reFresh SSDs: Enabling High Endurance, Low Cost Flash in Datacenters

[1]Vidyabhushan Mohan

[2]Sriram Sankar

[1]Sudhanva Gurumurthi

[1]Department of Computer Science, University of Virginia

[2]Microsoft

Storage Systems in Datacenters

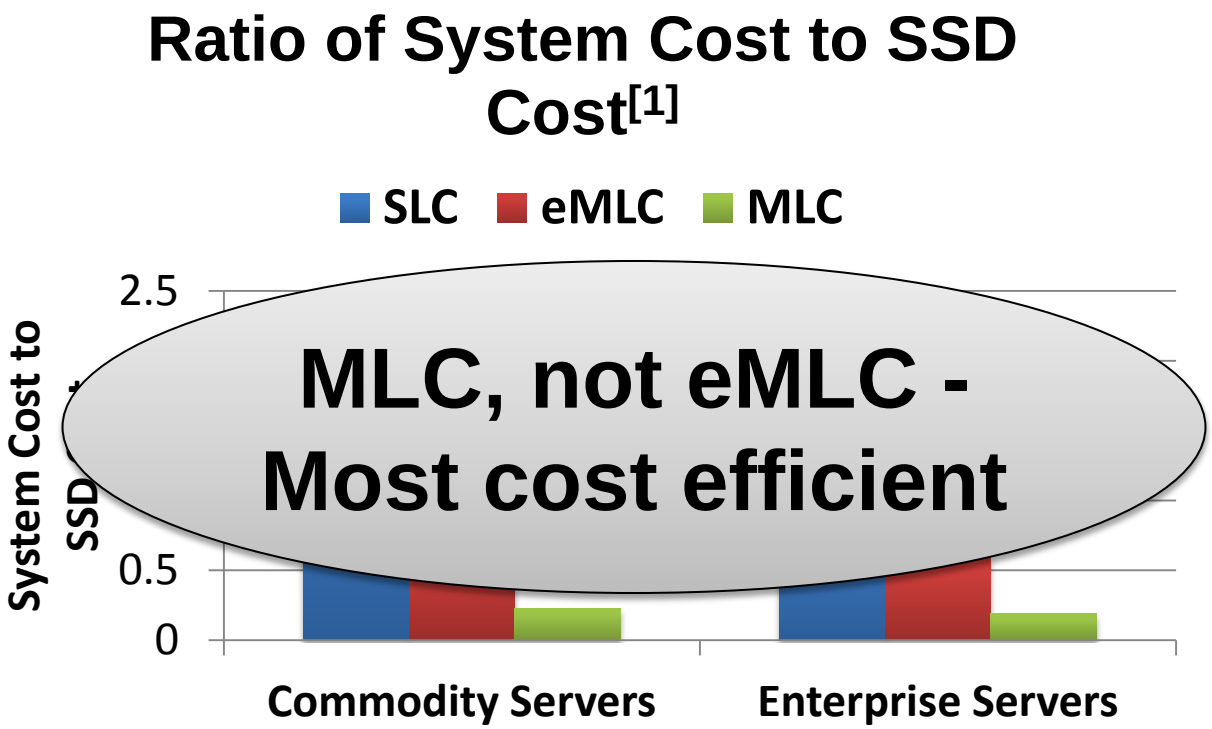
Solid State Disks

Fast - Yes

Cheap – Yeah, kind of…

Reliability - Depends

Cost of SSDs in Datacenters



SSD Capacity

Commodity Servers – 300GB

Enterprise Servers – 450GB

[1] Amazon.com. As of September 2011

Relative Cost of SSDs

Type of SSD	$/GB [1]	Relative Endurance @ 3xnm [2]

This talk: How to make MLC SSDs usable in Datacenters?

W **LC?**

- ❑ More
- ❑ Rigorous qualification
- ❑ Special firmware
- ❑ Better controller

eMLC flash cells use lower operating voltages for writes and erasures (lowers performance)

