



**Running NoSQL natively on flash**

**Thomas Rochner**

**[trochner@fusionio.com](mailto:trochner@fusionio.com)**



Fusion-io Confidential—Copyright © 2013 Fusion-io, Inc. All rights reserved.



# Topics – NoSQL Zurich 2013

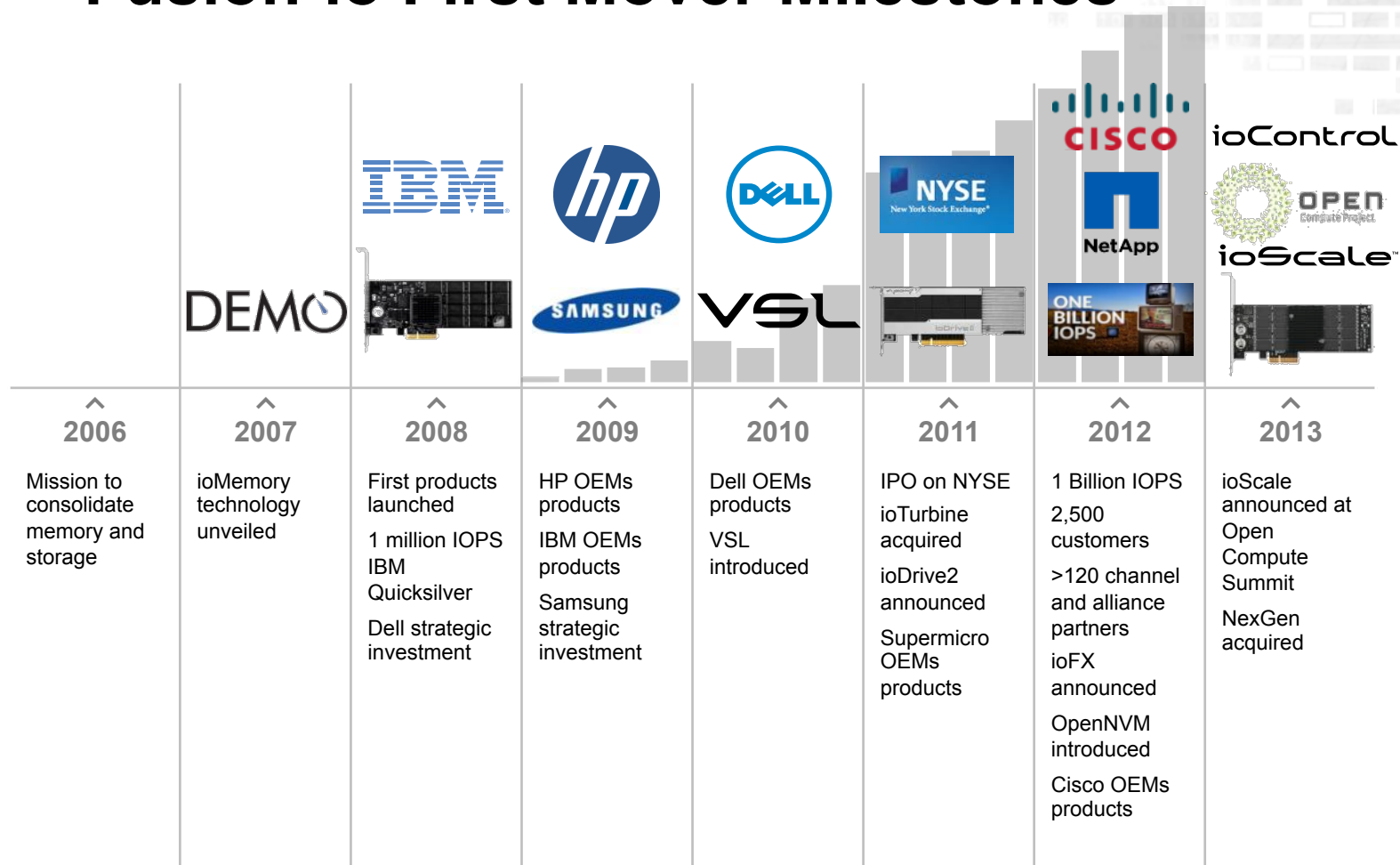
FUSION-io®

1. What are we building ?
2. Why are we building it?
3. OpenNVM
4. Use Cases
5. Where are we headed?



# Fusion-io First Mover Milestones

FUSION-io®





# Fusion-io Accelerates

FUSION-io®

## Databases

ORACLE®

MySQL®

SQL Server

IBM DB2® INFORMIX

SAP®

SYBASE®

INGRES®

PostgreSQL

## Virtualization

vmware®

Windows Server  
Hyper-V

XenDesktop 5

KVM

## Search

fast

A Microsoft® Subsidiary

Autonomy®

Lucene®

ORACLE® Text

## Analytics

AccessData®

MarkLogic®

LexisNexis®

## Big Data

hadoop

mongoDB

## Collaboration

Microsoft® Exchange

Microsoft® SharePoint 2010

IBM Lotus

## HPC

FLUENT

MagmaSoft

NX

NASTRAN

lustre

IBM GPFS

## Messaging

IBM MQ

TIBCO®  
The Power of Now®

software AG

## Workstation

Autodesk®

SolidWorks

Adobe®

## Development

PERFORCE

## Caching

Powered by  
SQUID

VARNISH  
SOFTWARE

## Security/Logging

ArcSight®

splunk™

## Web

LAMP

Microsoft® .NET





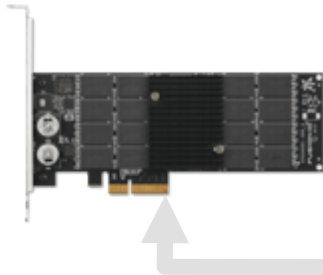
# Fusion-io Product Portfolio

FUSION-io®

## DIRECT

### Max Acceleration

- Closest to CPU
- Fastest deployment
- Scale with servers



ioMemory®

## CACHING

### Max Interoperability

- Virtualize more
- Consolidate server
- Offload SAN

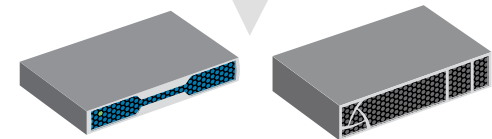


ioTurbine™

## SOLUTIONS

### Max Control

- Allocate across servers
- Scale independently
- Hybrid and QoS options



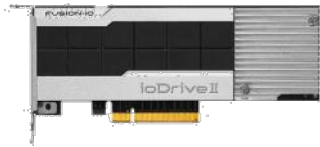
ioN | DATA ACCELERATOR

ioControl | HYBRID STORAGE



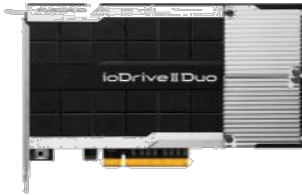
# Direct Acceleration

FUSION-io®



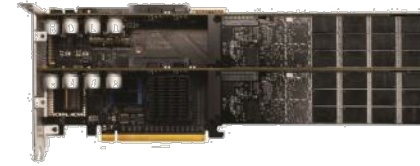
## ioDrive II

Up to 3.0TB of capacity



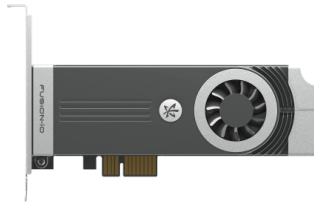
## ioDrive II Duo

Up to 2.4TB of capacity  
per x8 PCI Express slot



## ioDrive Octal

Up to 10.24TB to maximize  
performance for large data sets



## ioFX

Up to 1650GB of workstation  
acceleration for digital content  
creation



## ioDrive II MEZZANINE

Up to 1.2TB for maximum  
performance density



## ioScale™

Up to 3.2TB of low-latency,  
high-performance flash per PCI  
Express slot



# Caching Acceleration

FUSION-io®

## ioTurbine™



- Purpose-built performance for virtualized applications and databases
- Unparalleled low latency performance
- Increase VM density and consolidate servers

## ioCache®



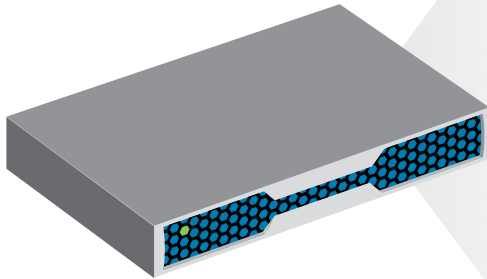
- Turbo Boost virtualization
- Transform ioMemory into a powerful, easy-to-manage, intelligent cache
- Unleash the potential of virtualized systems



# Solutions Acceleration

FUSION-io®

ioN | DATA ACCELERATOR™



Software

- 25-40x performance
- 100µs response time for cluster applications



ioMemory

- PCIe performance
- Adaptive Flashback reliability



Platforms

- SAN Connected
- Qualified servers

**25x** DATABASE PERFORMANCE

**40x** INDEX REBUILD

**35x** XPLORE QUERY TIMES



# ioMemory Solutions Platform

FUSION-io®

Enterprise

Hyperscale

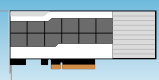
Workstation



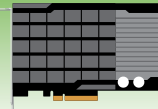
Small to Medium  
Enterprise

Flash-aware  
Applications

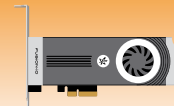
Flash-aware  
Applications



ioDrive II



ioScale™



ioFX™

ioSphere® — ioMemory® — VSL



# Comprehensive Solution Portfolio

FUSION-io®

## ENTERPRISE SCALE UP

- Databases
- Server Virtualization
- Virtual Desktop Infrastructure
- Mixed Workloads

## HYPERSCALE SCALE OUT

- Web Apps
- Big Data
- SaaS

## WORKSTATION SINGLE USER

- Visual Computing
- Digital Content



# Topics – NoSQL Zurich 2013

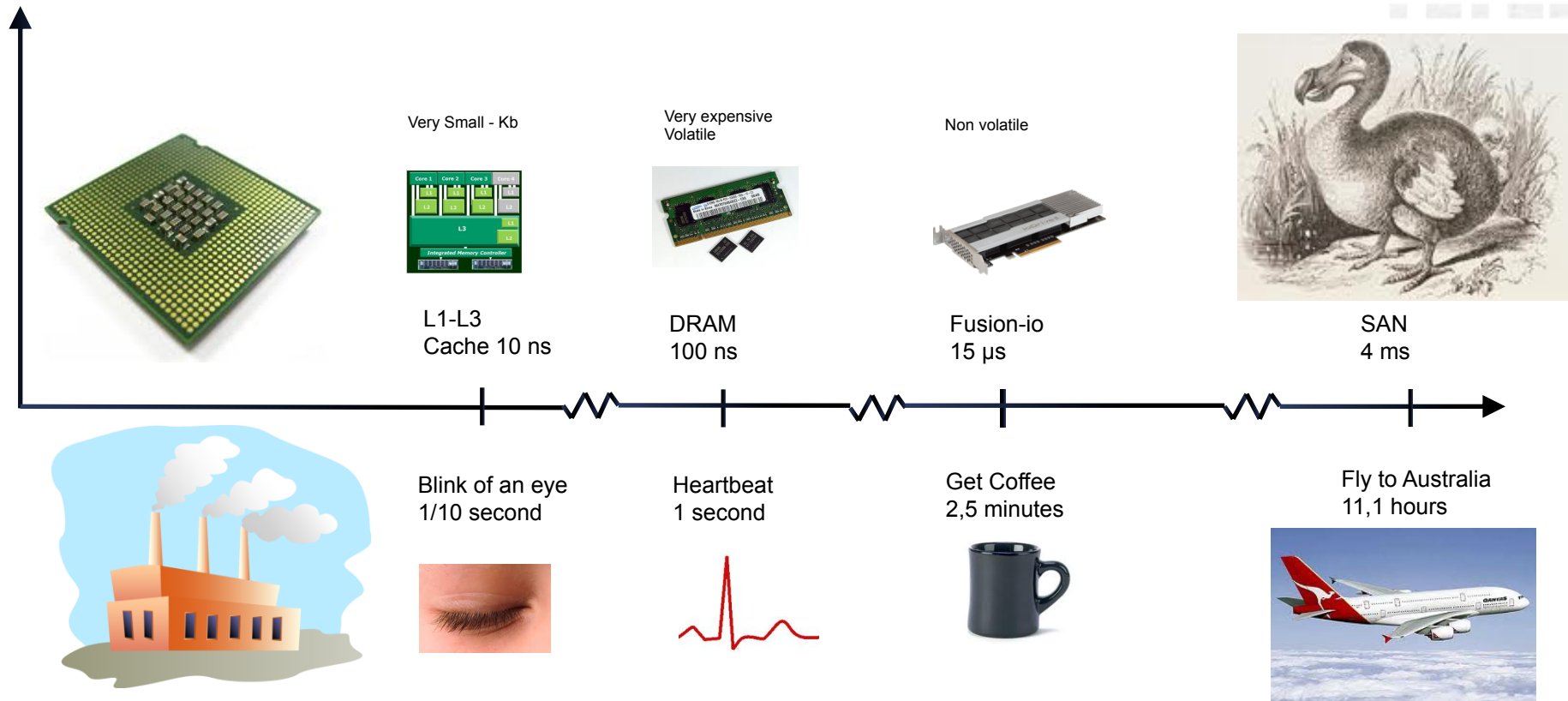
FUSION-io®

1. What are we building ?
2. Why are we building it?
3. OpenNVM
4. Use Cases
5. Where are we headed?



## The landscape of sub second Timings. How fast do you get data to the factory?

FUSION-io®



Multiplier is 10m - 10.000.000

September 24, 2013



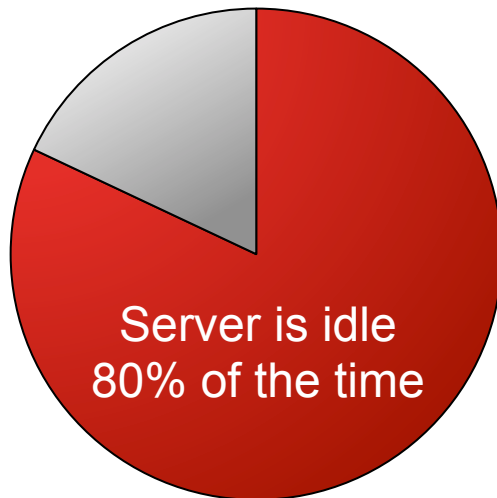


## SLOW STORAGE LEADS TO IDLE CAPACITY

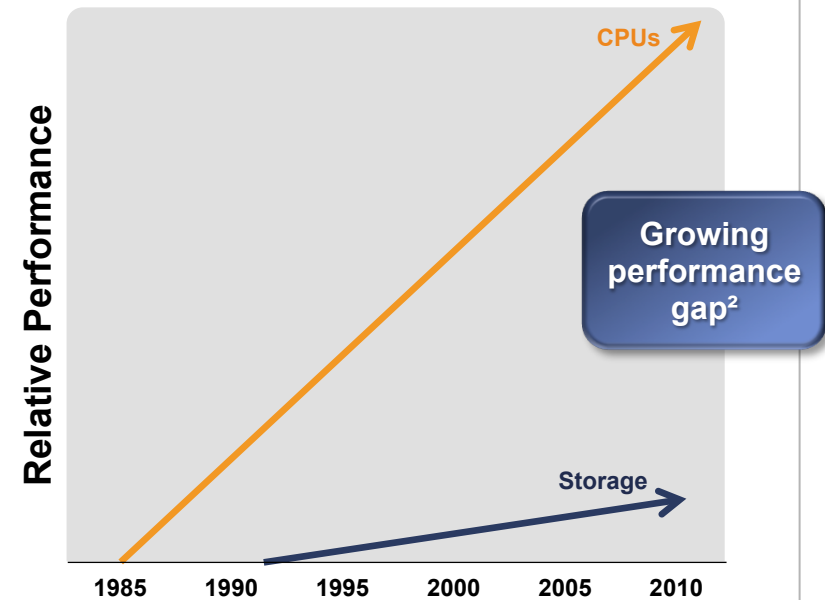
FUSION-io®

According to Moore's Law, processing performance doubles every 18 months

37% of servers are massively underutilized<sup>1</sup>...



