FUSION-iO

Running NoSQL natively on flash

Conference and Trainings

APRIL 16, 2013 MUNICH

NoSQL Search roadshow

trochner@fusionio.com

Topics – NoSQL Amsterdam 2013

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

2006

Mission to consolidate memory and storage

2007

ioMemory technology unveiled

2008

First products launched

1 million IOPS IBM Quicksilver

Dell strategic investment

2009

HP OEMs products

IBM OEMs products

Samsung strategic investment

2010

Dell OEMs products

VSL introduced

2011

IPO on NYSE

ioTurbine acquired

ioDrive2 announced

Supermicro OEMs products

2012

1 Billion IOPS

2,500 customers

>120 channel and alliance partners

ioFX announced

OpenNVM introduced

Cisco OEMs products

2013

ioScale announced at Open Compute Summit

NexGen acquired

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

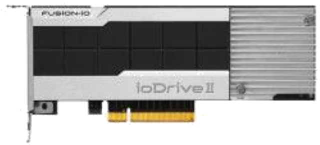
Development
PERFORCE

Caching
memcached, Powered by Squid, VARNISH SOFTWARE

Security/Logging
ArcSight, splunk>

Web
LAMP, Microsoft .NET

Direct Acceleration

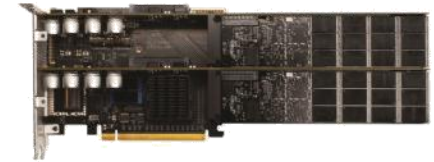


ioDrive II

Up to 3.0TB of capacity

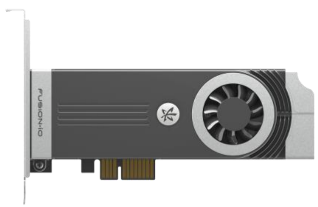
ioDrive II Duo

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

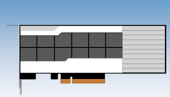
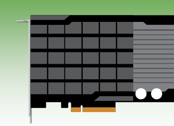
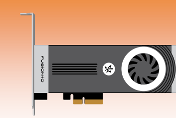
Up to 1.2TB for maximum performance density

ioScale™

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

ioMemory Solutions Platform

FUSION-io

Enterprise	Hyperscale	Workstation
ioTurbine ioN ioControl 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

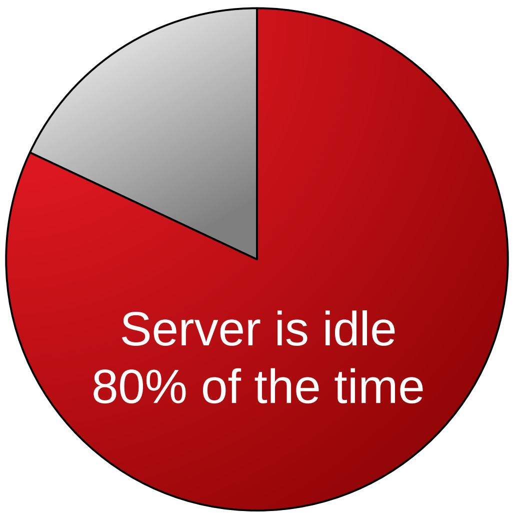
Topics – NoSQL Amsterdam 2013

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

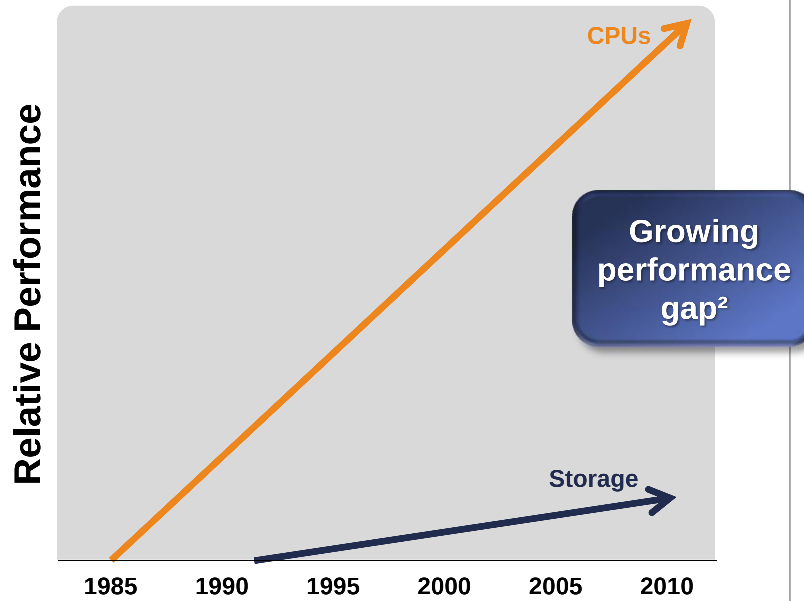
SLOW STORAGE LEADS TO IDLE CAPACITY

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

37% of servers are massively underutilized[1]…



…because the performance gap continues to grow



[1] *Source: IDC's Server Workloads 2010, July 2010*
[2] *Source: Taming the Power Hungry Data Center, Fusion-io White Paper*

