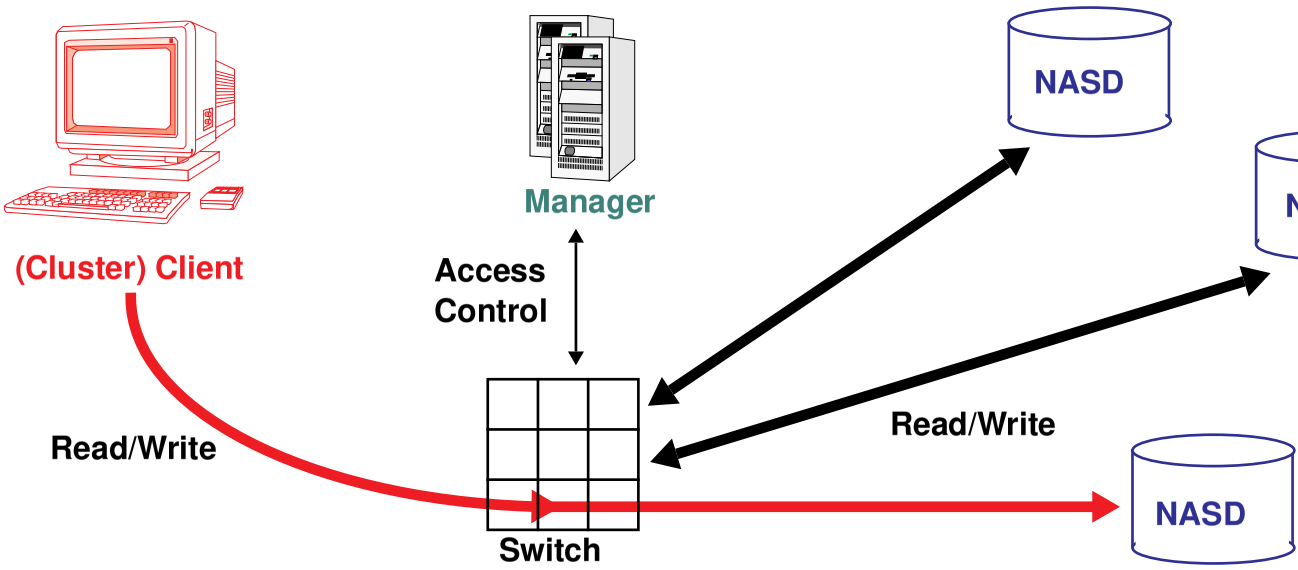
Network-Attached Storage Networking

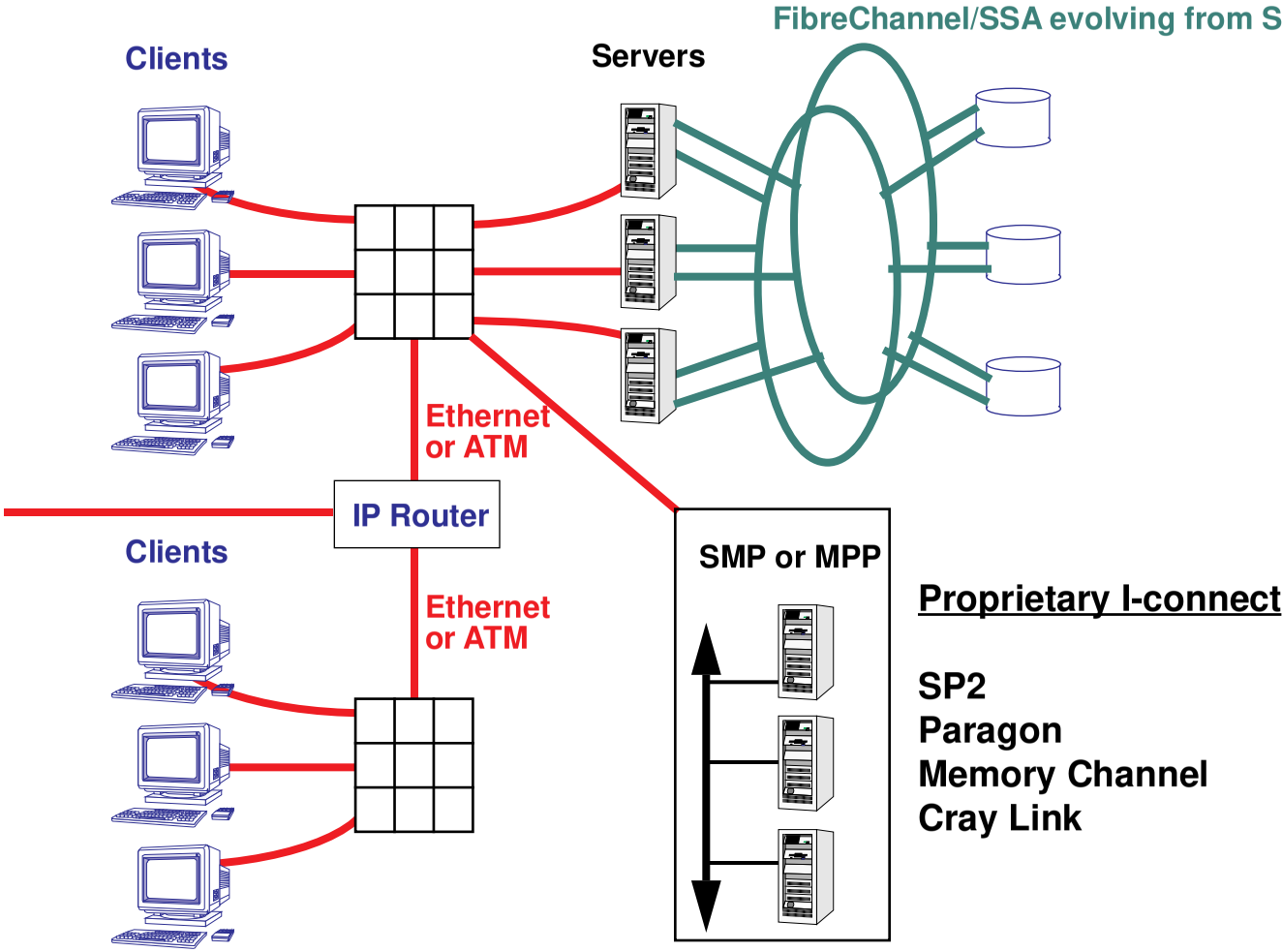
Garth A. Gibson, CMU, http://www.pdl.cs.cmu.edu/NAS