...because the performance gap continues to grow



<sup>1</sup> Source: IDC's Server Workloads 2010, July 2010

<sup>2</sup> Source: Taming the Power Hungry Data Center, Fusion-io White Paper



**SPINNING MEDIA**  
OVER 150 YEARS OLD



# SSD treats memory like disk

FUSION-io®

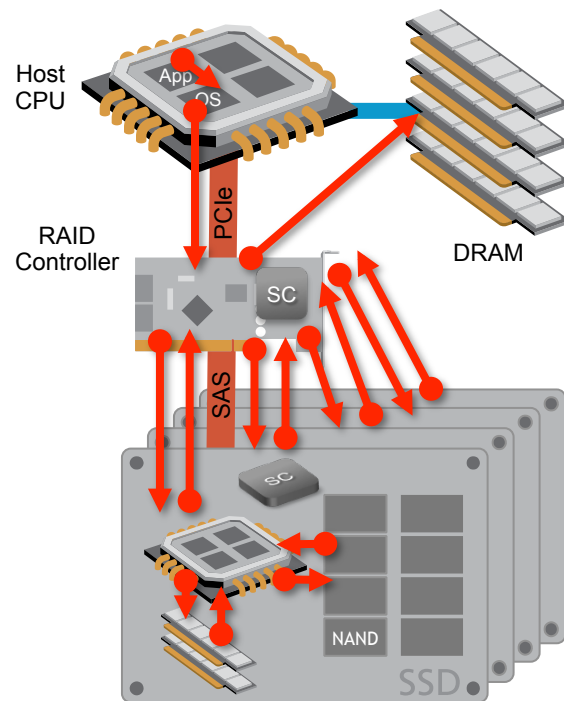




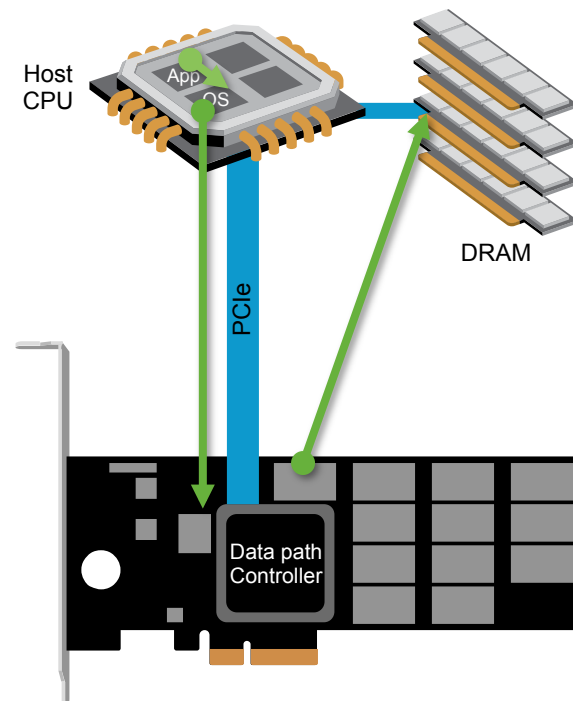
# Flash Architectures

FUSION-io®

## FLASH AS DISK



## FLASH AS MEMORY

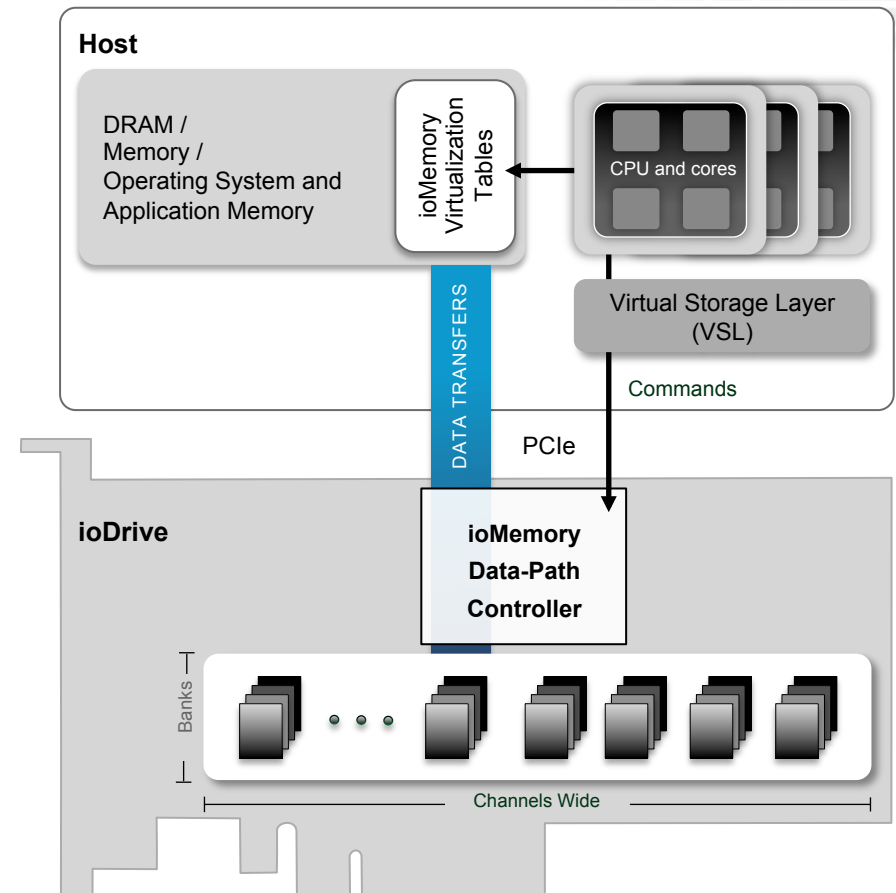
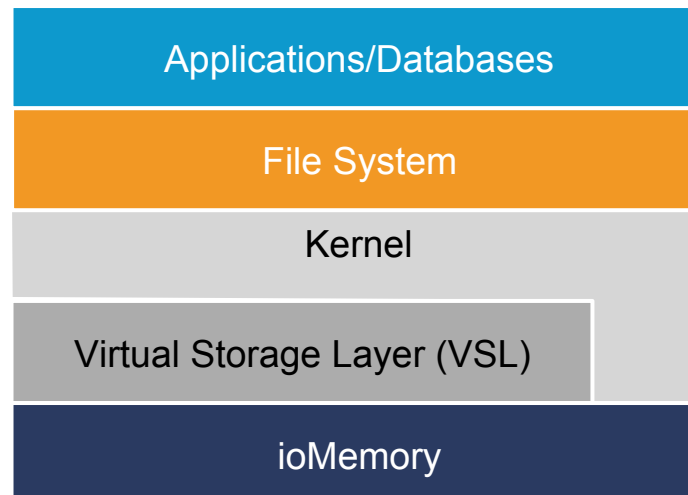




# Cut-through Architecture and VSL

FUSION-io®

- ▶ Sophisticated architecture
  - maximum performance
- ▶ Intelligent software
  - advanced features



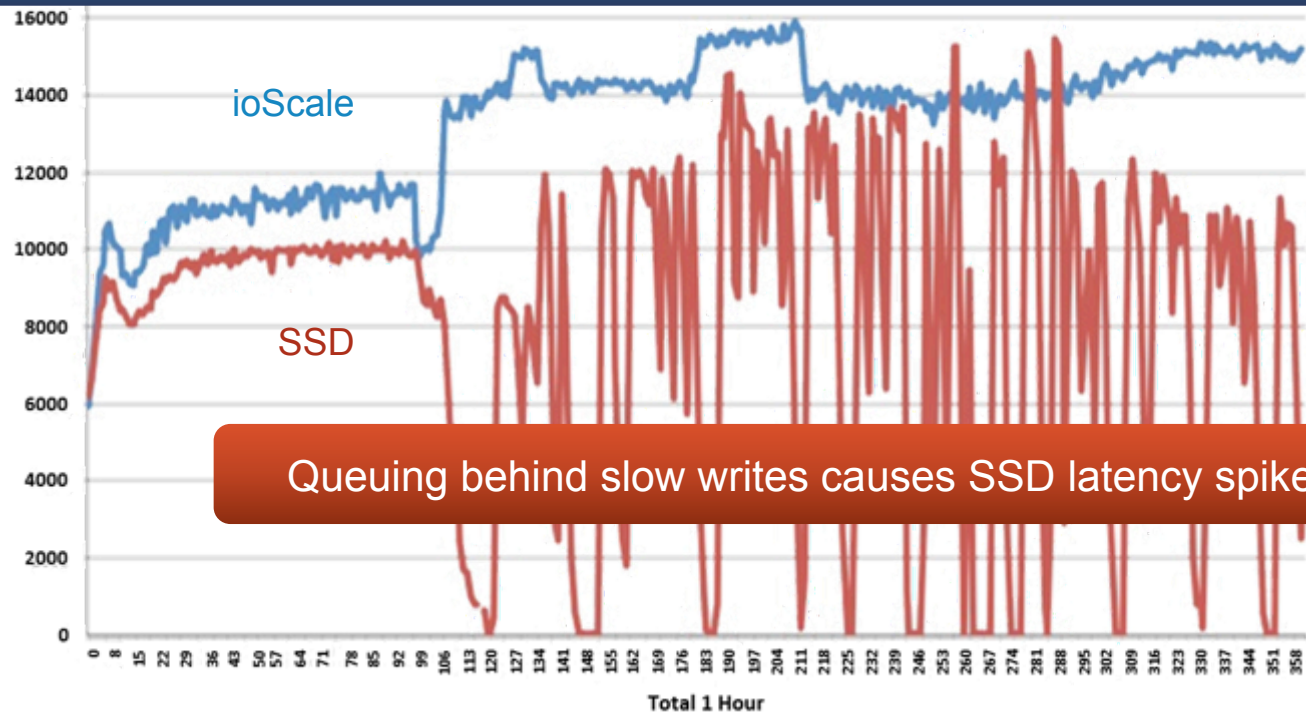




# Balanced Performance Affects Throughput

FUSION-io®

ioMemory balances read/write performance for consistent throughput





## ioDrive2 specs – visit [fusionio.com](http://fusionio.com)

FUSION-io®

ioDrive2 Capacity	365GB MLC	785GB MLC*	1.2TB MLC*	3.0TB MLC
Read Bandwidth - 1MB	910 MB/s	1.5 GB/s	1.5 GB/s	1.5 GB/s
Write Bandwidth - 1MB	590 MB/s	1.1 GB/s	1.3 GB/s	1.3 GB/s
Ran. Read IOPS - 512B	137,000	270,000	275,000	143,000
Ran. Write IOPS - 512B	535,000	800,000	800,000	535,000
Ran. Read IOPS - 4K	110,000	215,000	245,000	136,000
Ran. Write IOPS - 4K	140,000	230,000	250,000	242,000
Read Access Latency	68µs	68µs	68µs	68µs
Write Access Latency	15µs	15µs	15µs	15µs
Bus Interface	PCI-Express 2.0 x4			
Weight	6.6 ounces			9.5 ounces
Form Factor	Half-height, half-length			Full-height, half-length
Warranty	5 years or maximum endurance used			



# Topics – NoSQL Zurich 2013

FUSION-io®

1. What are we building ?
2. Why are we building it?
3. OpenNVM
4. Use Cases
5. Where are we headed?





# NoSQL Software challenges

FUSION-io®

- Keeping NoSQL software simplicity with data persistence
- Transforming in-memory structures to block I/O
- Tiering data between DRAM and persistent storage
- Keeping latency low with data persistence
- Scaling up



# OpenNVM - <http://opennvm.github.io>

FUSION-io®

OpenNVM

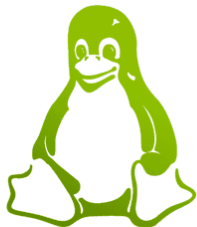
Get Started

Examples

Documentation

Benchmarking Tools

## Current OpenNVM Repositories



### Flash-aware Linux swap

When working set size exceeds the capacity of DRAM, demand page from a flash-aware virtual memory subsystem.

[Repository](#)

[Learn More](#)



### Key-value interface to flash

Create NoSQL databases faster. Automate garbage collection of expired data.

[Repository](#)

[Learn More](#)



### Flash programming primitives

Use built-in characteristics of the Flash Translation Layer to perform journal-less updates (more performance and less flash wear = lower TCO)

[Repository](#)

[Learn More](#)

- Native programming interfaces
- Access flash as a memory
- Eliminate legacy software layers
- Simplify application authoring
- Accelerate time-to-market



# NVM Software interfaces

FUSION-io®

- ▶ Industry-first, direct API access to non-volatile memory's unique characteristics.
  
