"Space: The Final Frontier"

Christopher R. Hertel

ubiqx consulting, inc

March, 2008

Introduction

Who am I?

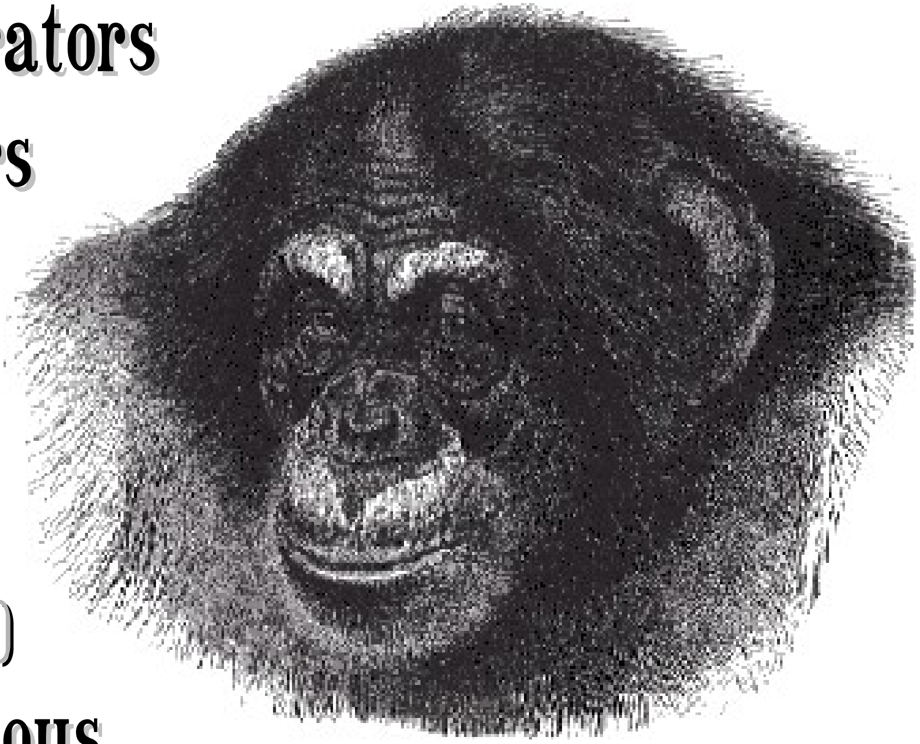
- ☆ Network Geek
- ☆ Storage Geek
- ☆ Samba/CIFS Geek
- ☆ Author (shameless plug)
- ☆ Incurable Idealist

A ruminant mammal (Geekus geekus) with long legs, humped shoulders, and broadly palmated antlers.

Introduction

Who are You?

- System Administrators
- Network Managers
- Security Geeks
- Students
- Coders
- Hackers (per RFC 1392)
- The Morbidly Curious



Introduction

Where are we going?

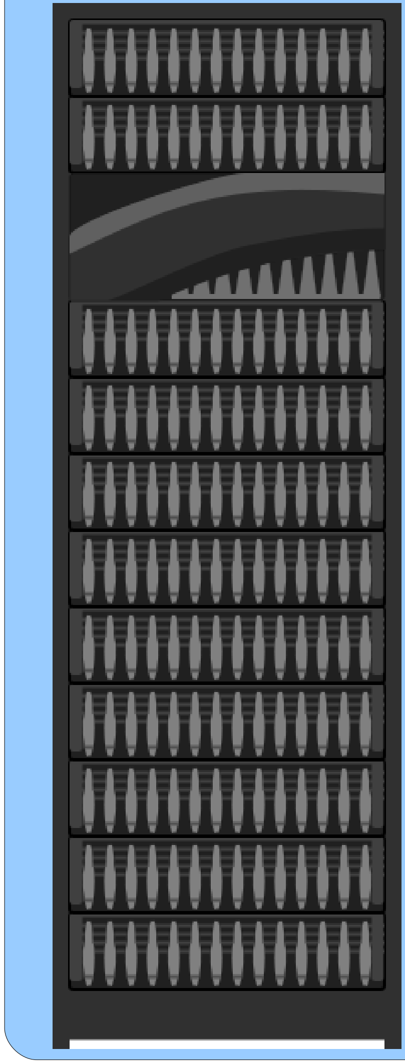
A Tour of Storage Technologies:

- Disk– 51.5 Years Young
- SAN – Shared Block Storage
- NAS – Networked File Systems
- Other Things You Will Encounter in your Travels.