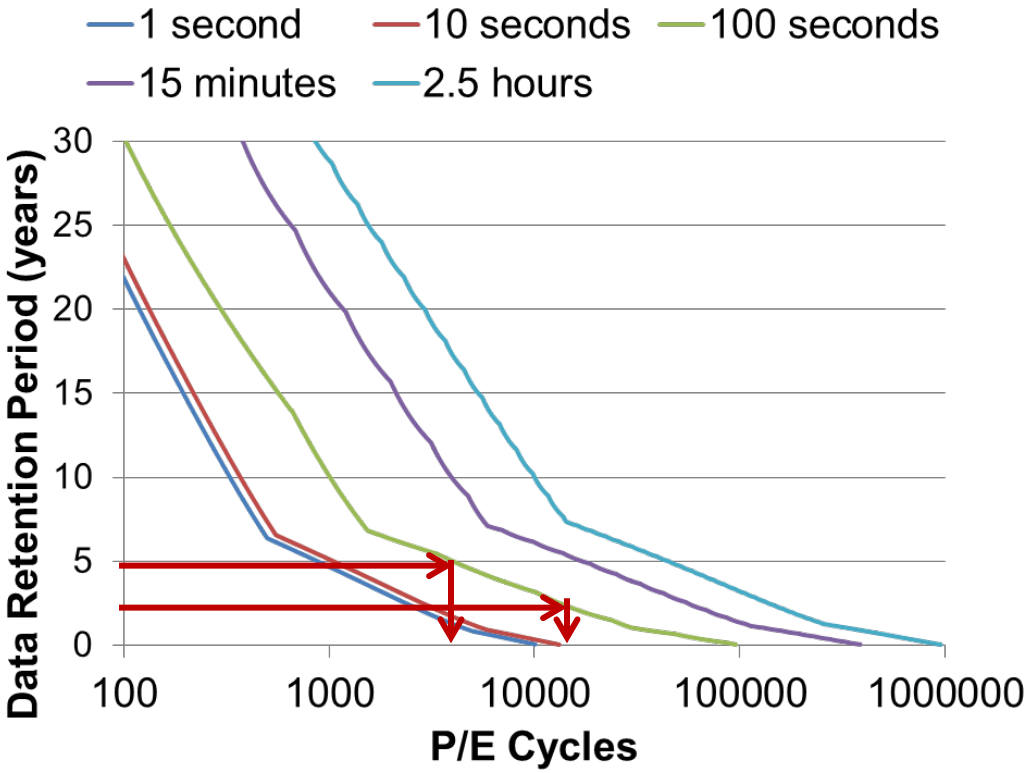
[1] Amazon.com. As of September 2011
[2] http://www.flashmemorysummit.com/English/Collaterals/Proceedings/2011/20110809_F2C_Wu.pdf

Agenda

- Tradeoff between endurance and data retention
- SSDs and datacenter workloads
- reFresh SSDs – Architecture and Operation
- Design and Evaluation

