ECE 5730

Memory Systems

Spring 2009

Disk Case Study

Disk Power Management

Cornell University

Announcements

- **No class tonight**

- **Quiz 13 average = 6.25**

- **Exam II**
  - May 7, 7:00-10:00pm, Hollister 314
  - Covers material from 3/10-4/28 but excluding 4/22 (Lectures 14-21, 23-24)

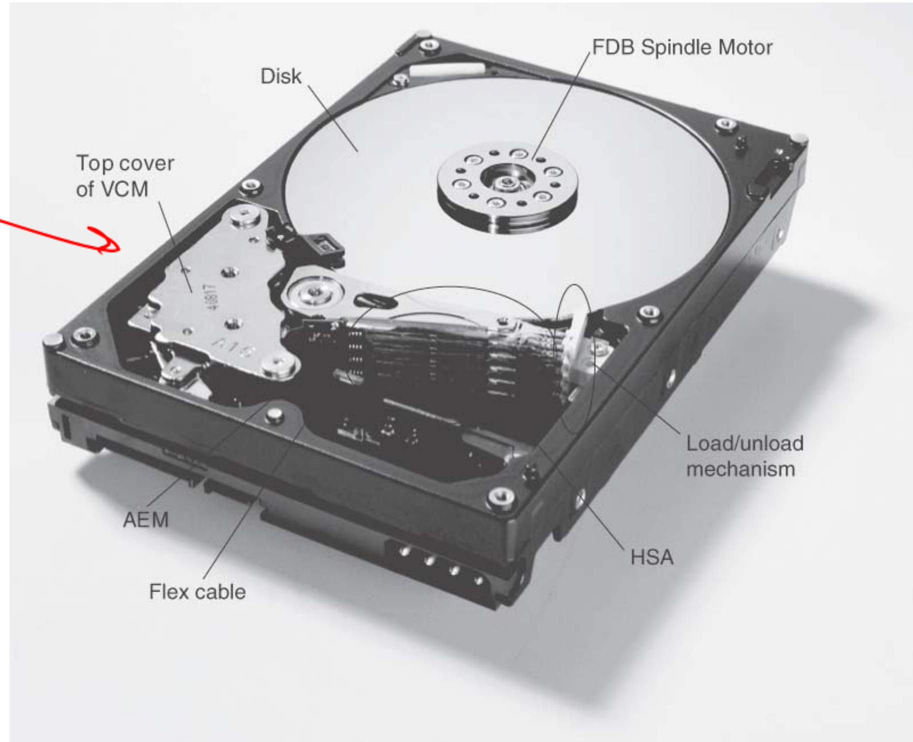
- **Final report (15-25 double-spaced pages)**
  - Email Word or PDF to me by 11:59pm on May 1
  - 20 points off final project grade if late

Hitachi Deskstar 7K500/E7K500

- **3.5” 500GB hard drive targeting desktops, video recorders, gaming**

- **100GB platters**
  - **50GB/surface**
  - **Areal density = 76 Gb/in^2**
  - **bpi = 720Kb/in**
  - **tpi = 105Ktracks/in**

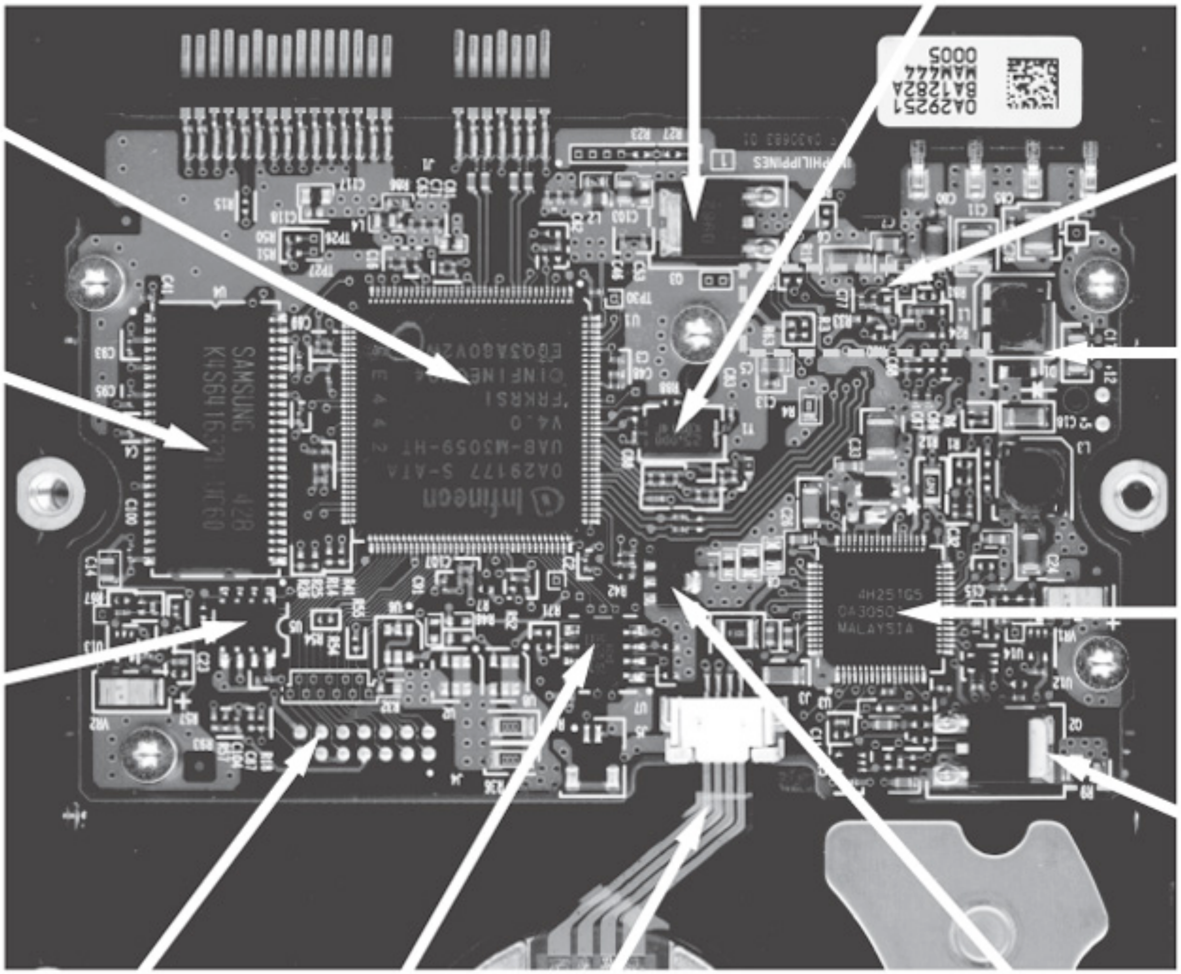
Voice Coil Motor



- **7200 rpm**

- **ATA/SATA interface**

Electronics Card



Disk Cache

- 8MB or 16MB DRAM
  - 271KB for controller firmware (loaded from EEPROM on boot)
  - Remainder for cache

- Circular buffer with variable size segments

Data Layout

- **Cylinder mode formatting across 10 heads** (5 platters, 10 surfaces)
- **Note: the 7K500 specification (Rev 1.5) lists 30 zones**

Zone	Start Cyl. No.	End Cyl. No.	Sectors Per Track
0	0	1999	1242
1	2000	3999	1215
2	4000	8999	1188
3	9000	14999	1170
4	15000	22499	1147
5	22500	30499	1125
6	30500	34499	1080
7	34500	38999	1026
8	39000	43499	1012
9	43500	47999	990
10	48000	52499	972
11	52500	56999	945
12	57000	61499	918
13	61500	65999	900
14	66000	69999	855
15	70000	73499	843
16	73500	80499	810
17	80500	83499	756
18	83500	85999	742
19	86000	88499	720
20	88500	90999	702
21	91000	93999	675
22	94000	96999	648
23	97000	99999	630
24	100000	103182	594

← outer diameter

smaller # of sectors per track w/ decreasing circumference

← inner diameter

[26.1]

Command Overhead

- Read/Write: Time from writing command to the register to DRQ = 1 (excluding seek and RL)
  Drive Request
  - DRQ: ready to transfer data to/from host

Command type (Drive is in quiescent state)	**Time (typical) (ms)**	**Time (typical) for queued command (ms)**
Read (cache not hit) (from Command Write to Seek Start)	0.5	0.5
Read (cache hit) (from Command Write to DRQ)	0.1	0.2
Write (from Command Write to DRQ)	0.015	0.2
Seek (from Command Write to Seek Start)	0.5	not applicable

[7K500v1.5]

Sustained Data Rate

- $T_{HS} = T_{CS} = 1.5\text{ms}$
- $N = 10$
- rpm = 7200

$$SDR = \frac{N \times SPT \times 512}{(N \times 60)/rpm + (N-1) \times T_{HS} + T_{CS}}$$

constant angular velocity
⇓
variable linear velocity
⇓
variable bitrate

Zone	Start PBA	End PBA	Sustained data rate (MB/s)[a]
0	0	24839999	64.67
1	24840000	49139999	63.26
2	49140000	108539999	61.86
3	108540000	178739999	60.92
4	178740000	264764999	59.72
5	264765000	354764999	58.58
6	354765000	397964999	56.23
7	397965000	444134999	53.42
8	444135000	489674999	52.69
9	489675000	534224999	51.55
10	534225000	577964999	50.61
11	577965000	620489999	49.20
12	620490000	661799999	47.80
13	661800000	702299999	46.86
14	702300000	736499999	44.52
15	736500000	766004999	43.89
16	766005000	822704999	42.17
17	822705000	845384999	39.36
18	845385000	863934999	38.63
19	863935000	881934999	37.49
20	881935000	899484999	36.55
21	899485000	919734999	35.15
22	919735000	939174999	33.74
23	939175000	958074999	32.80
24	958075000	976982019	30.93

outer diameter

less linear speed

less data rate

inner diameter

[26.1]

Seek Time

- Full stroke seek time (OD→ID or ID→OD)

Function	Typical (ms)	Max (ms)
Read	14.7	17.7
Write	15.7	18.7
Read (Quiet Seek mode)	32.5	35.5
Write (Quiet Seek mode)	33.5	36.5

