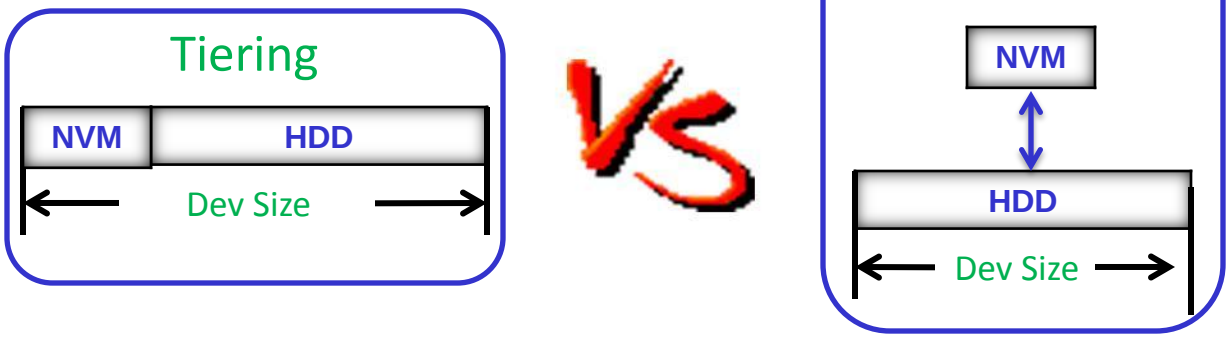
On the Importance of Evaluating Storage Systems' $Costs

Zhichao Li, Amanpreet Mukker, and Erez Zadok

http://www.fsl.cs.stonybrook.edu

Stony Brook University

Hybrid Drive (PCM/SSD/HDD/etc)



Fusion Drive.



DELL compellent

Stony Brook University

Problem Statement

- Performance alone is not enough and dollar cost matters
- Lacking empirical TCO study when systems deploy SSD

Our work?

- We built a cost model to justify gained performance
- We built and evaluated two empirical systems: tiering vs. caching

Cost Model

- Upfront purchase
  - Price per capacity ($/GB)
- Total cost of ownership (TCO)
  - Energy and power costs
    - Lookup local electricity authority
  - Endurance cost
    - ΔEndurance(dev) x Cost(dev)
  - Service and space cost



Scale TCO to predict long-term costs



Energy and Power Costs

Prices ($/)	Power <= 7KW		Power <= 145KW		Power > 145 KW	
	Energy	Power	Energy	Power	Energy	Power
offpeak	0.0863	0	0.0191	0	0.0218	0
peak	0.1052	0	0.0340	48.78	0.0446	28.76
intermediate	0.0863	0	0.0317	5.94	0.0356	8.13

PSE & G

(per KWh and per KW)

Stony Brook University

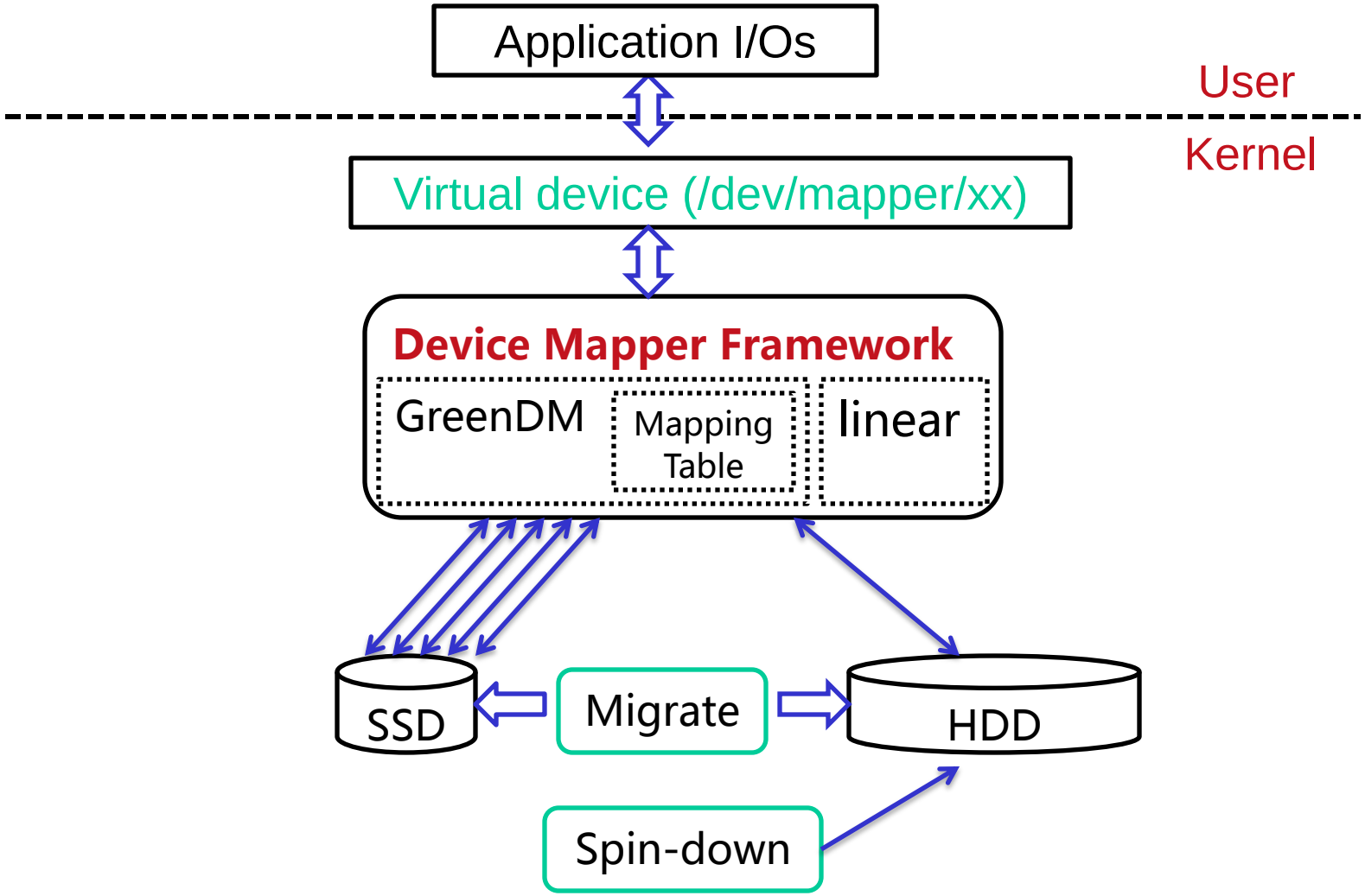
Device Endurance Model

- Model :
  - Track writes to the SSD (convert reads)
    - $E_{SSD}(t) = 1 - \#writes(t)/Limit_{SSD}$
  - Count the HDD start-stop cycles
    - $E_{HDD}(t) = 1 - \#start\text{-}stop(t)/Limit_{HDD}$
  - $\Delta E_{SSD} = E_{SSD}(t_2) - E_{SSD}(t_1)$
  - $\Delta E_{HDD} = E_{HDD}(t_2) - E_{HDD}(t_1)$
  - Endurance Unit: **eu**
    - [0, 1,000,000] = [dead, new]

