SALSA

Flash-Optimized Software-Defined Storage

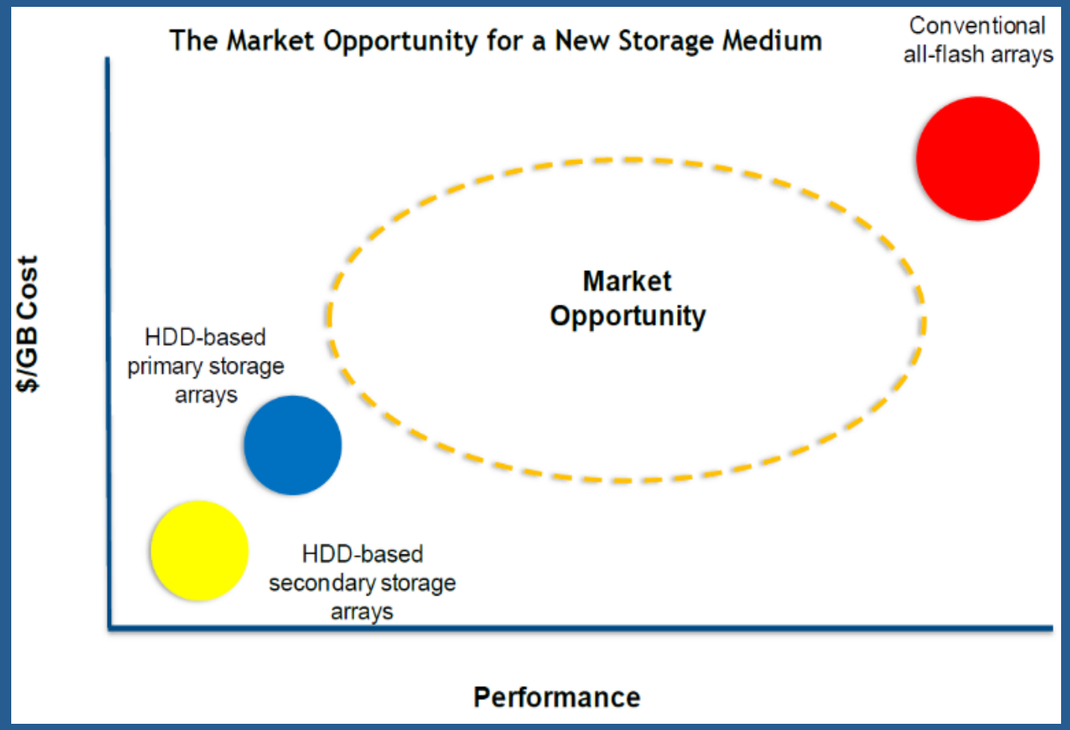
Nikolas Ioannou, Ioannis Koltsidas, Roman Pletka, Sasa Tomic,Thomas Weigold

IBM Research – Zurich

Flash Memory Summit 2015 Santa Clara, CA

New Market Category of **Big Data Flash**

- Multiple workloads **don't really need the write performance** and **endurance** of "good' Flash
  - In certain environments data is actually immutable
- What matters is **high density, low cost, and good read performance**
  - Current Flash architectures are not a good fit




Forbes
New Posts
+37 posts this hour
Most Popular
Buffett's Billion Dollar Bracket
Lists
TECH | 8/14/2013 @ 12:30PM | 2,186 views
Cold Flash And A New Dimension For Flash Memory
The Register
Data Centre Software Networks Security Policy Business Jobs Hardware Science Bootnotes
Servers HPC Cloud Storage Data Networking Virtualisation BOFH
Facebook's request to the flash industry: 'Make the worst flash possible'
Counterintuitive, perhaps – but eminently sensible

eBay: "*We could live with 1/3rd the number of writes that normal flash supports as long as we could get it for 1/4th the price.*"

- IDC just introduced a new market category **of Big Data Flash** (March 2015)
- Content repositories, media and streaming services, Big Data and analytics, NoSQL, Object storage, Web infrastructure.

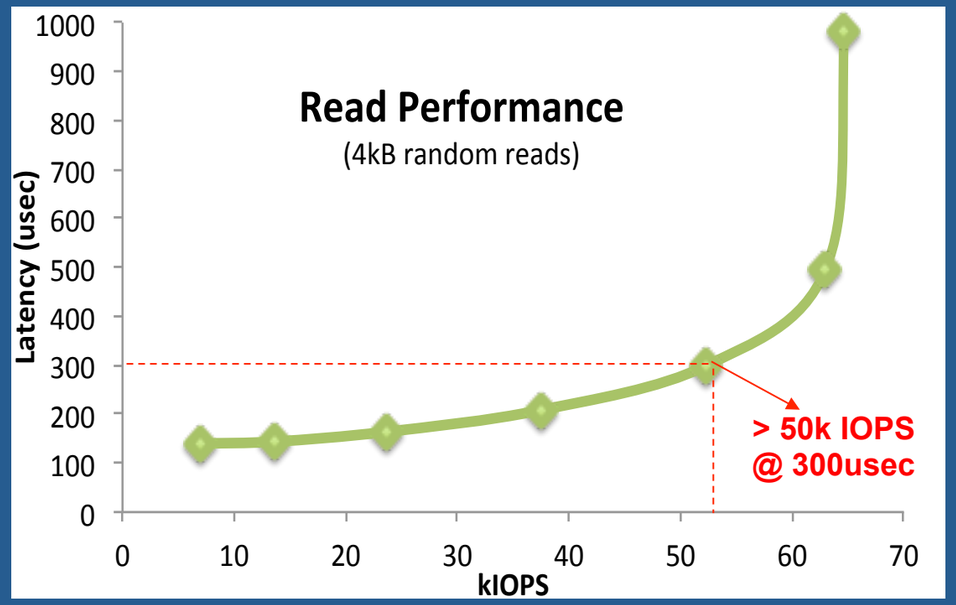
At <1$/GB for raw Flash, total acquisition cost becomes the same as an HDD-based solution, with much lower TCO.
- IDC

