Active Storage For Large-Scale Data Mining and Multimedia

Erik Riedel

Parallel Data Laboratory

Carnegie Mellon University

www.cs.cmu.edu/~riedel

CALD Seminar

Center for Automated Learning and Discovery

April 3, 1998

Outline

Network-Attached Disks

Industry Trends

Active Disks

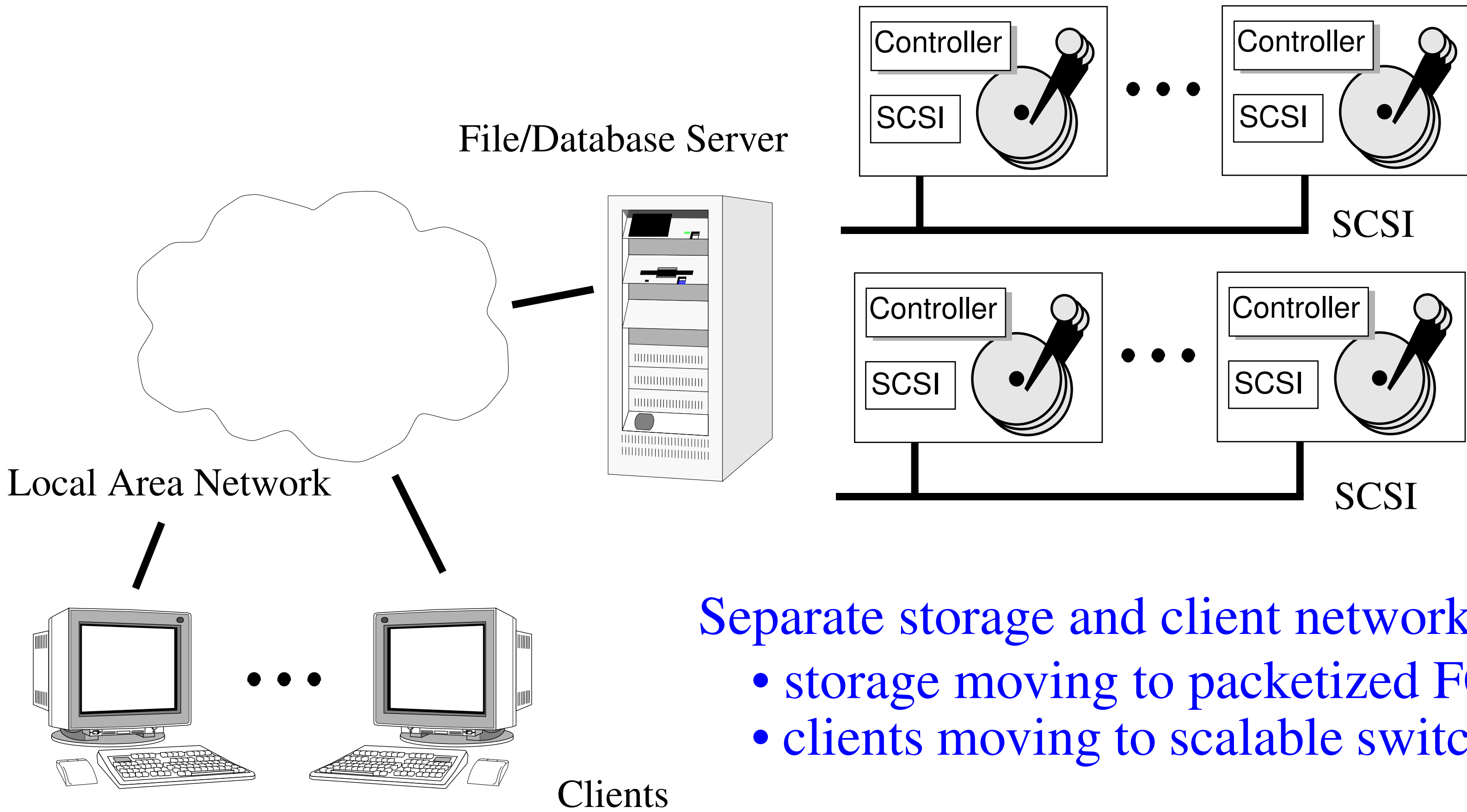
Applications

Speedups

Ideal Application

Today's Server-Attached Disks