A Place for your Stuff

(That's really what disk drives are all about.)

A Place for your Stuff



Sidebar

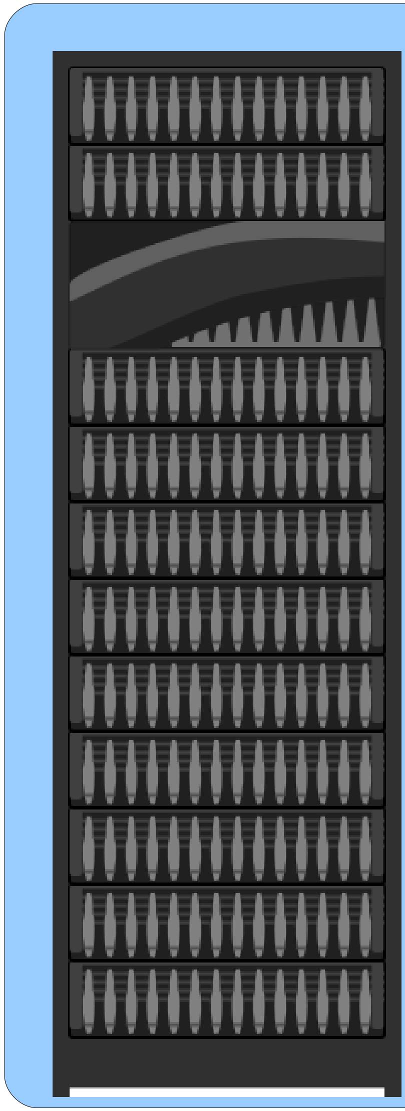
Disk-o-matic Math

Drive makers measure by 1000, not 1024.

1PB	= 1000TB	= 909.5 “real” TB
1TB	= 1000GB	= 931.3 “real” GB
1GB	= 1000MB	= 953.7 “real” MB
1MB	= 1000KB	= 976.5 “real” KB
1KB	= 1000B	

Operating Systems typically use powers of 2 (e.g., $2^{10} = 1024$). One “real” Petabyte = 2^{50} bytes.

A Place for your Stuff



Sidebar...continued

Disk-o-matic Math

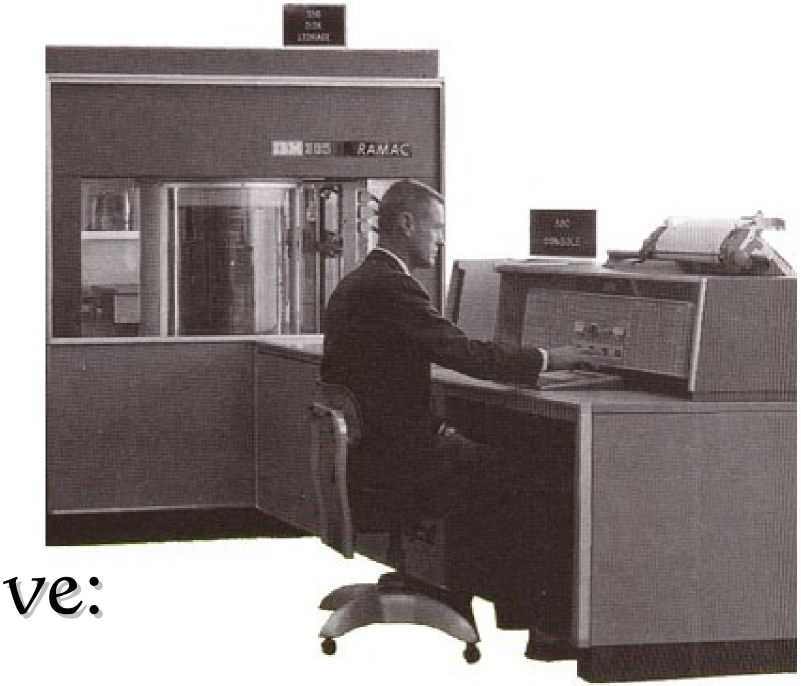
Redundancy further reduces “real” capacity:

RAID 1 (mirrored)	n/2
RAID 5 (parity)	(n-1)/n
RAID 6 (2xparity)	(n-2)/n

Be careful with your calculations!
Know what you're really getting.