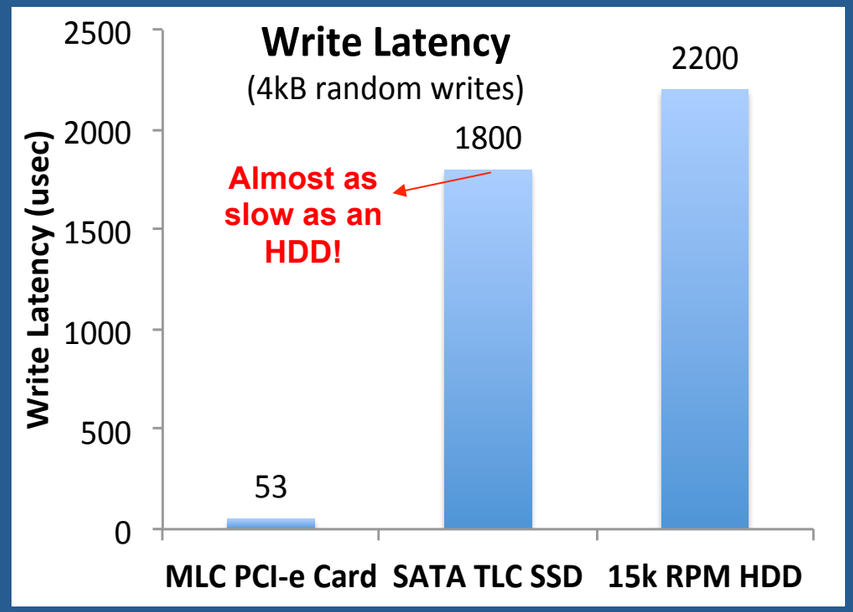
Low-cost Flash technology (c-MLC, TLC)

Can't we just use low-cost SSDs?

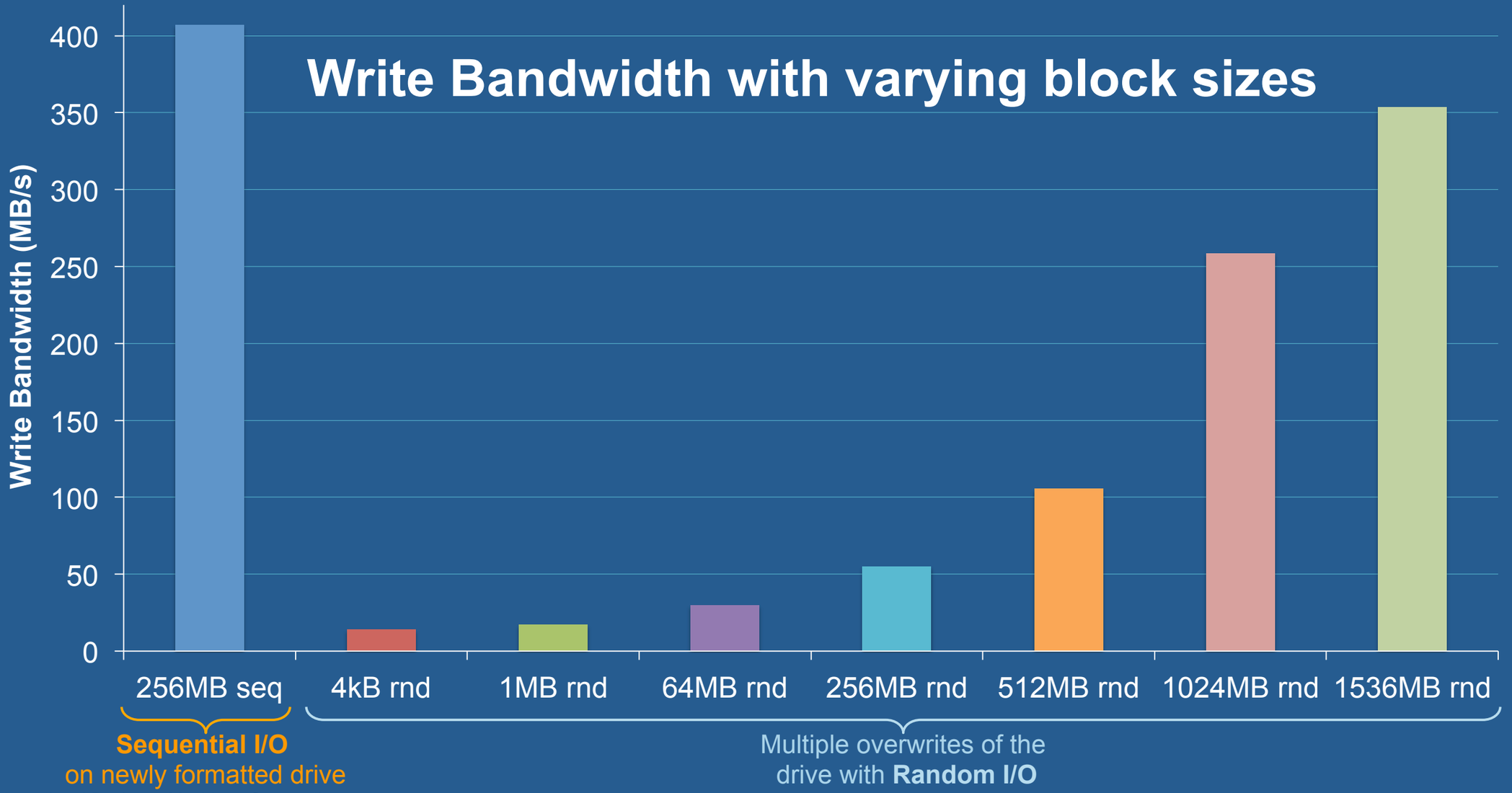
- Low-cost Flash suffers from high write latency, low endurance
  - E.g., TLC, 3D-NAND, c-MLC
- Low-cost SSDs have limited resources, simple controllers to keep the cost as low as possible (~ $0.4 /GB!)
- Therefore, they only employ simple Flash management
  - Sufficiently good read performance
  - But, **limited write endurance, terrible write performance**