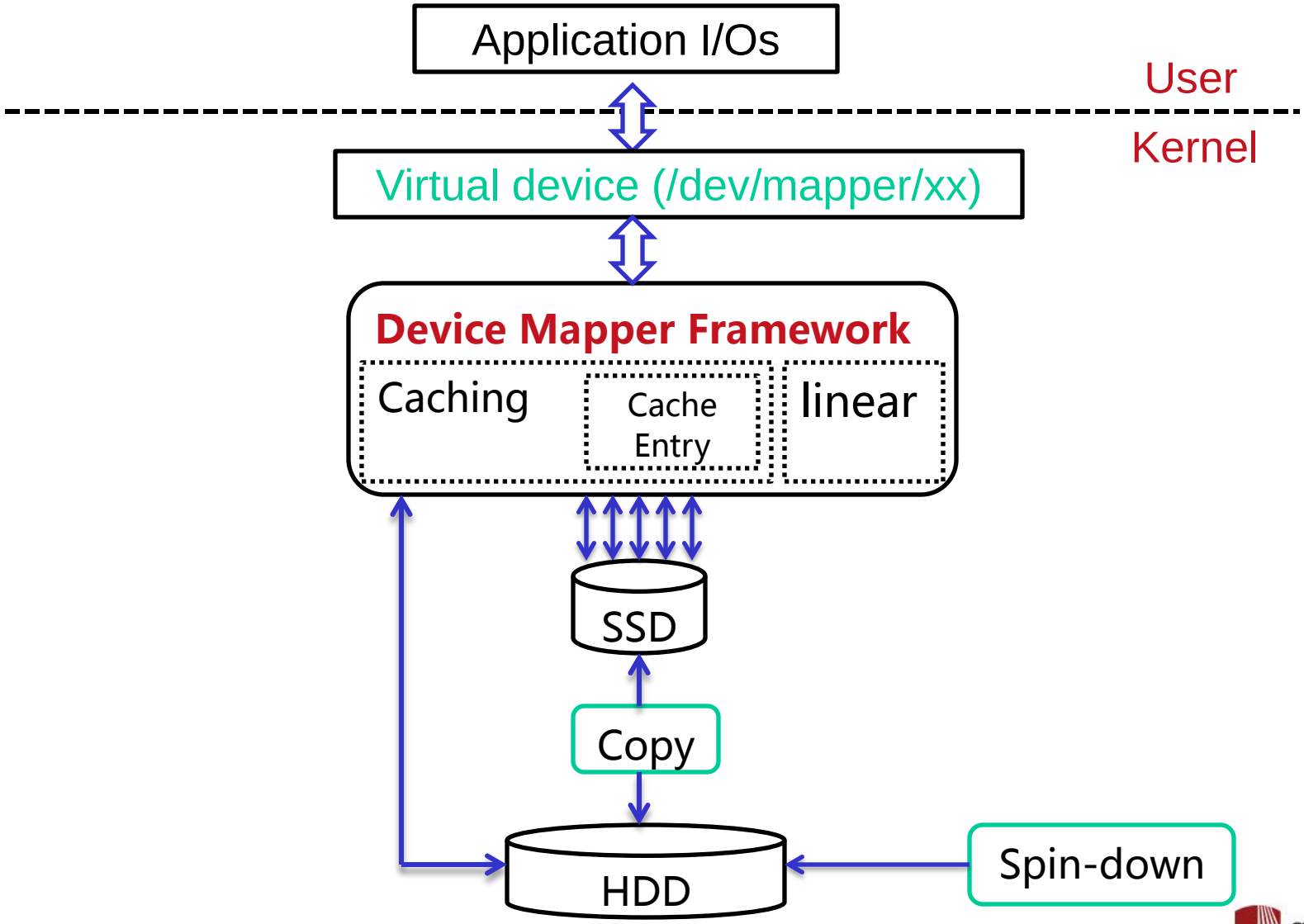


Architecture

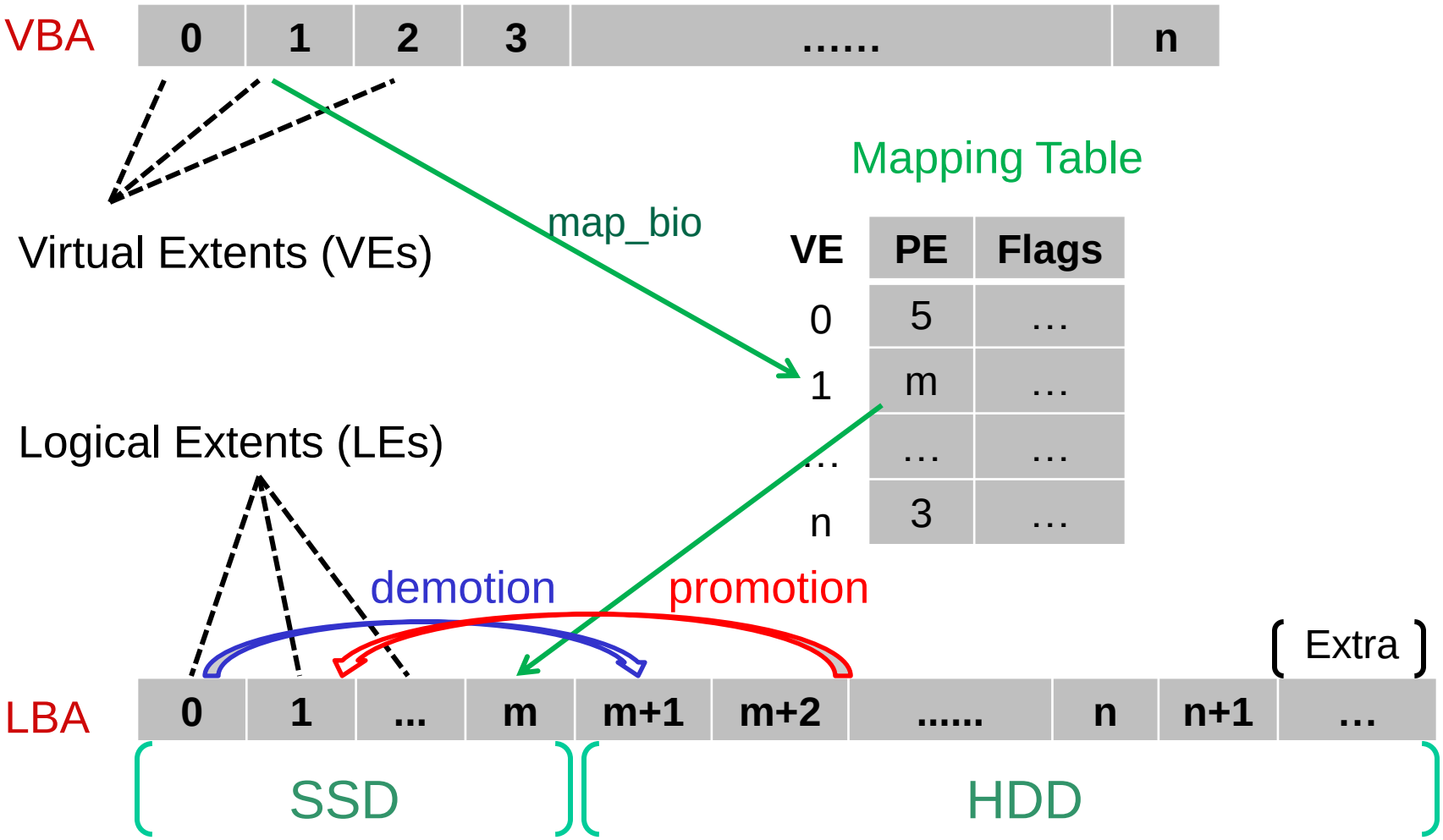
Tiering Architecture



Caching Architecture

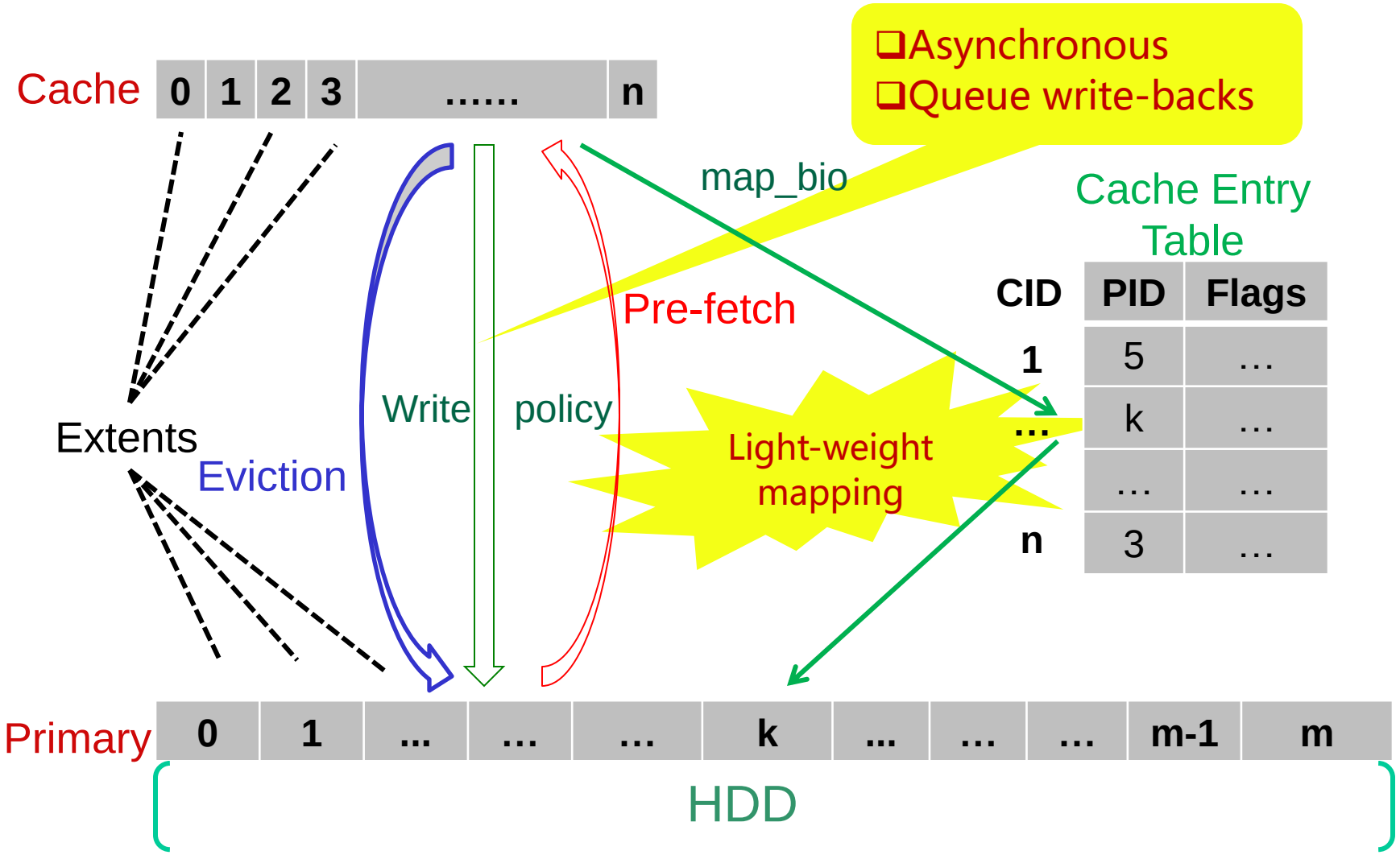


Tiering Data Management



Stony Brook University

Caching Data Management



Stony Brook University

Evaluation

Experimental Setup

- Lenovo Think-Center
  - Core 2 Quad 2.66GHz CPU and 4GB RAM
  - Intel 300GB SSD ($529) and Seagate 2TB HDD ($200)
- Wattsup Pro ES for power/energy measurement
