

Weilin Peng  
Sept. 28<sup>th</sup> 2009

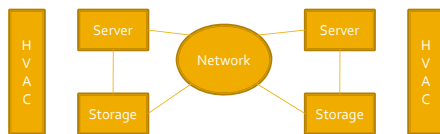
## Storage Power Management

## OUTLINE

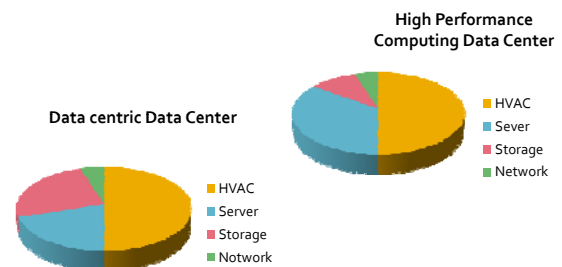
- 1. Introduction
- 2. Existing Methods
- 3. Future Research Issues
- 4. Existing works
- 5. My Research plan

## INTRODUCTION

- What is Data Center
  - Concentrated clusters of compute and data storage resources that are connected via high-speed networks and routers.



## ENERGY COST cont.



## STORAGE FACILITIES

- High-end SCSI/SATA drivers operating at speed of 10-15K RPM
- Amount from a few tens to several thousands
- Disks are always connected with network
- Disk operational states:
  - Active
  - Idle
  - Standby/Sleep
  - OFF

## DISK POWER CONSUMPTION

- Three parts in disk consume power:
  - Rotating Spindle
  - Head assembly
  - Buffers
- Rotating Spindle consumed majority of power

$$P_{spindle} \propto \omega^2$$

$\omega$ : angular velocity

## DISK POWER CONSUPTION

- For a typical SCSI disks rotating at 15K rpm
  - Power cost by spindle: 5-6 W
  - Power cost by head assembly: 2-4W
  - Array controller, Enclosures and Network : 1-2 W
- Totally power consumption
  - Active state: 8-11 W
  - Idle state: 6-7 W
  - Standby state: 1-2 W

## RESEARCH FOCUS

- Adaptive spin down or sleep transitioning
  - Spindle motor consumes much energy
- Dynamic modulation of rotation speed
  - Remember DVFS?
- Caching aid
  - For enterprise storage solutions
- New media types
  - Flash and solid state media

## OUTLINE

1. Introduction
2. Existing Methods
3. Future Research Issues
4. Existing works
5. My Research plan

## DYNAMIC ROTATION CONTROL

- Assumption: Disks can change its rotational speeds on the fly
- Idea is the same as DVFS
  - Response time is faster than threshold is a waste
  - Limit the waste
- Result
  - The idea can save the power up to 60% IF multiple speeds disk appear

## ALTERNATIVES APPROACHES

- Pinheiro et al. study on the following four alternatives approaches:
  - Powering down during idle periods
  - Replacing high performance SCSI disks with a set of lower power disks
  - Combine SCSI and laptop disk such that only one is active/ON
  - Multi-speed disks
- Result: only No. 4 can save energy

## OPTIMIZING DATA LAYOUTS

- Basic idea: Control disk accesses
- PDC (Popular Data Concentration)
  - Classify the data based on file popularity
  - Migrate the most popular file to a subset of disks
  - Maximize the idle time
  - Accessing to an unpopular file takes more 8-12 seconds

## MAID

- Massive Array of Idle Disks
  - Copy files based on their temporal locality
  - Used a subset of disk as cache
  - Suffer from the long access time of unpopular files
  - Useful for online library or backup solutions
- Improvement
  - GreenStor
  - Cache replacement algorithm

## READ/WRITE CACHE

- Reading/writing cache in each disk
- Use these cache efficiently to increase idle time
- Make improvement on cache replacement algorithm: LRU (Least Recently Used)

## PALRU & PBLRU

- PALRU: Partition Aware LRU
  - Classify all disks into two classes: Priority and Regular
  - Maintain two LRU queues. Always replace the first element in Regular queue till it is empty.
- PBLRU: Partition Based LRU
  - Divided total cache into partitions and allocate them to each disks
  - Dynamic change the size of partitions base on the workload

## RAID ADAPTATIONS

- RAID(Redundant Array of Inexpensive Disk) is wildly used in enterprise storage solution
- RAID 0 to RAID 5 do not save energy but for reliability
- RAID controllers or engines are developed to minimize energy consumption of RAID disks

## NEW STORAGE MEDIA

- NAND-based flash memory
  - Small size, low weight, low power consumption high shock resistance and fast read performance
  - Phone, digital cameras and sensor devices
- Flash based Solid State Drives (SSDs)
  - Minimum R/W unit is page
  - Cache data between memory and hard disk
  - NAND-based flash: fast reading, slow writing
  - NOR-based flash: fast writing, slow writing

## OUTLINE

- 1. Introduction
- 2. Existing Methods
- 3. Future Research Issues
- 4. Existing works
- 5. My Research plan