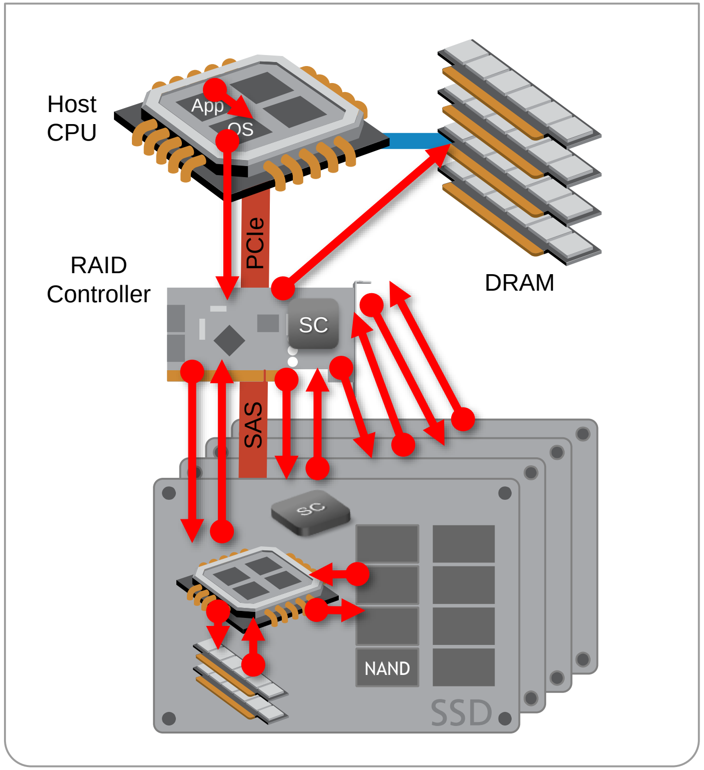
SPINNING MEDIA
OVER 150 YEARS OLD

SSD treats memory like disk

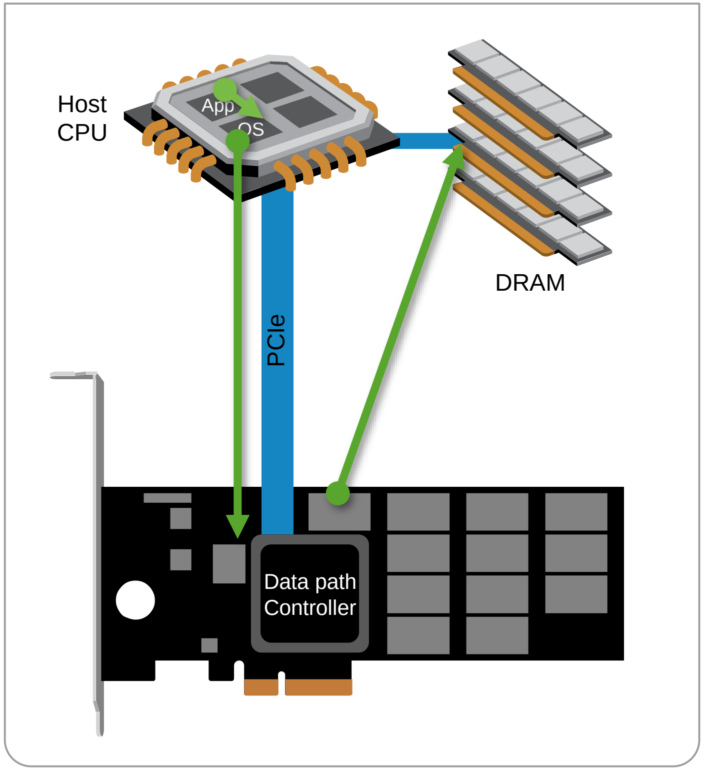
Flash Architectures

FUSION-IO

FLASH AS DISK

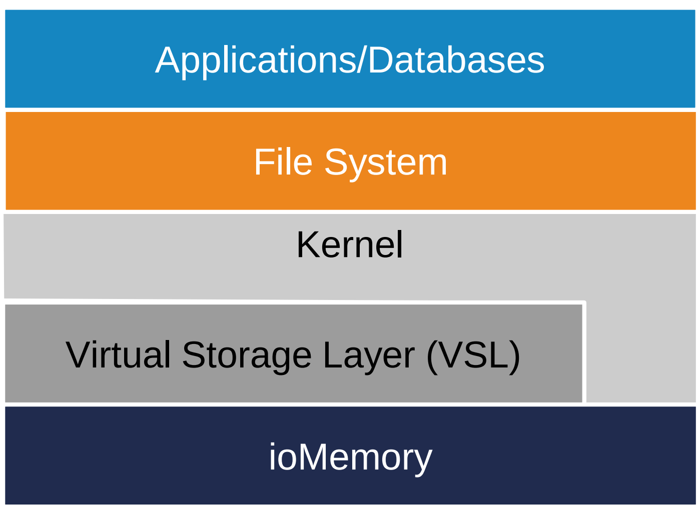


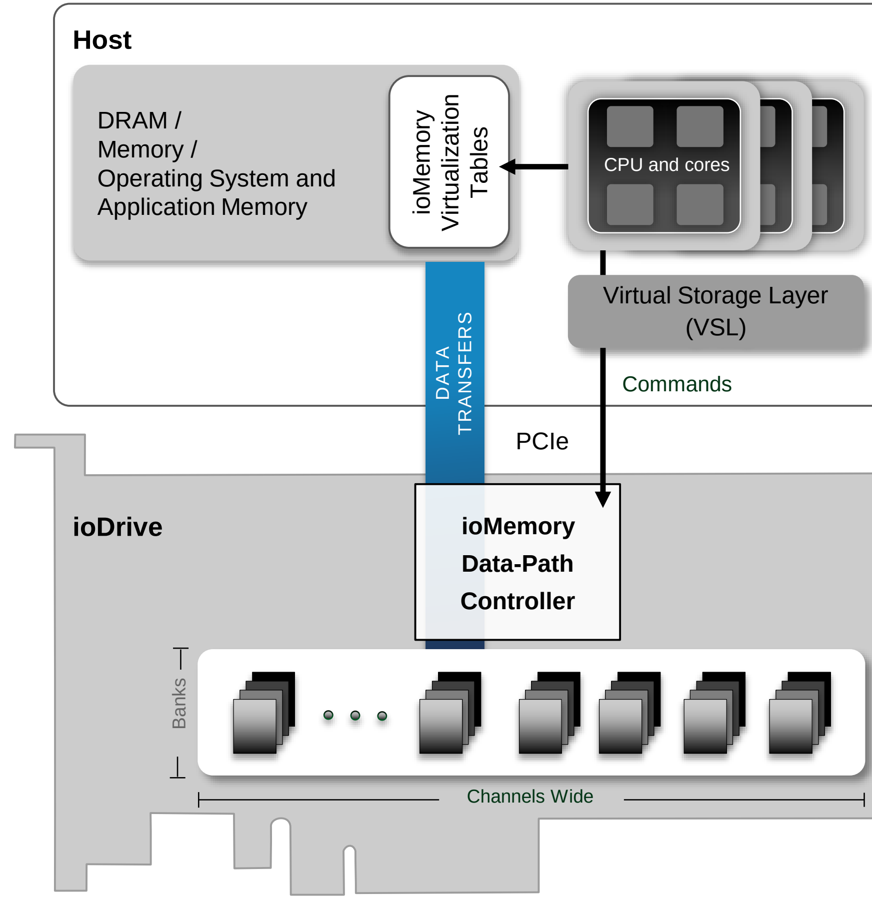
FLASH AS MEMORY



Cut-through Architecture and VSL

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

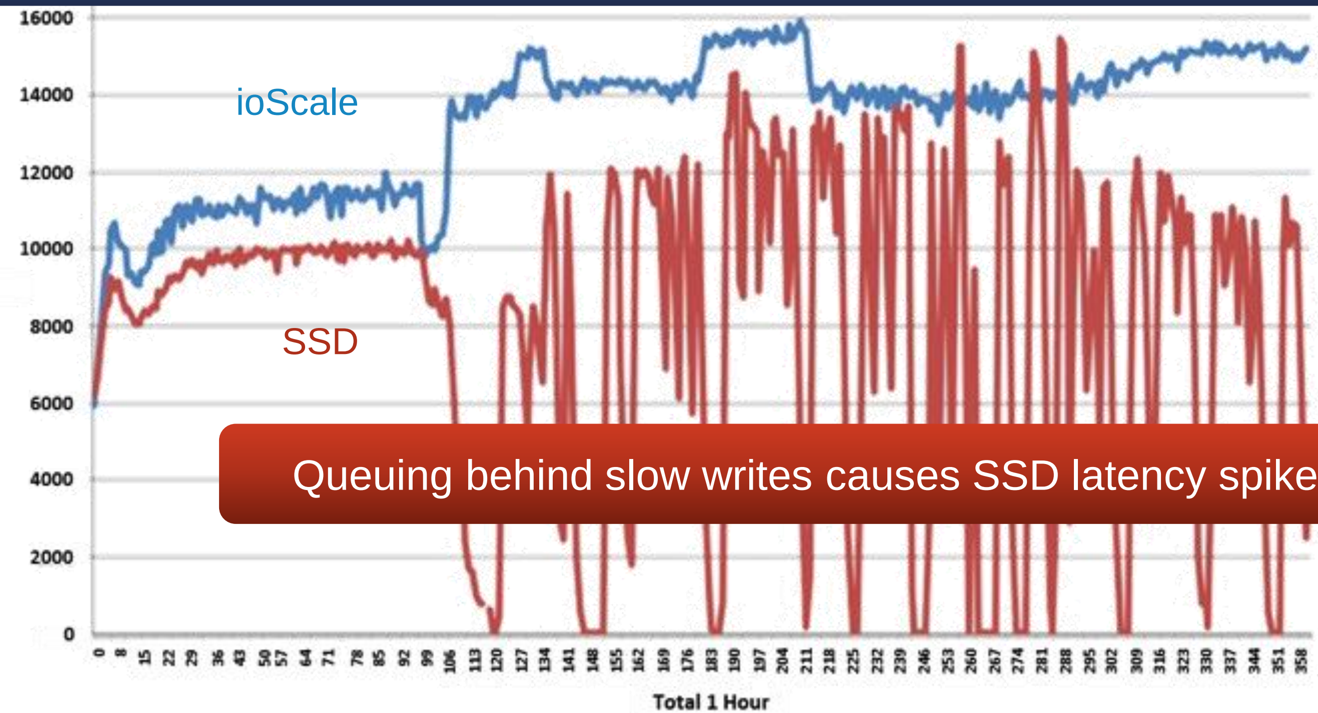




Balanced Performance Affects Throughput

FUSION-iO

ioMemory balances read/write performance for consistent throughput



Queuing behind slow writes causes SSD latency spikes