Meet scaling compute needs with storage strip over scalable client network



The World A Few Years Ago

Switched LANs promise scalable client/server



Resulting in CMU's NASD Perspective

Addressability and direct transfer essential

- all clients and drives share same network
- client LAN techn suite? or peripheral I-connect s

CMU starts with LAN for wide addressability

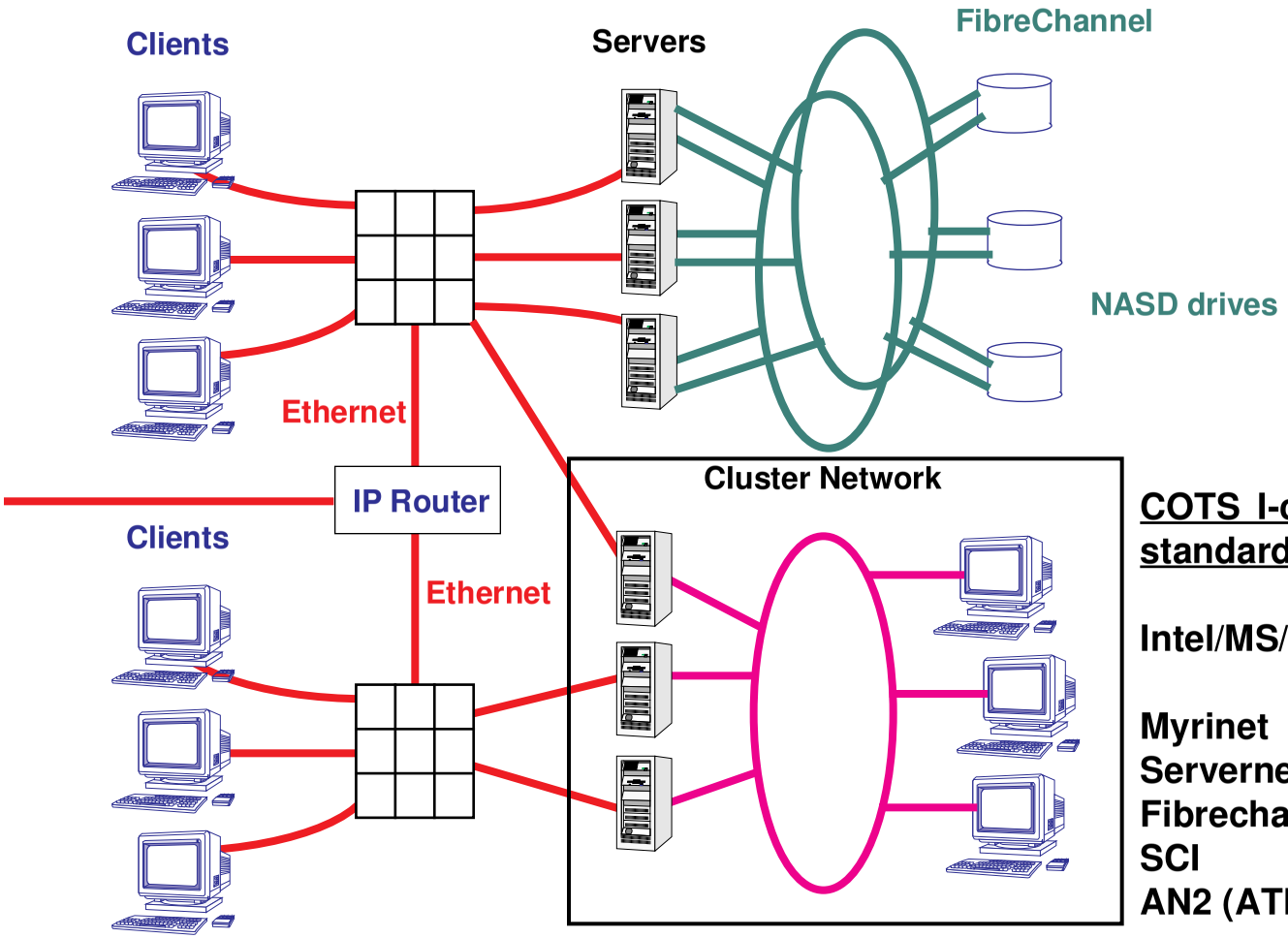
- first CMU prototype Alpha workstation, switche
- second CMU prototype Intel PC board, switched
- RPC/UDP/IP protocol suite; DCE so far