## FUTURE RESEARCH ISSUE

- Disks using SATA interfaces
  - Low cost and less power consumption: 2-6 W
  - Tradeoffs between reliability, performance and power consumption
  - Mix type storage devices: STAT disks and SCSI with FC drives
- New type storage is always the hottest spot

## OUTLINE

- 1. Introduction
- 2. Existing Methods
- 3. Future Research Issues
- 4. Existing works
- 5. My Research plan

## GREENSTOR

- Paper Name: GreenStor: Application- Aided Energy-Efficient Storage
- Author: Nagapramod Mandagere, et al.
- Year: 2007
- Ultimate goal is that disks
  - Should be in Standby or OFF state as much as possible
  - The number of transitions between ON and OFF or Standby states should be minimum

## PRO & CON OF MAID

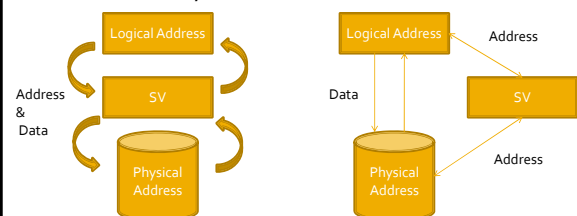
- Pro
  - Only 5% of data is popular and they are cached
  - Access to popular data is fast
  - A set of disks turn off means energy saved
- Con
  - It need to predict which 5% of data is popular
  - Access to uncached data has high performance penalty

## CHALLENGES

- Storage subsystem architecture to facilitate energy efficiency
  - Design a virtualized system to determine the physical placement of the virtual cache space
- Cache metadata management
  - Maintain something like lookup tables in memory
- Cache space usage management
- Their contributions are to address above challenges

## VIRTUALIZATION MECHANISM

- Storage Virtualization device (SV) separates logical address space from physical location of data (They choose Out-of-band one)



## EXTENT-BASE METADATA MANAGEMENT

- An one-on-one mapping lookup table is too large to be accepted
- Solution: Different granularities for movement of blocks from /to cache space
  - A set of contiguous logical blocks move in/out of cache together
- A monitor to guide the selection of extent size and cache access pattern

## I/O HINTS

- Research shows that HPC application can generate hints online
- At the beginning of application execution, they could disclose future I/O accesses
- The paper is focus on how to use the hints efficiently
- Assumption: Hints is about what and when the application need I/O accesses

## CACHE SPACE MODEL

- The size of cache space is proportional to the size of entire data set and is manifold
- Model
  - Prefetch and write request as consumer
  - Write destage, dirty flushing and read request as producer
- Only prefetch request scheduling is under control

## SCHEDULING OF PREFETCH

- First Come First Served (FCFS)
  - Don't consider deadline
- Earliest Deadline First
  - Multi-server and multi-application occur collisions
- Prefetch Horizon
  - No benefit in executing a prefetch request earlier than it is actually required

## DEADLINE-BASED PREFETCH SCHEDULER

- Objective
  - Achieve fairness
  - Give as much time as possible to execute prefetch request
- Energy saved on the second objective
  - Batch execution
  - Avoid multiple turning ON/OFF
- Perform deep prefetching: prefetch as far as possible based on resource constrain

## OUTLINE

- 1. Introduction
- 2. Existing Methods
- 3. Future Research Issues
- 4. Existing works
- 5. My Research plan

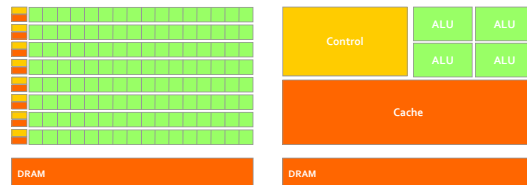
## MY RESEARCH PLAN

- Workload mapping scheme based on server with GPU
- GPU (Graphics Processing Unit)
  - Used in video cards
  - Hot in general computing area
- Multi-core programming
  - Derived from high performance computing and parallel computing

## GPU

GPU:  
A lot of components for computing  
Good for computing intensive tasks  
Good for numerical calculation or linear algebra

CPU:  
General purpose computing



## PROGRAMMING IN GPU

- Nvidia GeForce 8800 GPU
- CUDA programming model developed by Nvidia
- Able to Integrate with VS2005, language C/C++
- Multi-thread programming, the minimum unit in CUDA is thread

## CHALLENGES

- Different architecture between CPU and GPU
  - Inner organization: ALU, BUS etc.
  - Different feature in computing
- Granularity in workload mapping considering Heterogeneity
  - Large: mapping workloads to a group of servers with the same feature in some dimension: CPU, Disks, Energy Cost...
  - Small: Threads mapping to cores in a processor

## CHALLENGES

- How to implement DVFS in GPU
- Simulation and Experiment
  - Simulate GPU, workload pattern...
- How to measure the actual energy consumption of a server

## COMMERCIAL PRODUCT

- LSF from Platform Computing
  - HPC management software
  - Help to save energy by mapping workload into some of servers

**THANKYOU!**

QUESTION?

COMMENT?