HybridStore: A Cost Efficient, High Performance Storage System Combining SSDs and HDDs

July 26th 2011

Youngjae Kim, Aayush Gupta, Bhuvan Urgaonkar, Piotr Berman, and Anand Sivasubramaniam

Oak Ridge National Laboratory
Pennsylvania State University

Enterprise-scale Storage Systems

Enterprise-scale Hard Disk Drive

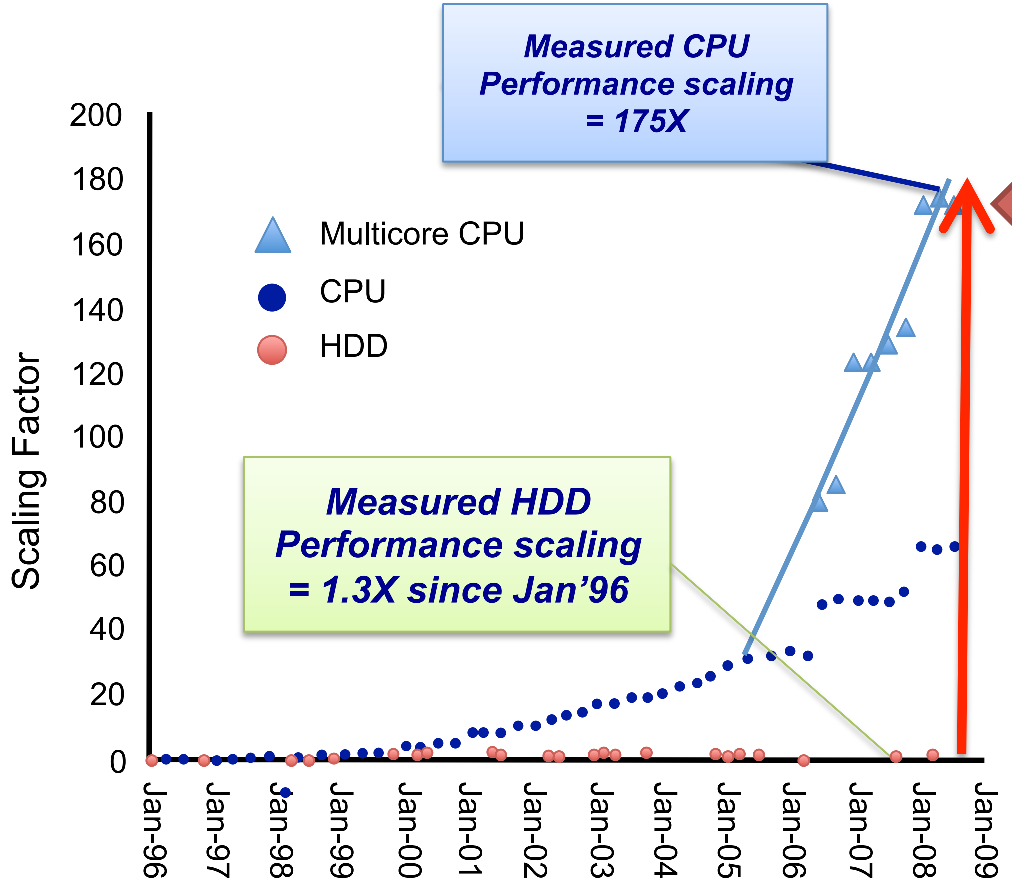
- **Enterprise-scale Storage Systems**
  - Information technology focusing on storage, protection, retrieval of data in LARGE-SCALE environments
- **Data-centric services**
  - File, web & media servers, transaction processing servers

Google's massive server farms

A Persistent Hurdle in Enterprise Computing

Huge Performance Discrepancy Between CPU and HDD

- Normalized CPU Performance and Media Access Time



Measured CPU Performance scaling = 175X

Measured HDD Performance scaling = 1.3X since Jan'96

Source: Intel Measurements

- HDD Performance (Ra as been sta

Flash SSD Potential !!

- I/O bottleneck has become increasingly worse over time.

4KB Random Read:

35,000 IOPs

Intel® X25-E

4,000 IOPs

Seagate Cheetah 15K.6

Emergence of NAND Flash

Embedded, Desktop, and Enterprise

- ○ **Embedded Storage**
  - PDAs, mobile phones, digital cameras
- ○ **Desktop storage**
  - MacBook Air, One Laptop Per Child (OLPC), game consoles, Intel's X25-E Extreme SATA Solid-State Drive
- ○ **Enterprise scale storage**
  - Fusion-io's ioDrive, Texas Memory System's RamSan-500, Symmetrix DMX-4 from EMC

$219 / 120GB

Intel 320 MLC Series
(38K IOPS for Reads,
1.4K IOPS for Writes)

$8,335 / 320GB

Fusion-IO's ioDrive Duo (MLC)
(100KIOPS for Reads, 141KIOPS for Writes)

Unknown

Violin Memory Inc – Violin 1010

Scalable Memory Architecture (VXM)
84 VIMMs (Violin Intelligent memory Modules)
(1M random IOPs, PCIe x4/x8 I/F, DRAM/Flash SSD)

Contents

- Introduction

- **Background**
  - NAND Flash based SSDs versus HDD
  - Motivation for HybridStore and Related Works

- Overview of HybridStore
  - Capacity Planner
    - Workload Analyzer
    - Storage Optimization Solver

