



# A SURVEY ON POWER-REDUCTION TECHNIQUES FOR DATA-CENTER STORAGE SYSTEMS

Tom Bostoan  
December 16, 2013

# ACM COMPUTING SURVEYS ARTICLE

- Article: Bostoен, T., Mullender, S., and Berbers, Y. Power-reduction techniques for data-center storage systems. ACM Comput. Surv. 45, 3 (June 2013), 33:1-33:38
- Me: researcher at Bell Labs in Belgium pursuing a Ph.D. in collaboration with KU Leuven.
- Supervisors: Sape Mullender (Bell Labs) and Yolande Berbers (KU Leuven).
- Website: <http://belllabs.be/people/tombostoen>
- Contact: [tom.bostoen@alcatel-lucent.com](mailto:tom.bostoen@alcatel-lucent.com)

## Power-Reduction Techniques for Data-Center Storage Systems

TOM BOSTOEN and SAPE MULLENDER, Alcatel-Lucent Bell Labs  
YOLANDE BERBERS, Katholieke Universiteit Leuven

As data-intensive, network-based applications proliferate, the power consumed by the data-center storage subsystem surges. This survey summarizes, organizes, and integrates a decade of research on power-aware enterprise storage systems. All of the existing power-reduction techniques are classified according to the disk-power factor and storage-stack layer addressed. A majority of power-reduction techniques is based on dynamic power management. We also consider alternative methods that reduce disk access time, conserve space, or exploit energy-efficient storage hardware. For every energy-conservation technique, the fundamental trade-offs between power, capacity, performance, and dependability are uncovered. With this survey, we intend to stimulate integration of different power-reduction techniques in new energy-efficient file and storage systems.

Categories and Subject Descriptors: C.4 (Computer Systems Organization): Performance of Systems; D.4.2 (Operating Systems): Storage Management; D.4.3 (Operating Systems): File Systems Management; E.2 (Data): Data Storage Representations; E.5 (Data): Files; H.3.2 (Information Storage and Retrieval): Information Storage—File Organization

General Terms: Design, Algorithms, Performance, Reliability

Additional Key Words and Phrases: Cloud storage, data center, disk drive, energy efficiency, power reduction

### ACM Reference Format:

Bostoен, T., Mullender, S., and Berbers, Y. 2013. Power-reduction techniques for data-center storage systems. ACM Comput. Surv. 45, 3, Article 33 (June 2013), 38 pages.  
DOI: <http://dx.doi.org/10.1145/2480741.2480750>

### 1. INTRODUCTION

With the advent of data-intensive, network-based applications and services, data centers around the world consume a significant and rapidly growing amount of electricity. In the U.S. alone, all data centers together are projected to consume 100 TWh per year by 2011, which costs more than \$10billion at a common price of \$100 per MWh [Kaushik et al. 2010]. Since 2005, data-center power consumption in the U.S. has more than doubled, from 40 TWh [Zhu et al. 2005]. The energy consumed by data centers represents 1–2% of the total U.S. power consumption [Sehgal et al. 2010]. The cost of energy is a growing concern for data-center operators, as it may constitute half of the data center's total cost of ownership [Loukov and Sipek 2008]. Data storage alone is responsible for about 25 to 35% of data-center power consumption [Kim and Rotem

This work is supported by the Flanders Agency for Innovation by Science and Technology (IWT), grant IWT 100690.

Authors' addresses: T. Bostoен and S. Mullender, Alcatel-Lucent Bell Labs, Copernicuslaan 50, B-2018 Antwerp, Belgium; email: {Tom.Bostoен, Sape.Mullender}@alcatel-lucent.com; Y. Berbers, Computer Science Department, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee (Leuven), Belgium; email: yolande.berbers@kuleuven.be

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).  
© 2013 ACM 0360-0300/2013/06-ART33 \$15.00  
DOI: <http://dx.doi.org/10.1145/2480741.2480750>

ACM Computing Surveys, Vol. 45, No. 3, Article 33, Publication date: June 2013.

33



Alcatel-Lucent



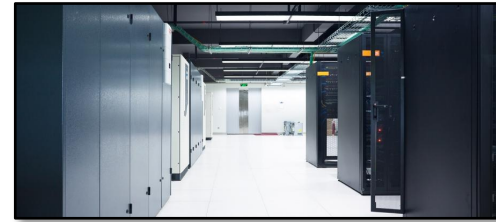
Alcatel-Lucent



Alcatel-Lucent 

# INTRODUCTION

- Electricity consumption by data centers surges due to proliferation of data-intensive network-based applications and services
- Data storage accounts for 25% to 35% of power consumed by data centers
- Because storage subsystem consists of hard disk drives, which require mechanical movement for their operation
- Research on power-reduction techniques for data-center storage systems started beginning of 2000s
- Survey on power-reduction techniques for data-center storage systems:
  - Any storage-stack layer
  - Any workload
  - Software rather than hardware
  - Focus on power-reduction techniques but also performance-improvement techniques that save energy



(1)



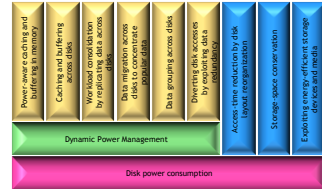
(2)

(1) © [ITP.net](http://ITP.net)  
(2) © [Western Digital Technologies, Inc.](http://Western Digital Technologies, Inc.)

# AGENDA

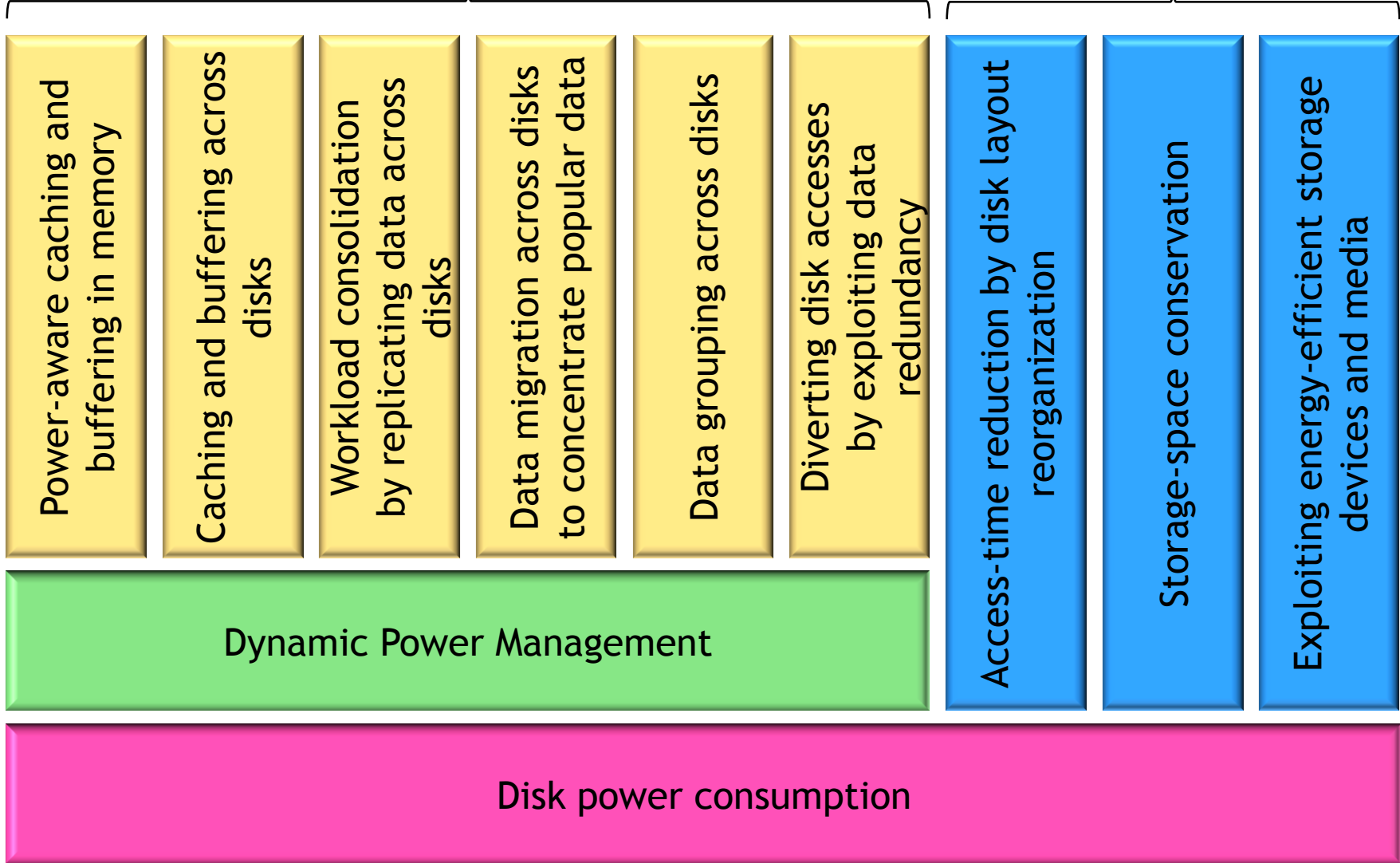
1. Introduction
2. Disk Power Consumption
3. Dynamic Power Management (DPM)
4. DPM-Enabling Workload-Shaping Techniques
5. Access-Time Reduction by Disk-Layout Reorganization
6. Storage-Space Conservation
7. Exploiting Energy-Efficient Storage Devices and Media
8. Conclusion