Implications of RPC/UDP/IP

- most impls not well optimized for local transfers
- clients burns a lot of cycles protocol processing
- hardware support for RPC/UDP/IP on-drive

Meanwhile the World Changes

COTS cluster networks: cost-effective scalable



COTS I-conn
standard API

Intel/MS/Com

Myrinet
Servernet
Fibrechannel
SCI
AN2 (ATM w/

COTS Clusters == System Area Network

Cost-effective parallel workstation-cluster

- **interconnect must be commodity as well as nodes**

Cluster network is

- **not yet well-defined, but some definitions include**
- **low latency for small messages**

  asynch reception, put/get memory model,
  applic access to NIC; ie. not sockets/TCP/IP
- **high bandwidth for large transfers**

  zero copy, scatter/gather, pre-posted receive buffe
  per transfer interrupt enable
- **need more efficient sockets interface**
- **need protocol impl to exploit network characteris**

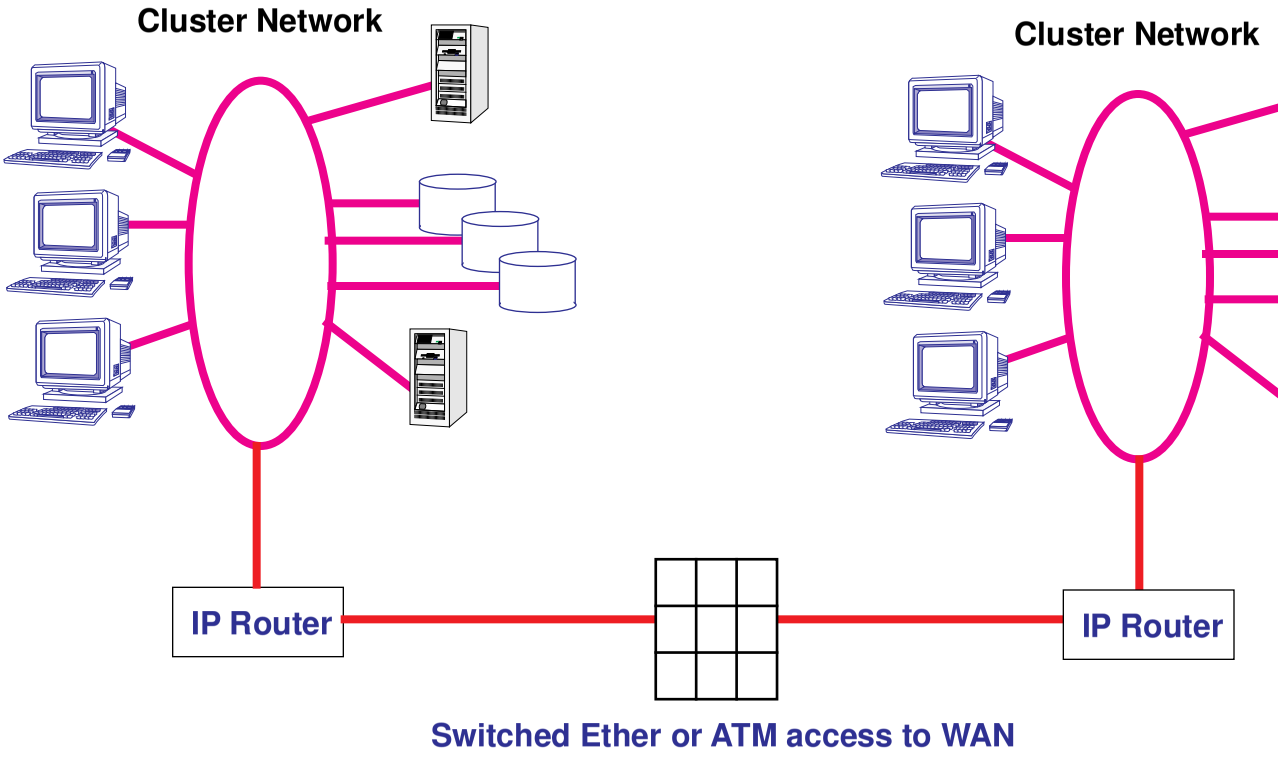
Cluster network volume drives down adapter