- Average seek time

Sometimes you want a quiet HD: Reduces acoustic noise at the cost of seek time →

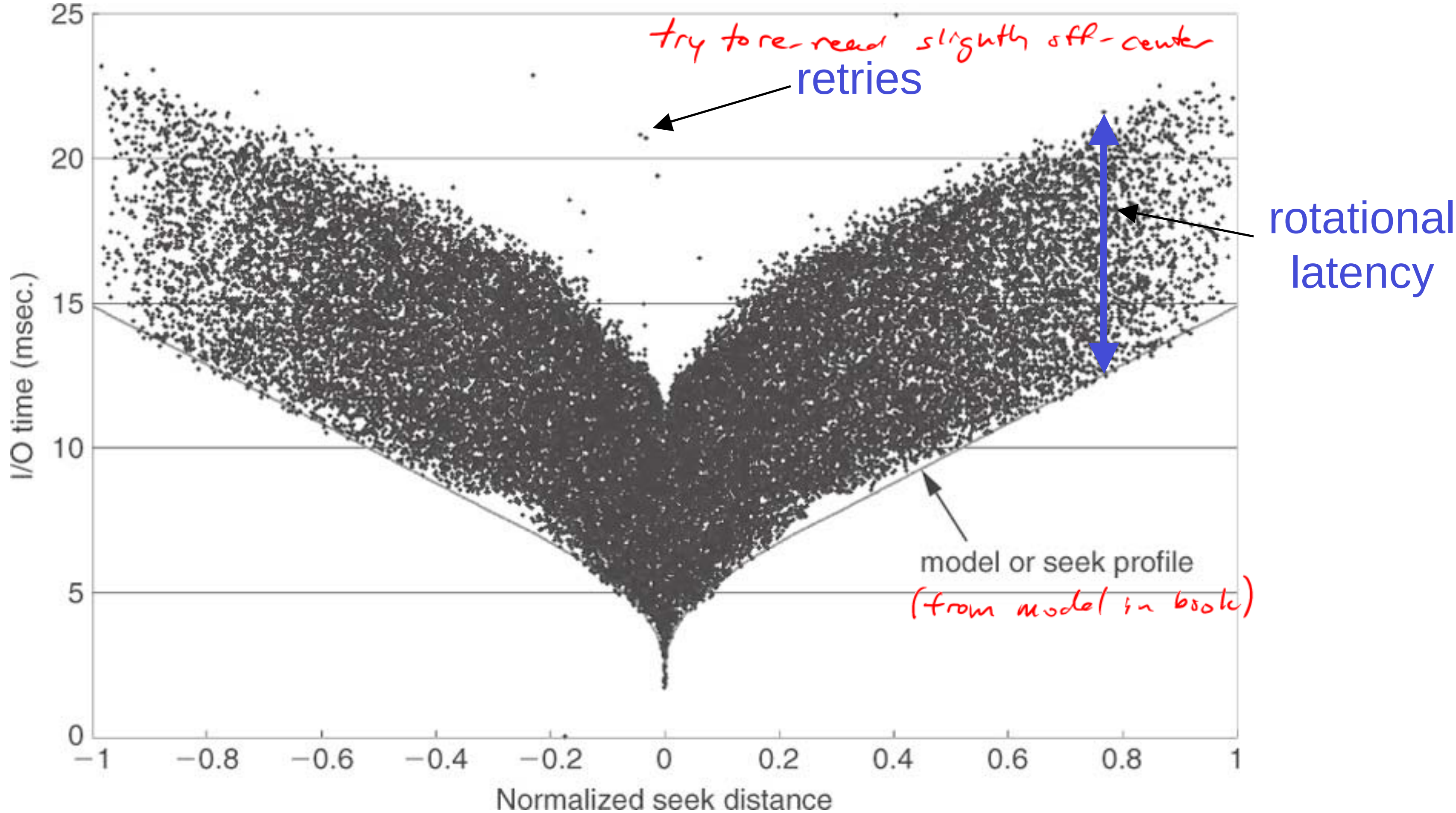
Command type	Typical (ms)	Max (ms)
Read	8.2	9.2
Write	9.2	10.2
Read (Quiet Seek mode)	19.5	20.5
Write (Quiet Seek mode)	20.5	21.5

[7K500v1.5]

Seek Time

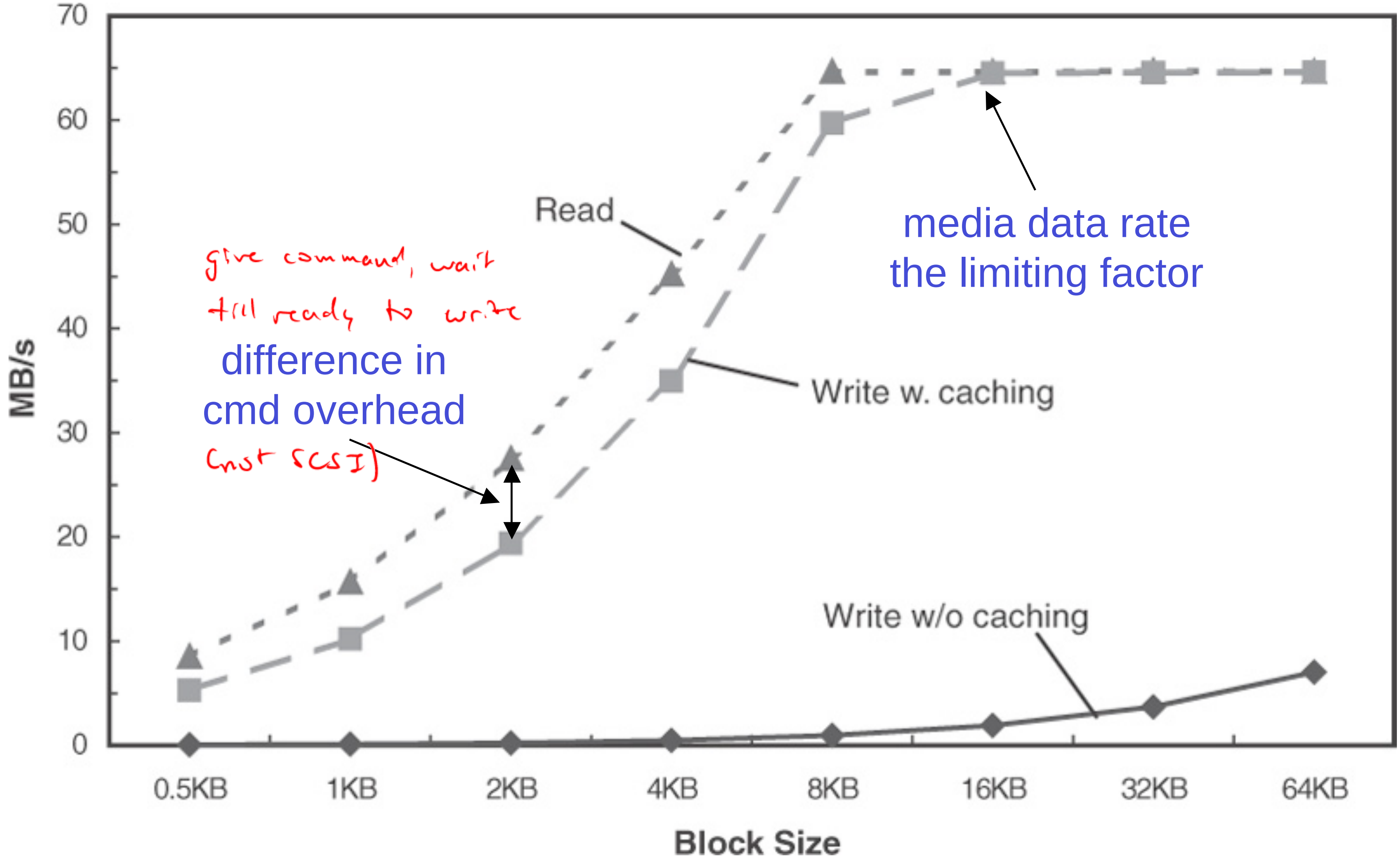
- **Whale tail plot**

seek distance = (current LBA - previous LBA)/max LBA



[26.3]

Sequential Access Performance



[26.7]