IBM RAMAC (4-Sept-1956)

Random **A**ccess **M**ethod of **A**ccounting and **C**ontrol



Original Disk Drive:

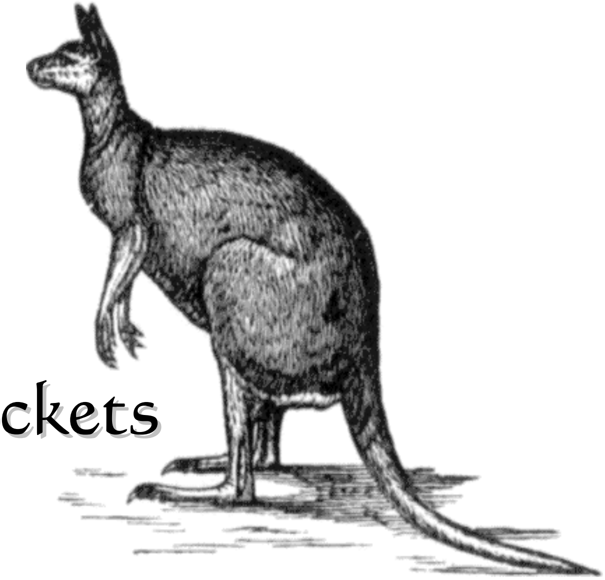
- Fifty 24″ Platters
- Less Than Five Megabytes (4.4MB)

A Place for your Stuff

25 years ago: 10MB was *a lot* of disk space.

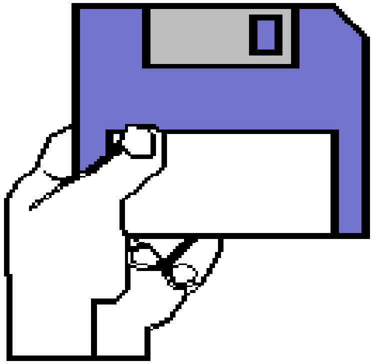
Today: I've got at least 2TB at home.

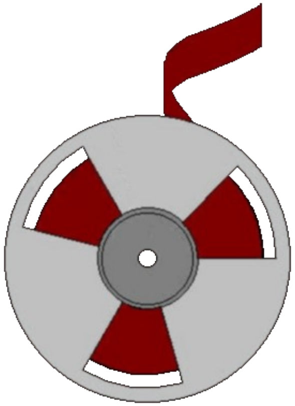
- 3.5″ Drives are < 20¢/GB
- Enterprise Storage is measured in Petabytes
- We carry Gigabytes in our pockets



Storage capacity, like computing power, has grown such that we can now hold in our hands what used to require a computer room *and* a team of experts.

A Place for your Stuff





In our increasingly digital world:

- We keep getting more Digital Stuff (data)
- Our Digital Stuff keeps getting bigger (Gigs)
- We worry about keeping our Digital Stuff safe
- We have trouble keeping track of Digital Stuff

A Place for your Stuff

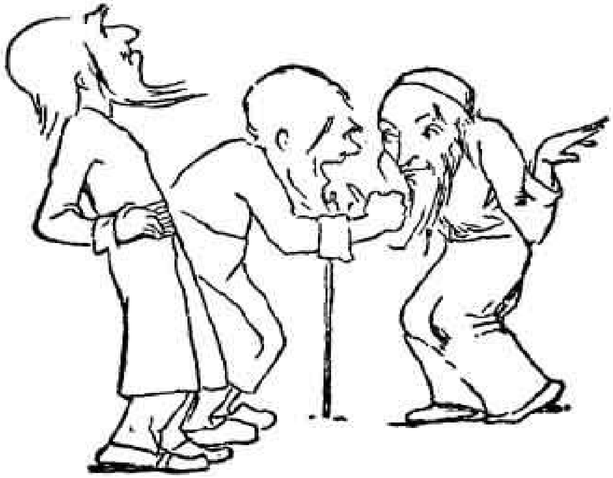
All of that storage...

...scattered all over the home

...scattered all around the office

...scattered all across the Internet

How do we handle it all?



A Place for your Stuff

"The problems that the Lunatic Fringe is working on today are the problems that the mainstream storage industry will face in 5-10 years."

➔ Tom Ruwart,
Storage on the Lunatic Fringe

(He's right, you know.)

