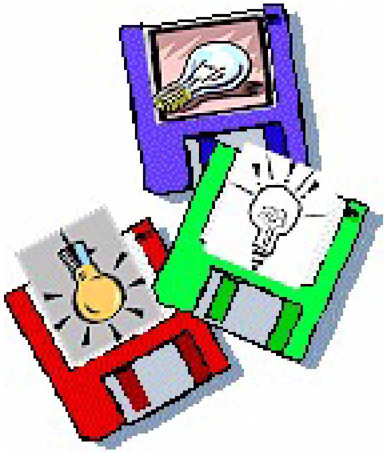


Active Disks - Remote Execution for Network-Attached Storage

Erik Riedel

Parallel Data Laboratory,
Center for Automated Learning and Discovery
Carnegie Mellon University
www.pdl.cs.cmu.edu/Active

UC-Berkeley Systems Seminar
8 October 1998

Outline

Network-Attached Storage

Opportunity

Active Disks

Applications

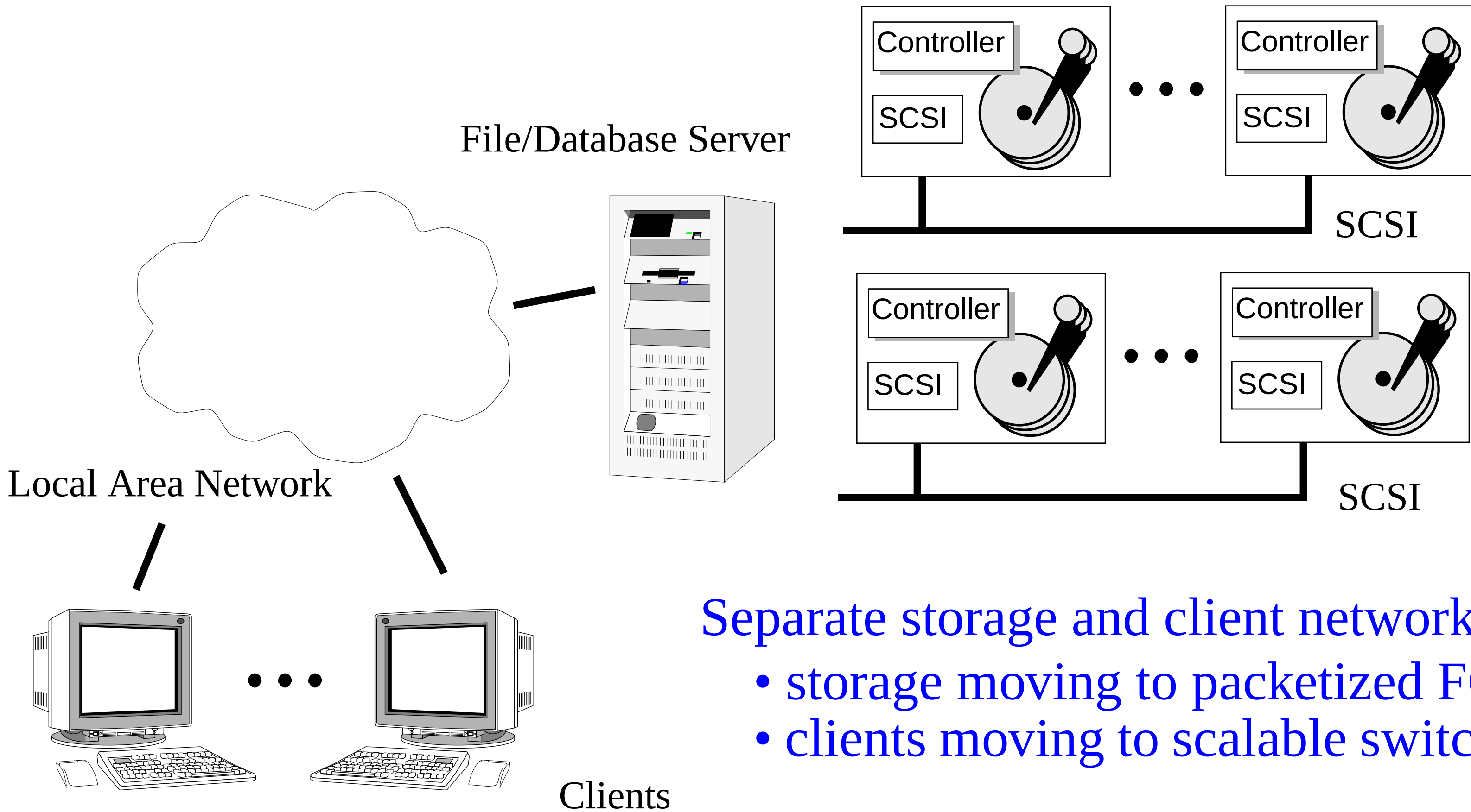