# SURVEY ORGANIZATION



## DPM-enabling techniques

## Other



Disk power consumption

Dynamic Power Management

Power-aware caching and buffering in memory

Caching and buffering across disks

Workload consolidation by replicating data across disks

Data migration across disks to concentrate popular data

Data grouping across disks

Diverting disk accesses by exploiting data redundancy

Access-time reduction by disk layout reorganization

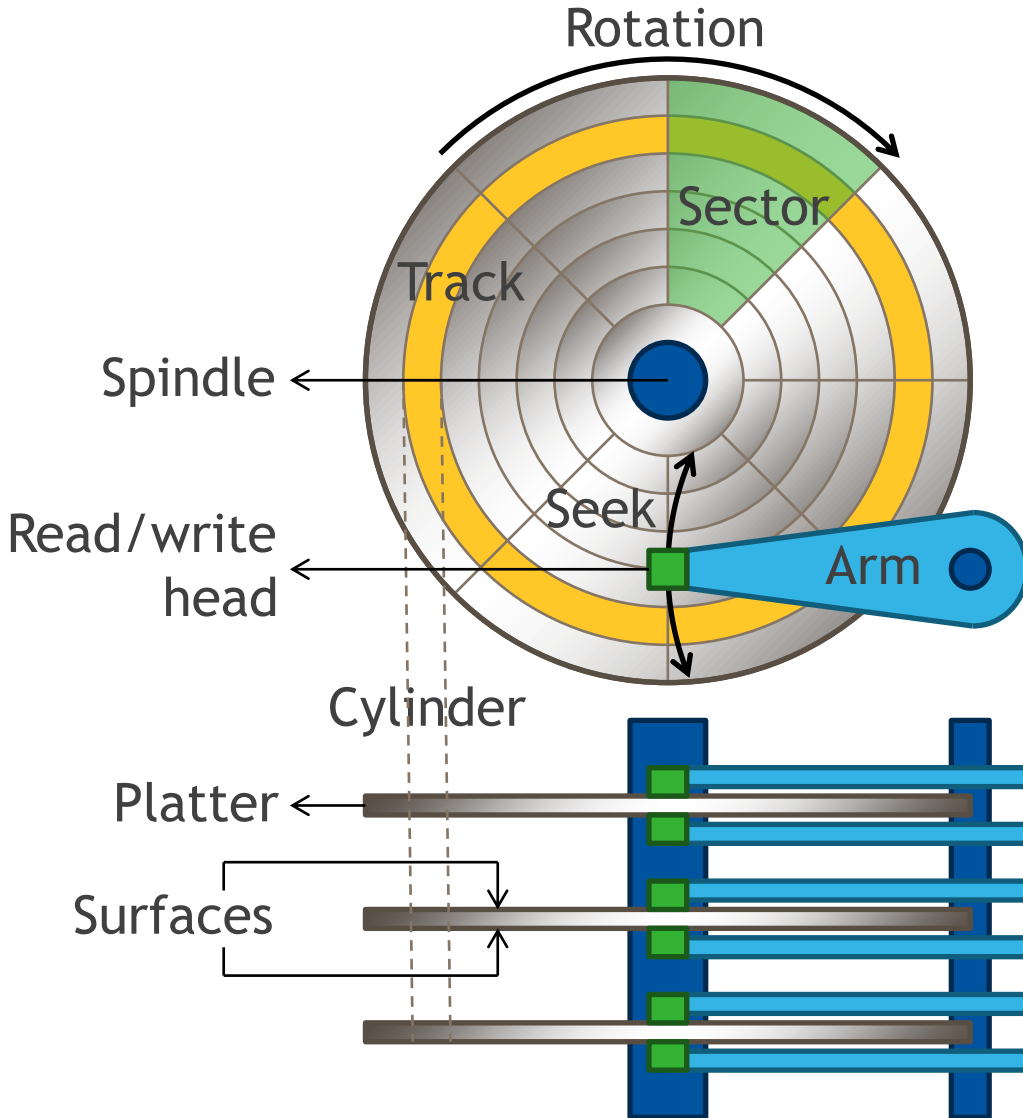
Storage-space conservation

Exploiting energy-efficient storage devices and media



# HARD DISK DRIVE

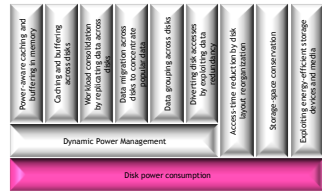
Power-aware caching and buffering in memory	Caching and buffering across disks	Maximal consolidation by replicating data across disks	Thin provisioning: use disks to concentrate popular data	Data grouping across disks	Diverting disk access by moving data	Access time reduction by file layout, reorganization	Storage space conservation	Exploiting energy-efficient storage devices and media
Dynamic Power Management								
Disk power consumption								



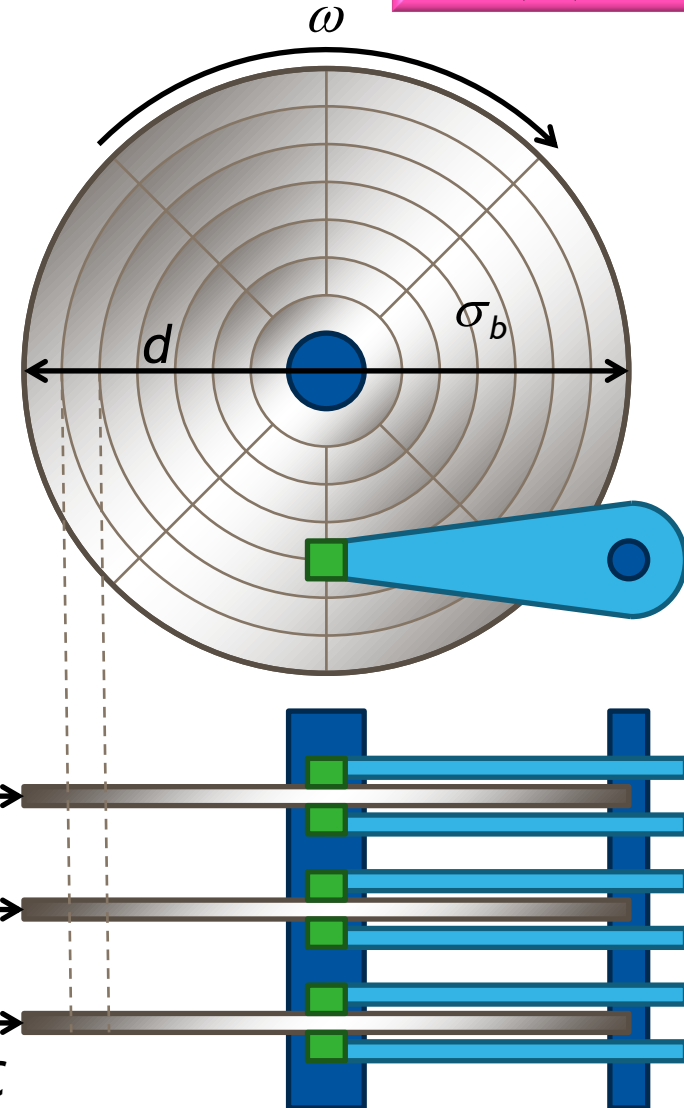
Copyright © Western Digital Technologies, Inc.