- Experimental Results

- Conclusion

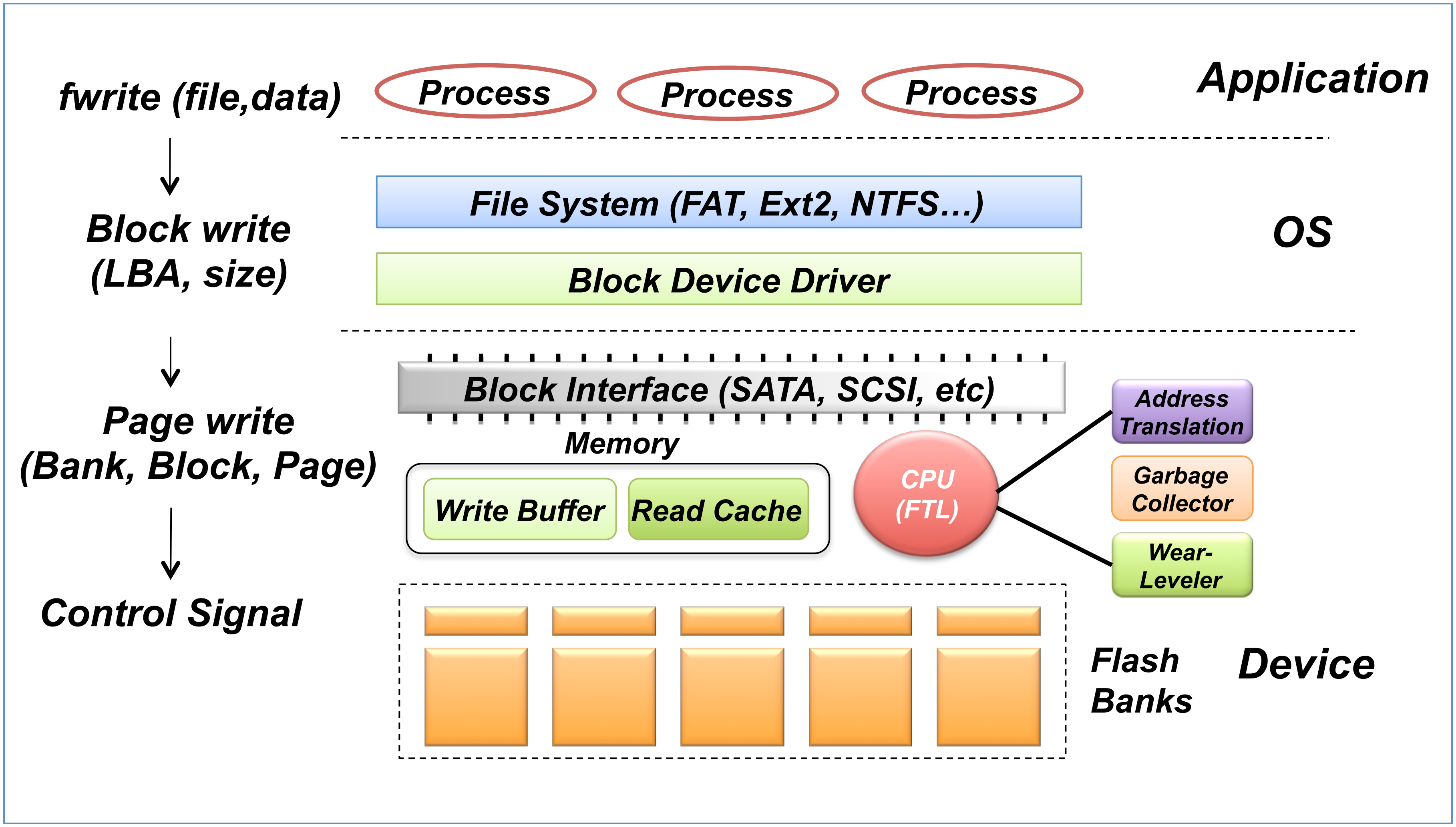
Emergence of NAND Flash based SSD

- **NAND Flash vs. Hard Disk Drives**
  - Pros:
    - Semi-conductor technology, no mechanical parts
    - Offers lower and more predictable access latencies
      - Microseconds (45us Reads / 200us Writes) vs. Milliseconds for Hard Disks
    - Lower power consumption
    - Higher robustness to vibrations and temperature

  - Cons:
    - Limited lifetime
      - 10K - 1M erases per block
    - High cost
      - About 8X more expensive than current hard disks
    - Random writes can be sometimes slow

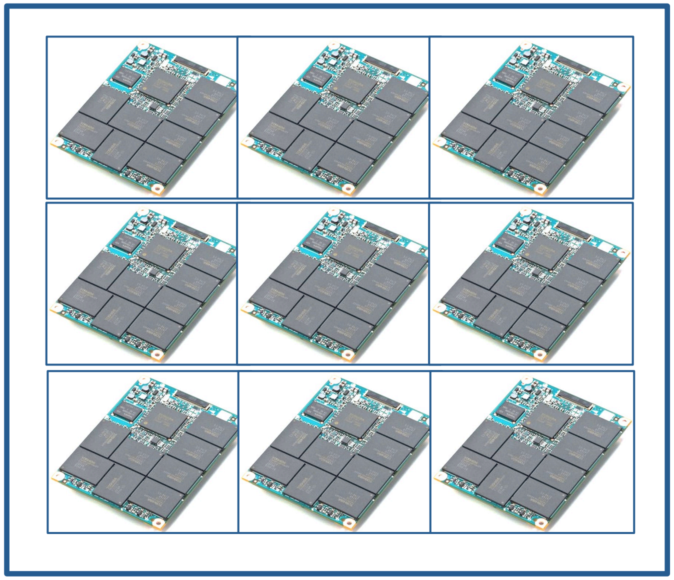
NAND Flash based SSD

System Architecture



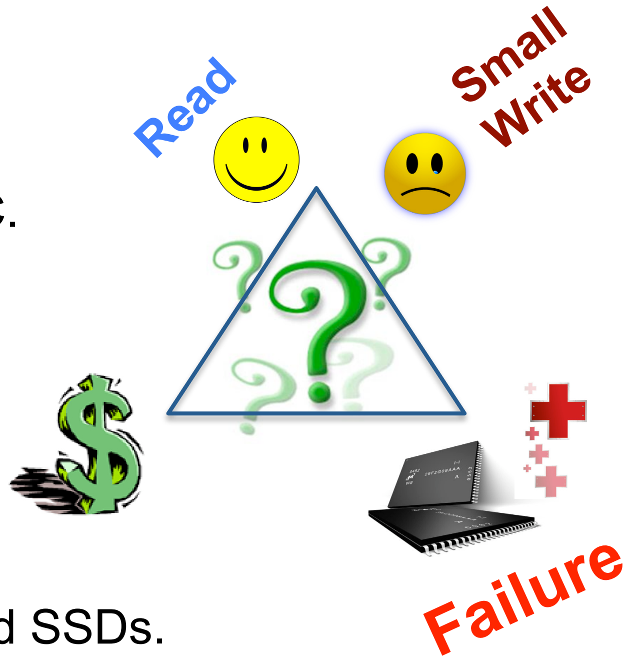
Existing Storage Server Platform

- **Examples of Storage Server Platform**
  - Various network interface
    - Fibre Channel, SAS etc
  - Various types of hard disk drives
    - 2.5” SAS drive, 3.5” SATA drive, etc



Can SSDs replace HDDs?

- **Challenges**
  - Unique performance characteristics of SSD
    - SSD may become worse than HDD due to GC.
  - Reliability Concerns
    - Lifetime of SSDs is limited by the write rates.
  - Cost Concerns
    - NAND Flash is still expensive over HDD.
- **HybridStore**
  - Hybrid storage systems that combine HDDs and SSDs.



Homogeneous HDD based Storage Server Platform

Homogeneous SSD based Storage Server Platform

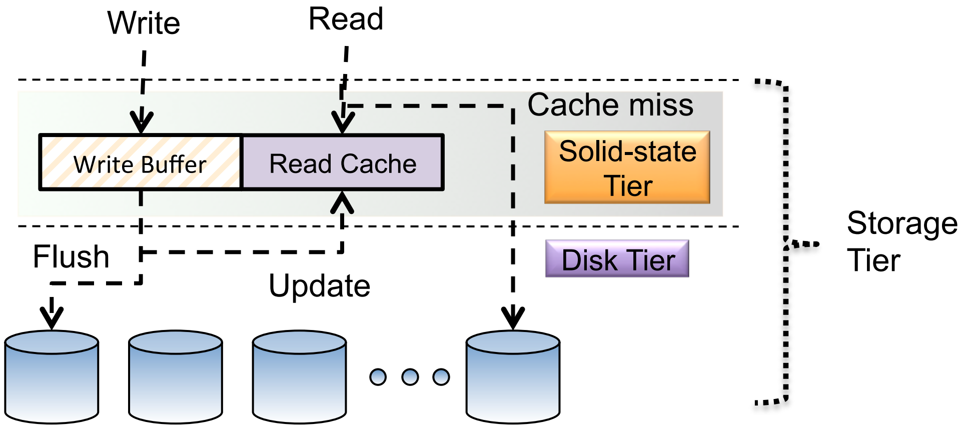
HybridStore: Heterogeneous Storage Server Platform

Existing Proposals in Enterprise

- **Hybrid Hard Disk**
  - NAND Flash is on-bard cache in HDD.
- **Intel Turbo Memory (ITM) [ACM TOS'08]**
  - Support for the ReadyBoost and Ready Drive of Microsoft
- **Two-tier Architecture from Microsoft [Eurosys'09]**
  - Use SSDs as Long-Term read Cache and Short Write Buffer
- **ZFS (designed by Sun)**
  - ReadZilla & LogZilla (Implementation of read cache and write buffer)