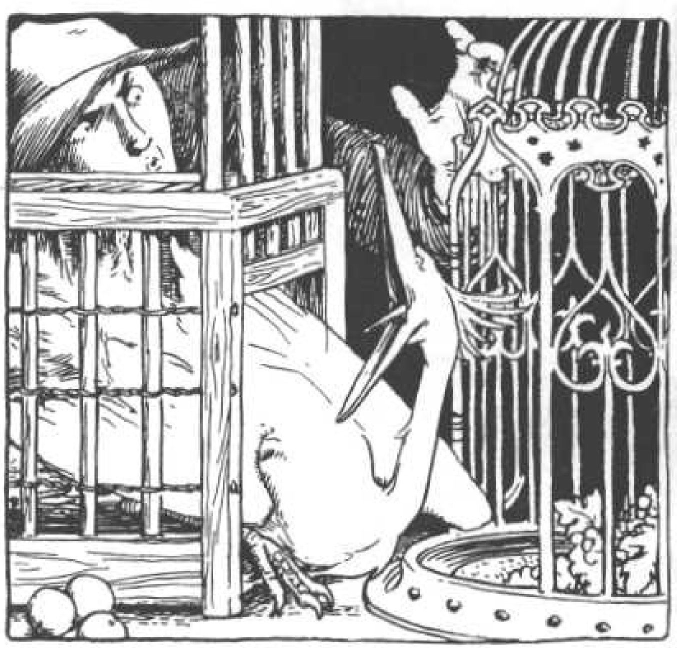
Storage on the Lunatic Fringe
http://www.dtc.umn.edu/resources/ruwart.ppt

A Place for your Stuff

Hertel's Corollary: The large-scale storage problems of yesterday afternoon have already become the home office / small office storage problems of early this morning.

Storage subsystems supporting 1, 2, or 4 drives are now common and available at commodity prices.

What's good for the goose...



Benefits of consolidated storage for small-end users:

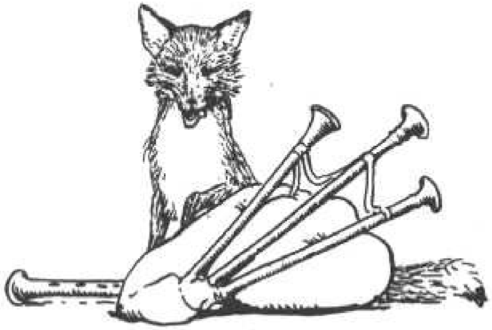
- Centralized management
- Efficient use of resources
- Data protection (RAID / Backup / Archive)
- Failure isolation

There are problems with centralization, so a mix of local and central storage is often the most workable choice.

User Interface is critical!

If it's not automagical, are you really going to use it?

- Automatic backup & archive
- File categorization & search
- Privacy & security
- Semi-Automatic Update
- Service Alerts
- Worldwide access



SCSI vs. IDE

The Epic Struggle

SCSI vs. IDE

- PC Revolution of the 1980's

  Several disk drive interfaces make their debut:

  - SMD, SMD-E
  - ESDI
  - ST506 ==> IDE

  - SASI ==> SCSI

As these evolved, Controller functionality was moved from the Host Adapter to the Drive.

Standards lead to interoperability.

- SCSI & IDE...

  are managed by the same standards body:

  - SCSI ==> T10
  - ATA ==> T13

IDE = Integrated Drive Electronics
ATA = AT Attachment
SCSI = Small Computer Systems Interface

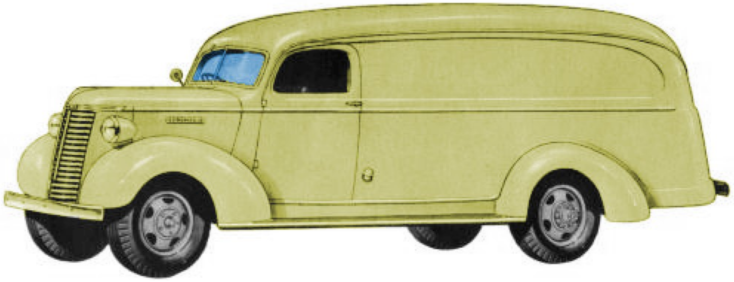
SCSI vs. IDE

IDE (ATA) only supports disk drives

- One-bit bus address space (master/slave)

SCSI has a wider command set

- Bus address space depends on SCSI version (minimum 4-bit)
- Multiple Logical Units (LUs or LUNs) per target
- Support for:
  - CD-Rom Drives
  - Tape Robots and Drives
  - More stuff



SCSI vs. IDE

Just to confuse things:

ATAPI = ATA with Packet Interface

- SCSI commands can be wrapped inside ATAPI commands
- Devices other than disk drives can be controlled (e.g., ATAPI CD-ROM/DVD drives)

This brings up an interesting point:

ATA and SCSI commands carried over other transports

- Fibre Channel == SCSI over Fiber (T11)
- iSCSI == SCSI over TCP/IP (IETF)
- ATAoE == ATA over Ethernet (Coraid Corp.)