ioDrive2 specs – visit 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			

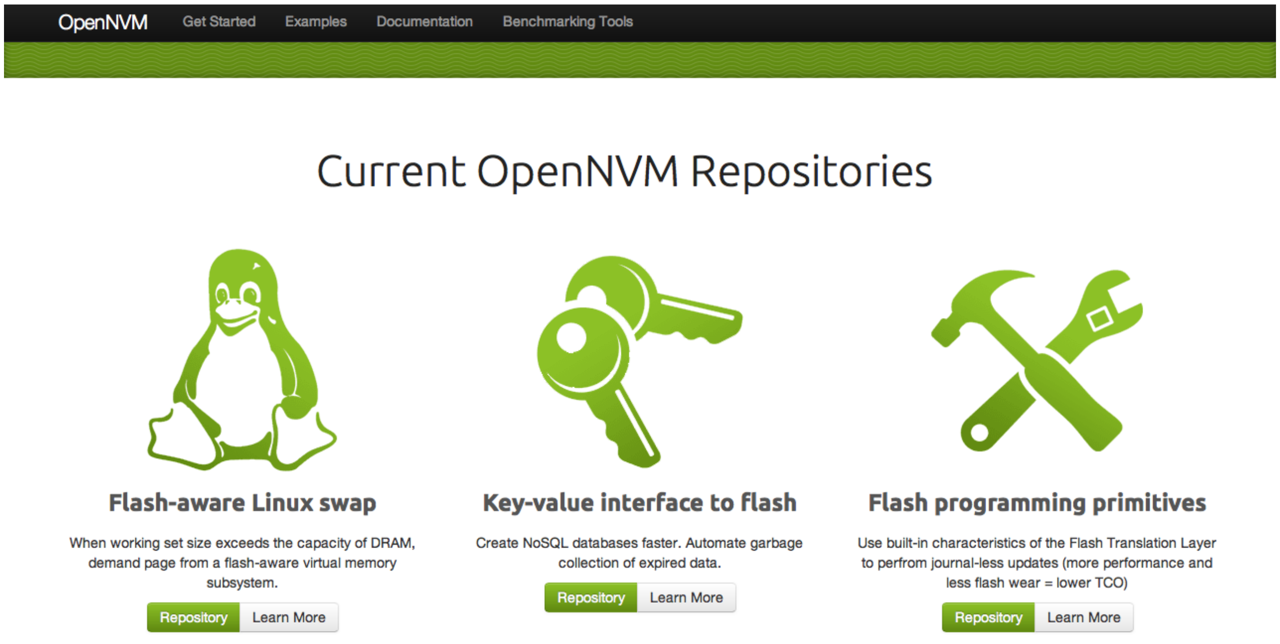
Topics – NoSQL Amsterdam 2013

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

NoSQL Software challenges

- 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



- 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

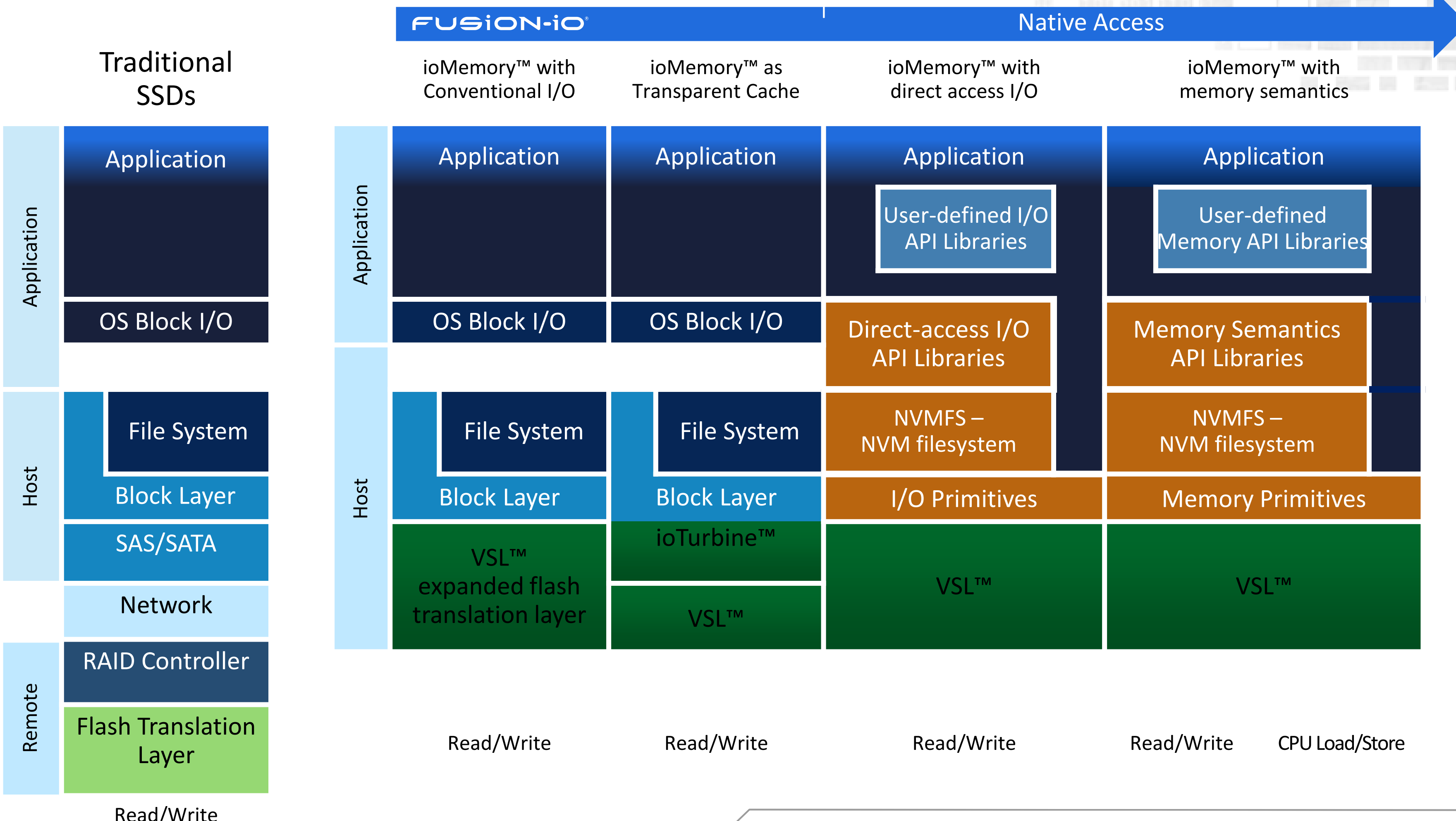
Direct-Access to non-volatile memory is now emerging

- Developers are beginning to manipulate data

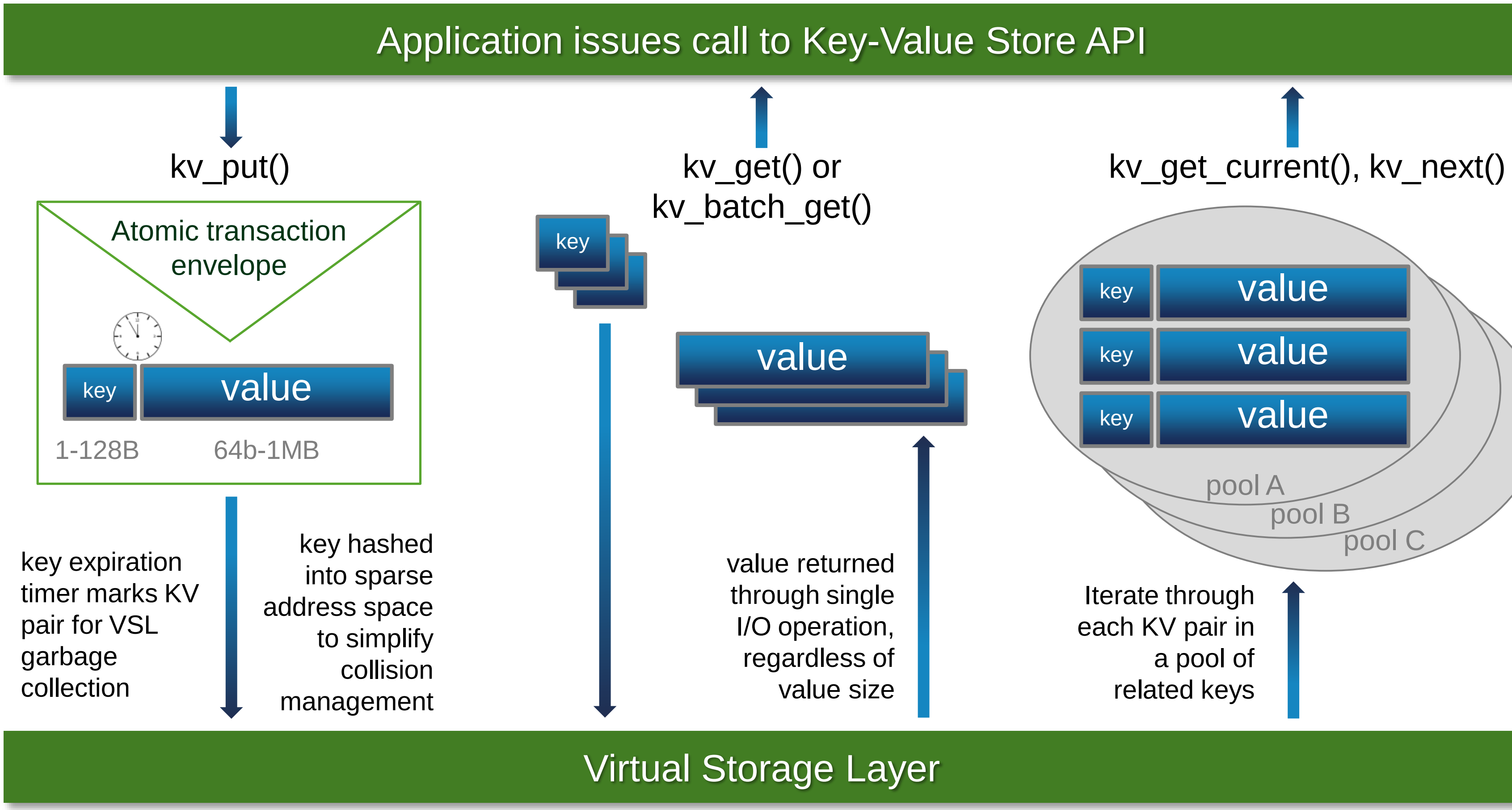
  directly in Non-Volatile Memory (NVM)

  without converting to basic block I/O.