- Three baselines
  - 1) Mylinear; 2) SSD-only; 3) HDD-only
- Various parameters
  - PT, MCML, and ES, for example

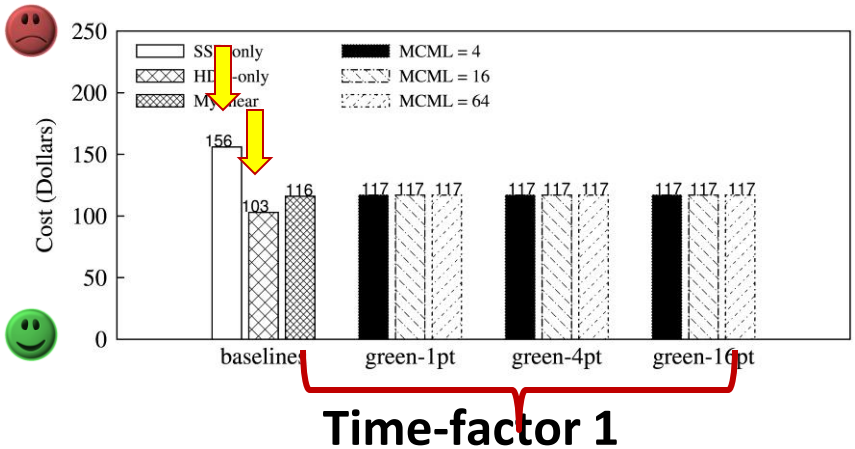


Web-Search Trace Replay

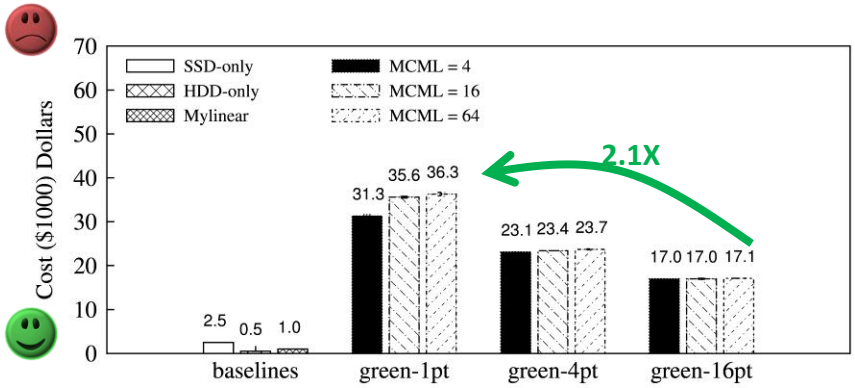
- Block trace from UMass trace repository
- Drive size: 32GB
- Average read size: 16KB (num: 1,055,236)
- Average write size: 8KB (num: 212)



Web-search Cost Results



Time-factor 1



Time-factor 100,000

([min, avg, max]=[0.2, 2.1, 7.7] yrs)