Tradeoff between Endurance and Data Retention for 2-bit MLC

Impact of P/E Cycle Time on Data Retention



Important Parameters
Feature Size (F) – 80nm
Temperature – 30 C

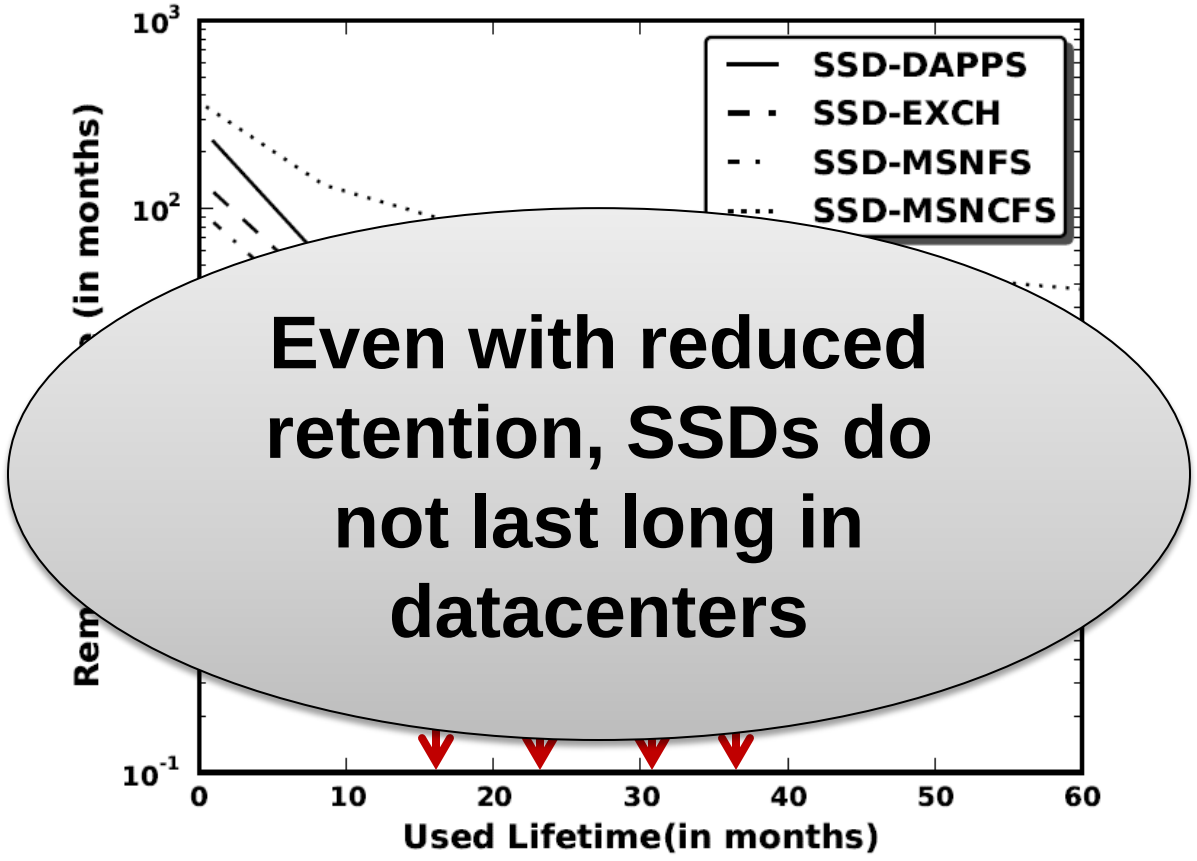
Workload Traces

Workload [3]	Total I/Os (millions)	Read/Write Ratio
Display Ads Platform Payload Server (SSD-DAPPS)	10.9	1:1.2
Exchange Server (SSD-EXCH)	22	1:2.2
MSN File Server (SSD-MSNFS)	15.54	1:1.2
MSN Metadata Server (SSD-MSNCFS)	7.8	1:0.64

SSD traces extrapolated from HDD I/O traces of enterprise workloads

[3] HDD Traces from IOTTA Trace Repository from SNIA - http://iotta.snia.org/

How Long Do Enterprise SSDs Last?