Store-and-forward data copy through server machine

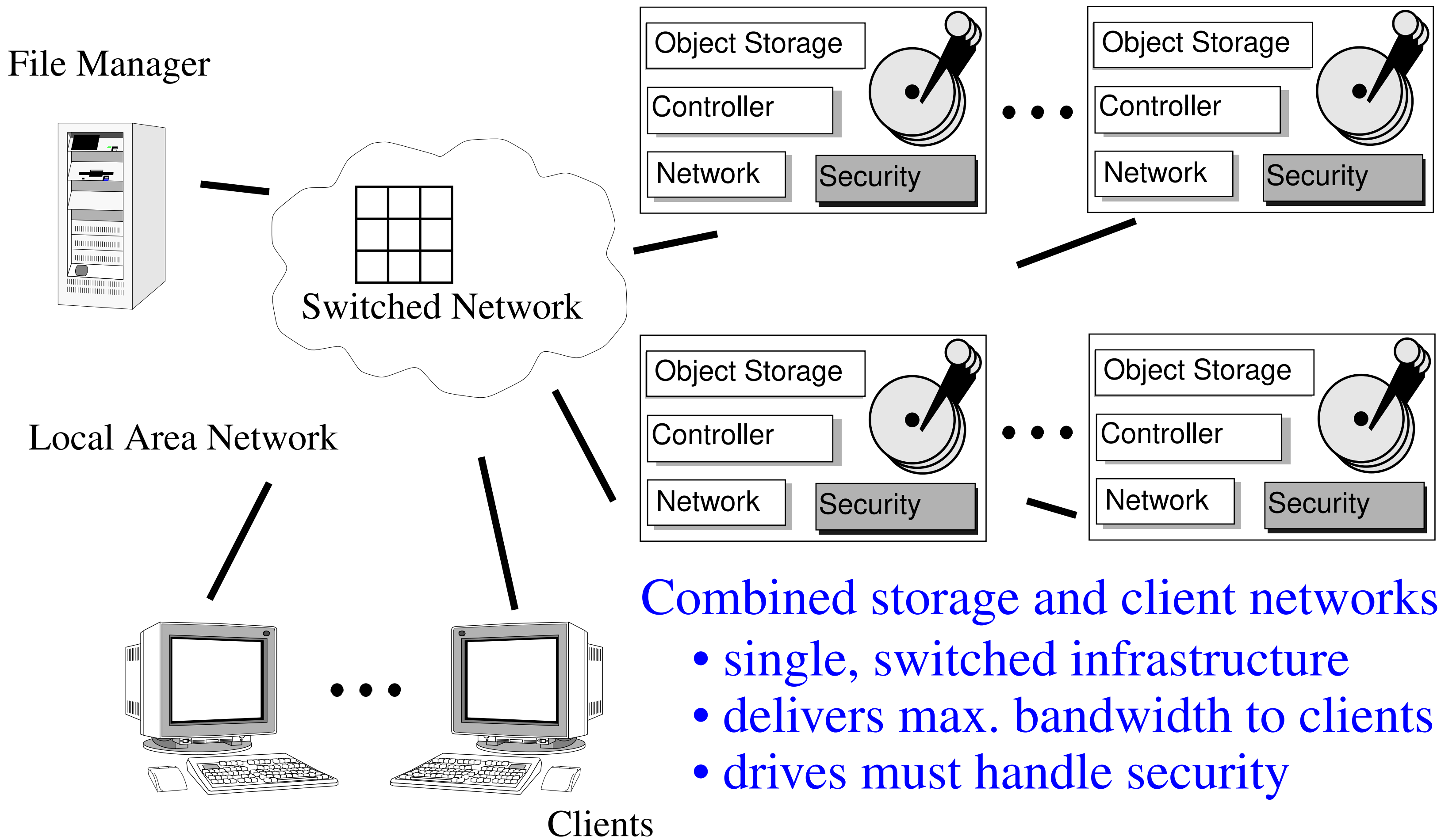


Separate storage and client networks

- 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

Storage Industry Trends

Drive interface is changing

- Drive bandwidth - now 15 MB/s and rising at 40% per year
- Disk-embedded, high-speed, packetized SCSI
- E.g. 100-1000 Mb/s Fibrechannel interconnect