- ▶ The OpenNVM was introduced to help developers:
  - **Write less code** to create high-performing apps
  - **Tap into performance** not available with conventional I/O access to SSDs
  - **Reduce operating costs** by decreasing RAM while increasing NVM



## **Direct-Access to non-volatile memory is now emerging**

FUSION-io®

- ▶ Developers are beginning to manipulate data

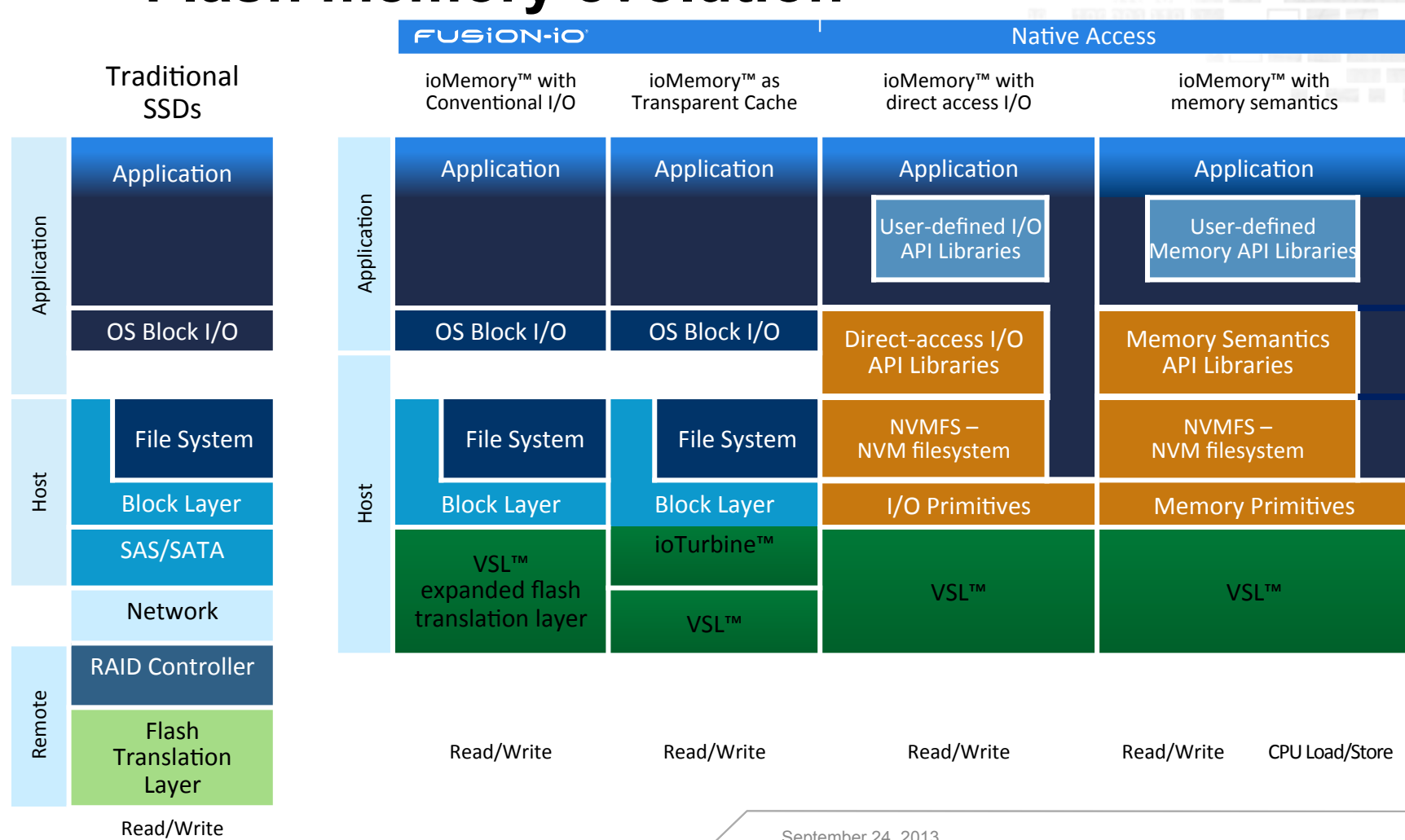
directly in Non-Volatile Memory (NVM)

without converting to basic block I/O.



# Flash memory evolution

FUSION-io®



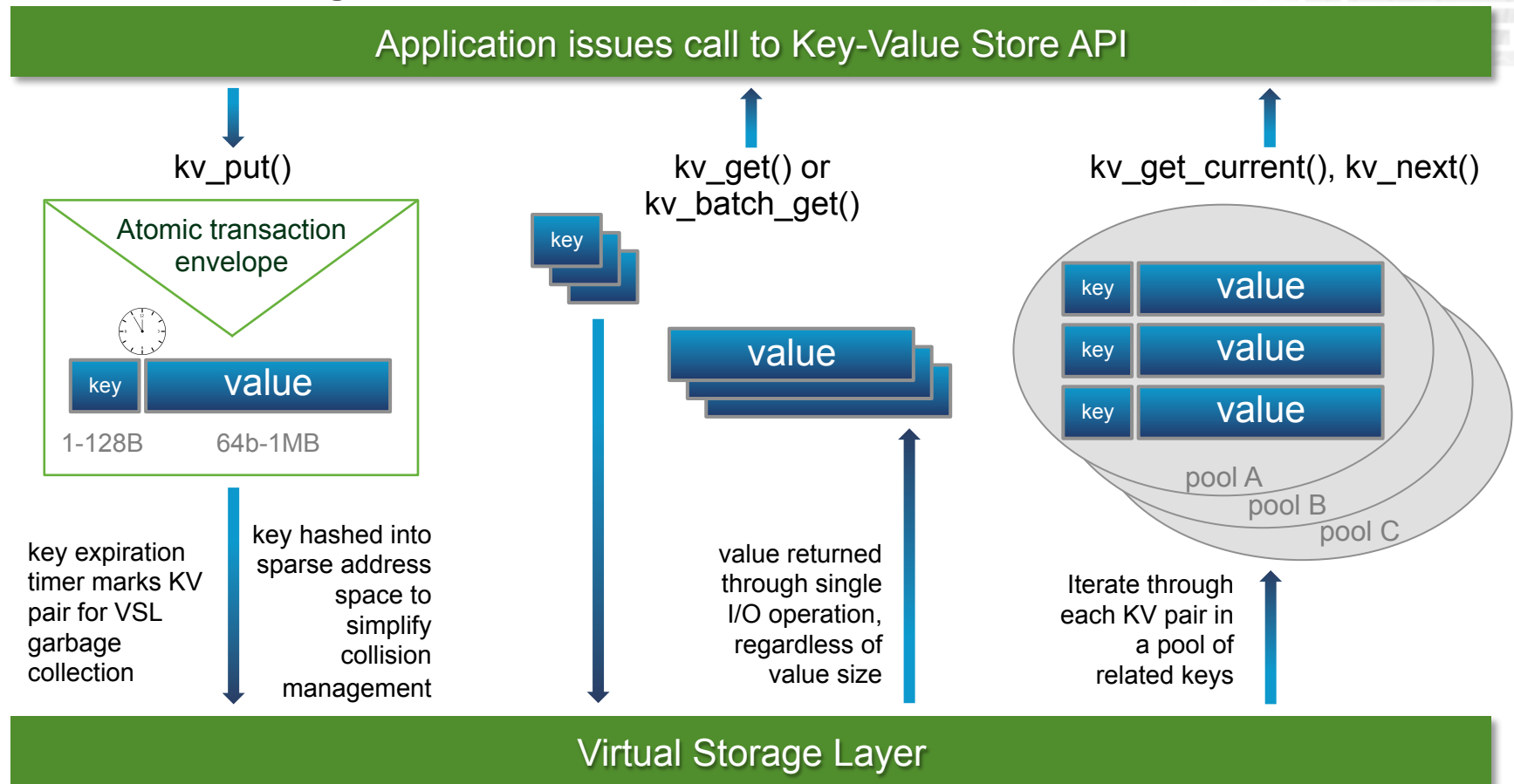
September 24, 2013

25



# Example: Key-Value Store API Library

FUSION-io®

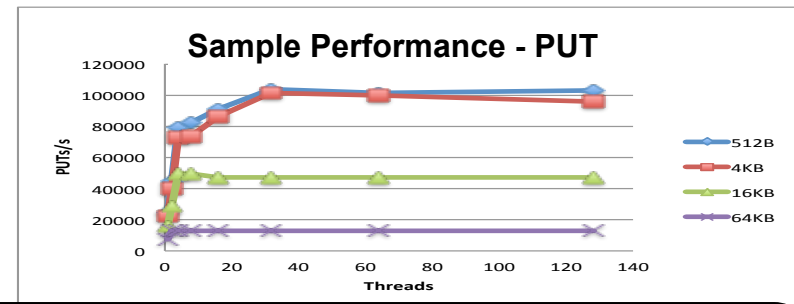
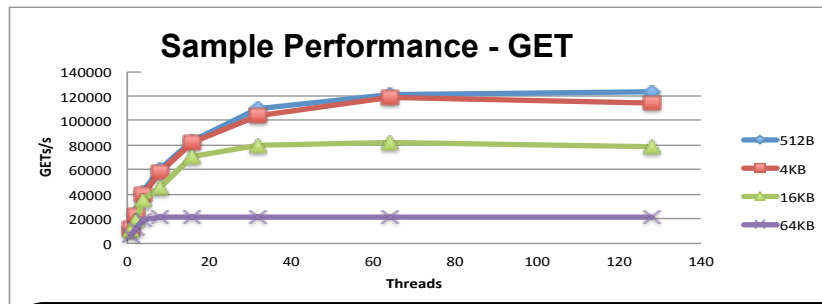




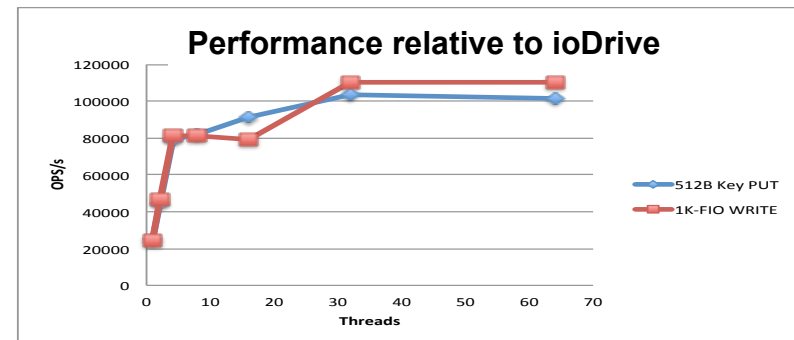
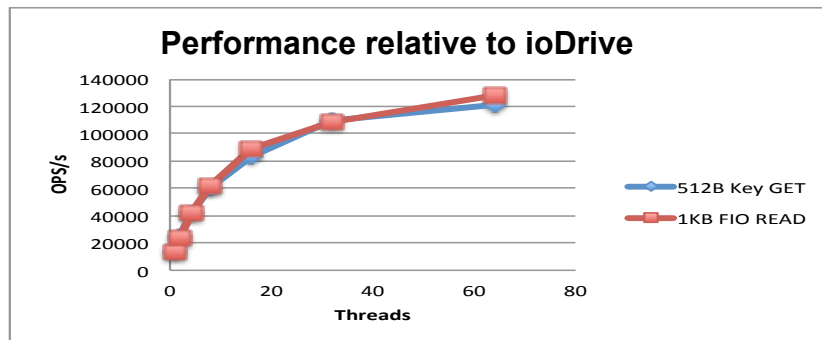
# Key-Value Store API Library Benchmarks

(native KV Get/Put vs. raw reads/writes)

FUSION-io®



**SIGNIFICANTLY MORE FUNCTIONALITY WITH NEGLIGIBLE PERFORMANCE COST**



1U HP blade server with 16 GB RAM, 8 CPU cores - Intel(R) Xeon(R) CPU X5472 @ 3.00GHz with single 1.2 TB ioDrive2 mono



## Key-value store API Library Benefits

FUSION-io®

**95%** performance of raw device

Smarter media now natively understands a key-value I/O interface with lock-free updates, crash recovery, and no additional metadata overhead.

**Up to 3x** capacity increase

Dramatically reduces over-provisioning with coordinated garbage collection and automated key expiry.