☞ The current ATA specification is ATA/ATAPI-7.

☞ The current SCSI specification is SCSI-3.

SCSI vs. IDE

Personal Storage (ATA)

- Consumer Prices
  - "Low cost dominates the design[2]"
- Commodity Parts
  - Simple to buy and to replace
- Individual operation
  - Generally used one at a time, not in groups

Enterprise Storage (SCSI)

- Enterprise Prices
  - Customers willing to pay for higher reliability and performance
- Multiuser / Multi-disk environments
  - Server Farms and Disk Arrays
- I/O tends to be more random
  - Small chunks of larger objects (RAID stripes)

SCSI vs. IDE

Higher rotational speed means lower latency

- Smaller platter sizes support faster spindles & lower seek time
- More platter mass means more energy used

More platters provide higher capacity

- Increased spindle mass requires more power to spin
- Increased actuator mass slows down seeks
- Tracks are too fine for “cylinders” to align

Higher bus bandwidth improves throughput

- Increased complexity to disk-side electronics

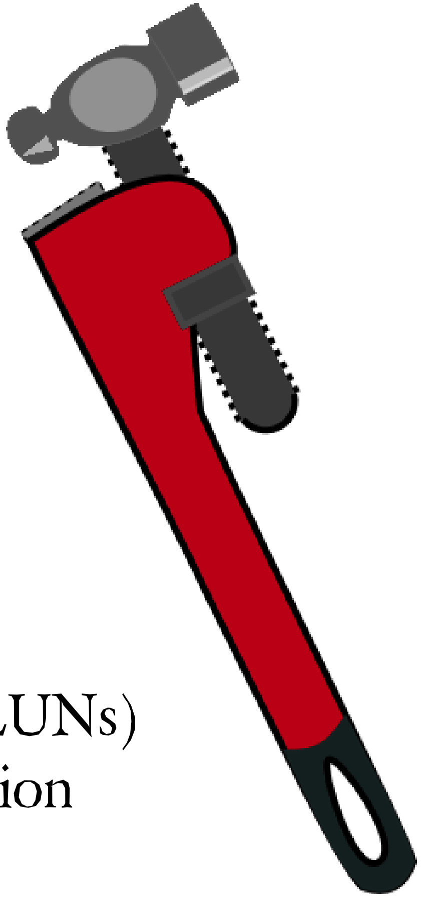
There are always trade-offs.

Environmental Hazards

- Servers and Large Arrays
  - Adjacent drives annoy one another with vibrations
  - Heat
  - 24x7 operation
  - “Hot spindle” rebuilds
- Desktop Systems
  - On/Off operation
- Laptops
  - Shock

Harsher environments create a requirement for higher-quality parts in order to maintain reliability.

SCSI vs. ATA Command Sets



- SCSI
  - Supports many devices (including graphics!?)
  - Designed for many-to-many operation
  - Robust Diagnostics
- ATA
  - Handles disk drives only
  - Very limited address space (master/slave, no LUNs)
  - Designed for one-to-one or one-to-two operation
  - Limited Diagnostics

It's all about choosing the right tool for the job.

SCSI vs. IDE

- Enterprise Drives
  - More expensive electronics
  - Lower capacity, higher performance
  - Better protected (against heat, vibration)
  - Longer-lasting parts
  - New features introduced to meet demand
- Personal Drives
  - Cheap electronics
  - Higher capacity, lower performance
  - Commodity parts
  - Trickle-down technology

The interface is only one difference.

