor of an equation is 10^{-3} , shaving instances of 10^{-6} off the total does not materially affect the result.

Performance Scalability

Facilitating the addition of more servers supports the scaling of an application.

Information Lifecycle Management

Keeping track of all that data – currently the largest single challenge facing storage customers.

THE ANATOMY OF A SAN

Virtual SCSI Cables

Modern SANs in the open-systems world revolve around the SCSI physical transport protocol. Specifically, they emulate the original host <--> SCSI environment, in which one host connects to one disk via a cable (running the SCSI Bus protocol). Variants on the original SCSI Bus (aka SCSI Narrow Bus) include SCSI Wide, SCSI Fast, SCSI Ultra, and the combinations (SCSI Wide/Fast, etc). All use the same SCSI Bus media access approach (Arbitrate ID, Select ID, Identify LUN).

All the various storage solutions which followed ... RAID, Serial Attached SCSI, Fibre Channel, iSCSI ... they all create a virtual 'SCSI Bus' across their respective fabrics. As far as the host's SCSI driver and the disk's SCSI driver are concerned, the two are connected via a simple SCSI Bus cable ... and the fact that complex things may happen in between is hidden from them.⁴⁶

From a plumbing point of view, the function of a SAN is to build a virtual relationship between a SCSI initiator and a SCSI target.

The speaker argues that all of these, from SCSI Bus through iSCSI, create Storage Area Networks (SANs). Dissenters may argue that anything less than Fibre Channel and iSCSI are merely direct-attached storage and don't merit the label of 'SAN'. The speaker argues that:

- Point-to-point topologies (one host connected to one disk) are still networks, albeit simple, two-node networks
- Modern data (switched & routed) data network consist solely of point-to-point links, i.e. solely of a collection of two-node networks ... and we seem to have no problem calling these point-to-point links 'networks' ... so why not extend that name to point-to-point SCSI configurations?
- All these topologies allow for multiple hosts accessing multiple devices across them, even the original SCSI Bus ... sounds like a network to me!
- The vast majority of Fibre Channel and iSCSI networks consist of a one-to-one mapping between hosts and their storage chunks.⁴⁷

⁴⁶ Notice, in fact, that SAN builders are merely building crummy SCSI cables ... crummy *virtual* SCSI cables ... taking short, simple, reliable parallel SCSI cables, with latencies measured in a few nanoseconds ... and replacing them with long, complex multi-pathed glass and/or copper cabling plants ... introducing latency measured in microseconds along with multiple complex, powered points of failure. Such is the price of progress. ☺

⁴⁷ In the data networking world, we install a network to create many-to-many connectivity and then boast about it... in the storage networking world, we install a network to create many-to-many connectivity ... and then promptly

Partitioning

Because operating systems tend to want exclusive access to a disk, SAN administrators put effort into partitioning their network, to create one-to-one mappings between hosts and chunks of storage. This process is called 'zoning'. Zoning can be implemented in many places:

- Operating systems (filters and masks)
- Host Bus Adapters (LUN mapping)
- Server Middleware
- Switches (hard zoning or soft zoning)
- Storage devices (LUN masking)

Hard zoning involves mapping one port to another port. Soft zoning relies on world-wide-naming, which is the storage world's version of DNS. Hard zoning requires more effort but is more secure; soft zoning requires less effort but is vulnerable to spoofing.

A more recent approach to solving the same problem is 'V-SANning', aka Virtual SANs; this is implemented in the switch. While the T11 group has blessed this approach as an optional part of the Fibre Channel standard, as of this writing, only Cisco implements it.⁴⁸

CHOOSING STORAGE SOLUTIONS

Design process

GUIDELINES

- Continually remind yourself of the answer to the following question: What am I trying to accomplish?
- Recognize that requirements embody many business and technical goals
- Address human factors (comfort with technology, turf wars, jockeying for promotions)
- Employ traditional network design methodologies: storage is just another byte of data on the wire
- Start with application requirements first and from those derive session and data transport requirements
- Only then move onto plumbing design

OUTLINE

- Characterize current environment⁴⁹

slice it up to restrict everybody to one-to-one connectivity ... and then crow about how great our network is. The commonality is how much humans like to show off what they've done.

