NASD: Network-Attached Secure Disks

Garth Gibson

garth.gibson@cmu.edu

also David Nagle, Khalil Amiri,Jeff Butler, Howard Gobioff, Charles Hardin,
Nat Lanza, Erik Riedel, David Rochberg, Chris Sabol, Marc Unangst, Jim Zelenka

a DARPA funded project of the Parallel Data Lab, 1995-1999

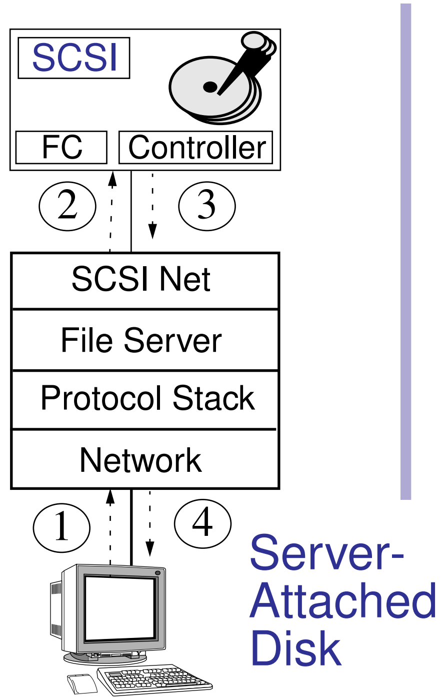
with support from: Seagate, HP, IBM, Intel, Quantum,
Compaq, Infineon, Hitachi, Clariion, LSI Logic,
3Com, MTI, ProCom, EMC, Novell, Storage Tek