Raw low-cost SSDs are **practically unusable** in a real datacenter





The characteristics of write performance



SoftwAre Log-Structured Array

What?
A **Flash-optimized** I/O stack that elevates the performance and endurance of consumer-level SSDs to enterprise standards.

Why?
Offer **cost-effective all-Flash** storage in public and private clouds, mainly for read-dominated workloads, complementing our high-end FlashSystem offerings.

How?
1. Use high-density, low-cost, off-the-shelf Flash SSDs
2. Move complexity from hardware to software to reduce cost
3. Optimize **end-to-end** for low Write Amplification
4. Employ aggressive Data Reduction
5. Natively support Object Storage

Squeeze the most capacity out of Flash

SALSA

- ✓ Implements the state-of-the-art Flash Management **in software**
- ✓ Runs on **Linux**, exposes **standard interfaces**
  - File-systems and applications run unmodified on top of SALSA
- ✓ Is ideal for cost-optimized scale-out storage systems like GPFS, CEPH
  - SALSA enables SDS on low-cost SSDs, offering high performance and endurance

SALSA Overview

SALSA Software Stack

Interfaces

- Block
- Object
- I/O memory (RDMA)

Logical Layer

- Storage Virtualization
- Quality of Service
- Data Reduction

- Workload Isolation
- Thin Provisioning
- **Compression**
- **De-duplication**
- Recurring Pattern Detection

Physical Layer

- Log-Structured Array
- Capacity Management
- Traffic Shaping
- Load Balancing
- I/O handling

- **Log-structured organization**
- Flash-friendly access patterns
- **State-of-the-art Garbage Collection**
- **Zero Read-Modify-Writes**
- RAID5-equivalent protection
- Small footprint

