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

Tom Bostoen
December 16, 2013

• 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
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
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

- 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

Other

- Access-time reduction by disk layout reorganization
- Storage-space conservation
- Exploiting energy-efficient storage devices and media

Dynamic Power Management

Disk power consumption
HARD DISK DRIVE

- Sector
- Track
- Spindle
- Read/write head
- Arm
- Cylinder
- Platter
- Surfaces

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 ~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
  
  \[ E_{id}(T_{id}) = P_{id}T_{id} \]

• Option 2: spin disk down
  
  \[ E_{sb}(T_{id}) = P_{sb}T_{id} + E'_{du} \]

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

\[
\begin{align*}
\text{Idle} & \quad P_{id} \\
\text{Active} & \quad P_{act} \\
\text{Standby} & \quad P_{sb} \\
\end{align*}
\]

\[
\begin{align*}
\text{Idle} & \quad P_{id} \\
\text{Seek} & \quad P_{sk} \\
\text{Active} & \quad P_{act} \\
\text{Spin up} & \quad E_{up} \\
\text{Spin down} & \quad E_{dn} \\
\end{align*}
\]
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

idle

Active

\[ \omega = 12000 \text{ RPM} \]

\[ N_q = 0 \]

\[ \Delta T_{rp} > UT \]

\[ \omega_{\text{min}} \]

Active

\[ \omega = 11400 \text{ RPM} \]

\[ N_q = 0 \]

\[ \Delta T_{rp} > UT \]

\[ \Delta T_{rp} < LT \]

Active

\[ \omega = 3600 \text{ RPM} \]

Seek

Seek

Seek

\[ \omega = 11400 \text{ RPM} \]

\[ \omega = 12000 \text{ RPM} \]

\[ \omega_{\text{min}} \]
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

BEFORE

AFTER
## DPM-ENABLING WORKLOAD SHAPING
### CLASSES OF TECHNIQUES

<table>
<thead>
<tr>
<th>Class</th>
<th>Diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power-aware caching and buffering in memory</td>
<td><img src="cache_diagram.png" alt="Cache Diagram" /></td>
</tr>
<tr>
<td>Caching and buffering across disks</td>
<td><img src="cache_diagram.png" alt="Cache Diagram" /></td>
</tr>
<tr>
<td>Workload consolidation by replicating data across disks</td>
<td><img src="replication_diagram.png" alt="1/3, 1/3, 1/3, 1" /></td>
</tr>
<tr>
<td>Popular data concentration by migrating data across disks</td>
<td><img src="data_concentration_diagram.png" alt="Popular, Unpopular" /></td>
</tr>
<tr>
<td>Data grouping across disks</td>
<td><img src="data_grouping_diagram.png" alt="Data Grouping Diagram" /></td>
</tr>
<tr>
<td>Diverting disk accesses by exploiting data redundancy</td>
<td><img src="redundancy_diagram.png" alt="Redundant" /></td>
</tr>
</tbody>
</table>
• 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
  - 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,…

![Diagram of data grouping across disks]
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
• 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 parallellism: \( D_k A_l S_m H_n \)

- 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

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

<table>
<thead>
<tr>
<th>Impact</th>
<th>Cause</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced disk reliability</td>
<td>Disk spin-up required to serve I/O request if disk is spun down</td>
<td>Traditional power management</td>
</tr>
<tr>
<td>Limited reduction of disk/data reliability</td>
<td>Fewer disk spin-ups required to serve I/O requests than for plain TPM</td>
<td>Write off-loading</td>
</tr>
<tr>
<td></td>
<td>Disk spin-ups or speed-ups only for increasing bandwidth to accommodate higher load</td>
<td>Power-aware RAID</td>
</tr>
<tr>
<td></td>
<td>Volatile memory used for buffering</td>
<td>Power-aware buffering</td>
</tr>
<tr>
<td></td>
<td>Number of erase cycles of NAND flash memory may exceed erase cycle rating for write-intensive workload</td>
<td>Solid-state disk</td>
</tr>
<tr>
<td>No impact</td>
<td>No impact</td>
<td>Free-space file system</td>
</tr>
</tbody>
</table>
## Classification of Power-Reduction Techniques According to Capacity Impact

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

**Legend:**
- Much more space required
- More space required
- No impact
- Less space required
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