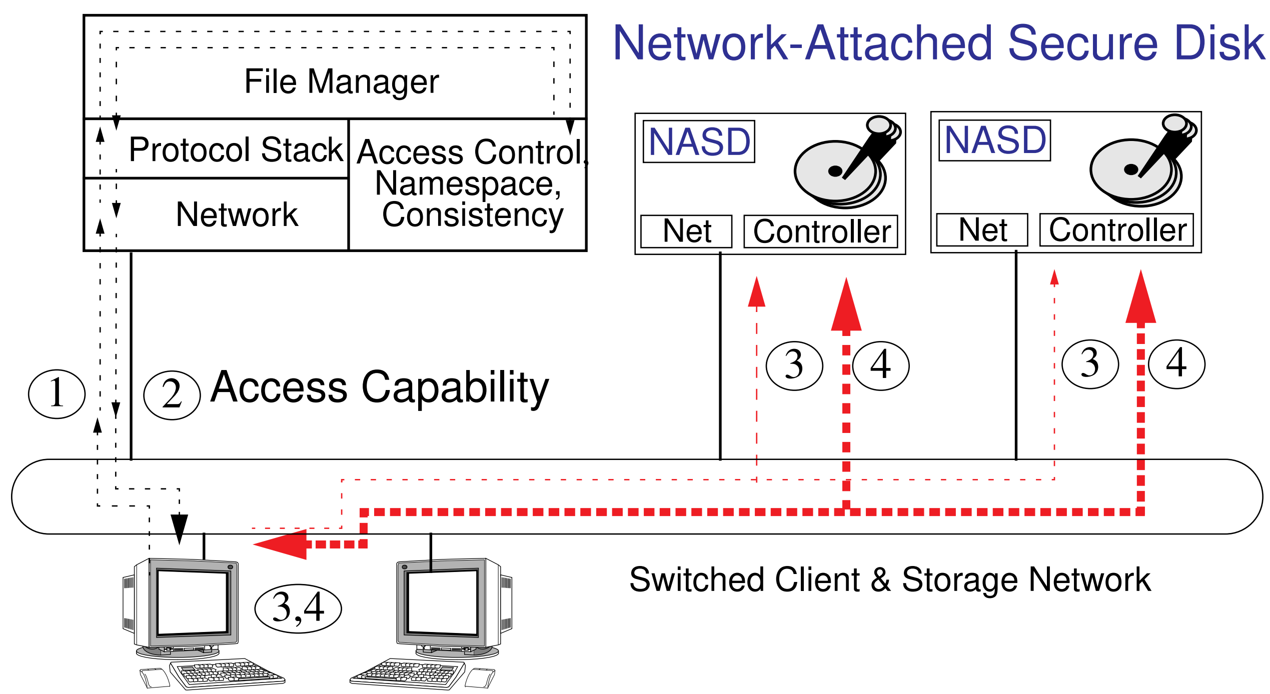
NASD Evolves Storage Architecture

Store & forward Server-Attached Storage is SAD

CMU NASD research issues:

- interface definition, prototyping, porting file systems, affordable security, aggregation and sharing, industry participation





Adapting Filesystems to NASD Storage

Primitives become drive responsibility

- data transfer and data layout
- attribute-based quality of storage specialization

Policy remains manager responsibility

- namespace navigation
- access control policy
- client cache management
- multi-access atomicity

CMU97-118 - interface definition
Sigmetrics97 - scaling mgmt
ASPLOS98 - scaling bandwidth
ExtremeLinux99 - code release
ISHPC99 - affordable security
ICDCS00 - locking & RAID

Scaling Manageability

Offloading work from the management to the workers