SSD-only: large purchase, small TCO

HDD-only: small purchase, small TCO

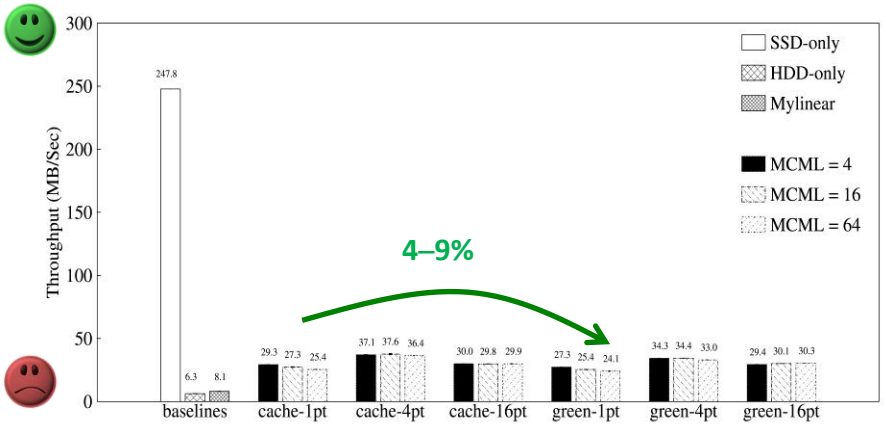
Hybrids: medium purchase, large TCO

GreenDM costs more than Mylinear

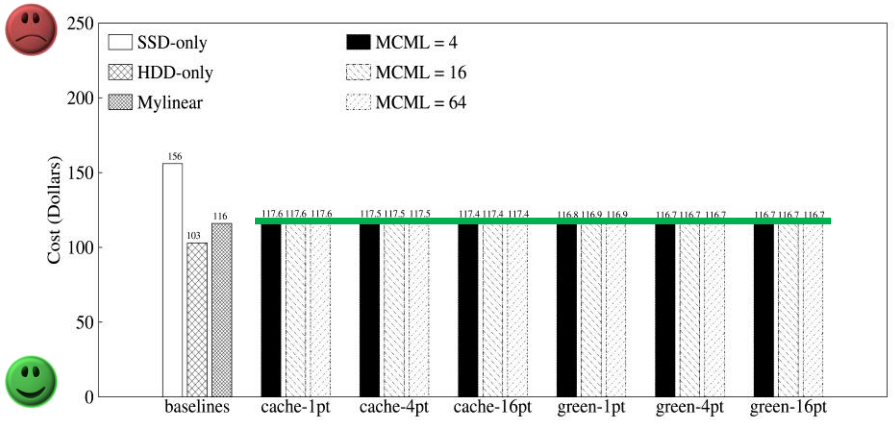
Larger PT, smaller TCO

Stony Brook University

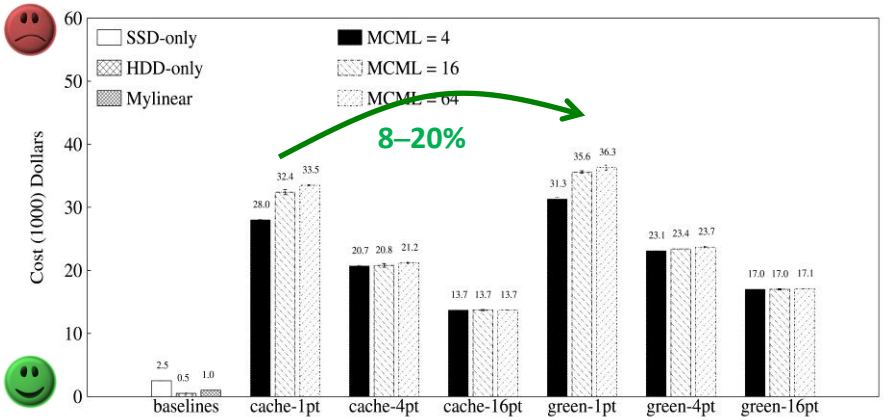
Web-search Tiering vs. Caching



Throughput



Time Factor 1



Time Factor 100,000

([min,avg,max]=[0.2, 2.1, 7.7] yrs)

- Read-intensive workload
- **Similar throughputs**
- **Similar short-term cost**
- **Tiering incurs larger TCO**
  - Aggregated primary SSD I/Os

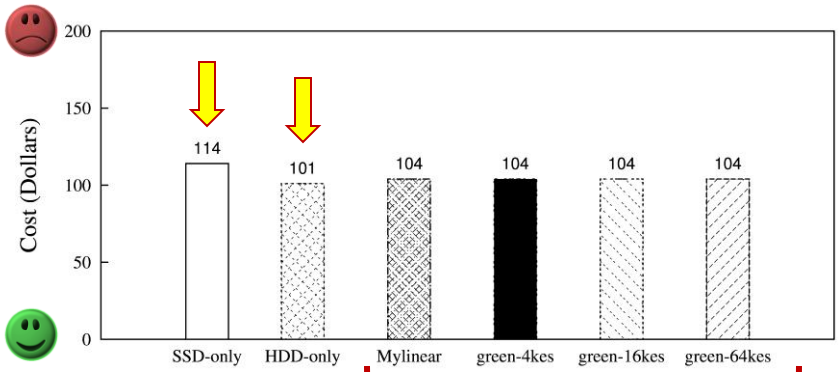
Stony Brook University

FIU Online Trace Replay

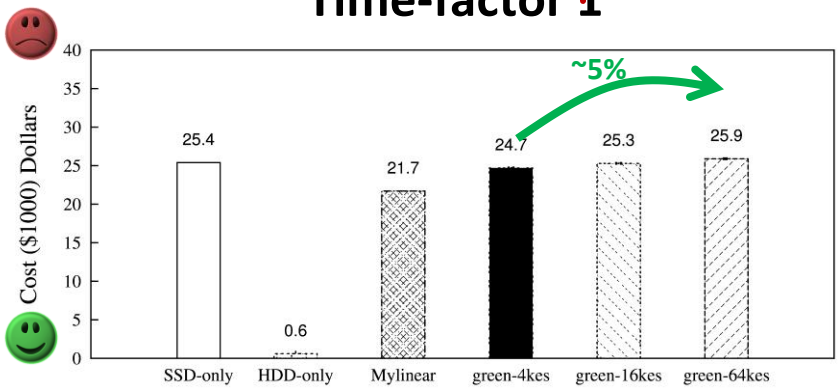
- Block trace from FIU trace repository
- Drive size: 8GB
- Average read size: 8KB (num: 655,526)
- Average write size: 4KB (num: 4,211,786)



FIU Online Cost Results



Time-factor 1



Time-factor 100,000

([min,avg,max]=[0.7, 3.3, 9.8] yrs)