⁴⁸ This is typical in the Fibre Channel world ... 'optional' standards which only one vendor implements.

⁴⁹ Identifying current (and desired) I/O behavior ends up being key to storage solution design. Check out free tools like IOMeter from Intel, HIMON from ???, PerfMon from Microsoft. And not-at-all-free tools like NetWisdom from Finisar.

- Analyze requirements: business and technical needs and goals
- Identify what resources I own currently the value of which I want to leverage
- Identify what technologies meet those requirements
- Design a selection of solutions to meet those requirements while leveraging those resources (logical design, identify topologies, addressing, naming, security, management, physical design)
- Product selection
- Optimize
- Document, implement, test, etc.

Typical Storage Requirements

- Support short & bursty traffic
- Support long & streaming traffic
- Expand space without incurring downtime or operator time spent shuffling files
- Facilitate backup & restore
- High availability
- Scaling: multiple servers accessing the same data
- Block-level access (database managers)
- File-level access (most applications)
- Write-heavy, read-light (Federal Express and UPS are constantly updating their databases ... hardly anyone is looking at the result)
- Read-heavy, write-light (typical data mining applications)

Technology Choices

Small and Skinny
 Small and Wide
 Big and Wide
 Huge and Wider

ATA, SCSI Bus, ESCON, Serial ATA (SATA)
 Serial Attached SCSI (SAS)
 Fibre Channel, iSCSI, iFCP, FCIP, NFS/TCP/IP/Ethernet
 InfiniBand

APPENDIX

Future Trends

FC-BASE T

This is Fibre Channel over Category 6. The FC-BaseT group plans to copy the phy from the IEEE 802.3an (10GigE over UTP) group, which should allow NIC and switch manufacturers to produce product with a price point lower than traditional glass-based Fibre Channel gear. In February 2007, the 'final candidate' proposal went to committee members for approval. Unclear when first product would ship. Product would likely be targeted at existing FC customers who want to save money by re-using existing copper cabling, rather than installing new glass cabling purely to support FC.

FC OVER ETHERNET

This is Fibre Channel Layer 2 over Ethernet, rather than over Fibre Channel Layer 1. This would require replacing FC NICs and switches with Ethernet NICs and switches but allow Fibre Channel vendors and customers to leverage their existing product and expertise in Fibre Channel, merely substituting a simpler, and cheaper, physical infrastructure, taking advantage of Ethernet volume and expertise. [This would not allow re-using existing Ethernet infrastructure, as existing Ethernet switches do not support the myriad of functions which Fibre Channel requires, functions like the FC equivalents of DHCP and DNS.] Product likely targeted at customers with existing FC infrastructures who are reluctant to migrate to iSCSI but who are looking for ways to reduce costs by buying cheaper NICs and switches. This would also permit consolidating IP traffic and FC traffic across a single NIC and a single switched infrastructure, reducing the number of NICs and switches in the data center. Cisco is pushing this, having just purchased a company which claims to be developing product in this space.

FC OVER CONVERGED ENHANCED ETHERNET

This requires a modification to Ethernet to support QoS parameters, something which Ethernet does not have and which Fibre Channel has in principle but not in practice (defined in specification but not implemented in product). With that in place, its supporters envision consolidating IP, FC, and InfiniBand traffic across a single (likely 10GigE) NIC and a single switched infrastructure. As of April 2007, this is vision is in the hype stage, with no one working on specifications, let alone product.

Disk Engineering

Disk manufacturers build two types of platters – SCSI and ATA – with the difference residing in manufacturing techniques which trade-off cost and reliability. SCSI platters are stuffed into disk assemblies engineered for reliability while ATA platters are stuffed into disk assemblies engineered for reduced cost. The driving factor is vibration: rapidly spinning platters vibrate, and containing this vibration (SCSI disks) requires costly engineering choices.

The disk manufacturer then attaches interfaces to the newly manufactured disk. P-SCSI, SAS, or FC in the case of SCSI disks; P-ATA, SATA, or FC in the case of ATA disks.⁵⁰

Real World Numbers

I acquired these numbers from a storage VAR; I have not verified them myself ... take them with salt. These are the best numbers which the sales engineer working with us says he has actually seen, all of them in the field.

- High-end x86 servers can sustain 85 Mb/s of iSCSI, NFS, or Fibre Channel traffic.