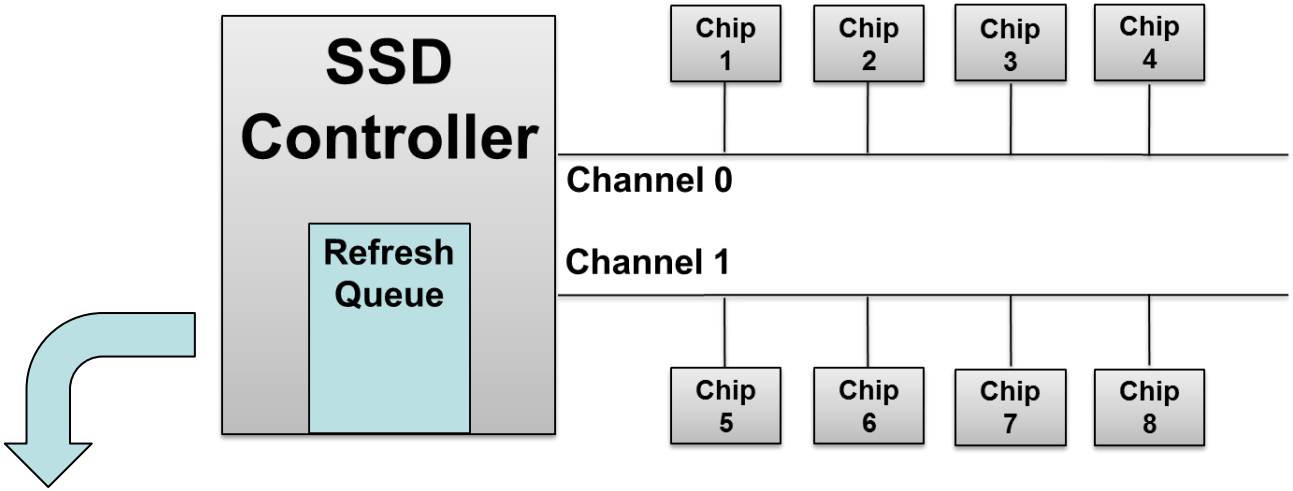


Endurance

reFresh SSDs: Making MLC SSDs Usable in Datacenters

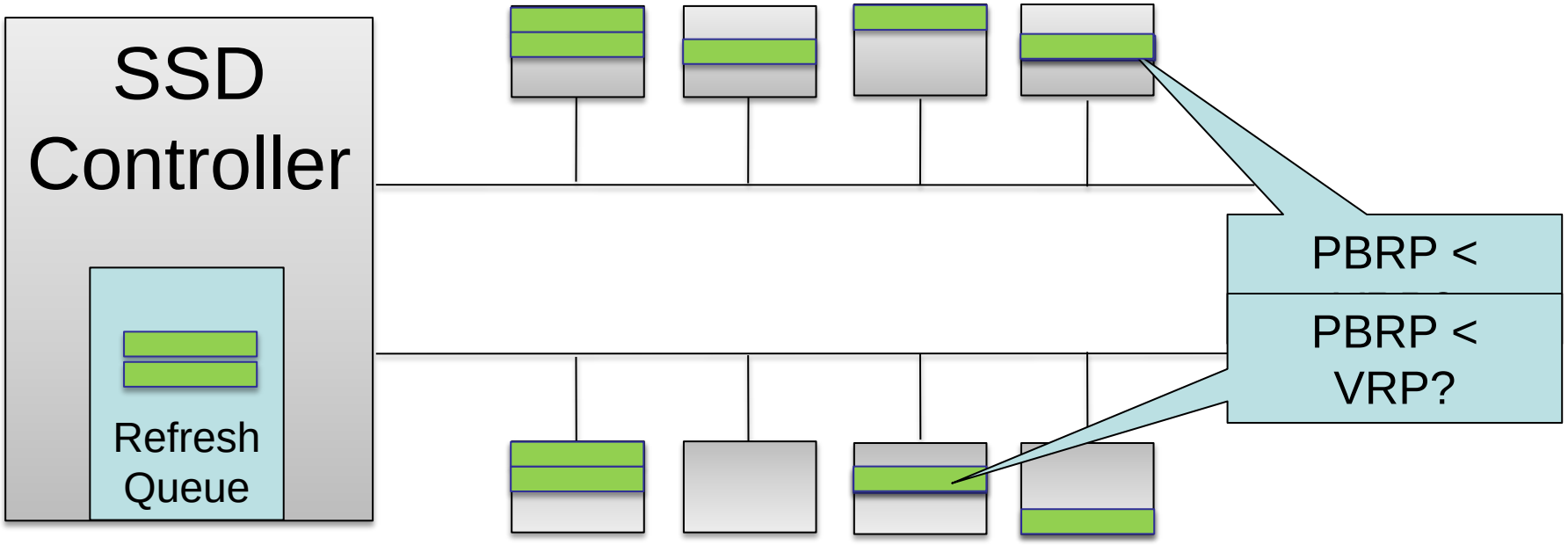
- Uses low endurance MLC flash.
  - Low cost, high performance (compared to eMLC)
- Useful for enterprise applications which do not require high data retention.
  - Tradeoff retention for higher endurance
- Exploit and Export application's knowledge of data lifetime to increase SSD lifetime.
  - Applications with different lifetime requirements can co-exist

reFresh SSDs: Architecture



- Refresh Queue
  - Managed by the SSD controller
  - Queue entries – **Pointers** to physical flash blocks that have valid data
  - Priority queue – Sorted by block lifetime
    - Most important blocks to be refreshed are at the head