- **Can we wire only once?**

A Hopeful Vision of Networking

Cluster network is LAN & peripheral interco

WAN protocols not used for intra-LAN traffic



NASD in the Cluster Network

IF clients get high bandwidth with little CPU c

Storage will be required to match this standar

Critical component is "improved sockets inter

- **Virtual Interface Arch: Intel, MS, Compaq, Tand**
- **queues of pre-posted message buffers: new socke**
- **virtual control&status registers: applic access to**
- **what reliability? security? WAN interoperation?**

Question: does NASD need low latency?

Storage access time is long, so latency is secon

- false: most file accesses involve small chunks of n
- low latency -> thin codepath -> cycles used by ap

NFS Operation	Count in top 2% by work (thousd)	SAD		NASD	
		Cycles (billions)	%of SAD	Cycles (billions)	%of SAD
Attr Read	792.7	26.4	11.8%	0.0	0.0%
Attr Write	10.0	0.6	0.3%	0.6	0.3%
Block Read	803.2	70.4	31.6%	0.0	0.0%
Block Write	228.4	43.2	19.4%	0.0	0.0%
Dir Read	1577.2	79.1	35.5%	0.0	0.0%
Dir RW	28.7	2.3	1.0%	2.3	1.0%
Delete Write	7.0	0.9	0.4%	0.9	0.4%
Open	95.2	0.0	0.0%	12.2	5.5%
Total	**3542.4**	**223.1**	**100.0%**	**16.1**	**7.2%**

Question: what does NASD need from cluster n

Nearly no packet loss - retransmit not "norma

Low bit error rate

- possibly net HW NACK on bit error

No packet loss at cluster switch

- switch congestion resolved by network, not appli
- INCAST: striped file access causes deterministic
- prefer link-by-link credit-based flow control

Pre-posting of receive buffers

- endpoint congestion causes received packets to b
- resolved by pre-transfer application-level buffer
- end-to-end reliability based on exception applica
- much easier if NACK comes back in nearly all d

Question: do applics need to know network topo

Destinations across ANY IP routers different

- intra-cluster transfers don't use WAN suite; don'

Clients outside cluster need access to cluster

- storage access protocols can be encapsulated in W

Application should not know location of IP r

- protocol selection done UNDER "improved socke
- e.g. Argonne NEXUS project lists possible protoc application-level "address" metadata
- protocol selection at message, not connection, gra

Unresolved: how many protocols does storage s

- cluster network specific and WAN encapsulated (
- alternative is indirect access for WAN encapsulat ie., file manager is application-level address for W
- de-encapsulation is network protocol; can router

Question: which cluster net?

Can NASD delay binding?

- application depends on “improved sockets interfa

Drive hardware depends on cluster net details

- vertical integration of protocols for buffer manag
- protocols making this harder have big impact on

Drive industry prefers FC

- how well does FC perform as MPP interconnect?
- how would systems vendors be convinced?
- how could systems vendors influence drive proto

Conclusion

Cluster networks because WANs inefficient lo

Opportunity for one (not 3) leaf networks

New application interface means higher stand

NASD does need low latency for metadata ops

Cluster net should do link-level flow control

Improved interface should defer protocol bind

Is FC a cluster network for all apps?