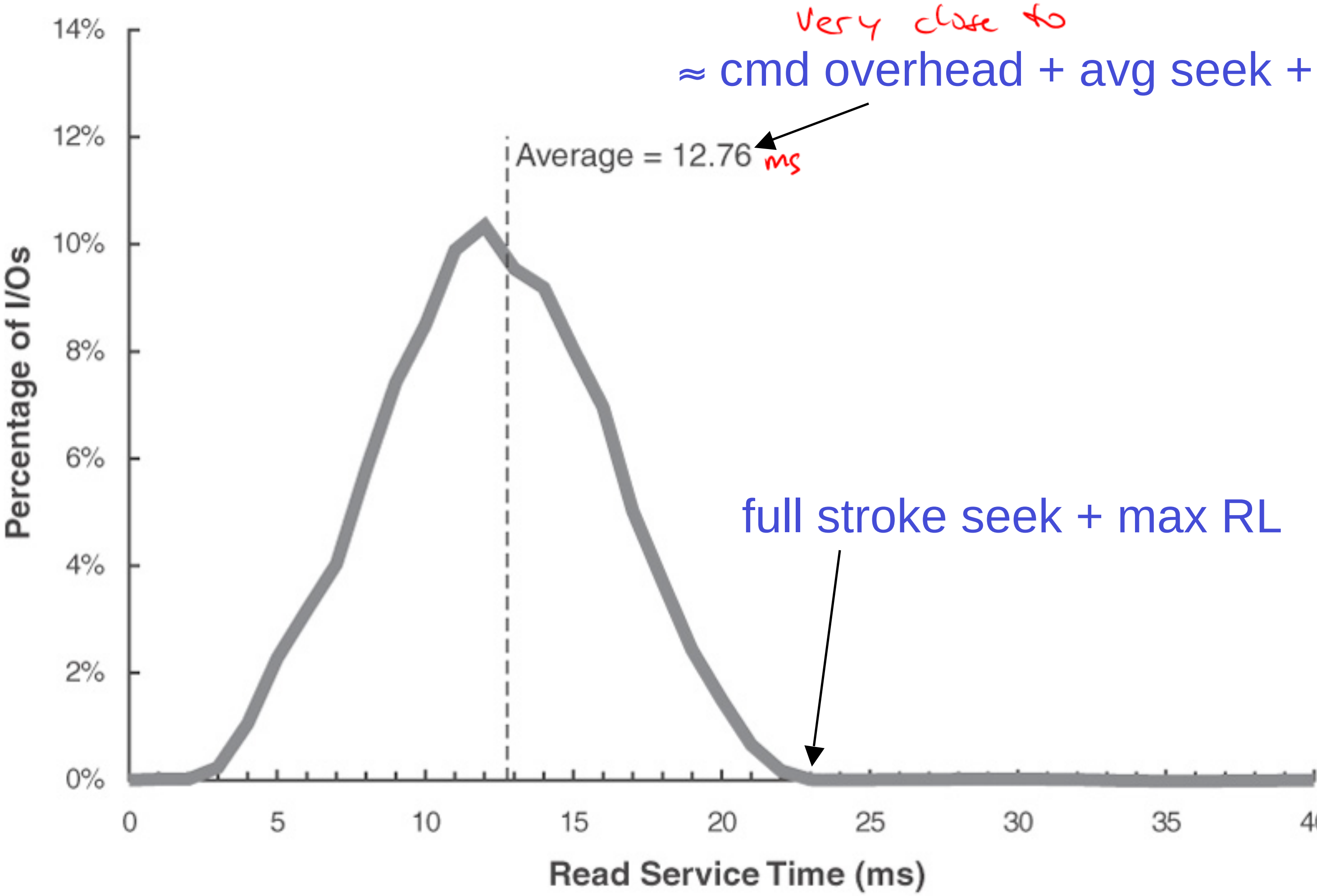
Random Read Performance

- **25,000 random requests**

histogram plot of read service time

very close to

≈ cmd overhead + avg seek + avg RL



[26.8]

Marketing Lingo: "Smooth Stream"

Streaming Feature Set

- In AV applications, drive need to supply data at a required rate → better to retain rate, allow occasional data errors, as humans won't notice small errors in pixels in video, for example
- Better to maintain constant rate and have a few bad pixels than perfect data with a long delay

→ tries to correct for ECC, but will retain rate and report errors.

- Streaming commands give constant rate for a given read or write stream and log any errors that occurred that could not be corrected
- Controller can read logs and access correct data if necessary

Hard Drive Power Management

- Hard drive can consume >20% of the total power in PC ~ 10's of Watts
- Disks can easily be the largest power component in a server
- Dynamic power management
  - Exploit disk low power modes while still delivering good throughout or response time

7K500 Low Power Operating Modes

- **Idle**
  - Spinning but not processing a command
- **Low RPM idle** ↙ quadratic relationship between rpm and power
  - Spinning at 4500 rpm but not processing a command
- **Unload idle**
  - Spinning with heads unloaded → not over platter surface
- **Standby**
  - ↙ 0 rpm
  - Heads unloaded, spindle motor stopped, commands can be received immediately
- **Sleep**
  - Heads unloaded, spindle motor stopped, reset required to move to standby

→ interface logic is powered down, waiting for reset
→ hard reset (pin)
→ soft reset (command)

Total

7K500 Average Power Consumption

~30-40W at full blast

active
- 13W normal random R/W seeks, 40% duty cycle
- 11W quiet random R/W seeks, 40% duty cycle
- 9W idle
- 6.8W unload idle
- 4.4W low RPM idle
- 1W standby
- 0.7W sleep

Transitioning Between Modes

- Time cost to move in/out low power modes

From	To	RPM	Transition time (sec)	
			Typical	Maximum
Standby	Idle	0 ---> 7200	15	31
Idle	Standby	7200 ---> 0	Cut Motor Power Immediately	Immediately
Standby	Sleep	0	Immediately	Immediately
Sleep	Standby	0	+ Reset Command Overhead Immediately	Immediately
Unload idle	Idle	7200	1	31
Idle	Unload idle	7200	0.7	31
Low ~~RMP~~ RPM Idle	Idle	4500->7200	7	31

[7K500v1.5]

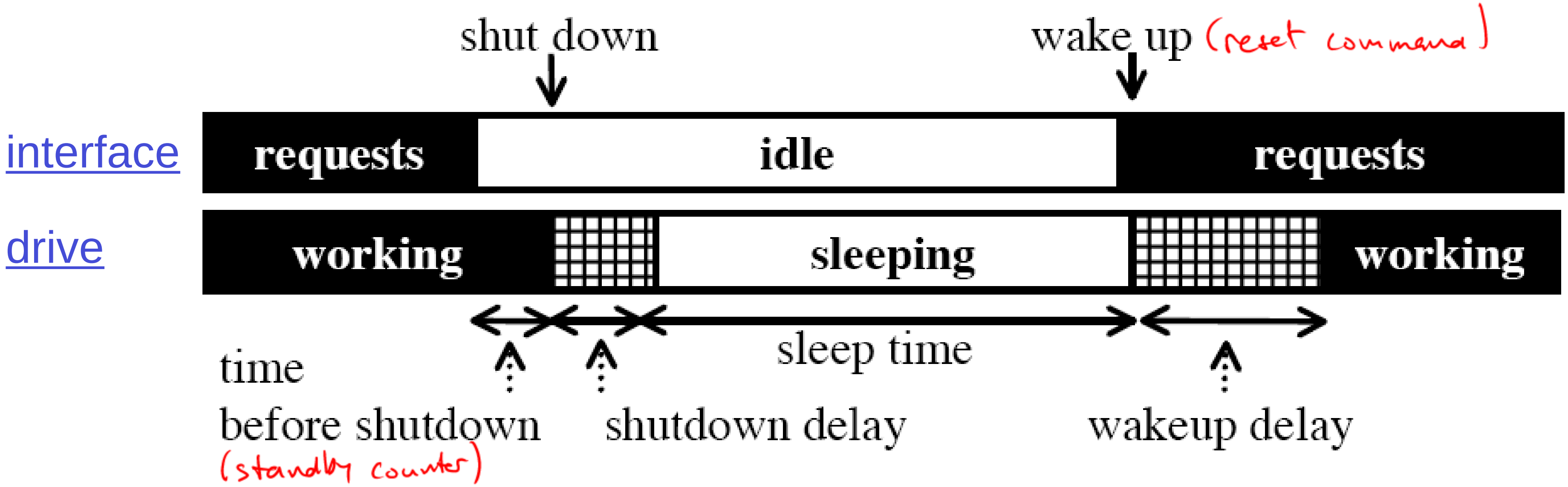
Enabling Low Power Modes

- Power management commands
  - Idle: moves to idle mode; optionally sets standby timer and starts standby count down (before switching to standby)
  - Standby: moves to standby mode; optionally sets standby timer and starts standby count down (before mode switch)
  - Sleep: moves to sleep mode
  - Reset: required to exit sleep mode and enter standby

- Standby timer → idle time counter
  - Counts down every consecutive cycle that no cmd is received
  - Drive enters standby mode when count = 0
  - Timer is reinitialized if a cmd is received

Idle Time Management

- **Enter low power mode when idle for long time**



- ***Minimum sleeping time*: Minimum sleep time to compensate for shutdown and wakeup overhead**
- ***Break even time*: Minimum idle time required to save energy** (to amortize costs of wakeup, etc)

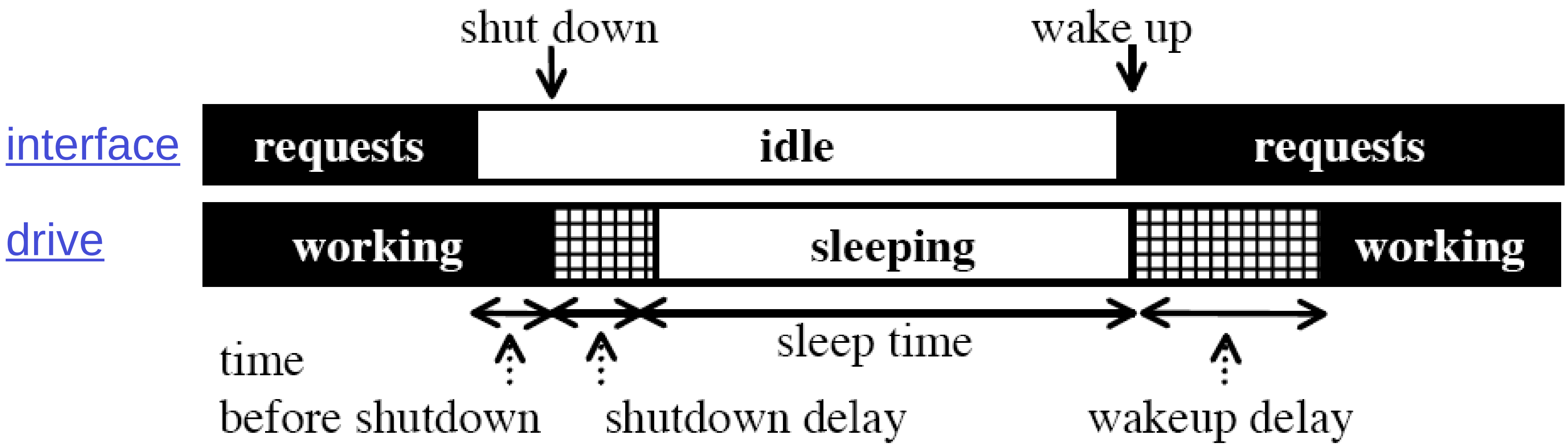
[Lu00]

Terminology

Symbol	Meaning
T_{sd}	shutdown delay
T_{wu}	wakeup delay
T_{be}	break-even time for an idle period
T_{ms}	minimum sleeping time to save energy
T_{bs}	time before shutdown
T_{ss}	average time in sleeping state
$T_{idle}[i]$	current idle period, candidate for shutdown
$t_{idle}[i]$	predicted value of $T_{idle}[i]$
$T_{busy}[i]$	busy period before $T_{idle}[i]$
$W_S[i]$	starting time of a waiting period
$W_E[i]$	ending time of a waiting period
τ	timeout value
E_{sd}	energy to shut down
E_{wu}	energy to wake up
P_s	power in sleeping state
P_w	power in working state
N_{sd}	number of shutdowns
N_{wd}	number of wrong shutdowns

[Lu00]