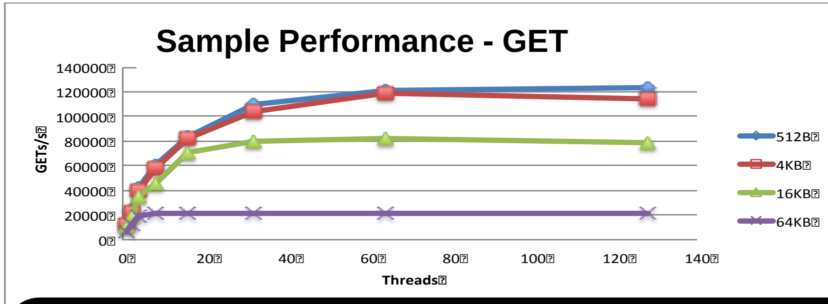
Flash memory evolution

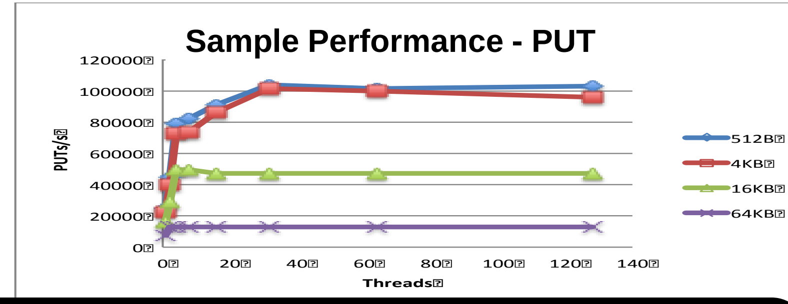


Example: Key-Value Store API Library

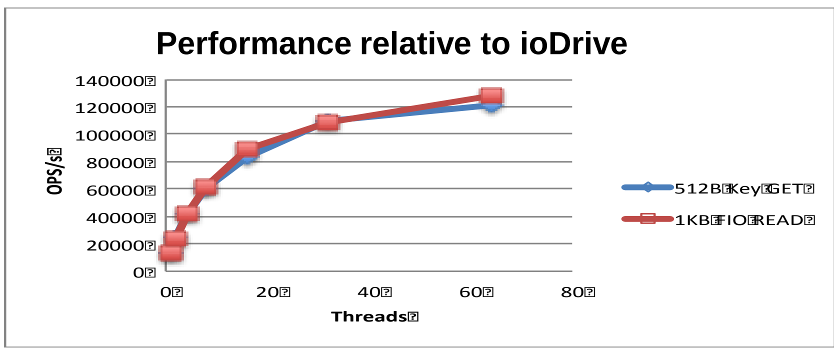


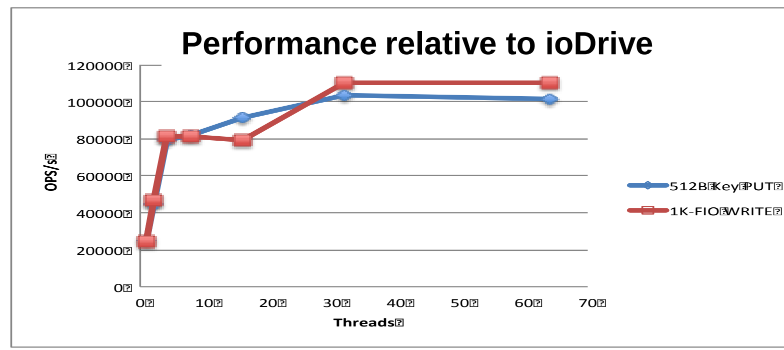
Key-Value Store API Library Benchmarks
(native KV Get/Put vs. raw reads/writes)





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.