- drives perform layout, data transfer, command processing

- offload over 90% of manager load; managers support 10X scale
- concurrent device-to-device work: backup, XOR, migrate, restripe
- extensive, transparent, long-term self-monitoring

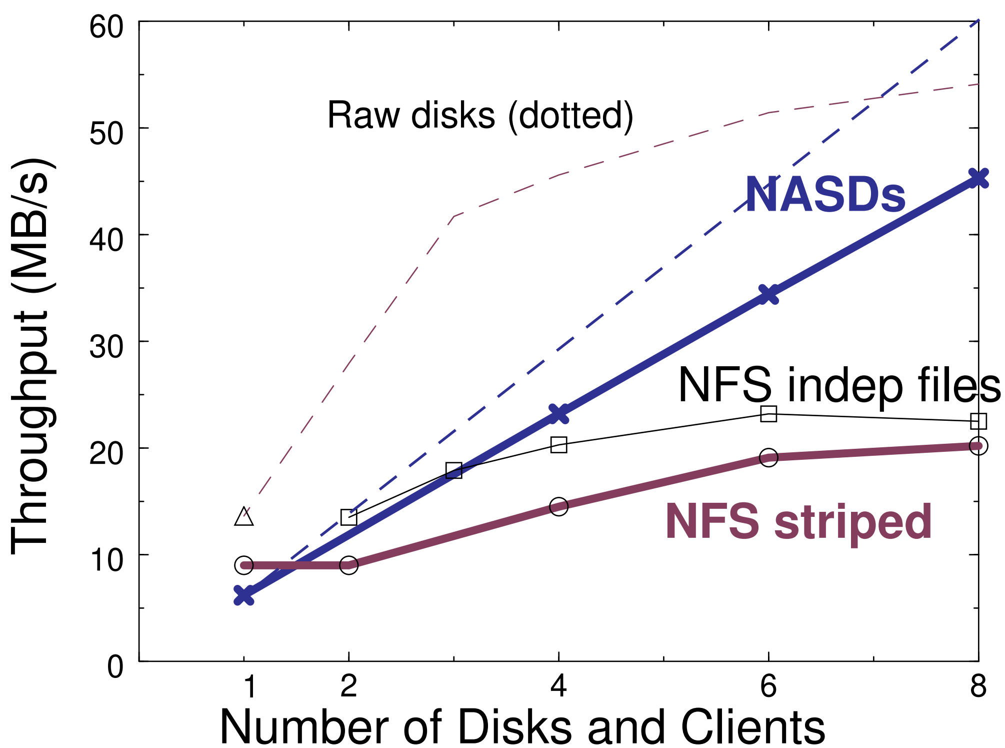
NFS Operation	Count in top 2% by work (thousd)	SAD		NetSCSI		NASD	
		Cycles (billions)	%of SAD	Cycles (billions)	%of SAD	Cycles (billions)	%of SAD
Attr Read	792.7	26.4	11.8%	26.4	11.8%	0.0	**0.0%**
Attr Write	10.0	0.6	0.3%	0.6	0.3%	0.6	0.3%
Block Read	803.2	70.4	31.6%	26.8	12.0%	0.0	**0.0%**
Block Write	228.4	43.2	19.4%	7.6	3.4%	0.0	0.0%
Dir Read	1577.2	79.1	35.5%	79.1	35.5%	0.0	**0.0%**
Dir RW	28.7	2.3	1.0%	2.3	1.0%	2.3	1.0%
Delete Write	7.0	0.9	0.4%	0.9	0.4%	0.9	0.4%
Open	95.2	0.0	0.0%	0.0	0.0%	12.2	**5.5%**
Total	**3542.4**	**223.1**	**100.0%**	**143.9**	**64.5%**	**16.1**	**7.2%**