Performance Model

Prototype

Summary

Today's Server-Attached Disks

Store-and-forward data copy through server machine

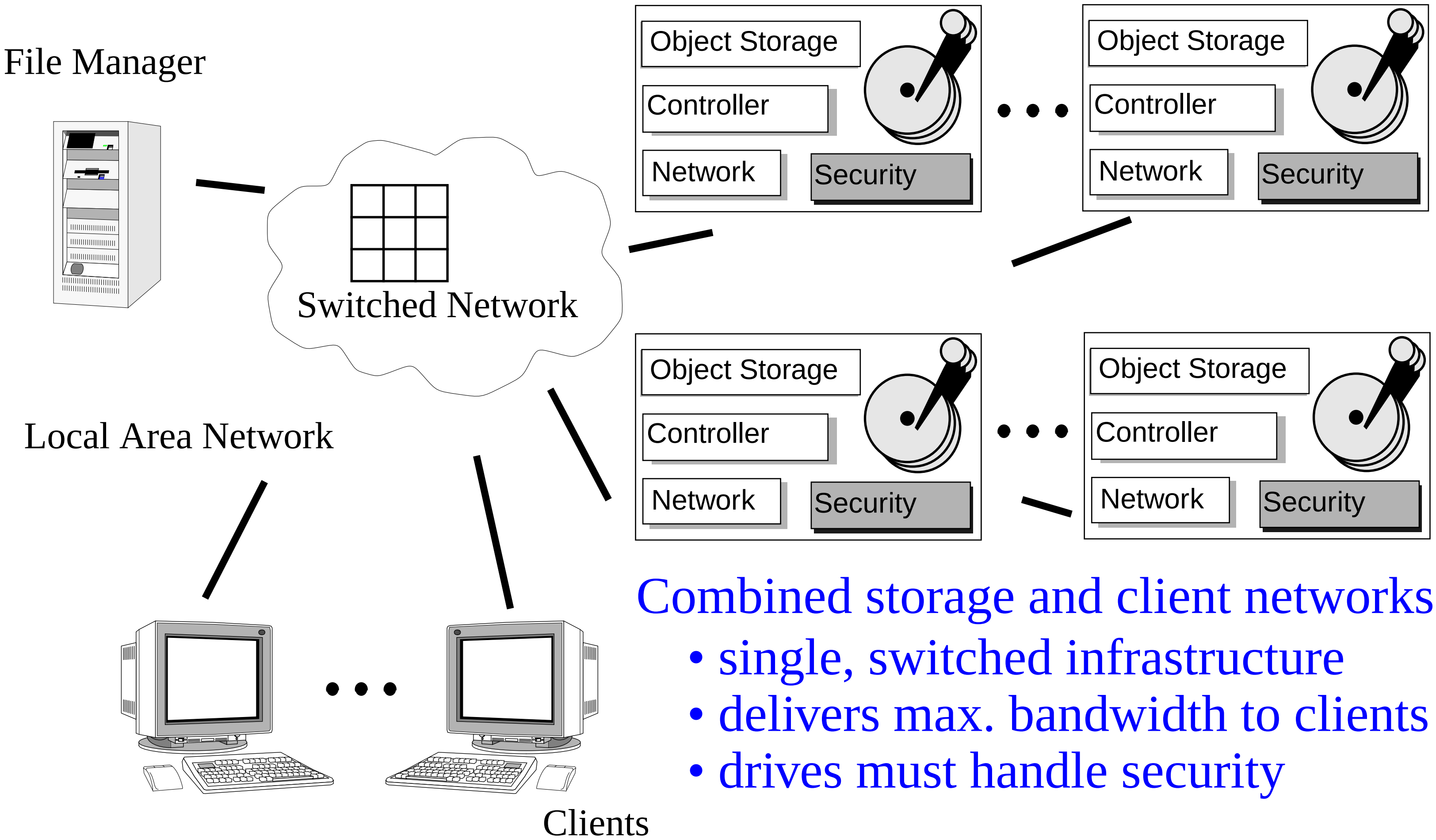


Separate storage and client networks

- storage moving to packetized FC
- clients moving to scalable switches

Network-Attached Secure Disks

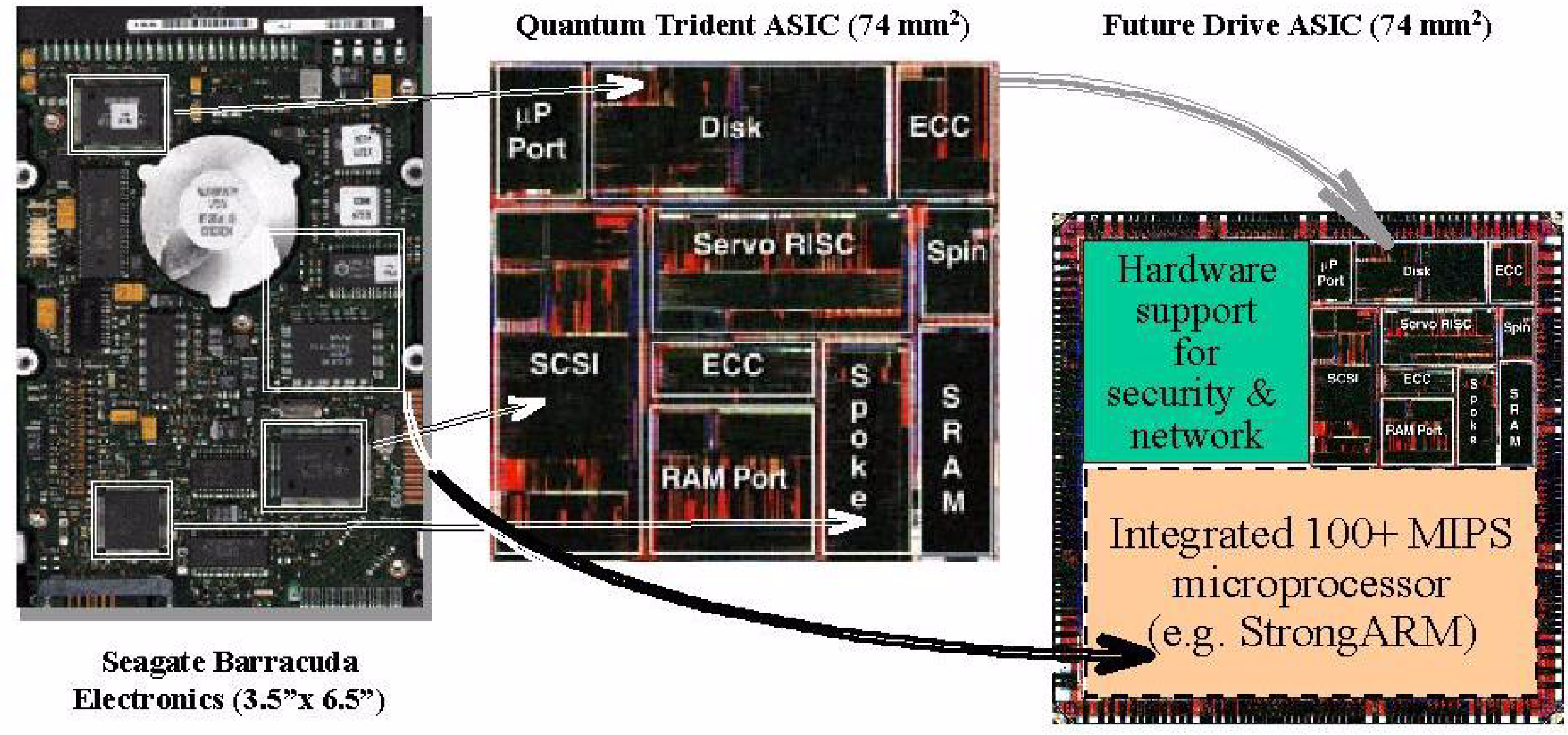
Eliminate server bottleneck w/ network-attached



Combined storage and client networks

- single, switched infrastructure
- delivers max. bandwidth to clients
- drives must handle security

Excess Device Cycles Are Coming



Higher and higher levels of integration in drive electronics

- specialized drive chips combined into single ASIC
- technology trends push toward integrated control processor
- 100 MHz, 32-bit superscalar w/ 2 MB on-chip RAM in '98