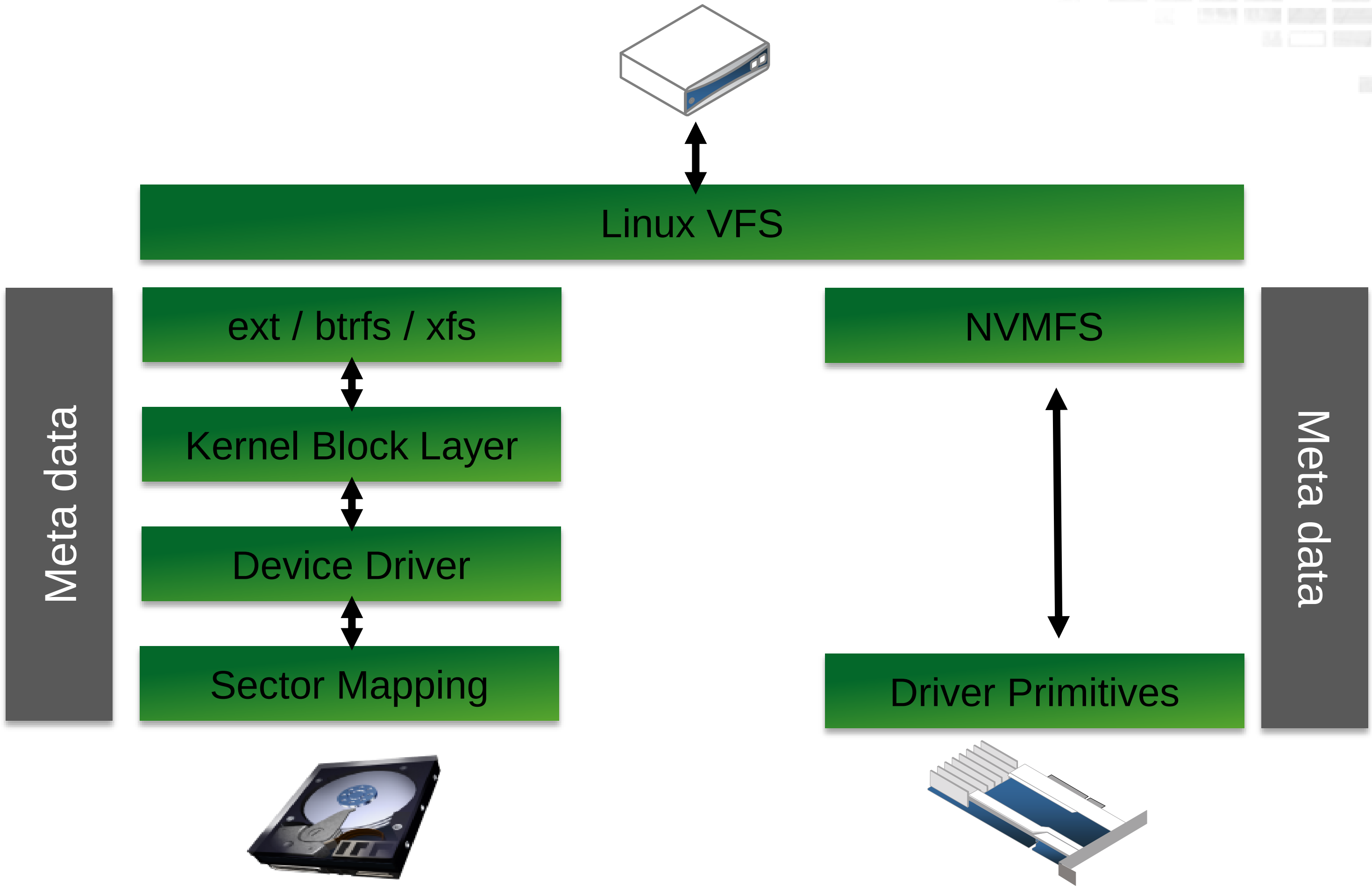
Up to 3x capacity increase

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

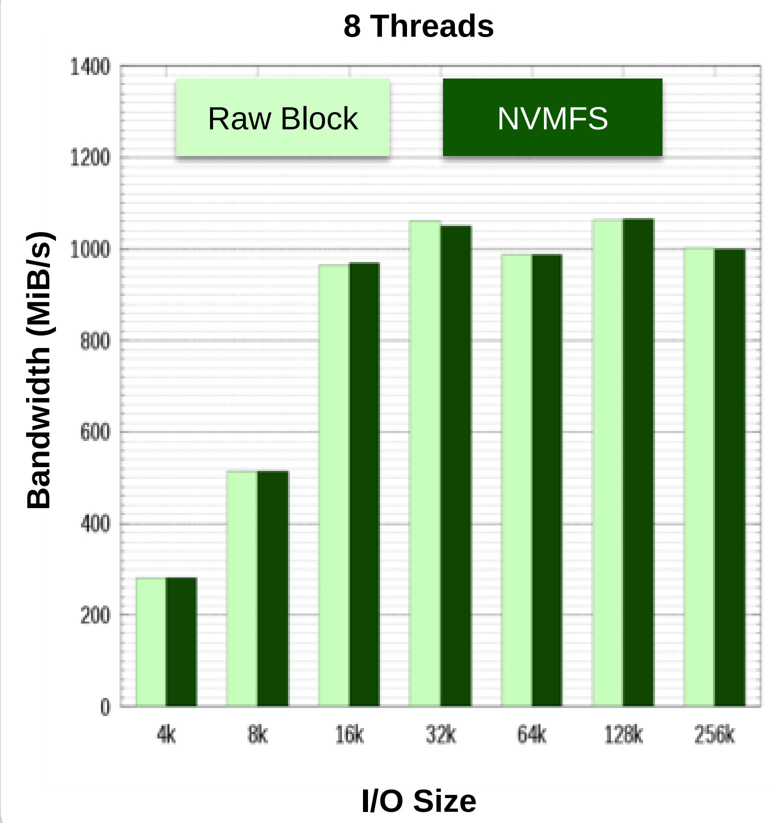


NVMFS – Benefits in Eliminating Duplicate logic

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

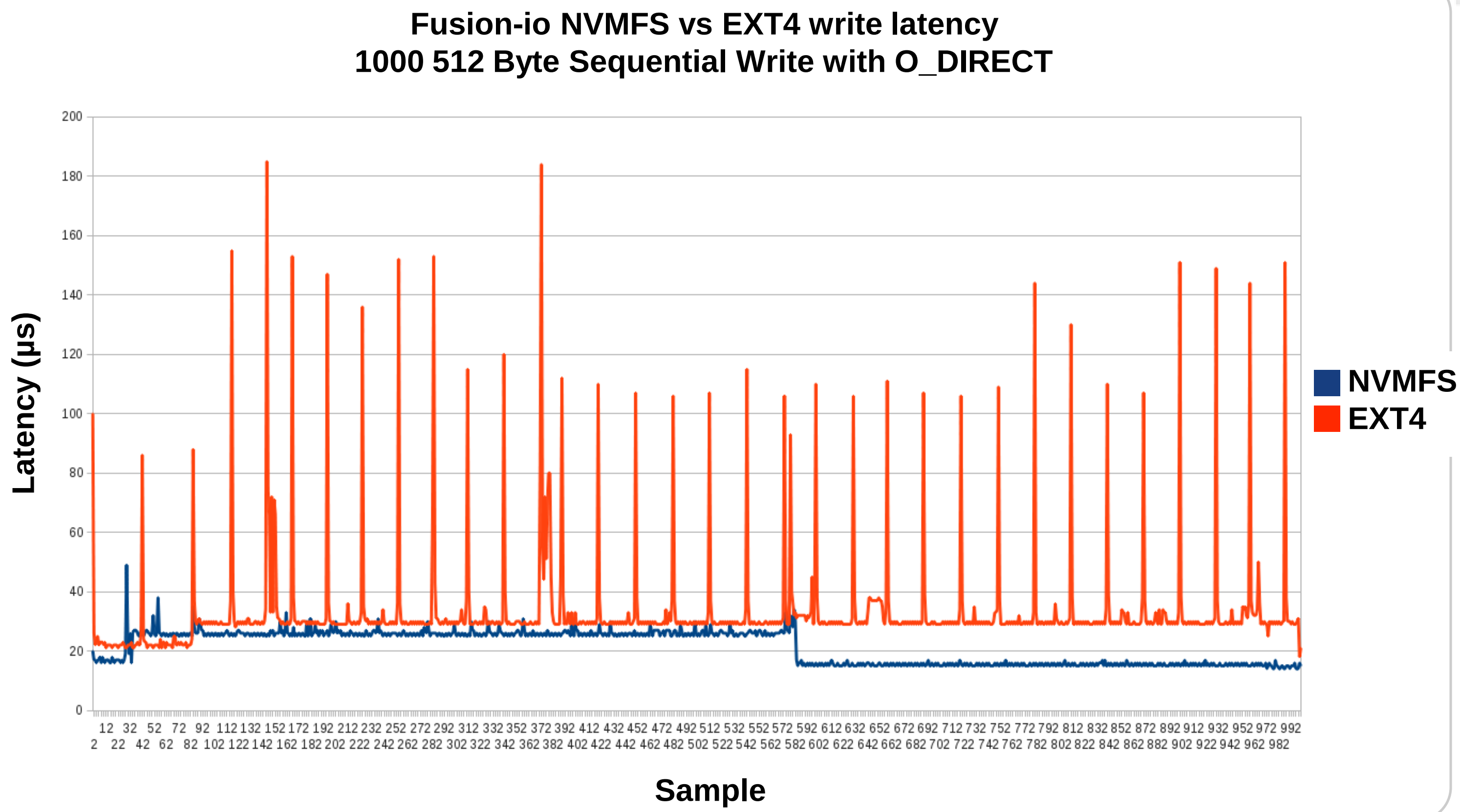
NVMFS: Native Flash Filesystem

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

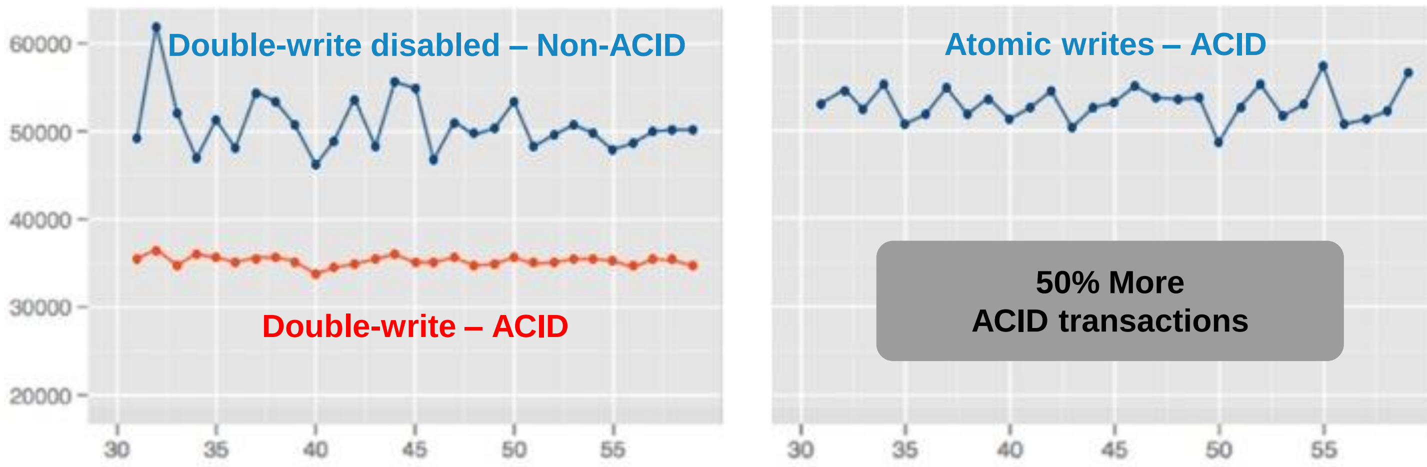


NVMFS: Consistent Low Latency

FUSION-iO



MySQL: NVMFS and Atomic Writes



XFS

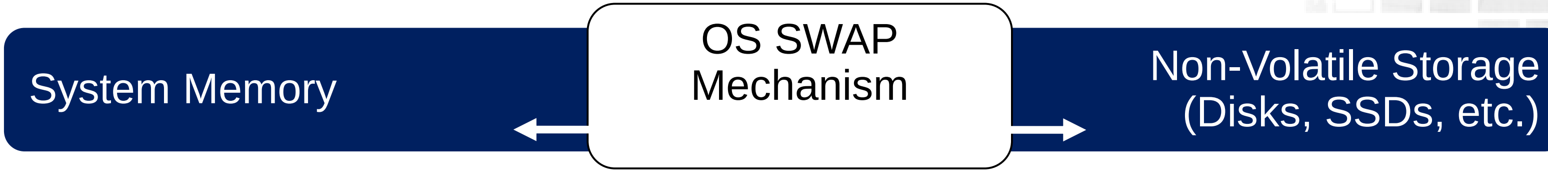
NVMFS with Atomic I/O

Range of memory-Access Semantics

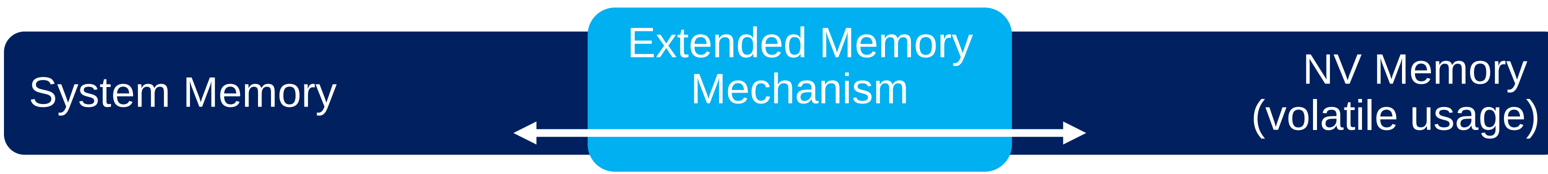
FUSiON-iO

Extended Memory	**Volatile**	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™	**Non-volatile**	Region of application memory automatically persisted to non-volatile memory and recoverable post-system failure