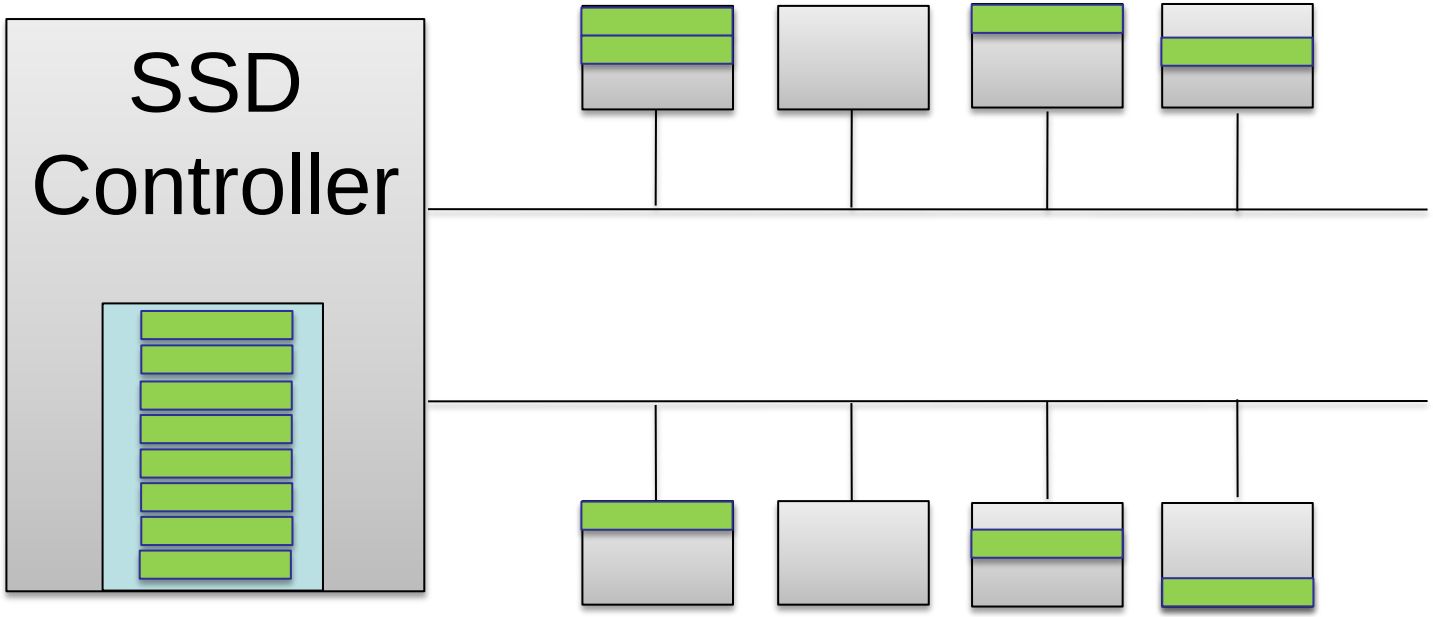
reFresh SSDs: Operation



PBRP – Block lifetime (Physical Block Retention Period)
VRP – Application specified lifetime (Virtual Retention Period)

reFresh SSDs: Operation

Refresh operation invoked at regular intervals on blocks in the refresh queue

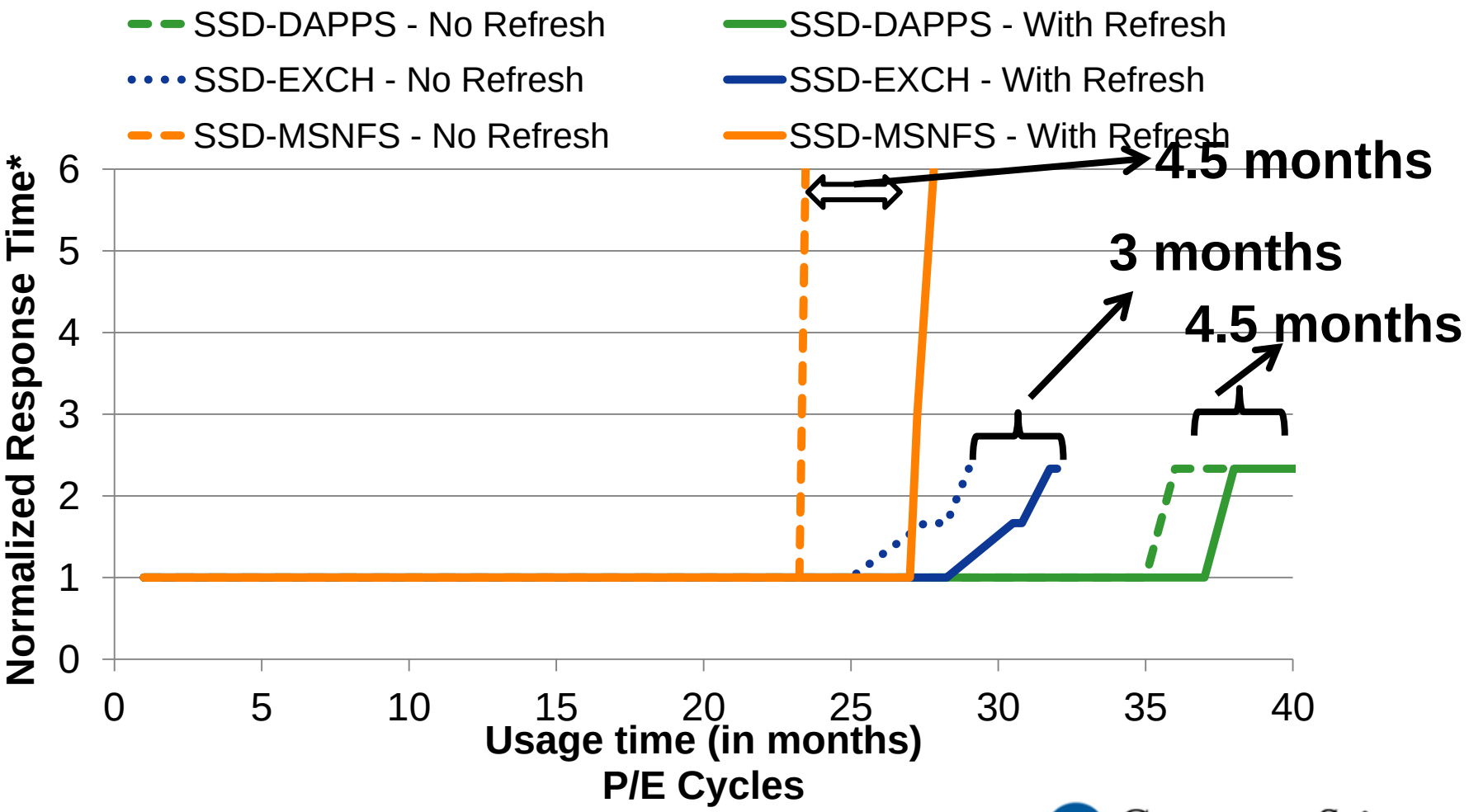


Unlike wear leveling, refresh operations are triggered to handle a immediate deadline (PBRP < VRP)