Minimum Sleeping Time Calculation

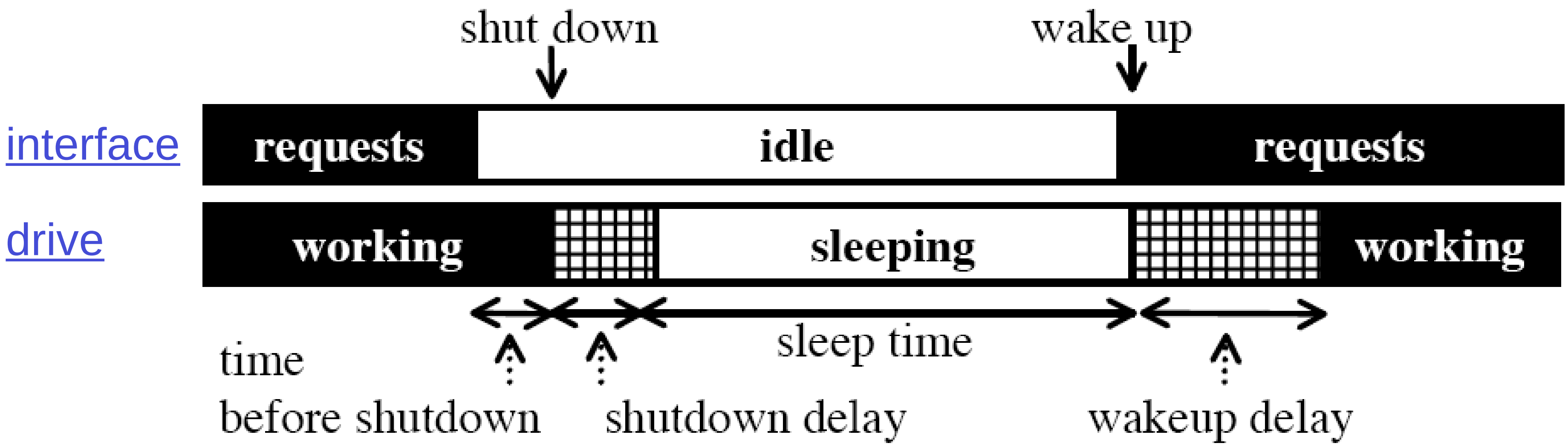


$$E_{sd} + E_{wu} + P_s \times T_{ms} = P_w \times (T_{ms} + T_{sd} + T_{wu})$$

$$\Rightarrow \quad T_{ms} = \frac{E_{sd} + E_{wu} - P_w \times (T_{sd} + T_{wu})}{P_w - P_s}$$

[Lu00]

Break Even Time Calculation



$$T_{be} = T_{ms} + T_{sd} + T_{wu}$$

$$= \frac{E_{sd} + E_{wu} - P_s \times (T_{sd} + T_{wu})}{P_w - P_s}$$

$$T_{idle}[i] > T_{bs} + T_{be}$$

[Lu00]

Performance Metrics

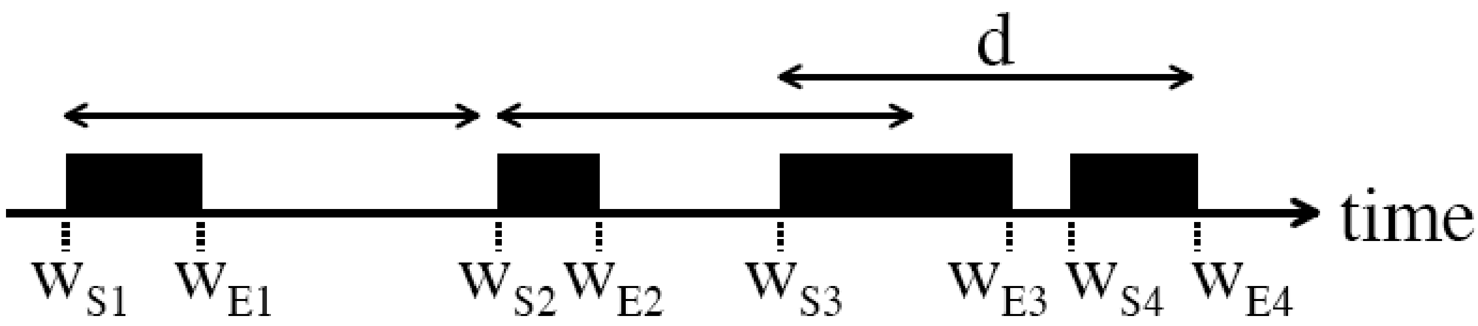
- Total or average wait time misleading
- Users bothered by high fraction of wait time over a short period than many short wait bursts
- Example
  - 60s total wait time in 10 minutes, but longest wait time in 1 minute is 6s (worse in total wait time)
  - 50s total wait time in 10 minutes, but longest wait time in 1 minute is 30s (worse to the user!)

Performance Metric #1

- **Largest total wait time in a duration of time d**

$$W_d = \max_t \sum_{\substack{i \text{ such that} \\ W_S[i] \geq t \\ W_E[i] \leq t+d}} (W_E[i] - W_S[i])$$

  - **Example**



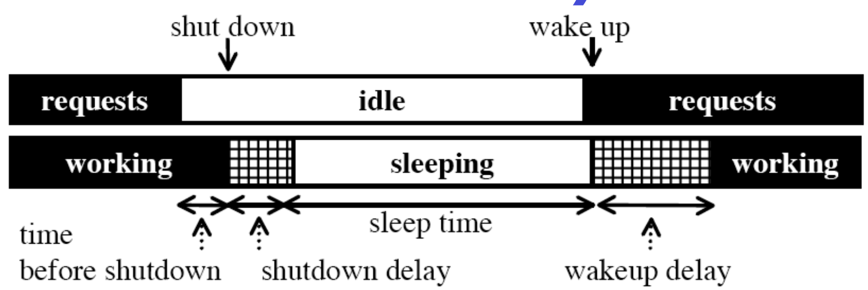
[Lu00]

Performance Metric #2

- Longest wait time sequence where the time between each wait period < a threshold
  - Long sequence of repetitive bursts of wait time

Adaptive Timeout Algorithms

- $T_{bs} = \tau$ (shutdown after idle for τ seconds)

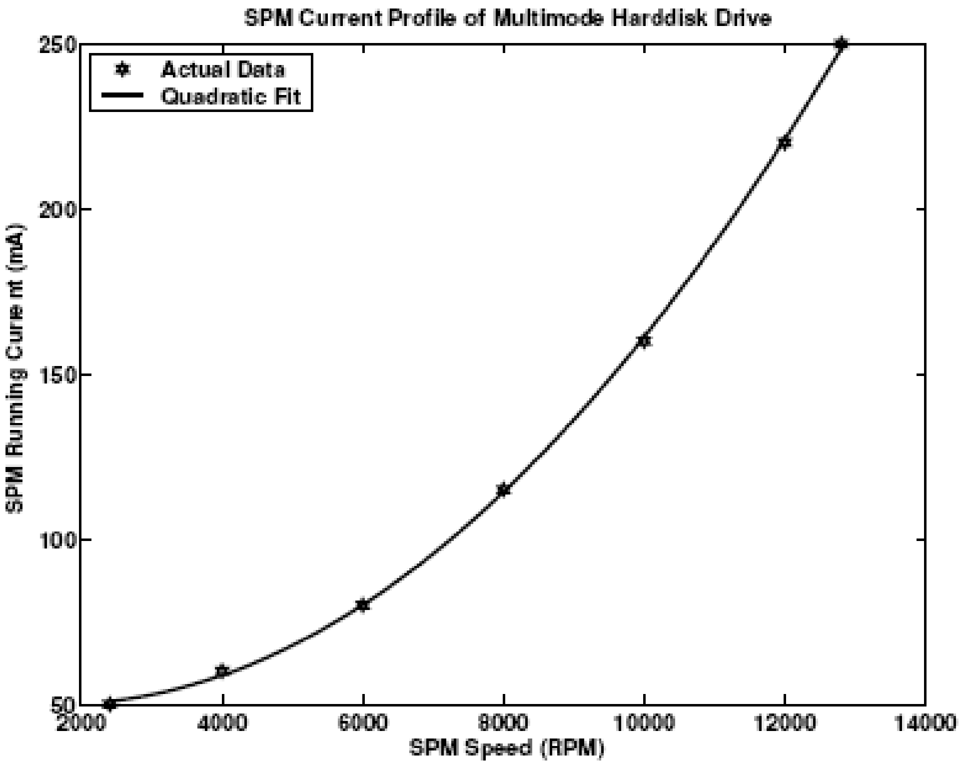


- Change τ dynamically
  - Example 1: $T_{idle}[i-1]/T_{wu}$
    - When small, increase τ; when large, decrease τ
    - May be done asymmetrically
  - Example 2: $T_{busy}[i]$
    - When small, decrease τ; when large, increase τ
    - how busy was I?
- Many other possibilities
  - Detect short burst followed by long idle
  - Exponential weighted average of previous idle periods
  - FSM

[Lu00]

What if the Disk is Busy?

- In server or streaming media environments, disk may not be put to sleep very often
- Spindle motor power
  - Can account for 80% of the total idle power
  - May increase exponentially with rotational speed



[Gurumurthi03]