OS Swap vs. Extended Memory



- 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 Amsterdam 2013

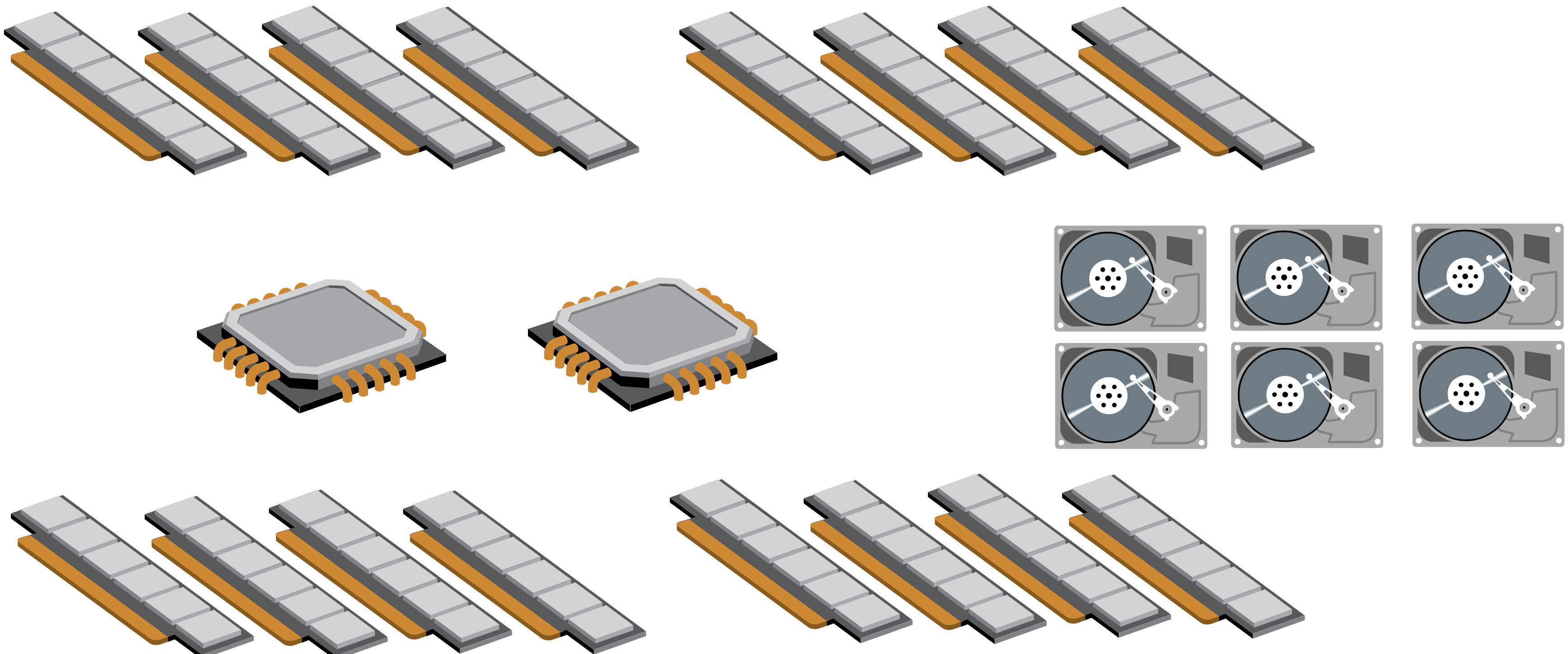
FUSION-IO

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

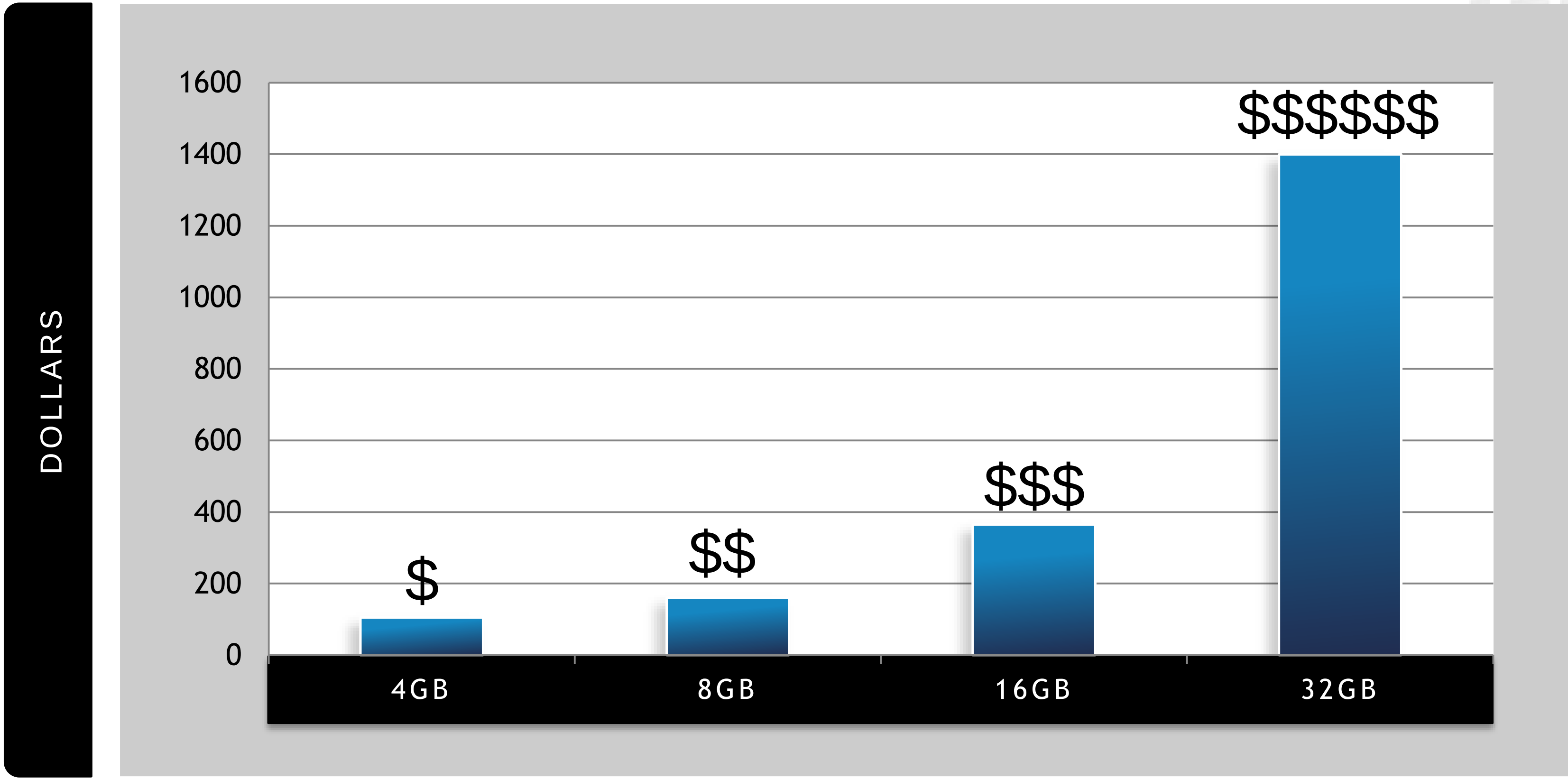
DRAM Dictates NoSQL Scaling

FUSION-iO

- Key Design Principle:
- Working Set < DRAM



Cost of DRAM Modules

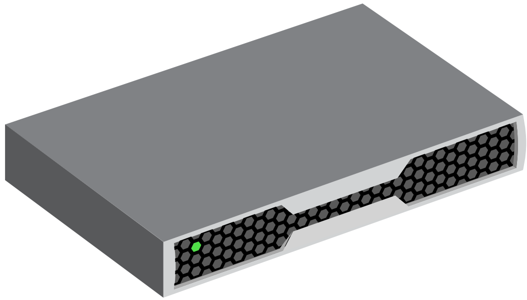


When do we scale out?

- A typical server…

CPU Cores: 32 with HT

Memory: 128 GB



…is your working set > 128GB?

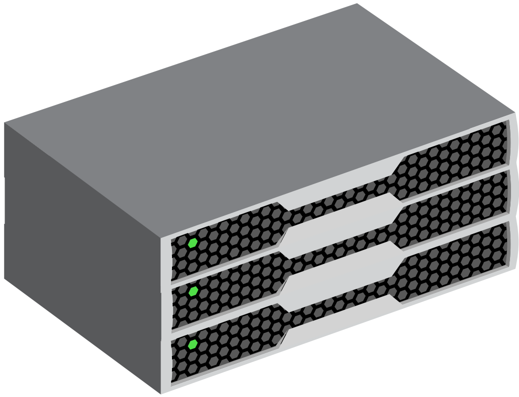
Is there a better way?

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

Combined Resources

CPU Cores: 96

Memory: 384 GB



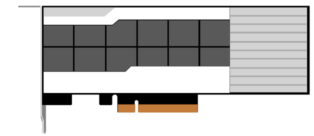
More cores than needed to serve reads and writes.