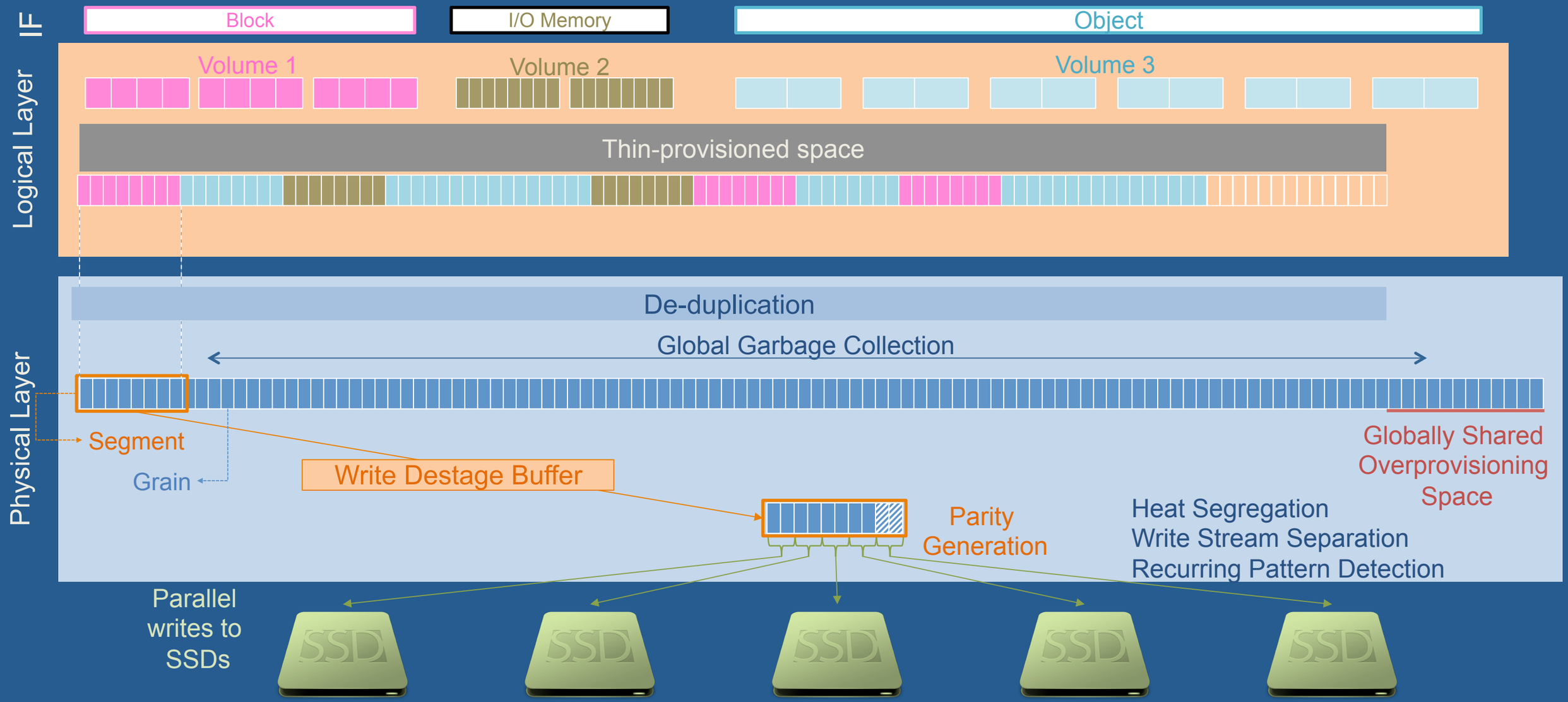
SATA

TLC **3D NAND** **c-MLC**

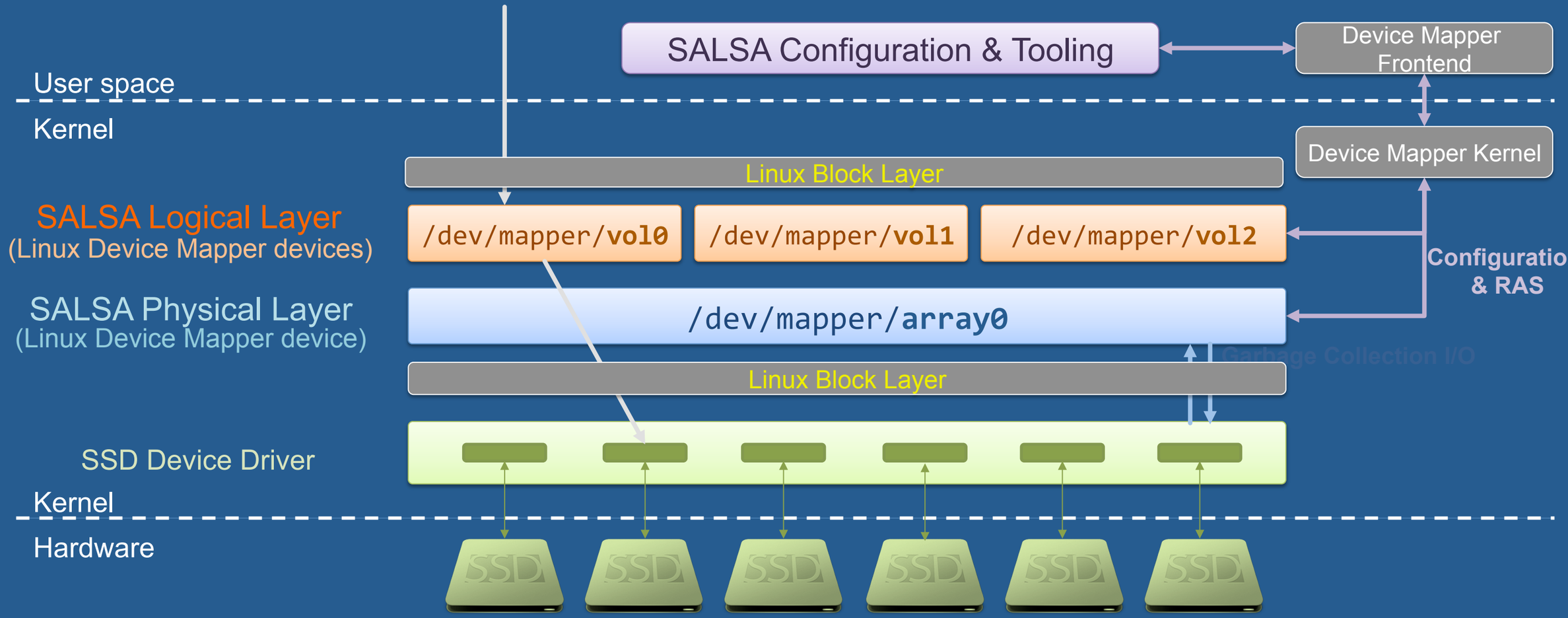
Low-cost, high-density consumer SSDs
- Limited resources (FPGA, CPU, RAM)
- Light Flash Management, simple GC