SSD-only: large purchase, large TCO

HDD-only: small purchase, small TCO

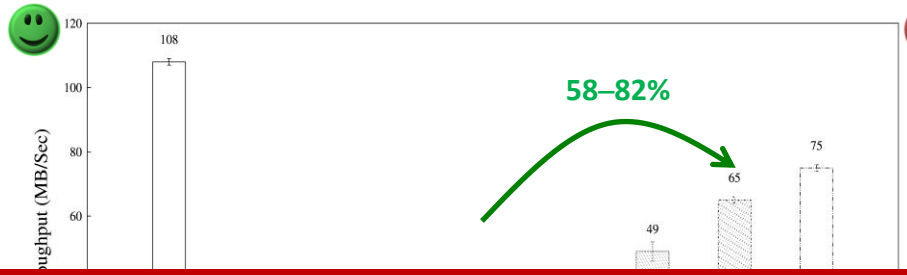
Hybrids: medium purchase, large TCO

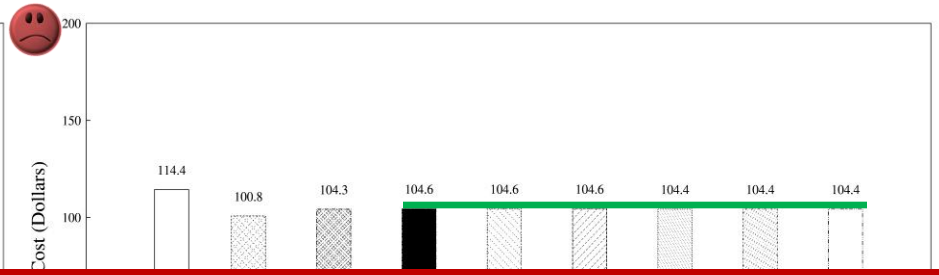
GreenDM costs more than Mylinear

Smaller ES, smaller TCO

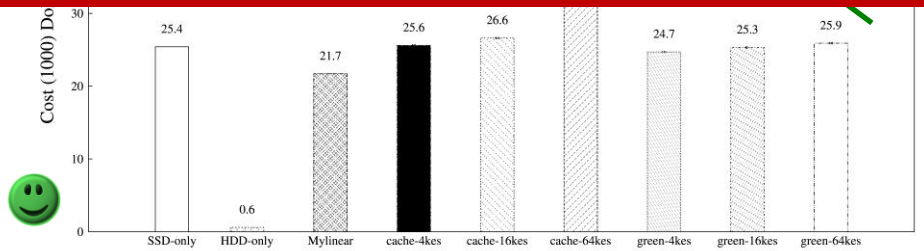
Stony Brook University

FIU Online Tiering vs. Caching





Evaluate storage systems across performance, long-term cost, and workloads



Time Factor 100,000

([min,avg,max]=[0.7, 3.3, 9.8] yrs)

- Caching write back

❑ **Similar short-term cost**

❑ **Caching incurs larger TCO**

- Write-back I/Os

Stony Brook University

Observed Trends Summary

- Parameters matter for TCO



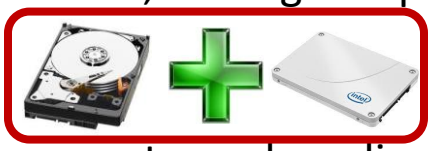
  - Read-intensive Web-search: larger PT
  - Write-intensive Online: smaller ES
- HDD-only drive



  - Least initial and long-term costs; but lowest performance
- SSD-only drive



  - Highest initial cost; various long-term costs; but highest performance
- Hybrids (two modes)



  - Medium initial costs; various long-term costs; and medium performance
  - Hybrids incurs more costs than Mylinear; but higher performance
- Versatility effects



  - Different system configurations lead to various long-term costs

Stony Brook University

Limitations

- Costs did not include several factors:
  - Computer hardware, air conditioning
  - Labor power, financing
- Simplified several conditions:
  - Data center hardware setup more complex
  - Data center workloads more complex
- Coarse-grained endurance model
- Journaling support



Related Work

- Simulation
  - **[HybridStore 2011]**
- No SSD endurance
  - **[EDT 2011, LSM 1996]**
- No TCO with SSD
  - **[Floyer 2009]**



Conclusion

- Cost model to justify performance gained
  - Upfront purchase cost

Hope for more long-term cost models of not only hybrids, but also all future storage systems

Q&A

On the Importance of Evaluating Storage Systems' $Costs

Zhichao Li, Amanpreet Mukker, and Erez Zadok

http://www.fsl.cs.stonybrook.edu

Stony Brook University

SSD Replacement Statistics

	Web-search					Online					File-server				
	SSD-only	HDD-only	My-linear	GreenDM	Cache	SSD-only	HDD-only	My-linear	GreenDM	Cache	SSD-only	HDD-only	My-linear	GreenDM	Cache
Time (yrs)	0.22	7.73	6.18	1.72	1.64	0.68	9.8	2.29	1.2	4.53	0.28	1.46	1.33	0.81	1.1
# SSD	4.3	0	0.96	46.76	41.13	47.7	0	40.6	47.43	52.06	25.1	0	8.36	60.12	34.62
Days per SSD lasts	18.67	∞	2349.7	13.43	14.55	5.2	∞	20.59	9.23	31.76	4.07	∞	58.07	4.92	11.6

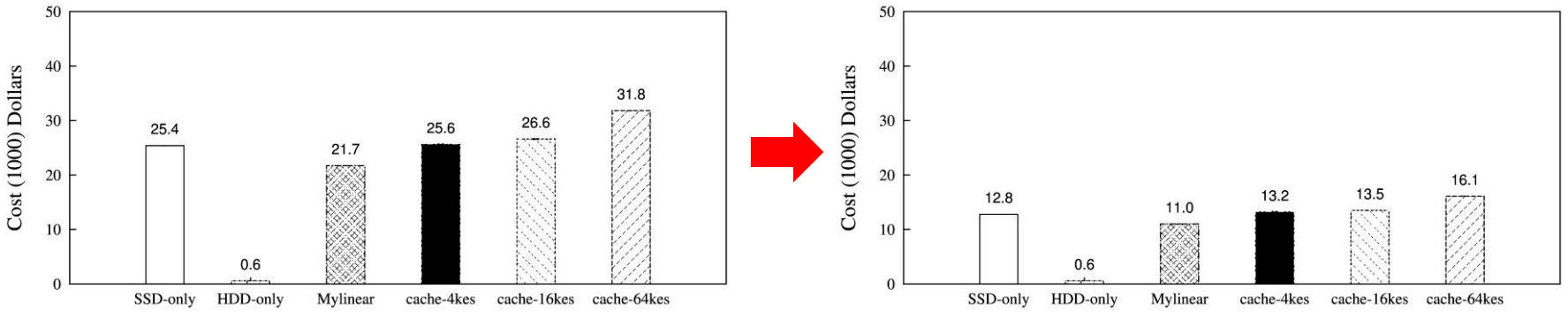
Statistics with time factor 100,000

More failures as storage systems become complex

[Schroeder CFDR'07]

Stony Brook University

What if Unit Prices Change?



- Online with time factor 100,000:
  - HDD/SSD prices are reduced by half
  - **Different absolute values, but similar trends**

Comments on Flash-Array?

- Products: Dell Compellent, NetApp EF550, Pure Storage FA-400, etc
- Super fast performance, but with challenges:
  - Mis-matching read and write performance
  - Co-related failures
  - Parity Drive ages faster
  - The stripe size matters
  - The aging factor

[Jeremic SYSTOR'11]

Others?