**3x** throughput on same SSD

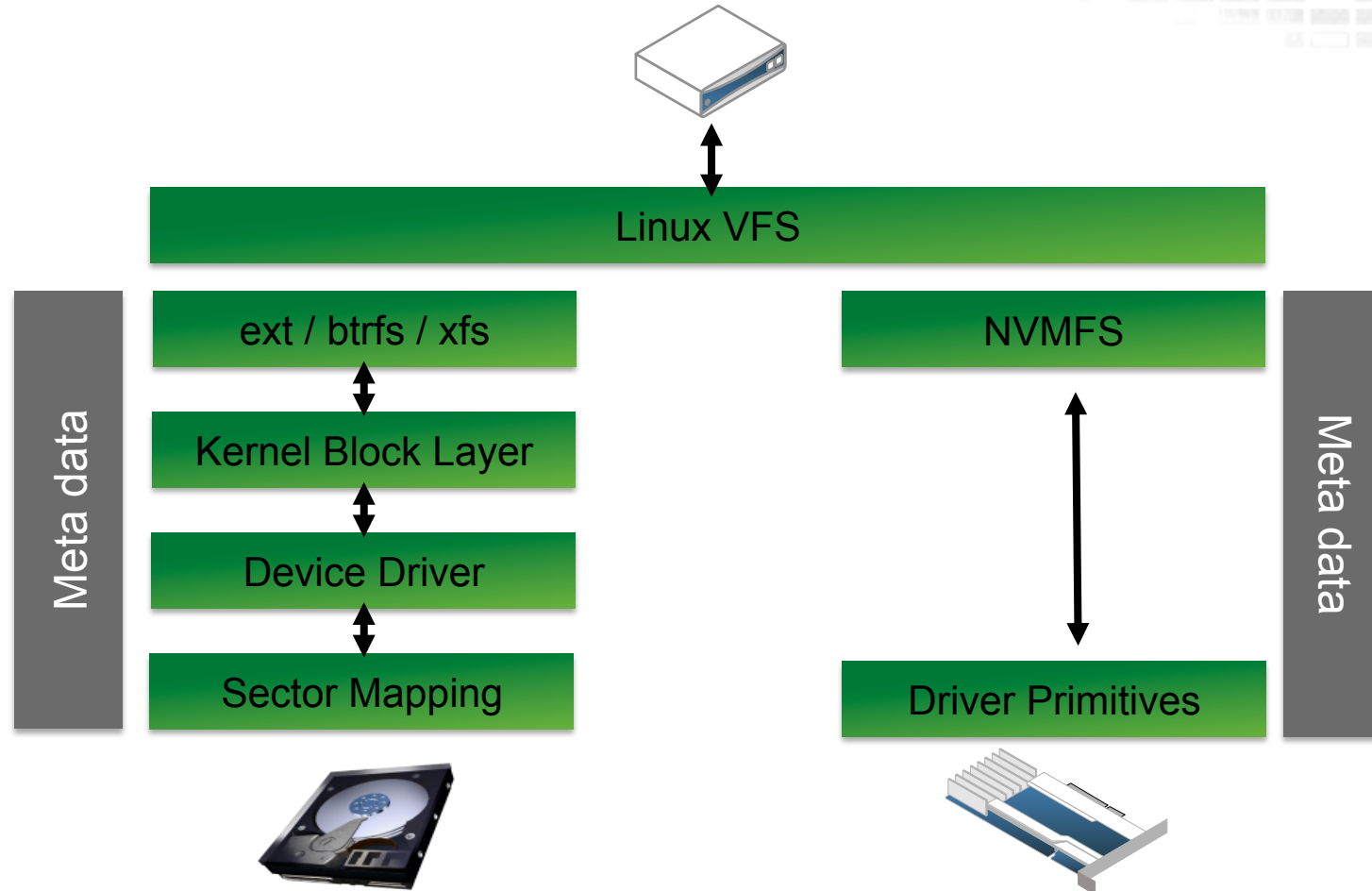
Early benchmarks comparing against memcached with BerkeleyDB persistence show up to 3x improvement.





# NVMFS – Eliminating duplicate logic

FUSION-io®





## NVMFS – Benefits in Eliminating Duplicate logic

FUSION-io®

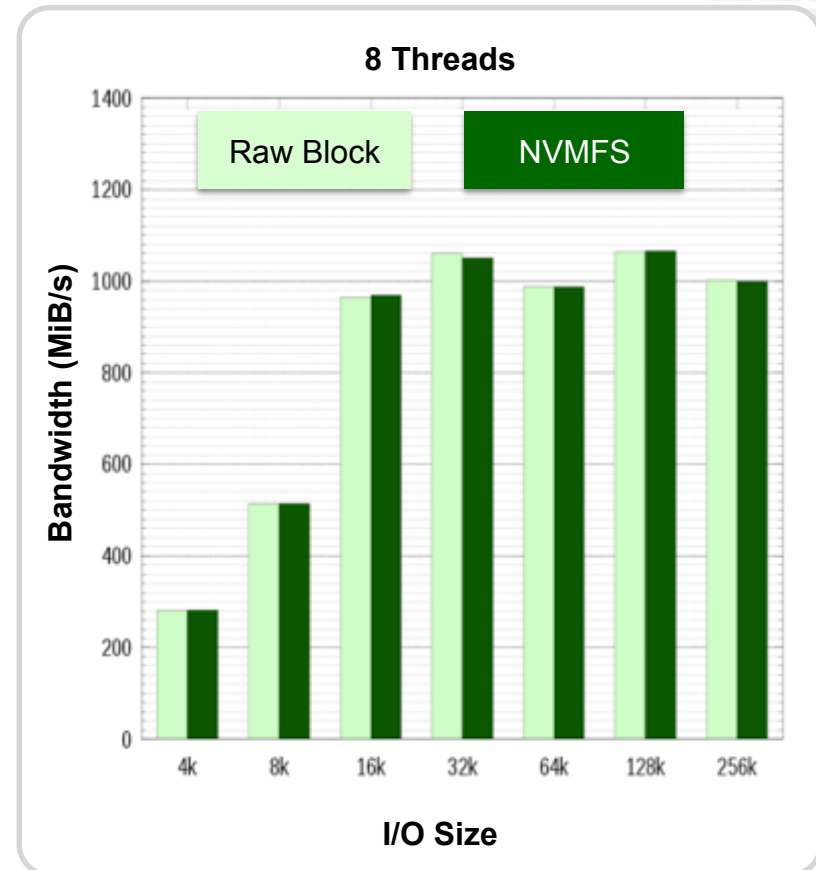
File System	Lines of Code
NVMFS	6879
ReiserFS	19996
ext4	25837
btrfs	51925
XFS	63230



# NVMFS: Native Flash Filesystem

FUSION-io®

- ▶ File system convenience
- ▶ Raw block performance
- ▶ No compromise necessary

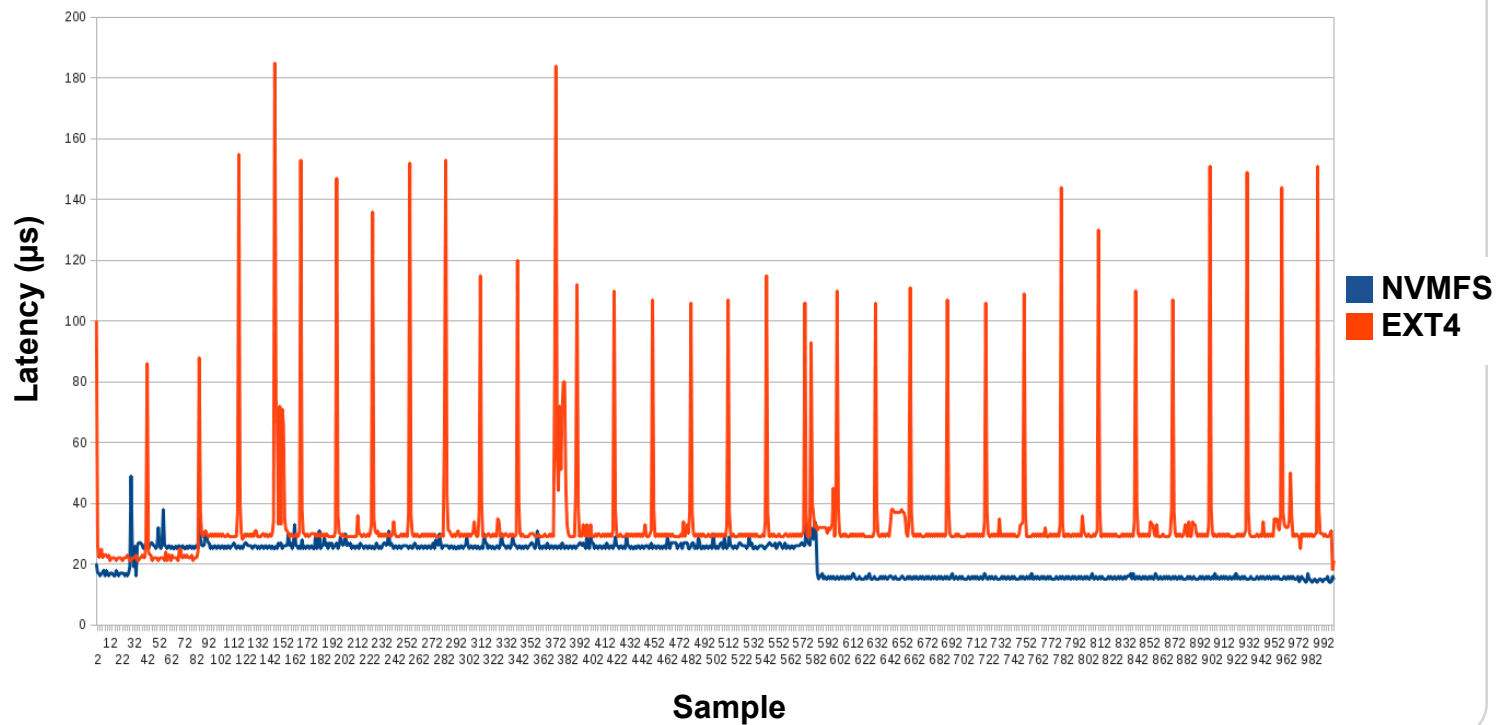




# NVMFS: Consistent Low Latency

FUSION-io®

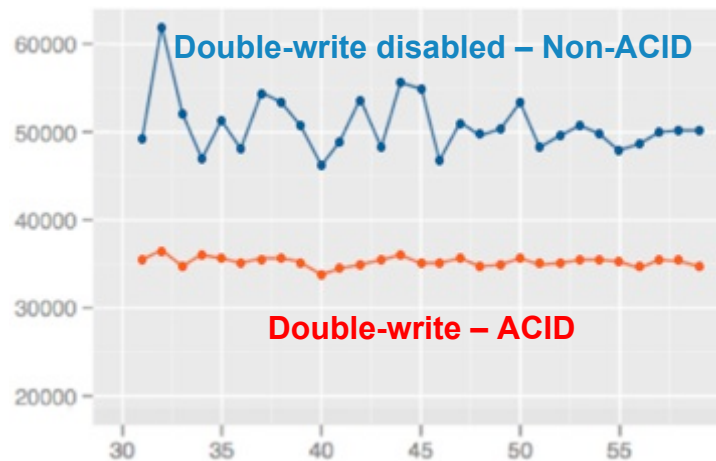
Fusion-io NVMFS vs EXT4 write latency  
1000 512 Byte Sequential Write with O\_DIRECT



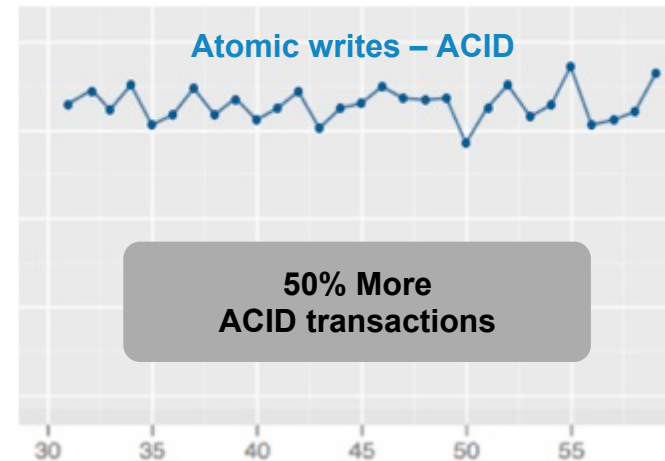


# MySQL: NVMFS and Atomic Writes

FUSION-io®



XFS



NVMFS with Atomic I/O



# Range of memory-Access Semantics

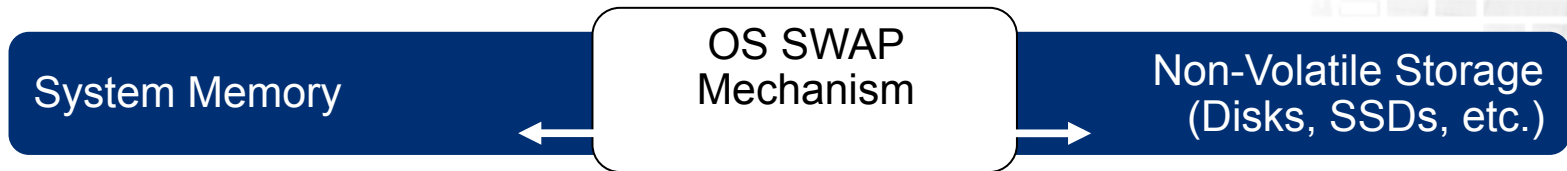
FUSION-io®

Extended Memory	<b>Volatile</b>	Transparently extends DRAM onto flash, extending application virtual memory
Checkpointed Memory	Volatile with non-volatile checkpoints	Region of application virtual memory with ability to preserve snapshots to flash namespace
Auto-Commit Memory™	<b>Non-volatile</b>	Region of application memory automatically persisted to non-volatile memory and recoverable post-system failure



# OS Swap vs. Extended Memory

FUSION-io®



- Originally designed as a last resort to prevent OOM (out-of-memory) failures
- Never tuned for high-performance demand-paging
- Never tuned for multi-threaded apps
- Poor performance, ex. < 30 MB/sec throughput



- No application code changes required
- Designed to migrate hot pages to DRAM and cold pages to ioMemory
- Tuned to run natively on flash (leverages native characteristics)
- Tuned for multi-threaded apps
- 10-15x throughput improvement over standard OS Swap



