Saving Power Consumption of Dependable Storages in a Cloud Center

Haruo Yokota
Tokyo Institute of Technology
yokota@cs.titech.ac.jp

Outline

1. Increase of data & power in a cloud center
2. Approaches for saving power of the center
3. Focus on power & reliability of a storage system with HDD properties
4. Our approach of adopting an asynchronous-update primary-backup configuration
5. Comparison with MAID (Massive Arrays of Idle Disks) modified to keep its reliability

Data Amount and Power Consumption

- The amount of data is increasing rapidly
  - Created and replicated data surpassed **2.7 ZB** in 2012
- Data is moving from on-premises to cloud
  - Large scale storage systems are required in a cloud center
  - Massive disk drives consume huge power
- It is important to save power consumption of storage systems

Approaches for Saving Power

- There are many approaches for saving the power consumption of a cloud center
  - Consideration of cooling the systems
    - Control air flow for cooling
    - Container center locating in a cold district
  - DC (Direct Current) power supply
    - To reduce AC/DC and DC/AC conversion loss in UPS
  - Stop parts of circuits, servers, racks, and so on

- Focus on storage systems for handling the large amount of data

Approaches for Saving Power of a Storage System

- Using Non-volatile RAMs
  – Flash Memory
  – MRAM (Magneto-resistive RAM)
  – ReRAM (Resistance RAM)
  – FeRAM (Ferroelectric RAM)
  – PRMA (Phase-change RAM)
  – …
- They are still expensive or not yet available for storing the large amount of data
  – We still need to assume the use of HDDs
  – We can use the non-volatile RAMs as cache

Approaches for Saving Power of Storage with HDDs

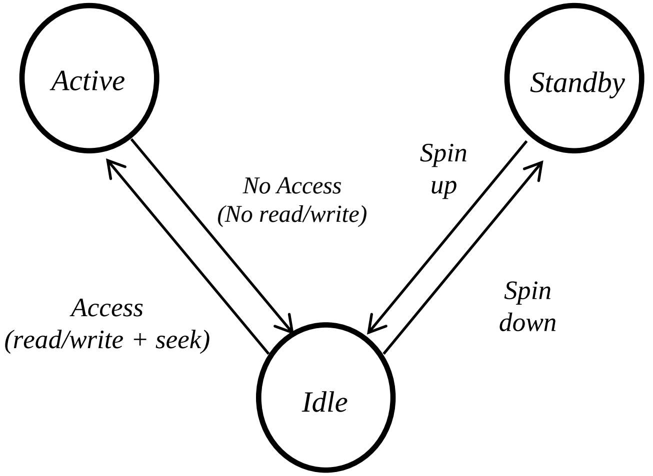
- Stop rotation of HDDs
  - Because the spindle motor occupies a large part of the power consumption of a disk drive

- Problems
  - Large power is consumed during spin-up or spin-down periods
  - Disk access is delayed until disk rotation speed reaches the maximum during spin-up periods

HDD States and Power Consumption

- An HDD has three states

State	I/O	Rotation (RPM)	Head Possision	Power Consum.
Active	Processing	Max Speed	On Platters	Large (P_{active})
Idle	No	Max Speed	On Platters	Middle (P_{idle})
Standby	No	0	Outside of Platters	Small ($P_{standby}$)