- Less aggressive caching write-back?
  - Absolute value may change
  - Negative effects of write-back won't change
- Counting metadata flush?
  - Fairly small to be ignored (<<1%)
  - Not closely related to the trade-off study
  - Presumably store in the SSD reserved space
- SSD/HDD ratio?
  - Example (1/4) and looking into (1/8)
  - Such SSD size is small compared with WL size



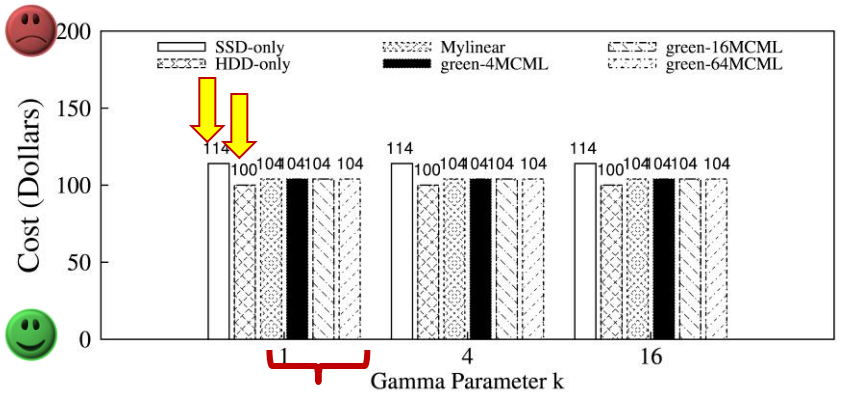
Stony Brook University

File-Server Workload

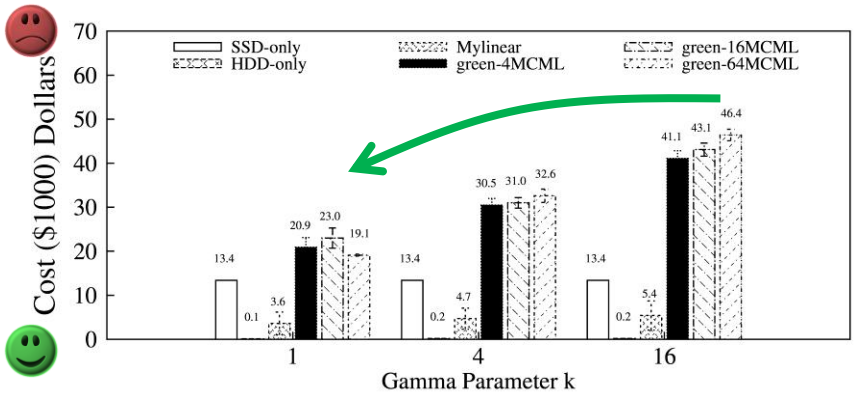
- Filebench with Gamma distribution
- Drive size: 8GB
- RAM size: 1GB
- Average I/O size: 128KB



File-server Cost Results



Time-factor 1



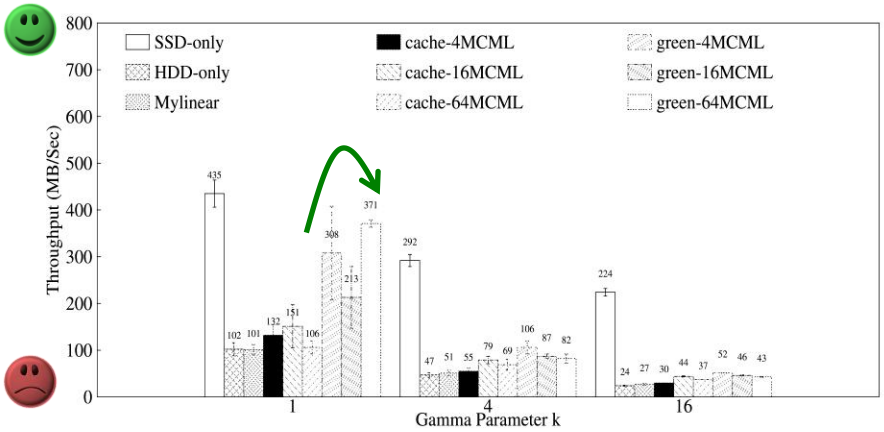
Time-factor 100,000

([min,avg,max]=[0.2, 1.0, 2.4] yrs)

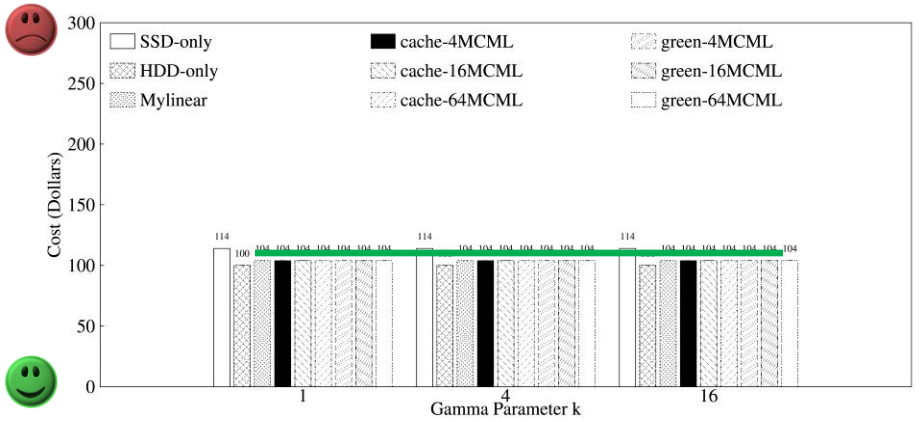
- SSD-only: large purchase, medium TCO
- HDD-only: small purchase, small TCO
- Hybrids: medium purchase, large TCO
- GreenDM costs more than Mylinear
- Smaller Gamma, smaller TCO

Stony Brook University

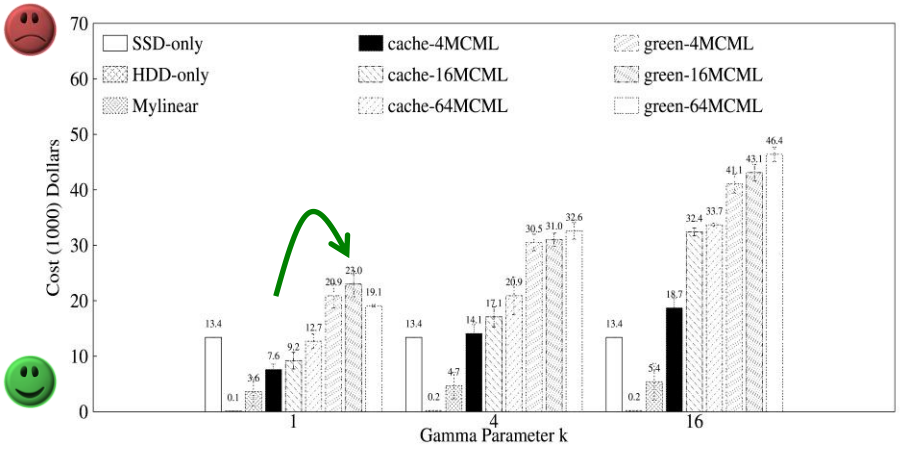
File-server Results



Throughput



Time Factor 1



Time Factor 100,000

([min,avg,max]=[0.2,1.0,2.4] yrs)

- Reads/writes workload
- **Tiering higher throughputs**
  - Caching write-back
- **Similar short-term cost**
- **Tiering incurs larger TCO**
  - Aggregated primary SSD I/Os

Stony Brook University

All about Storage Devices

Device	Performance	Price	Capacity	Endurance	Power
Tape	Low+	Low++	Large+	Large	Low
HDD	Low	Low	Large	Large	High
SMR	Low	Low+	Large++	Large	High
SSD	High	High	Small	Small+	Low
PCM	High+	High+	Small+	Large	Low
DRAM	High+			Large	Low
SCM	N/A			N/A	N/A

- SSD: SLC, ML

Pick your



by yourself!