# Topics – NoSQL Zurich 2013

FUSION-io®

1. What are we building ?
2. Why are we building it?
3. OpenNVM
4. **Use Cases**
5. Where are we headed?

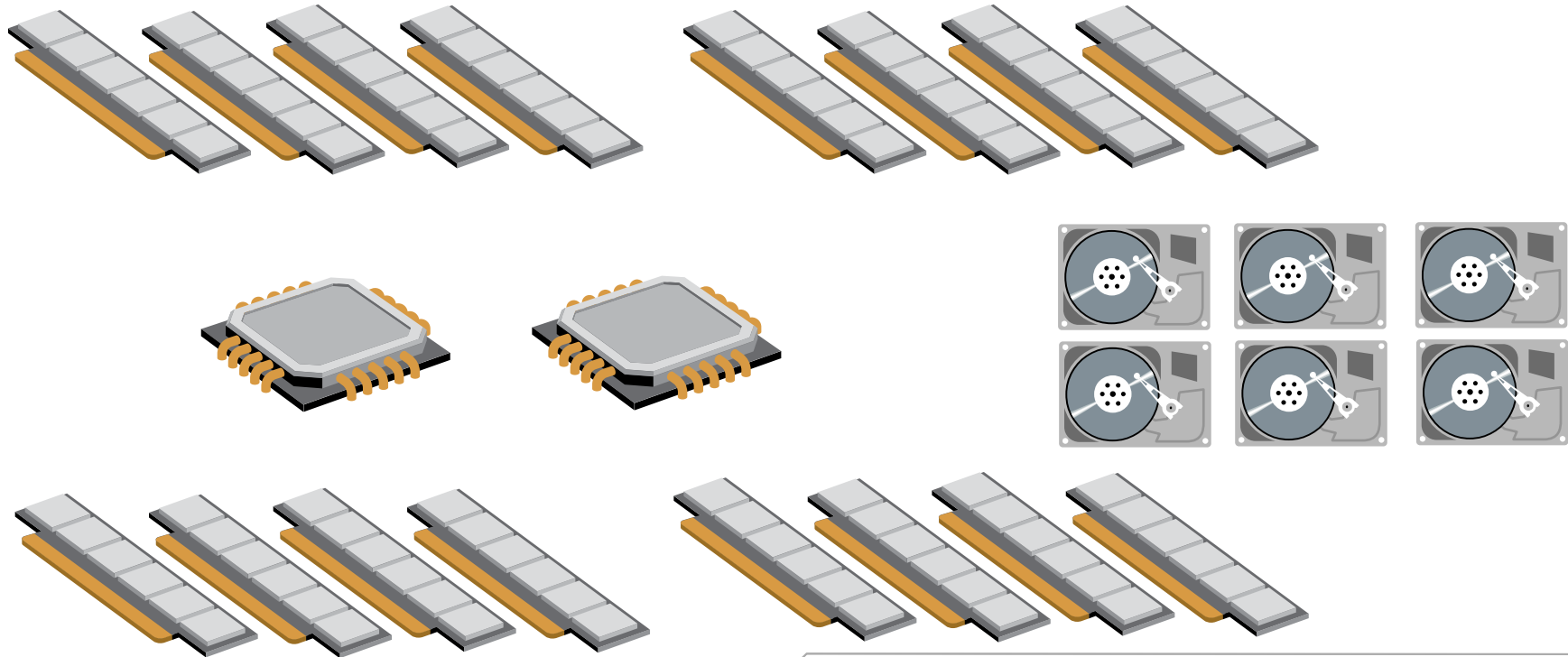




# DRAM Dictates NoSQL Scaling

FUSION-io®

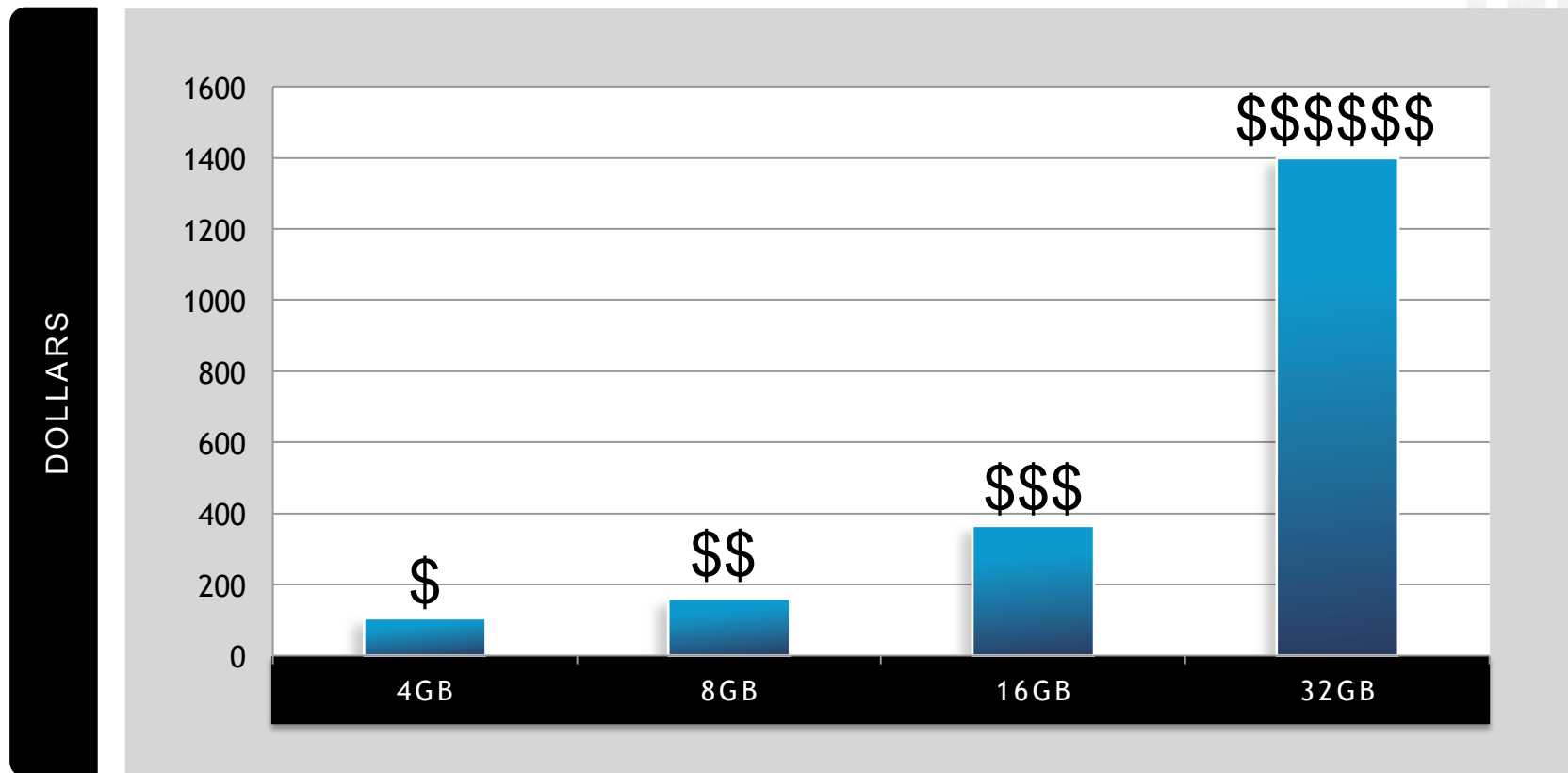
- ▶ Key Design Principle:
- ▶ Working Set < DRAM





# Cost of DRAM Modules

FUSION-io®





## When do we scale out?

- ▶ A typical server...

CPU Cores: 32 with HT  
Memory: 128 GB



...is your working set > 128GB?



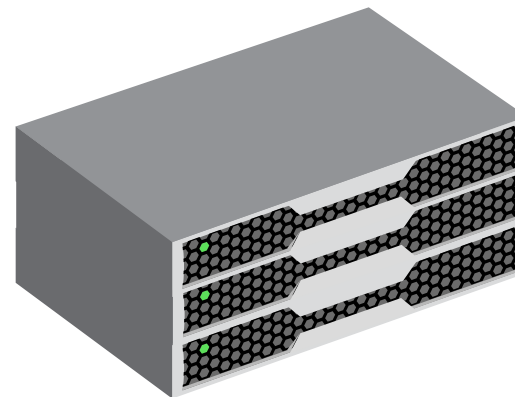
## Is there a better way?

- ▶ With NoSQL Databases, we tend to scale out for DRAM

Combined Resources

CPU Cores: 96

Memory: 384 GB



FUSION-io®

More cores than needed to serve reads and writes.



# Three Deployment Options

FUSION-io®

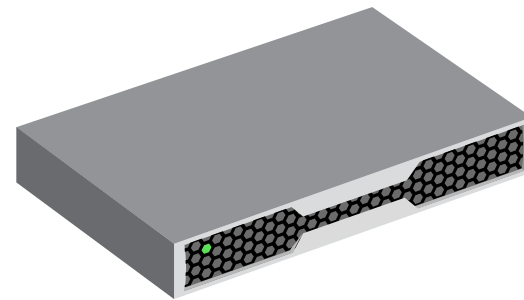
1. All Flash
2. Data Placement (CASSANDRA-2749)
3. Use Logical Data Centers



# Cassandra with All-Flash Storage

FUSION-io®

Step 1: Mount ioMemory at /var/lib/cassandra/data  
Step 2:





# Data Placement

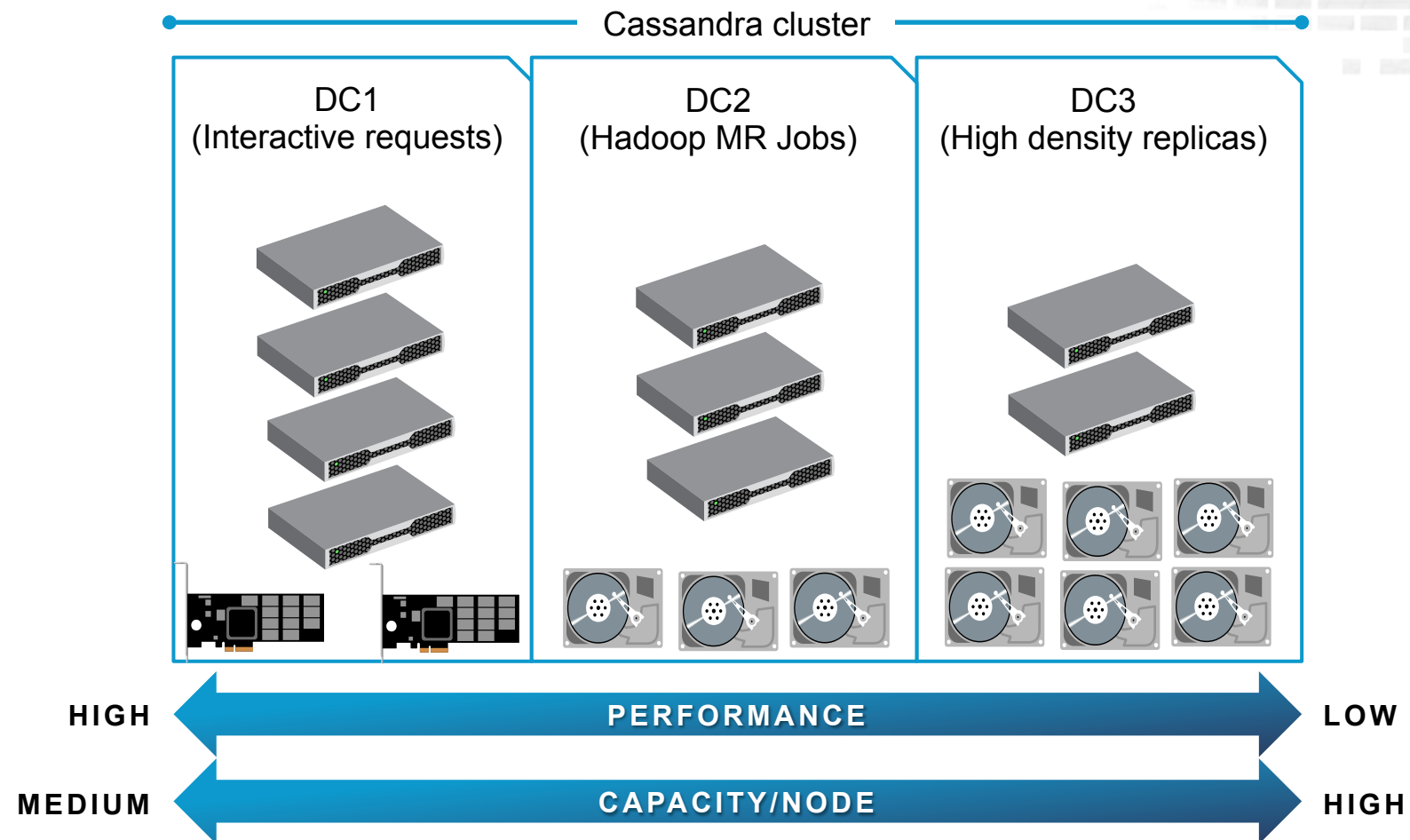
FUSION-io®

- ▶ <https://issues.apache.org/jira/browse/CASSANDRA-2749>
  - Thanks Marcus!
- ▶ Takes advantage of filesystem hierarchy
- ▶ Use mount points to pin Keyspaces or Column Families to flash:
  - `/var/lib/cassandra/data/{Keyspace}/{CF}`
- ▶ Use flash for high performance needs, disk for capacity needs