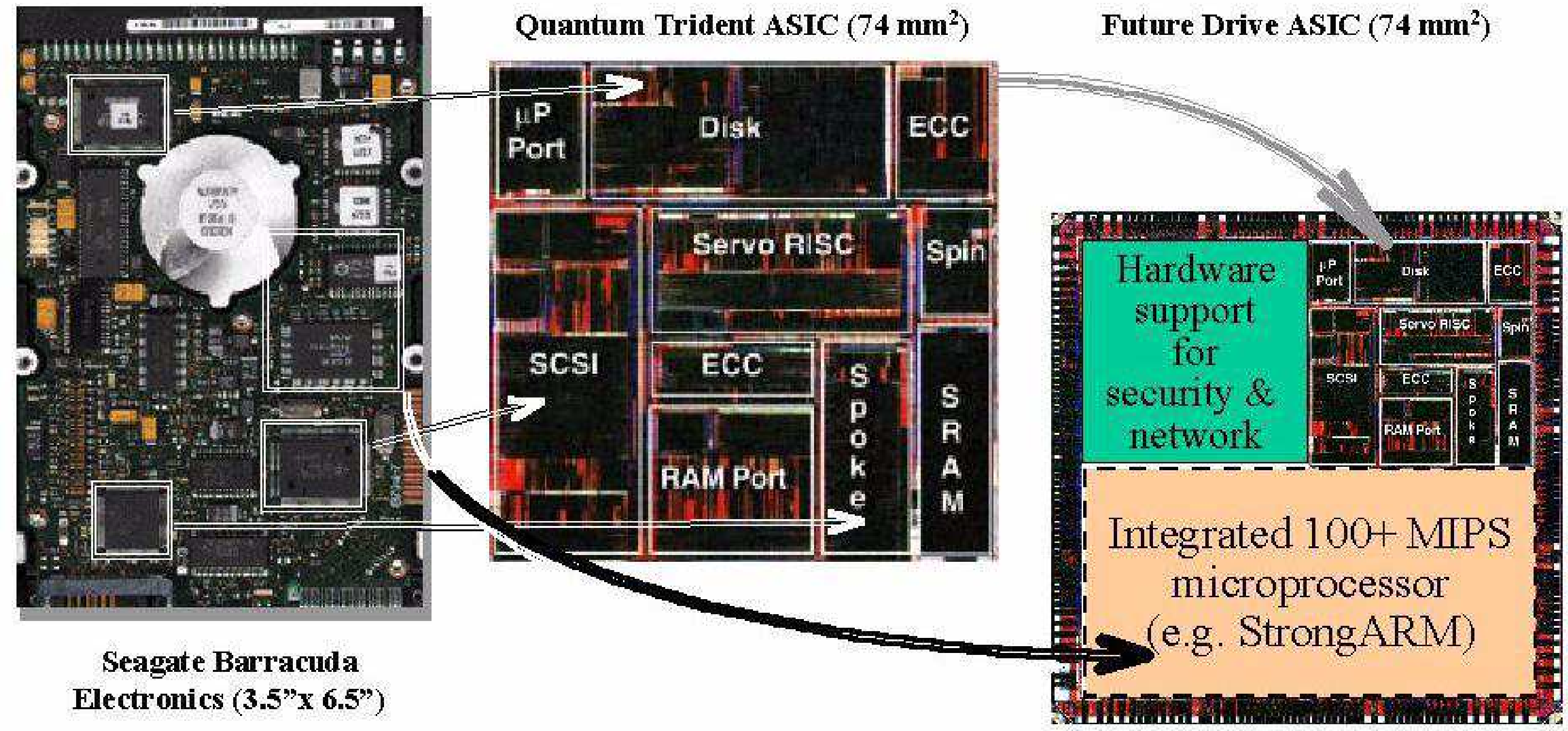
Competition is increasingly based on code in the drive