Hybrid HDD, FlashON™

Intel Turbo Memory



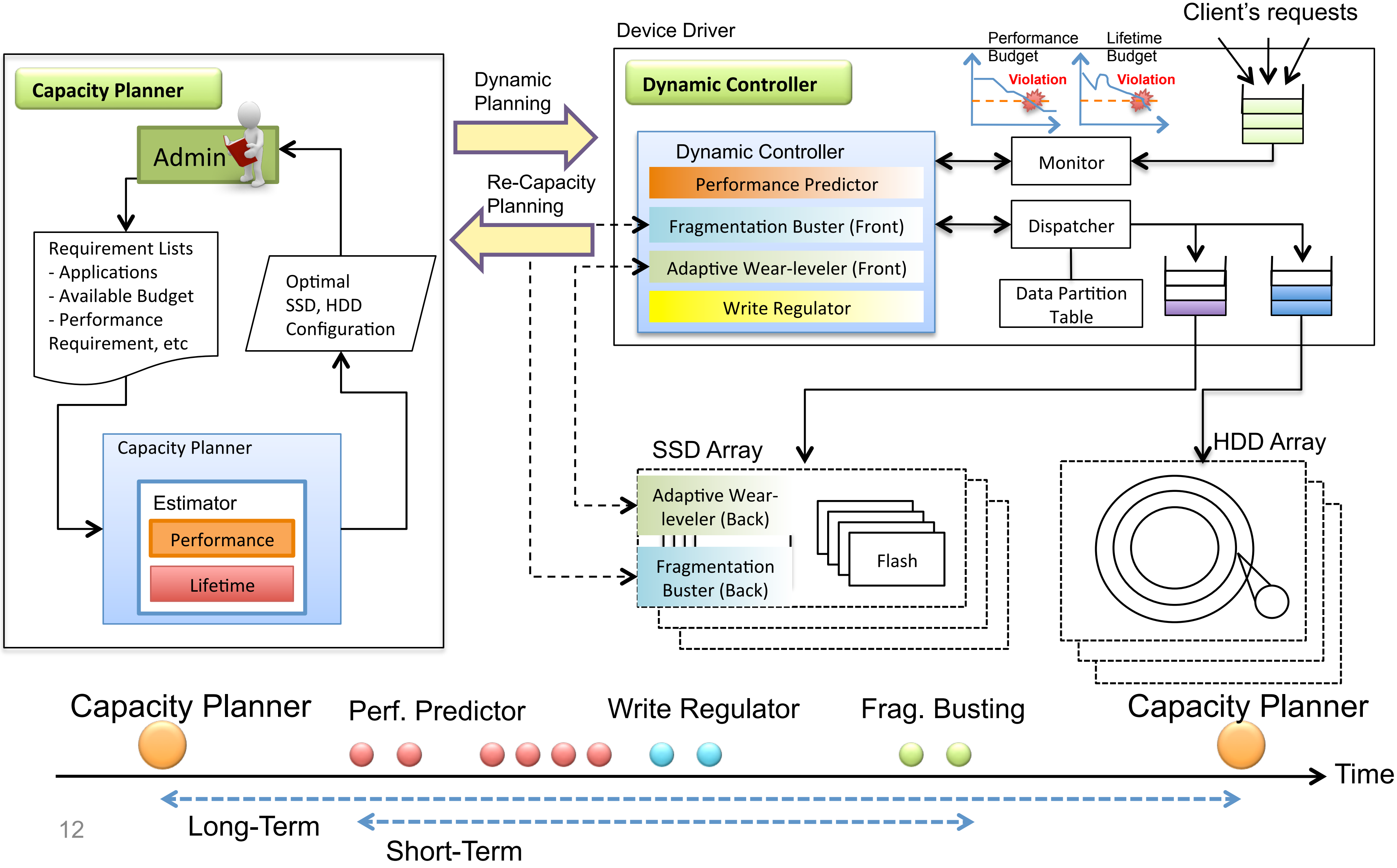
Two-tier storage architecture

Contents

- Introduction
- Background
  - NAND Flash based SSDs versus HDD
  - Motivation and Related Works
- **HybridStore**
  - Overview of HybridStore
  - Capacity Planner
    - Workload Analyzer
    - Storage Optimization Solver
- Experimental Results
  - Synthetic and Realistic Workloads
- Conclusion

Overview of HybridStore

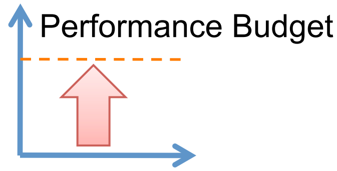
Capacity Planner and Dynamic Controller

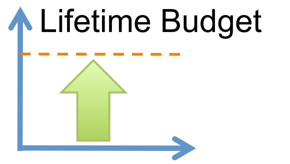


Capacity Planning

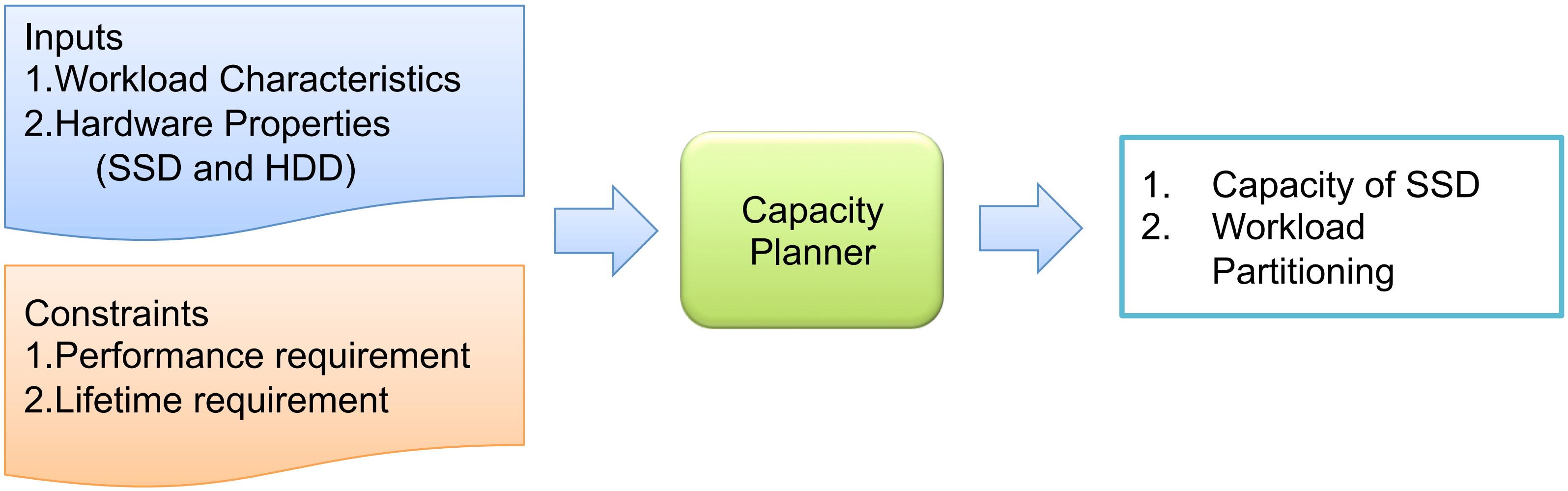
Problem Formulation: Goal and Constraints?