- Berkeley NFS traces [Dahlin94] (230 clients, 6.6M reqs)

Scaling Bandwidth

NASD PFS aggregates deliver raw disks' bandwidth

- Parallel association rule discovery experiment on 300 MB of sales records
- NASD-based middleware fetches 4 x 512KB blocks in parallel
- NFS server delivers 20% raw disk BW (60% net BW) @ 8 pairs



- **133Mhz NASDs 6 MB/s drive's max**
- **233Mhz clients**
- **MPI + SIO LLAPI**
- **switched OC3 ATM**
- **500 Mhz NFS server 14 MB/s drive's max dual OC3 links**

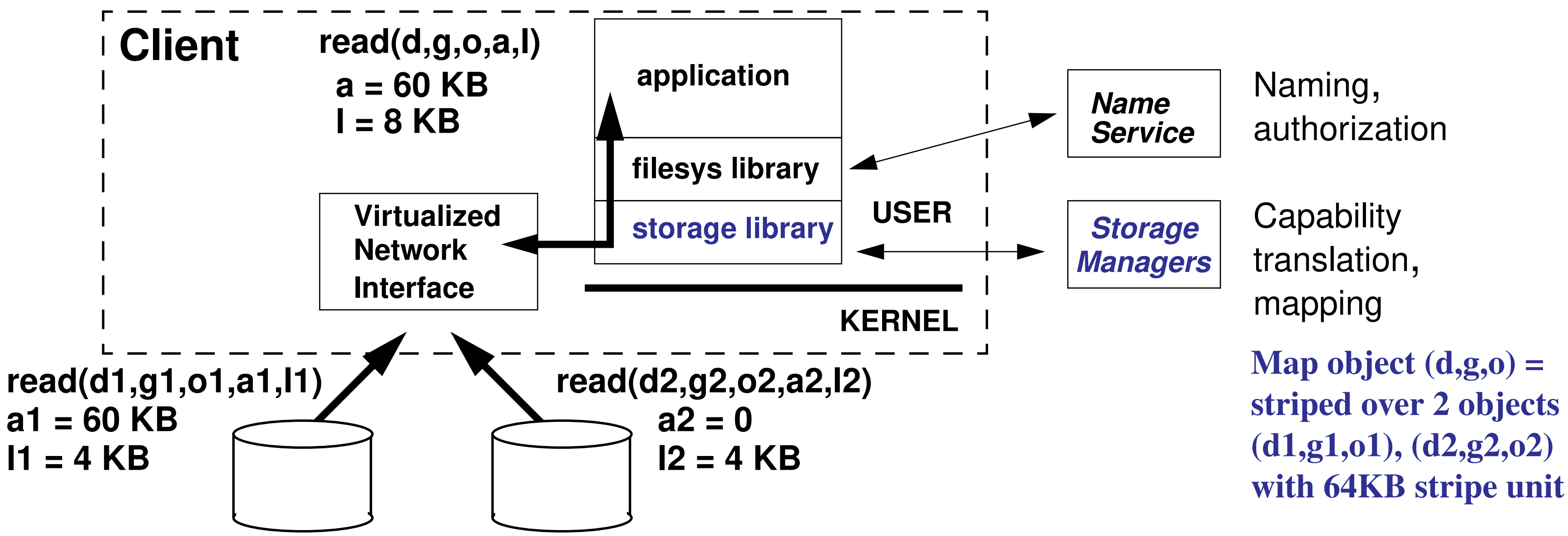
Minimize Operating System Interference

Synergy with user-level network access (Intel VIA)

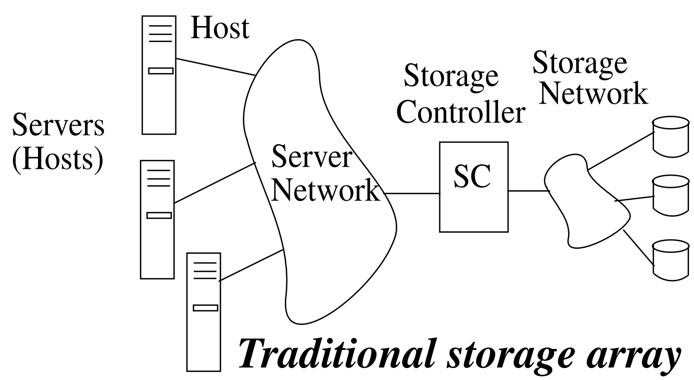
- security (integrity) not dependent on OS assurances/checking
- NIC protocol processing leaves client to run application
- defines client-processed virtual volume interface

Asynchronous storage management oversight

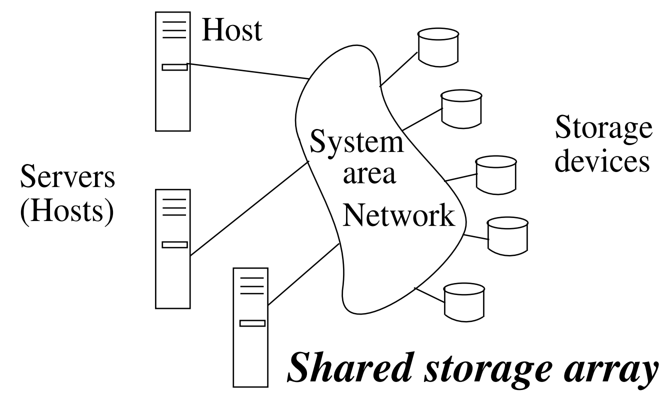
- file (name) management of aggregate unaware of components
- first storage access installs maps in client for aggregate object