Large database systems - lots of disks, lots of power

System	Processing (MHz)		Data Rate (MB/s)	
	CPU	Disks	I/O Bus	Disks
Compaq Proliant TPC-C	4 x 200=**800**	*113* x 25=**2,825**	**133**	**1,130**
Microsoft Terraserver	4 x 400=**1,600**	*320* x 25=**8,000**	**532**	**3,200**
Digital AlphaServer 500 TPC-C	1 x 500=**500**	*61* x 25=**1,525**	**266**	**610**
Digital AlphaServer 4100 TPC-D	4 x 466=**1,864**	*82* x 25=**2,050**	**532**	**820**

- **assume disk offers equivalent of 25 host MHz**
- **assume disk sustained data rate of 10 MB/s**

Lots more cycles and MB/s in disks than in host

Advantage - Active Disks

Basic advantages of an Active Disks system

- **parallel processing** - lots of disks
- **bandwidth reduction** - filtering operations common
- **scheduling** - little bit of computation can go a long way

Appropriate applications

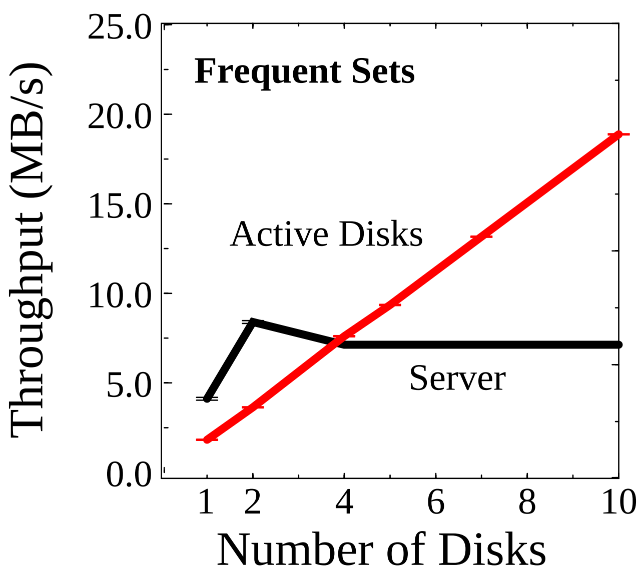
- execution time dominated by data-intensive core
- allows parallel implementation of core
- small memory footprint
- small number of cycles per byte of data processed

Example Application

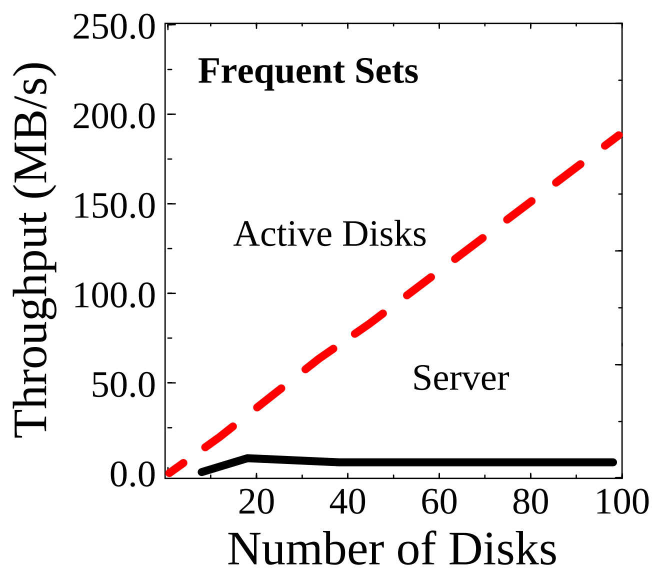
Data mining - association rules [Agrawal95]

- **frequent sets summary counts**
- **count of *1-itemsets* and *2-itemsets***
- **milk & bread => cheese**
- **diapers & beer**

Prototype



Scaling Up



Basic Performance Model

Execution = `max`(processing, transfer, disk access)

- **`selectivity` is `#bytes-input` / `#bytes-output`**
- **assume fully overlapped pipeline (avoids Amdahl's law)**

Processing time per byte

- **Host: `#cycles/byte` / `host-cpu-speed`**
- **Disks: `#cycles/byte` / (`disk-cpu-speed * #disks`)**

Transfer time per overall byte

- **Host: `1` / `interconnect-data-rate`**
- **Disks: (`1` / `selectivity`) / `interconnect-data-rate`**