Runs on Linux,
Intel x86 and Power8

SALSA Stack

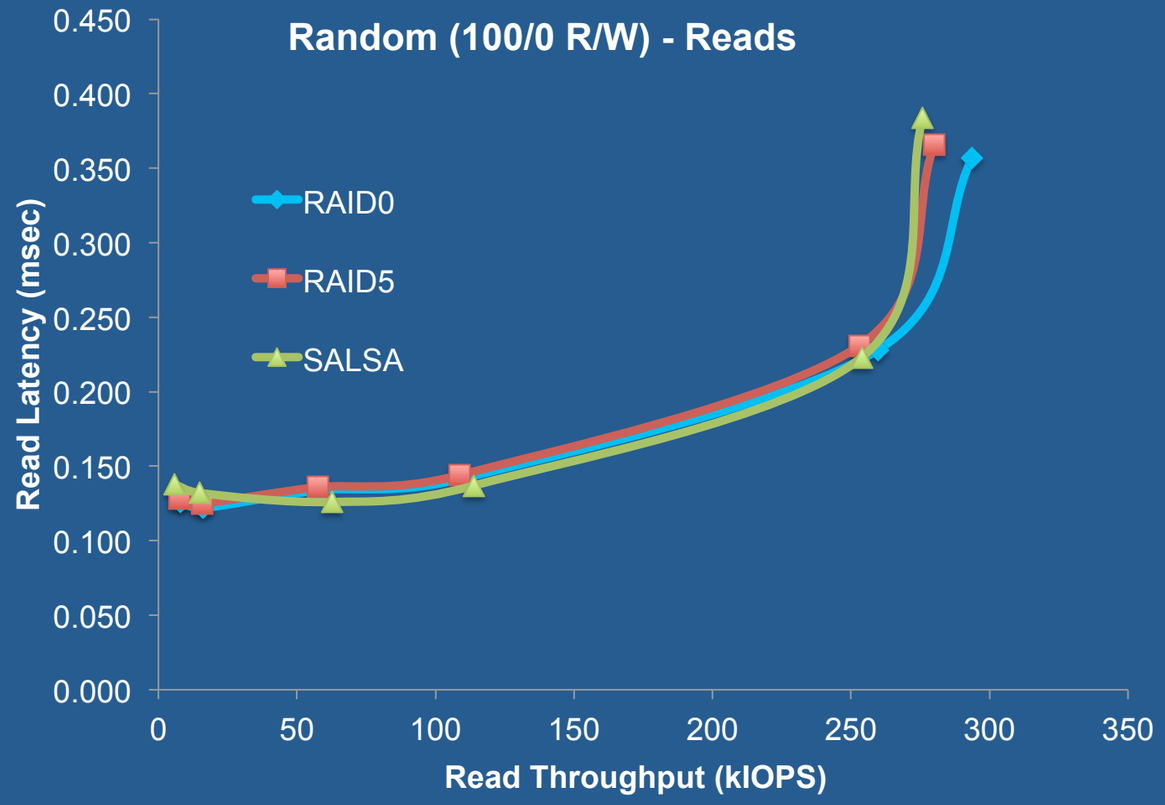


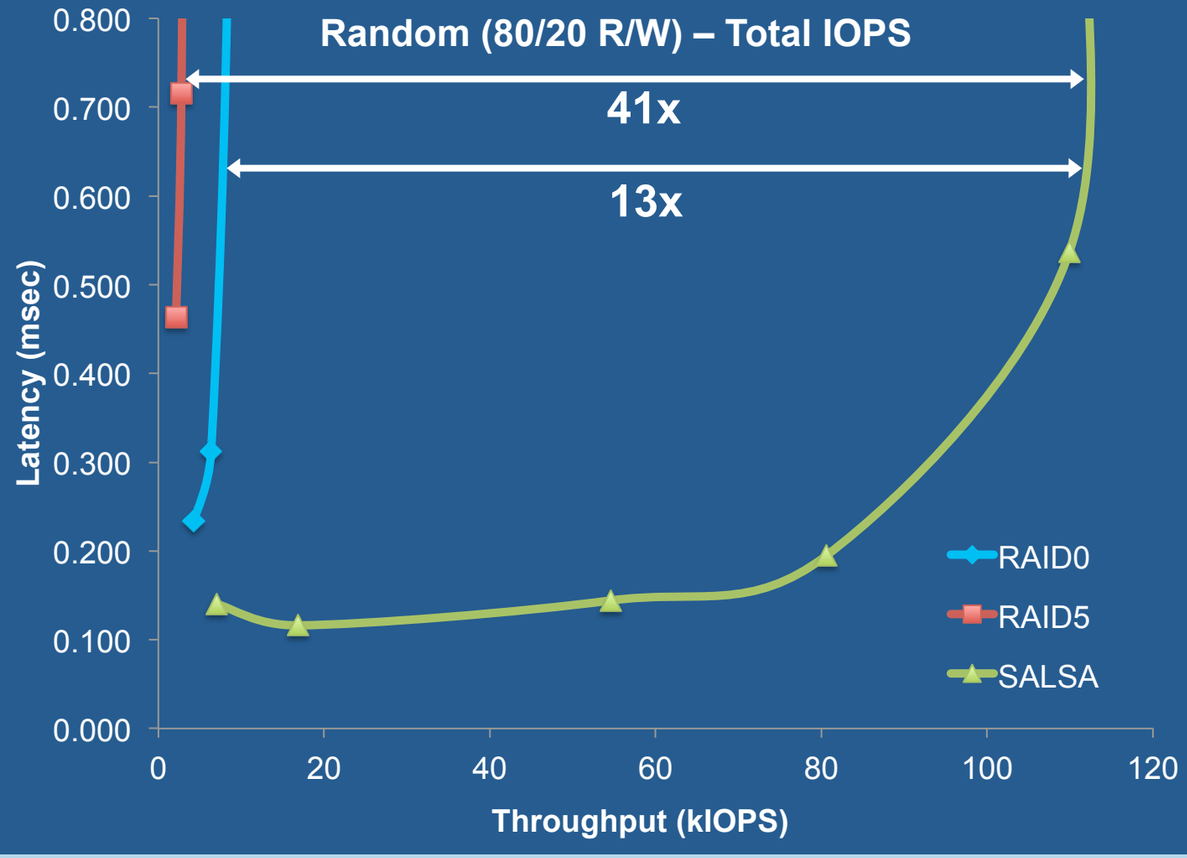
SALSA Stack in Linux



Experiments – Block Storage

- Using SALSA in a commodity Linux server to create an array out of 5 SSDs
  - With RAID5-equivalent parity protection
- Comparing against RAID0, RAID5 on the same SSDs