⁵⁰ SAS and FC-ATA disks are the newest arrivals on the disk scene.

- Open systems ‘big iron’ (specialized NAS Heads, high-end Sparc or PowerPC based systems) can sustain 115Mb/s of iSCSI/NFS over Gigabit Ethernet; 140Mb/s over 2 Gb Fibre Channel
- Hardware off-load iSCSI cards increase bandwidth utilization by ~10% and save ~5% on CPU performance.
- By default, most operating systems will crash and burn if they don’t receive a response back from a SCSI LUN after 60 seconds. Oracle will crash and burn after 90 seconds. Microsoft Exchange will start logging error messages after 10 seconds – the speaker did not know at what point Exchange crumps. In general, these thresholds are configurable.
- CIFS is slow, lots of overhead, chatty. Some storage vendors are implementing CIFS accelerators; the speaker was unclear on how these things worked their magic.
- On our network, using four Windows boxes concurrently employing iSCSI to access a high-end NAS Head (all located on the same subnet): 115 MB/s. We did not have jumbo frames enabled.

Lab Numbers

The same sales engineer as above says that his contacts at corporate headquarters believe that the following numbers are accurate, at least in the lab.

- High-end storage arrays can sustain 400 MB/s – doing this requires employing multiple 2Gb Fibre Channel ports on the array, lots of spindles, and lots of cache. And, of course, multiple hosts emitting read requests.
- At the high-end, fancy arrays can beat this 400 MB/s number – arrays equipped with as many as 192 ports, some or all of which are 4 Gb Fibre Channel.

The Case for Fibre Channel SANs

Remember, customers want to leverage what they already own, i.e. increase **asset utilization**.

EXISTING FC NETWORKS

Customers with existing Fibre Channel networks, having invested in the capital acquisitions, the management applications, the software licensing fees, the vendor relations, the in-house expertise, will tend to add additional hosts and storage arrays using Fibre Channel, leveraging their existing investment.⁵¹

⁵¹ Of course, customers in this position will also consider FCIP extension (extending their Fibre Channel connectivity to remote hosts using an existing IP network) as well as FC-iSCSI gateways. In this way, they can conserve dollars when adding hosts with low-end requirements or low-end capabilities to their SAN, e.g. x86-based hosts, which can’t go fast enough to take advantage of Fibre Channel performance, hosts which don’t need the extra bandwidth, hosts to which glass connectivity isn’t available, or hosts which are too low-end to be worth the expense of a Fibre Channel HBA plus glass connectivity.

HIGH-END PERFORMANCE

Customers who have invested in the open systems version of ‘big iron’ (typically Sparc or PowerPC based Unix boxes, specialized NAS Heads like NetApp and BlueArc, or high-end storage arrays) may have spent hundreds of thousands or even millions of dollars on each device and may have business requirements driving high throughput. In these environments, the cost of installing a Fibre Channel network is small compared to the dollars already spent on the servers, and pushing throughput above the 115Mb/s ceiling of GigE may be critical to business needs.⁵²

UNRELIABLE OR OVERLOADED LAN

Customers who have unreliable or overloaded Ethernet/IP networks will off-load their storage applications to a parallel SAN. Depending on the application requirements and the gear involved, an iSCSI SAN might make sense ... but then again, a Fibre Channel SAN might make sense also.⁵³

The Case for iSCSI SANs

Remember, customers want to leverage what they already own, i.e. increase **asset utilization**.

EXISTING ETHERNET/IP NETWORKS

Customers with an existing Ethernet/IP network, having invested in the capital acquisitions, the management applications, the vendor relations, the in-house expertise, will tend to add storage applications to the LAN.⁵⁴

Here are exceptions:

- Environments running applications and/or operating systems which only support Fibre Channel.
- Performance-conscious environments, particularly those driven by ‘big iron’ servers which can push past the 1 GigE barrier.

Designing iSCSI SANs

Our organization is currently debating how to overlay iSCSI on our existing Ethernet/IP network. Current ‘best practices’ recommend creating a private VLAN on existing data center switches and dedicating ports to iSCSI traffic (i.e. dual NICs per end-station, one carrying iSCSI traffic; the other carrying commodity traffic).