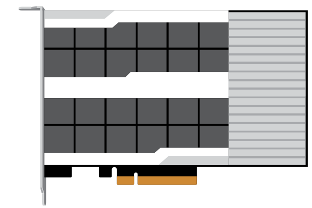
Three Deployment Options

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

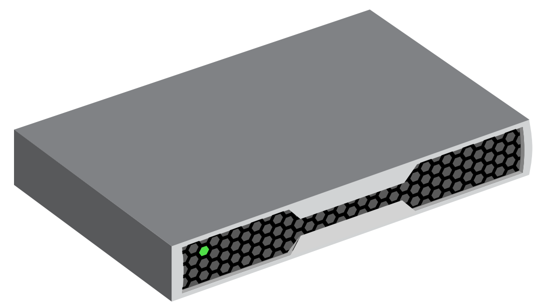
Cassandra with All-Flash Storage

FUSION-iO





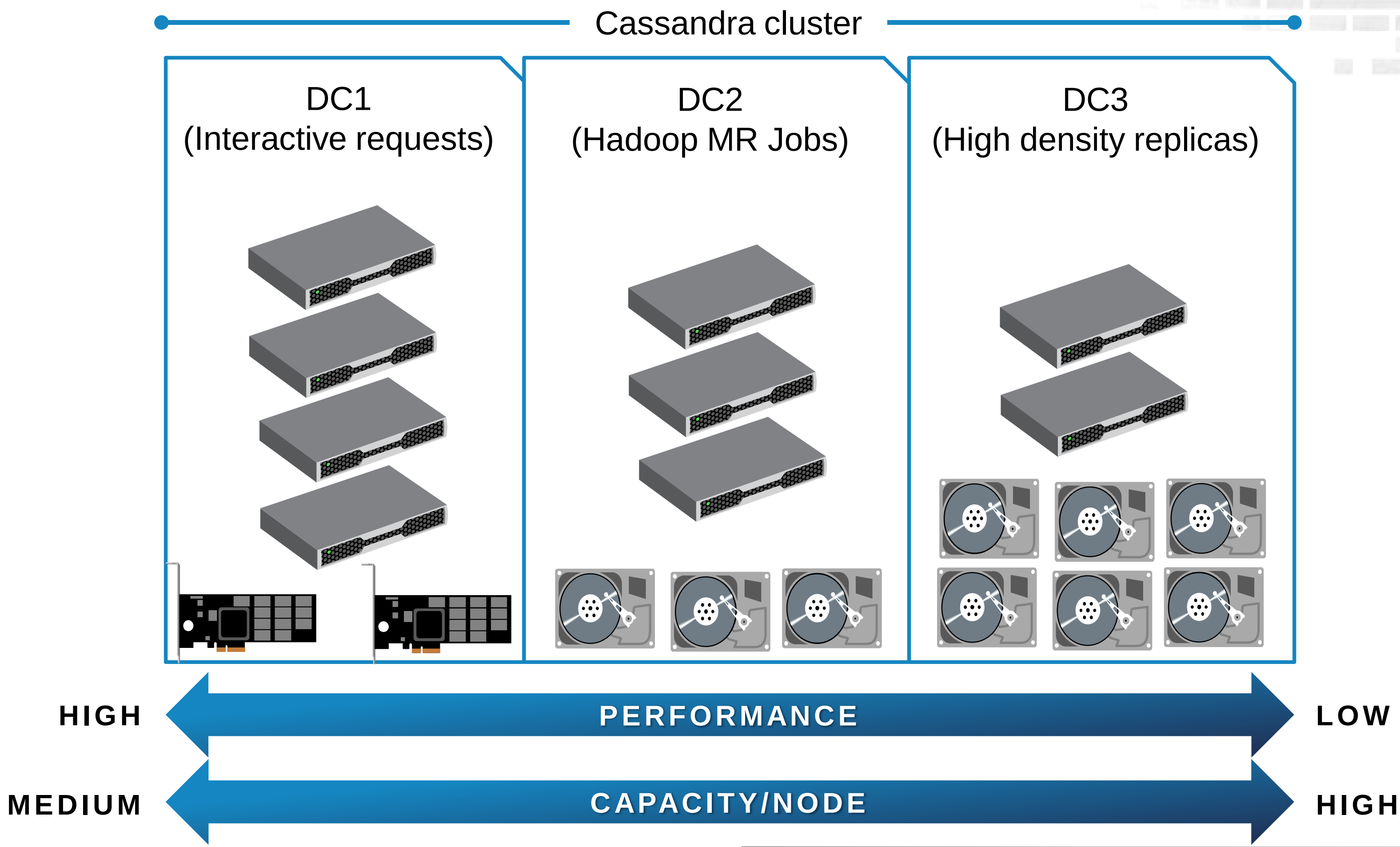
Step 1: Mount ioMemory at /var/lib/cassandra/data
Step 2:



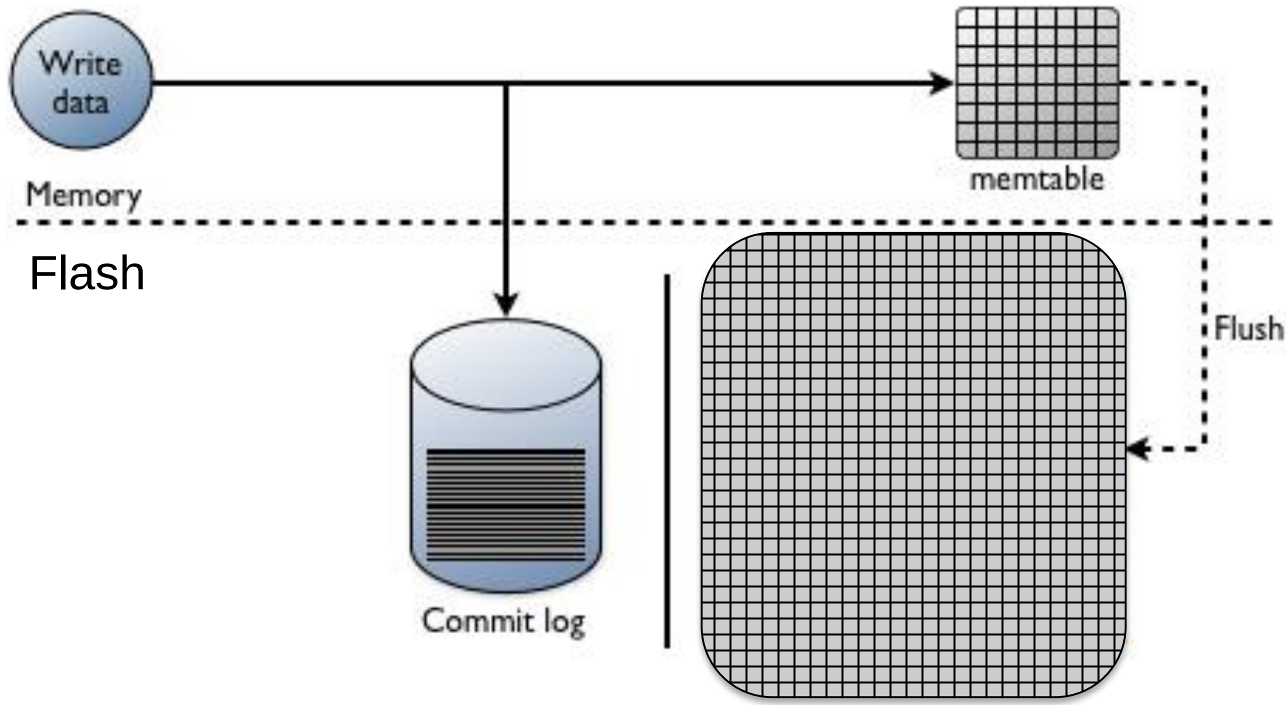
Data Placement

- 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



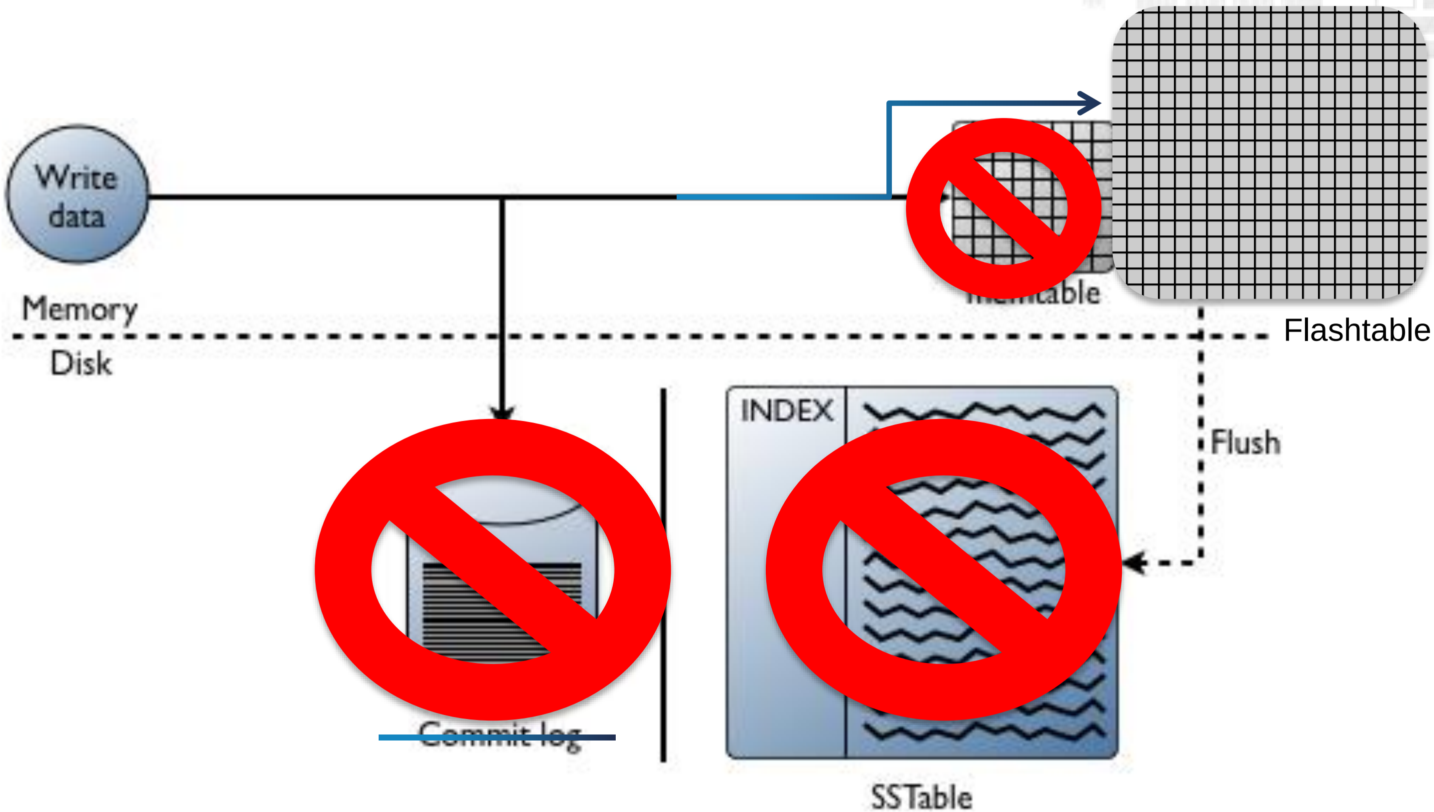
Rethinking Cassandra I/O



http://www.datastax.com/docs/1.2/dml/about_writes

Rethinking Cassandra I/O

FUSION-iO



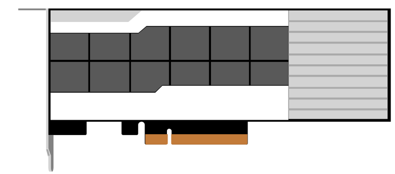
http://www.datastax.com/docs/1.2/dml/about_writes