# DISK POWER CONSUMPTION

## TOTAL DISK POWER AND SPIN POWER

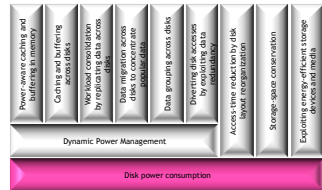


- Hard disk drive consumes power
  - To keep platters spinning:  $P_{sp}$
  - To displace actuator arm:  $P_{sk}$
  - For interface and control logic:  $P_{ct}$
- Total disk power:  $P_{dk} = P_{sp} + P_{sk} + P_{ct}$
- Spin power:  $P_{sp} \propto N_{pl} d^{4.6} \omega^{2.8}$
- Reduction of number of platters  $N_{pl}$ 
  - Leads to proportional reduction of disk capacity  $C \propto N_{pl} d^2 \sigma_b$
  - Unless compensated for by increase of  $\sigma_b$
- Reduction of platter diameter  $d$ 
  - Leads to reduction of  $P_{sp} / C$  but increase of  $P_{ct} / C$

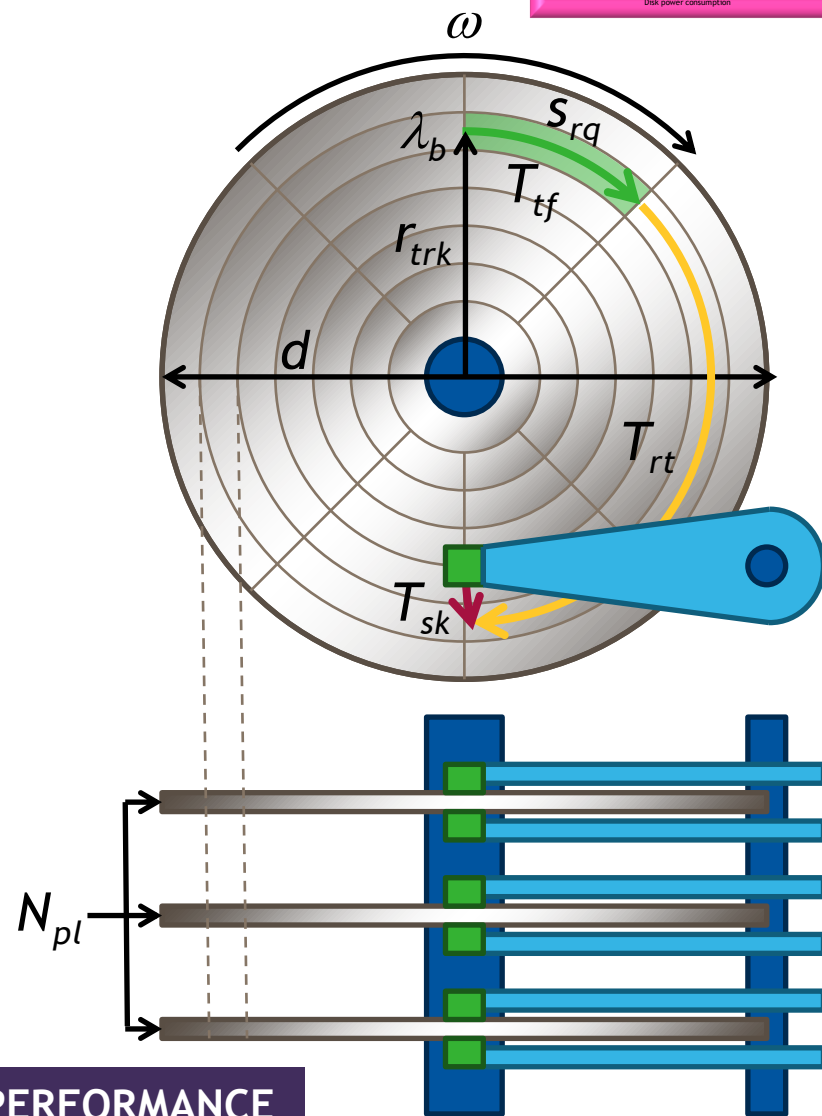


# DISK POWER CONSUMPTION

## REDUCTION OF ROTATIONAL SPEED



- Spin power:  $P_{sp} \propto N_{pl} d^{4.6} \omega^{2.8}$
- Reduction of rotational speed  $\omega$ 
  - Transfer rate  $R_{tf} = \frac{\omega}{60} \frac{2\pi r_{trk} \lambda_b}{8 \times 1024 \times 1024}$
  - Transfer time  $T_{tf} = \frac{s_{rq}}{1024} \frac{1000}{R_{tf}} \propto \frac{1}{\omega}$
  - Rotational latency  $T_{rt} = \frac{1}{2} \frac{60 \times 10^3}{\omega}$
  - Access time  $T_{acs} = T_{sk} + T_{rt} + T_{tf}$
  - Queueing time  $T_q = \rho T_{acs} / (1 - \rho)$
  - Response time  $T_{rp} = T_q + T_{acs}$
  - Throughput  $R_{rq} = 1000 / T_{acs}$

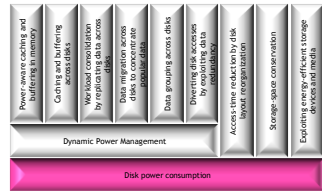


WHEN POWER REDUCTION TAKES PRIORITY OVER PERFORMANCE

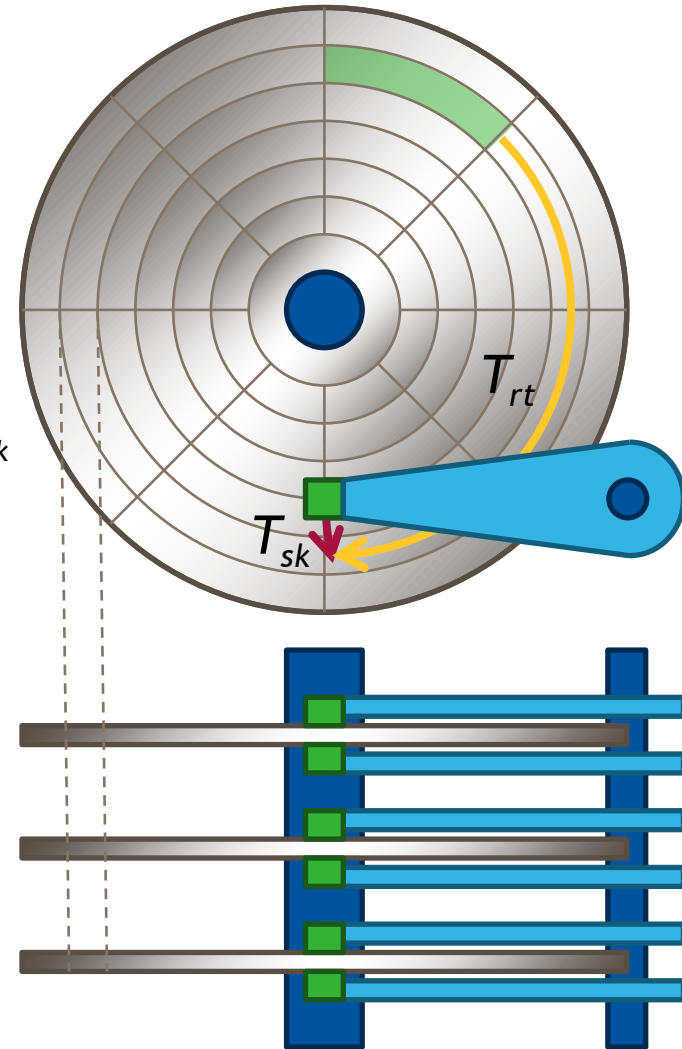


# DISK POWER CONSUMPTION

## SEEK POWER



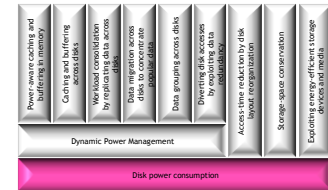
- Total disk power:  $P_{dk} = P_{sp} + P_{sk} + P_{ct}$
- Seek power:  $P_{sk} = P_{acl} + P_{dcl} + P_{st}$ 
  - No coasting for average seek
  - Assumption:  $a_{acl} = a_{dcl} = a$
  - Acceleration/deceleration power:  $P_{acl} = P_{dcl} \propto ad_{sk}$
- Reduction of seek acceleration
  - Leads to increase of seek time  $T_{sk} = 2\sqrt{\frac{d_{sk}}{a}} + T_{st}$
  - Leads to decrease of same size of rotational latency
  - Seek acceleration may be reduced until  $T_{rt} = 0$



