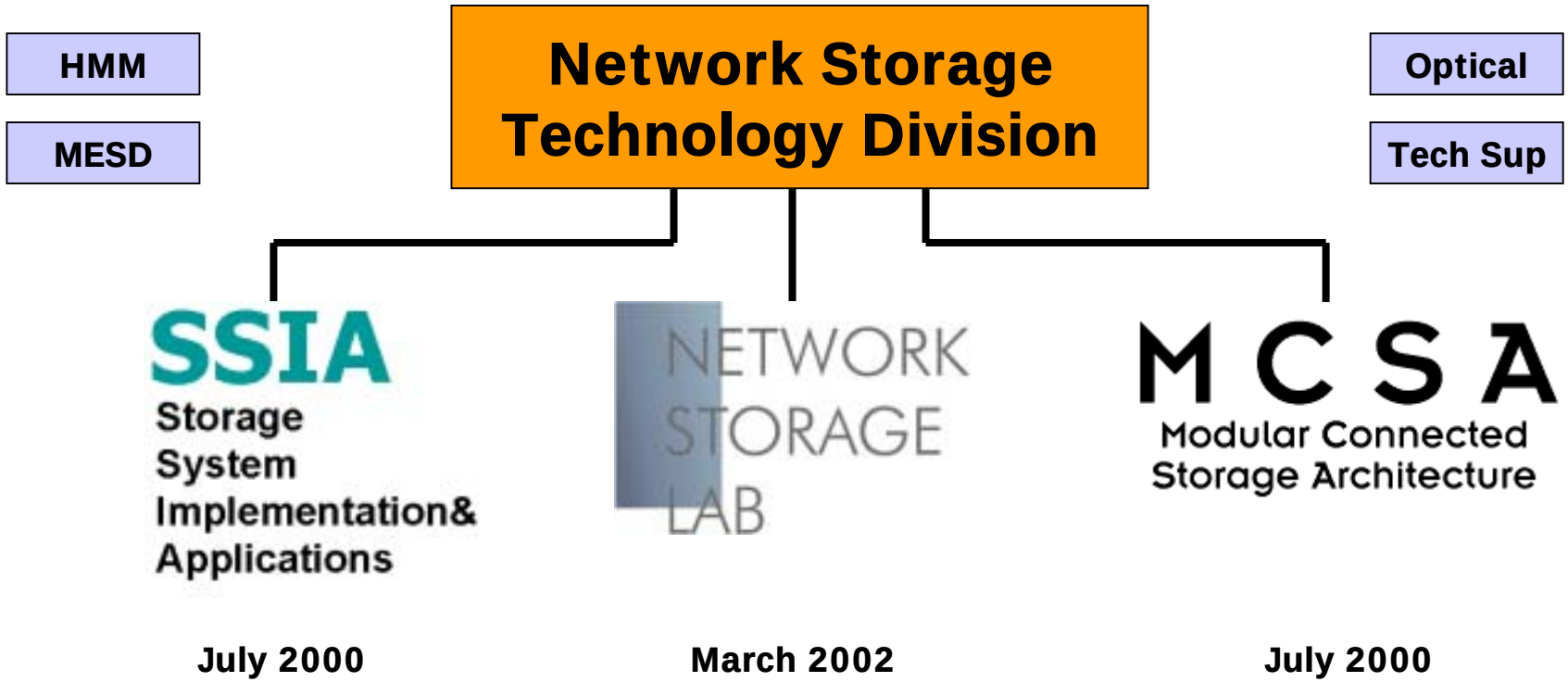
Ethernet Network Storage: **Yes, you can have it now!**

By Patrick Khoo
Program Manager
Modular Connected Storage Architecture Group
Network Storage Technology Division
Data Storage Institute

Data Storage Institute
A*STAR

MCSA
Modular Connected Storage Architecture

About DSI and NST Division



DSI is a nationally funded non-profit R&D institute focusing on data storage technologies and industries

Data Storage Institute
A*STAR

Can I do More for Less?

- University of California at Berkeley - 2001
  - 12 Exabytes in mankind's history to date
  - 12 **more** Exabytes in next two and a half years alone!

115 Video Tapes

100Gbit/in^2 (2002)

1 HDD

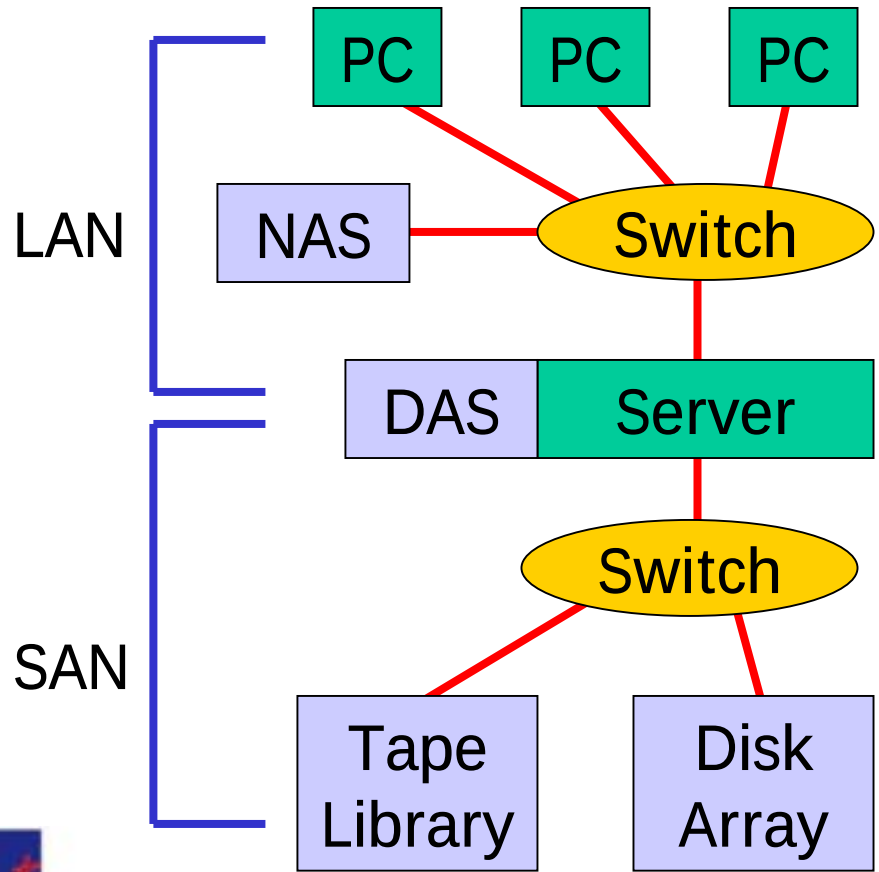
250 hours of video

or

80,000+ songs

Network Storage to the Rescue!

But Does Network Storage **REALLY** Help?



Definitions

LAN - Local Area Network

DAS - Direct Attached Storage

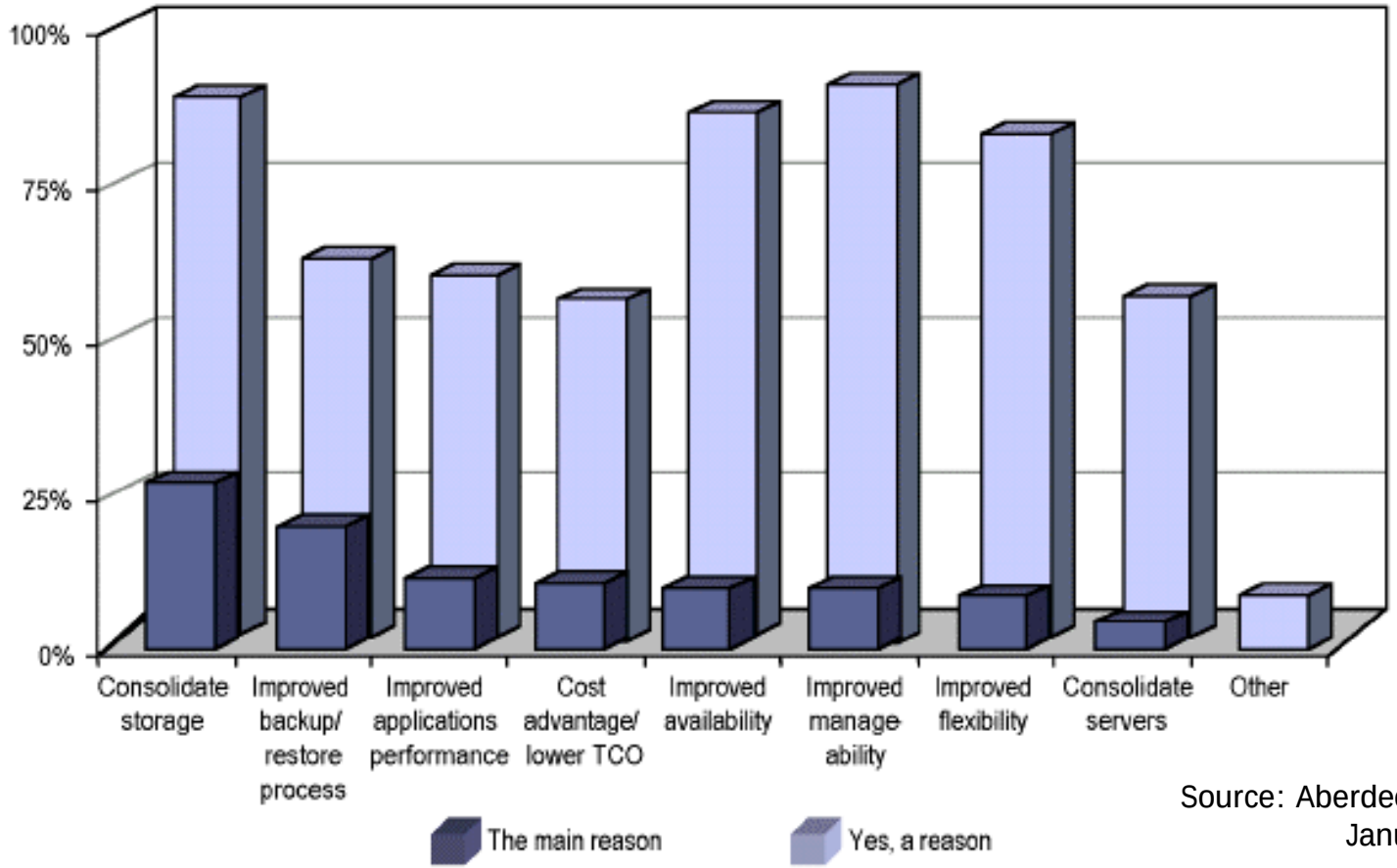
NAS - Network Attached Storage

SAN - Storage Area Network

Components

- Servers
- Storage systems (eg. disk arrays, tape libraries, etc)
- Interconnect technologies (eg. fibre optic cables, switches etc)
- Host-bus Adapters (HBA), Network Interface Cards (NIC)
- System and Data Management Software

Why Build a SAN?



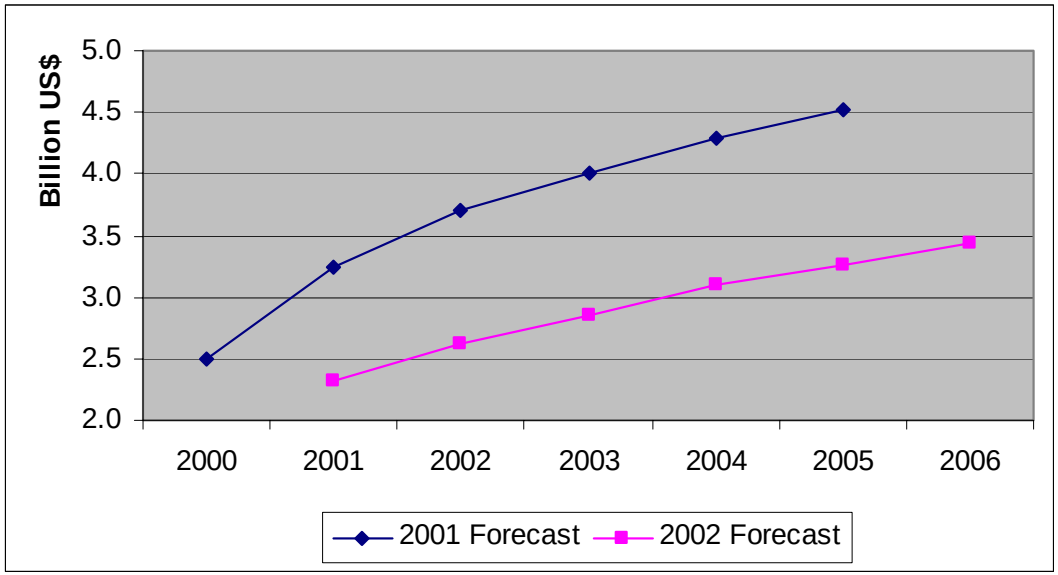
Source: Aberdeen Group, January 2002

The truth is, there is **NO** killer app, so stop waiting for one!