Multi-host Shared Storage Arrays



Traditional storage array



Shared storage array

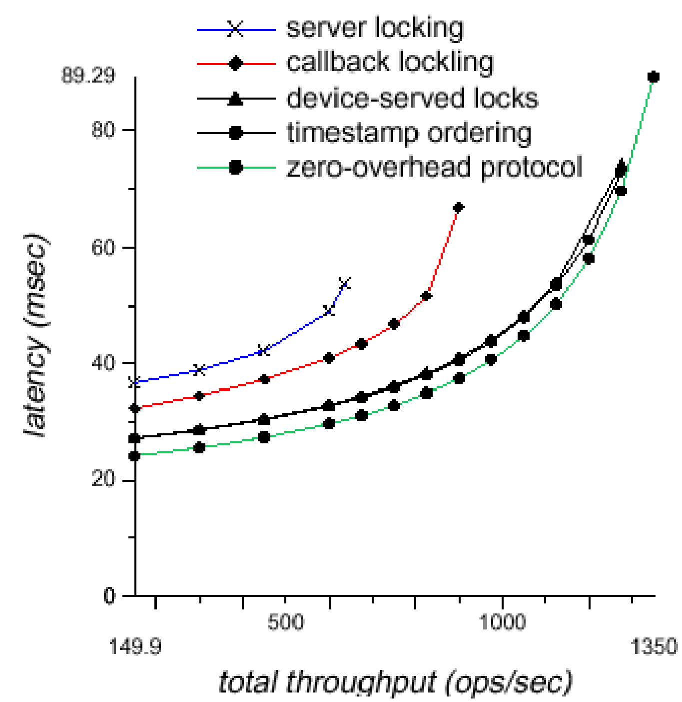
Serializability no longer easy

- not a panacea for reckless hosts
- databases do all serializability at host

Complex ordering uses locks

- central lock server
- callback, leased central lock server
- parallel, device-based locks
- device-based timestamp ordering