- RAID support to off-load parity update
- Dynamic mapping underneath SCSI
- Increasingly sophisticated prefetching/caching
- Cost of managing storage 3-7x storage cost per year

On-drive cycles are available

- RISC core coming in integrated function drive ASIC
- Control processor not on critical path

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 available in '98

Opportunity

Sampling of large-scale database systems

System	Processing (MHz)		Data Rate (MB/s)	
	CPU	**Disks**	**CPU**	**Disks**
Compaq TPC-C	4x200=**800**	*113*x25=**2,800**	**133**	**1,130**
Microsoft Terraserver	4x400=**1,600**	*320*x25=**8,000**	**532**	**3,200**
Digital 500 TPC-C	1x500=**500**	*61*x25=**1,525**	**266**	**610**
Digital 4100 TPC-D	4x466=**1,864**	*82*x25=**2,050**	**532**	**820**

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

More cycles and MB/s in disks than in host

Active Disk Fundamentals

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

Simple 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`

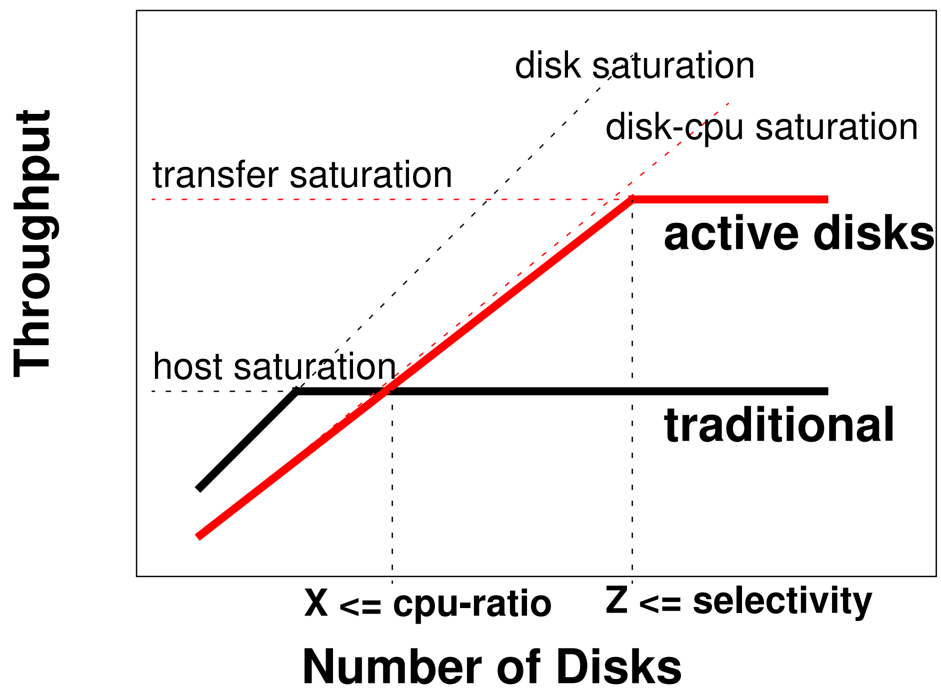
Disk access time per overall byte

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

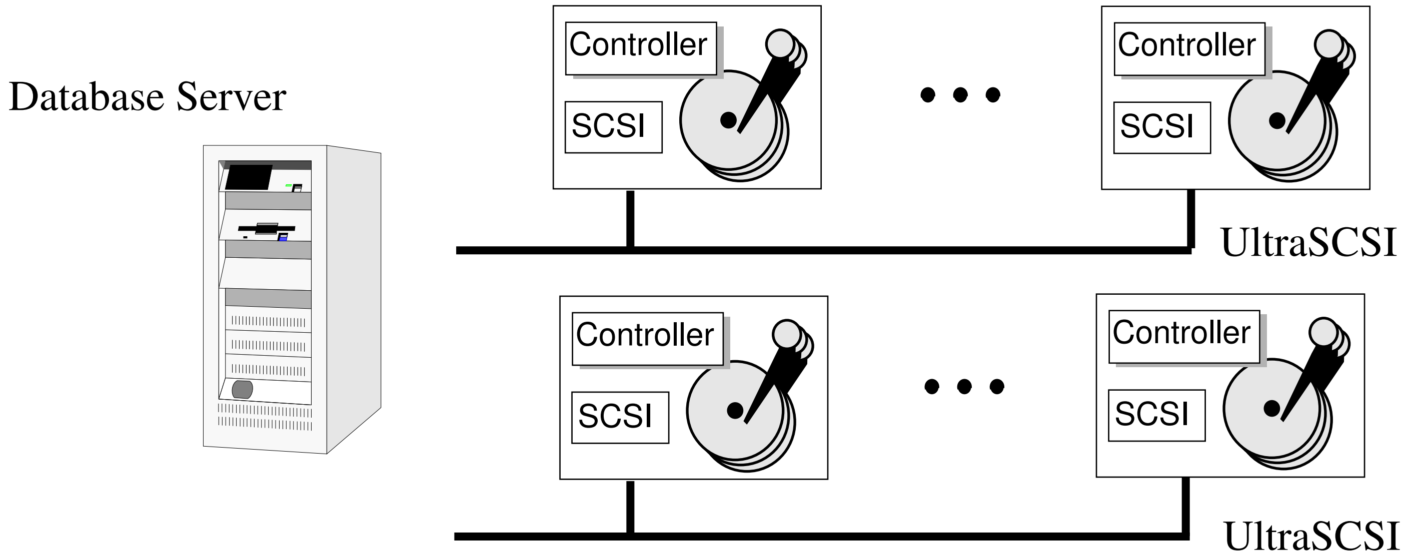
Throughput Model

Speedup

- (#disks*disk-cpu-speed) / host-cpu-speed   [X<#disks<Z]
- > selectivity*(host-cpu / disk-cpu-speed)   [#disks>Z]
- (host-cpu/disk-cpu-speed) ~ 5 per host cpu (2 generations)