But there are plenty of reasons to adopt Network Storage!

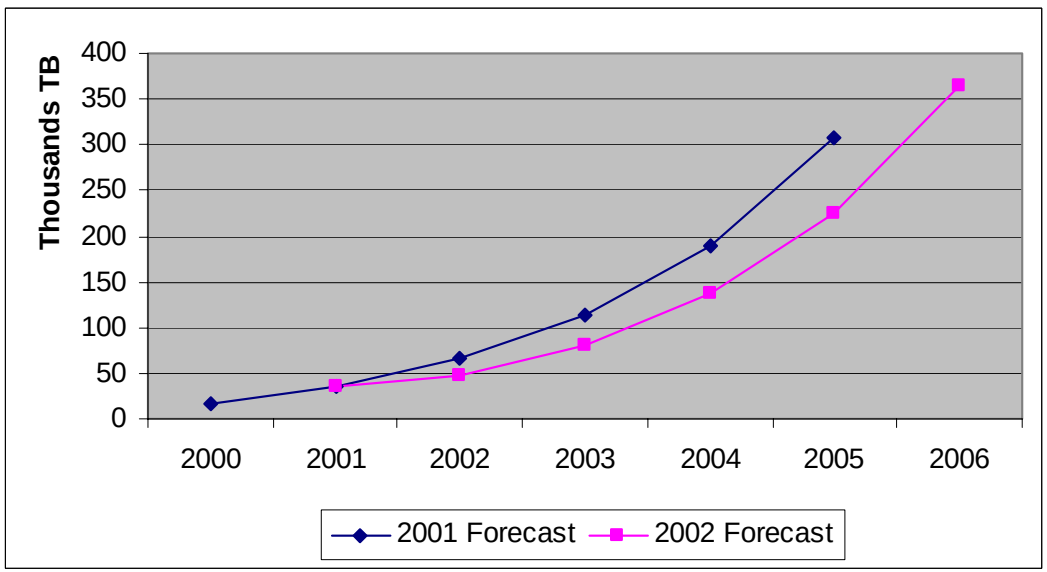
Data Storage Institute
A*STAR

Industry Forecasts – Comparisons



Asia/Pacific Disk Storage Revenue, 2000-2006

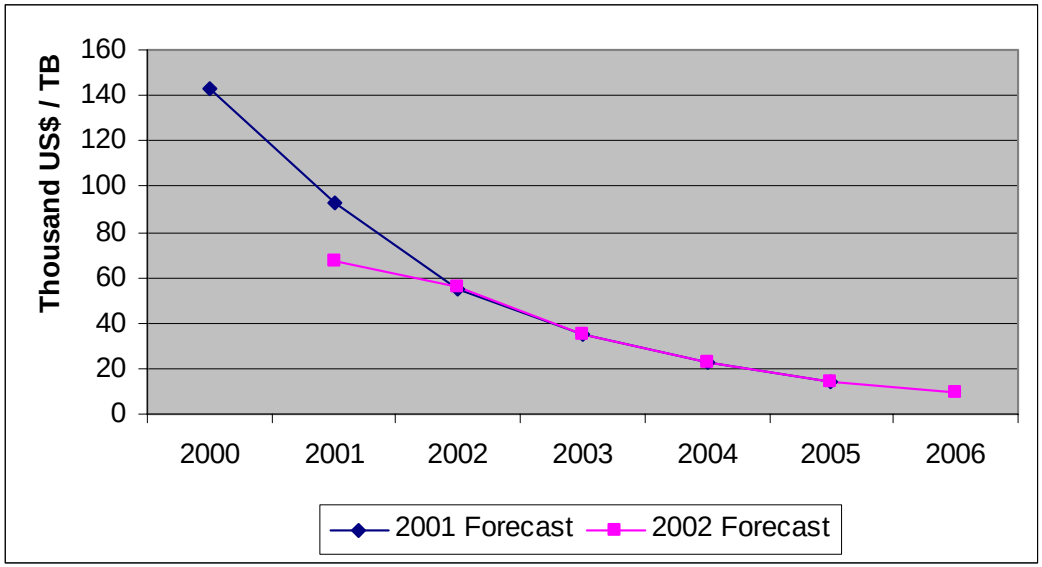
Source: IDC Asia/Pacific, 2001, 2002



Asia/Pacific Disk Storage Terabyte Shipments, 2000-2006

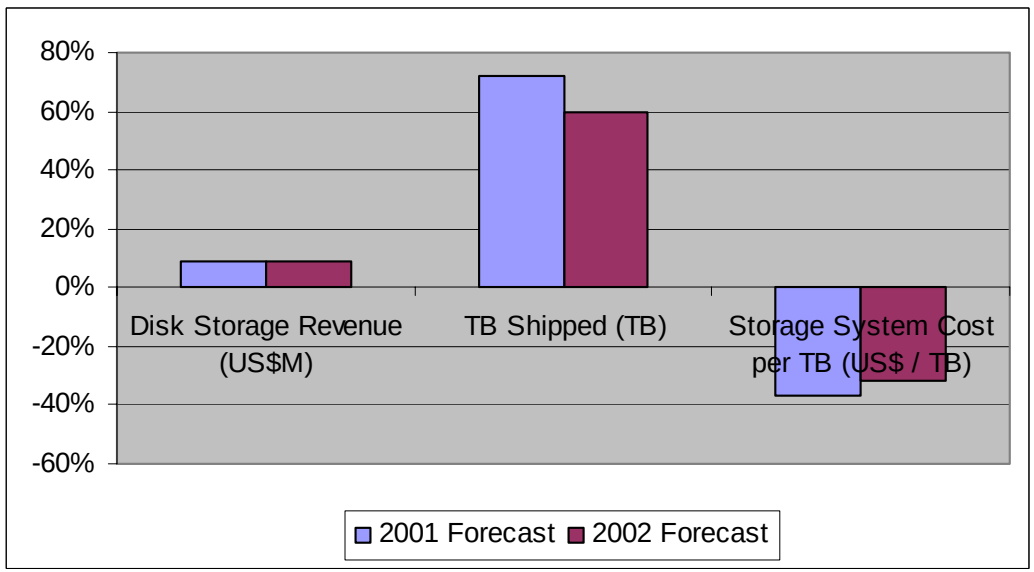
Data Storage Institute
A*STAR

Industry Forecasts – Comparisons



Asia/Pacific USD/TB Storage System Costs, 2000-2006

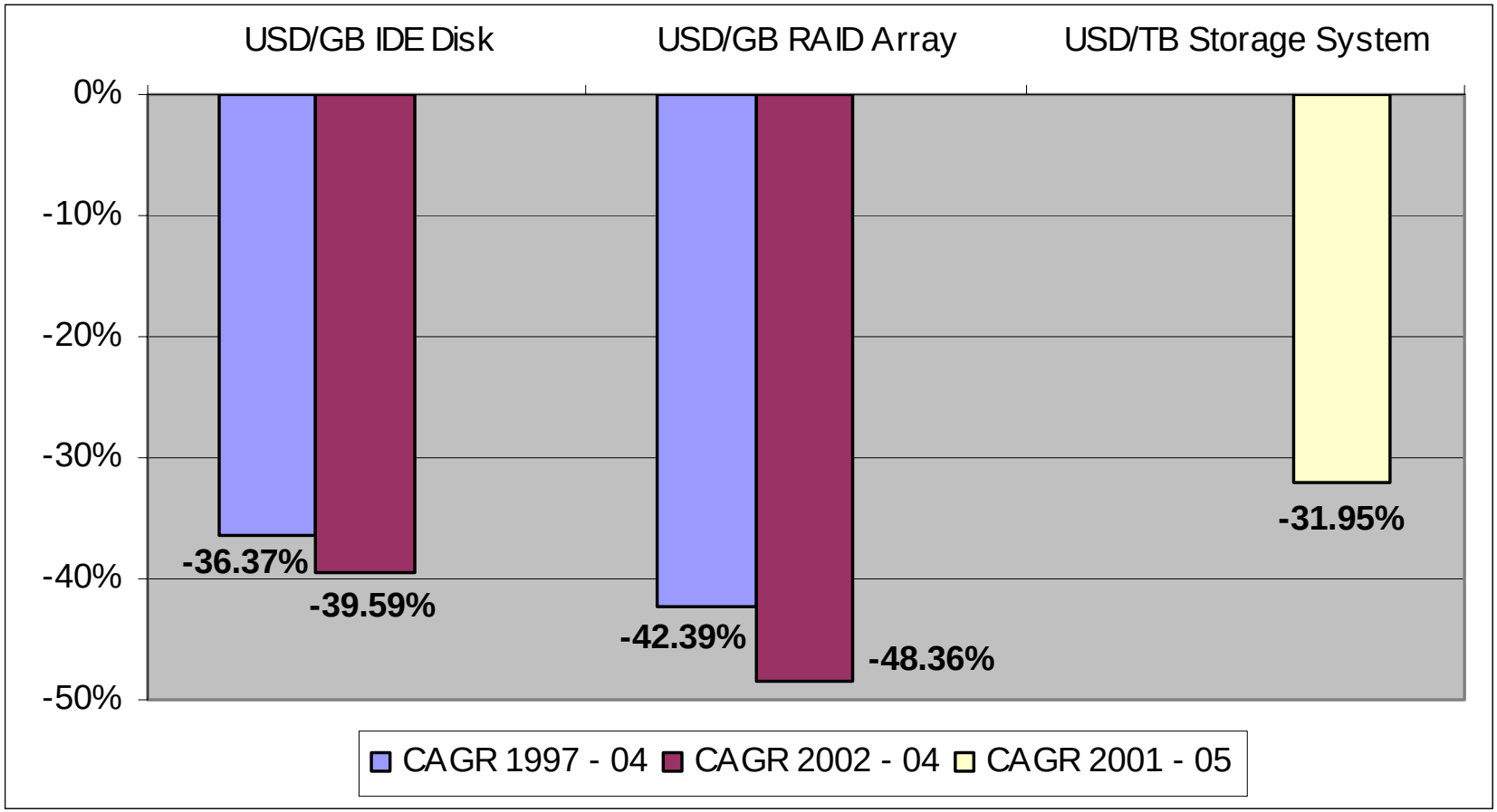
Source: IDC Asia/Pacific, 2001, 2002



Asia/Pacific CAGR Forecasts, 2001-2005

Data Storage Institute
A*STAR

Trends in Storage Price Erosion



Source: IDC Asia/Pacific, 2002

Data Storage Institute
A*STAR

Industry Forecasts – The Difference

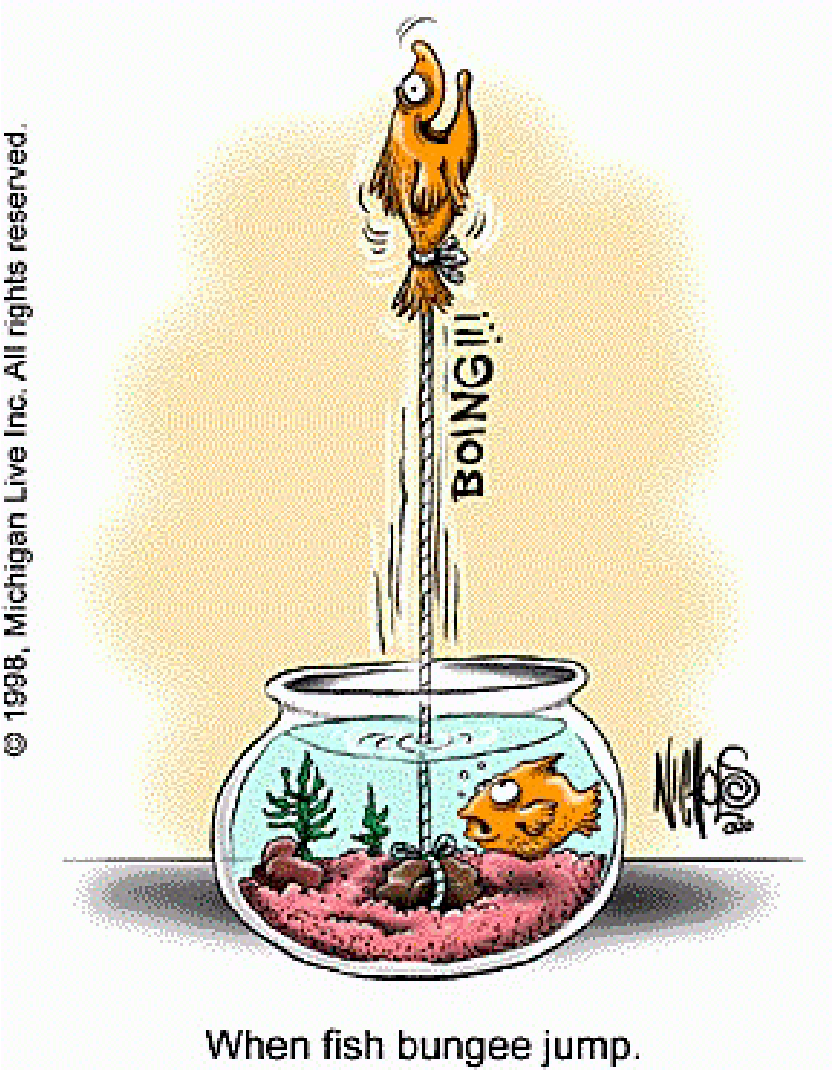
- What is the effect of the current economic climate on the storage business?
  - Storage growth has not been halted, merely delayed (by about one year)
  - At the same time, the cost of storage is dropping about 35% per year but will still stablise over the long term [IDC Asia Pacific 2002]

- Conclusion: The cost of storage <u>systems</u> has not really dropped (other than due to newer high density HDDs) – some kind of technological advancement is needed

So How?