Goal	Minimize Cost of HybridStore
Constraints	1) Perf. of HybridStore > Perf. Budget 2) Lifetime of HybridStore > Lifetime Budget





$$\text{Cost of HybridStore} = \text{Cost}_{\text{SSDs}} + \text{Cost}_{\text{HDDs}} + \text{Cost}_{\text{Recur}}$$



Inputs
1.Workload Characteristics
2.Hardware Properties
(SSD and HDD)

Constraints
1.Performance requirement
2.Lifetime requirement

Capacity Planner

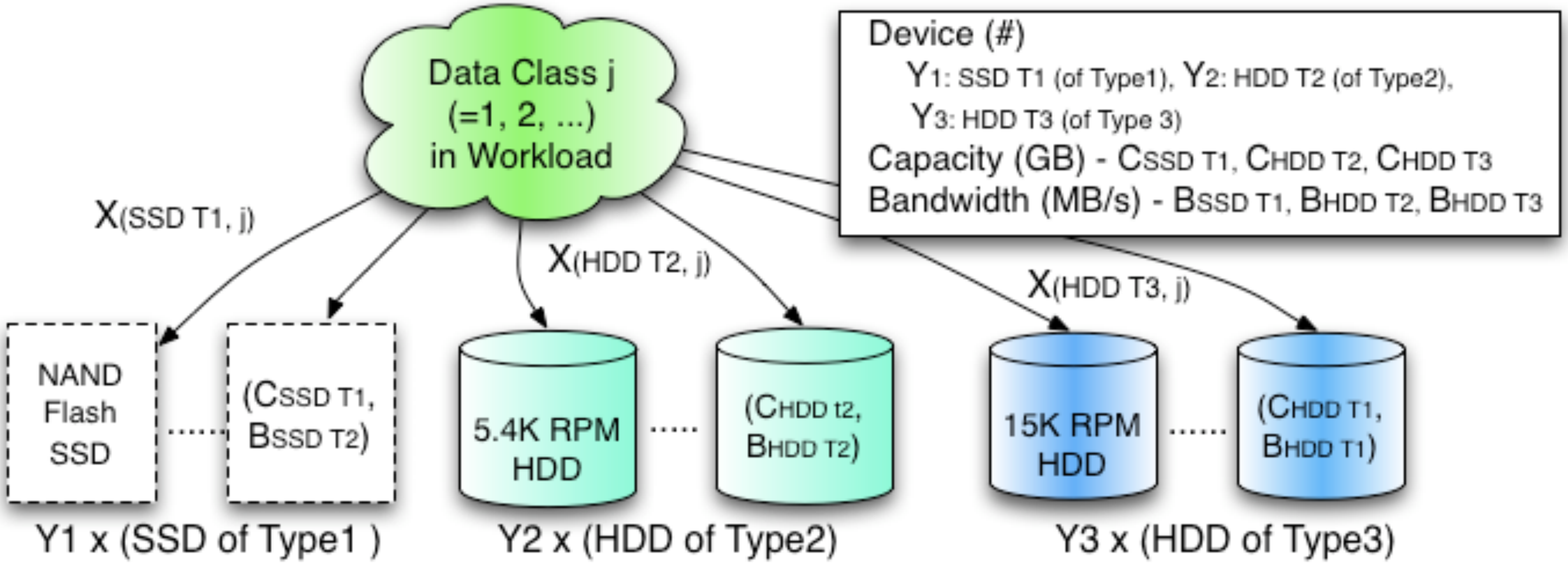
1. Capacity of SSD
2. Workload Partitioning

Capacity Planning

HybridStore Hardware Model

- **Provisioning SSDs**
  - Find storage capacity of SSDs and HDDs
  - Find out the amount of data partition sent to SSDs for a given workload

- **Storage Model & Data Partitioning**



Finding Workload Attributes

- **I/O workloads can be characterized by**
  - Hot (highly accessed) and cold (rarely accessed) data, Read/write ratio, Sequentiality, Request arrival rate, etc

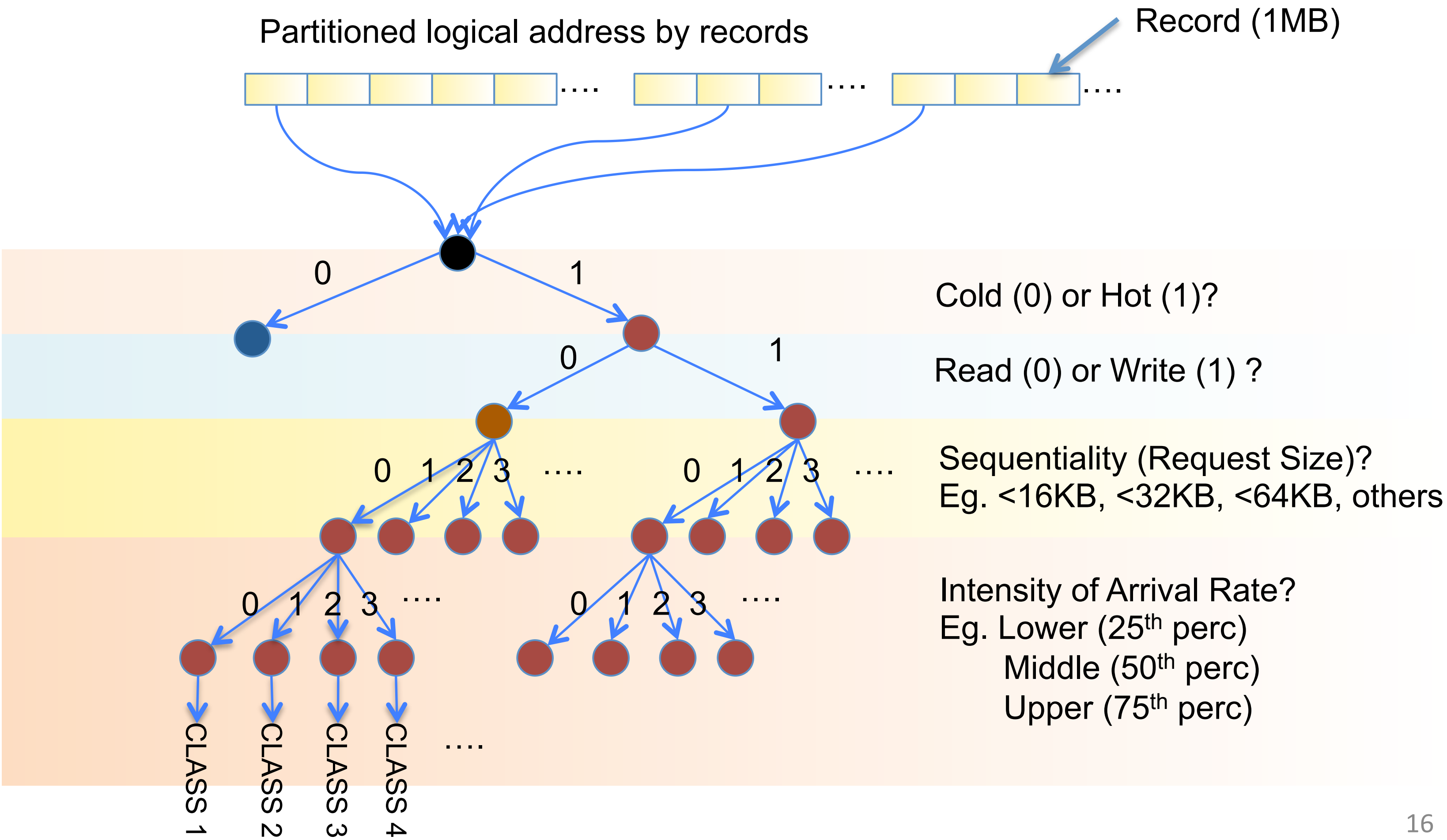
- **Data Classification**
  - A methodology to partition a workload into smaller subsets.

- **Finding workload attributes**
  - The entire logical address space of the workload is divided into fixed-size chunks (or records), then, mapped to different data classes.
    - 1MB record size is used because 1MB roughly corresponds to the granularity of data prefetching doe by HDDs/SSDs.
  - Each data record is represented by the following workload attributes
    - Temporality (frequency of accesses per unit time)
    - Read/write ratio
    - Request size (spatial locality) – sequential, partially sequential, partially random, and random.
    - Request arrival rate

Data Classification

Hierarchical Data Classification

- Tuples (Hot or cold, Read ratio, Sequentiality (request size), Arrival rate)



Cold (0) or Hot (1)?

Read (0) or Write (1) ?