Evaluating reFresh SSDs

- Metrics
  - Endurance
  - Variation of performance with age
- Input Parameters
  - Data lifetime (as specified by the application)
  - SSD properties
  - Enterprise application I/O traces

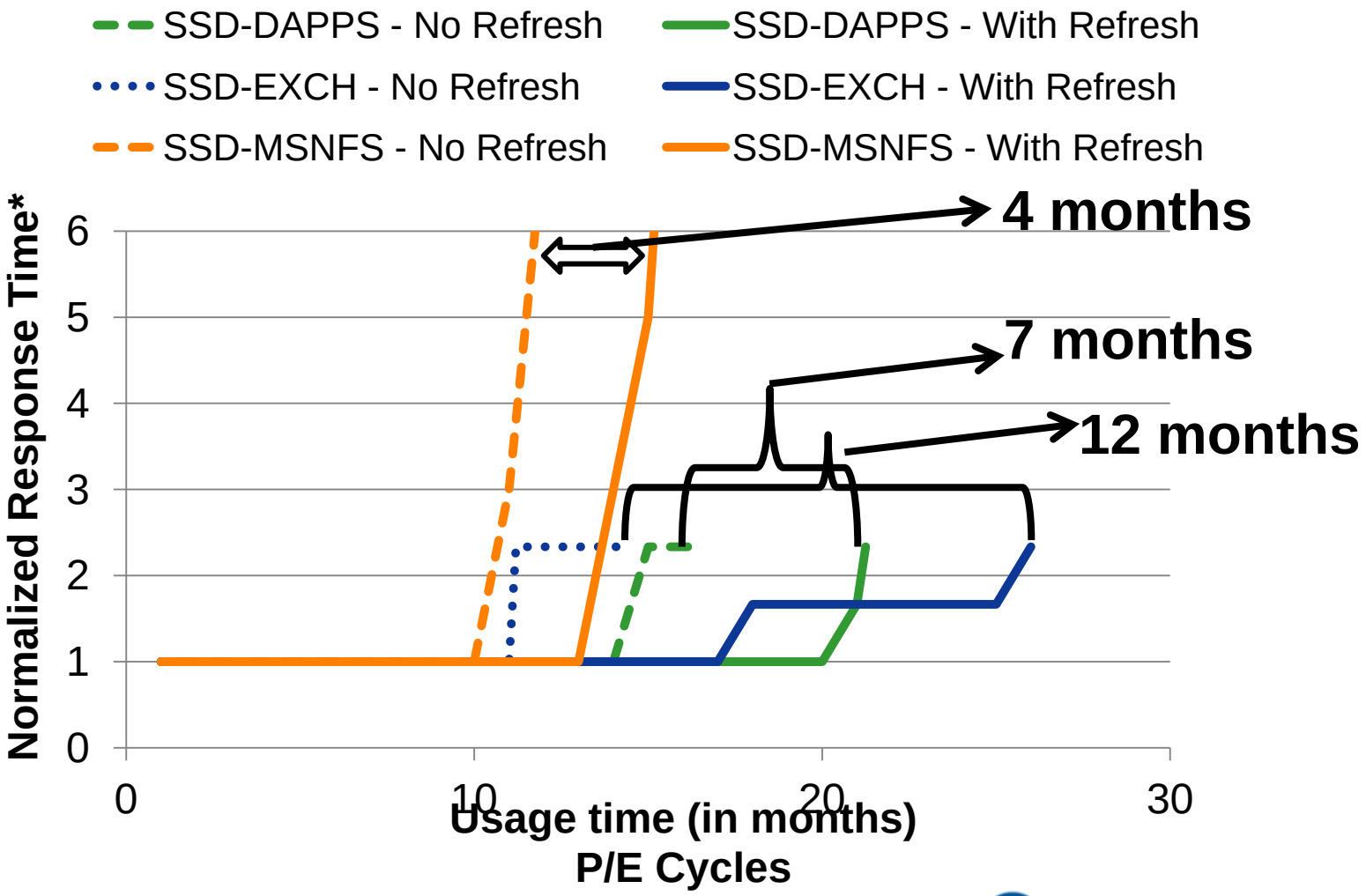
Evaluating reFresh SSDs with 1 month Retention