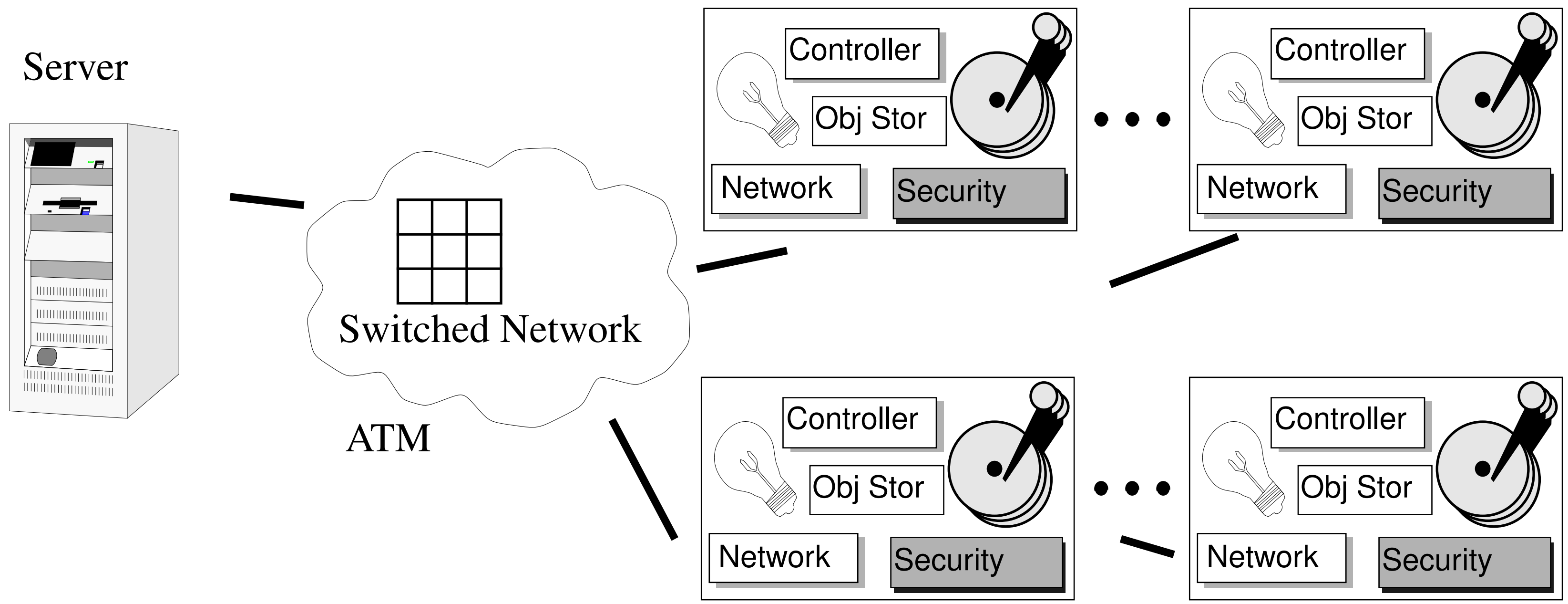


Traditional Server



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

Data-Intensive Applications

Database - nearest neighbor search

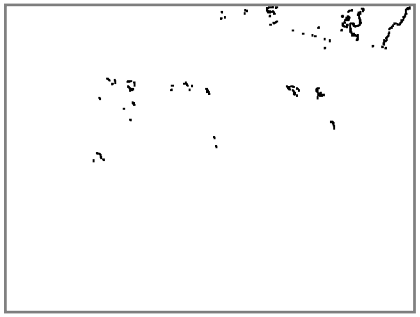
- *k* records closest to input record
- with large number of attributes, reduces to scan

Data mining - association rules [Agrawal95]

- count of *1-itemsets* and *2-itemsets*

Multimedia - edge detection [Smith95]