Sequentiality (Request Size)?
Eg. <16KB, <32KB, <64KB, others

Intensity of Arrival Rate?
Eg. Lower (25th perc)
Middle (50th perc)
Upper (75th perc)

Capacity Planning

Capacity Planner: Problem Formulation

Declaration of Variables

- Properties of *device type i*

$$C_i = Capacity\ \ of\ \ device\ type\ i$$

$$U_i = Utilization\ \ of\ \ device\ type\ i$$

$$B_i = Maximum\ \ Bandwidth\ \ of\ \ device\ type\ i$$

- Properties of *data class j*

$$S_j = Size\ \ of\ \ data\ class\ \ j$$

$$F_i = Frequency\ \ of\ \ data\ class\ \ j$$

$$W_{ij} = Weight\ \ factor\ \ for\ bandwidth\ of\ \ data\ cass\ \ j\ on\ \ y_i\ devices\ of\ \ device\ type\ i$$

Decision Variables

$$x_{ij} = data\ of\ \ class\ \ j\ on\ \ y_i\ devices\ of\ \ device\ type\ i$$

$$y_i = number\ \ of\ \ devices\ of\ \ device\ type\ i$$

← Integer variable

Capacity Planning

Mixed Integer Linear Programming

- Objective Function

$$Cost_{HybridStore} = Cost_{Installation} + Cost_{\mathrm{Re}\,curring}$$

$$= (\sum_{i=1}^{I} y(i) \times D_{\$}(i) \times C_i + (K_{\$} \times \sum_{i=1}^{I} y(i) \times \int_{t} P(t)dt))$$

- Constraints

$$\sum_{i} x_{ij} = S_j,\ (\forall j \in J)$$

$$\sum_{j} x_{ij} \le (U_i \times C_i) \times y_i,\ (\forall i \in I)$$

Space constraint

$$Expected\ lifetime = \frac{(Size\ of\ NAND\ flash\ \times \#\ of\ erase\ cycles)}{bytes\ written\ per\ day}$$

$$F_j \times \frac{x_{ij}}{S_j} \le B_{ij} \times y_i,\ (\forall i \in I,\ \forall j \in J)$$

Performance constraint

$$Lifetime(i,x) \le Useful\ Lifetime\ of\ HDD\ (i \in Flash\ based\ SSDs)$$

Evaluating HybridPlan

- **Solver development**
  - Developed a trace analyzer (lines of codes less than 500)
  - Developed the solver of HybridPlan using CPLEX
    - CPLEX, a well-regarded Integer Linear Programming (ILP) solver
- **Workloads**
  - Synthetic workloads
  - Realistic workloads
    - MSR Cambridge traces, and Microsoft Exchange server Traces
- **Devices**

Device	Type	Capacity (GB)	Per-GB ($)	Utilization	Read (MB/s)	Write (MB/s)	Latency (ms)	Erase (#)	Power (W)
Seagate Cheetah	15K HDD	146	1.80	0.8	171	171	3.6	-	12.92
Seagate Barracuda	7.2K HDD	750	0.17	0.8	125	125	4.2	-	9.4
Intel X 25-E	SLC SSD	32	11.96	0.5	230	200	0.125	100K	2
Intel X-25-M	MLC SSD	80	3.22	0.5	220	80	0.25	10K	2

Synthetic Workloads

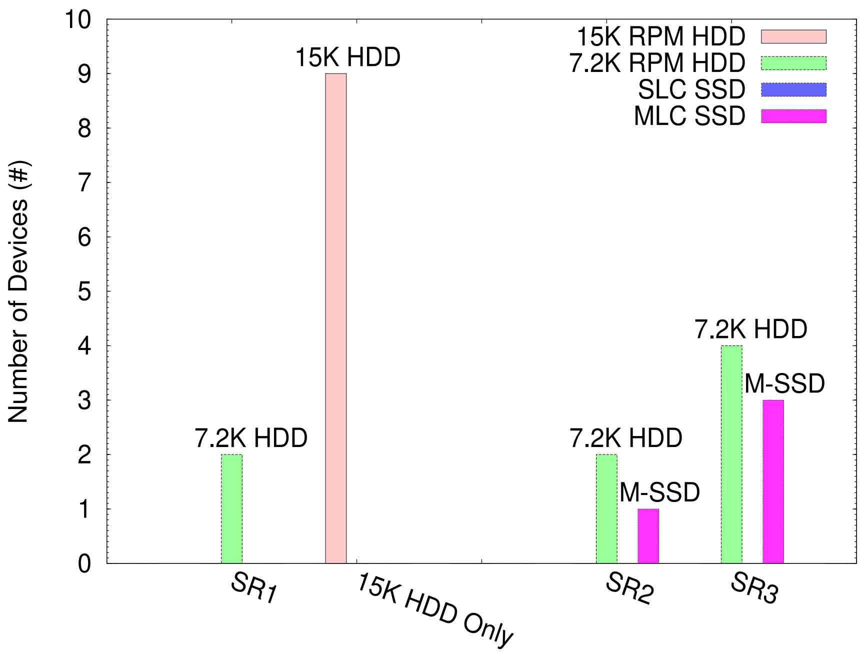
- Description of Synthetic Workloads

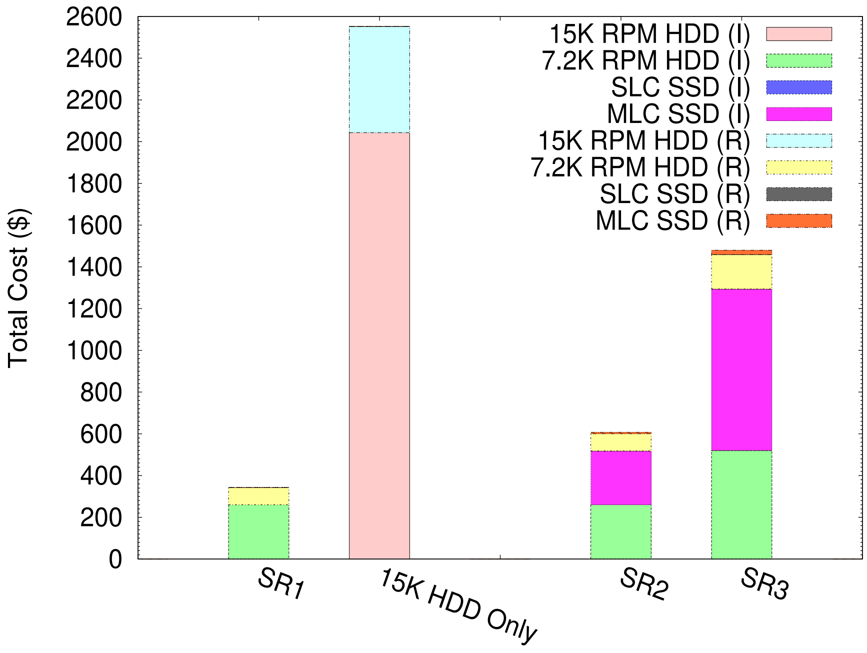
Workloads	Index	Read (%)	Size (KB)	Inter-Arrival	I/O Bandwidth	
				Time (ms)	MB/s	IOPs
Sequential Read	SR1	80	128	100 (L)	1.25	-
	SR2	80	128	2 (M)	62.5	-
	SR3	80	128	0.2 (H)	1,250	-
Random Read	RR1	80	4	100 (L)	-	10
	RR2	80	4	2 (M)	-	500
	RR3	80	4	0.2 (H)	-	10,000
Sequential Write	SW1	20	128	100 (L)	1.25	-
	SW2	20	128	2 (M)	62.5	-
	SW3	20	128	0.2 (H)	1,250	-
Random Write	RW1	20	4	100 (L)	-	10
	RW2	20	4	2 (M)	-	500
	RW3	20	4	0.2 (H)	-	10,000