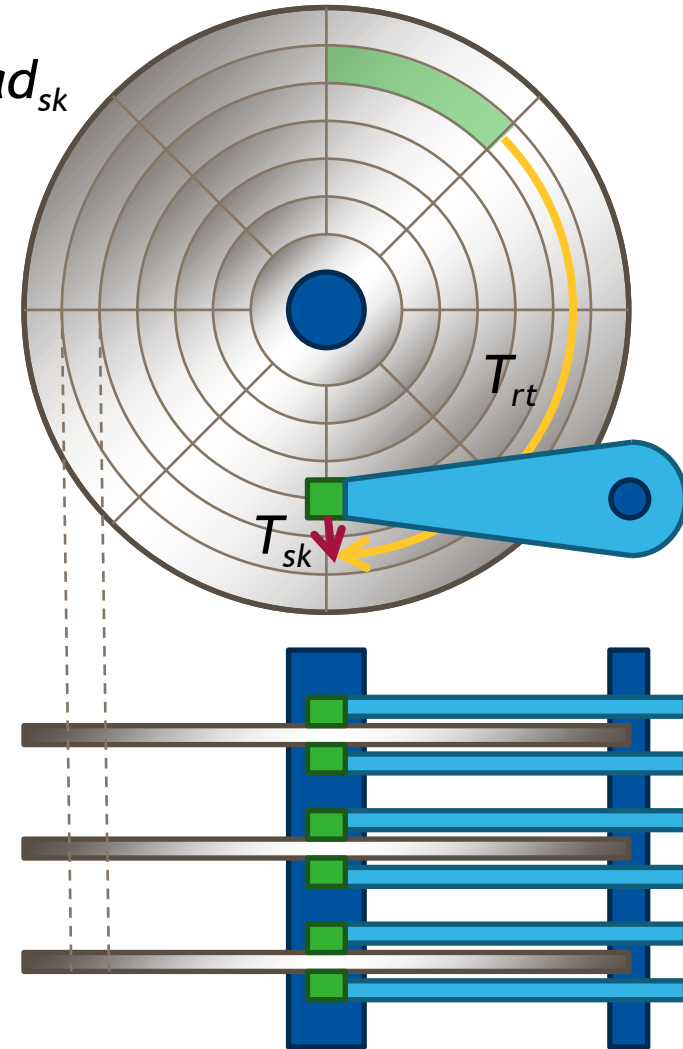
### JUST-IN-TIME SEEKING

# DISK POWER CONSUMPTION

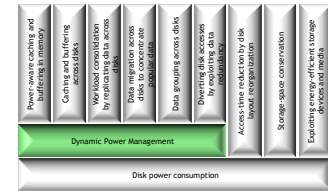
## REDUCTION OF SEEK DISTANCE AND CONTROL POWER



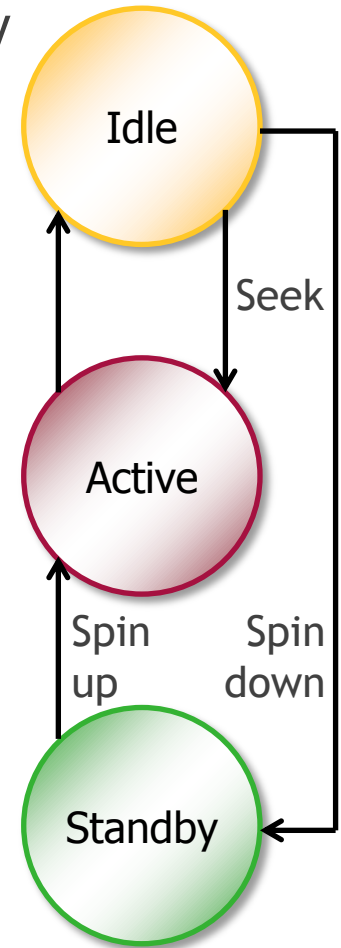
- Acceleration/deceleration power:  $P_{acl} = P_{dcl} \propto ad_{sk}$
- Reduction of average seek distance  $d_{sk}$ 
  - By reducing platter diameter  $d$
  - By increasing areal bit density  $\sigma_b$
  - By improving data placement on disk
- Total disk power:  $P_{dk} = P_{sp} + P_{sk} + P_{ct}$
- Control power  $P_{ct}$ 
  - For control and interface logic
  - Even when disk is idle but higher during data transfer
  - Data transfer power only significant for larger I/O requests



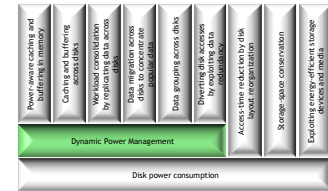
# DYNAMIC POWER MANAGEMENT



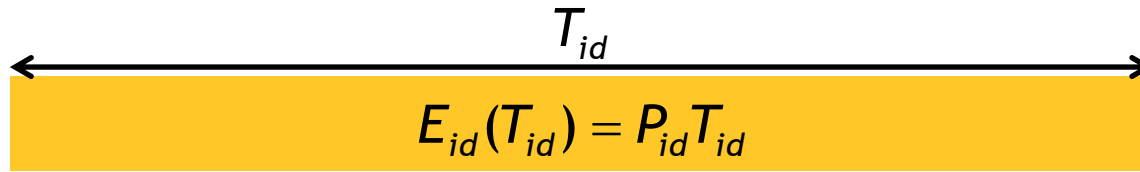
- Definition: technique for reducing power consumption by turning off system components or decreasing their performance when they are idle or underutilized
- DPM can make system energy-proportional
- Without DPM, disk is far from energy-proportional because  $\sim 2/3$  of maximum power consumed when idle
  - To keep platters spinning
  - For interface and control logic (excluding data transfer)
- DPM = Power-State Machine + Power-Control Policy
  - Idle: platters spin
  - Active: disk transfers data
  - Standby: platters at rest



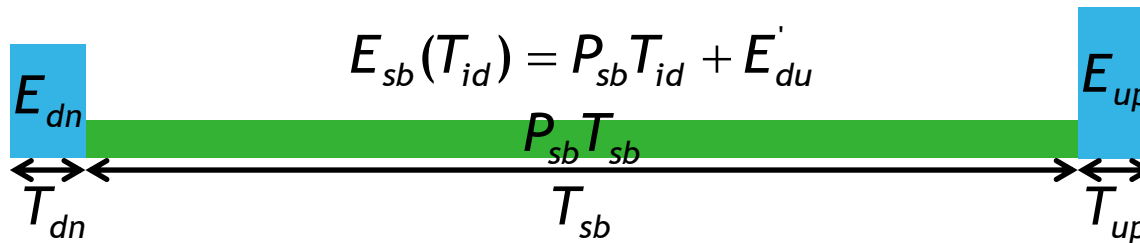
# DYNAMIC POWER MANAGEMENT BREAK-EVEN TIME



