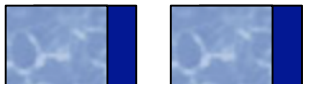
Energy Conservation Techniques for Array-Based Storage Servers

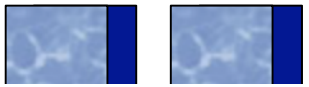
Eduardo Pinheiro & Ricardo Bianchini
Rutgers University
DarkLab
http://darklab.rutgers.edu

Vice-versa Meeting, Princeton Univ. 16/Apr/2004



Motivation

- Energy/power conservation important for mobile devices:
  - Power dissipation => casing, heat dissipation.
  - Energy => battery life.
- Also important for servers
  - Power planning.
  - Electricity bill.
  - Heat dissipation in highly-packed machine rooms.
- In this work: focus on array of disk drives on storage servers.

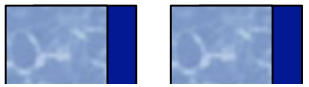


Goal

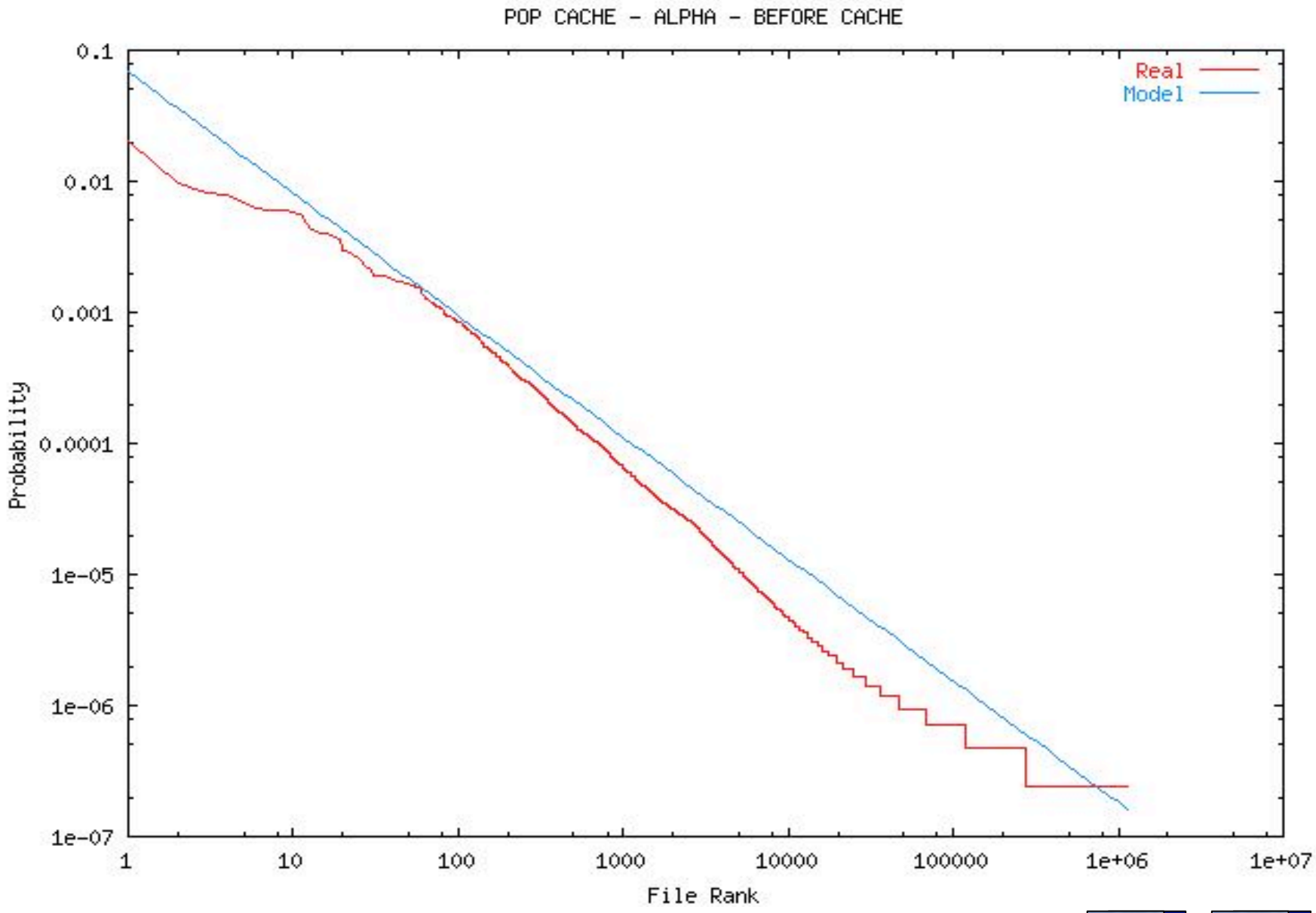
- Energy-Aware File Server.
- Rank popular files and cluster them in subset of disks.
- Keep HOT disks at high throughput/high power consumption.
- Send COLD disks to low throughput/low power consumption.
- Save energy overall.
- Periodic migration.

BUT

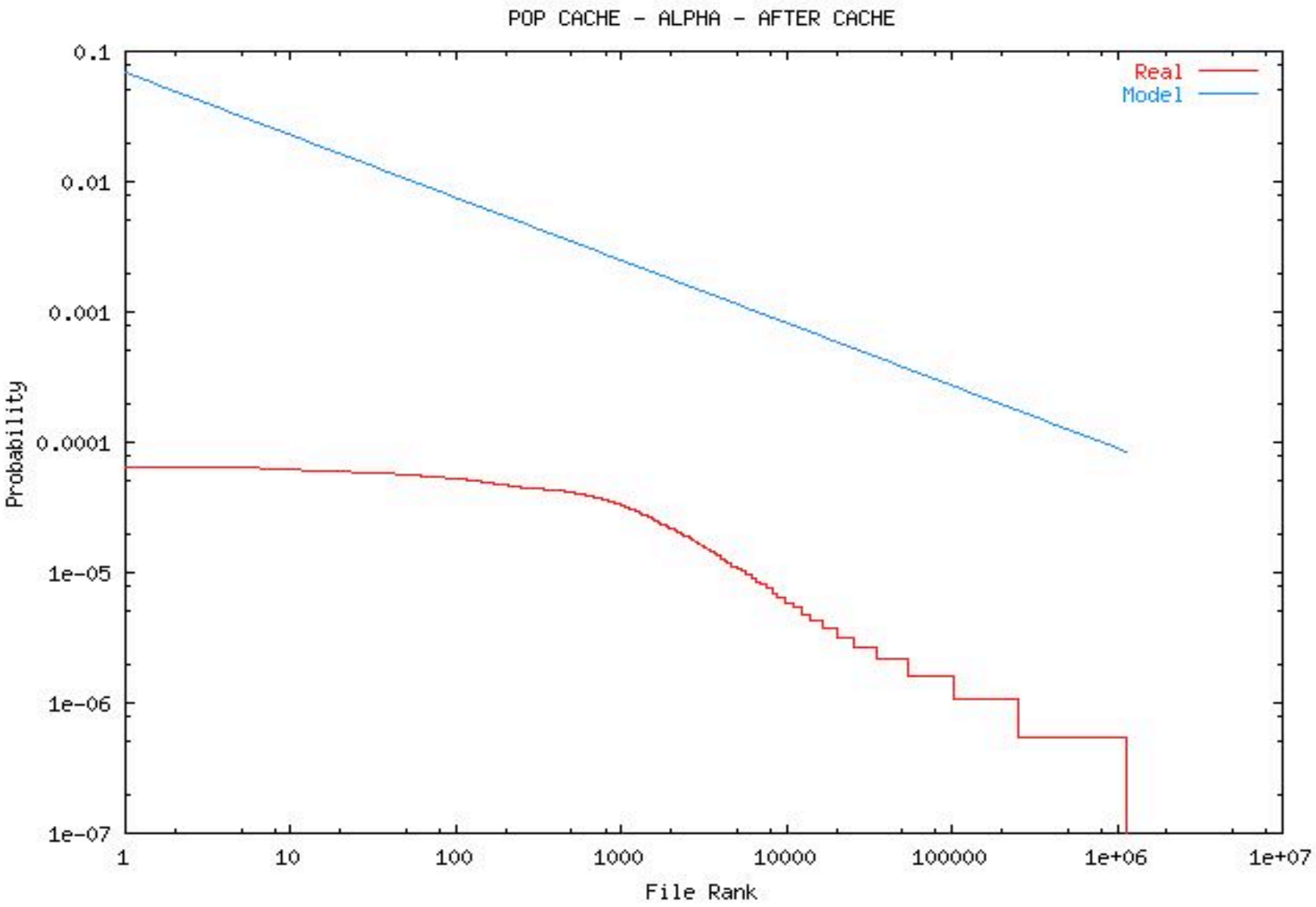
- Is there popularity in file server workloads? Zipf?