Impact of I/O Intensity

○ Results for Sequential Read Dominant Workloads

- SR1, SR2, SR3



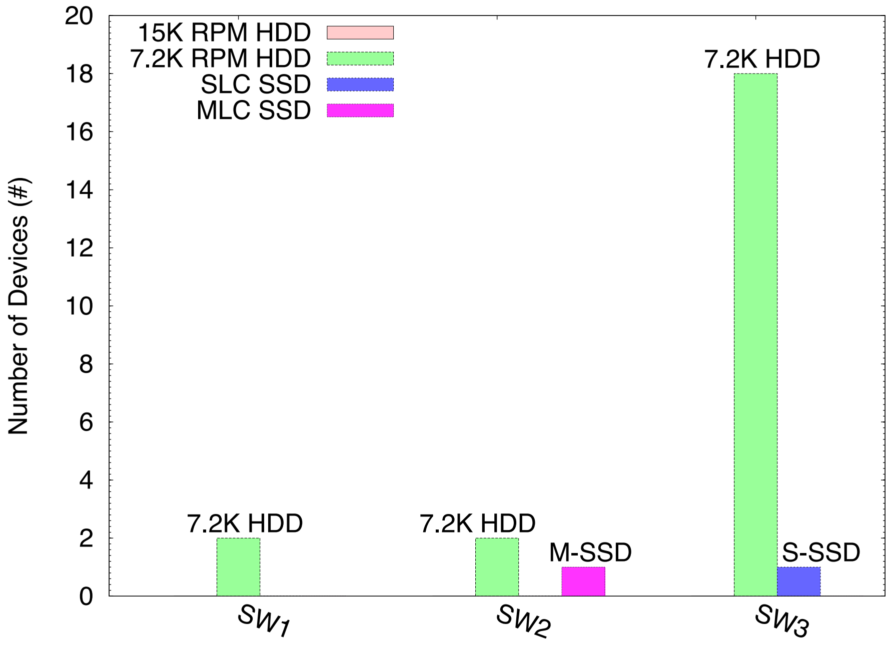


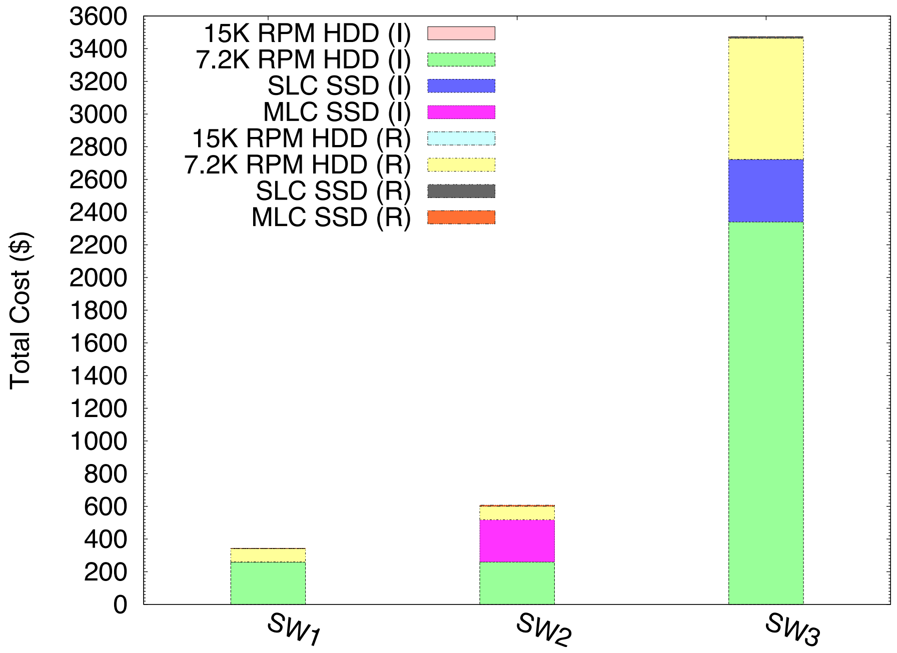
- SR1 only requires 2 slow 7.2K RPM HDDs whereas it requires 9 fast 15K RPM HDDs.
- Our solver determines the right devices to meet the capacity needs.
- As the arrival rate increases, we observe the need for MLC SSDs (considering $/GB for SLC SSD, it is not efficient to use compared to using MLC SSD).
- Recurring cost (Electricity cost) are quite small compared to device installation cost.

Impact of I/O Intensity

○ Results for Sequential Write Dominant Workloads

- SW1, SW2, SW3



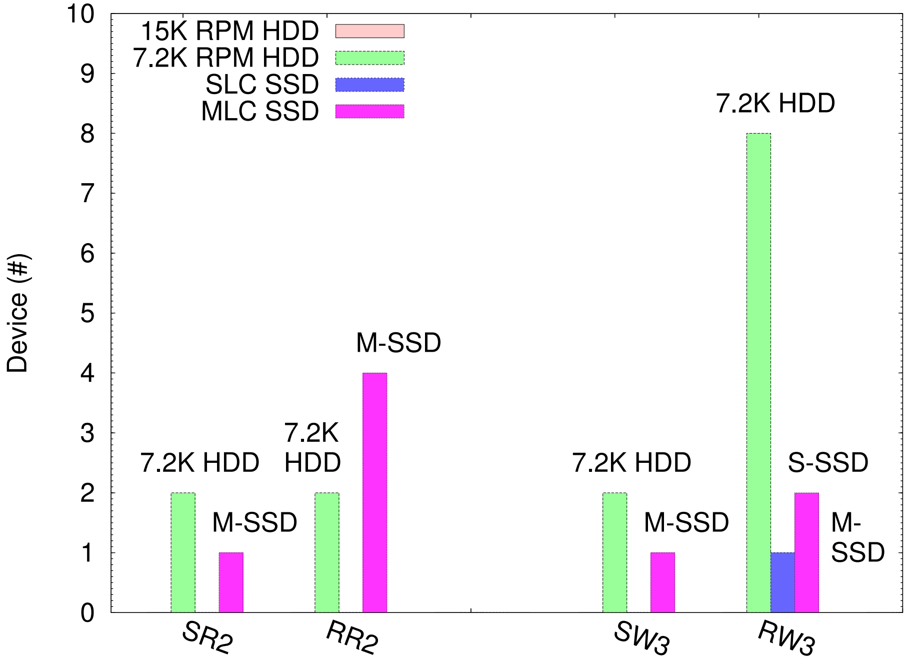


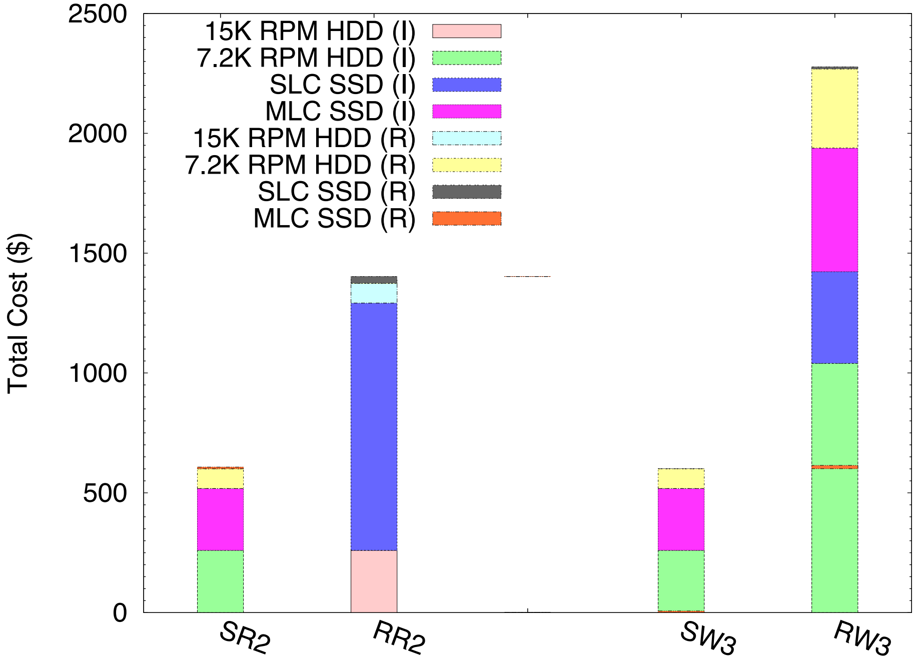
- For write dominant SW3, unlike observation from SR workloads, the solver suggests to use one SLC SSD instead of the MLC ones for its read-intensive counterpart (SR3). It's because SLC SSD that we use is 2.5 times faster than the MLC one.
- Also it needs a sharp increase in the number of slow HDDs because of the vast $/GB difference between SLC SSDs and slow HDDs.