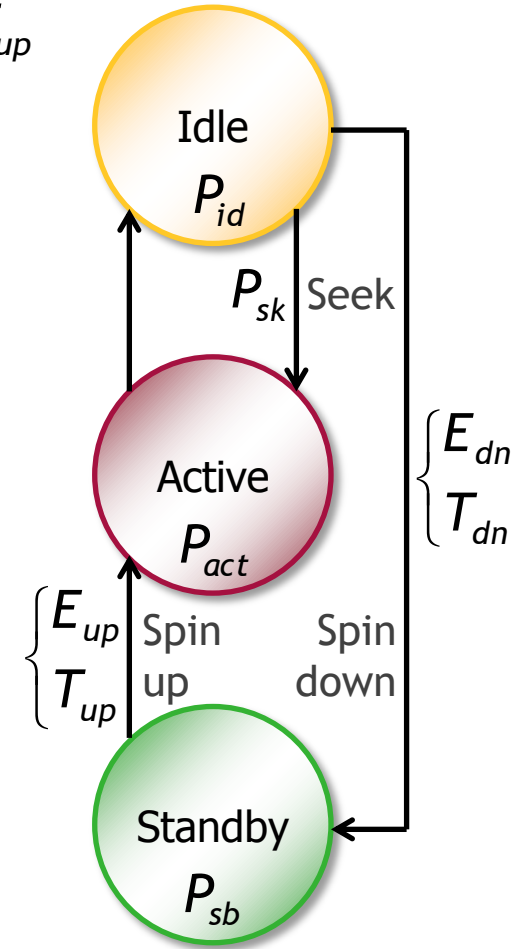
- Let  $E_{du} = E_{dn} + E_{up}$ ,  $E'_{du} = E_{du} - P_{sb}T_{du}$ , and  $T_{du} = T_{dn} + T_{up}$
- Assuming idle time  $T_{id}$ , when to spin down disk?
- Option 1: keep disk in idle mode



- Option 2: spin disk down

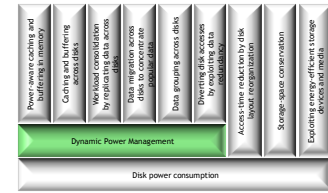


- Break-even time:  $E_{id}(T_{id}) = E_{sb}(T_{id}) \Rightarrow T_{be} = \frac{E'_{du}}{\Delta P}$
- With:  $\Delta P = P_{id} - P_{sb}$



# DYNAMIC POWER MANAGEMENT

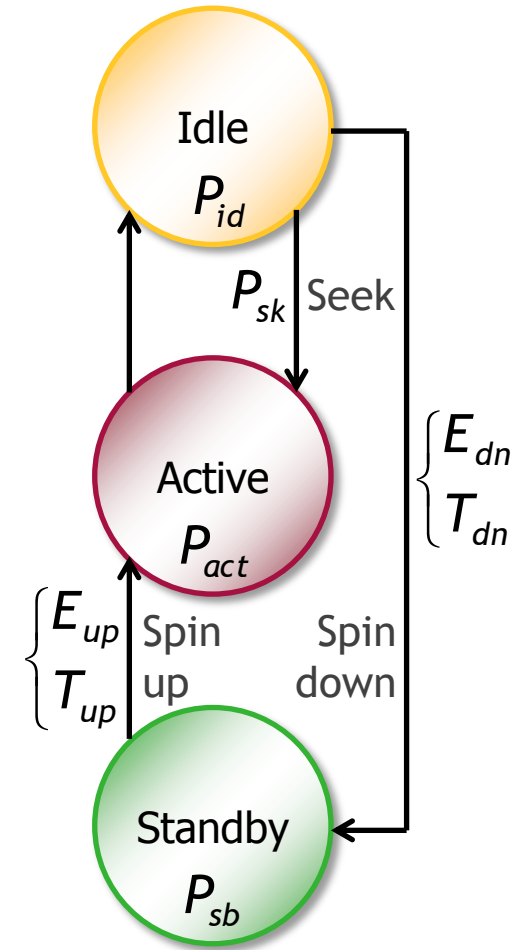
## POWER-CONTROL POLICY



- Threshold-based power-control policy
  - Spin down disk when idle time exceeds threshold
  - Spin up disk when new request arrives
  - = Traditional Power Management (TPM)

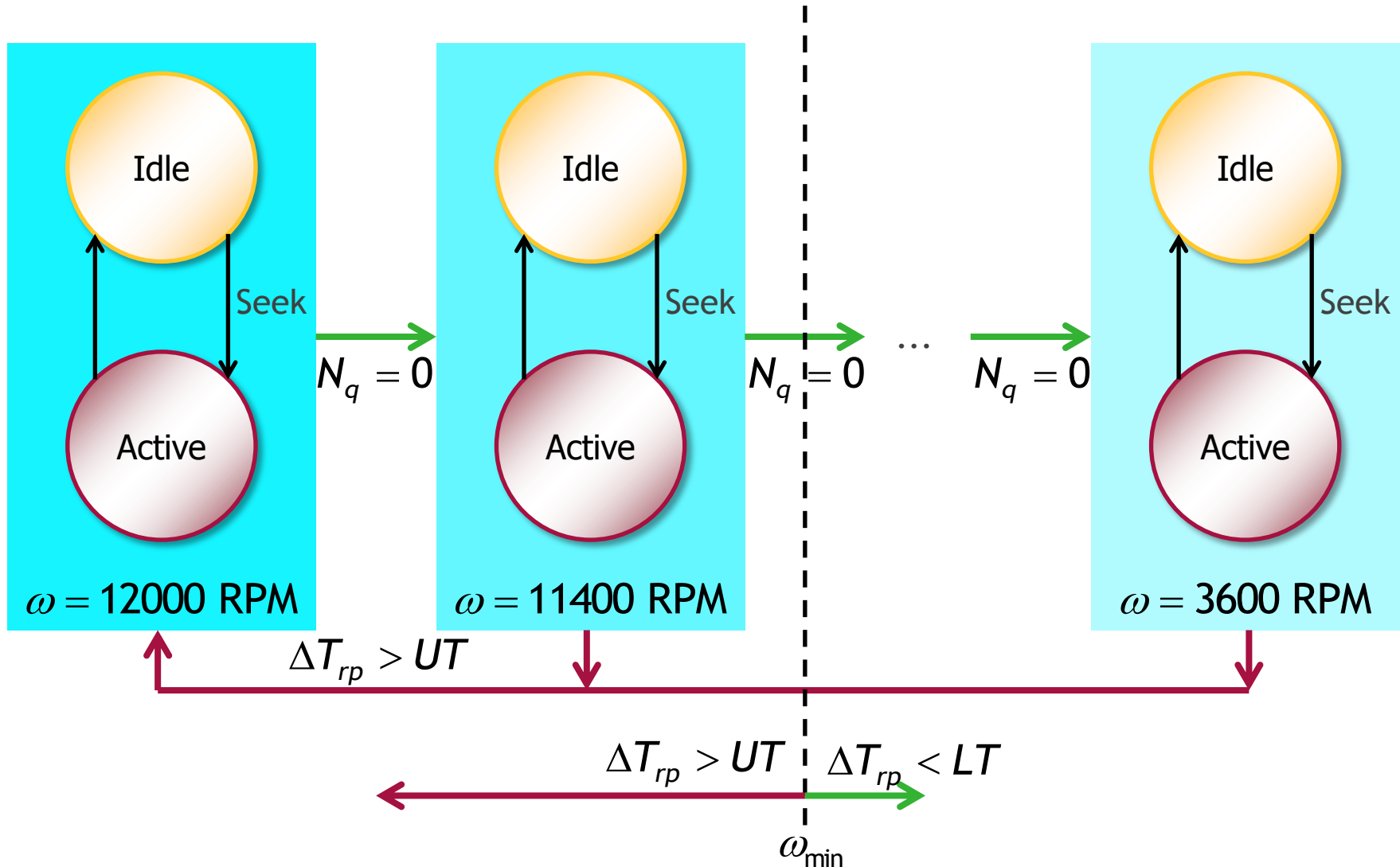
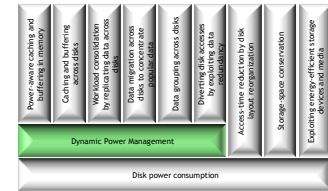
### BUT IN ENTERPRISE ENVIRONMENT