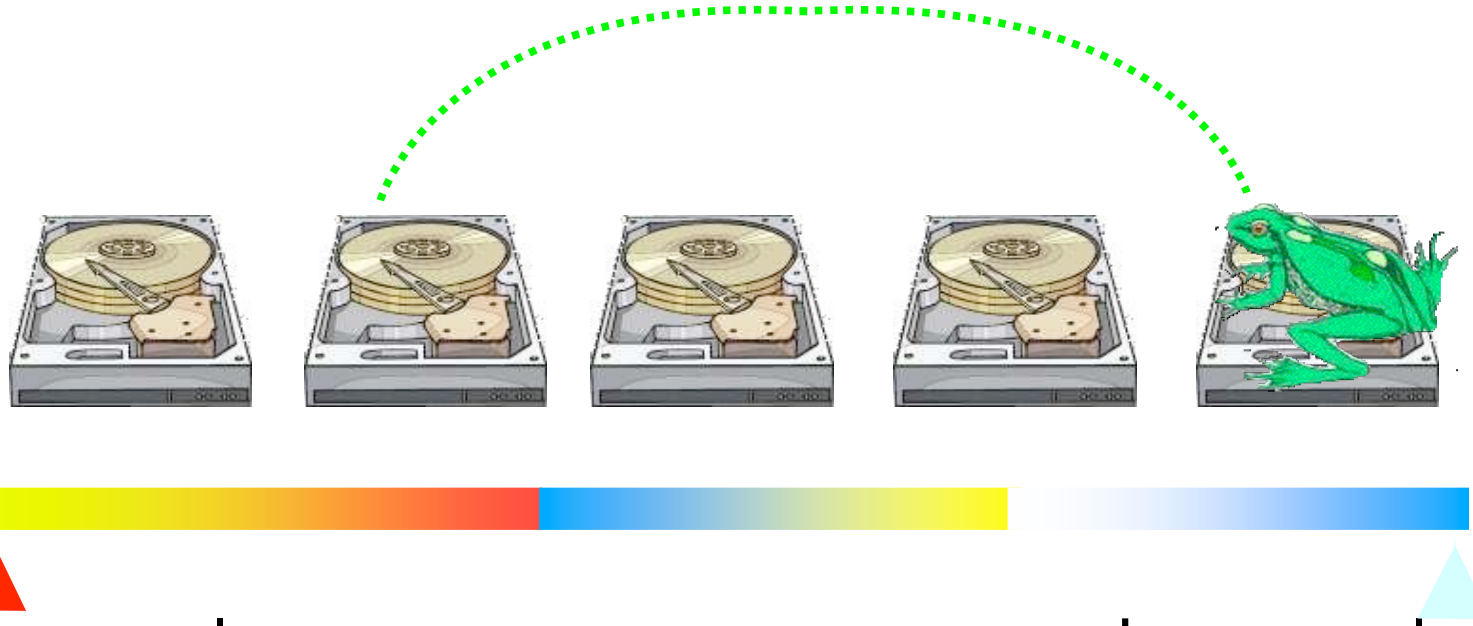
Key Observation



Key Observation



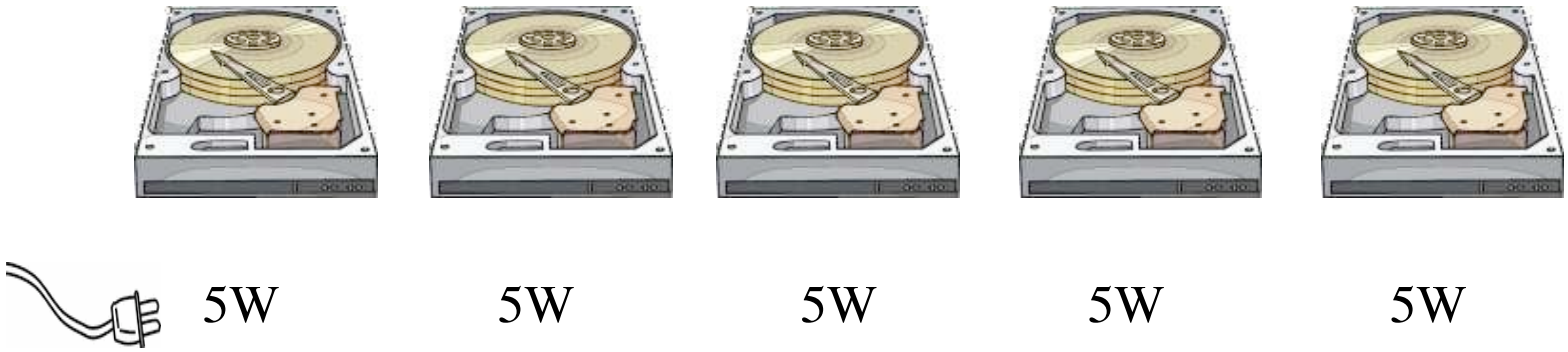
Popular Data Concentration



Periodically, rank data and migrate

- *Hot* ones one way (higher power modes)
- *Cold* ones the other way (lower power modes)

Illustration

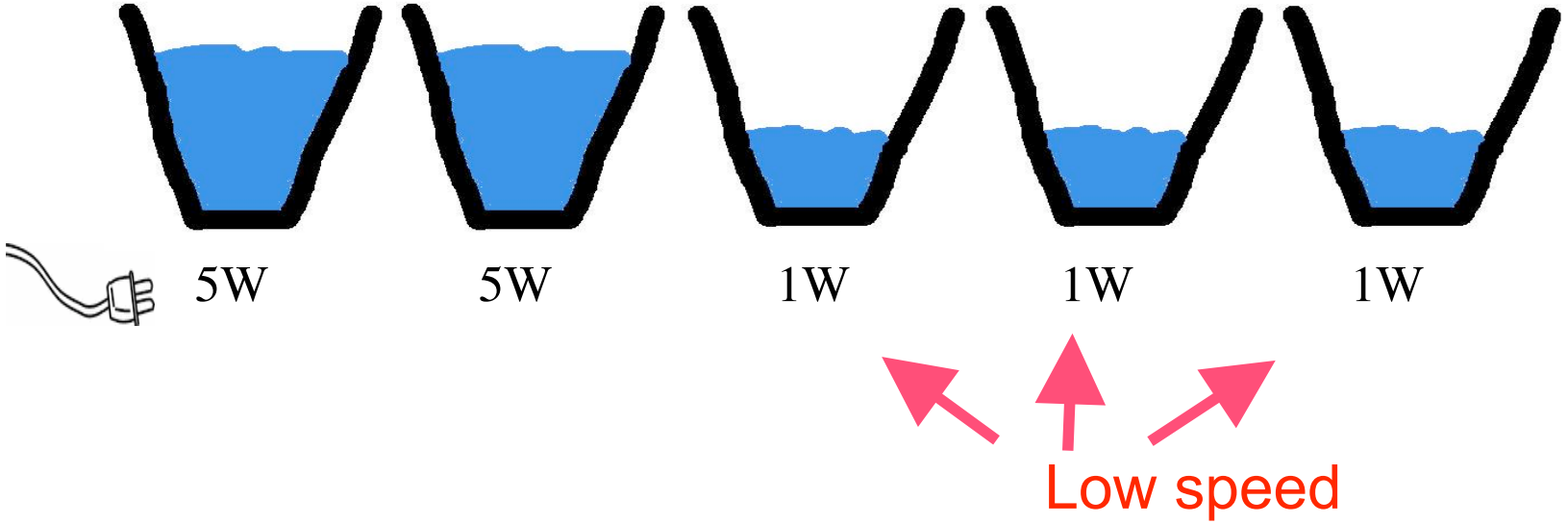


- Assume two-speed disks.
- Low speed allows accesses at higher latency, lower power.
- At low speed: 1 W