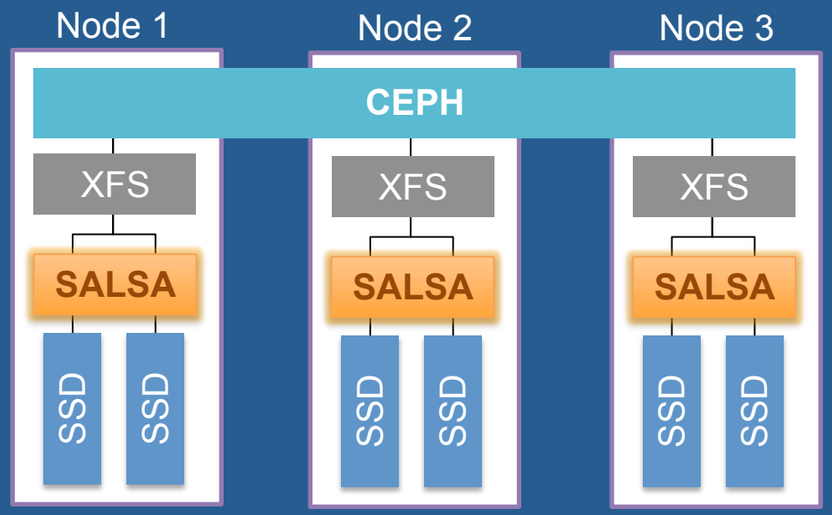


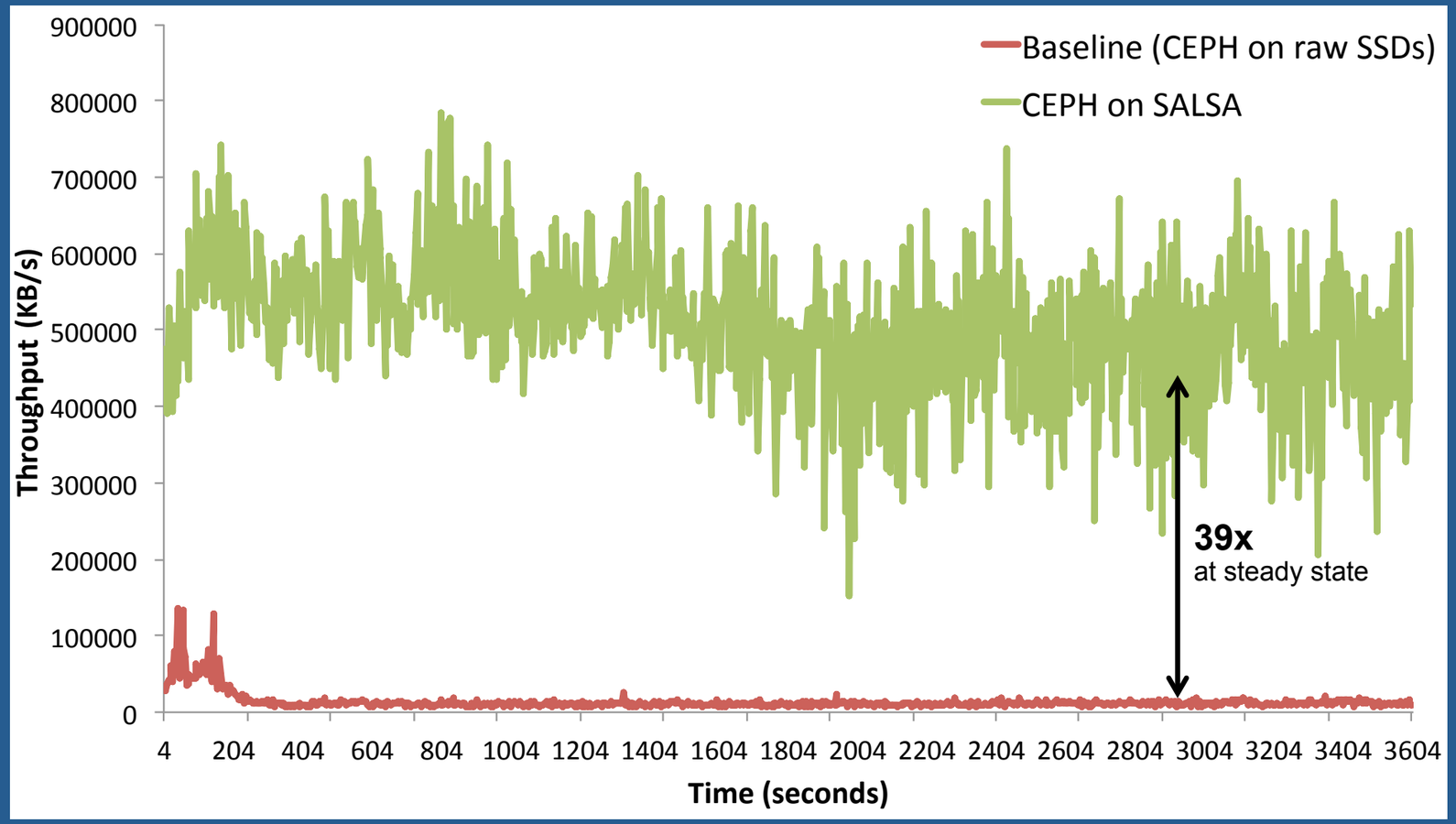


SALSA dramatically improves performance in the presence of writes

CEPH on SALSA

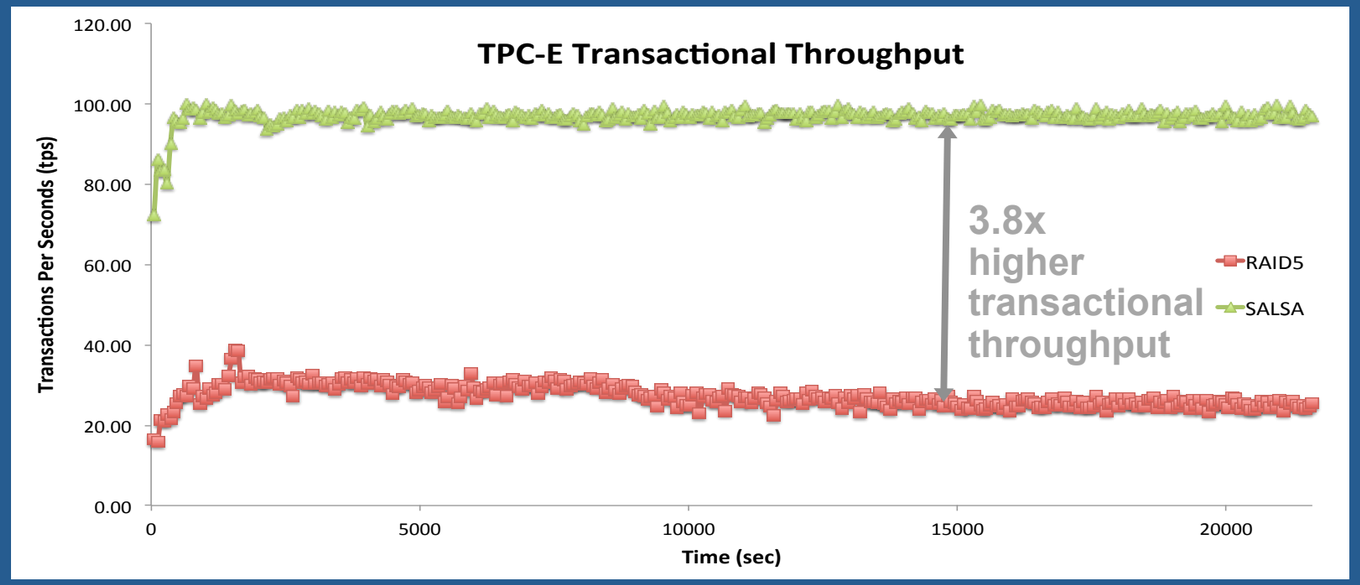
- 3-node x86 cluster
- 10 Gbit Ethernet network
- 2 x 1TB TLC SSDs per node
- Replication factor of 3
- Mixed read/write random I/O



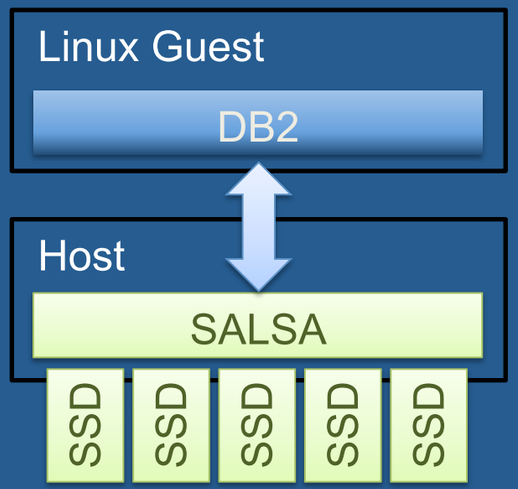


SALSA can enable CEPH on Flash with high performance at a low cost!