- IDLE TIME TOO SHORT
- SPIN-UP DELAY IN MOST CASES UNACCEPTABLE
- DISK DUTY-CYCLE RATING LIMITED

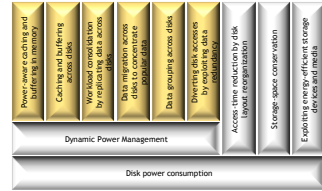


# DYNAMIC POWER MANAGEMENT

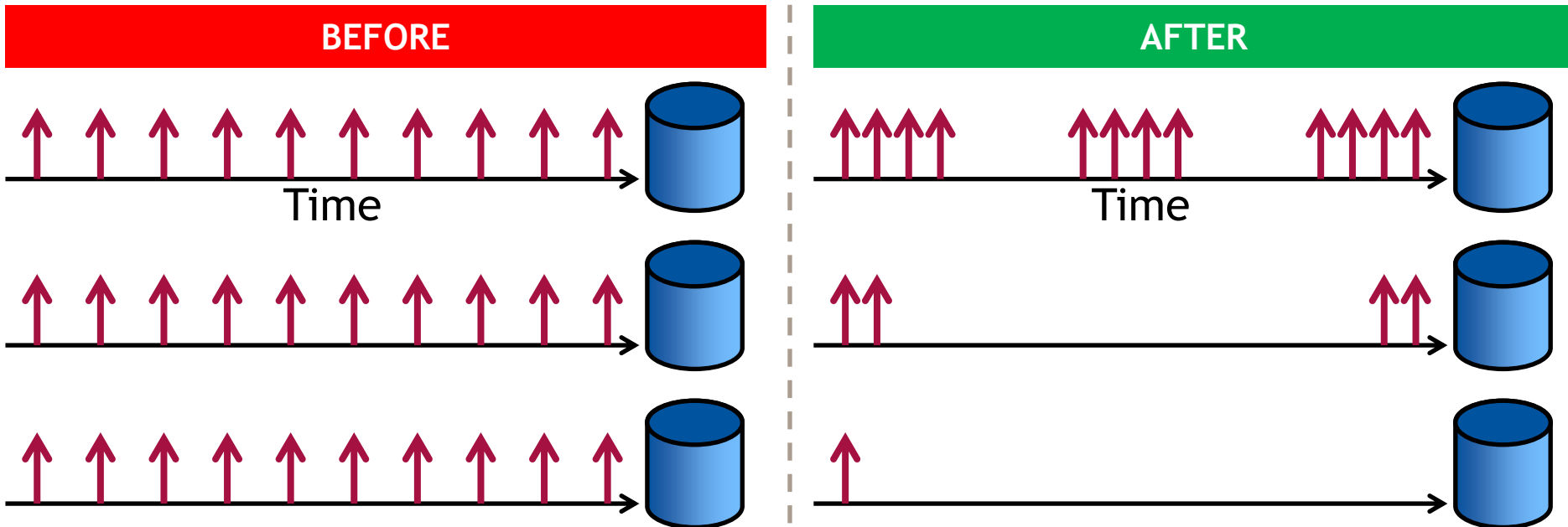
## MULTISPEED DISK = DYNAMIC ROTATIONS PER MIN



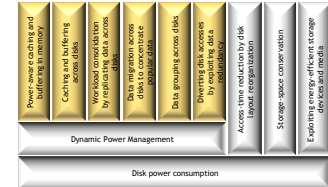
# DPM-ENABLING WORKLOAD SHAPING OBJECTIVE



- In enterprise environment, idle time too short to spin down disk
- Alternative solution to multispeed disk is workload shaping to
  - Increase mean idle time
  - Increase idle time variance over time and across disks



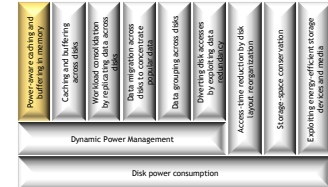
# DPM-ENABLING WORKLOAD SHAPING CLASSES OF TECHNIQUES



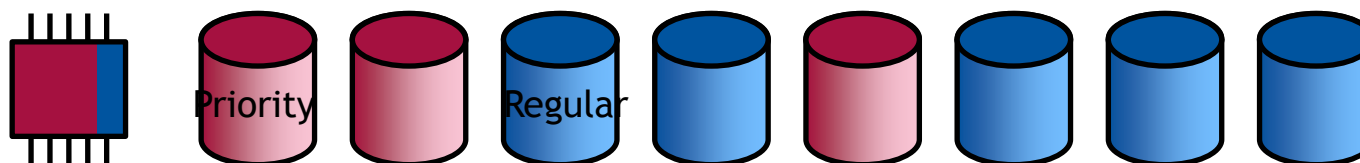
Power-aware caching and buffering in memory	
Caching and buffering across disks	
Workload consolidation by replicating data across disks	
Popular data concentration by migrating data across disks	
Data grouping across disks	
Diverting disk accesses by exploiting data redundancy	



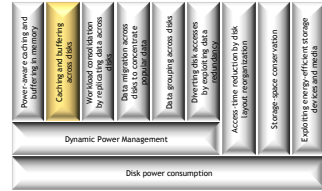
# POWER-AWARE CACHING AND BUFFERING IN MEMORY



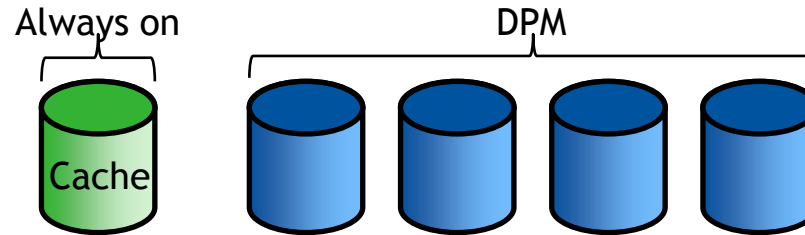
- Power-aware cache-replacement policies
  - Traditional cache-replacement policies minimize number of cache misses but don't consider distribution of such misses over time or across disks
  - Power-aware cache-replacement policies trade cache hit rate for energy savings
  - E.g. PA-LRU makes distinction between priority and regular disks
    - Priority disks: large percentage of long idle periods and high ratio of capacity misses to cold misses



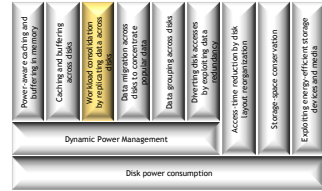
# CACHING AND BUFFERING ACROSS DISKS



- Massive Array of Idle Disks (MAID)
  - Replacement of tape libraries for archival storage
  - RAID's performance and dependability not required
  - Cache disks also buffer writes

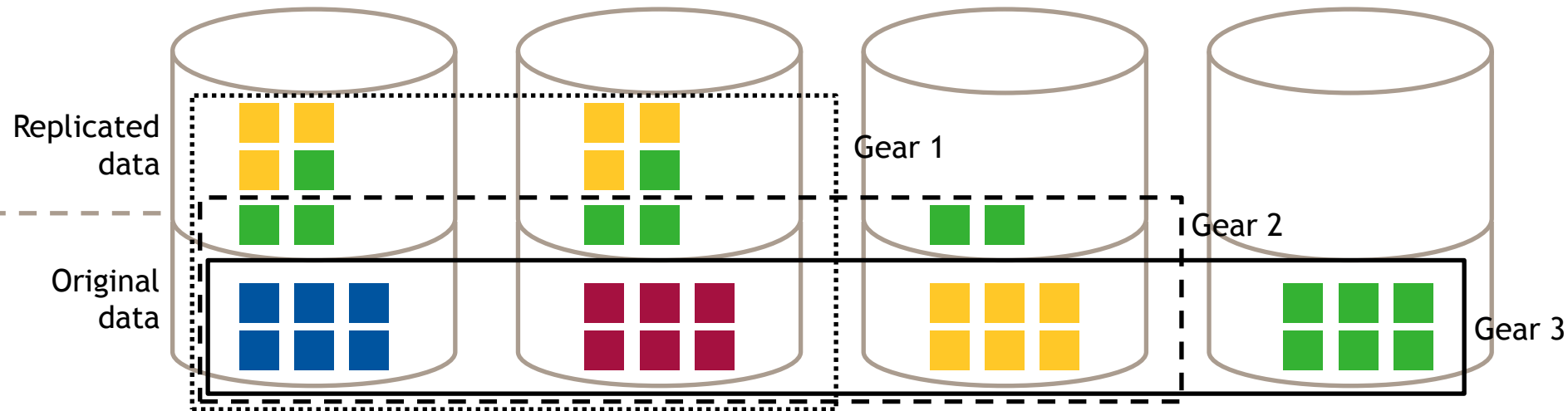


# WORKLOAD CONSOLIDATION BY REPLICATING DATA ACROSS DISKS

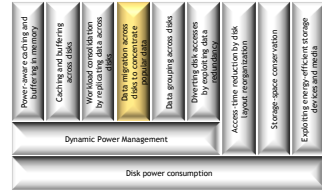


- Power-Aware RAID

- Exploits cyclic load fluctuations and unused storage space
- Load-directed power control
- Data replication results in skewed striping pattern