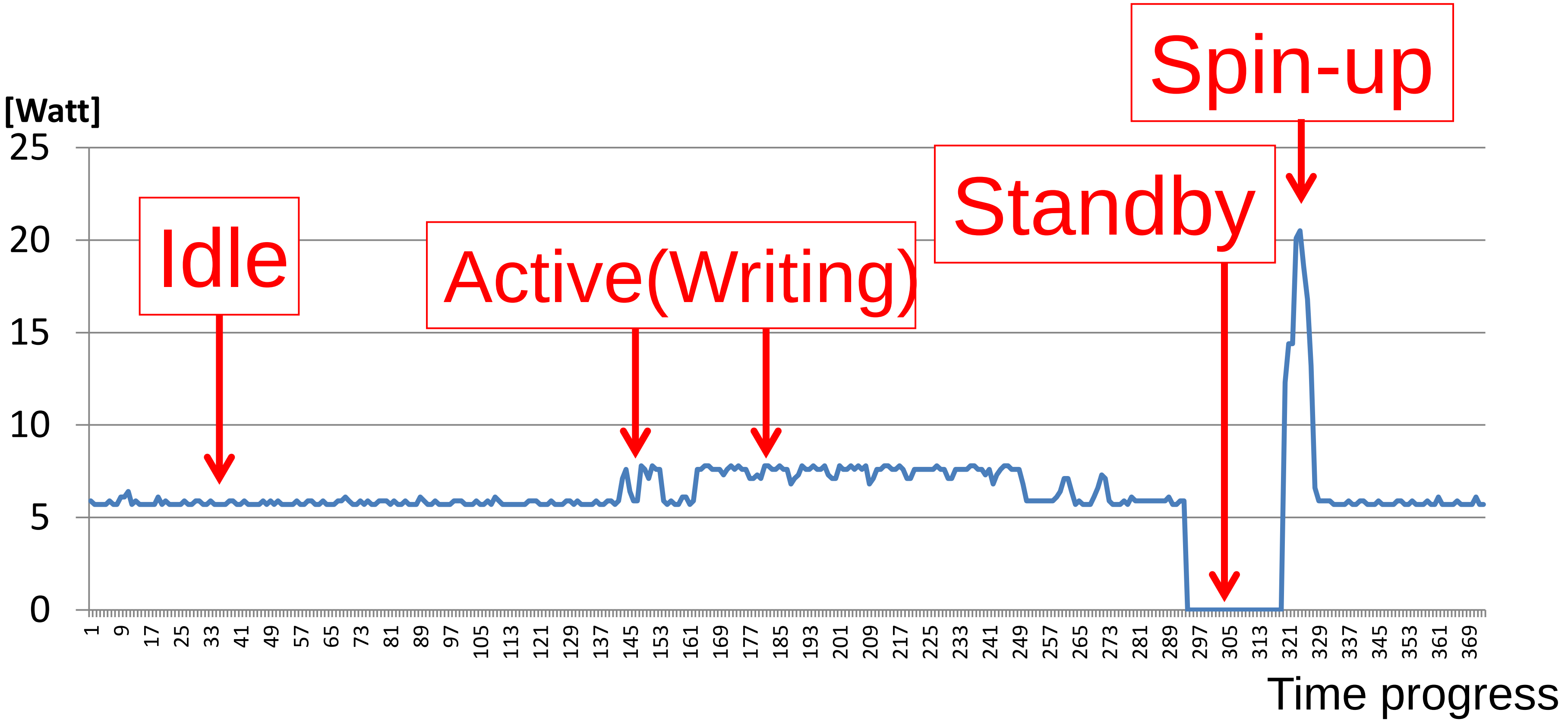


- $P_{active} > P_{idle} >> P_{standby}$

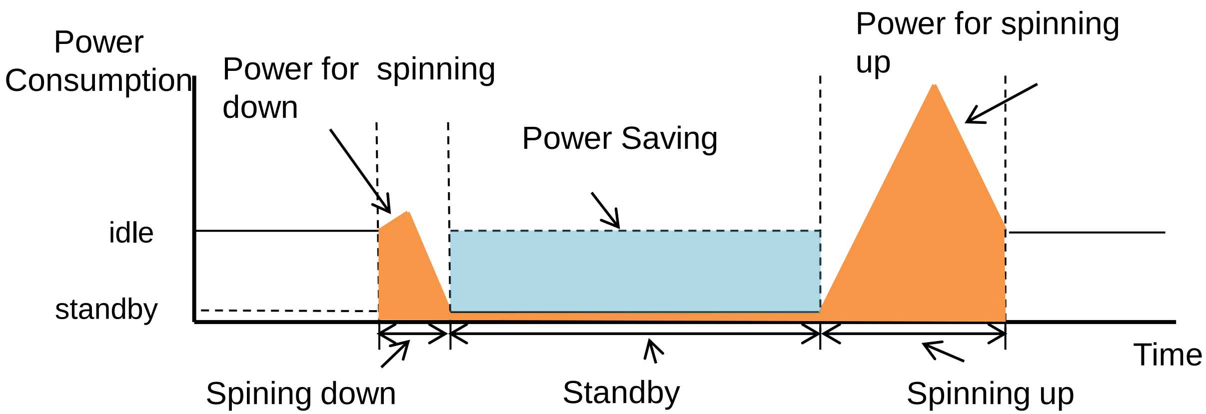
- It also consumes large power to spin up the HDD (to move from *Standby* state to *Idle* state)

Actual Power Consumption

- Measured HGST Deskstar 7K2000（2 TB）in our lab.



Break-even Time



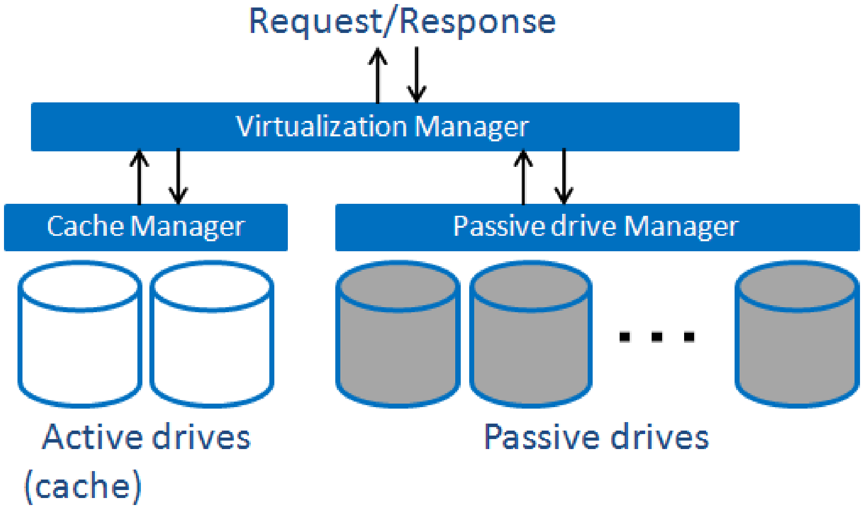
- If the power reduction by the difference in power of the standby and idle state exceeds the power for spinning up/down and standby state, the spinning down is effective. Otherwise, ineffective.