In our environment, we don’t have a single data center; instead, we have nine server rooms, of varying sizes and vintages, scattered around our campus. Server rooms are typically equipped with a pair of Layer 2 Ethernet switches (redundant access layer switches), dual-homed to a pair

⁵² In this situation, the customer wants to increase the utilization of the storage array, along with server CPU and Bus.

⁵³ In this case, the Ethernet/IP network is over-utilized; there is no room for increasing asset utilization there.

⁵⁴ This is why the iSCSI story is so attractive ... everyone has an existing Ethernet/IP LAN.

of Layer 3 Ethernet switches in the basement of the building (redundant distribution layer routers⁵⁵), which in turn are dual-homed to a pair of Layer 3 Ethernet switches in the campus core (redundant core layer). Interlinks between switches and between layers (access layer to distribution layer and distribution layer to core layer) consist of multiple Gigabit Ethernet pipes. We route heavily, i.e. we do not permit VLANs to spread beyond a single switch (or beyond the redundant pair of switches feeding a server room).

When we spread a private network past a single switch (or past a pair of redundant switches), we utilize ‘ships in the night’ routing (aka Virtual Route Forwarding) in our distribution and core layers to logically isolate the private network from other private networks and from the commodity network. We call these things ‘V-Nets’, part of a larger model for segregating end-stations which we call ‘Tiered Networking’ (see NAG session notes on this topic).

In this section, I want to focus on meeting our iSCSI needs for the next five years, i.e. for the life of the network we are currently installing, and I want to assume that iSCSI becomes a popular, though not exclusive, transport for SCSI frames. I am hoping that we can use this section to facilitate to develop consensus around how to design our support for iSCSI.

HOST-BASED ROUTING

The introduction of iSCSI has implications on host-based routing tables. With Ethernet NICs carrying commodity IP traffic and Fibre Channel NICs carrying storage traffic, host-based routing are unaffected – the IP and FC routing tables behave as ‘ships in the night’ processes and do not affect one another. With iSCSI, storage traffic can follow the same path as can all other IP traffic. For a host equipped with a single NIC, this presents no issue. But if sys admins want to dedicate one or more NICs to iSCSI traffic, then they must find a way to bind the iSCSI stack to one or more IP addresses, associated with one or more NICs.⁵⁶

MICROSOFT ISCSI INITIATOR

Microsoft has customers for whom NIC TEAMing does not work reliably; in an effort to reduce their iSCSI group’s support costs (they don’t want to trouble-shoot customer TEAMing problems), Microsoft has written a check into the initiator, such that it will refuse to run over TEAMed NICs. Thus, for highly available Windows servers, equipped with TEAMed NICs carrying commodity traffic, iSCSI traffic must be redirected across physically separate (non-TEAMed) NICs.⁵⁷

THE CASE FOR A PRIVATE V-NET

In this design, stations wishing to speak iSCSI are equipped with four NICs. Two of those NICs receive commodity network IP addresses and are accessible to end-users; one is plugged into

⁵⁵ The access-layer switches servicing users are also dual-homed to these redundant routers.

⁵⁶ The same issue arises if sys admins want to employ a single NIC but a separate IP address for iSCSI traffic.

⁵⁷ This has the effect of raising the costs of iSCSI implementations in Microsoft environments (doubling cabling, Ethernet NIC, and Ethernet switch port inventory).

switch 'a', the other into switch 'b'. The remaining two NICs receive private addresses specific to a proposed iSCSI V-Net and again are distributed amongst the two switches in the server room housing the device. This is a popular recommendation from vendors.⁵⁸

Pros

- Various over-the-wire attacks are mitigated. (a) For a worm exploiting an iSCSI vulnerability to attack other iSCSI nodes, it would have to find its way onto a device plugged into the iSCSI V-Net, a small subset of the total devices on our network, (b) same goes for man-in-the-middle attacks.
- Various sys admin initiated DoS events are mitigated: (a) assigning a duplicate IP address to one being used by an iSCSI device, (b) the installing sys admin accidentally leaving a storage device wide open and another sys admin accidentally assigns a LUN to it, (c) security nerds scanning the network with brutal tools, like Nmap or Nessus, from their workstations cannot reach a production iSCSI node.
- SCSI traffic is special because when it fails, the LUN corrupts, and a corrupted LUN means lost data (rolling back to the previous backup). Isolating LUNs to a private V-Net shields them, partially, from interruptions originating in the commodity network.
- The use of VLAN tagging permits deploying QoS for iSCSI traffic.⁵⁹
- Increases the likelihood that we could deploy the 'jumbo frames' feature.
- Reduces the chances of clear-text iSCSI traffic being flooded to other ports equipped with sniffers and hostile eyes.
- Smooths the use of NetApp tech support (who recommend the use of private networks to carry iSCSI traffic).