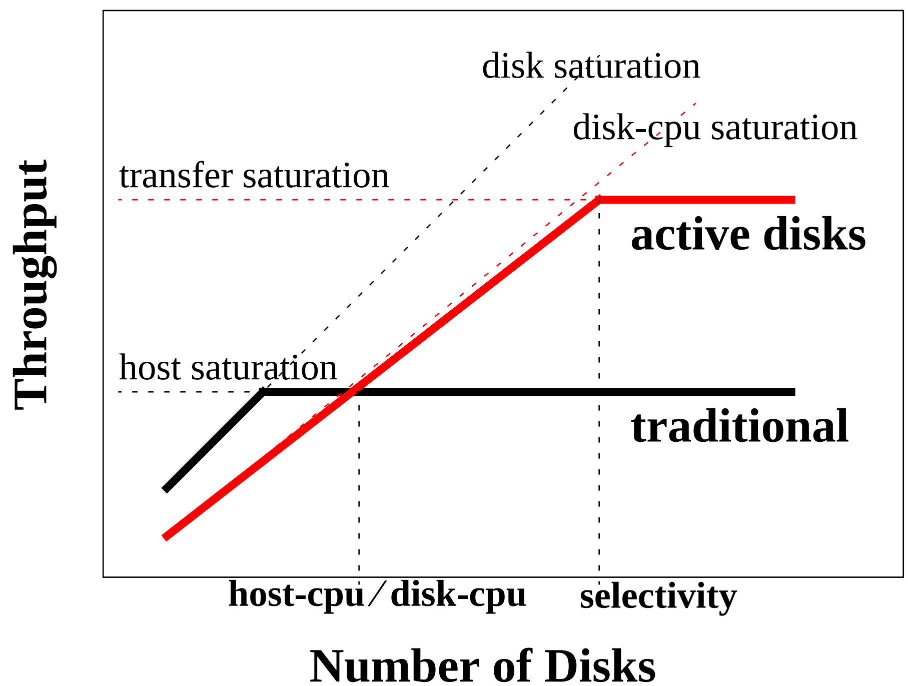
Disk access time per overall byte

- **Both: `1` / (`disk-data-rate * #disks`)**

Throughput Model

Scalable throughput

- speedup = (#disks)/(host-cpu-speed/disk-cpu-speed)
- (host-cpu/disk-cpu-speed) ~ 5   (two processor generations)
- selectivity = #bytes-input / #bytes-output



Additional Applications

Database - select

- extract records that match a particular predicate

Database - nearest neighbor search

- *k* records closest to input record
- with large number of attributes, reduces to scan

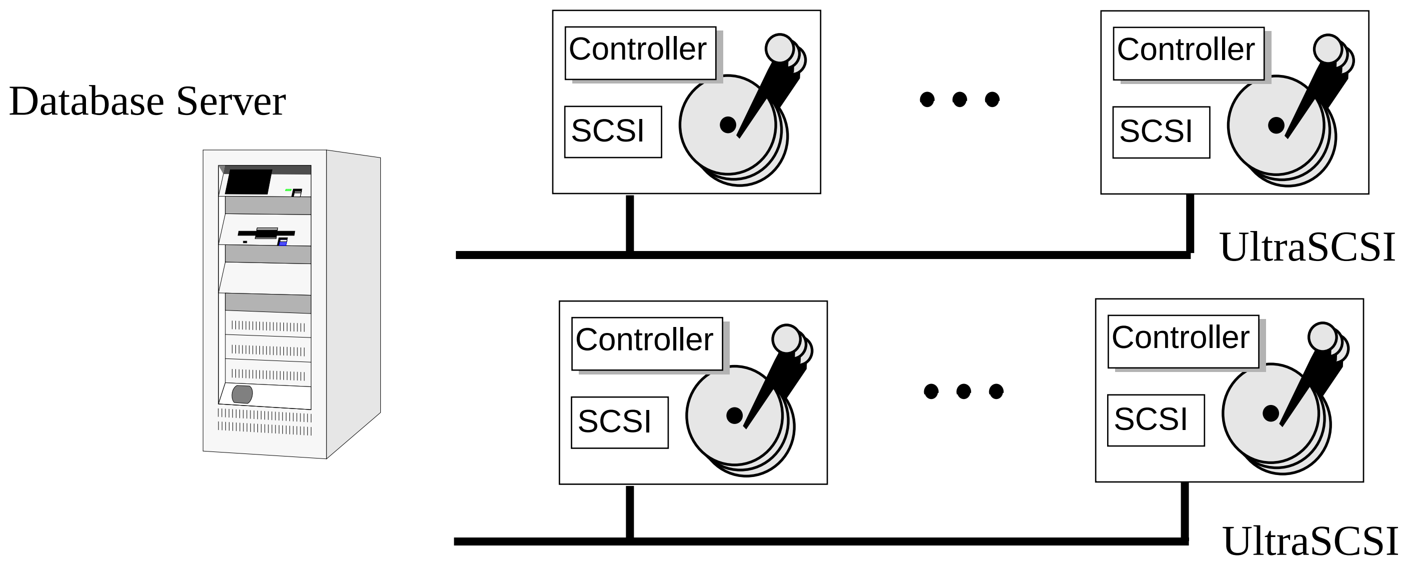
Multimedia - edge detection [Smith95]