Impact of Sequentiality

○ Results for Sequential and Random Workloads

- SR, SW



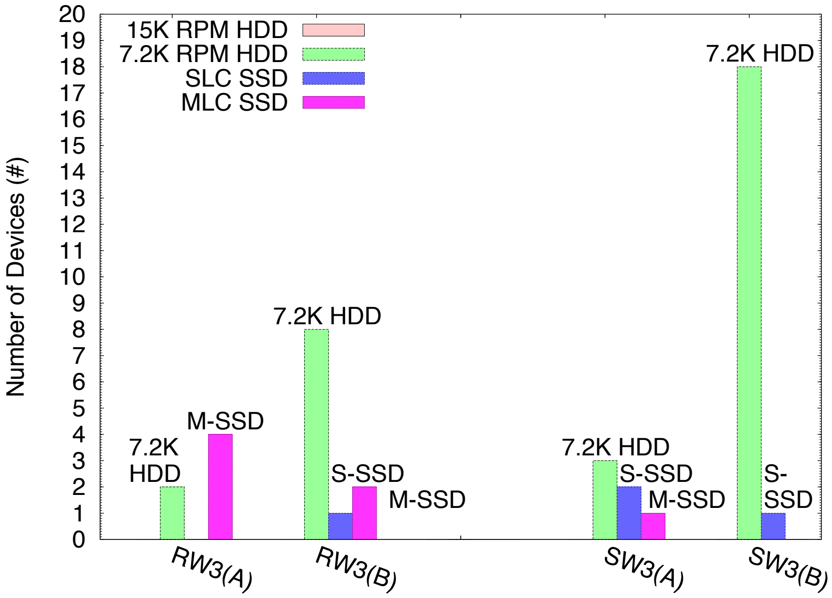


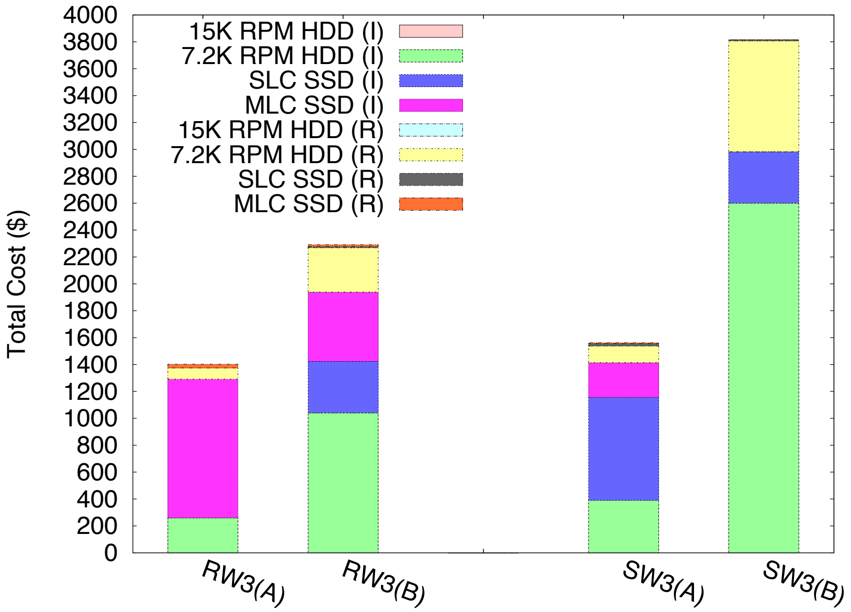
- We clearly see the needs of the larger number of SSDs as the workloads are random.
- For RW3, we observe the needs of SLC-SSDs to meet the high IOPS requirement.
- As a storage administrator, it is highly advisable to increase the sequentiality of incoming workloads so as not to employ expensive SSDs.

Impact of Lifetime Constraint

○ Results for without and with lifetime constraints

- denoted as (A) and (B) respectively





- Lifetime constraint is an important metric in capacity provisioning.
- Without lifetime constraint, we see a greater portion of SSDs being used than with the lifetime constraint.
- For SW3, without lifetime constraint, we may have lower number of devices as well as the overall cost compared to when the lifetime constraint is forced, however, the storage administrator needs to re-provision prematurely, eventually increase the overall costs over the initial estimated period.

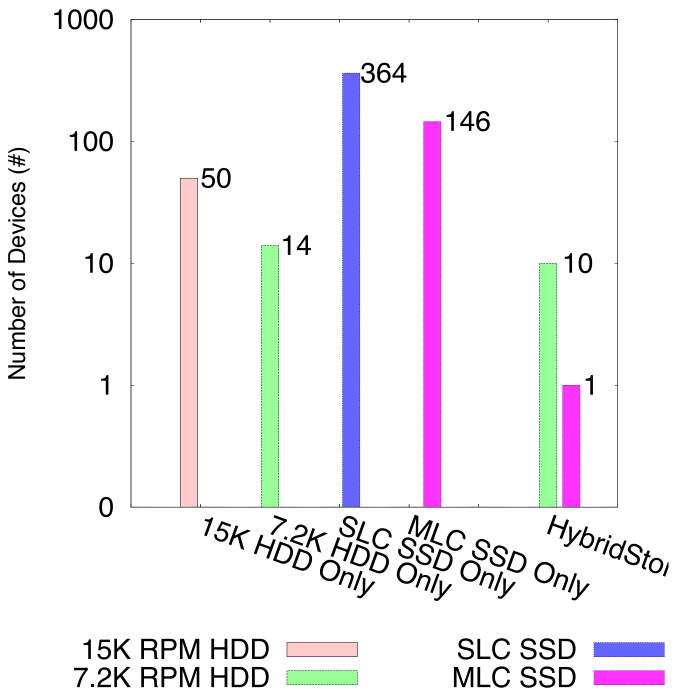
Realistic Workloads

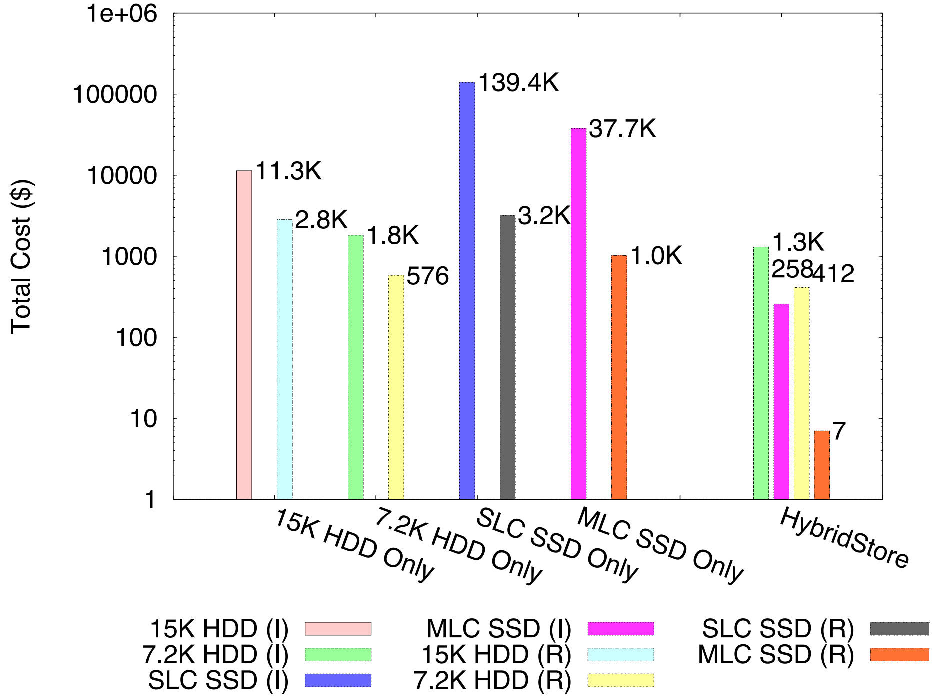
- **Description of Realistic Workloads**