Stony Brook University

Why not Separating Reads/Writes?

- Reads/writes are equal to HDD
- Reads/writes are unequal to SSD
- Yes, it is possible to distinguish
- Migration model will be more complex
- Our trade-off study will be complex as well
- Future work to look even deeper



Workloads Mainly Reside in DRAM?

- What previous systems want?
  - Performance and transaction
  - But it costs more
  - As the WL grows, colder data goes to cheaper device
- Data filtered out of DRAM
  - Hybrid systems better trade-off performance and cost



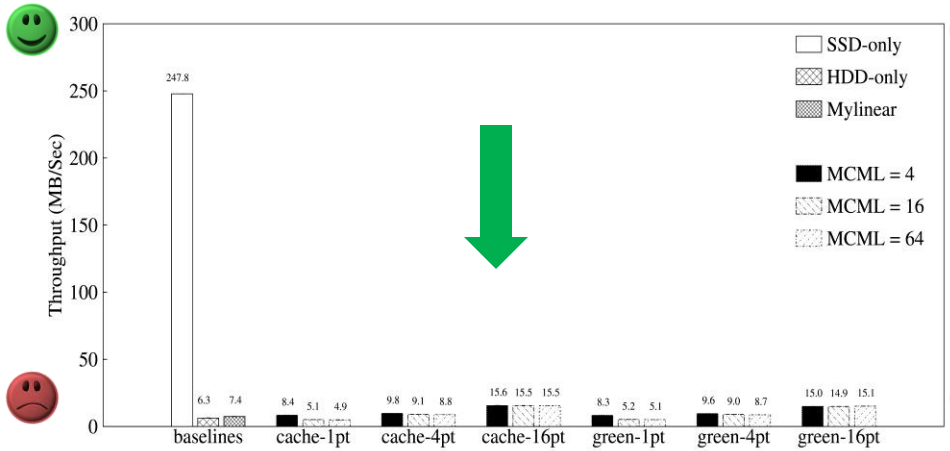
To Live or To Die? That is a Question

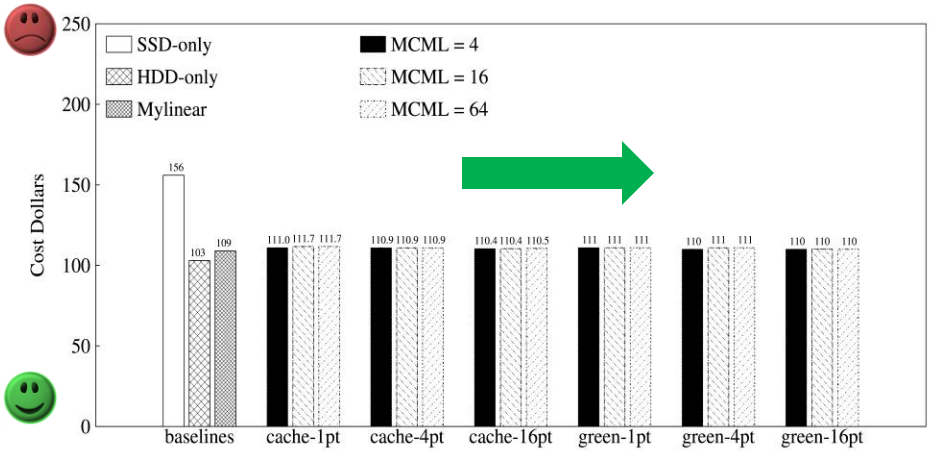
- DRAM?
  - Die
    - SCM (e.g., PCM) blurs the boundary between DRAM and Disk and may replace DRAM (cheaper)
    - SCM is comparably fast. Complex DRAM cache may turn out to hurt the performance, instead
  - Live
    - OS needs a clean state sometimes

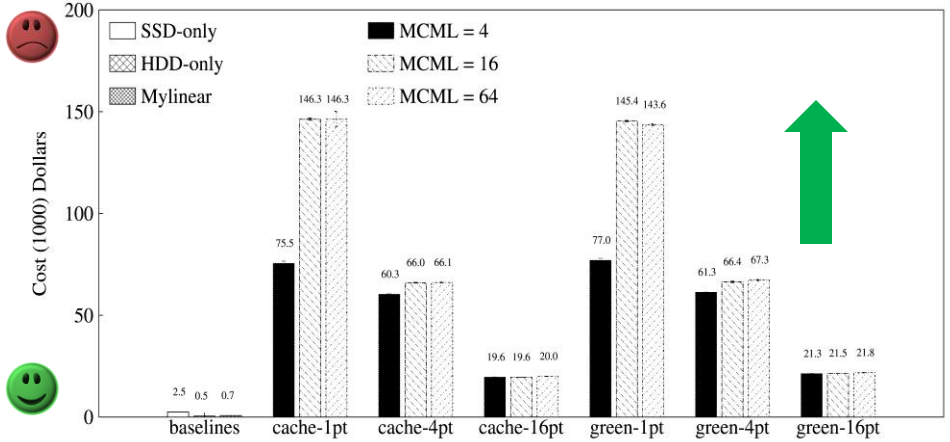


- What if die?
  - Then, all I/O goes directly to lower-level systems

Web-search (Capacity Ratio: 1/8)