- detect edges in an image

Multimedia - image registration [Welling97]

- find rotation and translation from reference image

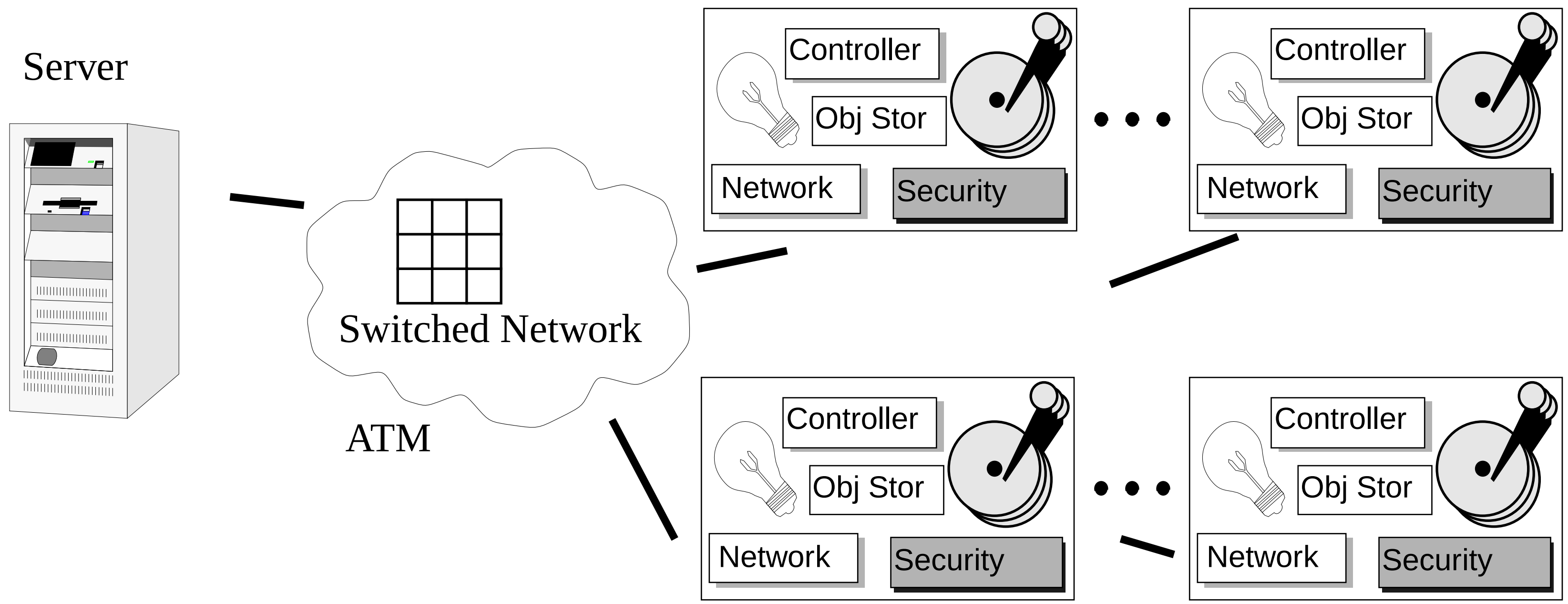
Traditional Server



Digital AlphaServer 500/500

- 500 MHz, 256 MB memory
- disks - Seagate Cheetah
- 4.5 GB, 10,000 RPM, 11.2 MB/s