# Data Centers for Storage Control

FUSION-io®

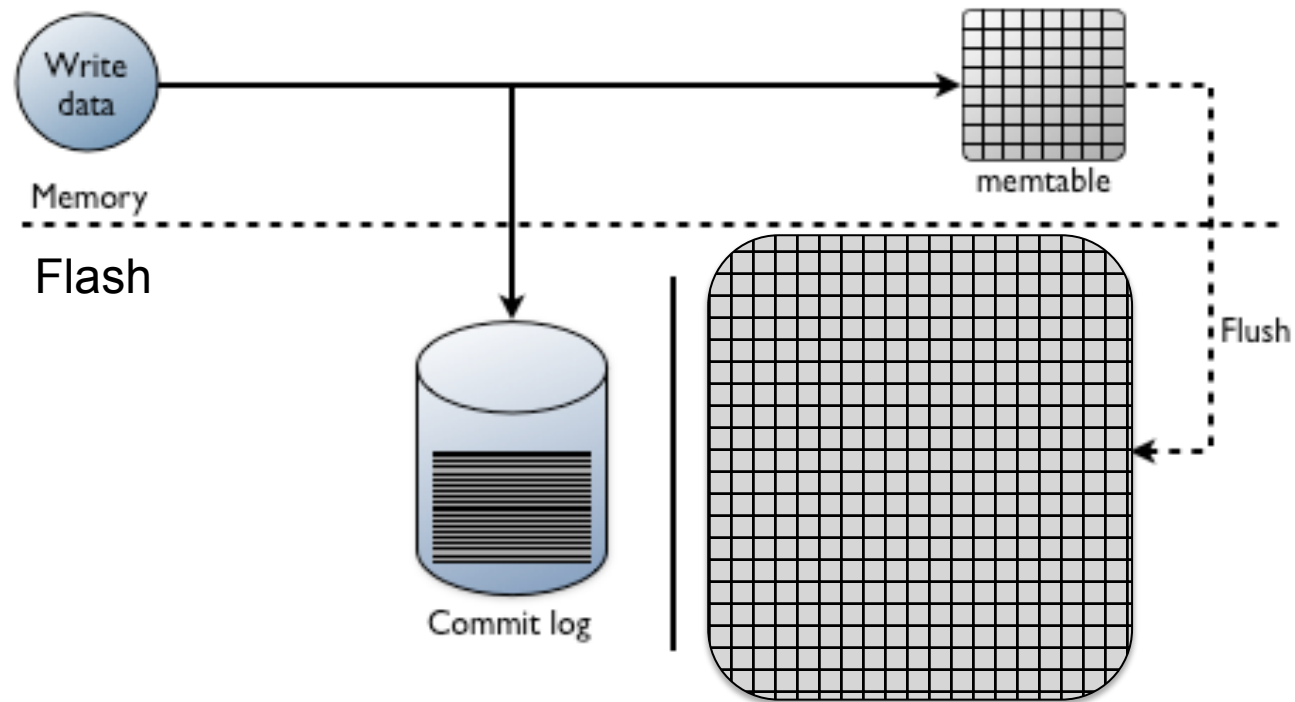






# Rethinking Cassandra I/O

FUSION-io®

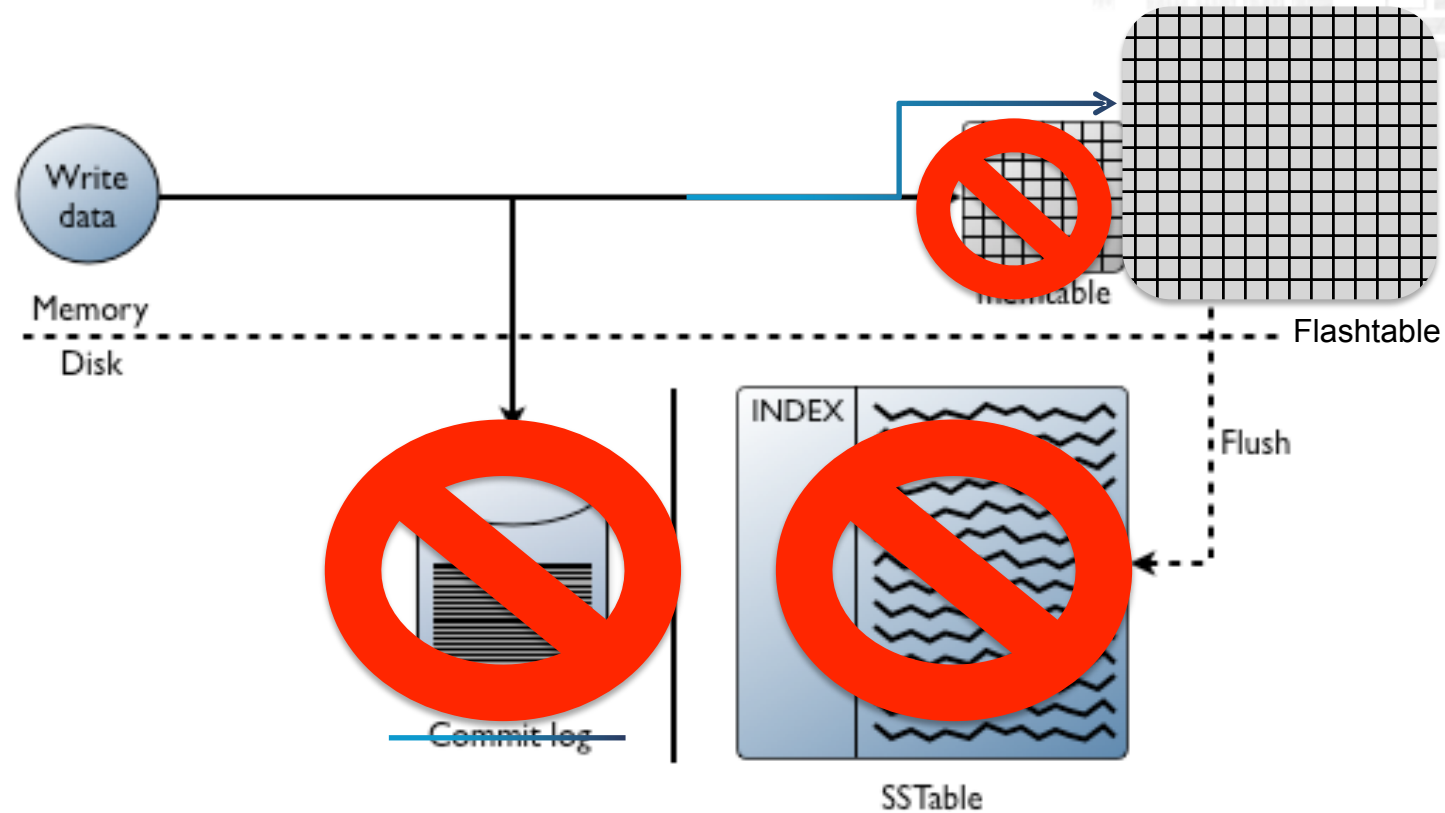


[http://www.datastax.com/docs/1.2/dml/about\\_writes](http://www.datastax.com/docs/1.2/dml/about_writes)



# Rethinking Cassandra I/O

FUSION-io®



[http://www.datastax.com/docs/1.2/dml/about\\_writes](http://www.datastax.com/docs/1.2/dml/about_writes)

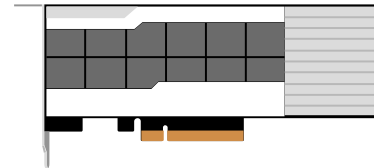
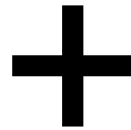


# Accelerating Cassandra With Flash

FUSION-io®



***Cassandra***



**NAND Flash Accelerator**

We have been talking with Cassandra about how we can solve some of these problems, so stay tuned.



# Real-World Cassandra on Fusion

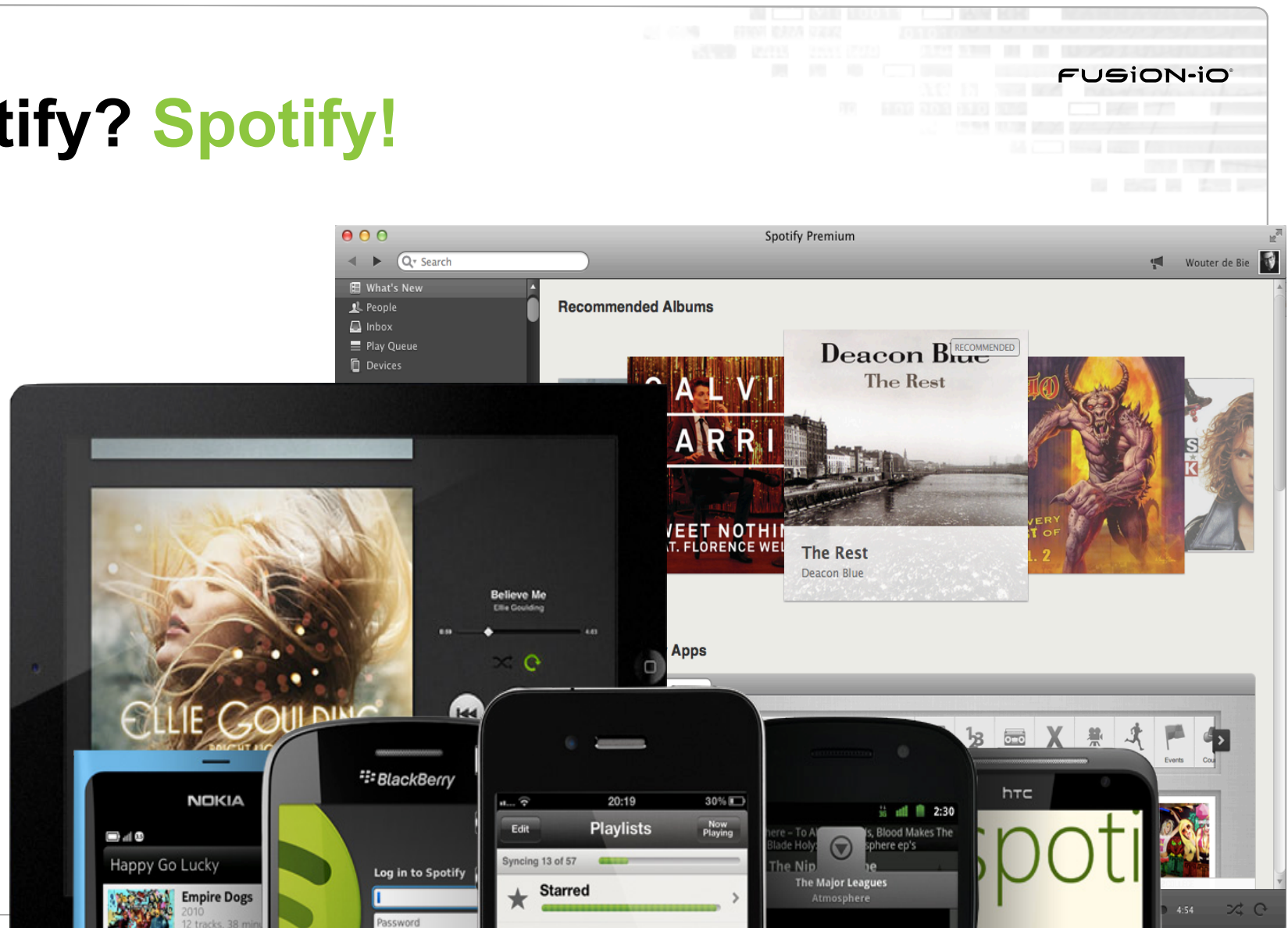
FUSION-io®

## SPOTIFY STRIKES A CHORD WITH CASSANDRA

GLOBAL ONLINE MUSIC LEADER ACCELERATES ITS  
MUSIC DATABASE WITH ioMEMORY



# Spotify? **Spotify!**





Spotify®

FUSION-io®

- ▶ Over 24 million active users
- ▶ Over 20 million songs available globally
- ▶ Over 6 million paying subscribers
- ▶ Over 1 billion playlists created
- ▶ Over \$500 million paid to rights-holders
- ▶ Over        employees
- ▶ Over        developers
- ▶ Available in: 28 countries - USA, UK, Australia, New Zealand, Germany, Sweden, Finland, Norway, Denmark, France, Spain, Austria, Belgium, Switzerland, The Netherlands, Ireland, Luxembourg, Italy, Poland, Portugal, Mexico, Singapore, Hong Kong, Malaysia, Lithuania, Latvia, Estonia and Iceland.





# Cassandra at Spotify

- Over 24 clusters and quickly growing
- Containing over 300 nodes
- Distributed over 4 data centers around the world
- Our main solution for scalable storage





## Why flash?

- It changes everything, is a step change going from spinning disks to flash
- Cassandra is page cache bound - flash moves scaling from memory to flash
- Allow us to both consolidate and scale our clusters at the same time
- Developers can focus on delivering products instead of optimizing for I/O





## Why Fusion-io?

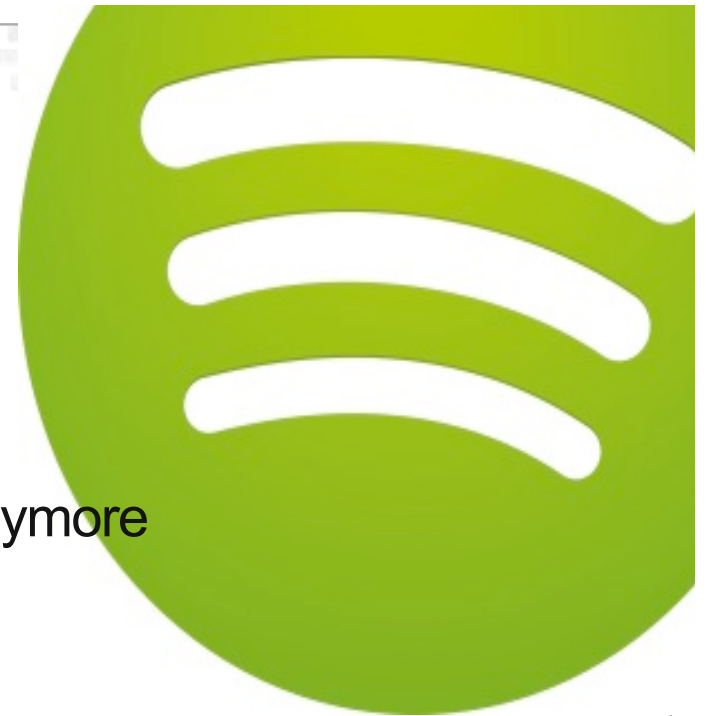
- Why attach flash to a **legacy** platform?
- It turns out that it's **easier** to get installed
- **Developer kit** allow direct access to flash
- Performance





## Early results

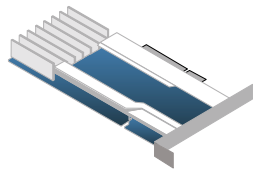
- 3-4x consolidation factor
- 3-6x reduction in latency
- Forcing SSTables to memory not needed anymore
- ROI so far is 2.2x
- Consolidation limited by Cassandra 1.1