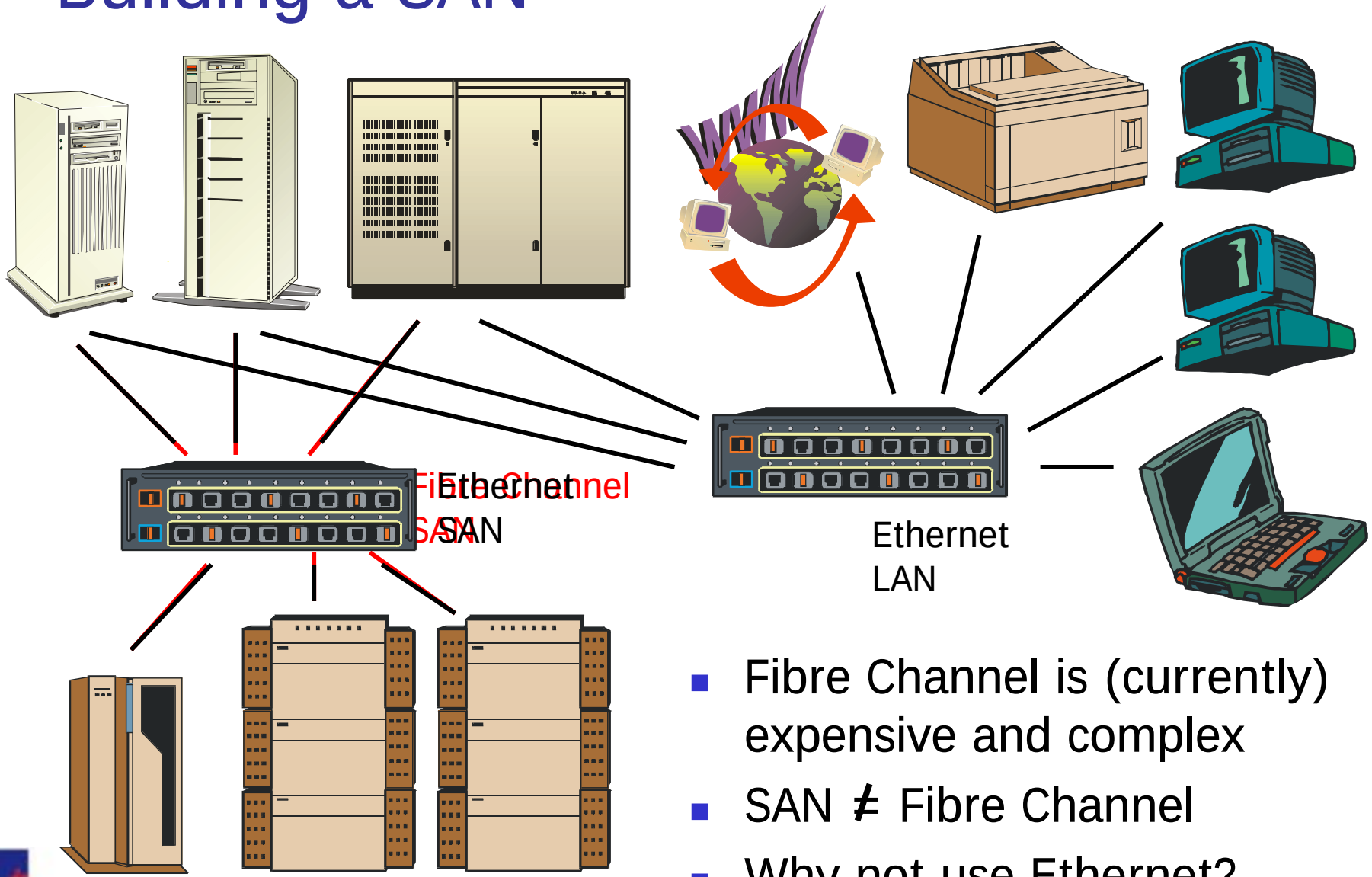
Innovate!

. . . and think of new ways to do old things.



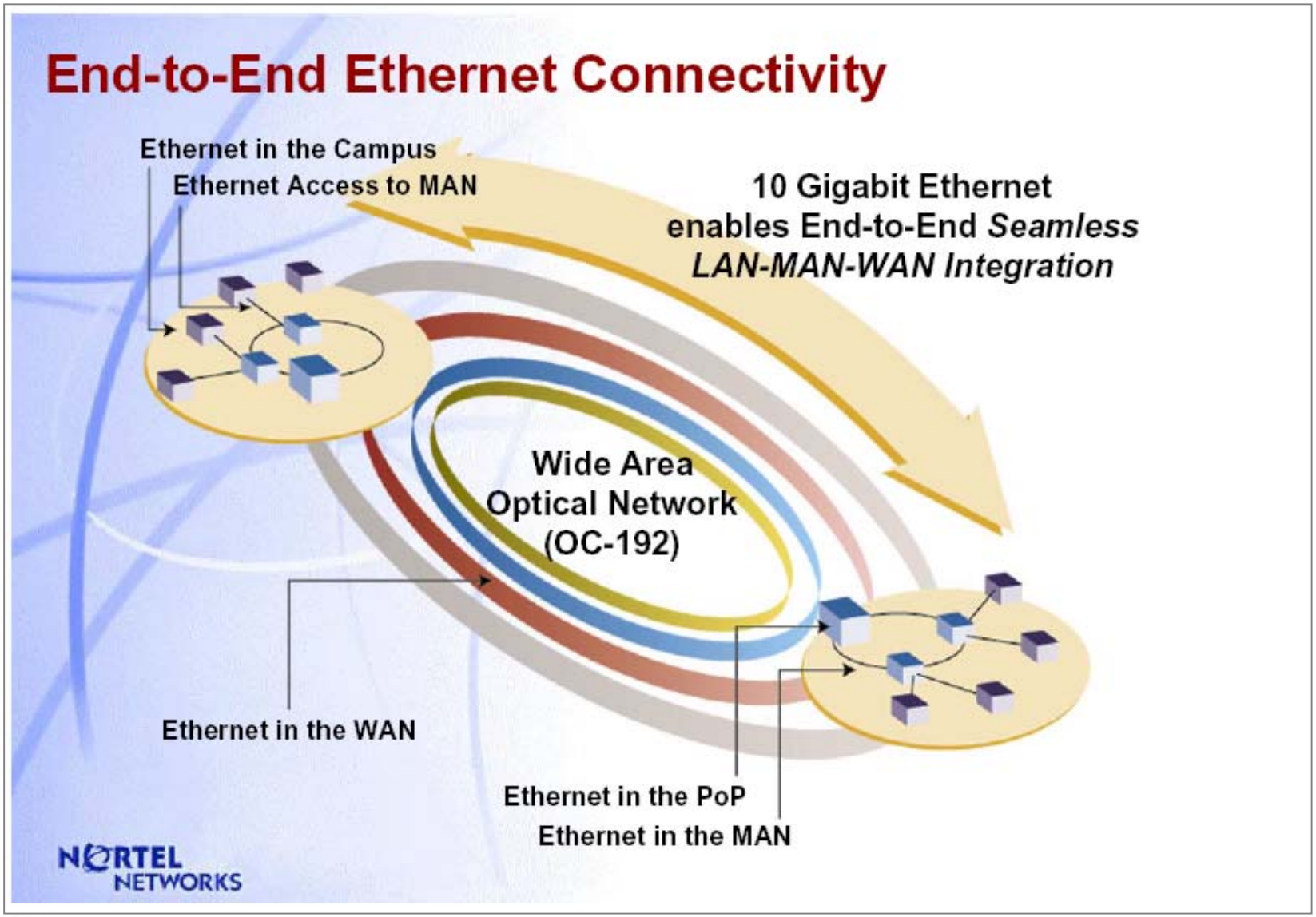
When fish bungee jump.

Building a SAN



- Fibre Channel is (currently) expensive and complex
- SAN ≠ Fibre Channel
- Why not use Ethernet?

“Ethernet the World!”



Network and Storage Differences

You can't compare an Ethernet cable with a SCSI cable, SCSI cables transmit data in parallel!

Overheard from a computer science professor

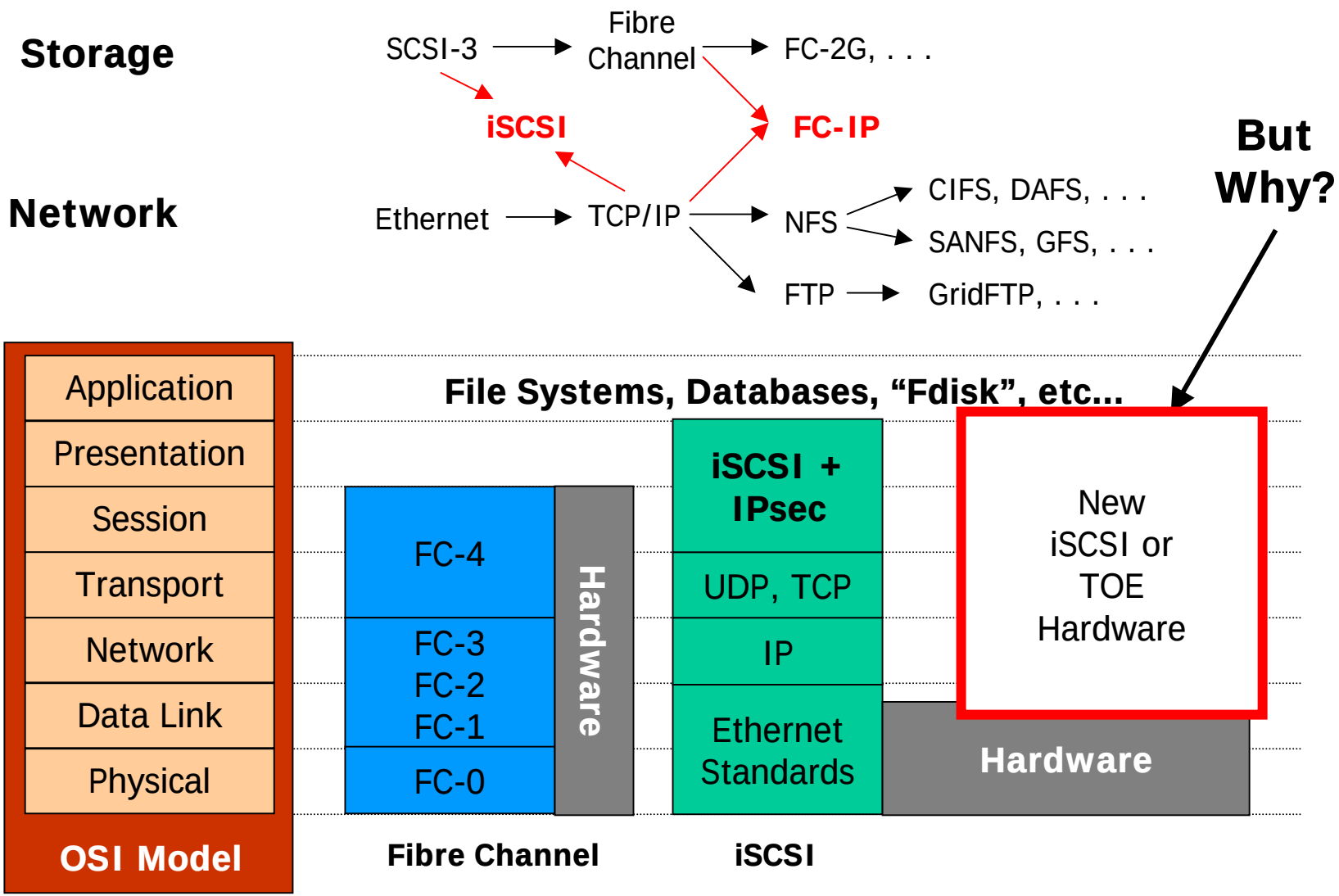
Conclusion:

Storage systems are very different from Network systems

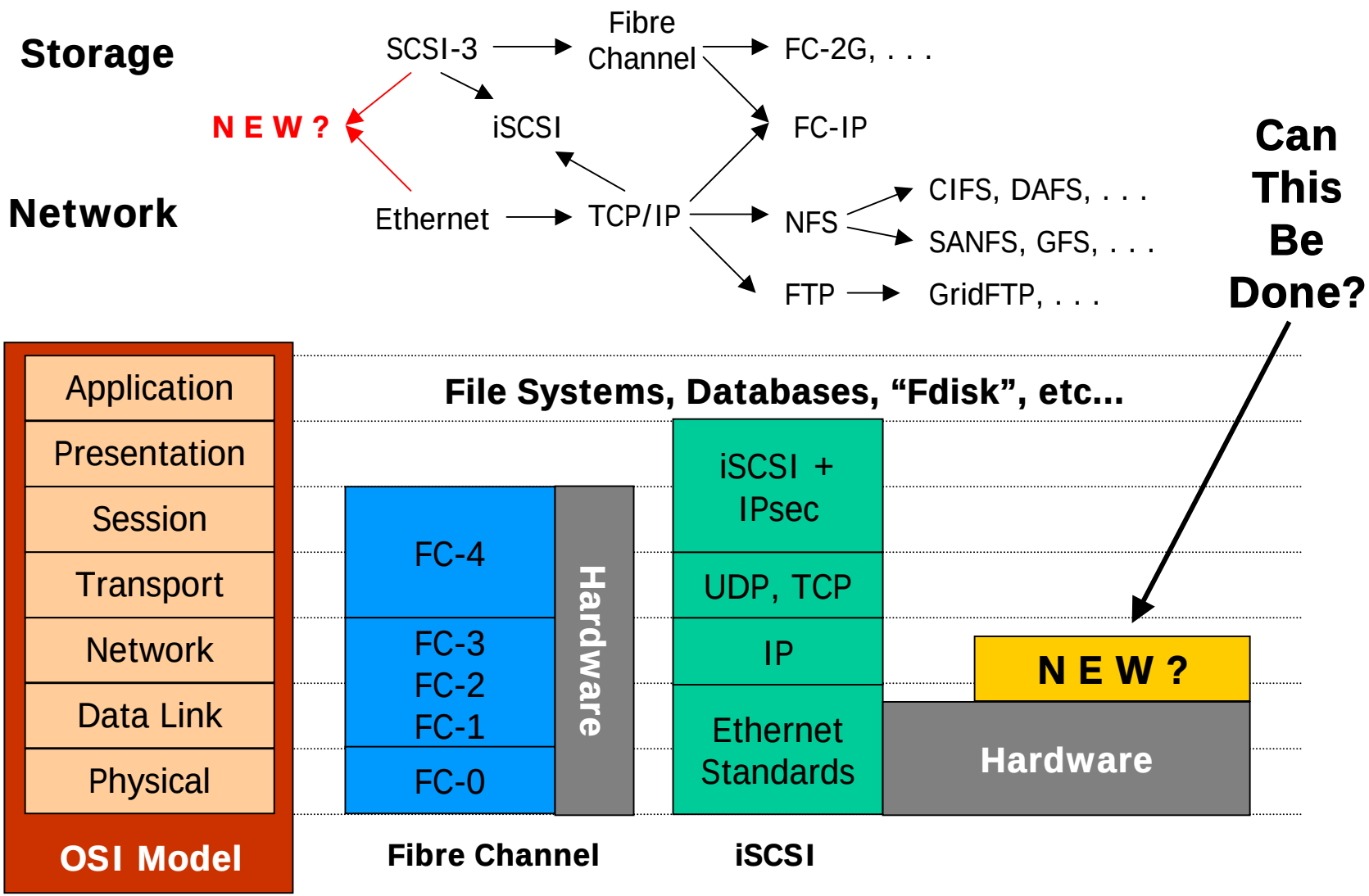
Corollary:

Network systems providing Storage must therefore be designed differently from <u>*normal*</u> Network systems

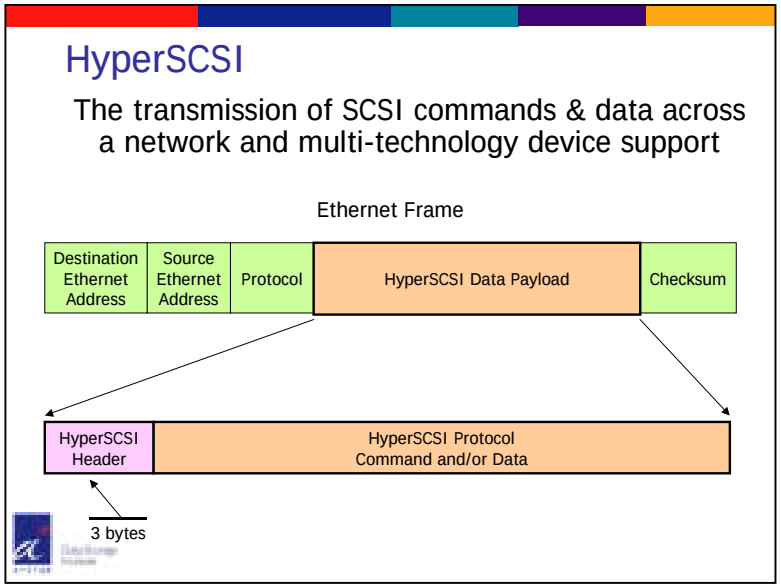
Combine Two Worlds . . .



. . . And Think Out of the Box

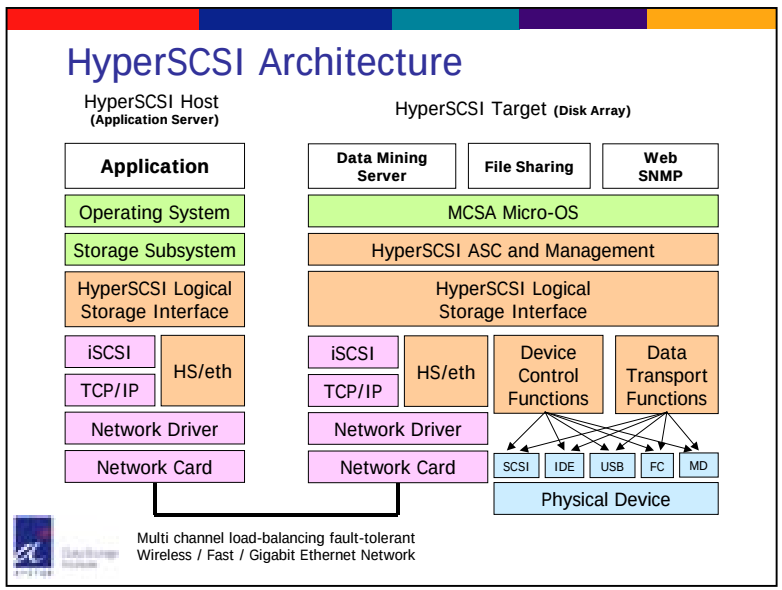


Yes, It Can Be Done!



- **HyperSCSI** is a new open source Network Storage Protocol
- Transmit SCSI commands and data over a network
- High performance, secure, simple, low cost solution
- Runs directly on Ethernet (No TCP/IP!)

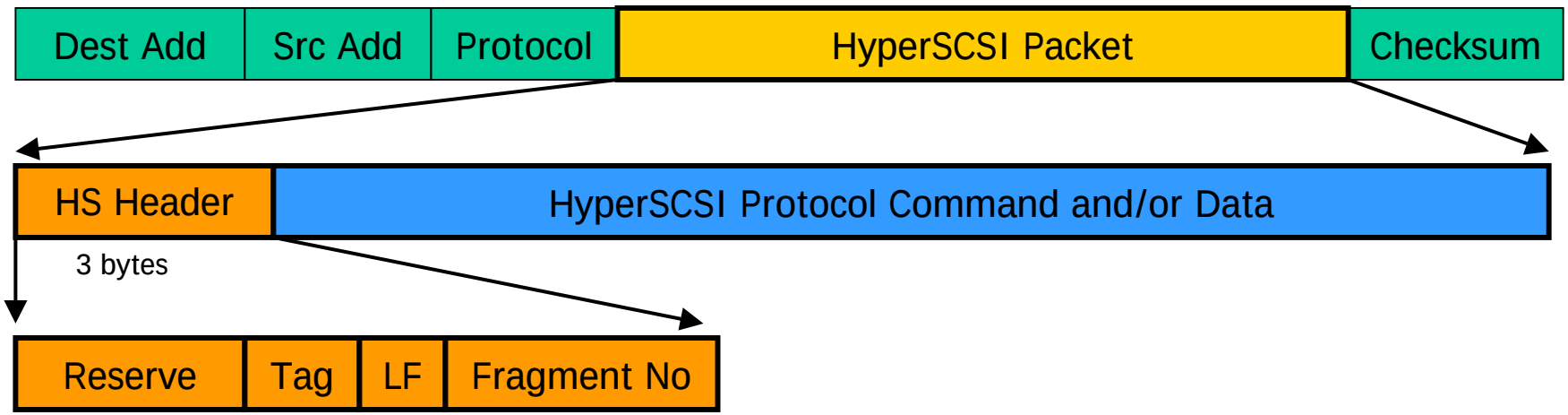
Access storage over a network



Data Storage Institute
A*STAR

The HyperSCSI Protocol

HyperSCSI Packet Framing / Encapsulation on Ethernet



No TCP/IP!

Routeable?
Secure?
Reliable?

HyperSCSI Command and Data Block

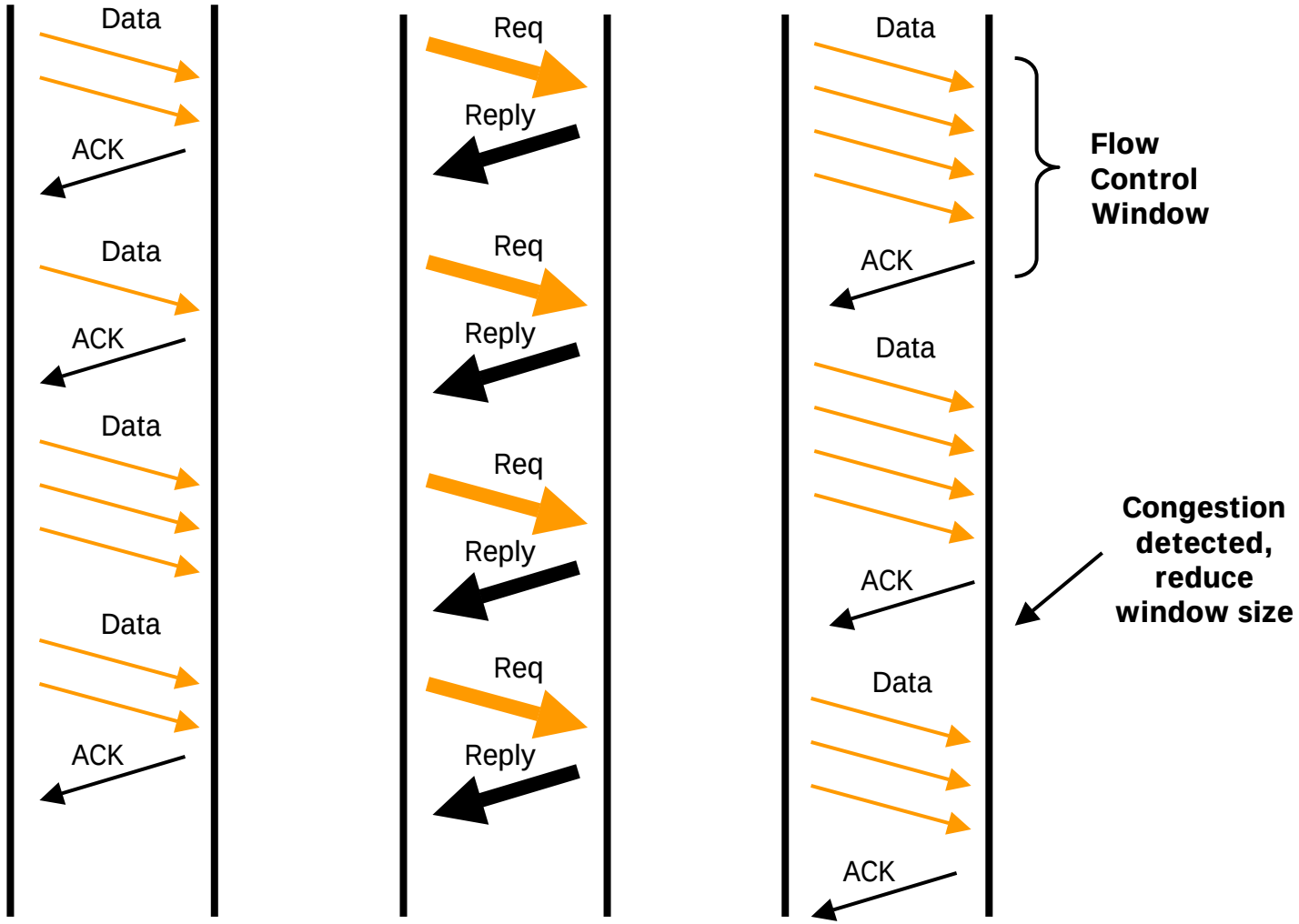
Ver	M	OpCode	HPB Length	P. Param
HS Serial No				
HS Digest				
HS Protocol Parameters (if any)				
Data Block				