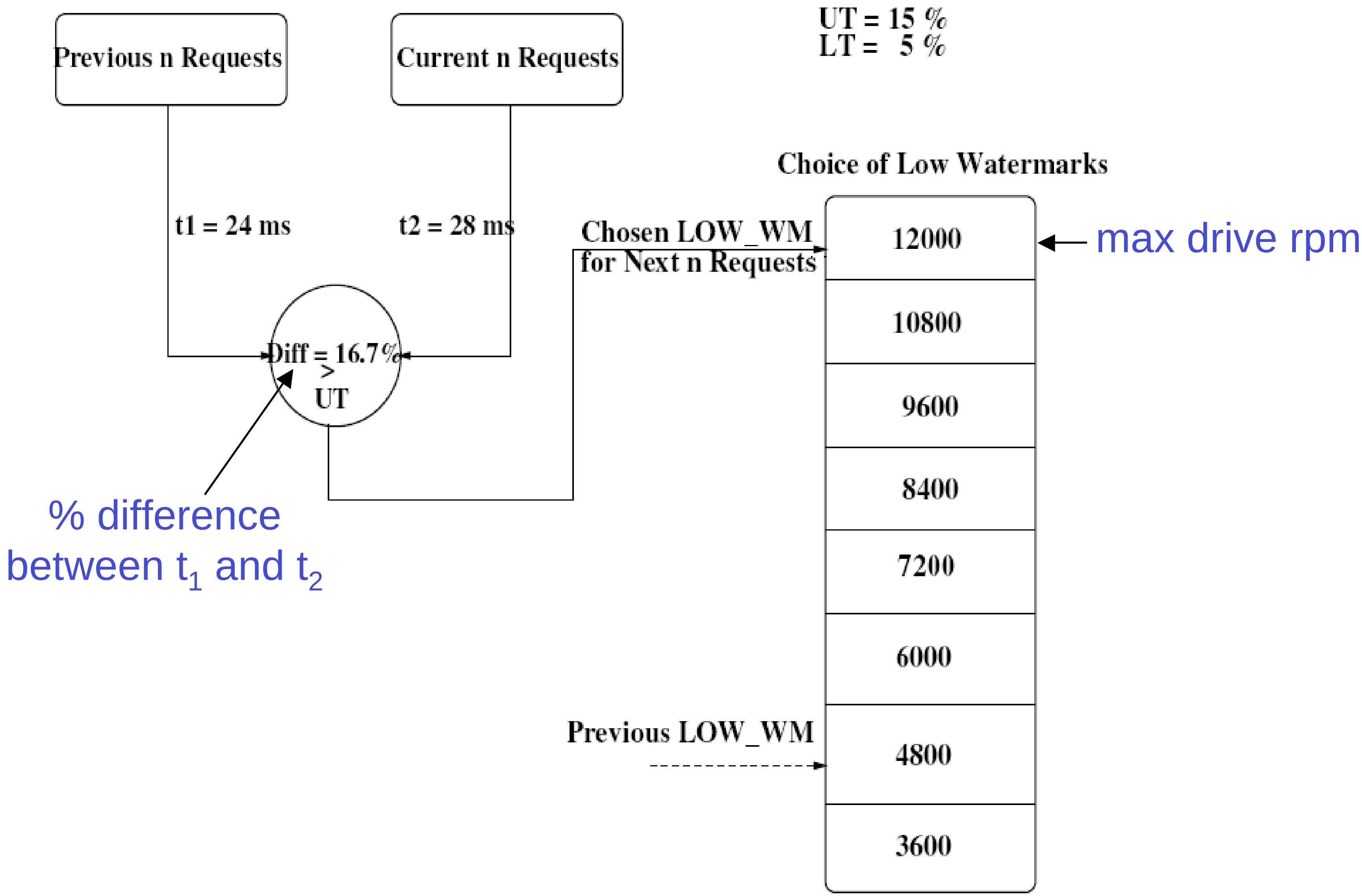
Dynamic Rotations Per Minute (DRPM)

- Dynamically match the disk rpm to user demand
- Analogous to how Intel SpeedStep and AMD PowerNow! match processor clock speed and voltage to application demand
- Can roughly equate the number of queued requests with the user load
- Assumes a RAID configuration with local disk controllers (DCs) and one array controller (AC)

DRPM Algorithm

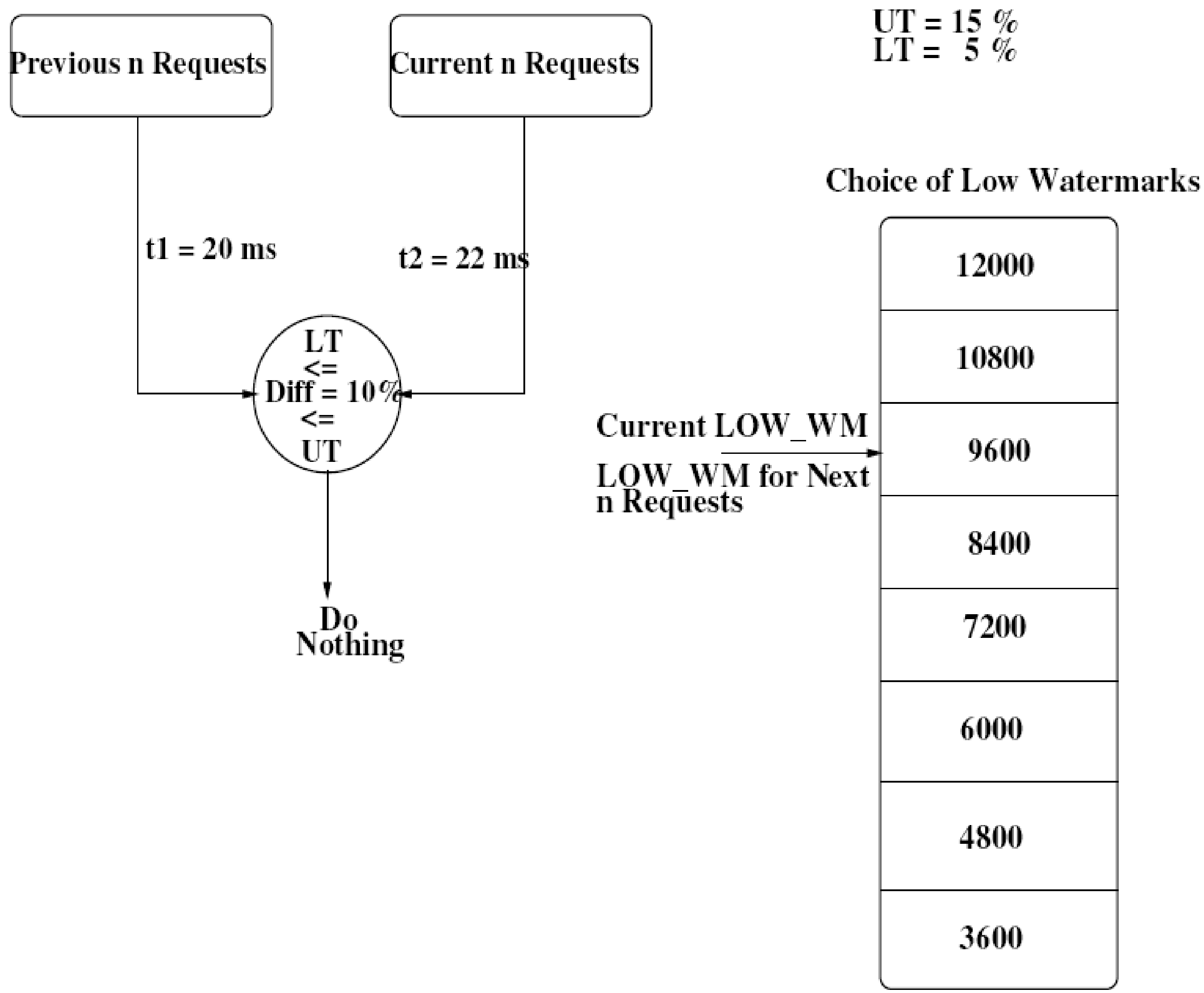
- local disk controller
- Each DC checks if its request queue occupancy is less than a minimum value
- If so, it reduces its rpm one step, *unless* it is already at its rpm *low water mark* (lowest rpm allowed)
- array controller
- The AC tracks the average response time over each window of n requests
- Depending on the change in response time over the last two windows, the AC will
  - (1) Force all disks to their maximum rpm if bad response time
  - (2) Make no changes if no change
  - (3) Lower the rpm low water mark for all disks if good response time

DRPM Scenario (1)



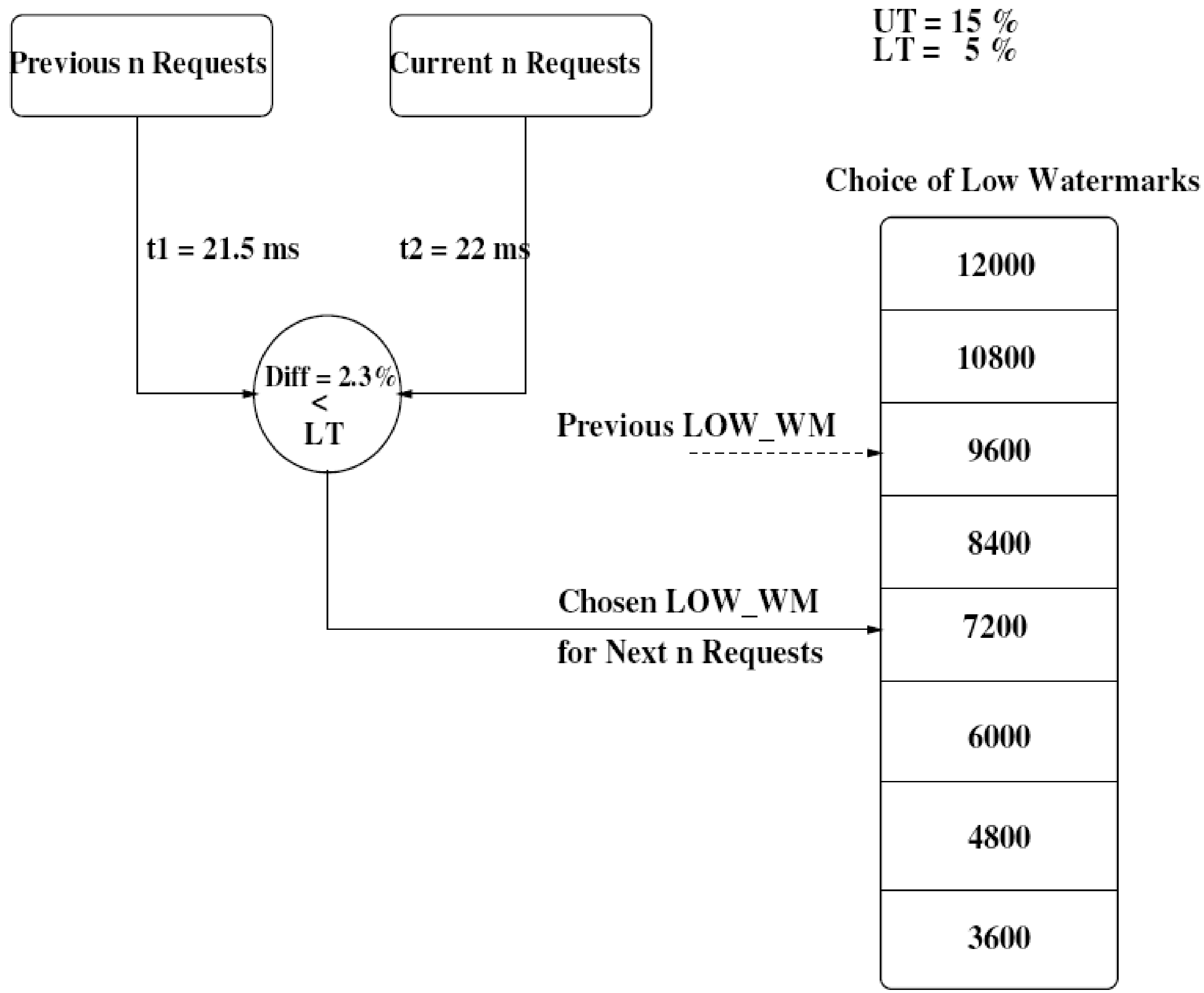
[Gurumurthi03]