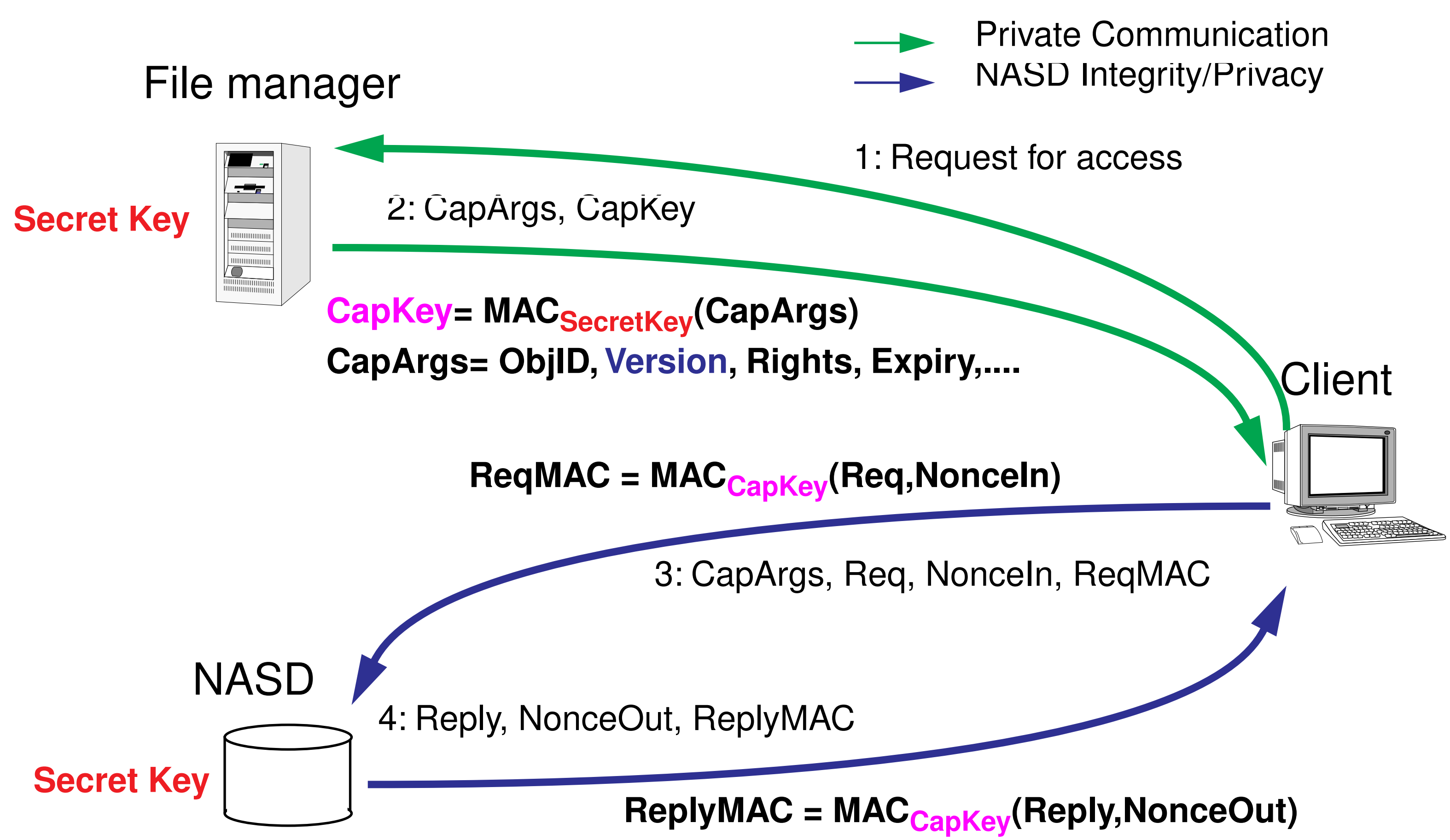
Device support effective



NASD Security Enforcement Protocol Detail

Based on digital signatures (MAC); manager revocable

- client's key is derived from manager's key; cacheable at client
- NASD need not record any per-client long-term state



NASD Technology Transfer

CMU Network-Attached Secure Disk (NASD)

- founded NSIC NASD WG (CMU, IBM, HP, Seagate, Quantum, StorageTek)

NSIC Network-Attached Storage Devices (NASD)

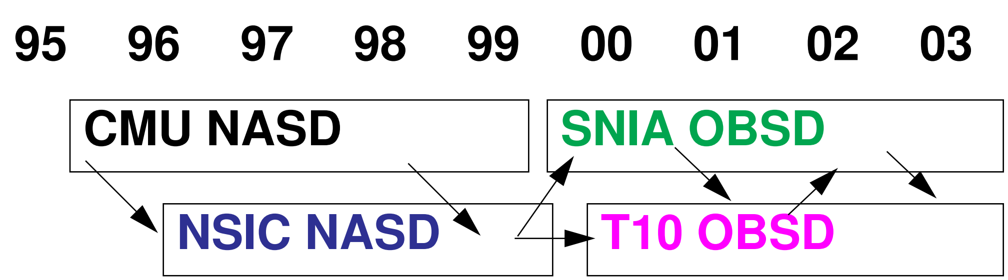
- produced Object-Based Storage Device (OBSD) interface proposal

ANSI X3 T10 committee (storage interace standards)

- 11/99 oversight board 17-0 for taking to plenary (IBM, Seagate, Quantum, Compaq, HP, SUN, DG, Adaptec, LSIL, ENDL, etc.); editor/chair volunteer

Storage Networking Industry Association

- 10/99 working group kickoff (IBM, Seagate, Quantum, Compaq, Adaptec, LSIL, STK, Auspex, PLC, Veritas, Fujitsu, Amdahl, Intel, 3Com, Gadzoox)



Carnegie Mellon
Parallel Data Laboratory

Object-Based Storage Devices (OBSD)

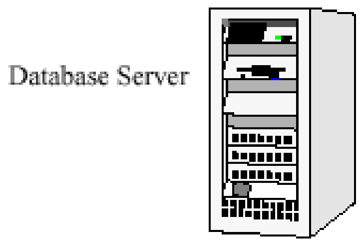
OLD: Sector-Based Storage Devices (SBSD)

- Fixed number of fixed-sized blocks (sectors)
- Naming represents physical allocation
- Format-time subdivision into set of contiguous-spaces (partitions)
- Access unit is one or more blocks
- Arbitrary accesses schedulable by host (queue tags)
- Physical zones differentiate peak bandwidth

NEW: Object-Based Storage Devices (OBSD)