# POPULAR DATA CONCENTRATION

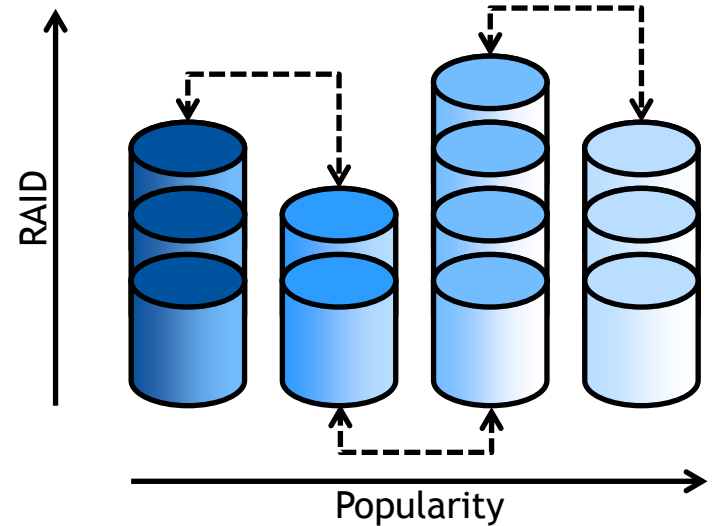


- Popular data concentration

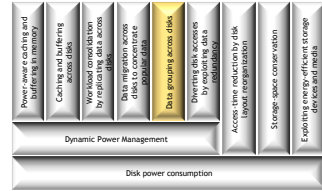
- File access frequencies follow Zipf distribution:  $p_i \propto 1/r_i$
- Files are placed across disks according to their access frequency
- Disks storing most popular files may not be filled to capacity to avoid disk contention

- Combined with RAID and DRPM: Hibernator

- Speed of disks in array periodically adapted
- Continuous small-scale reorganization: blocks are migrated across all disks according to their access frequency
- Large-scale reorganization upon disk migration: blocks are migrated across disks of their tier to even out average access frequency across disks

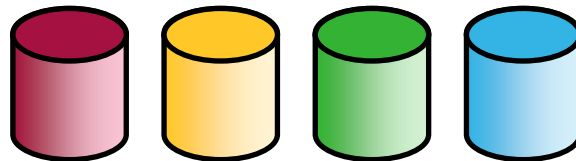


# DATA GROUPING ACROSS DISKS

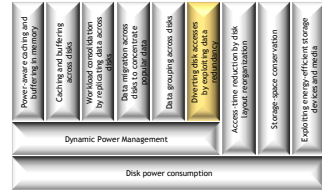


- Semantic data placement

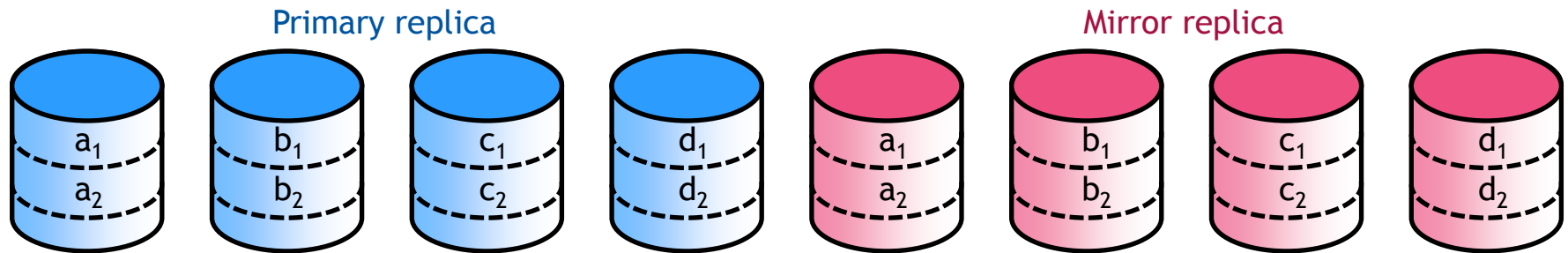
- Archival-by-accident workload: write-once/read-maybe except for changing hot area that exhibits large number of reads and overwrites
- Similar data are temporally related
- Data are grouped across disks according to semantic and incidental labels
- E.g. time stamp, file-system placement, author, file type,...



# DIVERTING DISK ACCESSES BY EXPLOITING DATA REDUNDANCY

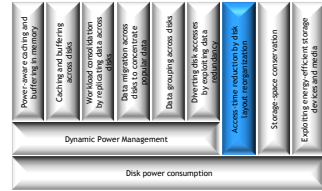


- Diverted accesses (DIV): separate redundant from original data and turn off disks storing redundant data when load conditions allow
- Application of DIV to RAID: EERAID
  - RAID-1: windows round-robin policy for dispatching reads and power and redundancy-aware flush policy for buffering writes

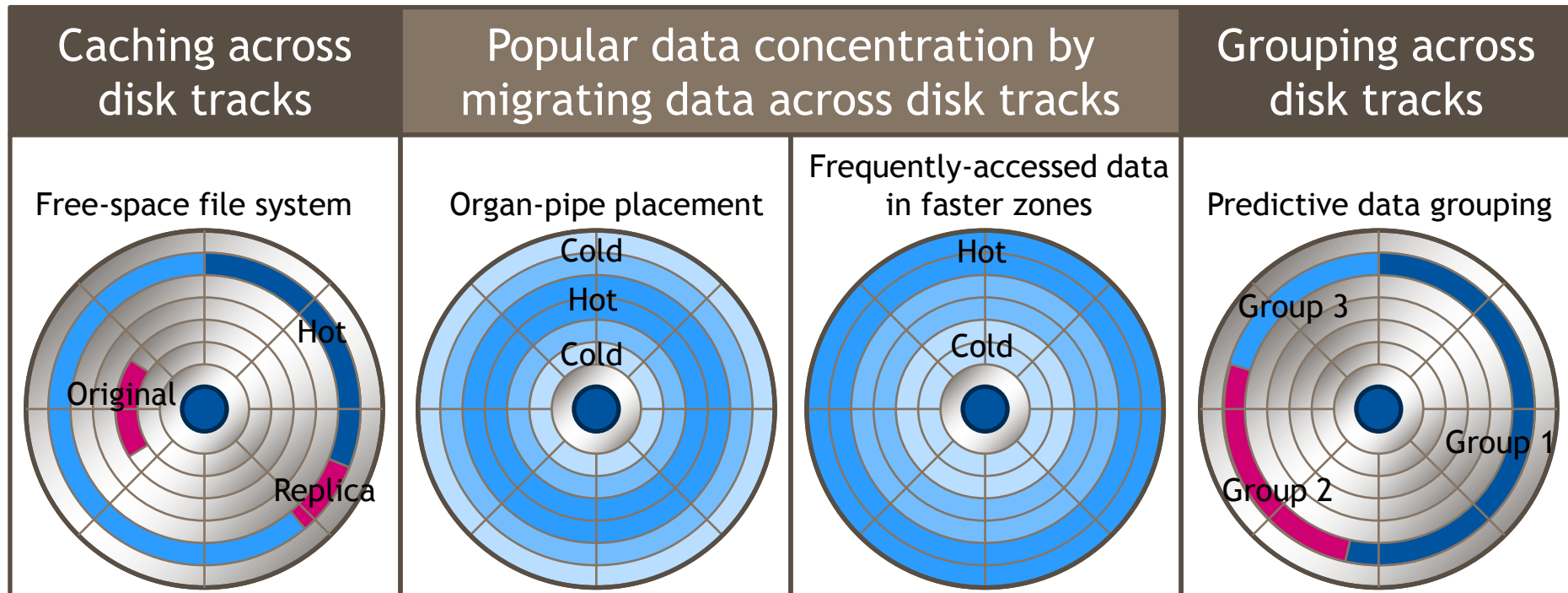


- RAID-5: transformable read policy for dispatching reads and power and redundancy-aware destage policy for buffering writes