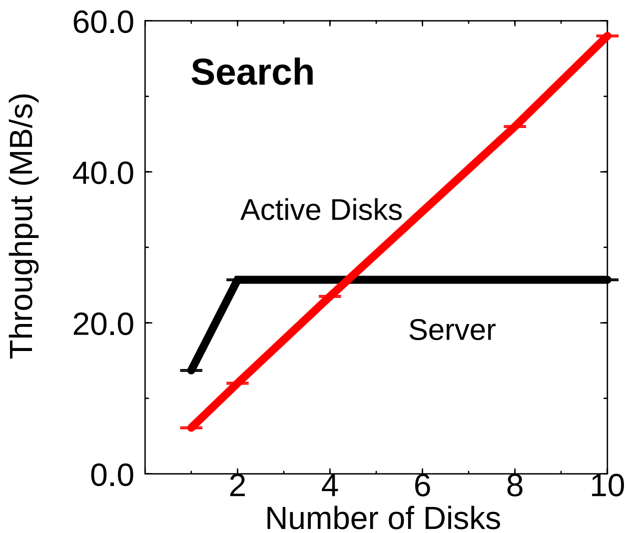
Server with Active Disks

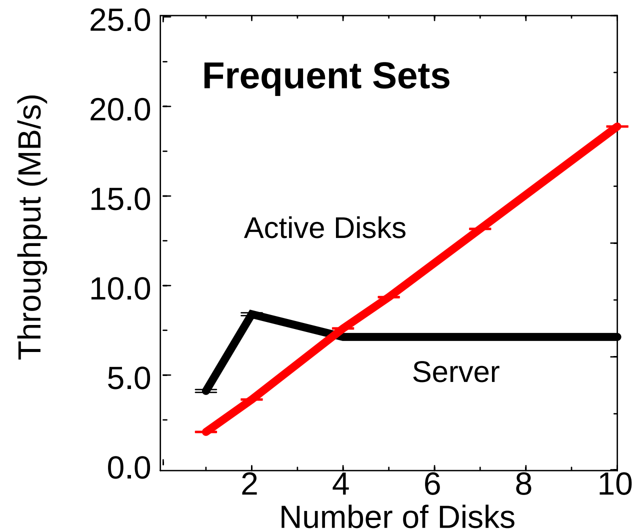


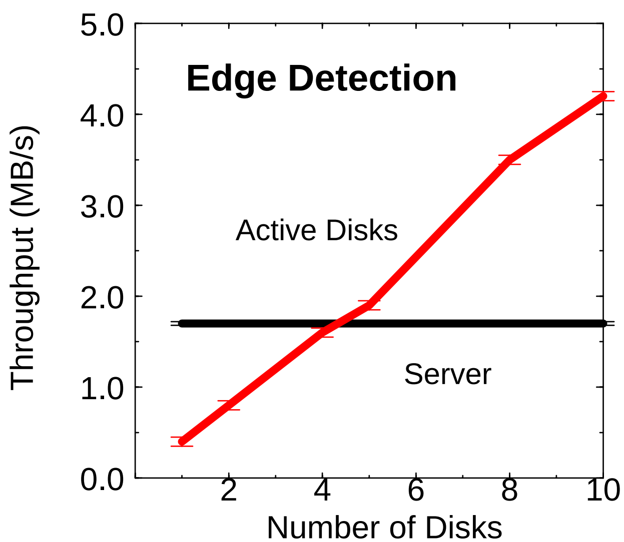
Prototype Active Disks

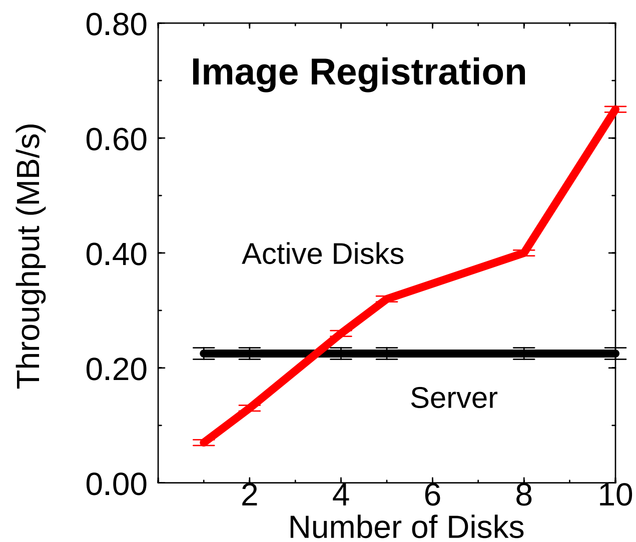
- Digital AXP 3000/400 workstation
- 133 MHz, software NASD prototype
- Seagate Medallist disks