Workload	Size (TB)	Read (%)	Request Size (KB)	IOPS
MSR Trace	5.7 TB	68.1	23.32	823
Exchange Server	750GB	38.3	16.54	3,692

Can SSDs replace HDDs?

○ Results for MSR Traces

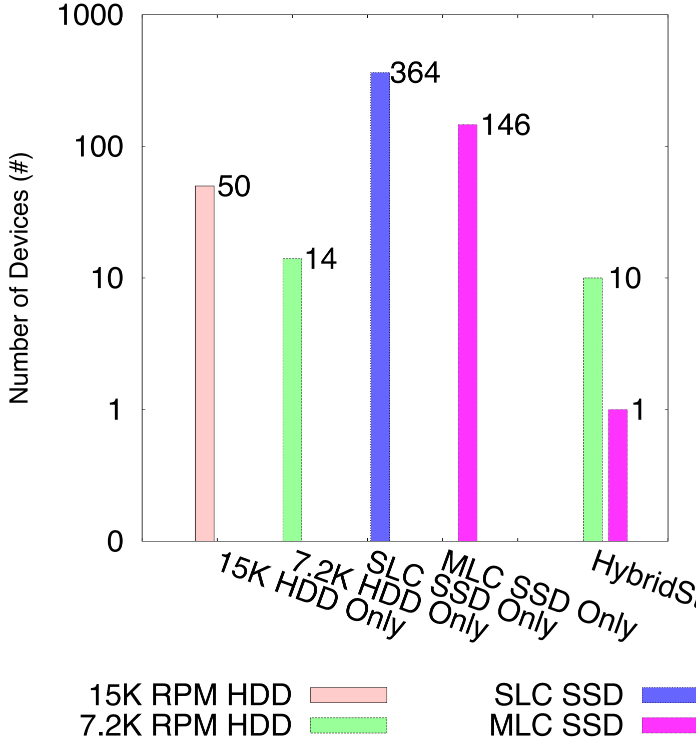


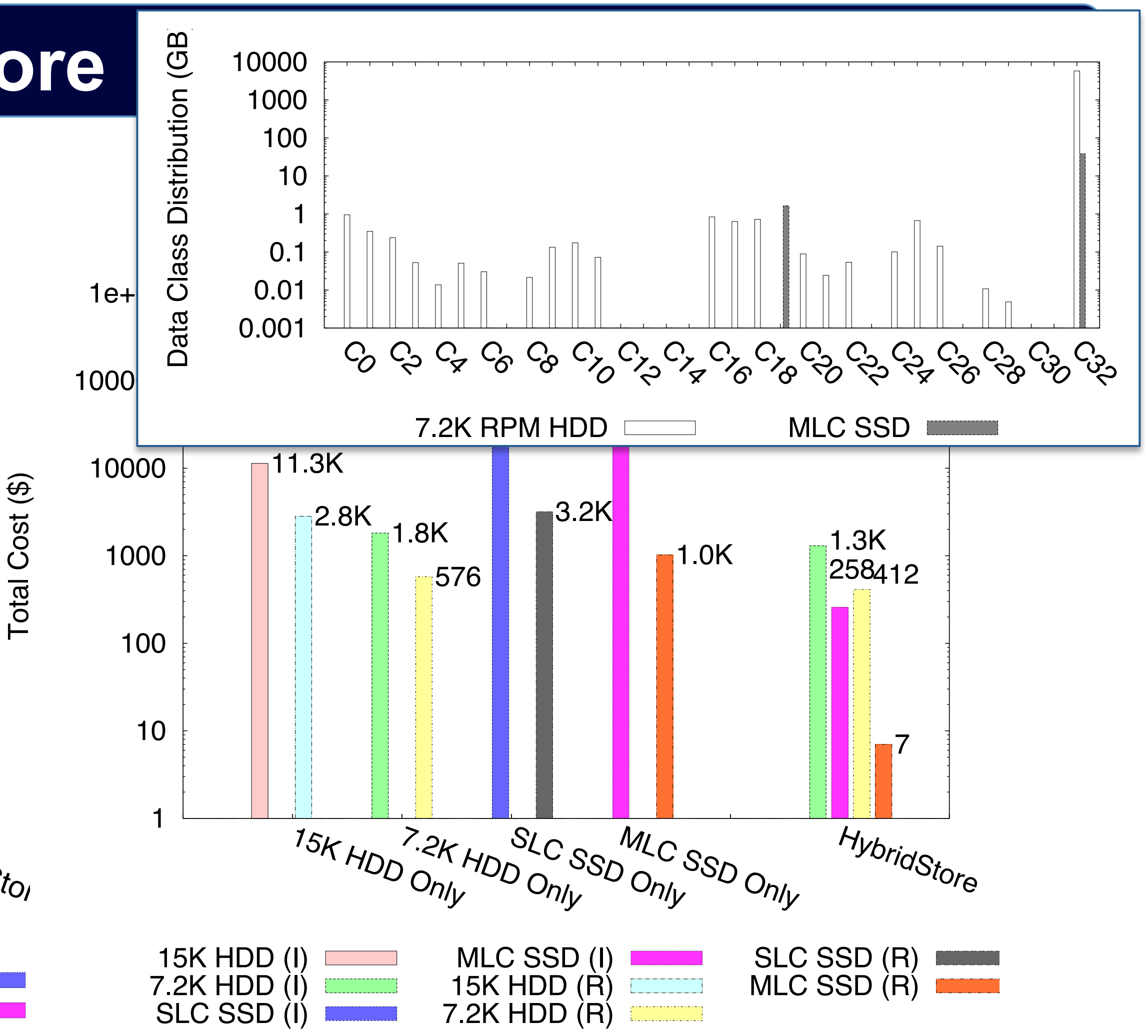


- Employing 7.2K RPM HDDs is more economically efficient than employing 15K RPM HDDs.
- In case of SSD systems, it requires several hundreds of SSDs to satisfy the capacity requirement.
- The bounding factor for decision-making of HybridPlan is not I/O bandwidth but storage capacity requirement.

Efficacy of HybridStore

○ Results for MSR Traces





- HybridPlan can find the most economic storage composition.
- HybridPlan suggests 2 x 7.2K RPM HDDs and 1 MLC SSD for MSR Trace.
- Total cost saving of HybridStore is about 85% compared to high-end HDD only system.
- 99% data are classified into C32, a data class storing data rarely accessed.

Lessons Learned

I. We developed an capacity planner that finds the most economically efficient storage configurations while meeting the performance and lifetime requirements of devices

II. We provided a general form of comprehensive methodology using a well-known technique for optimization problems, Mixed Integer Linear Programming (LP)

III. Experiments showed that our capacity planner is able to identify close to minimum SSD capacity needed to meet a specified performance goal for realistic workloads while ensuring similar performance as compared to a comparatively more over-provisioned system