Illustration

Total Consumption: 25 W

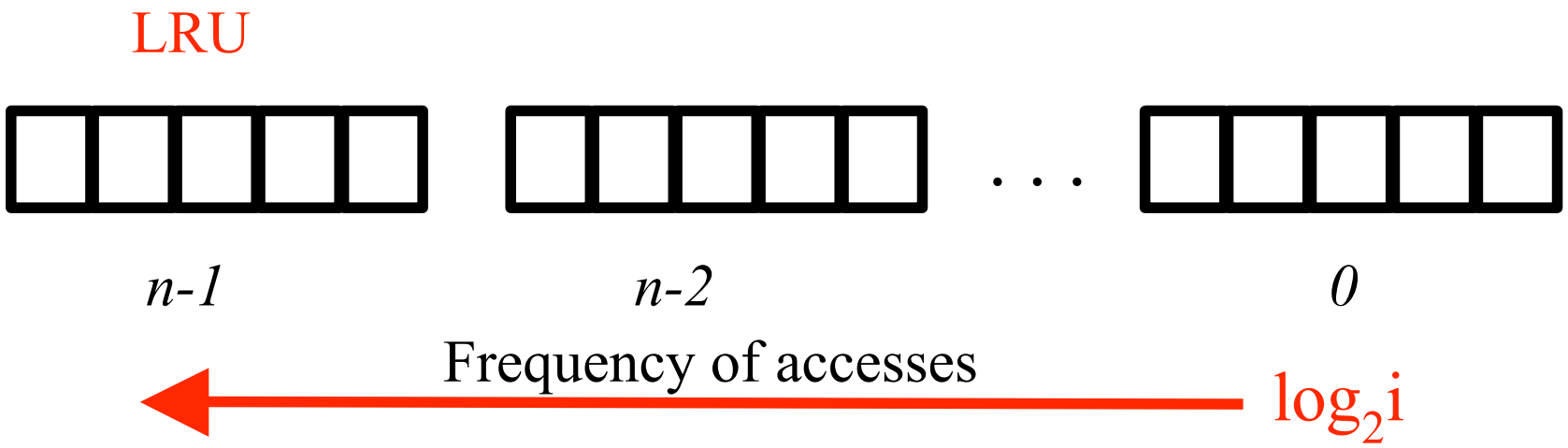
Illustration



Total Consumption: 13 W

MQ Cache -Design

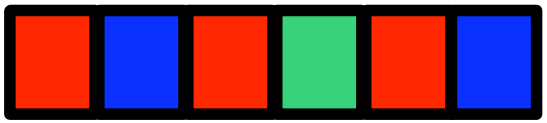
- MQ - second-level cache [Y.Zhou et al, USENIX'01]



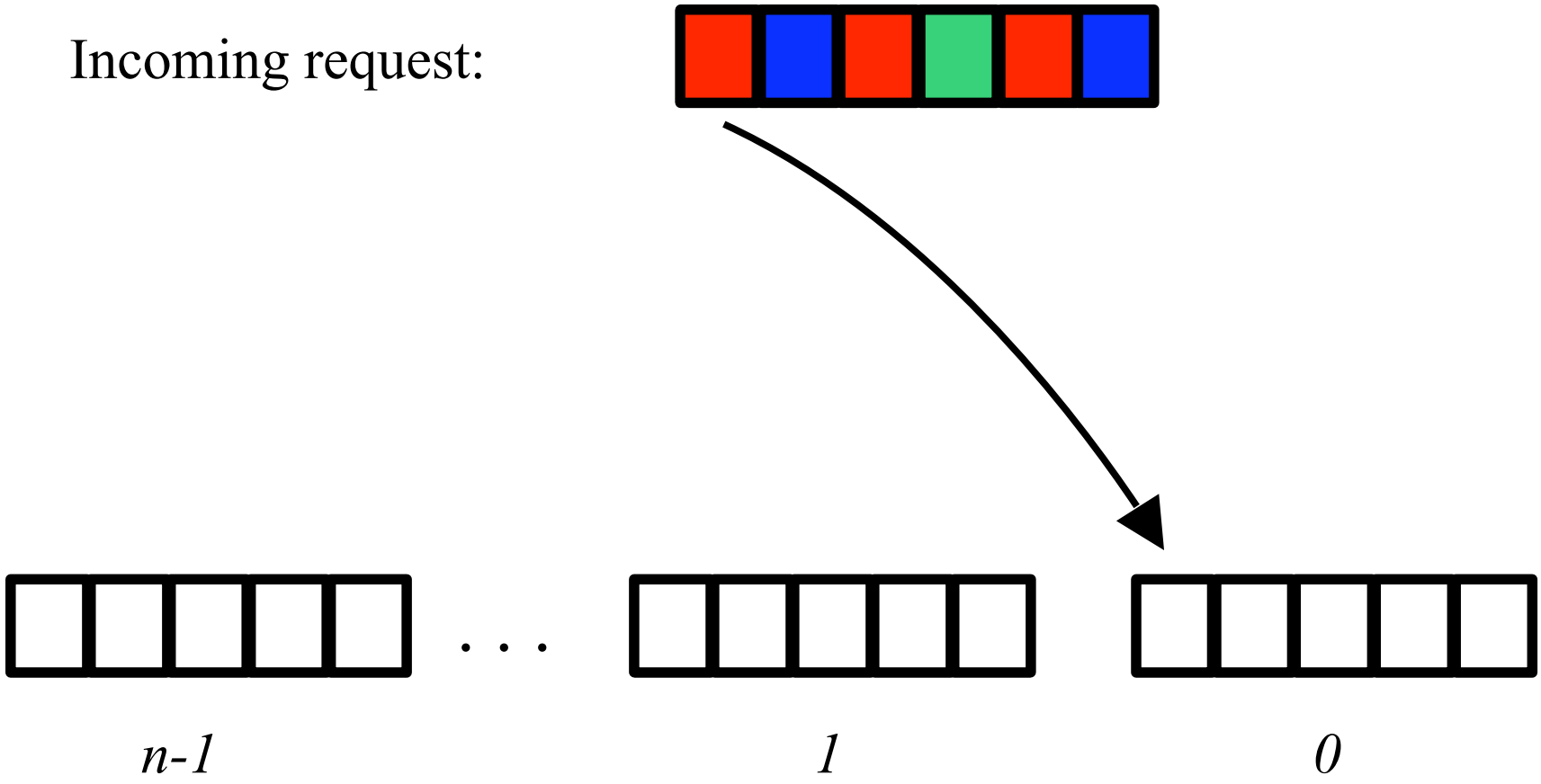
- LRU inside each queue
- Frequent accesses move entry up
- Prolonged inactivity moves entry down
- Re-use buffer for evicted entries

MQ Cache -Example

Incoming request:

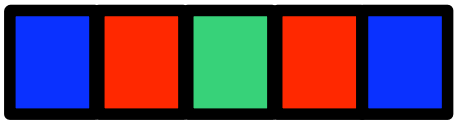


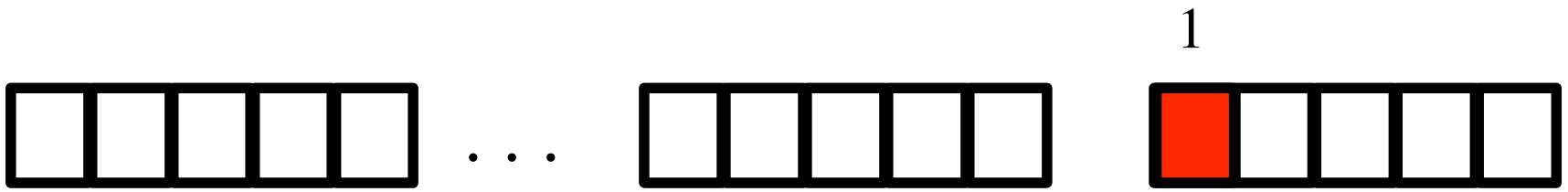
MQ Cache -Example



MQ Cache -Example

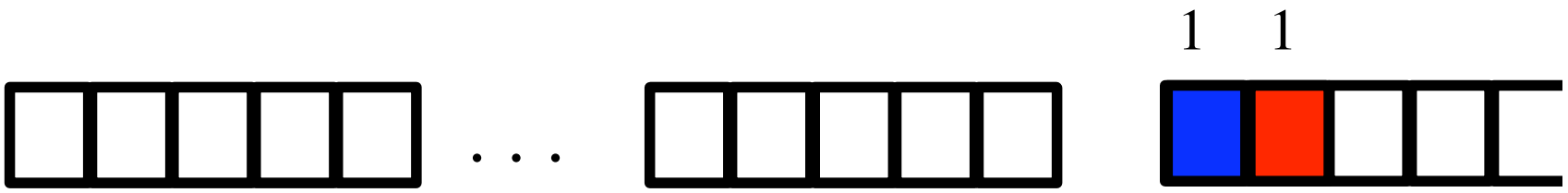
Incoming request:





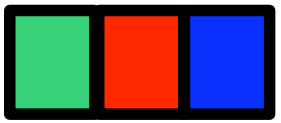
MQ Cache -Example

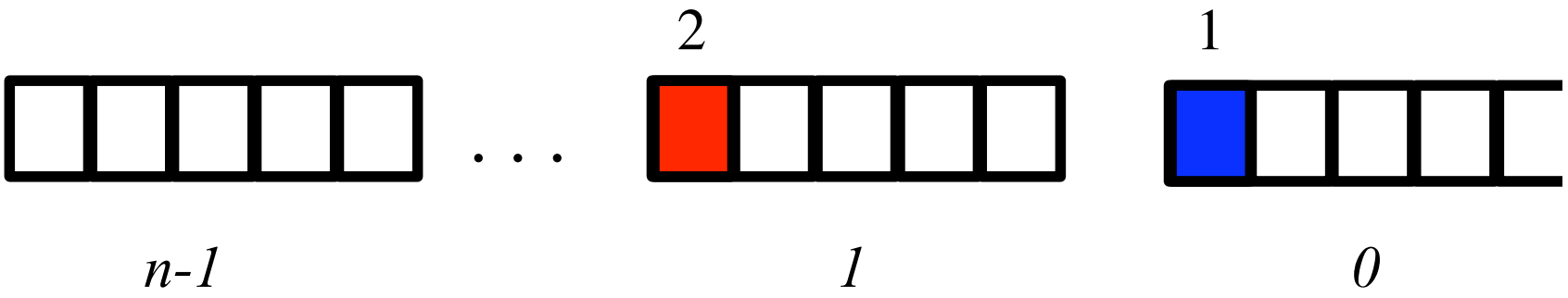
Incoming request:



MQ Cache -Example

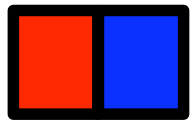
Incoming request:

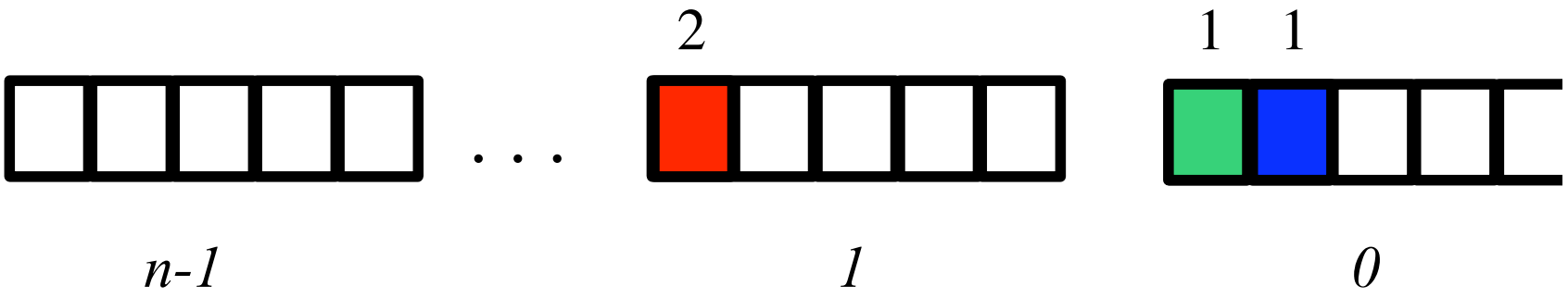


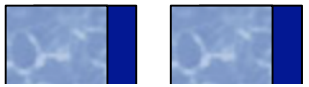


MQ Cache -Example

Incoming request:

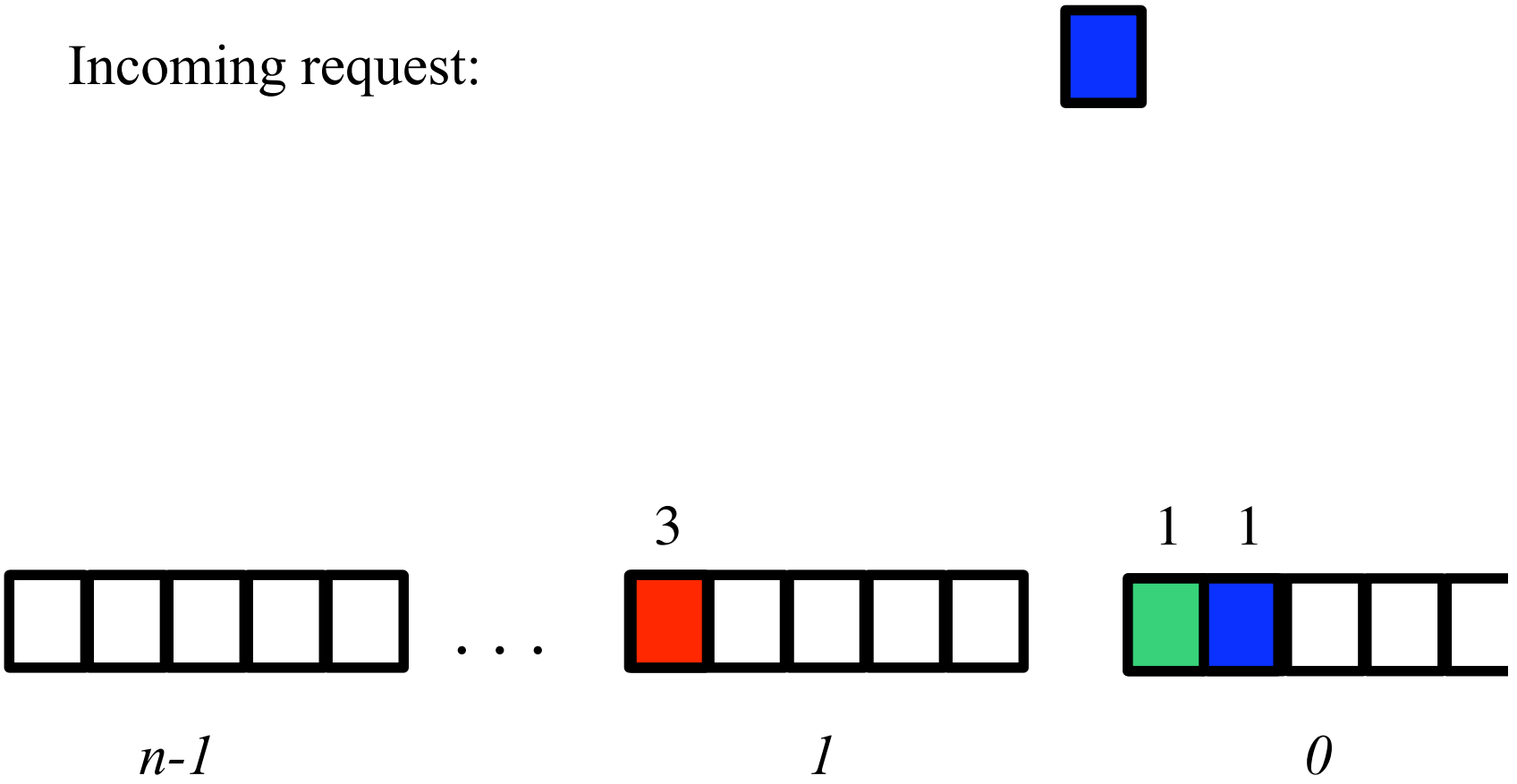




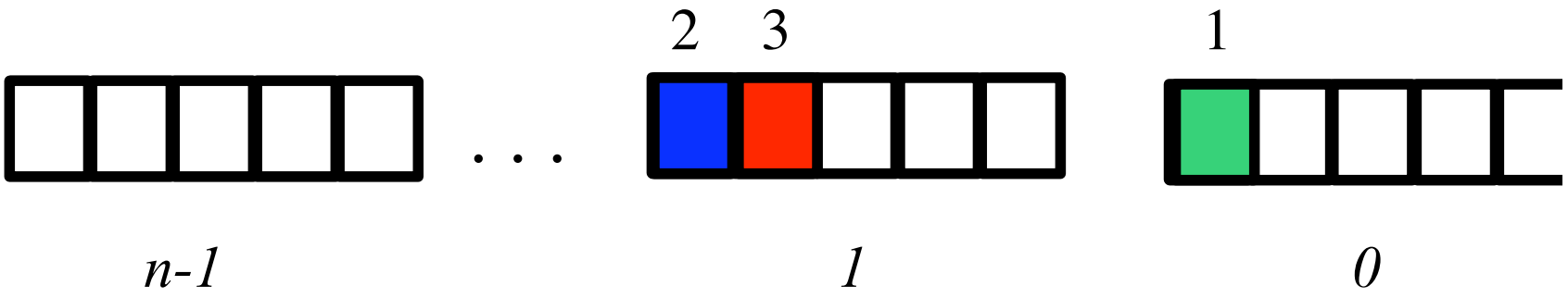


MQ Cache -Example

Incoming request:

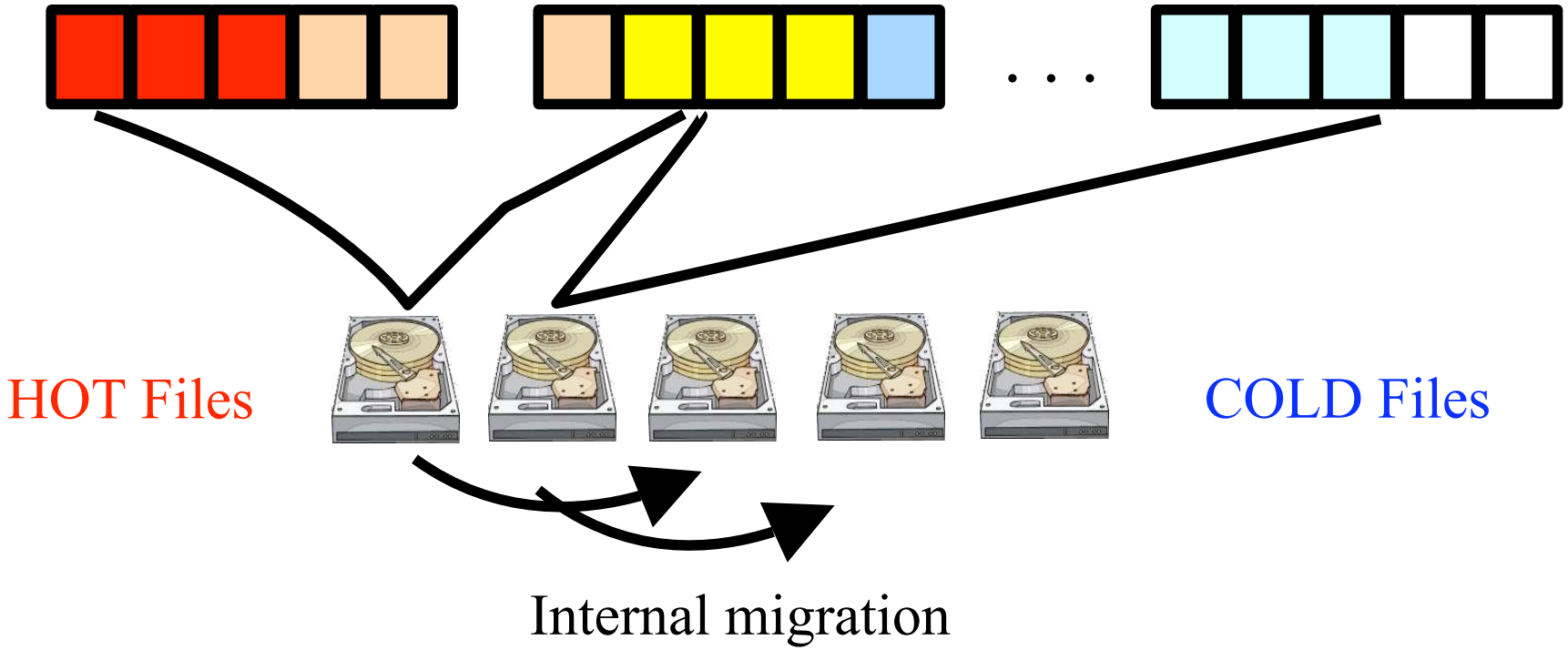


MQ Cache -Example

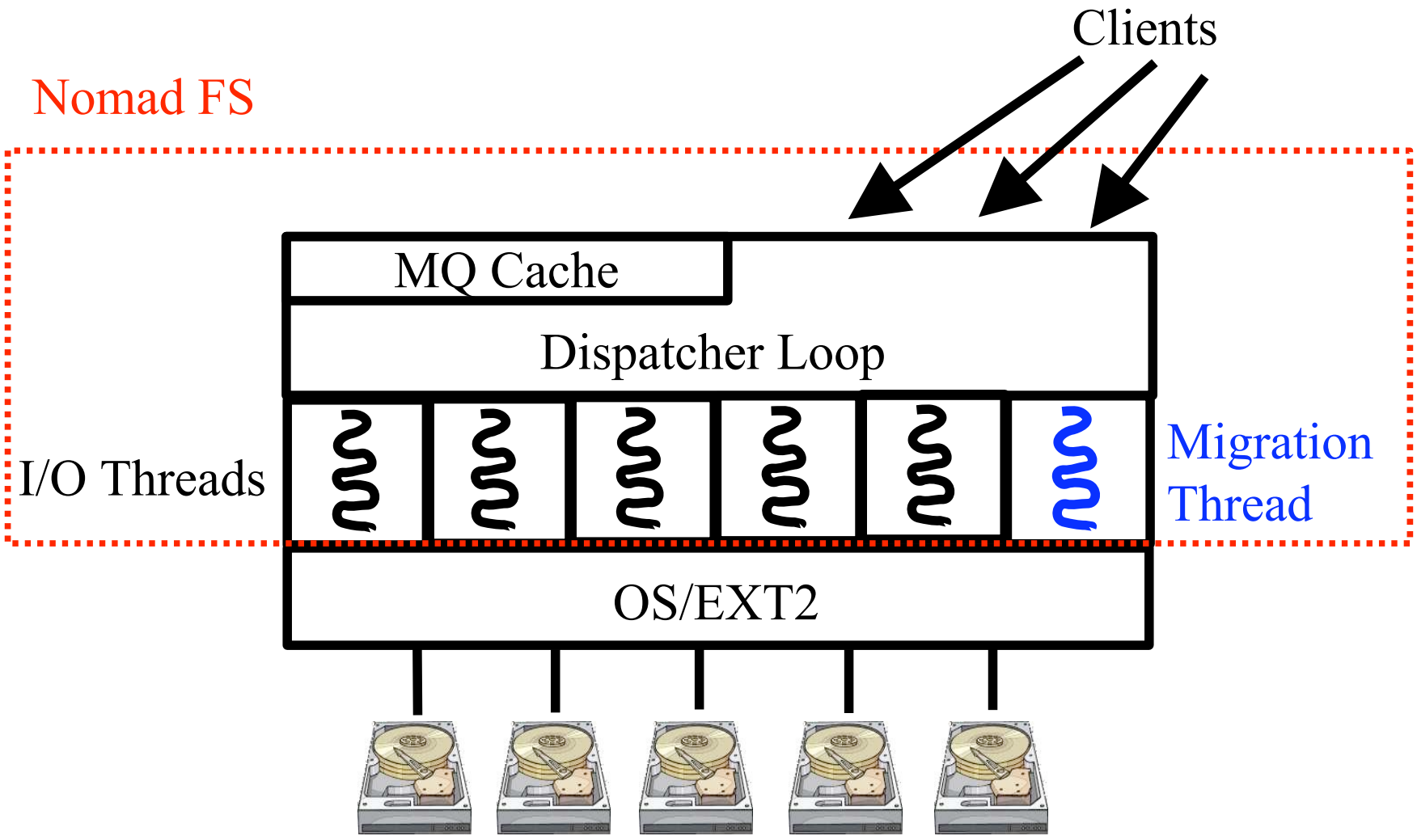


Migration Algorithm

- Re-use MQ cache on file handlers.
- Migrate "hotter" files to first disk, then second and so on.

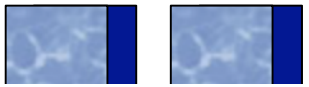


Design


Clients
Nomad FS
MQ Cache
Dispatcher Loop
I/O Threads
Migration Thread
OS/EXT2

Nomad FS

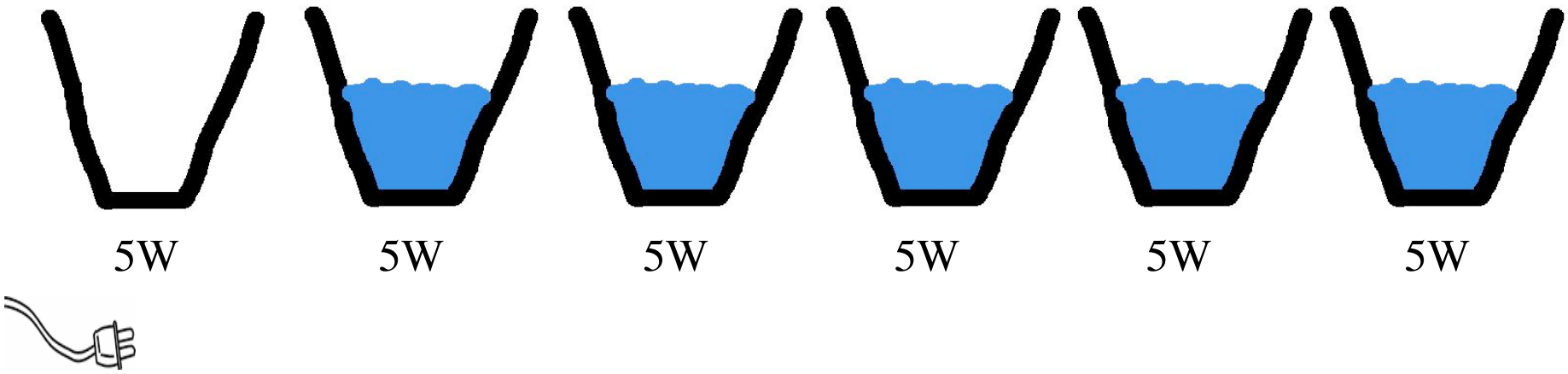
- Implements Popular Data Concentration (PDC).
- Uses MQ Cache for its second-level cache.
- Implements other algorithms for comparison:
  - MAID - Massive Array of Idle Disks [Colarelli et al. SC'02].
  - Naïve two-speed disks.