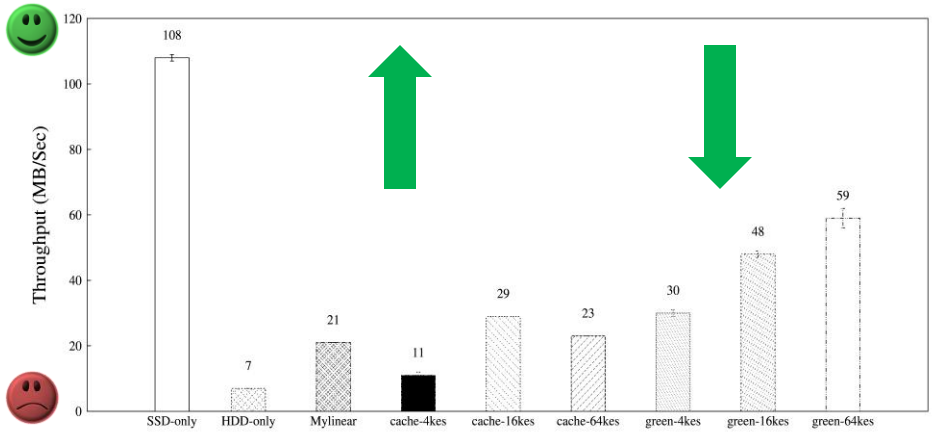


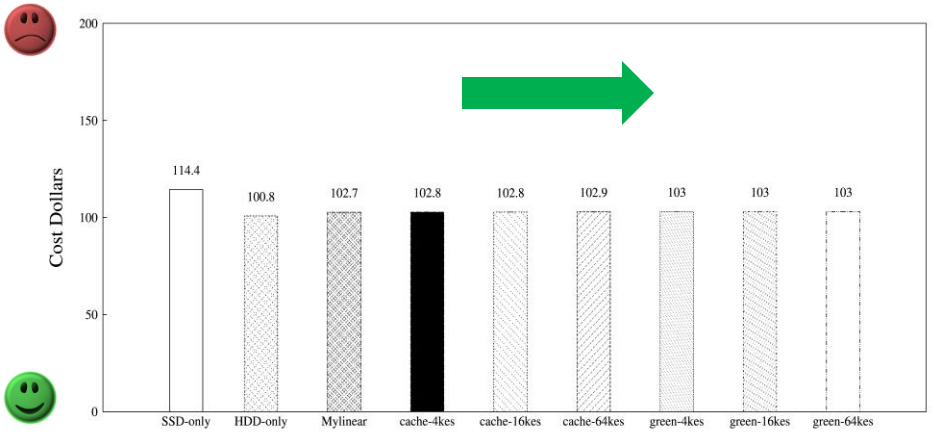


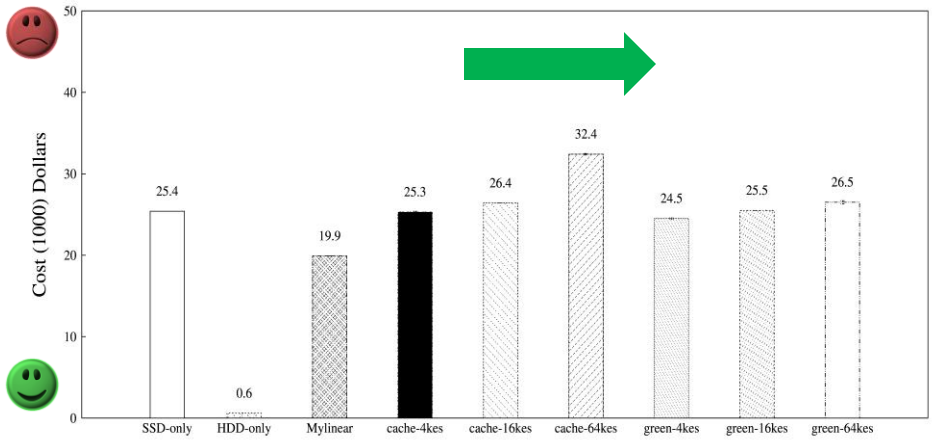
- Read-intensive workload
- **Decreased throughput**
  - More data movement
  - Less SSD hit
- **Similar short-term cost**
- **Increased long-term cost**
  - More SSD accesses

Stony Brook University

Online (Capacity Ratio: 1/8)



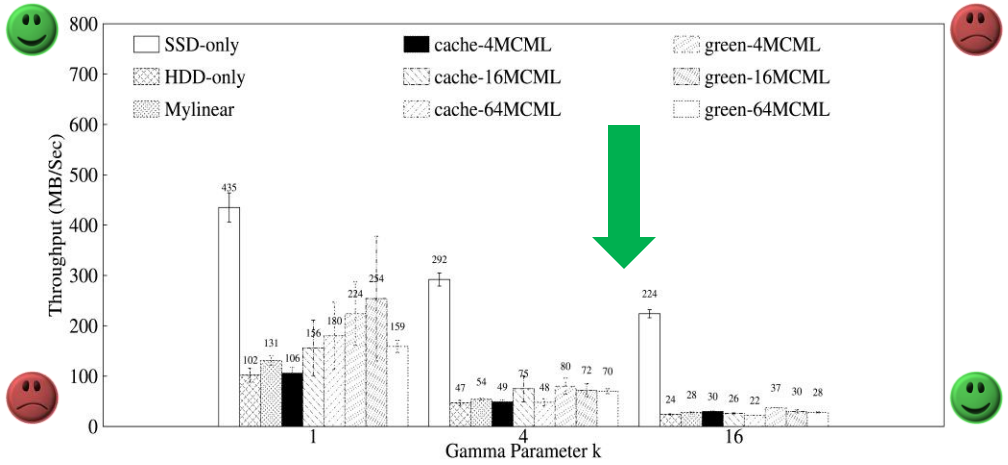


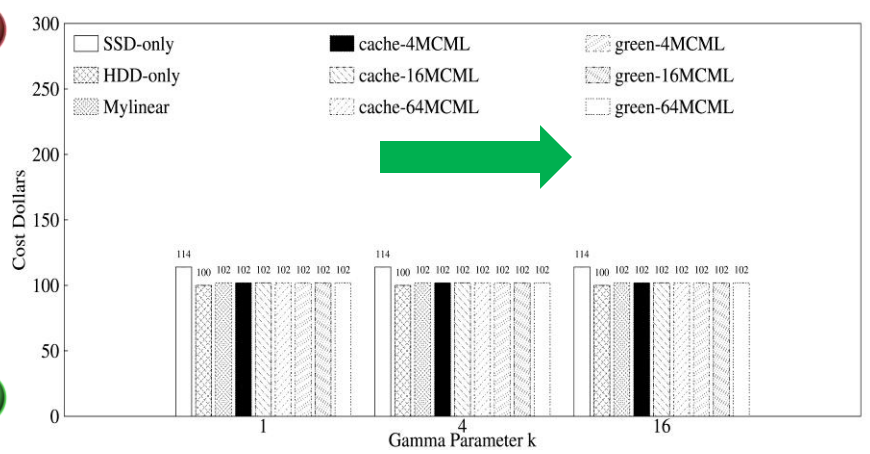


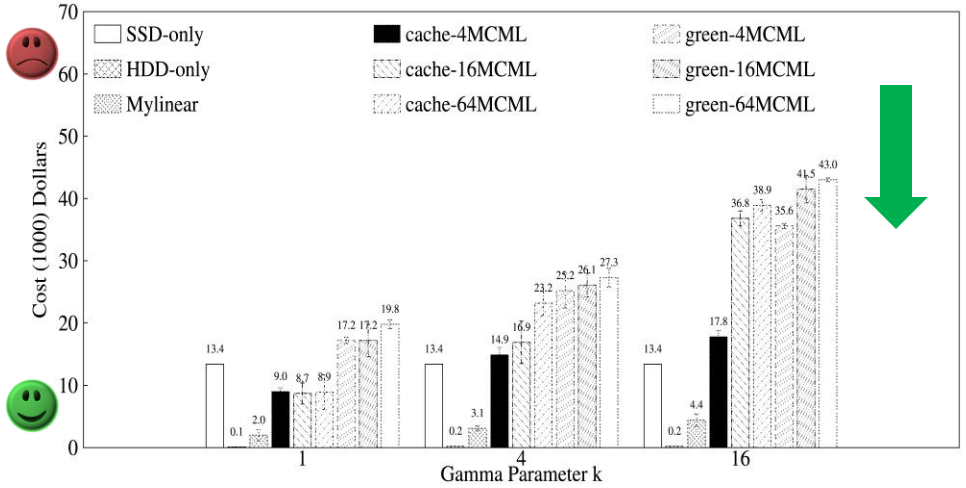
- Write-intensive workload
- **Increased caching throughput**
  - Less write-back I/Os
- **Decreased tiering throughput**
  - More Migrations and less SSD hit
- **Similar short-term cost**
- **Similar long-term cost**
  - Net effect

Stony Brook University

File-server (Capacity Ratio: 1/8)







- Read-intensive workload
- **Decreased throughput**
  - More data movement
  - Less SSD hit
- **Similar short-term cost**
- **Decreased long-term cost**
  - Less SSD accesses

Stony Brook University