“Stealing” Components

	Network	Storage	HyperSCSI
Flow Control	Sliding window	“Buffer Credit”-based *	Dynamically sized fixed window
Transmission	Stream-based	Block-based	Block-based
Data Delivery	Guaranteed	Guaranteed	Guaranteed
Channels	Single-channel (vendor specific trunking)	Parallel transmission	Vendor independent multi-channel
Addressing	“Unlimited”	Limited	Almost “Unlimited” **
Device Discovery	Lookup-based	Broadcast-based (Bus Scan)	Broadcast-based (Local-area)
Authentication	Multi-user challenge & response	Physical security or Zone/LUN Masking	Single-user challenge & response
Tx Security	(Add-on) Encryption	Physical security	(Built-in) Encryption
Scalability	“Unlimited”	Limited	Almost “Unlimited” **
Access	Wide-area	Local-area	Local-area

This is not meant to be a complete or accurate depiction of the components or mappings, but merely as an illustration of the differences between systems and components

Flow Control



IP-based protocols (Network) **SCSI-3 (Storage)** **HyperSCSI**

Easy Management

137.132.29.10 - PuTTY

```
hs-server/etc/hscsi: hs-server status
** Displaying Status . . .
HYPERSCSI server module - 20020725
Status: HyperSCSI server module has been initialized!
HS DEVICE CONFIGURATION SUMMARY:

 CONFIGURATION:
 Name Device-type  Host  Channel ID LUN  Capacity(KB)  Group_Name
 ---- -----------  ----  ------- -- ---  ------------  ----------
  hda    IDE-DISK     0

 CONNECTION:
 Mac-address  Name Device-ty
 -----------  ---- --------

** Displaying Recent Messag
Network device eth0 is goin
The network MTU is 1497
There is no SCSI Host.
hscsi_pkt_window: 5

HyperSCSI Server module ini

hs-server/etc/hscsi:
```

137.132.29.10 - PuTTY

```
hs-client/root: hs-client status
** Displaying Status . . .
HYPERSCSI client module - 20020725
Status: HyperSCSI client module has been initialized!
HyperSCSI DEVICE CONFIGURATION SUMMARY:

 CONNECTION:
 Mac-address  Device_ID  Name  Host  Channel  ID  LUN  Group_Name
 -----------  ---------  ----  ----  -------  --  ---  ----------
0001034476E7          0   sda     1        0   0    0  MCSA-HYPERSCSI

** Displaying Recent Messages . . .
Network device eth0 is going to be used.
hscsi_sg_tablesize: 16
[HS client] fc_window : 5
HyperSCSI Client module initialized!

hs-client/root:
```

Data Storage Institute
A*STAR

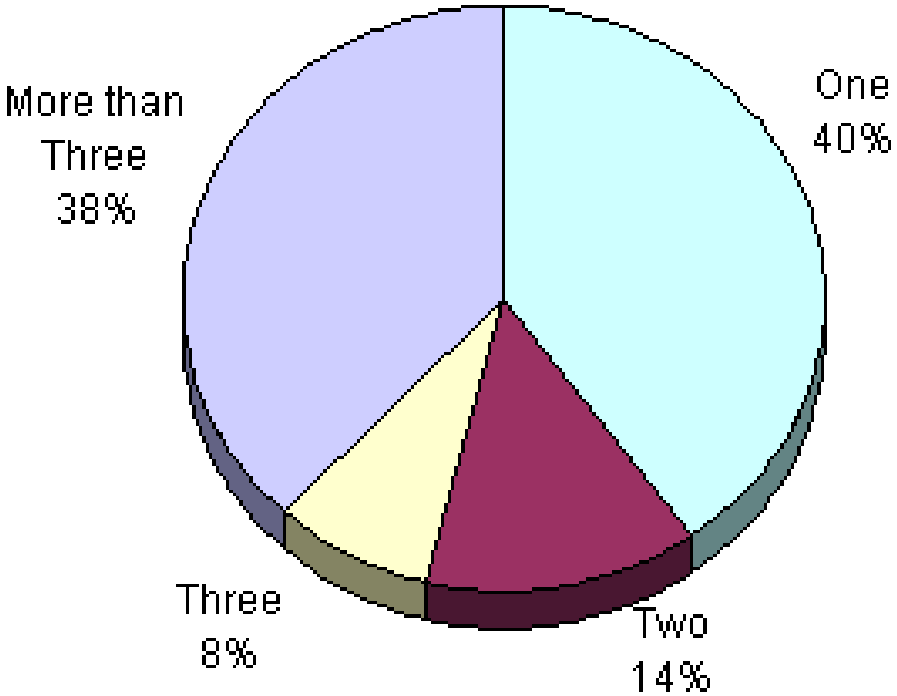
Easy Management

137.132.29.10 - PuTTY

```
hs-server/etc/hscsi: hs-server status
** Displaying Status . . .
HYPERSCSI server module - 20020725
Status: HyperSCSI server module has been initialized!
HS DEVICE CONFIGURATION SUMMARY:

 CONFIGURATION:
 Name Device-type  Host  Channel ID LUN  Capac
 ---- -----------  ----  ------- -- ---  -----
  hda    IDE-DISK     0        0  0   0      1

 CONNECTION:
 Mac-address  Name Device-type  Host  Channel
 -----------  ---- -----------  ----  -------
000102586BF7   hda    IDE-DISK     0        0
SI

** Displaying Recent Messages . . .
Network device eth0 is going to be used.
The network MTU is 1497
There is no SCSI Host.
hscsi_pkt_window: 5

HyperSCSI Server module initialized!

hs-server/etc/hscsi:
```

```
[HYPERSCSI-SERVER-CONFIG-VERSION-20021024]
# Sample Config File for HyperSCSI Server - Modify Before Use!
# Optimised for Fast Ethernet
# Last Updated - 24 July 2002

[ADD]

[MODULE_DEF]
# For GE, try PKT_WINDOW_SIZE 32
# PKT_WINDOW_SIZE:        32
PKT_WINDOW_SIZE:          5
MULTI_RCV_THREAD:         2
MULTI_XMIT_THREAD:        3
REXMIT_COUNT:             2
DIRECT_MC:                0

[VOL_DEF]
VOL_1:                    SDA

[NETWORK_DEF]
LAN_1:                    ETH0

[GROUP_DEF]
GROUP_NAME:               MCSA-HYPERSCSI
PASSWORD:                 0123456789
NET:                      LAN_1
IP_ON:                    0
VOL_NAME:                 VOL_1
VOL_OPT:                  0:0

[END]
```

SAN Islands are a Reality



Number of SANs Deployed in Respondents' Organizations

- Most organisations preparing or considering implementing SANs, should take into account adding new "independent" SANs in the future
- Most SANs are "Local" in nature
- Why do you need IP to go long distance? (for most people anyway)

Source: Aberdeen Group, January 2002

Secure Storage

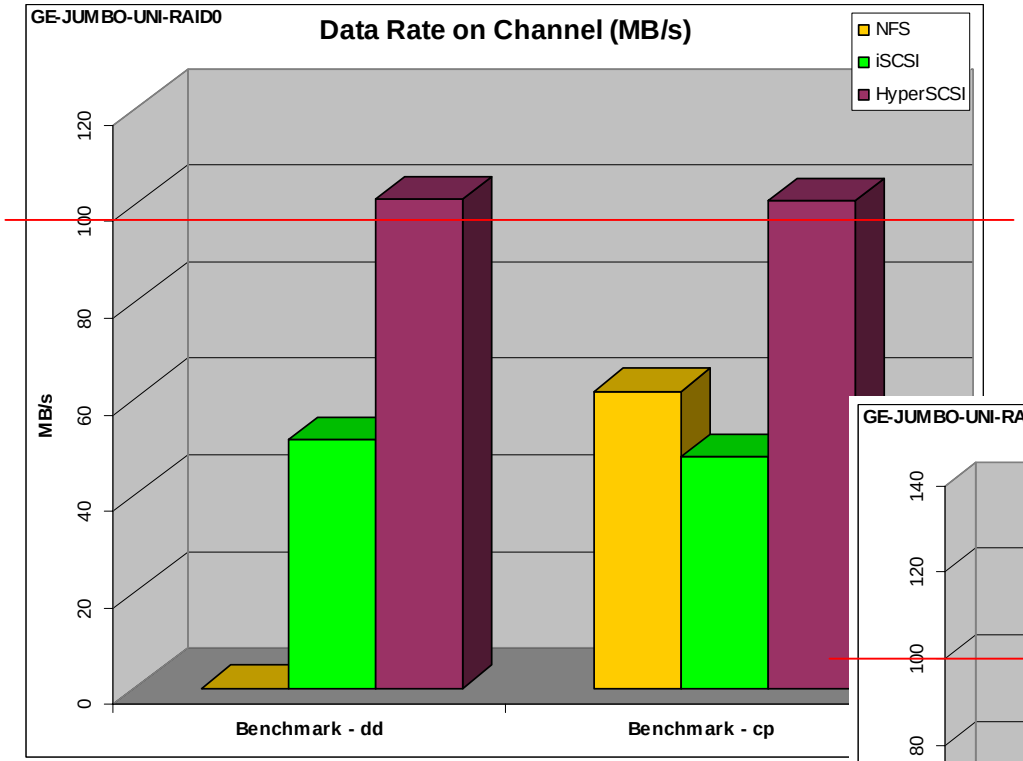
C:\My Documents\hacking-log.txt - Viewer

Files Edit Search View Options Coding Help

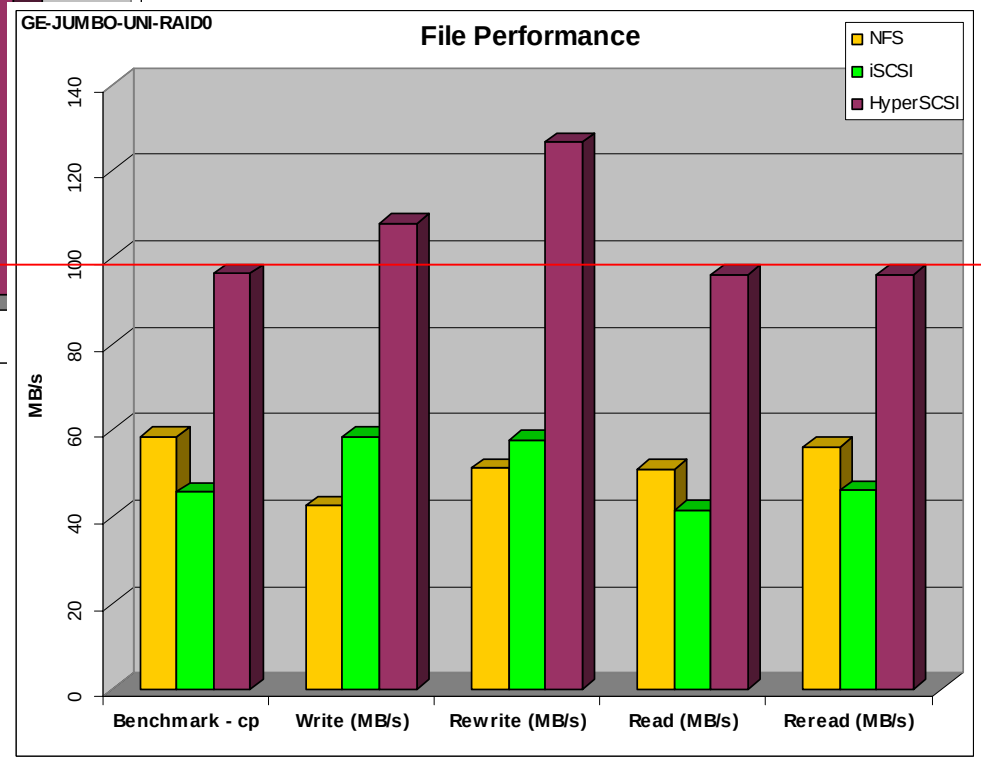
```
==============================================================
/var/log/messages
==============================================================
Nov 18 04:02:01 nst syslogd 1.4-0: restart.

----> Hacker Start
Nov 19 04:18:59 nst ftp(pam_unix)[20988]: check pass; user unknown
Nov 19 04:18:59 nst ftp(pam_unix)[20988]: authentication failure; logname= uid=0
euid=0 tty= ruser= rhost=AMarseille-101-1-1-243.abo.wanadoo.fr
Nov 19 04:19:01 nst ftpd: AMarseille-101-1-1-243.abo.wanadoo.fr: connected: IDLE
[20988]: failed login from AMarseille-101-1-1-243.abo.wanadoo.fr [193.252.177.243
]
Nov 19 04:19:12 nst ftpd: AMarseille-101-1-1-243.abo.wanadoo.fr: connected: IDLE
[20988]: lost connection to AMarseille-101-1-1-243.abo.wanadoo.fr [193.252.177.24
3]
Nov 19 04:19:12 nst ftpd: AMarseille-101-1-1-243.abo.wanadoo.fr: connected: IDLE
[20988]: FTP session closed
----> Hacker Stop

----> Hacker Start
Nov 20 03:40:01 nst ftpd[23171]: ACCESS DENIED (not in any class) TO AMarseille-1
01-1-1-243.abo.wanadoo.fr [193.252.177.243]
Nov 20 03:40:01 nst ftpd[23171]: FTP LOGIN REFUSED (access denied) FROM AMarseill
e-101-1-1-243.abo.wanadoo.fr [193.252.177.243], anonymous
Nov 20 03:40:02 nst ftpd[23171]: FTP session closed
----> Hacker Stop
```