Network Attached Storage

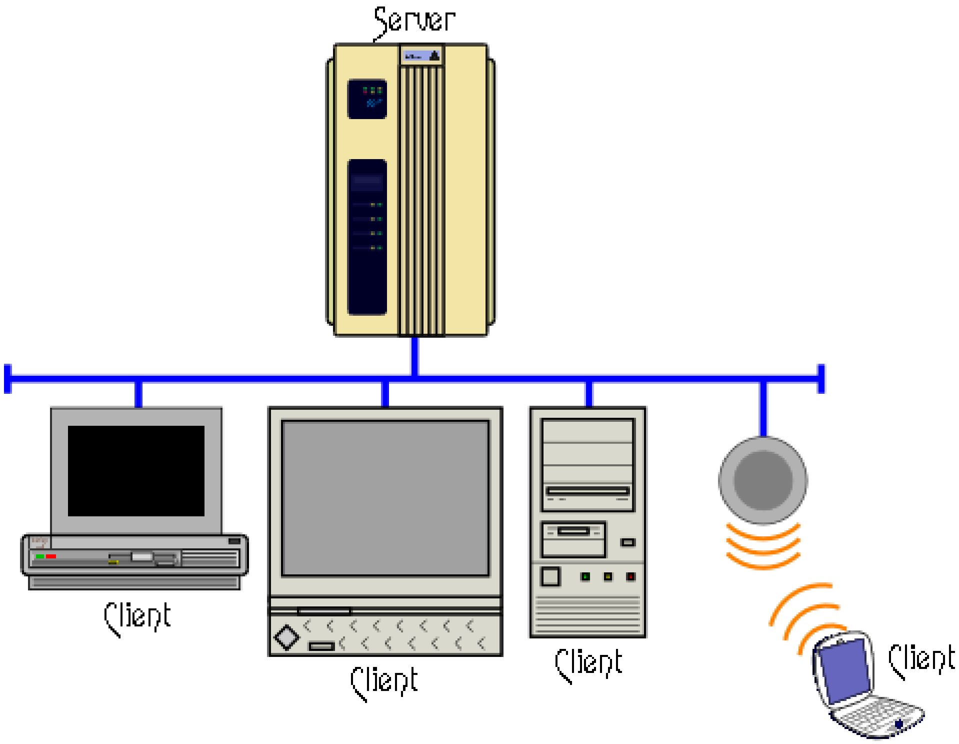
NAS

Familiar NAS Systems:

- IBM's (& MS's) SMB/CIFS Popular!
- Novell's NetWare Fading...
- Apple's Appleshare Fading...
- Sun's NFS Improved!
- IETF WebDAV New!

Local file systems on the server are shared with multiple hosts across a LAN or inter-network.

NAS



Typical client/server NAS

- Large server with local disk
- Multiple clients
- Shared access to files & directories

NAS Concerns:

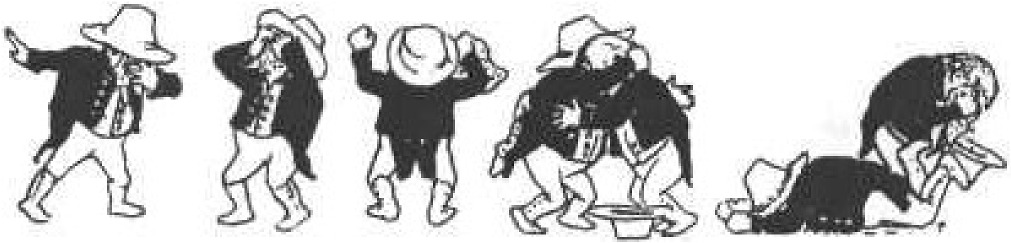
- Authentication, Authorization, & Access Management
- File Locking & Sharing
- Meta-data Semantics

DOS FAT	MacOS	Windows NTFS	Linux/Unix
System, ReadOnly, Hidden, & Archive bits No UID/GID 8.3 Format EOLN: <CR><LF>	Data and Resource Forks EOLN: <CR>	Extended Attributes File Streams SIDs NT ACLs EOLN: <CR><LF>	User, Group, World permission bits UID/GID POSIX ACLs EOLN: <LF>

NAS File Systems are “Vendor Biased”.

Case In Point: CIFS vs. NFS

- For a geek, NFS is easy:
  - Traditionally server-to-server
  - Traditionally geek-to-geek
  - Simple authentication model
- For a user, CIFS is easy:
  - Traditionally user-to-server or peer-to-peer
  - Non-technical user community
  - Specifications & protocol details are hidden



WebDAV

- An extension of HTTP
- Makes the web “read/write”
- Adds only seven new commands
- Messages passed in XML format

The use of XML
allows great flexibility
... and complexity.

“...as simple as possible, but no simpler.”

This is a picture of my cat.

THE NEWS

20-Dec-07: Samba Team Receives Microsoft Protocol Documentation
http://www.groklaw.net/article.php?story=20071220124013919

22-Feb-08: Microsoft Makes Strategic Changes in Technology and Business Practices to Expand Interoperability
http://www.microsoft.com/presspass/presskits/interoperability/default.mspx

- The documentation required by the US and EU anti-trust cases are now available on-line.
- There is still a lot of work to be done to understand what this all means.
- The jCIFS and Samba Teams are already busy reviewing the documentation.