# One iodrive vs a 10-disk raid-0 - MongoDB

FUSION-io®



VS



ioDrive 2 – 1.2TB

10 x 7200 RPM SAS disks in raid-0 mdarray



## YCSB Details

FUSION-io®

- 2 x 8-core 2.9GHz Xeon/64GB/2x10GbE
- XFS

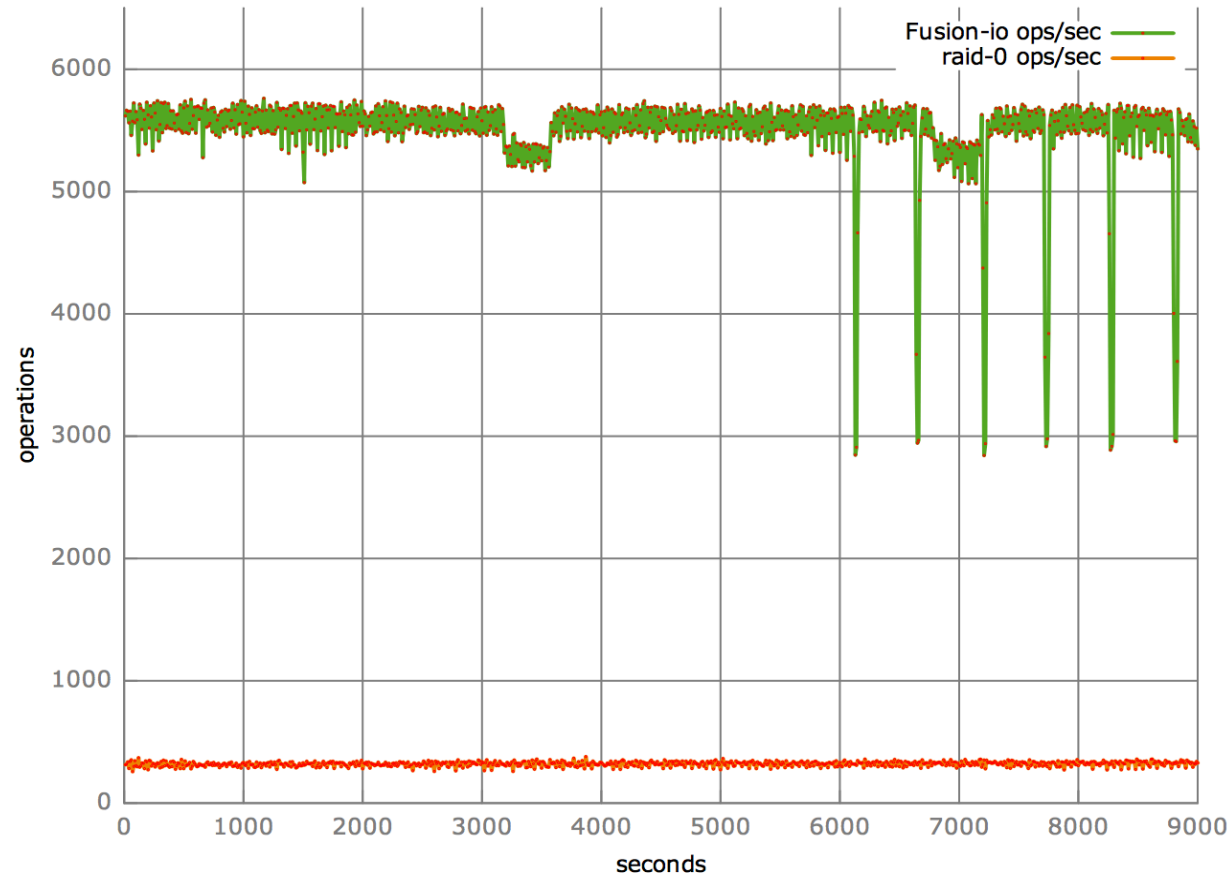
### Test

- Workload A: 50/50% read/write mix
- Workload B: 95/5% read/write mix
- Workload C: 100% read
- Workload F: 50/50% read/read+modify+write mix