This can't happen to your storage, **IF** your storage doesn't have TCP/IP

High Performance



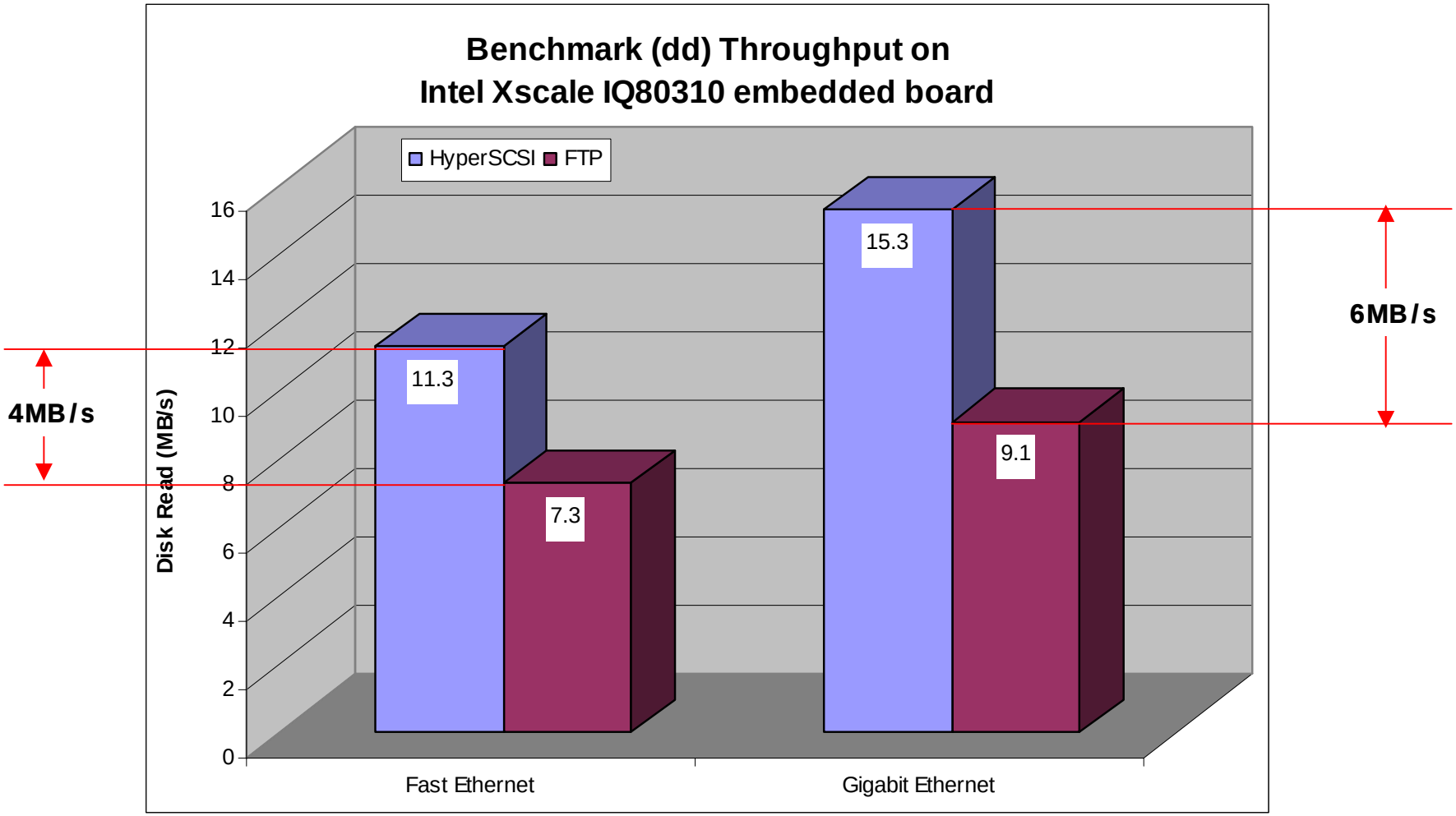
100MB/s sustained data transfer speeds for both block and file, more than 60% faster than NFS and iSCSI protocols in comparative benchmark tests



Tests conducted on a clean Gigabit Ethernet network with Jumbo frames, single initiator and target, 8 hard disks configured in RAID0 and using only common off-the-shelf hardware and software without special tweaks or optimisations.

Data Storage Institute
A*STAR

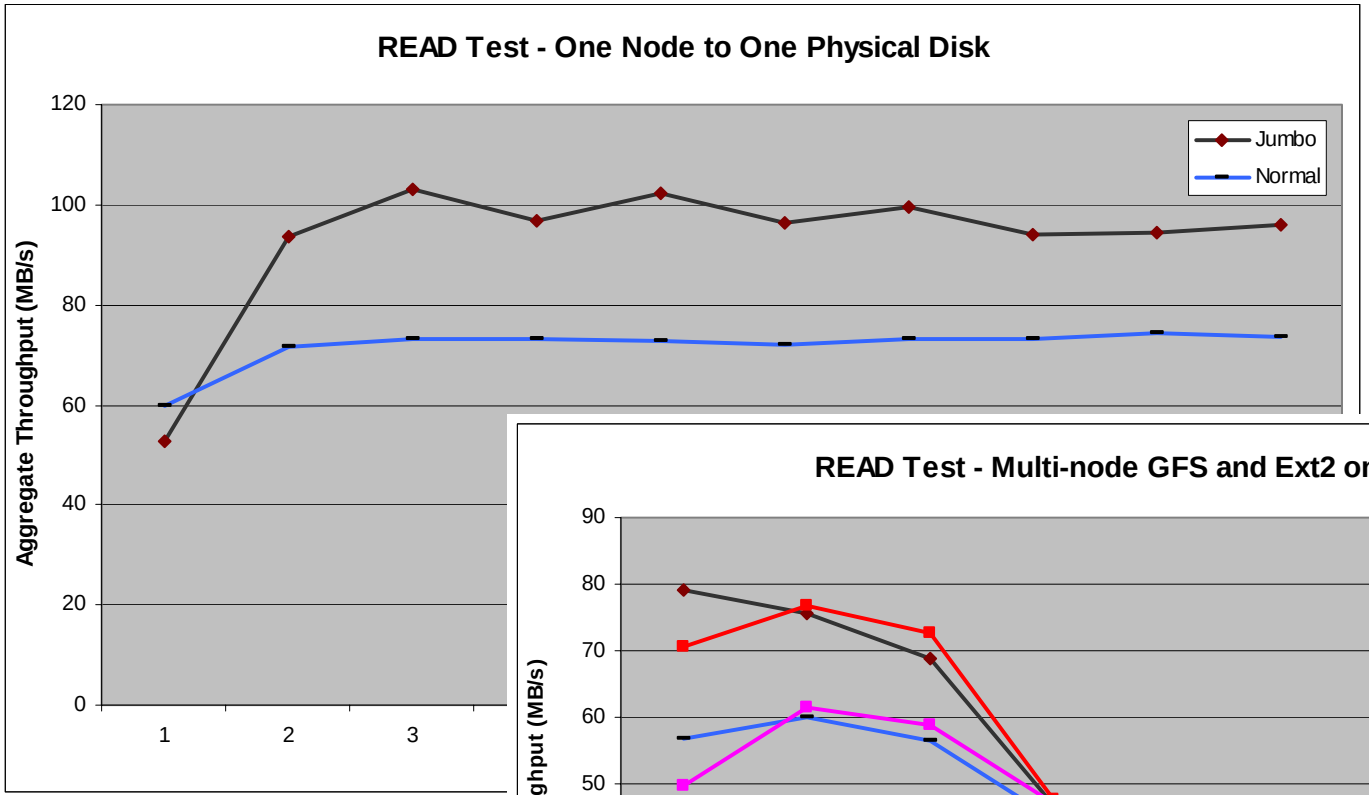
Lower Embedded Overheads



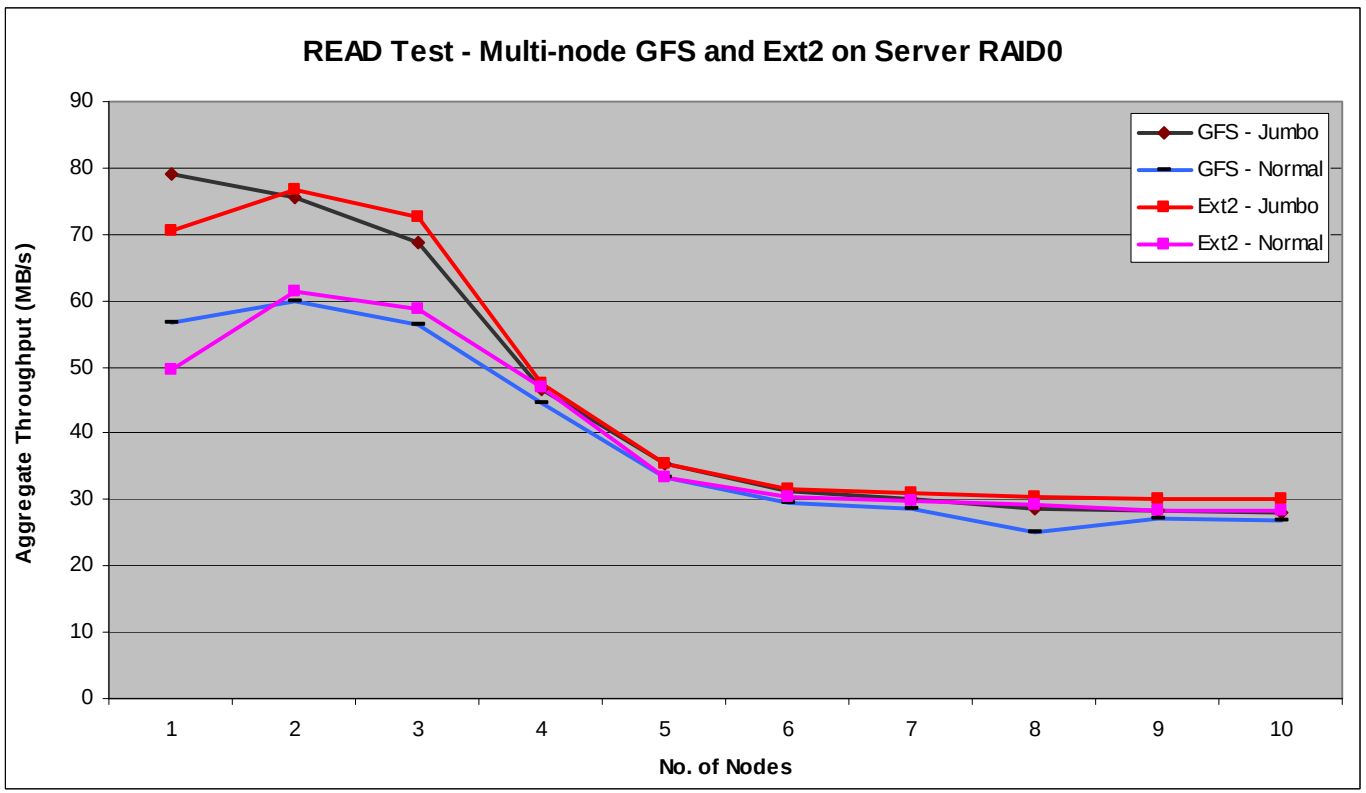
Single drive on Embedded Board running Embedded Linux
Slow GE Performance is noted due to poor memory speed on IQ80310 EVB (well documented HW bug)