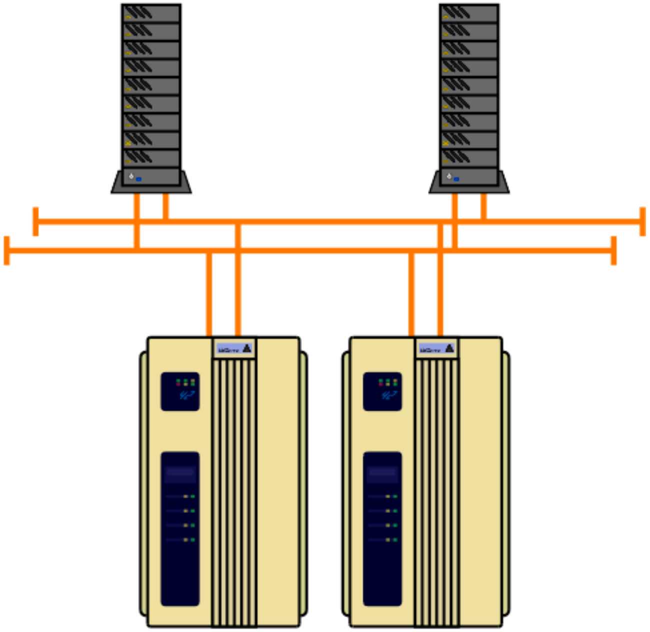
Storage Area Networks

SAN Overview

Precursor: Direct Attached Disk Arrays

- Redundant Array of Inexpensive Disk
- Expandable
- “Virtualizable” (Is that a word?)

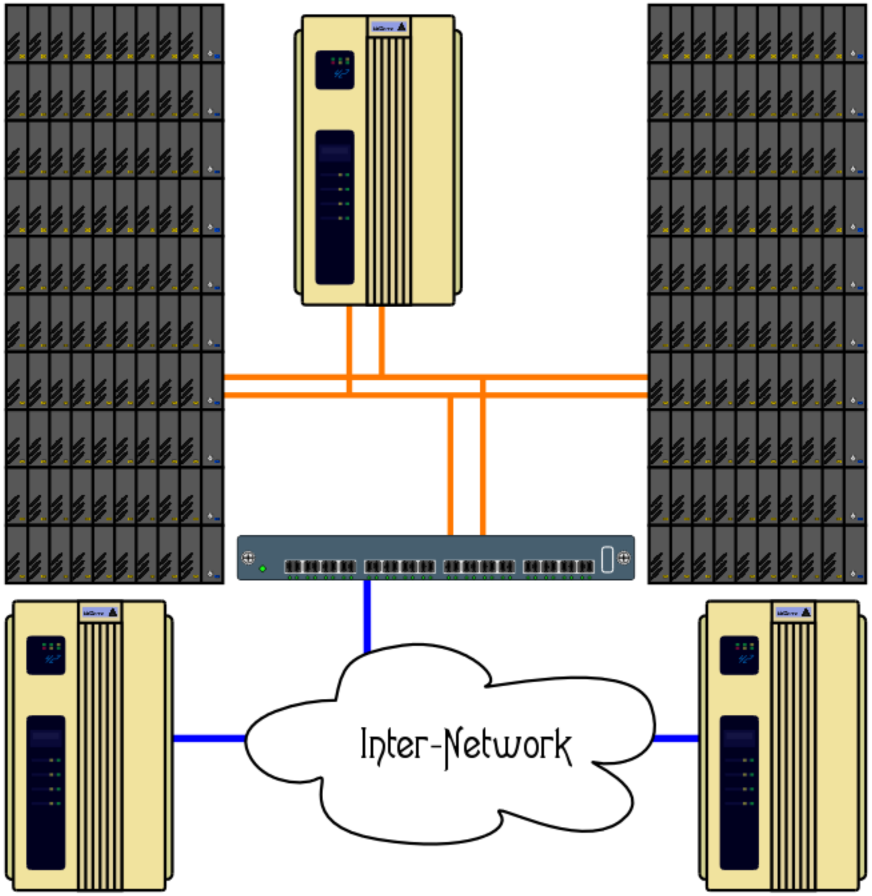
SAN



FibreChannel SANs

- SCSI over Shared/Switched Fiber
- Longer Distances
- 1, 2, 4, and soon 8 Gbps Speeds
- Redundancy