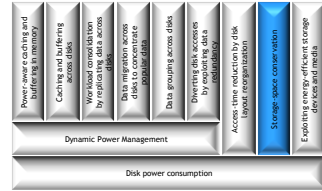
# ACCESS-TIME REDUCTION BY DISK-LAYOUT REORGANIZATION



- Saving energy while disk is active
- Reduction of average seek distance and rotational latency by improved I/O scheduling or reorganization of the disk layout



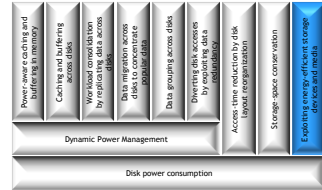
# STORAGE-SPACE CONSERVATION



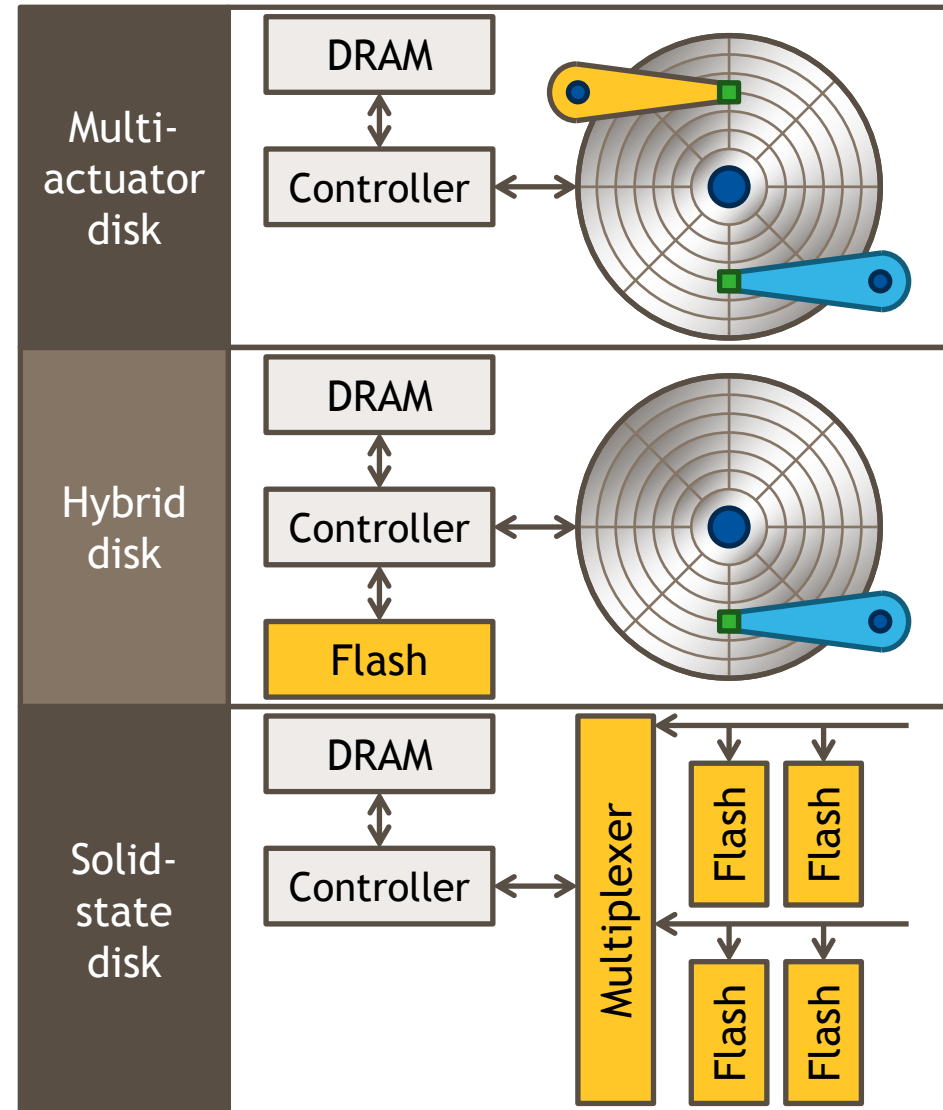
- By eliminating unnecessary redundancy, data can be stored on fewer disks and reading and writing data requires less time
- Space-conservation techniques that save energy on the side
- Redundancy elimination as opposed to exploitation in diverted accesses
- Opposite of replication-based energy-saving techniques
- Mainly for archival storage because size-intensive data rather than load-intensive
- Data compression and data deduplication



# EXPLOITING ENERGY-EFFICIENT STORAGE DEVICES AND MEDIA



- Multiactuator disk
  - Number of disks determined by requirements of performance rather than capacity
  - Avoid disk-space waste by improving performance by intradisk parallelism:  $D_k A_l S_m H_n$
- Hybrid disk
  - Flash serves as second-level cache
  - Similar as caching and buffering in memory but using low-power, non-volatile flash memory
- Solid-state disk
  - No spin or seek power
  - Higher throughput per Watt than HDD but similar or less capacity per Watt than HDD



# CLASSIFICATION OF POWER-REDUCTION TECHNIQUES ACCORDING TO PERFORMANCE IMPACT

Impact	Cause	Example
	Disk spin-up required to serve I/O request if disk is spun down	Traditional power management
	Fewer disk spin-ups required to serve I/O requests than for plain TPM	Write off-loading
	Disk spin-ups only for increasing bandwidth to accomodate higher load	Power-aware RAID
	Only disk speed-ups for increasing bandwidth but requests served at lower speed under lighter load	Dynamic rotations per minute
	Computational overhead	Data compression
	Reduced seek and rotational latency	Free-space file system
	Substitution of disk access by DRAM or SSD access	Solid-state disk

# CLASSIFICATION OF POWER-REDUCTION TECHNIQUES ACCORDING TO DEPENDABILITY IMPACT

Impact	Cause	Example
	Disk spin-up required to serve I/O request if disk is spun down	Traditional power management
	Fewer disk spin-ups required to serve I/O requests than for plain TPM	Write off-loading
	Disk spin-ups or speed-ups only for increasing bandwidth to accomodate higher load	Power-aware RAID
	Volatile memory used for buffering	Power-aware buffering
	Number of erase cycles of NAND flash memory may exceed erase cycle rating for write-intensive workload	Solid-state disk
	No impact	Free-space file system

# CLASSIFICATION OF POWER-REDUCTION TECHNIQUES ACCORDING TO CAPACITY IMPACT

Impact	Cause	Example
	Decreased capacity utilization to avoid disk contention	Popular data concentration
	Redundancy addition (more than working set)	Power-aware RAID
	Working-set replication	Massive array of idle disks
	No impact	Diverted accesses
	Redundancy elimination	Data deduplication
	Increased capacity utilization	Multiactuator disk

# ULTIMATE POWER-AWARE ENTERPRISE STORAGE SYSTEM

- Integrates different classes of power-reduction techniques
  - Dynamic power management
  - DPM-enabling workload shaping
  - Access-time reduction by disk lay-out reorganization
  - Storage-space conservation
  - Energy-efficient storage devices and media
- Covering all of the storage-stack layers
- Addressing diverse workloads
- Offering flexible trade-off between power consumption, performance, capacity, and dependability



(1)



(2)

(1) © [ITP.net](http://ITP.net)  
(2) © [Western Digital Technologies, Inc.](http://Western Digital Technologies, Inc.)

[www.alcatel-lucent.com](http://www.alcatel-lucent.com)