Performance with Active Disks









Application Characteristics

Critical properties for Active Disk performance

- **cycles/byte => maximum throughput**
- **memory footprint**
- **selectivity => network bandwidth**

application	input	computation (cycles/byte)	throughput (MB/s)	memory (KB)	selectivity (factor)	bandwidth (KB/s)
Select	m=1%	7	28.6	-	100	290
Search	k=10	7	28.6	72	80,500	0.4
Frequent Sets	s=0.25%	16	12.5	620	15,000	0.8
Edge Detection	t=75	303	0.67	1776	110	6.1
Image Registration	-	4740*	0.04	672	180	0.2
Select	m=20%	7	28.6	-	5	5,700
Frequent Sets	s=0.025%	16	12.5	2,000	14,000	0.9
Edge Detection	t=20	394	0.51	1750	3	170

Summary

Technology trends provide the opportunity

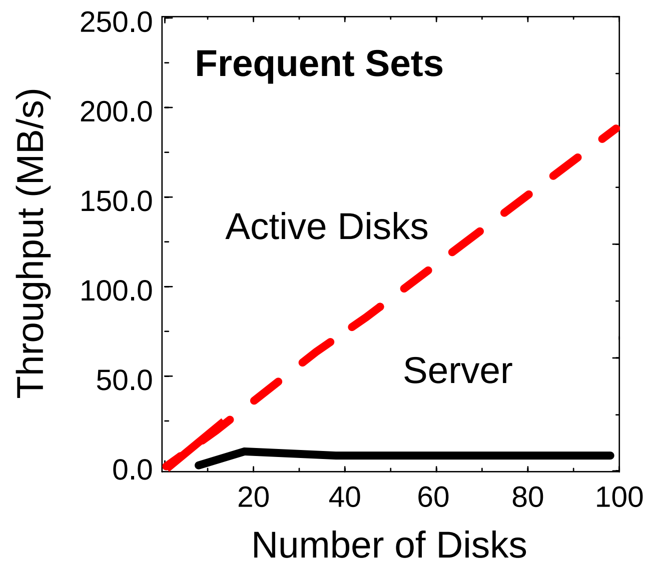
- “excess” cycles
- large systems => lots of disks => lots of power

Dramatic benefits possible

- *application examples* - data mining and multimedia
- *characteristics for big wins* - parallelism, selectivity
- *basic advantage* - compute close to the data

Prototype

- speedup of 2x on 10 disks
- scales to 15x in 60 disk system
- bottleneck can be above 1000s of disks



Conclusions & Future Work

Leverage for Active Disks

- **powerful drive chips available now**
  - **Siemens Tri-Core [announced March ‘98, first silicon Sept ‘98]**
  - **Cirrus Logic 3CI [announced June ‘98]**
- **higher-level storage interfaces & security architecture**
  - **NASD [Sigmetrics ‘97, ASPLOS ‘98]**
  - **Object-oriented disks [Seagate and X3 T10], NSIC, SNIA**
- **aggressive applications**
  - **data mining [Center for Automated Learning & Discovery]**
  - **multimedia [Informedia, Digital Libraries]**

Challenges

- ***programming model* - partitioning, mobility, interfaces**
- ***resources* - driven by cost, reliability, volume**
- ***management* - disk come in boxes of ten**
- ***additional application classes* - sort/join, storage management**

Related Work

Database Machines (CASSM, RAP, Gamma)

- higher disk bandwidth, parallelism
- general-purpose programmability

OS/Database Extensions

- application-specific specialization/extension (SPIN, VINO)
- data type extensions (Sybase, Informix)

Parallel Programming

- automatic data parallelism (HPF), task parallelism (Fx)
- parallel I/O (Kotz, IBM, Intel)

Other “Smart” Disks

- offload SMP database functions, disk layout (Berkeley)
- select, sort, images via extended SCSI (Santa Barbara)

Why Isn't This Parallel Programming?

It is

- parallel cores
- distributed computation
- serial portion needs to be small

Disks are different

- must protect the data
- must continue to serve demand requests
- memory/CPU ratios driven by cost, reliability, volume
- come in boxes of ten
- advantage - compute close to the data

Opportunistically use this power

- e.g. data mining possible on an OLTP system