- Variable number of variable sized blocks (objects)
- Naming implies no physical quality
- Anytime subdivision of objects into named sets (object groups)
- Access unit is arbitrary byte range
- Arbitrary accesses bound to specific Quality of Service (sessions)
- per-object tags for physical, logical, performance state (attributes)

NASD Followon Work: Active Disks

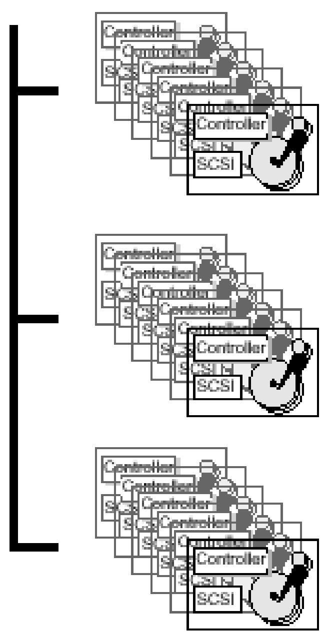


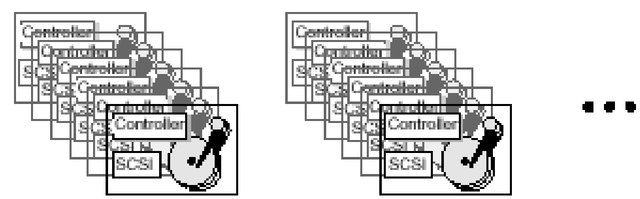
Digital AlphaServer 8400
- 12 x 612 MHz 21164
- 8 GB memory
- 3 64-bit PCI busses
- 29 FWD SCSI controllers

= 7,344 total MHz

3 x 266 = 798 MB/s

29 x 40 = 1,160 MB/s

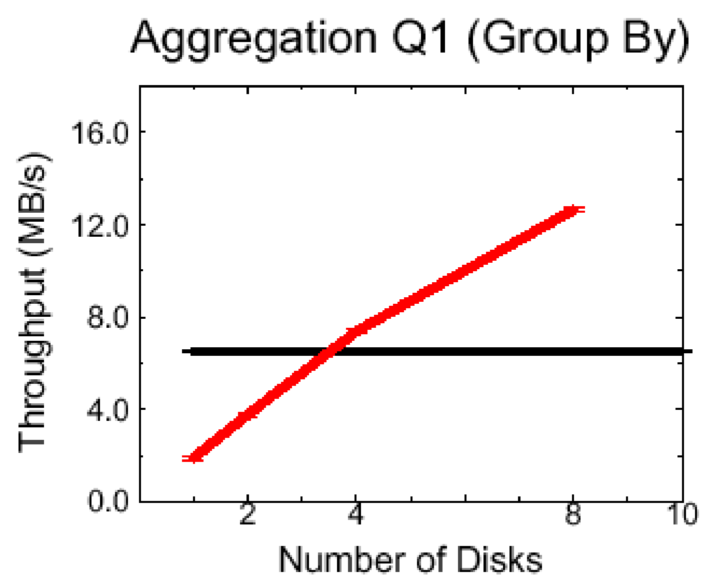




Storage
- 520 rz29 disks
- 4.3 GB each
- 2.2 TB total

= 104,000 total MHz
(with 200 MHz drive chips)