MAID

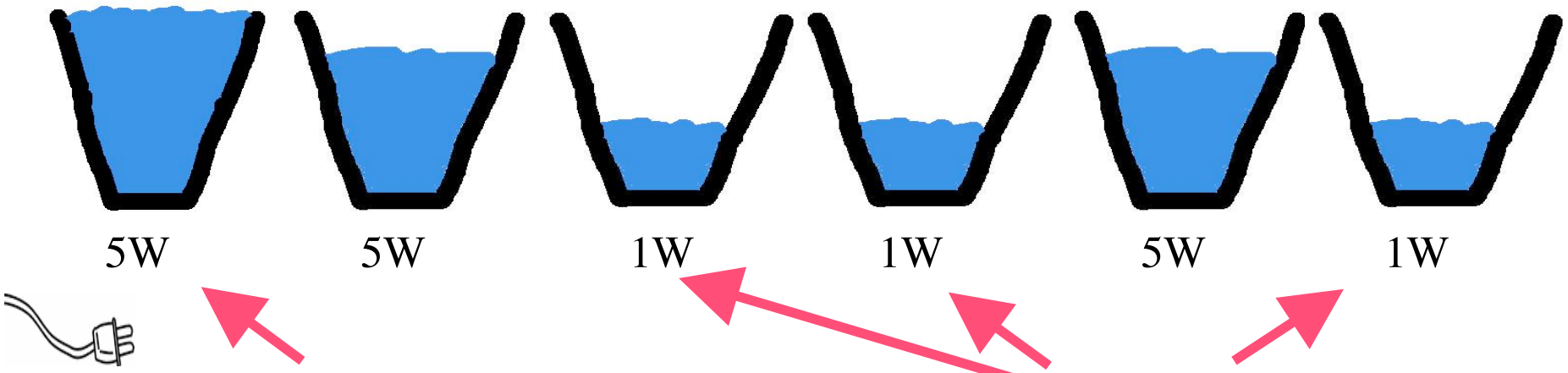
Total Consumption: 25 W

MAID



Total Consumption: 30 W

MAID



Total Consumption: 18 W

Implementation

Implementation

Description	Value
Disk model	Seagate Cheetah ST39205LC
Standard interface	SCSI
Storage capacity	9.17 GBytes
Number of platters	1
Rotational speed	10000 rpm
Avg. seek time	5.4 msecs
Avg. rotation time	3 msecs
Transfer rate	31 MBytes/sec
Idle power	5.26 Watts
Down power	1.86 Watts
Active energy (8-KB read)	61 mJoules
Spin up energy	65.91 Joules
Spin down energy	28.25 Joules
Spin up time	6.12 secs
Spin down time	11.24 secs
Idleness threshold	17.9 secs

Table 1: Main characteristics and measured power, energy, and time statistics of our SCSI disk.

Preliminary Results

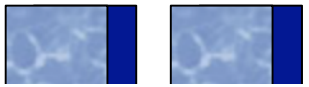
- System works with active and powerdown (sleep) modes.

HOWEVER,

- Not enough idle time to spindown disks on busy servers.
- So, how useful is that?

SOLUTION?

- Use multi-speed disks.
- But not available yet.
- Then simulate multi-speed disks.
- Carefully...



Validation

Description	Total energy consumed			Files moved		MBytes moved		Spin downs/ups		Delayed requests	
	Sim	Real	Error	Sim	Real	Sim	Real	Sim	Real	Sim	Real
PDC	152023.23	172842.68	12.0%	9715	10943	227	283	336	326	1.3%	1.2%
MAID	141349.20	156335.21	9.6%	8738	8716	204	204	300	304	1.1%	1.0%
FT	190374.14	200478.84	5.0%	n/a	n/a	n/a	n/a	10	13	0.2%	0.2%
EO	190345.68	200832.88	5.2%	n/a	n/a	n/a	n/a	n/a	n/a	0.0%	0.0%

Table 4: Summary of validation. Energy values are in Joules.

Parameter Space

- Disk Drive Parameters: *rotation speed, power consumption, conventional, two-speed disks.*

- File System Parameters: *cache size, migration period, number of disks, cache replacement policy.*

- Workload Characteristics: *coverage, popularity, request rate, % writes, temporal correlation.*

Results

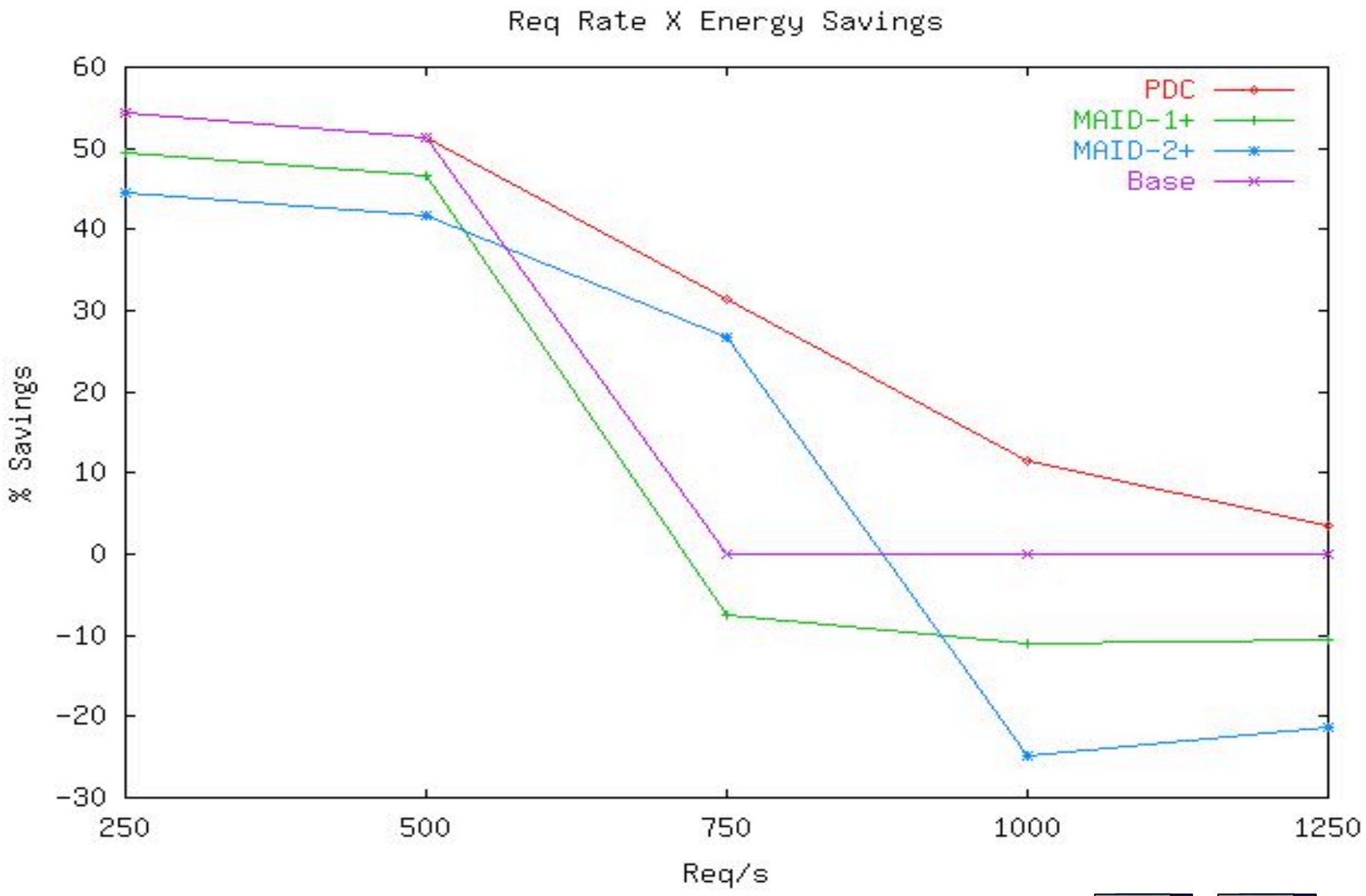
Synthetic traces (default values):

- Req Rate: 750 r/s
- File Size: 48kb
- Disk params: 9.17 Gb, 10k/3k rpm
- Alpha: 0.85
- Coverage: 40%
- Read-only
- Cache: 1 Gb