Cons

- Requires managing 'difference' amongst ports. Currently, all Ethernet ports are configured the same⁶⁰; this allows the physical layer people to punch down connections without knowing what sits behind them. Similarly, this allows the logical layer people to configure switches without paying attention to port-specific configuration. Creating port-specific configurations will lead to lowered reliability (physical layer and logical layer

⁵⁸ When asked privately about this, I have heard vendors explain to me that this approach reduces their tech support burden by simplifying, from a storage point of view, the customer's environment. In fact, these vendors prefer for customers to implement a private physical network for storage, an even simpler arrangement, from a storage point of view. Of course, this is a benefit for the vendor, not for the customer, who experiences an increase in either complexity or cost from these strategies.

⁵⁹ In our environment, the only frames which currently receive 'different' tags are VoIP frames and SCHARP frames (i.e. frames isolated to the SCHARP V-Net).

⁶⁰ This is inaccurate. Here are the current exceptions: Wireless Access Points (require additional VLAN trunking statement), Indigo (requires LACP statements), CSS internal network (requires private VLAN statements), SCHARP internal network (requires private VLAN statements).

people will make mistakes and trouble-shooting problems will take longer). How much? Don't know.

- Requires heating up a V-Net, with its associated complexity in switch & router configuration and the addition of yet another routing table. This affects the logical layer people. As usual, additional complexity leads to lowered reliability (more errors on install; increased Mean Time To Repair). The key question again is: how much? Don't know.

Counterarguments

- Yes, isolating iSCSI to a private V-Net mitigates various attacks but so too would isolating all our servers to a private V-Net ... and our Microsoft servers and aging Unix boxes are far more vulnerable to attack than any iSCSI stack. This argument can be extended to propose that we isolate all our IIS servers to a single V-Net ... and all our SQL Servers to another V-Net ... all our Apache servers to yet another V-Net ... and so forth. Why is iSCSI special?
- Yes, isolating iSCSI to a private V-Net mitigates DoS events triggered by sys admins ... but again, this argument can be used to apply to all our servers. Why is iSCSI special? And if an administrator doesn't use LUN-masking and/or authentication to protect his/her storage devices from marauding colleagues ... then isolating such a system to a private V-Net won't help much ... that's the V-Net populated exclusively by iSCSI devices ... replete with sys admins configuring initiators pointing at LUNs ... i.e. the very community most likely to direct an iSCSI initiator to the wrong LUN.
- OK, so here is why iSCSI is special. And here is some push-back: so SCSI traffic is special because it carries data and, when mangled, corrupts LUNs ... well, what happens when SQL Server crashes? Does SQL Server carry data, and doesn't mangled data lead to corruption of data? Or are you confident that transaction logs save the data in this case? How about CIFS or NFS servers?
- QoS: identify a candidate for such treatment, specifically, identify an application here which requires consistent sub-millisecond latency. If we have such applications, then we need to know about them – they could have a major impact on how we design storage transport services. And if we don't have such applications, then perhaps we don't need to tag iSCSI frames for higher priority transport.
- Jumbo frames: ok, as I understand it, jumbo frames matter when one performs sequential I/O ... and only 20% of the traffic in a typical environment is sequential. Seems to me that jumbo frames wouldn't make a big difference, in our environment. Furthermore, according to my memory, the NetApp folks tested sequential read/write in our environment, and came darn close to the maximum throughput of GigE, suggesting that implementing jumbo frames couldn't improve performance by much ... since our gear (NetApp plus current network) delivers close to the ceiling already.