*Normalized response time at 80th percentile
Lower the better

Computer Science at the University of Virginia

Evaluating reFresh SSDs with 1 year Retention



*Normalized response time at 80th percentile
Lower the better

Designing reFresh SSDs

- Controller Modifications
  - Manage a refresh queue to keep track of block lifetime
  - Store additional metadata for each page
    - Data lifetime, block lifetime
  - No hardware change required, just modify firmware
- Host/Interface Modifications
  - Applications provide data lifetime information to the SSD controller
  - NVM Express already provides dataset management commands
  - Extend the command set to provide data lifetime

Conclusion

- reFresh SSDs
  - Uses low endurance flash
  - Smart controller design to increase SSD lifetime
  - Uses application specified data lifetime.
  - Applications with different retention period requirements can co-exist
  - Increases SSD lifetimes by 6-56% for various enterprise workloads

Flash Memory SUMMIT

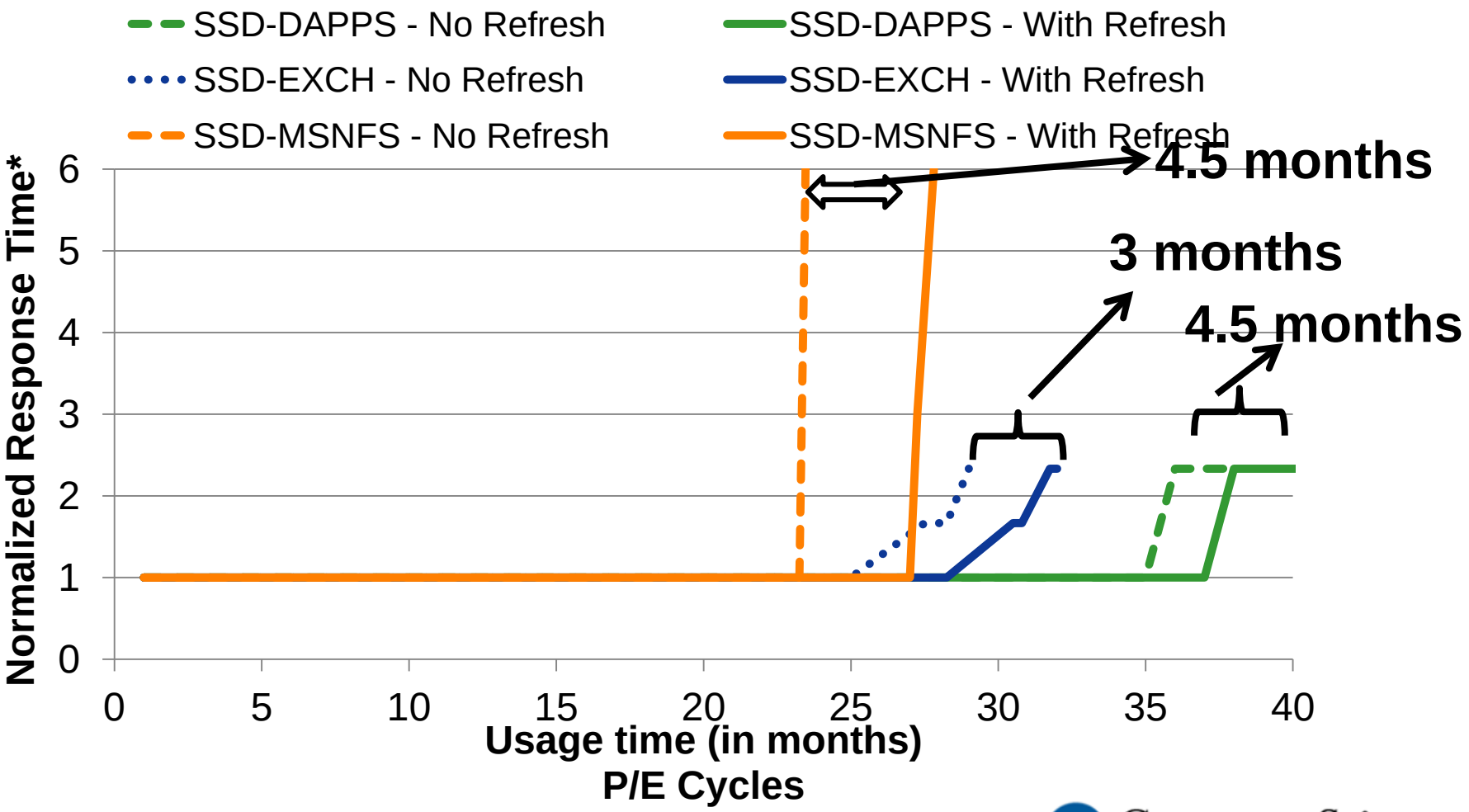
Questions?