DRPM Scenario (2)



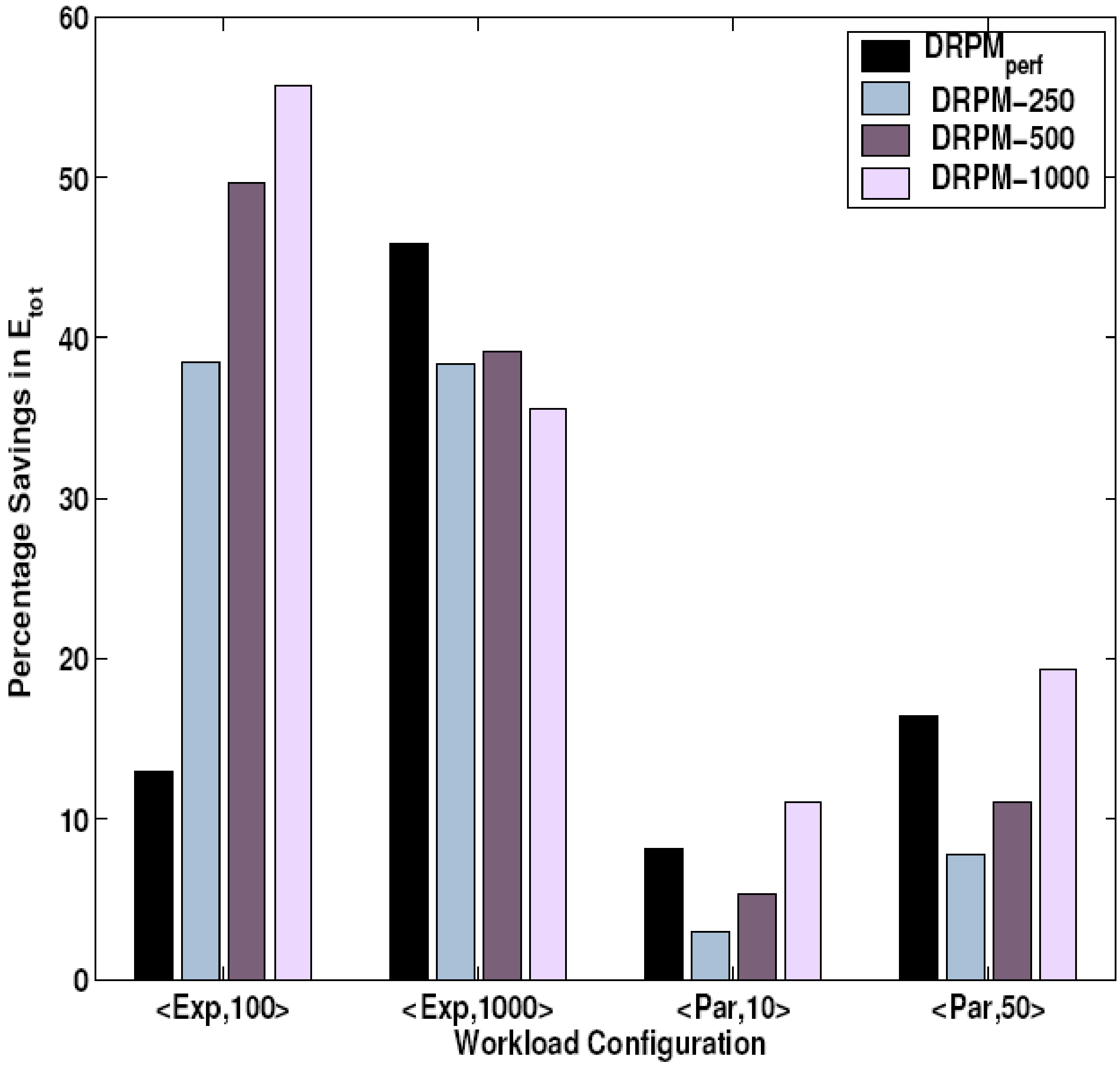
[Gurumurthi03]

DRPM Scenario (3)



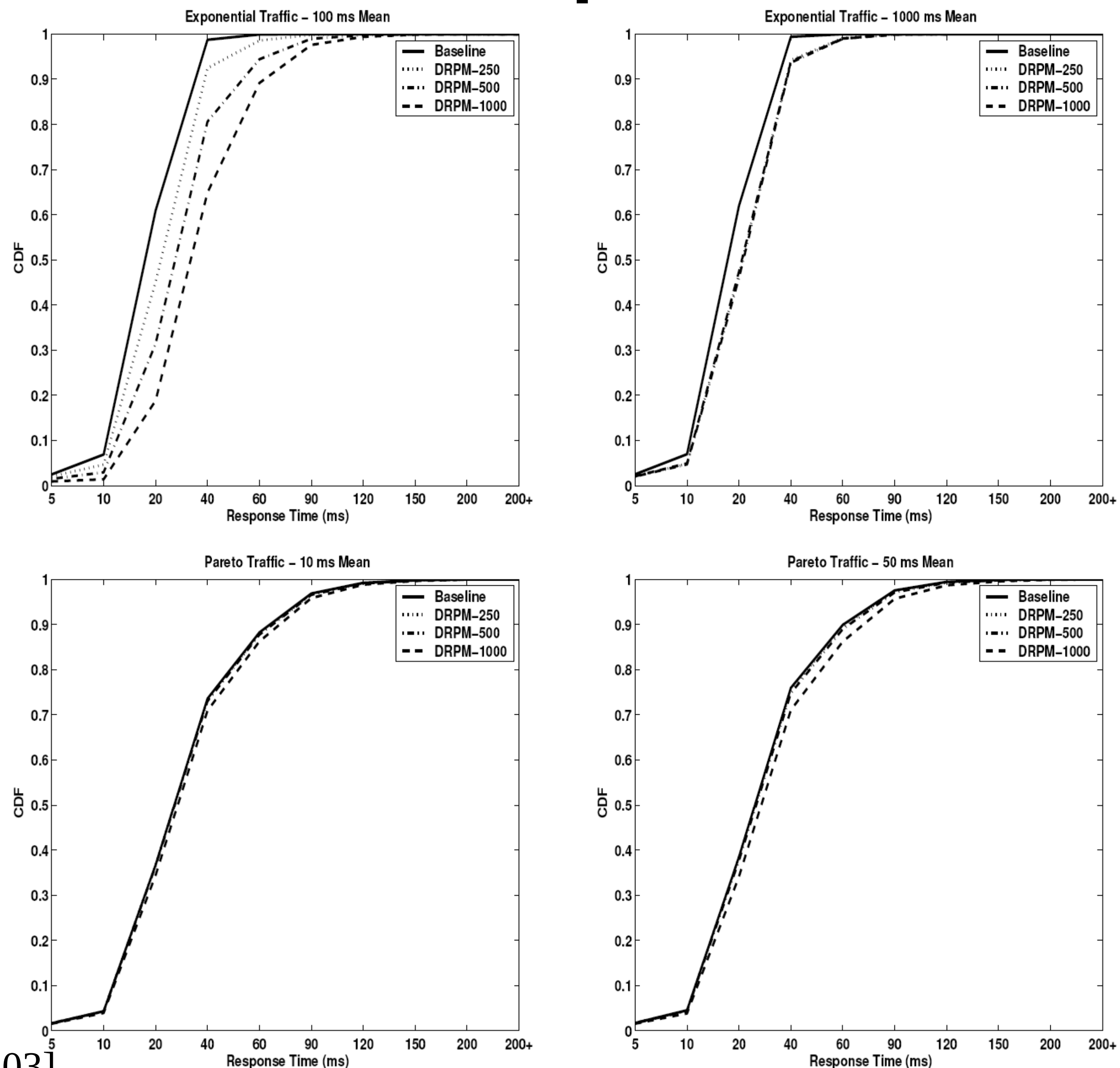
[Gurumurthi03]

DRPM Energy Savings



[Gurumurthi03]

DRPM Response Time



[Gurumurthi03]

Intradisk Parallelism

- **Should we return to multiple actuators?**

Disk Drive Characteristics	Disks From SIGMOD'88 RAID Paper [26]			Modern Disk Drive Technology	
	IBM 3380 AK4	Fujitsu M2361A	Conners CP3100	Seagate Barracuda ES	**Projection for 4-Actuator Intra-Disk Parallel Drive**
Areal Density (Mb/in^2)	12			128000	
Disk Diameter (inches)	14	10.5	3.5	3.7	3.7
Formatted Data Capacity (MB)	7,500	600	100	750,000	750,000
No. Actuators	4	1	1	1	**4**
Power/box (Watts)	6,600	640	10	13	**34**
Transfer Rate (MB/s)	3	2.5	1	72	***Explored in Section 7***
Price/MB (including controller)	\$18-\$10	\$20-\$17	\$10-\$7	\$0.00042-\$0.00034	***Explored in Section 9***