- Yup. Under some circumstances, generally pathological, switches can flood traffic, and if a host were compromised and running a packet capture program, the attacker could acquire clear-text data. However, (a) this is rare, and (b) plenty of applications are pushing clear-text data ... Microsoft's CIFS for example, arguably the most wide-spread carrier of sensitive data here, a protocol without support for encryption. Why spend a lot of effort protecting against an unusual and pathological case when that effort won't protect the vast majority of data flowing across our wires?
- The IArchive people argued against this, saying that NetApp, while recommending private networks, will support iSCSI carried on commodity networks too – their requirement is not private networks but rather IP connectivity.

THE CASE FOR COMMODITY MINGLING

In this design, stations wishing to speak iSCSI are equipped with two NICs or, optionally, four NICs. All NICs receive commodity addresses and are accessible to end-users. iSCSI frames are tagged just like any other frame. We would overlay a separate logical network (10.111.0.0/16) on top of our commodity IP space (140.107.0.0/16), permitting, though not requiring, sys admins to split commodity NICs and iSCSI NICs across separate IP networks. There is no security win from doing this, of course, as the routers would route both 140.107.0.0/16 and 10.111.0.0/16 networks within the same routing process.

Pros

- NIC proliferation becomes optional – the sys admin can choose to employ two NICs or can choose to employ four NICs.⁶¹
- Sustains the current level of reliability, from a cabling/switch/router management point of view.

Cons

- Does not meet 'best practices' as articulated by the leading player in our iSCSI environment (e.g. NetApp).

Counterarguments

- As we pile more and more services and applications on top of our infrastructure, we increase complexity, no doubt about it. But this is our direction, impeding progress with cries of 'no more complexity' doesn't serve our end-users needs. We are best served by constructing designs which slow the rise of complexity while delivering functionality.

⁶¹ If you believe NetApp peoples' numbers, then the best a high-end x86 box can do is 85MB/s ... and we already know, from the MRTG graphs, that our x86 end-user boxes see trivial throughput. Putting these details together ... I would argue that our applications don't push enough bits to warrant separate NICs (i.e. a NIC for iSCSI and a NIC for commodity).

- We employ many more sys admins than we do network admins ... by a factor of ten or more. If we increase complexity for network admins, *but reduce complexity for sys admins*, then we have saved staff time, overall.
- Keep the end-stations simple; push the complexity into the infrastructure in the middle.
- Is deploying iSCSI over a private V-Net a big win? Probably not. But every time we can reduce complexity and increase reliability for the sys admin (even at the expense of complexity for the network admin), we have made a step forward in reducing the overall complexity of our environment.

THE CASE FOR SERVER-WIDE V-NET

In this model, we take server-oriented security concerns seriously and migrate all servers to a private V-Net, protected by a firewall which permits end-user access only to necessary services; in our Tiered Networking model we call this 'HarshNet'. Simultaneously, we move all externally-accessible servers into 'The Pit', creating a V-Net to service it.

Pros

- Hardens the server environment, shielding it (partially) from incursions originating from user stations.
- Shields internal servers from hacks occurring on externally visible servers.

Cons

- That's a lot of work, both for set-up and for on-going support.
- How do sys admins reach their servers, once they are behind a firewall? Can we live with a firewall separating our end-users from our servers?

THE CASE FOR FUNCTION-SPECIFIC SERVER ROOMS

In this model, we dedicate rooms to functions. One server room becomes 'The Pit', another becomes 'iSCSI-land', etc. Servers located in The Pit are accessible from the outside world but cannot initiate connections to our internal network⁶². The switches in the 'iSCSI' server room support a private VLAN dedicated to iSCSI (and perhaps other private VLANs, dedicated to other functions). This model ignores the capabilities of V-Nets and employs physical proximity to reduce complexity. (i.e. the switches servicing the 'iSCSI' server room would contain port-specific configurations ... but no other server room switches would need to support this complexity).

⁶² Some folks call this function a 'DMZ' ... but I prefer to avoid this term, because the word is overloaded: networking geeks use it one way, sys admins use it another way. I use the term 'Meet Me Zone' to refer to a transport only network which interconnects separate business entities and the term 'Pit' to refer to a network which is firewalled from both the Internet and the corporate network.

Pros

- Simple to understand.

Cons

- Risks pushing capacity (i.e. what if the server room hosting ‘The Pit’ fills up?)
- Requires moving servers.