Performance – Virtualized TPC-E

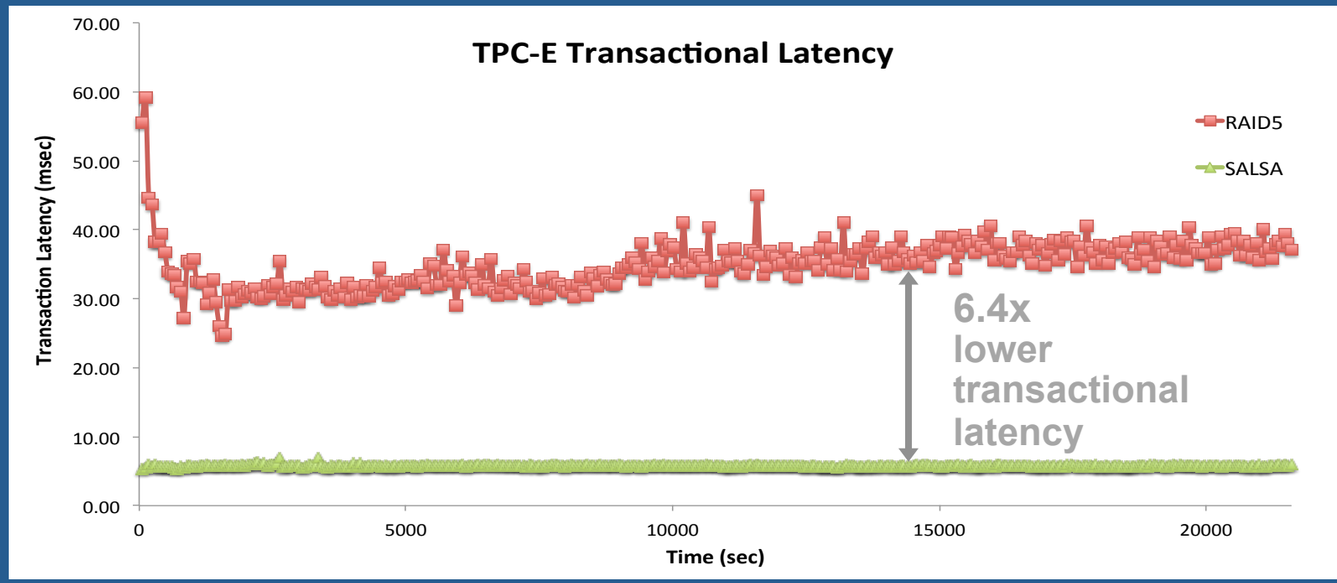


SALSA vs. RAID5
using 5 x 1TB TLC SSDs



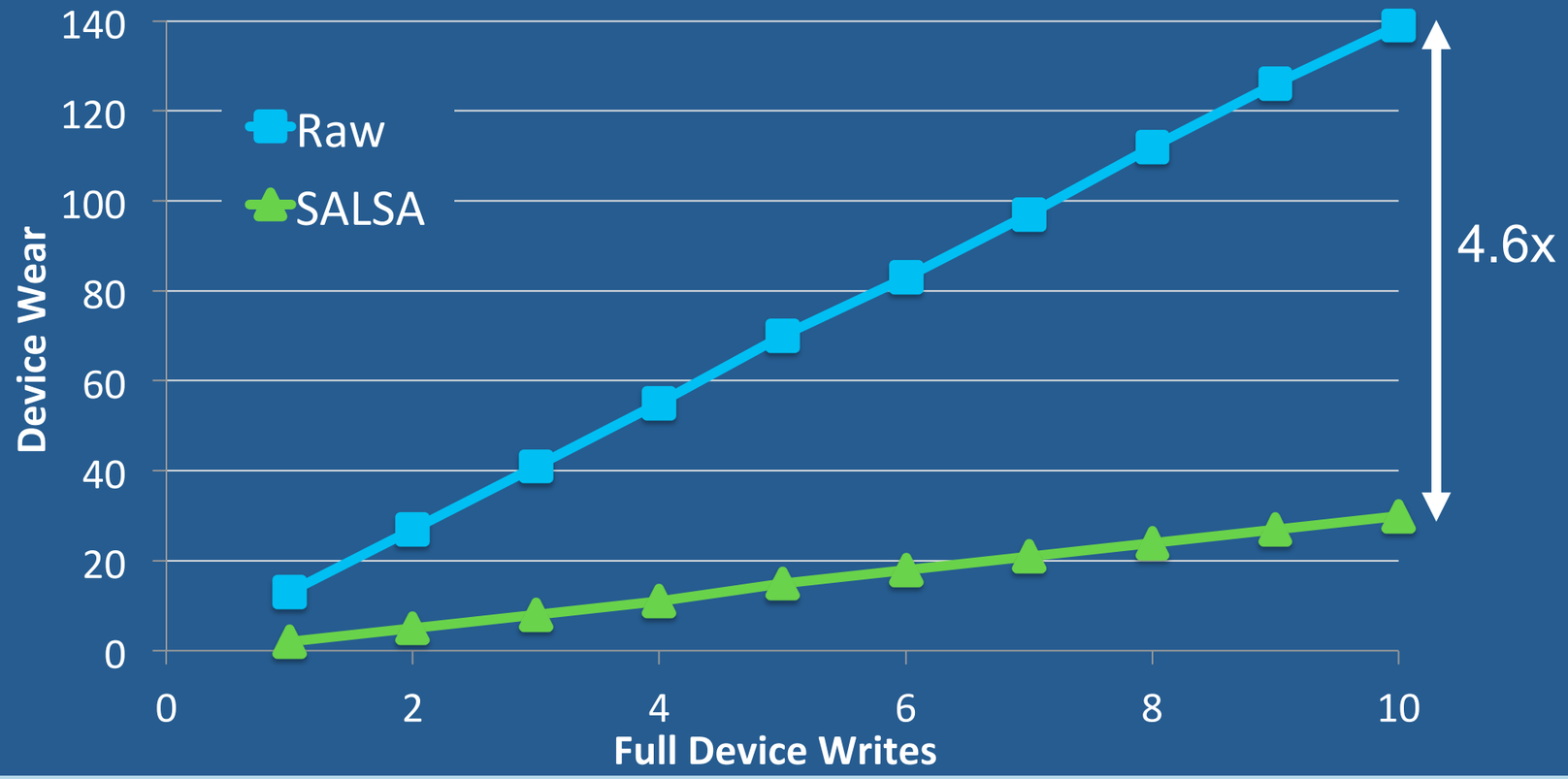
TPC-E

- OLTP benchmark that simulates the workload of a brokerage firm
- Running against **DB2 in KVM guest**
- 90% Reads / 10% Writes



Endurance

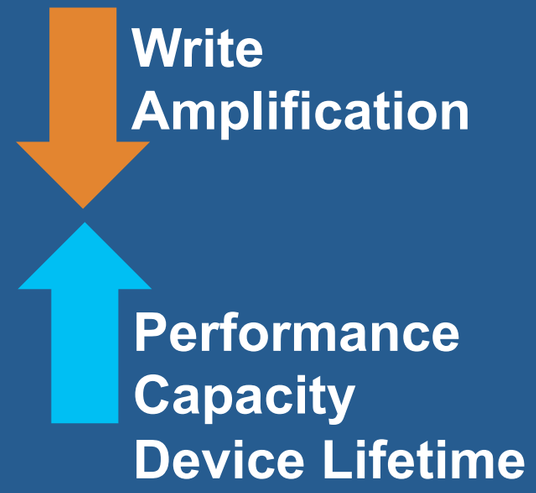
- Test using an off-the-shelf low-cost SSD (0.4 $/GB).
- We measured the wear of the device, as reported by vendor-specific S.M.A.R.T attributes.
- Comparing the wear incurred by SALSA to the wear incurred using the raw device



SALSA prolongs the SSD lifetime by 4.6 times!

Conclusion

- Low-cost Flash is in high demand
- Many workloads could benefit tremendously from capacity-optimized Flash
- **SALSA** is a Flash-optimized storage virtualization stack for Linux
  - Shifts the complexity of the FTL to software
  - Transforms user access patterns to be as Flash-friendly as possible
  - Elevates the performance and endurance of low-cost SSDs to enterprise standards
- File systems & applications do not need to be modified




Questions ?
www.research.ibm.com/labs/zurich/cci/