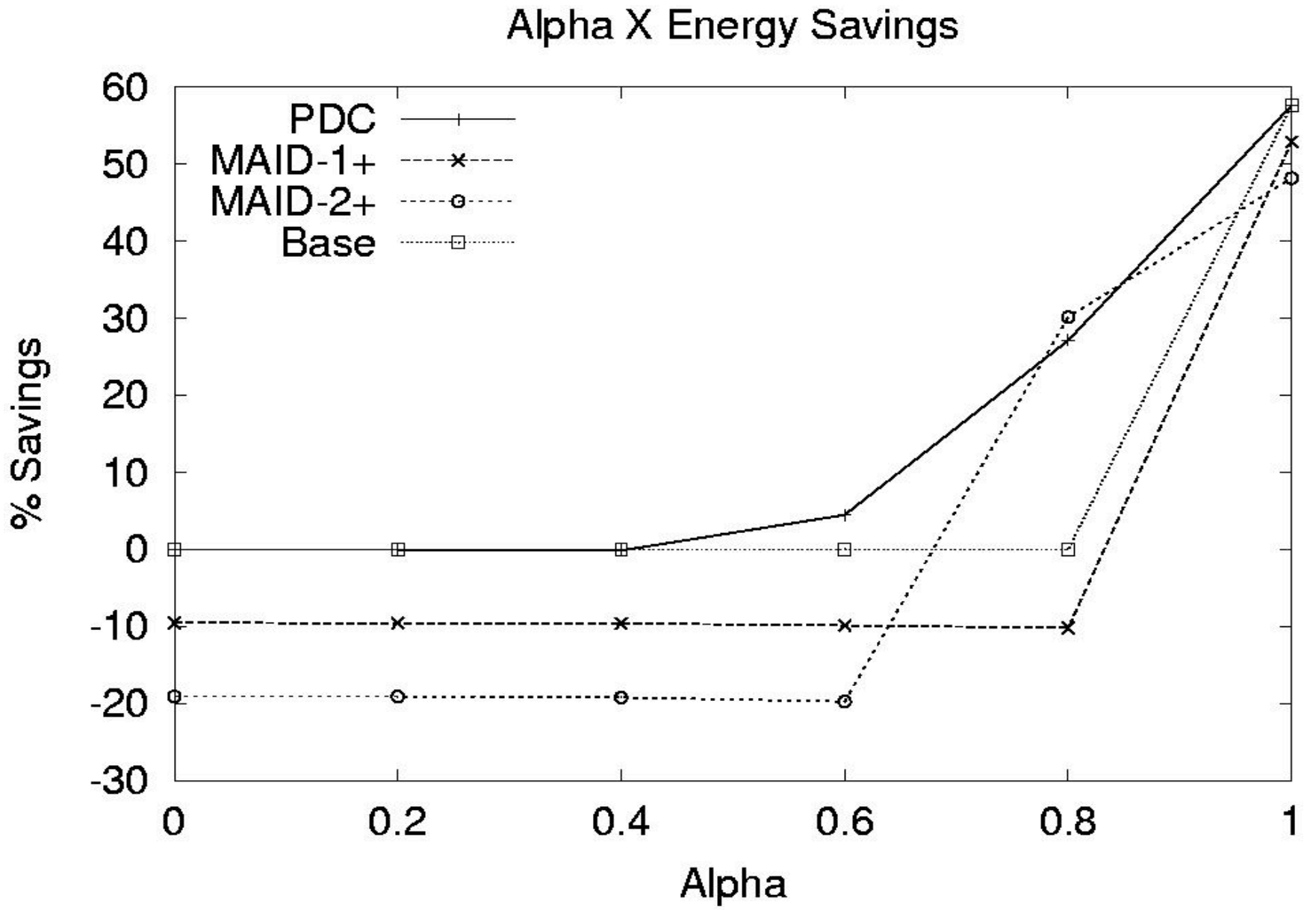
Real traces:

- Hummingbird (Proxy cache)
  - Req Rate: 241 r/s
  - Avg File Size: 18kb
  - Disk params: 11 Gb, 10k/3k rpm
  - Alpha: 0.70
  - Coverage: 93%
  - Writes: 35%
  - Cache: 1 Gb
- Pop Cache (filtered proxy cache)
  - Req Rate: 263 r/s
  - Avg File size: 20kb
  - Disk params: 6 Gb, 10k/3k rpm
  - Alpha: 0.93
  - Coverage: 55%
  - Read-only
  - Cache: 64 Mb

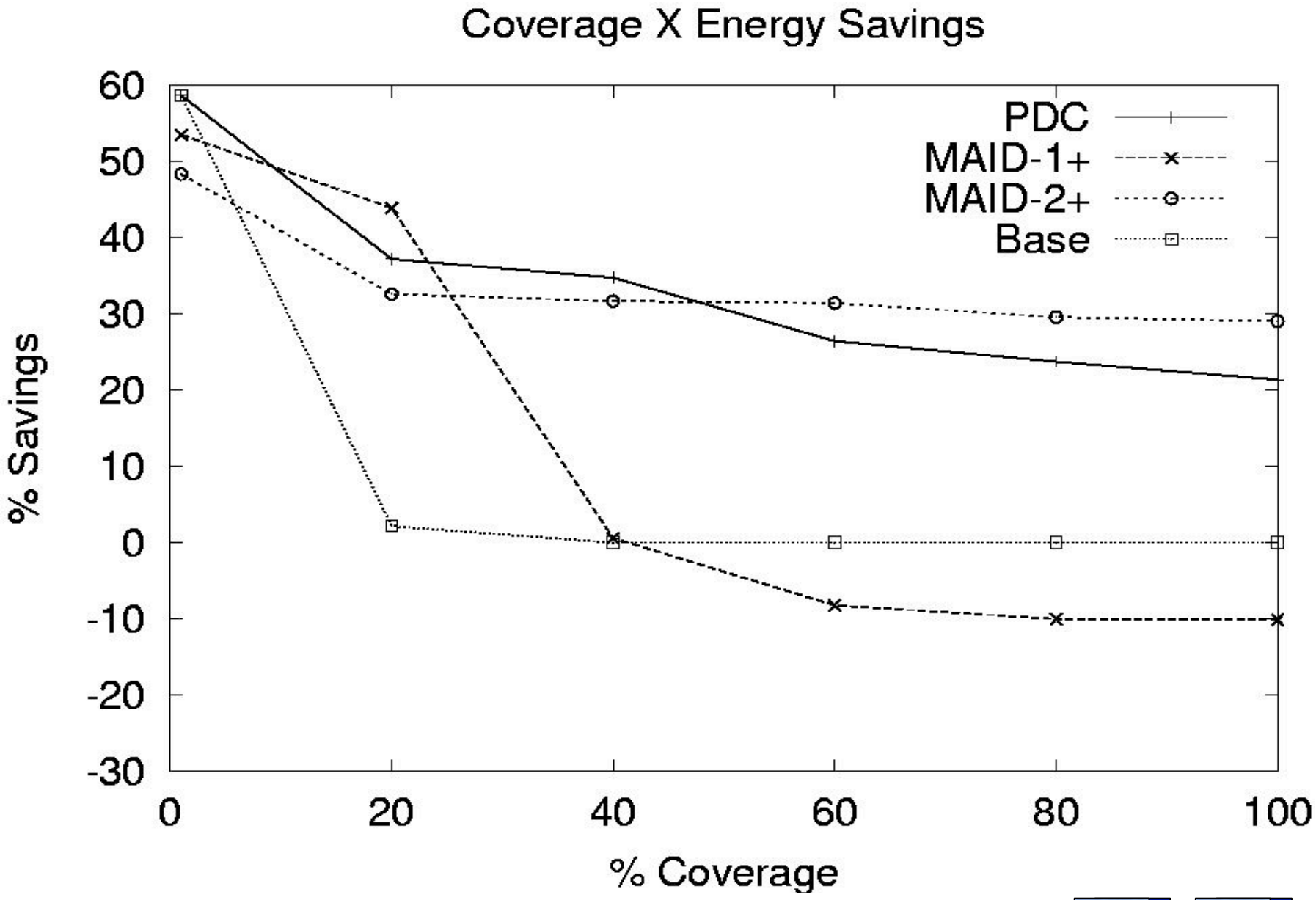
Results


Req Rate X Energy Savings
PDC
MAID-1+
MAID-2+
Base
% Savings
60
50
40
30
20
10
0
-10
-20
-30
250
500
750
1000
1250
Req/s