Data Storage Institute
A*STAR

File Systems and Multiple Clients



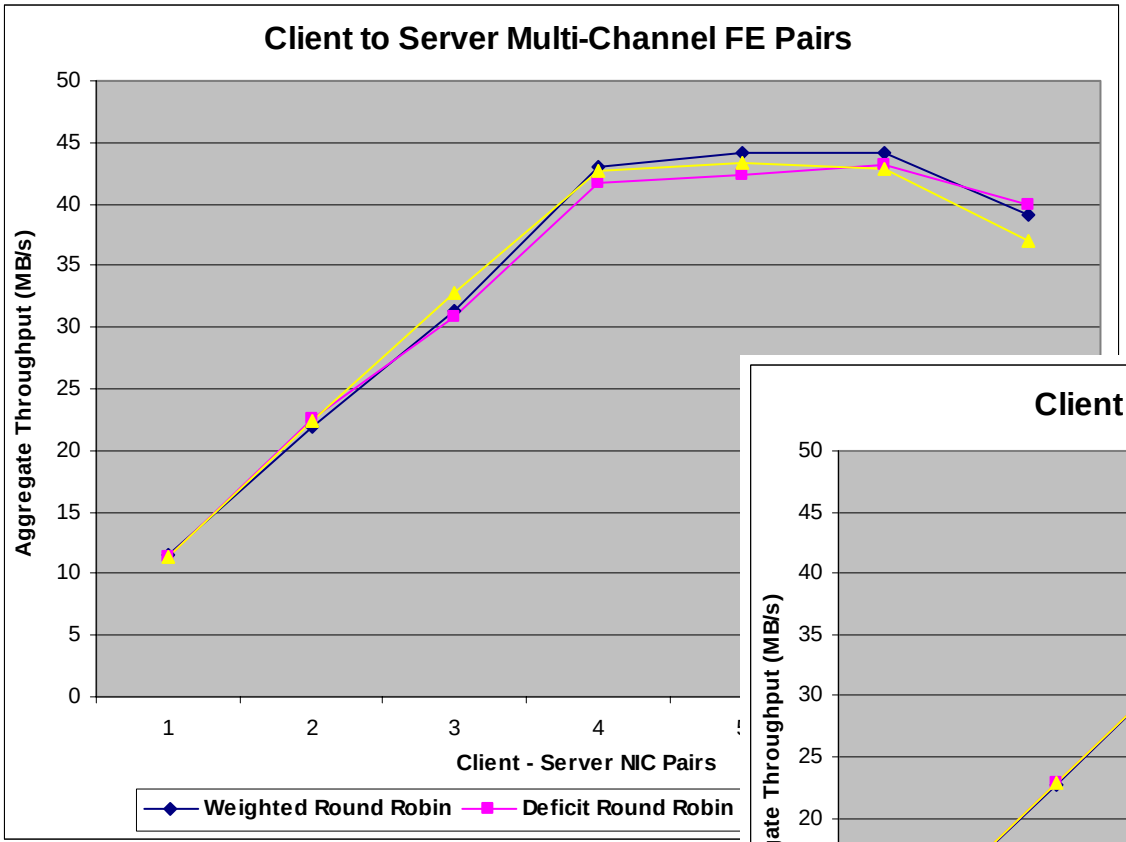
Each node read different file on same RAID0 volume on one server, Sistina GFS and Ext2 File Systems used



Each node read different file on different physical disk with Ext2 simultaneously, one node to one physical disk on one server

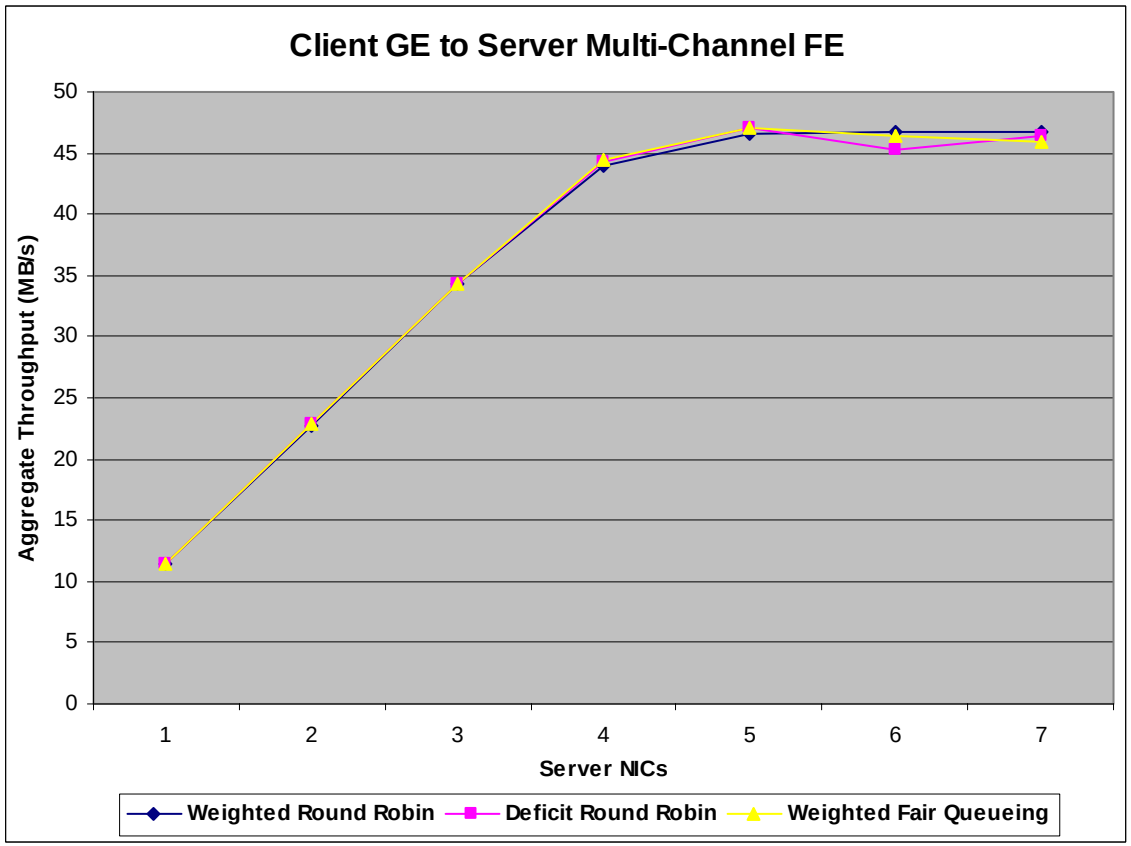
Data Storage Institute
A*STAR

Multi-Channel Technology



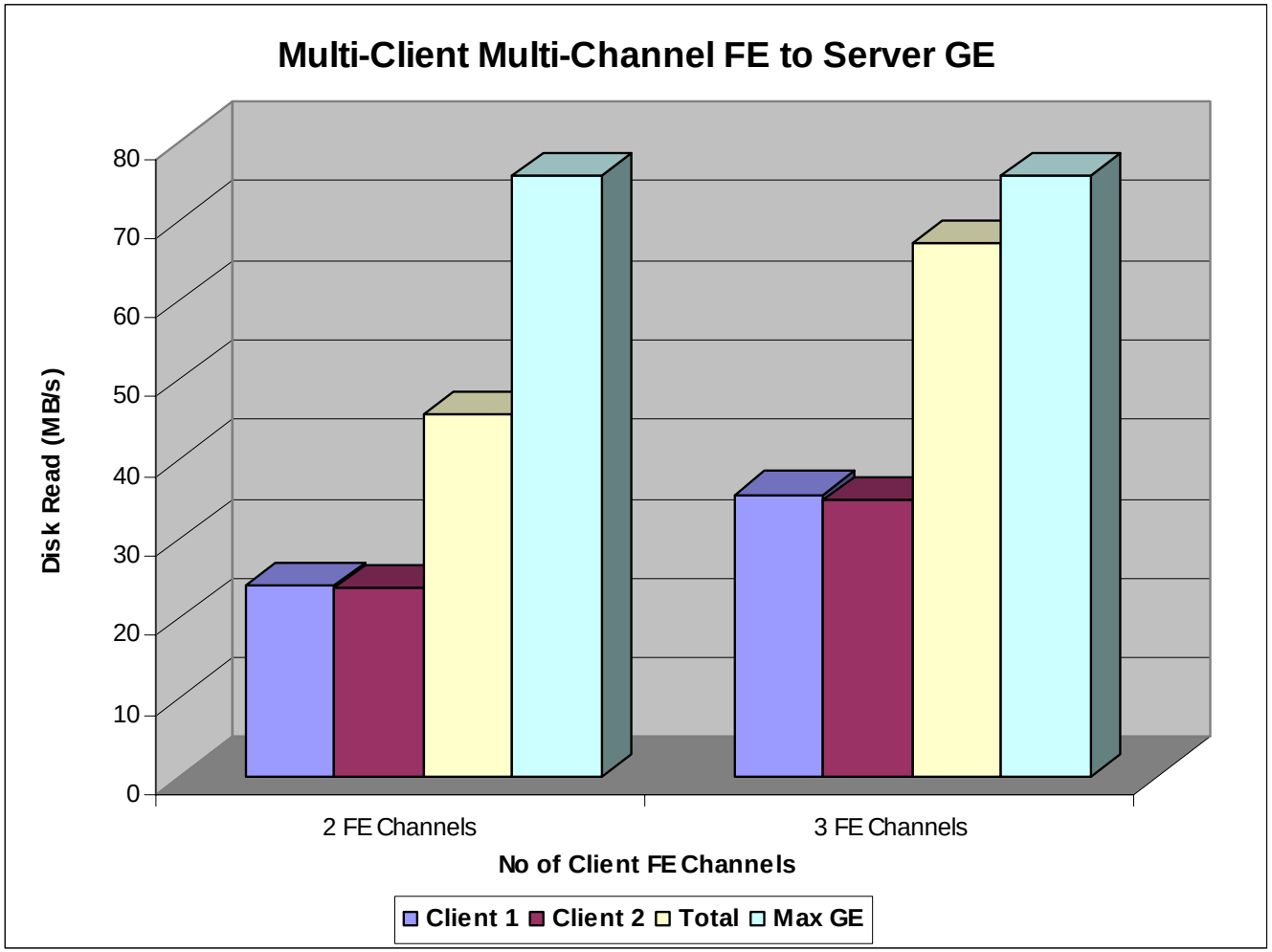
RAID0 on Client, Two HDDs on Server, One FE channel added on both client and server in pairs

RAID0 on Client, Two HDDs on Server, One GE channel on client, add one FE channel on server in succession



Data Storage Institute
A*STAR

Multi-Channel Multi-Client



RAID0 on Client, Two HDDs on Server, One GE channel on server, add one FE channel on each client in pairs
* Weighted Fair Queuing algorithm used
* Max GE Performance is measured from one GE channel between client and server

Data Storage Institute
A*STAR

Key Features

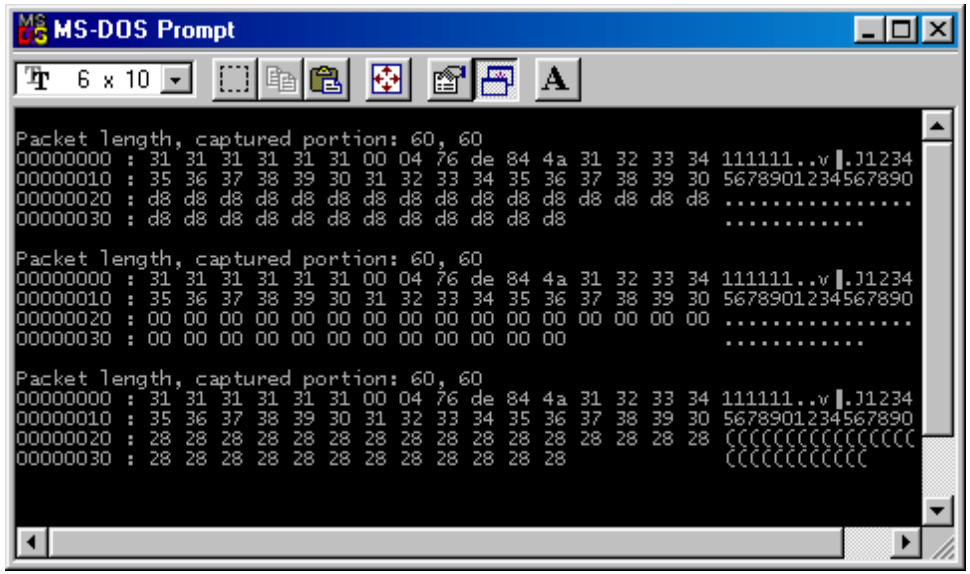
- Runs on raw Ethernet – does not use TCP/IP
- Supports various storage devices (including disk, optical, removable media and tape) and interface technologies (including SCSI, Fibre Channel, IDE and USB)
- Supports various network technologies (including Fast Ethernet, Gigabit Ethernet, GE + Jumbo Frames and Wireless LAN 802.11b)
- HyperSCSI on Ethernet (HS/eth) can be deployed on a multi-protocol network environment safely
- Includes built-in 128-bit Encryption
- Supports active device discovery for plug and play ad-hoc network storage
- Independent of hardware or vendor products - runs with wide variety of disk/tape/optical devices and arrays, SCSI/FC HBAs, NICs, and switches
- Designed for easy deployment, and above all, to provide users with the freedom of choice, to implement network storage the way they want to, with the options they need

What Others Have to Say

"That is a very good news that you managed to get HyperSCSI running over Wireless LAN. Sustaining ~100MB/s throughput is very impressive. **I am very impressed by your work.** We are rebuilding our cluster nodes to use 2.4.16 kernel and will install the HyperSCSI software soon. I cannot wait to see HyperSCSI working in our test bed."