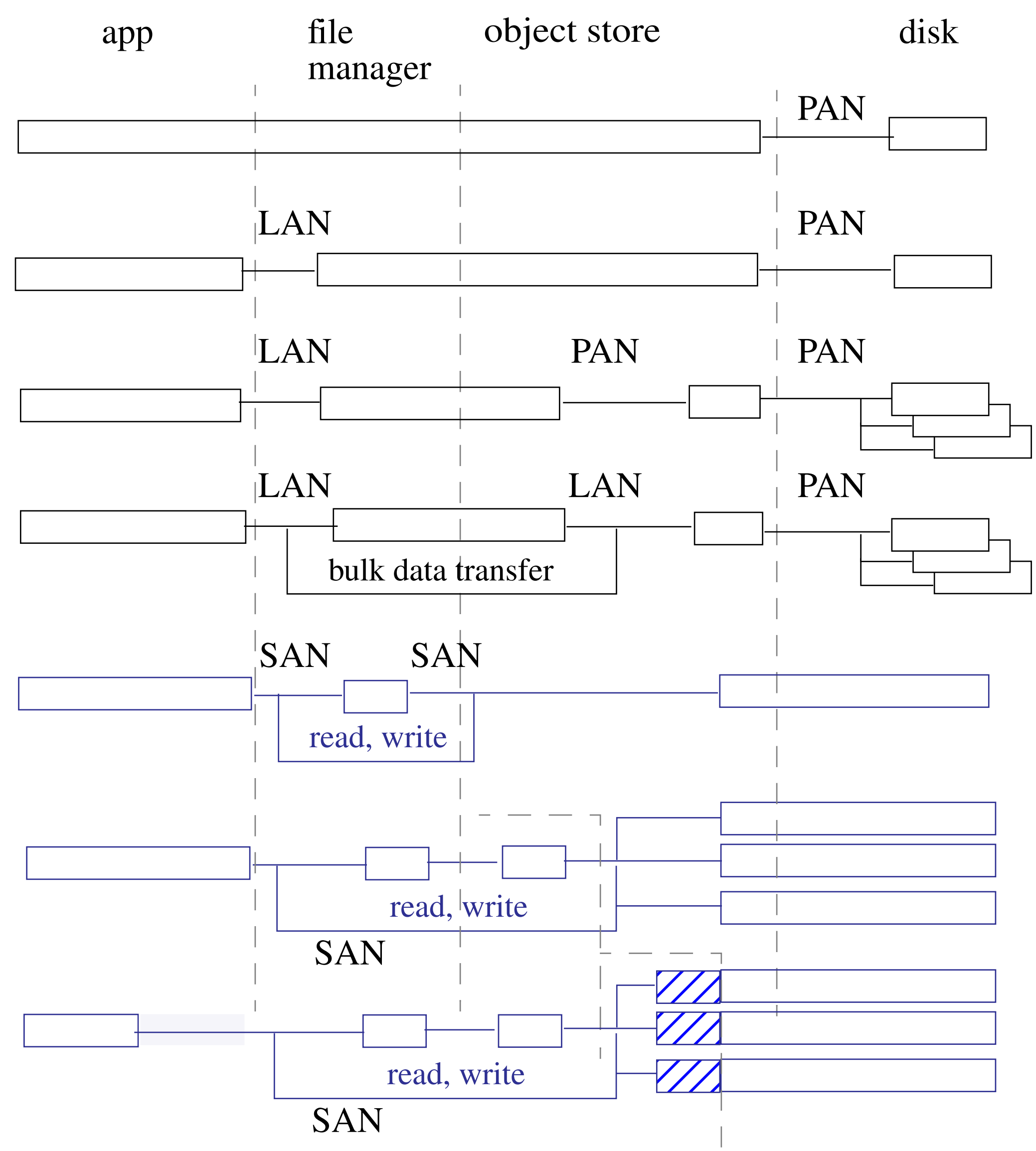
= 5,200 total MB/s
(at 10 MB/s per disk)



PostgreSQL 6.5

- drives (133MHz, 64MB)
  server (500MHz, 256MB)

Storage Interface Evolution Taxonomy



computers in data/control path	
2	Local Filesystem
3	Distributed FS (DFS)
4	DFS + RAID
3/4	DFS + HPSS bypass
2/2+	DFS w/ NASD
2/2++	DFS w/ NASD, Cheops
2/2++	DFS w/ NASD, Cheops and Active Disks