server

motivation
for RAID

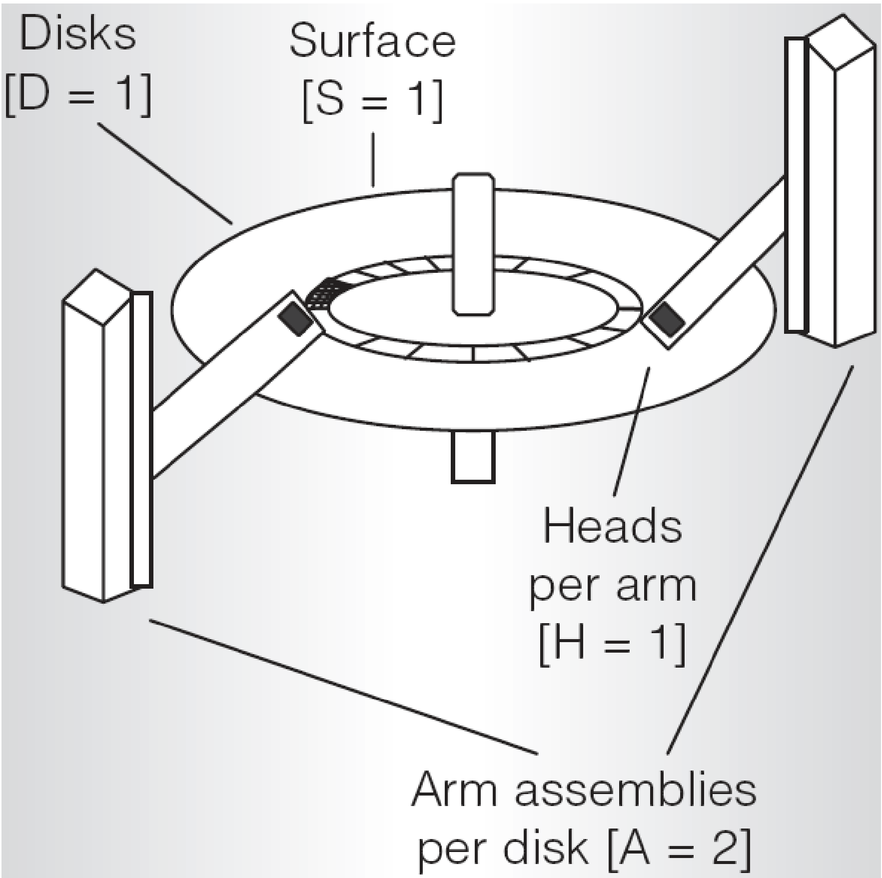
client

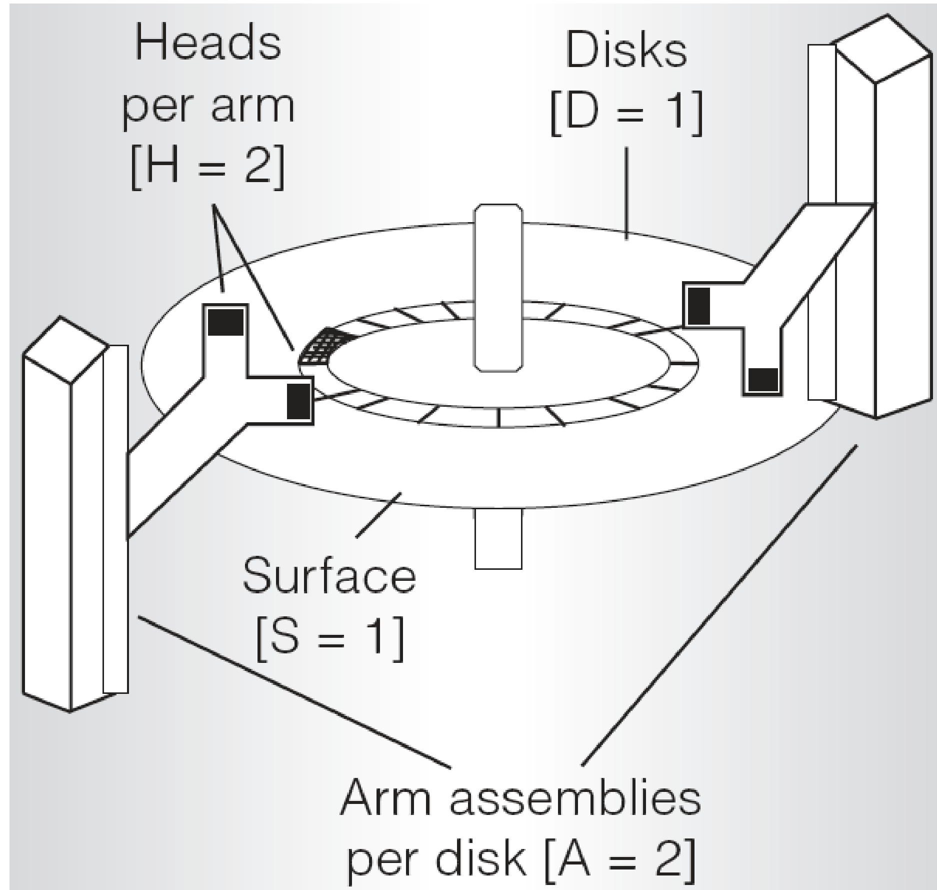
times
have
changed

the
future?

[Sankar08]

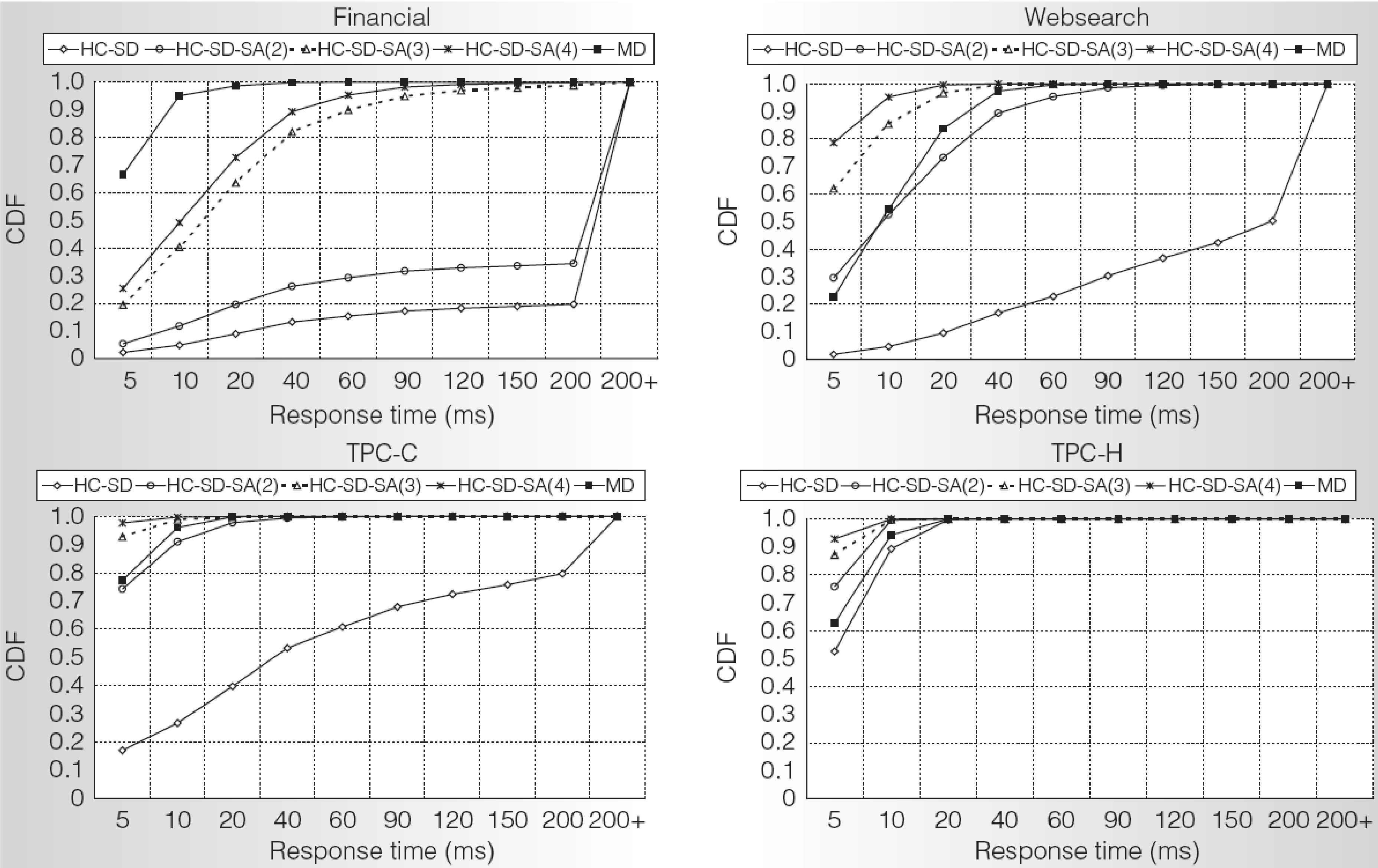
Multi-Actuator Drives





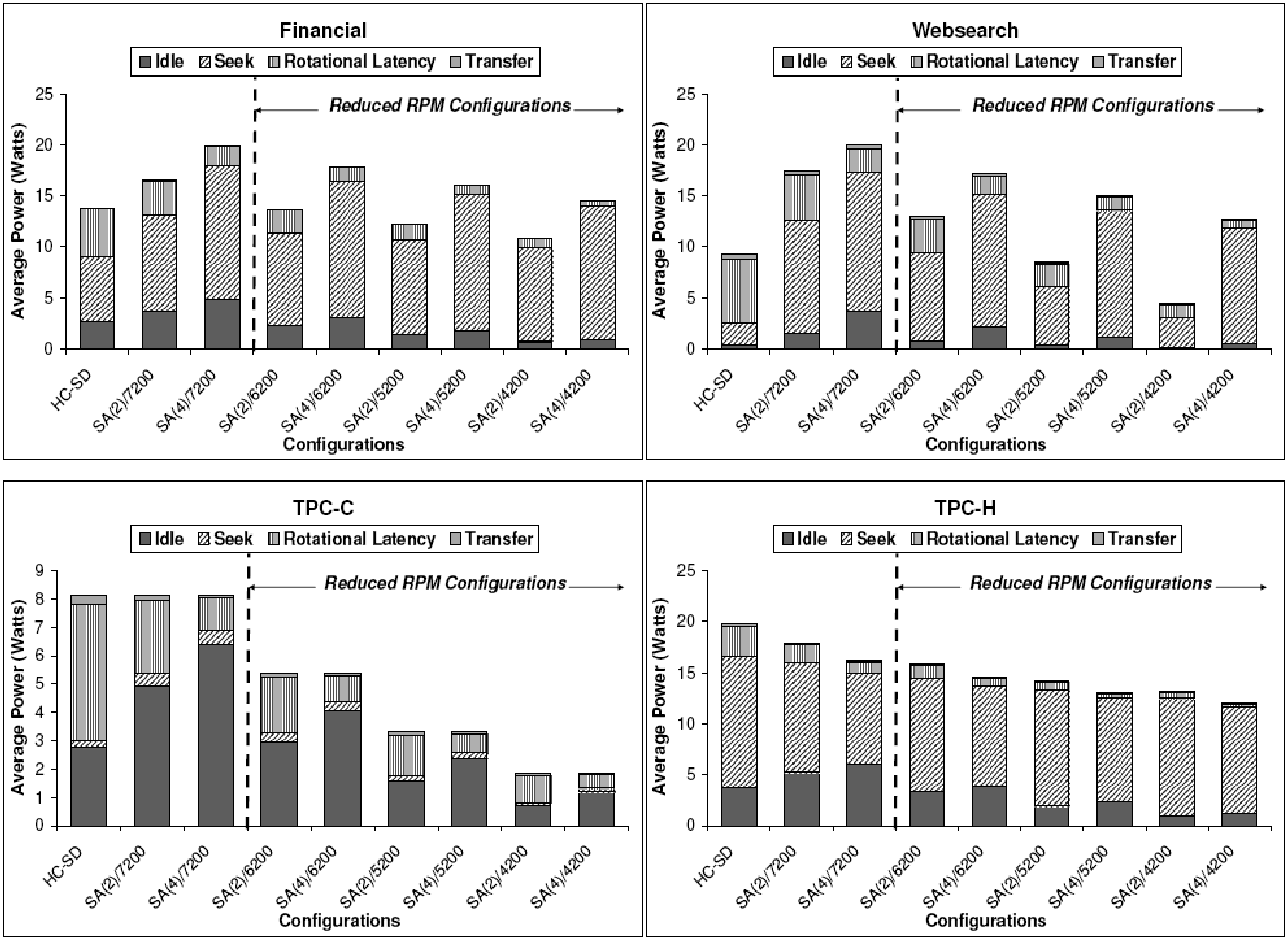
[Gurumurthi09]