"I had been researching a solution for about a month before I stumbled across HyperSCSI. I started out looking at FiberChannel but the protocol is cryptic, the Linux support is poor, and I object to the cost being 6-20x that of Gig-Ethernet when its basically the same technology. I looked at iSCSI but again the protocols are over engineered, the Linux support is poor, the cost of equipment is robbery, and there are a lot of research papers that suggest block protocols over TCP/IP are a poor choice. The Linux network block driver is a hack and not a very good one. I looked into doing sharing with a SCSI bus, but the lack of target mode support in Linux killed that idea quickly. So I started researching how to do raw Ethernet access in Linux with the intent of writing a Ethernet based protocol to allow machines concurrent access to raw disk. **And then I found HyperSCSI. Its simple. Its elegant. You can implement it using commodity hardware. It works today. End of search.**"

Porting to Windows 2000 / XP



- Virtual SCSI Card
- HyperSCSI client on Windows

Is This for Real? **Yes** for Consumers!

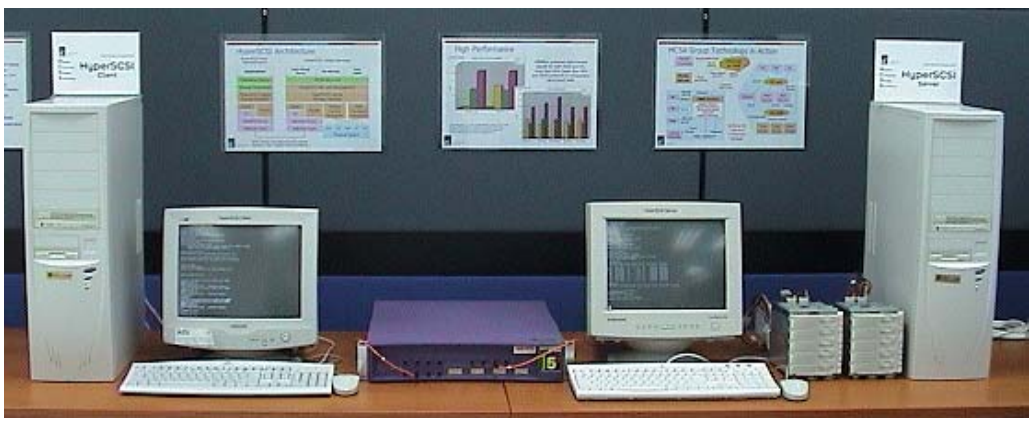
- Wireless Network Storage
- Personal and Home Networks
- Consumer Electronics
- Entertainment and Content Distribution

Is This for Real? **Yes** for Corporates!

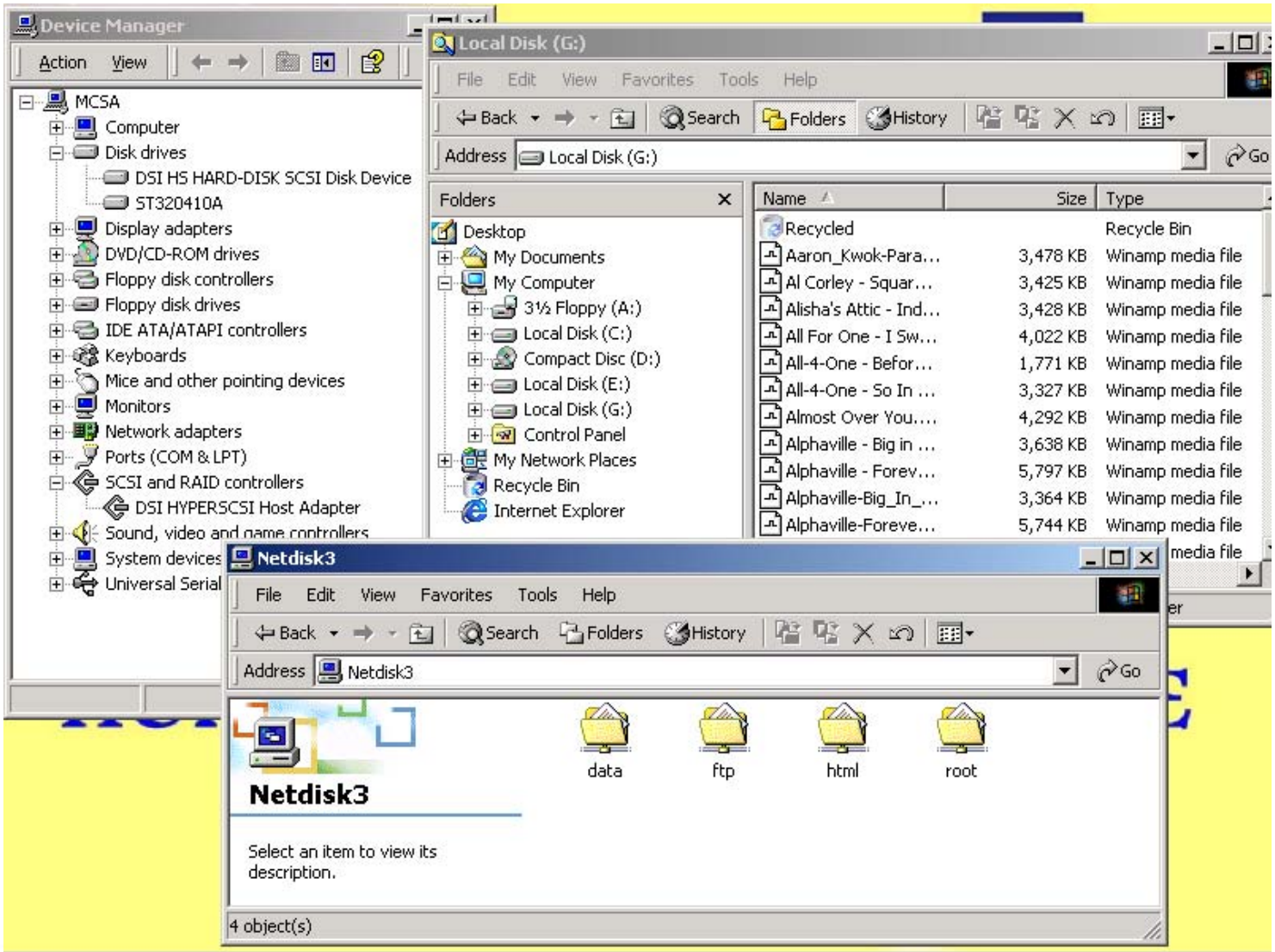
- Storage Area Networks (SAN) and Network Attached Storage (NAS)
- Information Continuance, Performance, Security and Reliability

Is This for Real? **Yes** for Clusters!

- Storage for HPC / Grid
- Remote Boot, Concurrent Access, Cluster File Systems

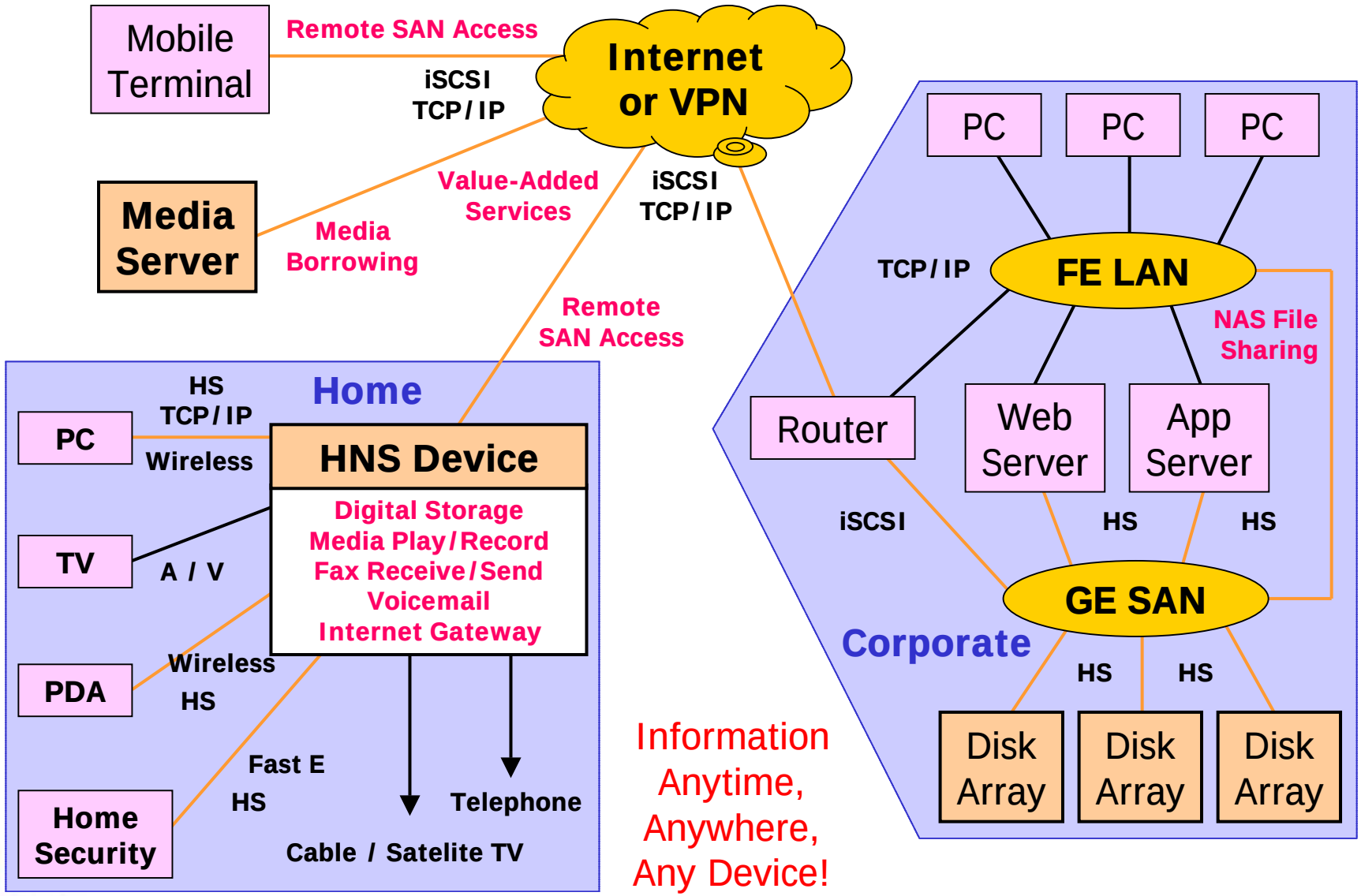


Is This for Real? **Yes** for Convergence!



Data Storage Institute
A*STAR

The Network Storage Vision



Information Anytime, Anywhere, Any Device!

Data Storage Institute
A*STAR

Conclusion

- Some key points:
  - New Technologies like HyperSCSI, iSCSI, TOE, iSCSI HBA, blah, blah, blah
  - New applications for Network Storage like Wireless LAN, Home Networks, Personal Networks
  - Lower costs, ease of use, network storage for everyone!

Ethernet Storage is **NOT** coming,

It's already here!

And yes, you **WILL** get more for less

Thank You

http://nst.dsi.a-star.edu.sg/mcsa/