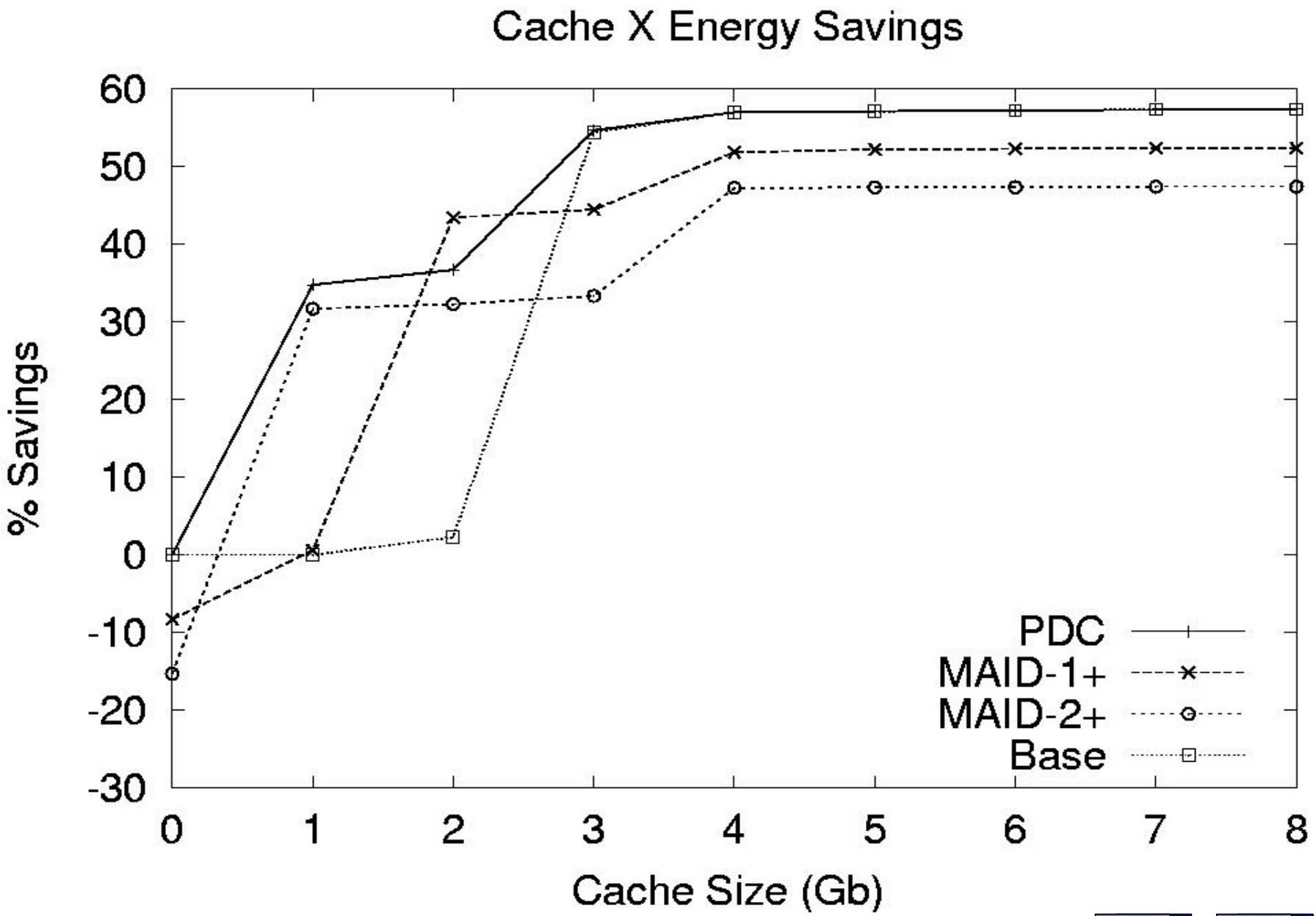
Results


Alpha X Energy Savings
PDC
MAID-1+
MAID-2+
Base
% Savings
60
50
40
30
20
10
0
-10
-20
-30
0
0.2
0.4
0.6
0.8
1
Alpha

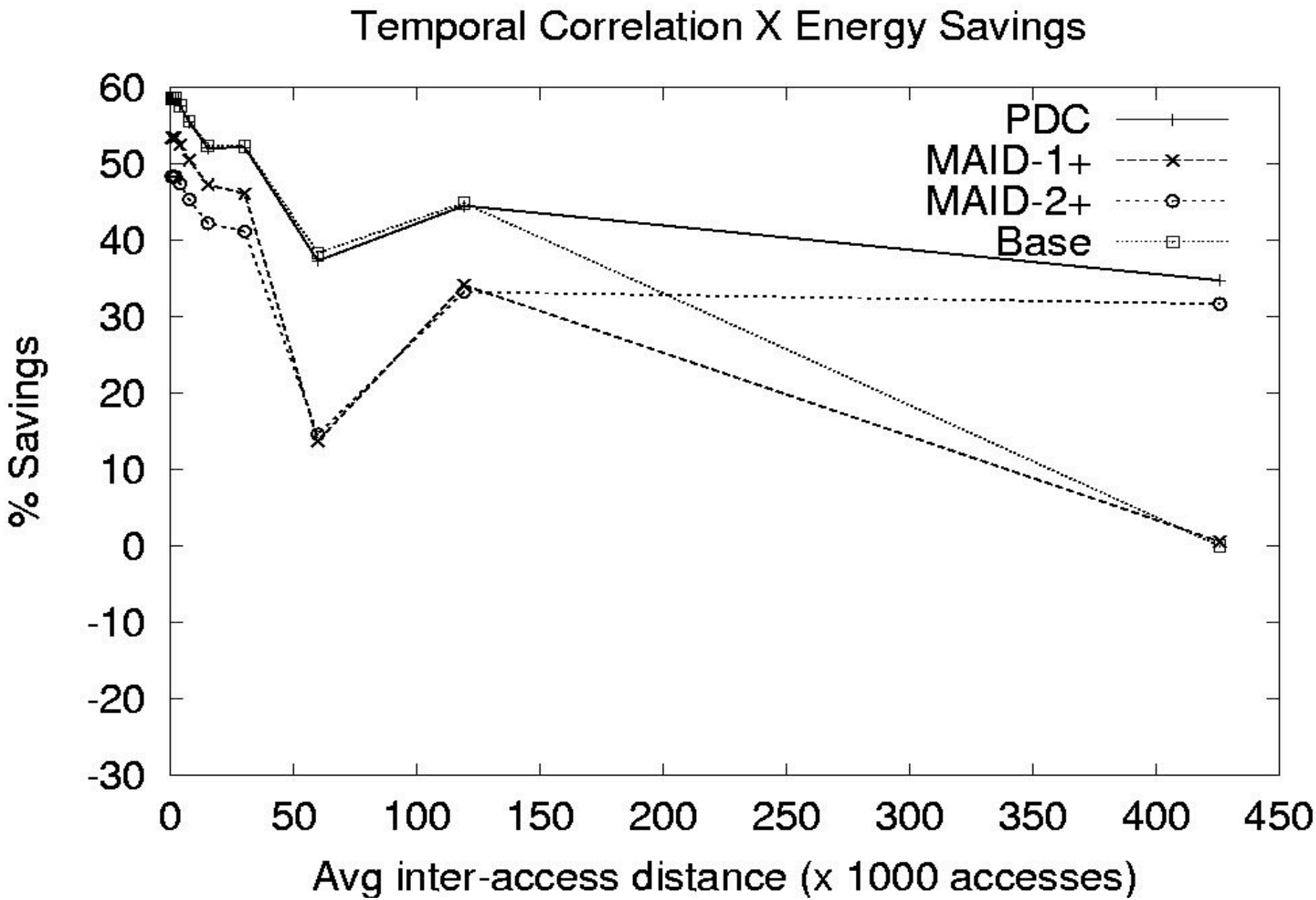
Results


Coverage X Energy Savings
PDC
MAID-1+
MAID-2+
Base
% Savings
60
50
40
30
20
10
0
-10
-20
-30
0
20
40
60
80
100
% Coverage

Results


Cache X Energy Savings
% Savings
60
50
40
30
20
10
0
-10
-20
-30
0
1
2
3
4
5
6
7
8
Cache Size (Gb)
PDC
MAID-1+
MAID-2+
Base

Results


Temporal Correlation X Energy Savings
% Savings
60
50
40
30
20
10
0
-10
-20
-30
0
50
100
150
200
250
300
350
400
450
Avg inter-access distance (x 1000 accesses)
PDC
MAID-1+
MAID-2+
Base

Results

Real Traces

Pop Cache (energy gains) (% latency)

- FT : 36% 7.8%
- PDC : 42% 8.5%
- MAID-1+ : 33% 8.4%
- MAID-2+ : 29% 8.1

HummingBird (energy gains) (% latency)

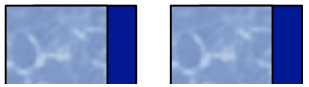
- FT : 23% 6.0%
- PDC : 26% 8.5%
- MAID-1+ : 20% 14.3%
- MAID-2+ : 15% 11.1%

Related Work

Inspired by Load Concentration [Pinheiro et al. COLP'01].

Energy conservation on storage servers:

- [Carrera et al. ICS'03] - two speed emulation.
- [Gurumurthi et al. ISCA'03] - multi-speed simulation.
- [Gurumurthi et al. ISPASS'03] - RAID w/ multi-speed setting.
- [Colarelli et al. SC'02] - MAID
- [Zhu et al. HPCA'04] - Cache repl. algo. for disk idleness.



Conclusions

- Introduced new energy-saving technique PDC.
- Implemented NomadFS.
  - PDC, MAID+, FT, Multi-speed disks.
- Tested and validated NomadFS simulator.
- Substation energy savings are possible under light load.
- PDC able to get more energy gains beyond naïve two-speed.
- PDC more robust/adaptable than MAID.