- It means the frequency of spin-down/up is important
  - To keep enough long standby time

A Well-known Proposal

- MAID (Massive Arrays of Idle Disks)[Colarelli, 2002]
  - Keeps a small number of disk drives rotating as cache disks
  - Many other idle drives are spin-downed
  - It is effective when access patterns have locality

- However, MAID does not consider its reliability



Reliability and Power Consumption

- To make a large storage system reliable
  - Straightforward approaches increase its power consumption
- RAIDs: increase the frequency of spin down/up
  - RAID 1: Mirroring
    - Access two disks simultaneously
  - RAID 4-6: Parity calculation approaches
    - Read-Modify-Write for both data and parity disk for each access

Our Goal and Approaches

- Goal
  - To save power consumption of a storage system with keeping its reliability and improving its performance

- Approaches
  - Employ asynchronous-update primary-backup configuration
    - To ensure the reliability of a storage system
  - Consider individual disk rotation
    - Choose suitable disks to serve requests in terms of energy efficiency

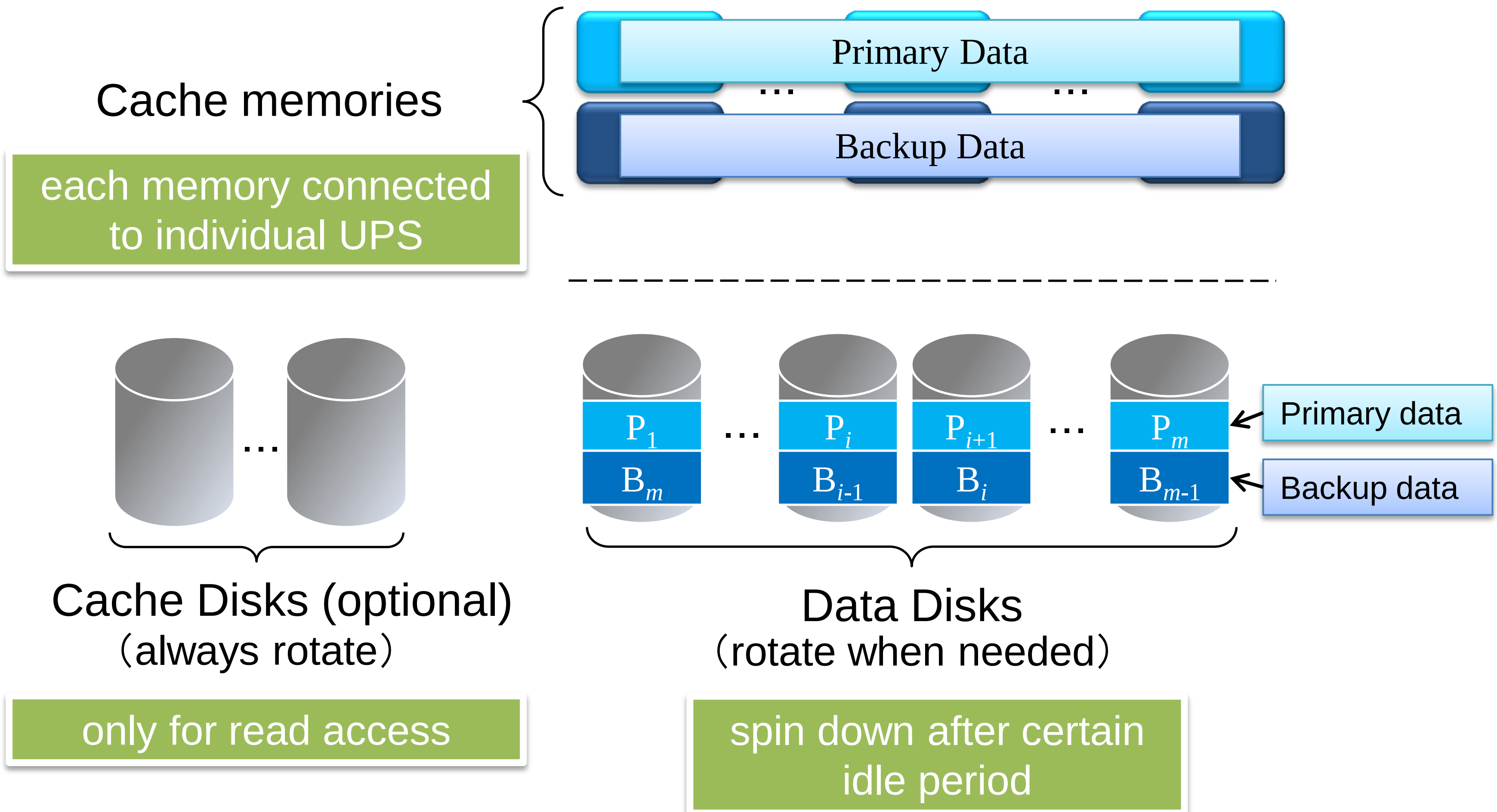
Other Related Proposals

- EERAID1/5 (Energy Effective RAID)[Li, 2004] introduces windows round-robin dispatch
  - Not take care of rotation state of each disk
- GRAID (Green RAID) [Mao, 2008] uses a dedicated log disk for RAID10
  - Access two disks simultaneously (mirroring)
- There are several power proportional approaches such as PARAID (Power-Aware RAID) [Weddle, 2007]
  - Power proportional data placement methods for HDFS are also proposed such as RABBIT[Hrishikesh, 2010]