## Workload A - 50/50% read/write mix

FUSION-io®



September 24,  
2013

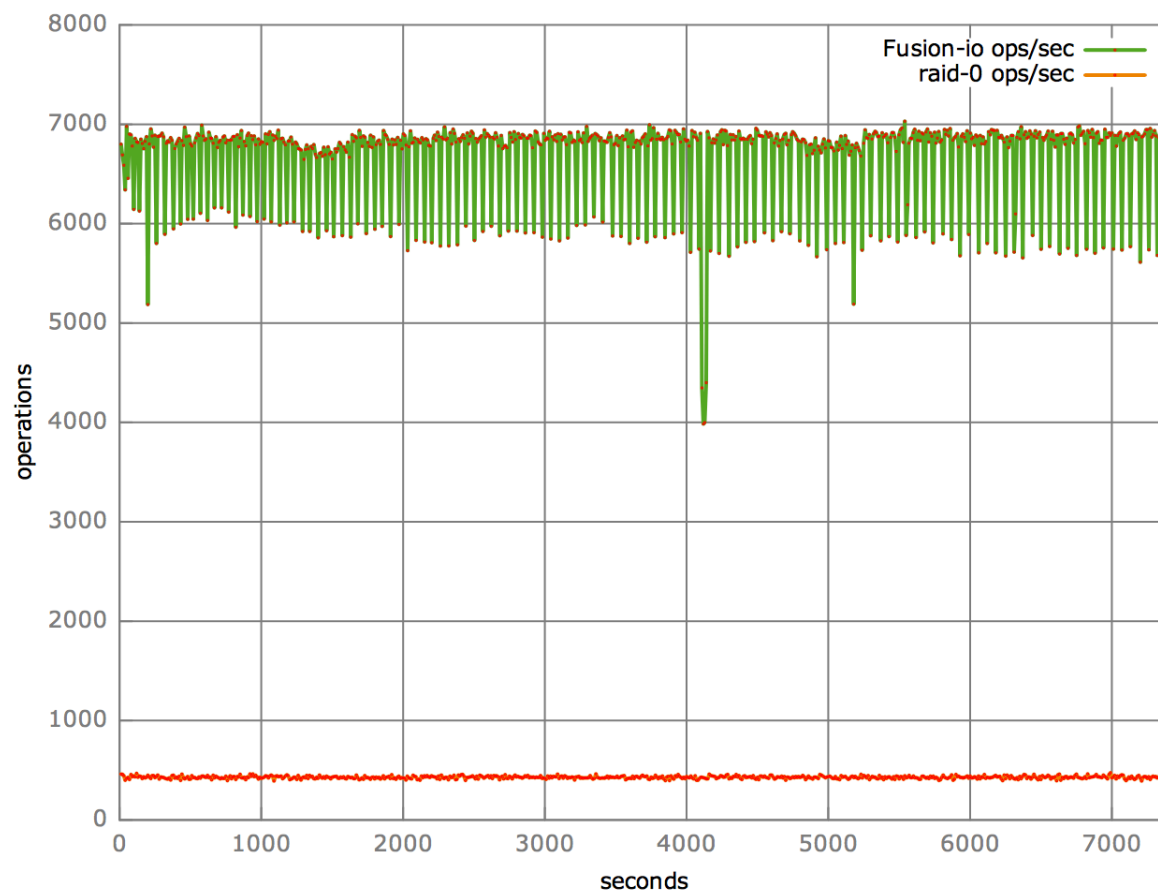
Fusion-io

57



## Workload B - 95/5% read/write mix

FUSION-io®



September 24,  
2013

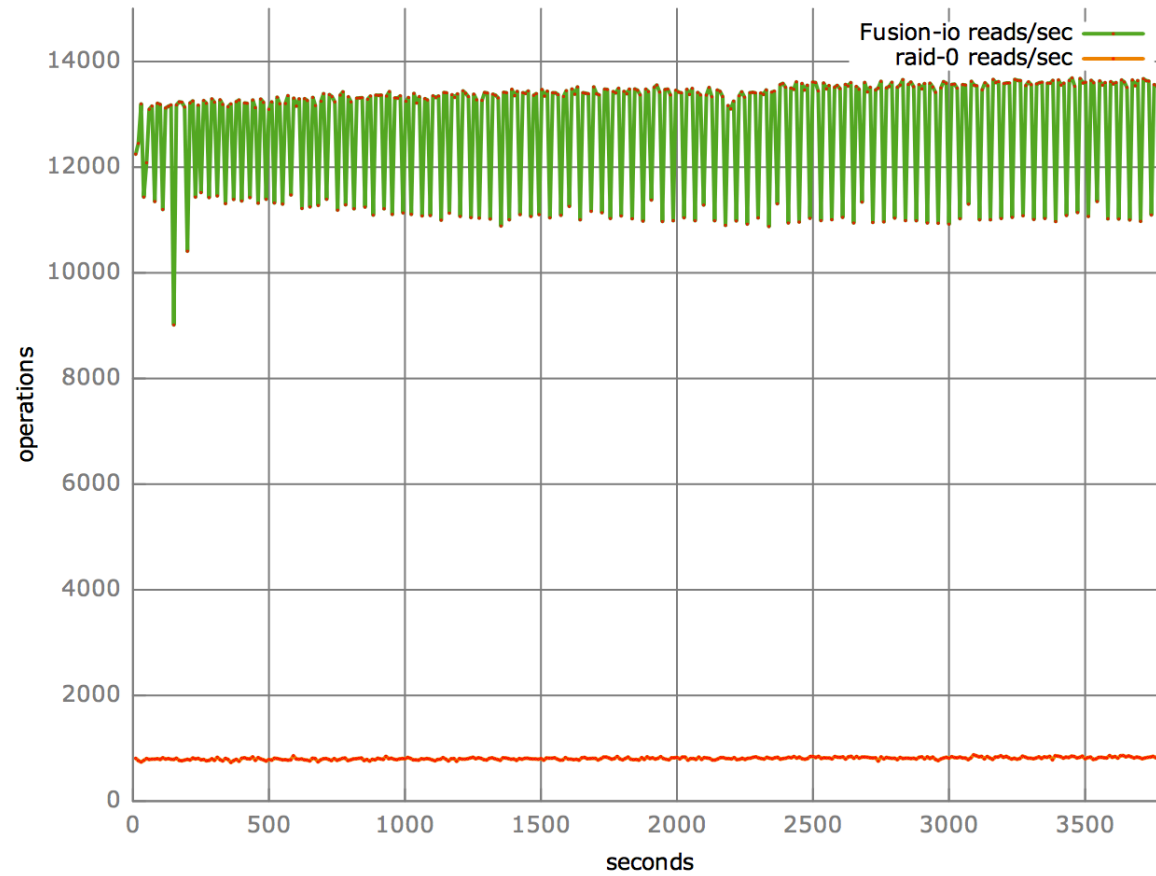
Fusion-io

58



## Workload C - 100% read

FUSION-io®



September 24,  
2013

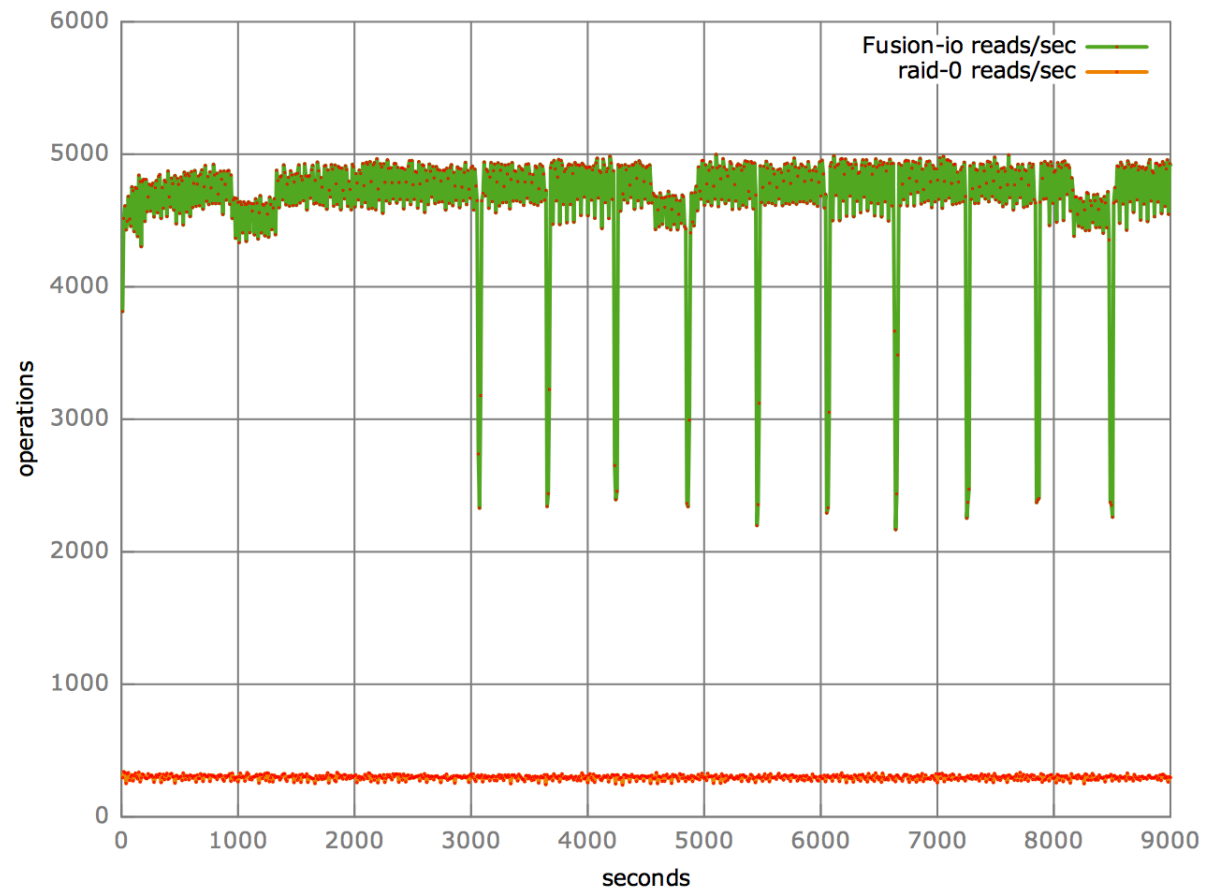
Fusion-io

59



## Workload F - 50/50% read/read+modify+write mix

FUSION-io®



September 24,  
2013

Fusion-io

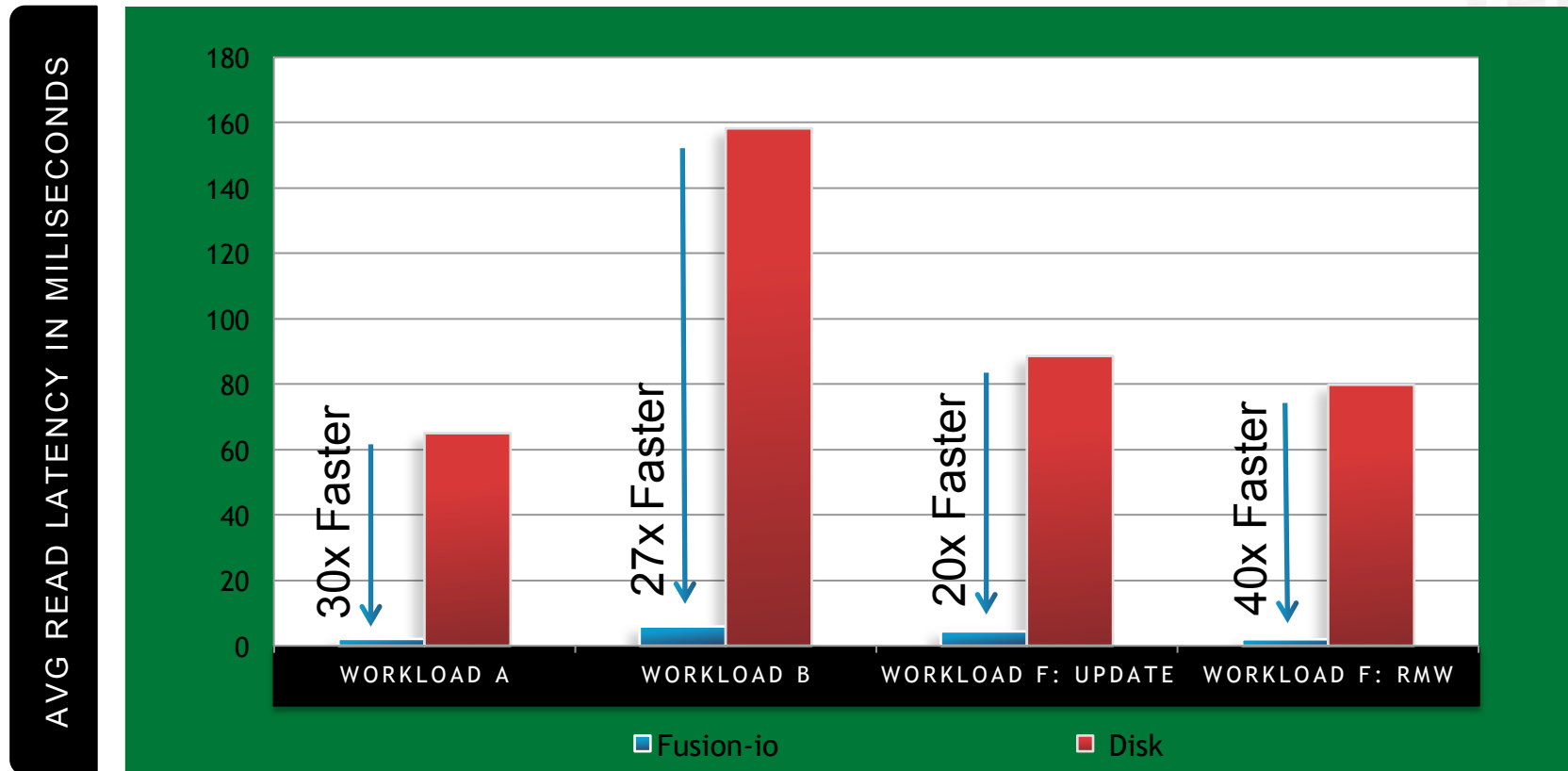
60





## YCSB: Lowest read latency wins

FUSION-io®



September 24,  
2013

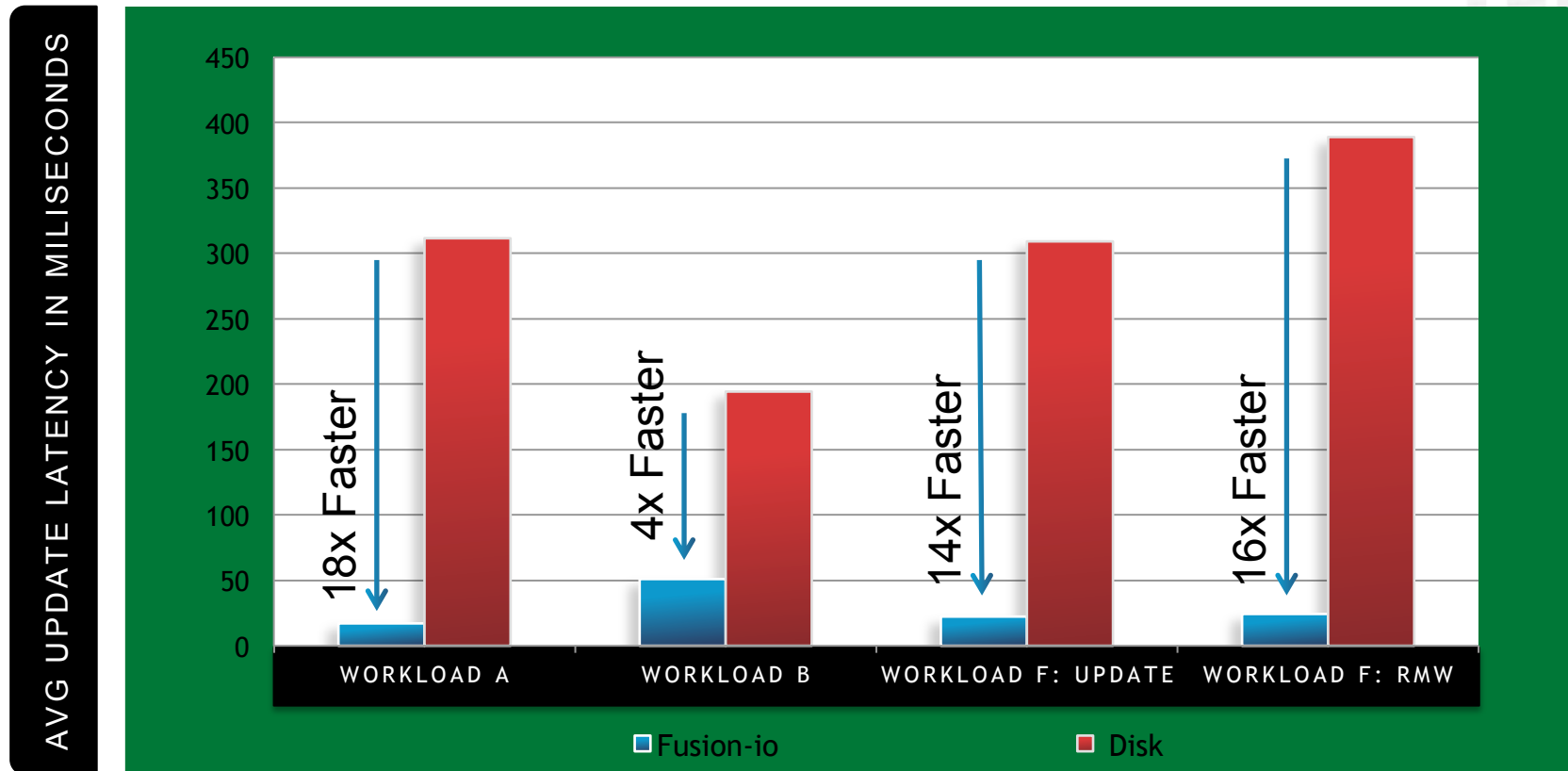
Fusion-io

61



## YCSB: lowest Update latency wins

FUSION-io®



September 24,  
2013

Fusion-io

62



# Architectural impact of Fusion-io

FUSION-io

- Avoid scaling out for DRAM
  - Nodes can handle higher transaction load
  - Terabytes of low latency persistent storage on single nodes
  - Potentially avoid sharding
- Use less DRAM per node
  - Lower cost servers
  - DRAM available for applications



# Topics – NoSQL Zurich 2013

FUSION-io®

1. What are we building ?
2. Why are we building it?
3. OpenNVM
4. Use Cases
5. Where are we headed?



## API Specs posted at opennvm.github.io

FUSION-io®

Direct-access to NVM is for developers whose software retrieves and stores data.

- ▶ Early-access to OpenNVM API specs and technical documentation (limited enrollment during early-access phase)
  - ▶ <http://opennvm.github.io>
- 
- **Write less code** to create high-performing apps
  - **Tap into performance** not available with conventional I/O access to SSDs
  - **Reduce operating costs** by decreasing RAM while increasing NVM



# Open Interfaces and Open Source

FUSION-io®

- NVM Primitives: Open Interface
- NVMFS: Open Source, POSIX Interface
- NVM API Libraries: Open Source, Open Interface
- INCITS SCSI (T10) active standards proposals:
  - SBC-4 SPC-5 Atomic-Write  
<http://www.t10.org/cgi-bin/ac.pl?t=d&f=11-229r6.pdf>
  - SBC-4 SPC-5 Scattered writes, optionally atomic  
<http://www.t10.org/cgi-bin/ac.pl?t=d&f=12-086r3.pdf>
  - SBC-4 SPC-5 Gathered reads, optionally atomic  
<http://www.t10.org/cgi-bin/ac.pl?t=d&f=12-087r3.pdf>
- SNIA NVM-Programming TWG active member



# Catalyst for top industry players to Accelerate pursuit of NVM programming

FUSION-io®



## SNIA Links:

- Webcasts
- Videos
- Certification
- Tutorials
- Multimedia
- e-Courses
- Standards
- Events
- News
- Membership
- Solid State Storage

## A Message from SNIA Technical Council

### SNIA CALL FOR PARTICIPATION NVM Programming Technical Work Group (TWG)

The SNIA Technical Council has recently approved a new technical work group. The NVM Programming TWG was created for the purpose of accelerating availability of software enabling NVM (Non-Volatile Memory) hardware. The TWG creates specifications which provide guidance to operating system, device driver, and application developers. These specifications are vendor agnostic and support all the NVM technologies of member companies. The NVM Programming TWG:

Dell, EMC, Fujitsu, HP, Intel, NetApp, Oracle, and QLogic have all communicated their support for this activity. Development teams at several other SNIA member companies have expressed support and are waiting for official company approval to state support.



# ...And Resonating through the Industry

FUSION-io®

## The Register®

Three questions Fusion-io's rivals face after flash API bombshell  
Apps bypassing OS and disk to store hot data - chaos or breakthrough?

By [Chris Mellor](#) • [Get more from this author](#)

Posted in [Blocks and Files](#), 20th April 2012 07:29 GMT

Storage array vendors are at a disadvantage here. They need three things to play in this area:

- To remain strategically important to their customers they need to get server-connected flash hardware, or shared flash array hardware connected to servers across links fast enough to provide a memory tier, meaning PCIe-class speed.
- Then they need to get cut-through software capability equivalent to that of Fusion-io.
- They would also require software to hook up their existing arrays to the server flash, bleeding off cooling data and loading up hotter data to keep app software direct disk I/O to a minimum.

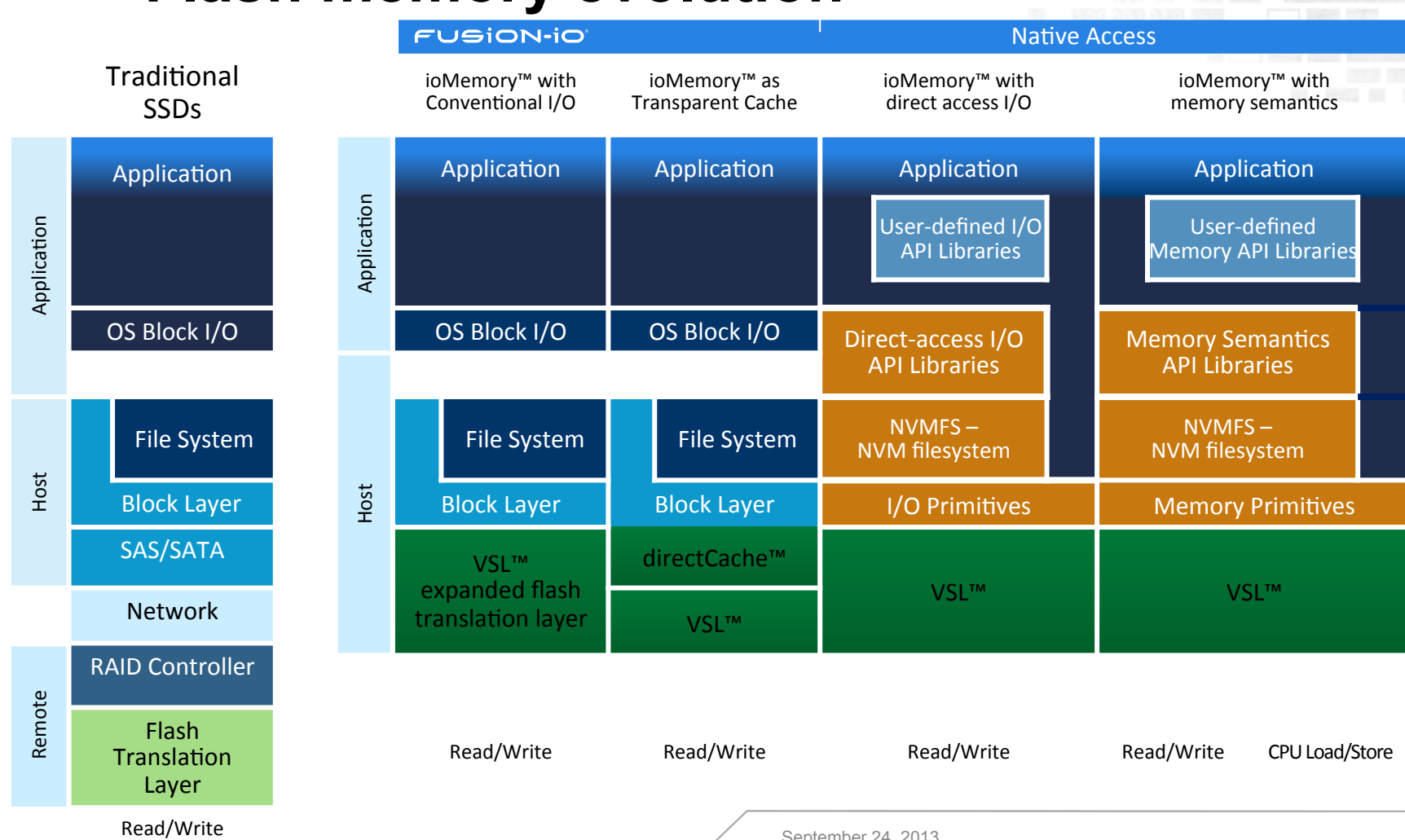
These are the table stakes I think are necessary for storage array vendors to play in the server flash application speed-up game. Getting the ability to accelerate applications by factors of 5X to 20X is going to place storage vendors in a whole new pecking order. Application acceleration glory days are there for the taking.





# Flash memory evolution

FUSION-io®



September 24, 2013

69



# Comprehensive Customer Success

FUSION-io®

FINANCIALS	WEB	TECHNOLOGY	RETAIL	MANUFACTURING/ GOVERNMENT
  <b>5x</b> FASTER DATA ANALYSIS	  <b>30x</b> FASTER DATABASE REPLICATION	  <b>40x</b> FASTER DATA WAREHOUSE QUERIES	  <b>15x</b> QUERY PROCESSING THROUGHPUT	  <b>15x</b> FASTER QUERIES

30+ case studies at <http://fusionio.com/casestudies>

September 24, 2013

70

**THANK YOU!**



[fusionio.com](http://fusionio.com) | REDEFINE WHAT'S POSSIBLE