Response Time Comparison



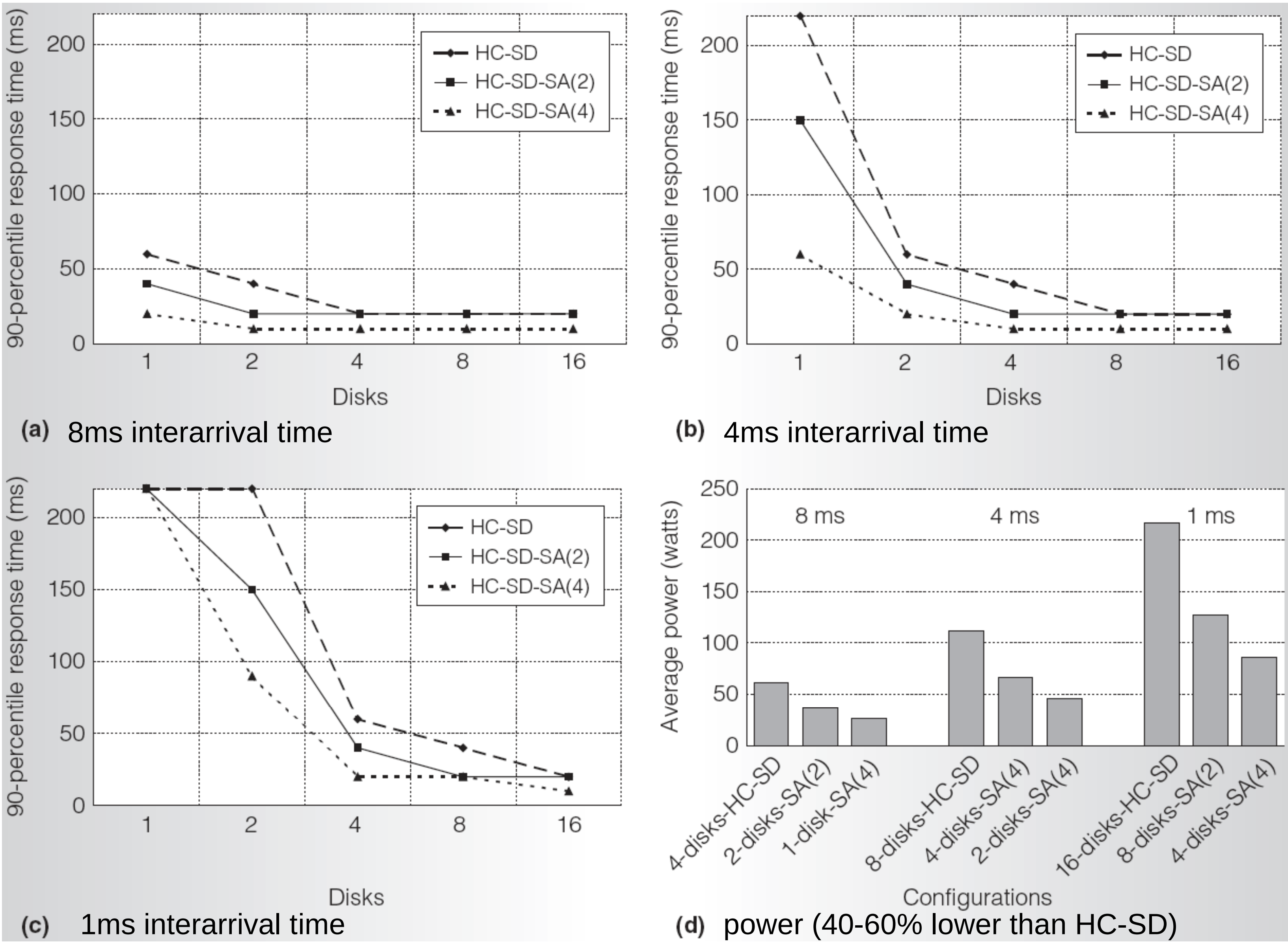
[Gurumurthi09]

Power Compared to HC-SD



[Sankar08]

A Better Way to Build RAID Systems?



[Gurumurthi09]

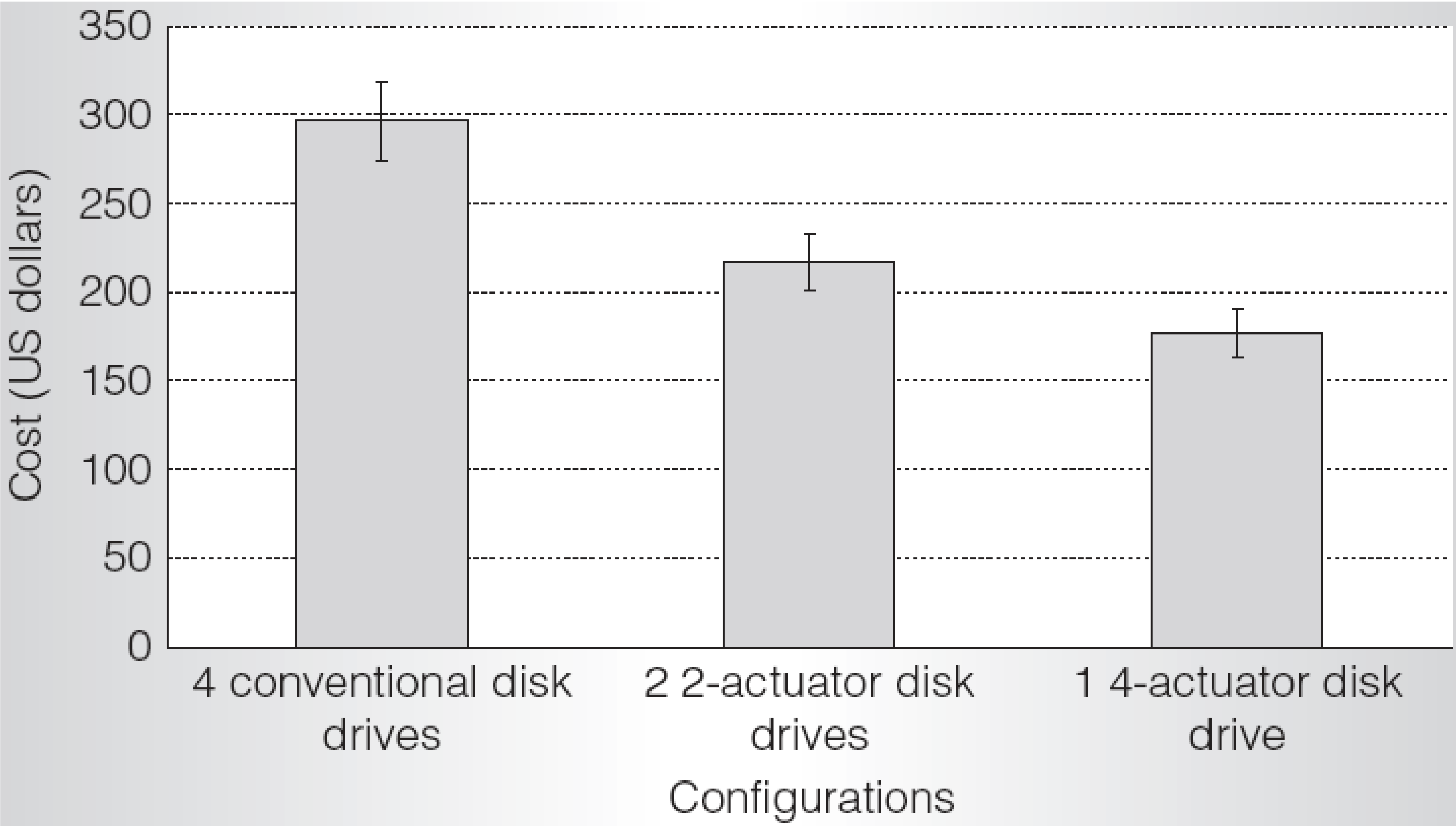
Cost Comparison

Table 2. Estimated component and disk-drive costs (in US dollars).

Component name	Component	Conventional disk drive	Two-actuator disk drive	Four-actuator disk drive
Media	6–7	24–28	24–28	24–28
Spindle motor	5–10	5–10	5–10	5–10
Voice-coil motor	1–2	1–2	2–4	4–8
Head suspension	0.50–0.90	2–3.6	4–7.2	8–14.4
Head	3	24	48	96
Pivot bearing	3	3	6	12
Disk controller	4–5	4–5	4–5	4–5
Motor driver	3.5–4	3.5–4	5–6	8–10
Preamplifier	1.2	1.2	2.4	4.8
Total estimated cost		67.7–80.8	100.4–116.6	165.8–188.2

[Gurumurthi09]

Cost Comparison



[Gurumurthi09]

Next Time

Exam II