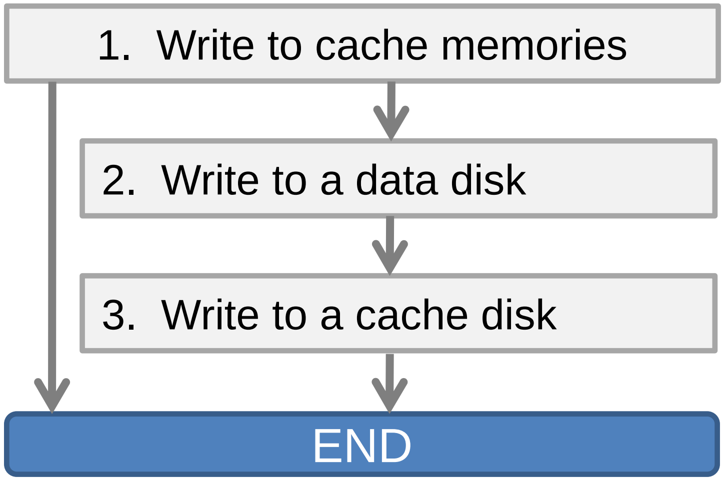
Our Basic Strategy

- Control the timing of spinning up/down of each disk by the asynchronous property
  - Different from the mirroring
- Group a number of write requests
  - Using RAM with buttery backup (UPS) or non-volatile RAM to keep write requests in cache
- The cache should also be reliable
  - Make the cache primary and backup, too

RAPoSDA
(Replica-Assisted Power Saving Disk Array)