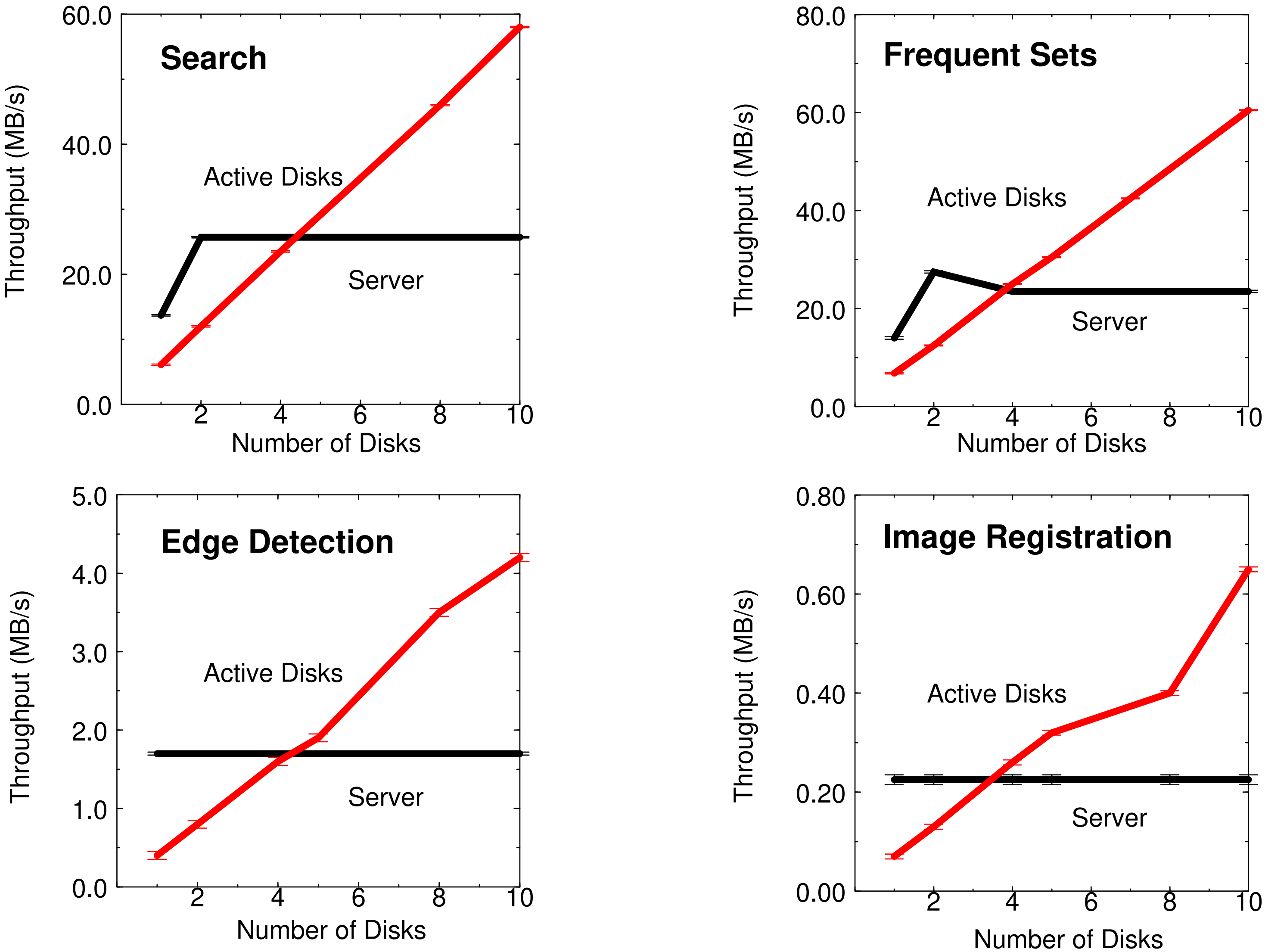
- detect edges in an image



Multimedia - image registration [Welling97]

- find rotation and translation from reference image

Performance with Active Disks



Application Characteristics

Critical properties for Active Disk success

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

application	input	computation (cycles/byte)	throughput (MB/s)	memory (KB)	selectivity (factor)	bandwidth (MB/s)
Select	m=1%	7	28.6	-	100	0.3
Search	k=10	17	11.8	0.6	100,000	0.0001
Frequent Sets	s=0.25%	15	13.3	220	14,000	0.001
Edge Detection	t=75	394	0.51	256	175	0.002
Image Registration	-	2387*	0.08	768	230	0.0003
Select	m=20%	7	28.6	-	5	5.7
Frequent Sets	s=0.025%	15	13.3	2,000	14,000	0.001
Edge Detection	t=20	394	0.51	256	3	0.2

Parallel Sample Sort

- **computation at drives saves one full network transfer**
- **data goes to the "right" place sooner**
- **instead of exchanging data among client nodes**

Step	Parallel Sample Sort	Sample Sort for Active Disks
1	Sample data	Sample data using `sample()` on drives
2	Create distribution histogram	Create distribution histogram
3	Read data into clients from local disks	Read data into clients using `scan()`
4	Distribute data among clients by histogram	
5	Sort locally at each client	Sort locally at each client
6	Write back to local disks in sorted order	Write back to drives in sorted order

Future Directions

Executables downloaded into drives

- safe, secure, controllable, continuous media

Applications: schedule, semantic extension

- sort, join, collective I/O, video, web, storage mgmt

Compiler-assisted "Disklet" definition

- library, framework support, automatic partitioning

Active networking for storage

- NASD capabilities extended to network components
- in network: protocol conversion, caching, dynamic routing