SAN



iSCSI SANs

- Leverage the IP Network
- Coexist with FibreChannel
- Run on Commodity Network Hardware

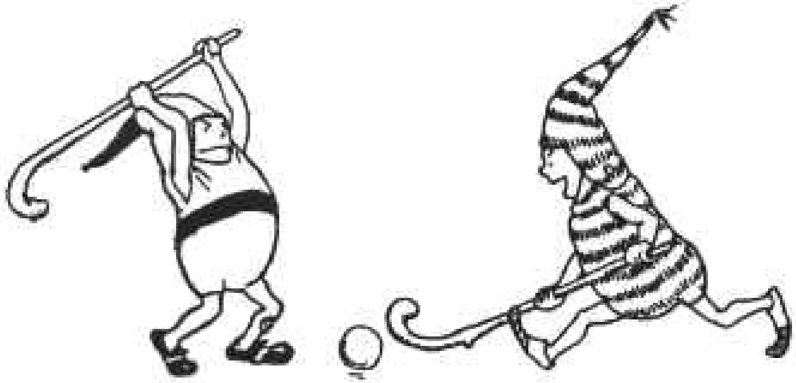
SCSI
is the Traditional SAN “Protocol”

- FibreChannel carries SCSI PDUs
- iSCSI is just SCSI PDUs over TCP/IP

The message is the same;
only the transport changes.

Rivals

- Network Block Dæmon (nbd) for Linux uses TCP/IP as a transport
- AoE (ATA over Ethernet) transports ATA commands over Ethernet frames
- FCoE (Fibre Channel over Ethernet)



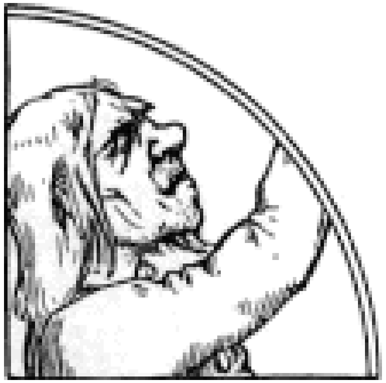
SAN vs. NAS

SAN

- Block Storage
- One-to-One Relationship
- Data-center Oriented
- Space is Not Shared

NAS

- File System Storage
- One-to-Many Relationship
- End-User Oriented
- Data Can Be Shared



Other Stuff

MAID: Massive Array of Idle Disks

- ☞ Cheap Disks (Commodity ATA)
- ☞ Densely Packed
- ☞ Mostly Powered Down
- ☞ Presented as (virtual) Tape Libraries

Idle drives are spun up from time to time to ensure that they don't get stuck.

Diagnostics keep track of “likely failures”.

Other Stuff

ILM: Information Lifecycle Management

- Identify different storage classes
  - high speed vs. low speed
  - high availability vs. high latency
  - expensive vs. cheap
- Monitor data access
- Migrate data (manually/automatically)

For example, migrate from RAID1+0 SCSI drives to RAID5 ATA to Tape.

Other Stuff

Linux: Your Storage Playpen

- Home SAN:
  - ATAoE and iSCSI
- FUSE: User Mode File System Interface
  - E.g.: SSH, FTP, and BitTorrent clients
- Logical Volume Manager (LVM)
- Software RAID
- Lots more cool toys

Other Stuff

Unusual Beyond the Strange

- Cluster File Systems
  - E.g.: Global File System (GFS)
- Distributed File Systems
  - E.g.: Google File System (GFS)
- Object File Systems
  - E.g.: Lustre and UofM T-10 OSD

References

[1] The SCSI Bus & IDE Interface
Friedhelm Schmidt. ISBN-13: 978-0201175141, Addison-Wesley Professional; 2nd Ed., June 17, 1999.

[2] More than an Interface--SCSI vs. ATA
Dave Anderson, Jim Dykes, Erik Riedel. Seagate Technology.
Proceedings of the 2nd Annual Conference on File and Storage Technology (FAST), March 2003
http://www.seagate.com/content/docs/pdf/whitepaper/D2c_More_than_Interface_ATA_vs_SCSI_042003.pdf

[3] Reference Guide – Hard Disk Drives
http://www.storagereview.com/guide/index.html

[4] Implementing CIFS – The Common Internet File System
Christopher R. Hertel. ISBN-10: 013047116X, Prentice Hall PTR, August, 2003. http://www.ubiqx.org/cifs/

The End

Slides available at: `http://ubiqx.org/presentations/`