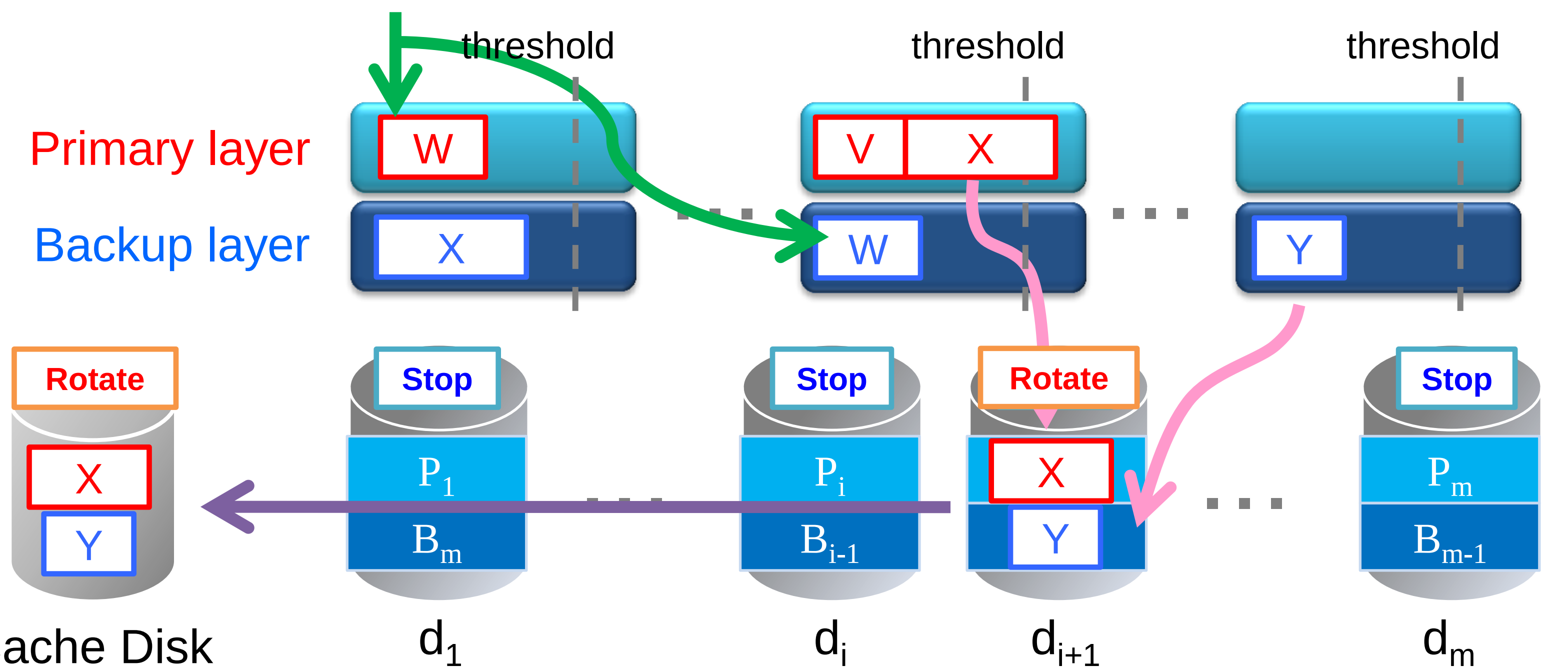
Handling write requests



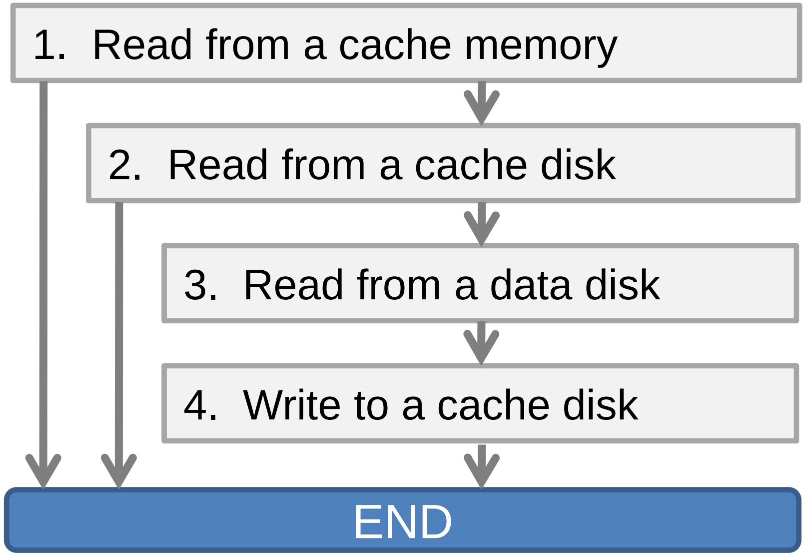
- Write primary and backup data
- If the data on the memory over the buffer threshold
- Write data which should be stored in primary or backup area on the target disk drive
- Copy the data to a cache disk (group writing)

Write requests:

W	V	X

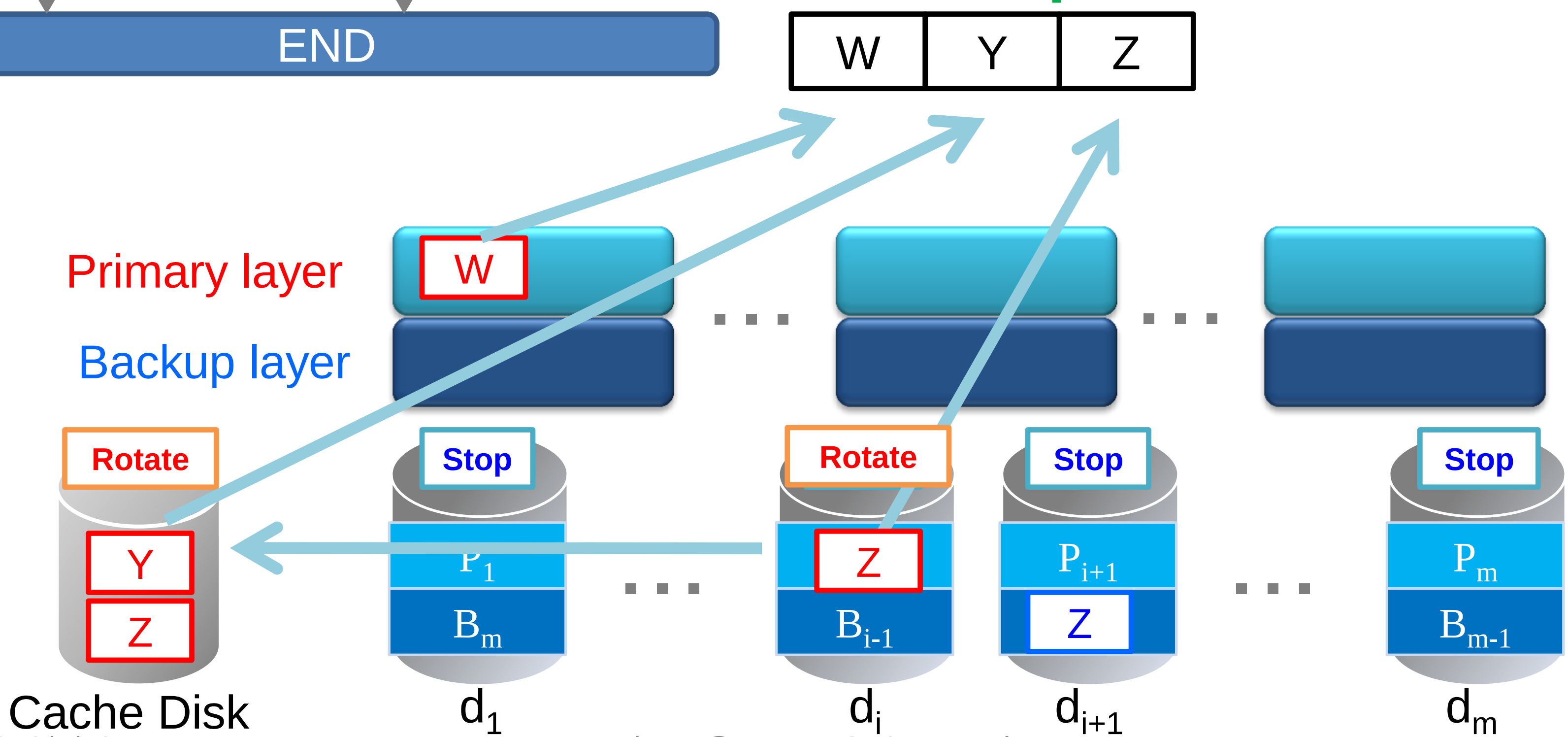


Handling read requests



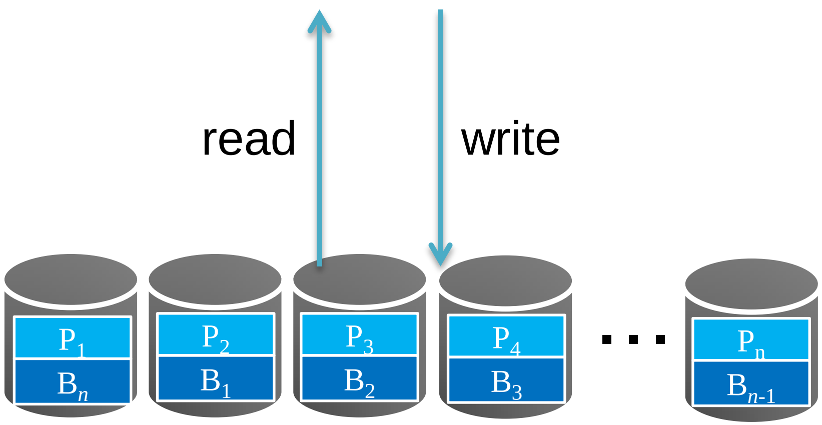
Choose an appropriate disk by considering rotation states
(i.e. rotating, stopping or length of standby periods)

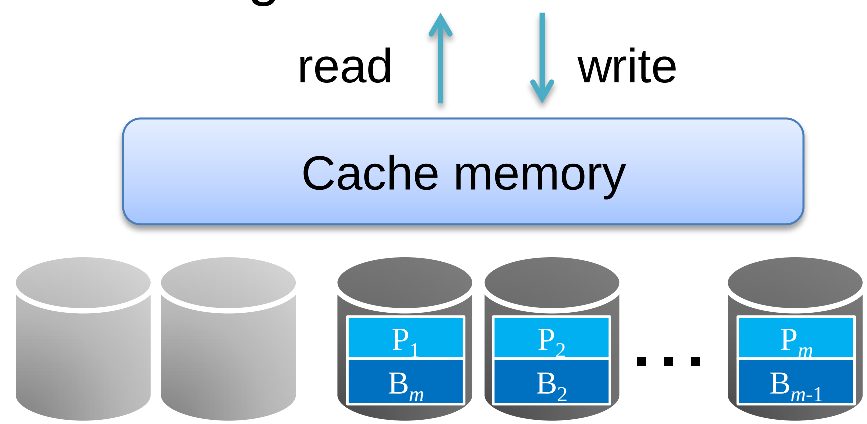
Read requests:



Comparative Configurations

- Normal
  - Only data disks
  - Never spin-down
  - Disks employ a primary backup configuration
- MAID (modified)
  - Includes cache memory
    - Write-through, no replication
  - Data disks employ a primary backup configuration





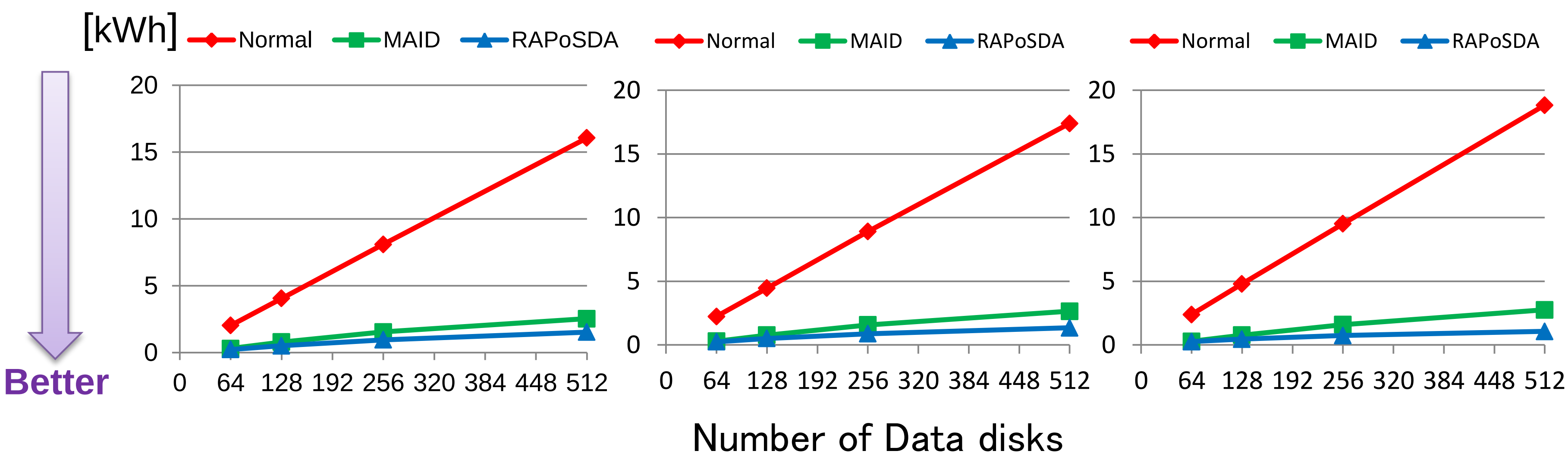
- Both systems randomly choose one of the replicated disks when a disk access is requested