mohan@cs.virginia.edu

Paper here - www.cs.virginia.edu/~vm9u

Backup slides

Evaluating reFresh SSDs with 1 month Retention



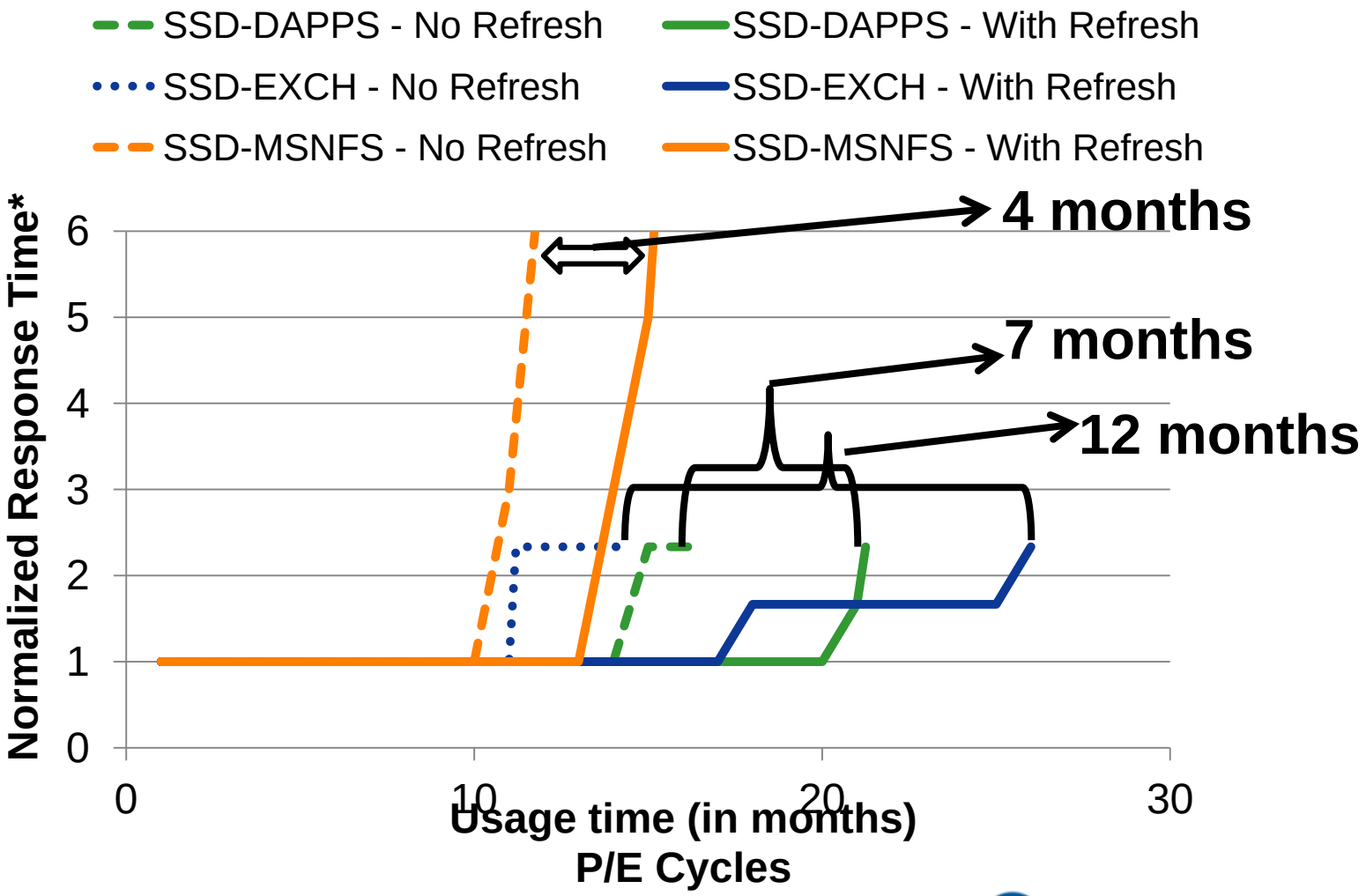
*Normalized response time at 80th percentile
Lower the better

Computer Science at the University of Virginia

Refresh vs Wear Leveling

Wear Leveling	Refresh Operations
Triggered to balance wear-out of blocks	Triggered when PBRP drops below VRP
Manages reliability over a long term	Triggered due to handle immediate deadlines
No knowledge about data lifetime	Takes data lifetime into account.

Evaluating reFresh SSDs with 1 year Retention



*Normalized response time at 80th percentile
Lower the better