Simulation parameters

- Synthetic workload（**5** hours）
  - Access skew: Zipf distribution
  - Arrival rate: Poisson distribution (**25** req/s)
  - read:write ratio: **7:3**, **5:5**, **3:7**
  - The number of files: **1,000,000**（**32** KB/file）
- Disk model: HGST Deskstar 7K2000（2 TB）
- Storage system parameters
  - The number of data disks: **64-512**
  - The number of cache disks in MAID/RAPoSDA: **10%**
  - Capacity of total cache memory: **#disk/4 (16-128)** GB

Power Consumption with Changing the number of disks

- Normal consumes power since it rotates all disks
- Difference between MAID and RAPoSDA comes from the asynchronous manner with considering the rotation state



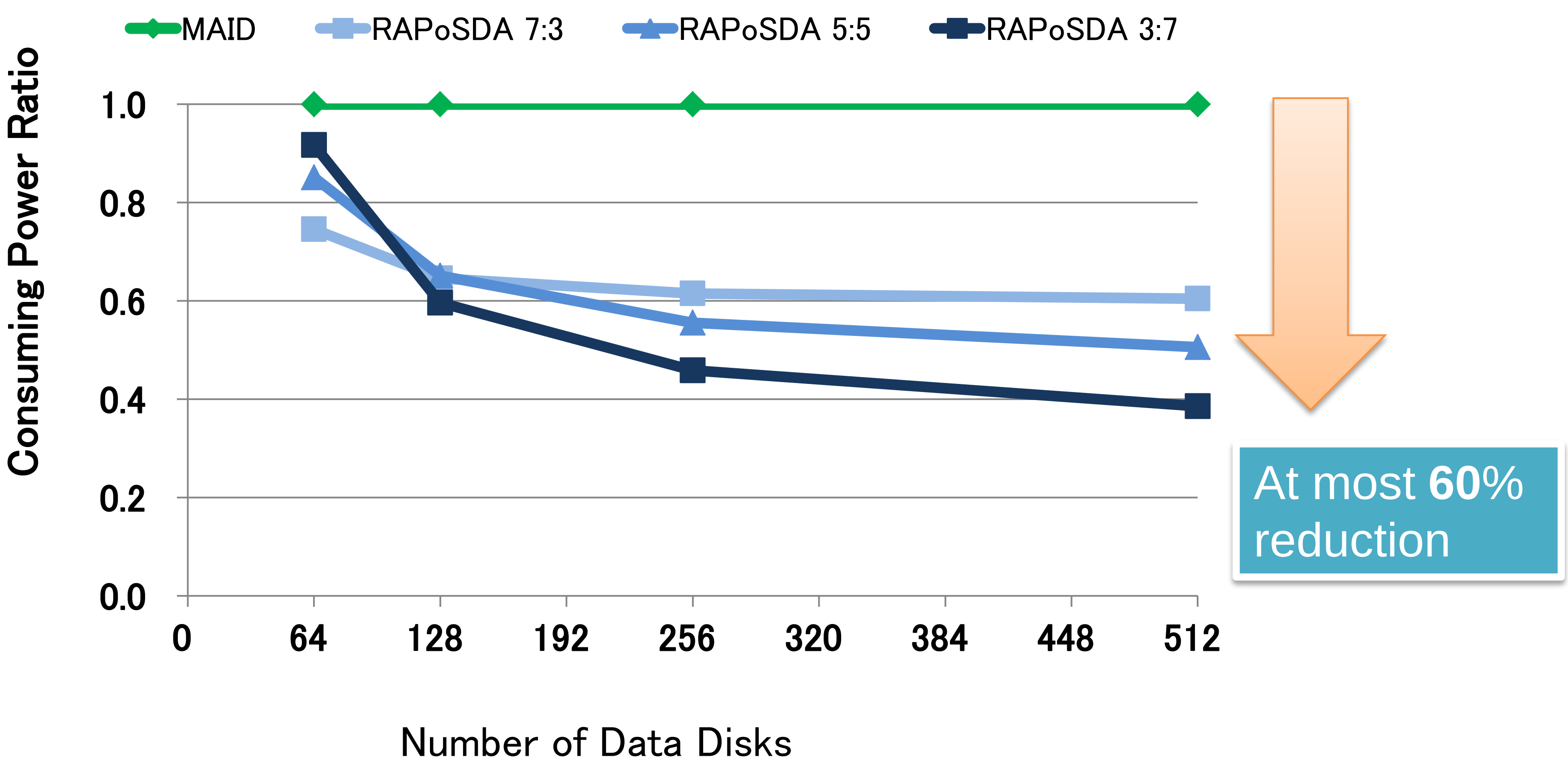
read : write

7:3

5:5

3:7

Power Reduction Ratio RAPoSDA vs MAID

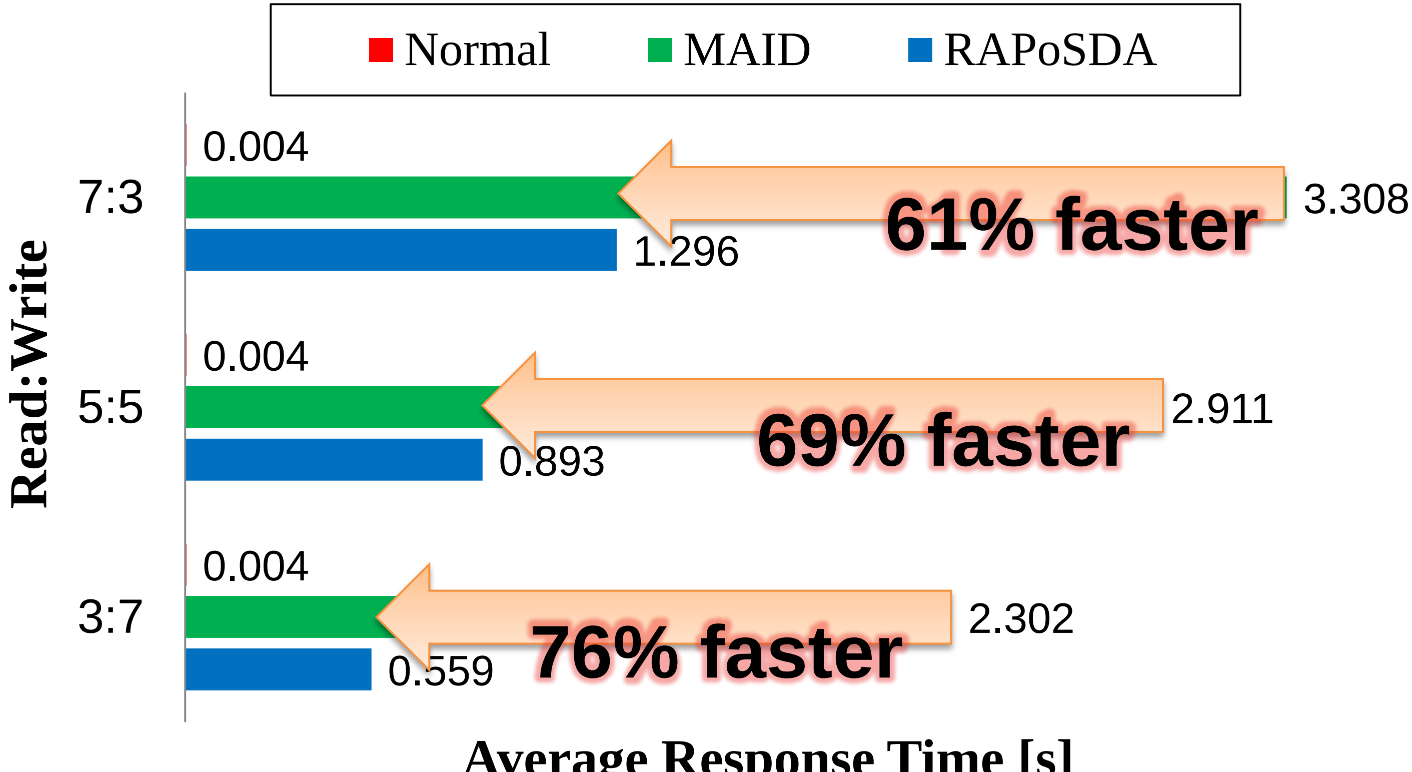


Reasoning of Power Reduction

- Group writing of RAPoSDA worked efficiently
  - Conserved unnecessary disk spin-ups/spin-downs by simultaneously writing primary and backup data
- MAID could not take long enough standby period to achieve power reduction
  - Due to random accesses in particularly heavy write workloads
  - It is difficult to transfer data disks to standby state due to the write-through policy at the cache memory

Average response times

- Normal is the fastest, but also the largest power consumer
- RAPoSDA is at most 76% faster than MAID by considering each disk rotation



Conclusion

- We propose RAPoSDA (Replica-Assisted Power Saving Disk Array)
  - Considers individual disk rotation states to achieve effective disk accesses
  - Ensures reliability by a primary-backup configuration in both cache memory and data disks
- Evaluated and compared the power reduction ratios and performance of three configurations
  - RAPoSDA provides superior power reduction and a shorter average response time compared with MAID

Working on

- Dynamic adjustment of cache threshold for P&B
- Efficient mapping between disks and caches
- Handling more than two replicas
  - Such as HDFS, GFS

- Applying the asynchronous update primary-backup (AUPB) configuration to manage security in a storage system
  - Reduction of re-encryption time for revocation
- Applying the AUPB configuration for keeping QoS of a storage system
  - Efficient data migration for handling access skews