Accelerating Cassandra With Flash

FUSION-IO

Cassandra

+



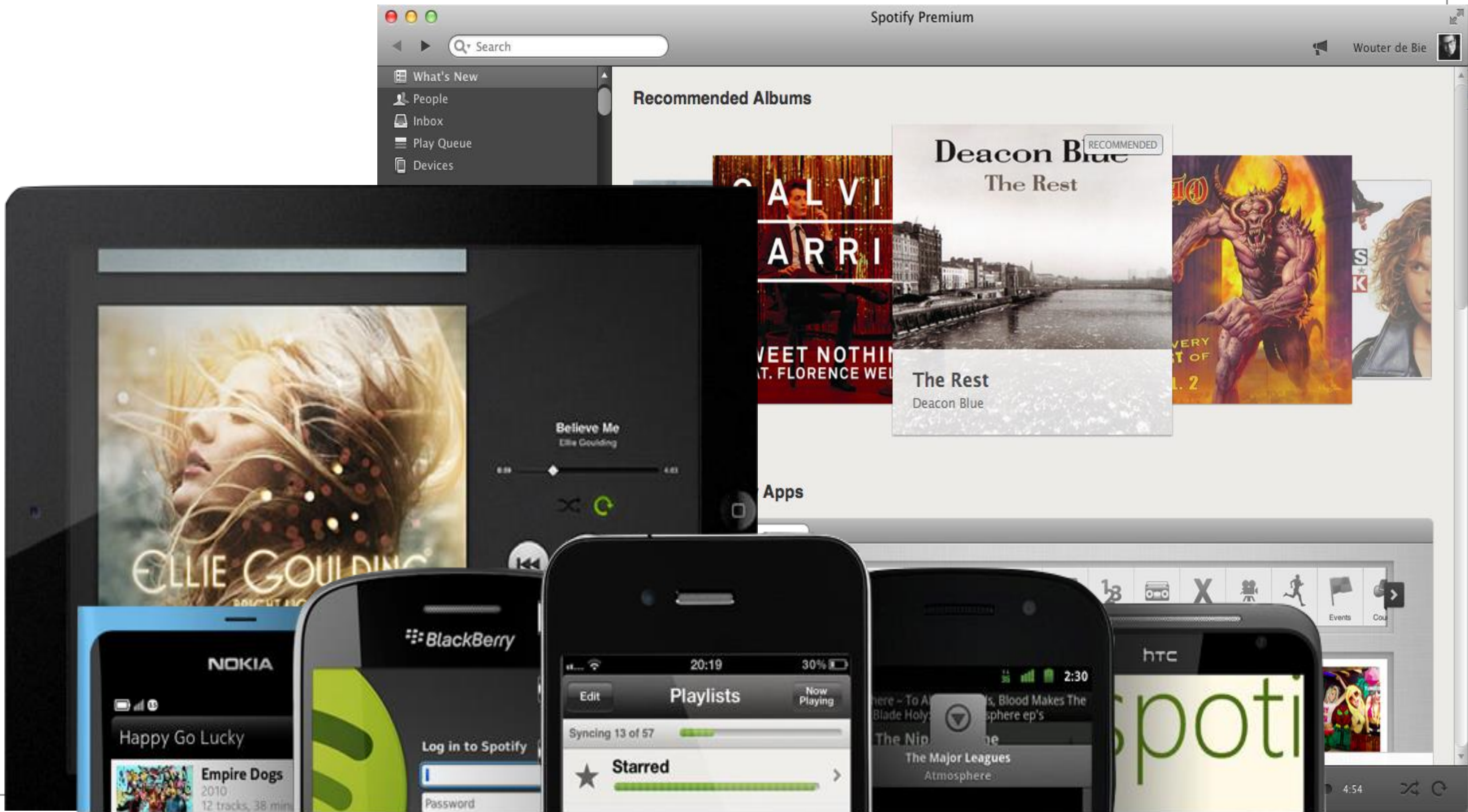
NAND Flash Accelerator

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? Spotify!



- 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

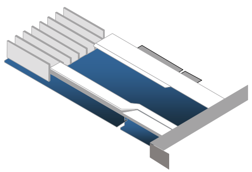
Why Fusion-io?

- 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

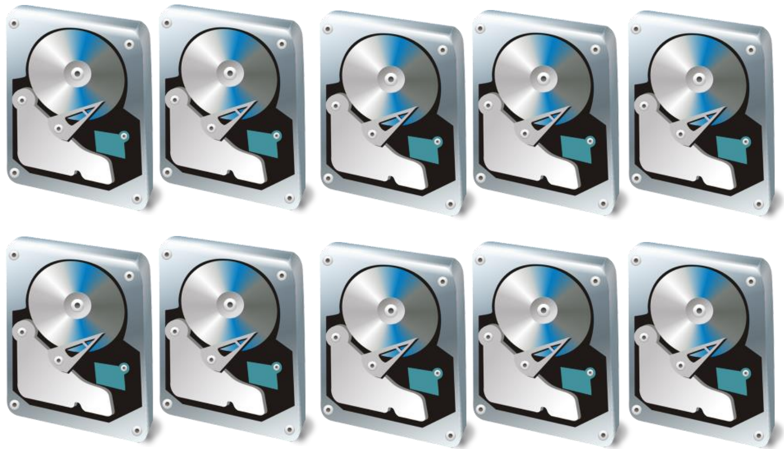
Early results

- 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



VS



ioDrive 2 – 1.2TB

10 x 7200 RPM SAS disks in raid-0 mdarray

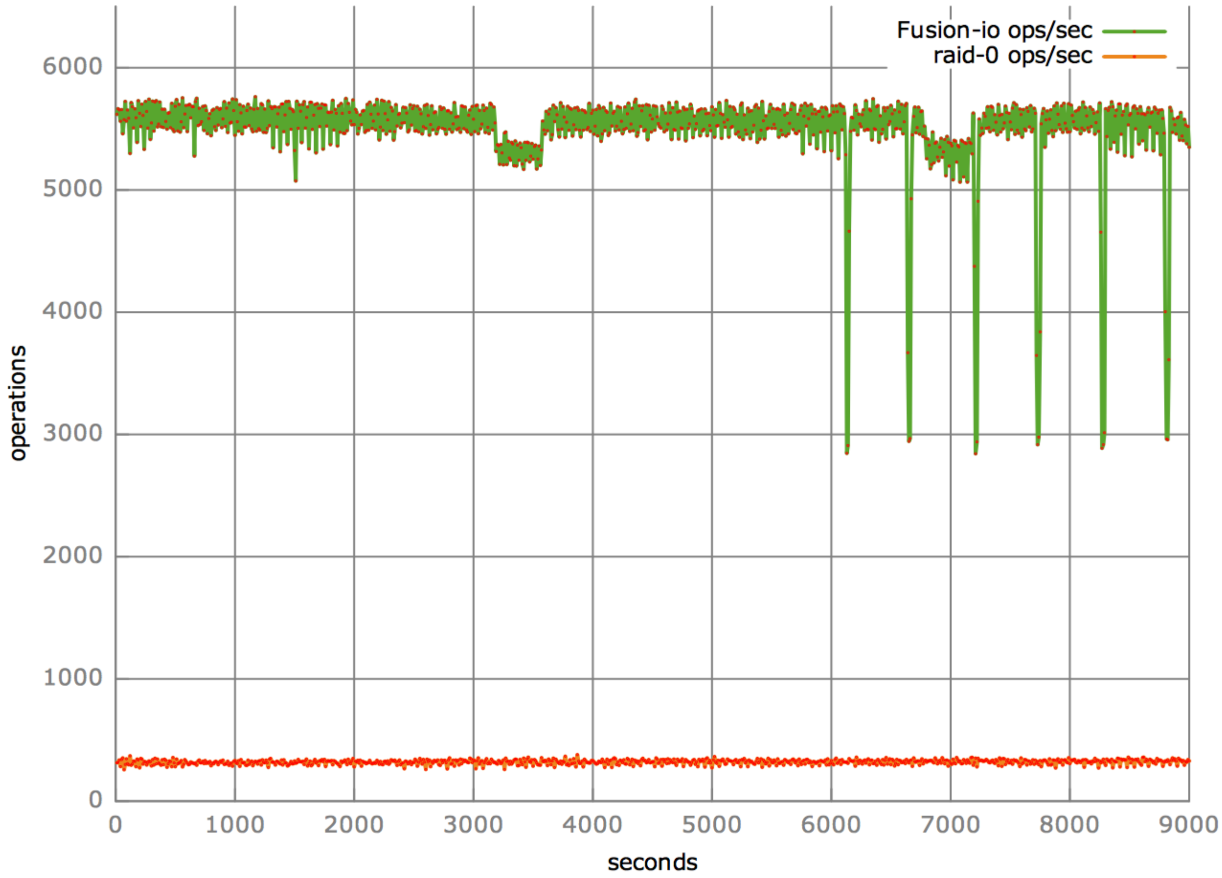
YCSB Details

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

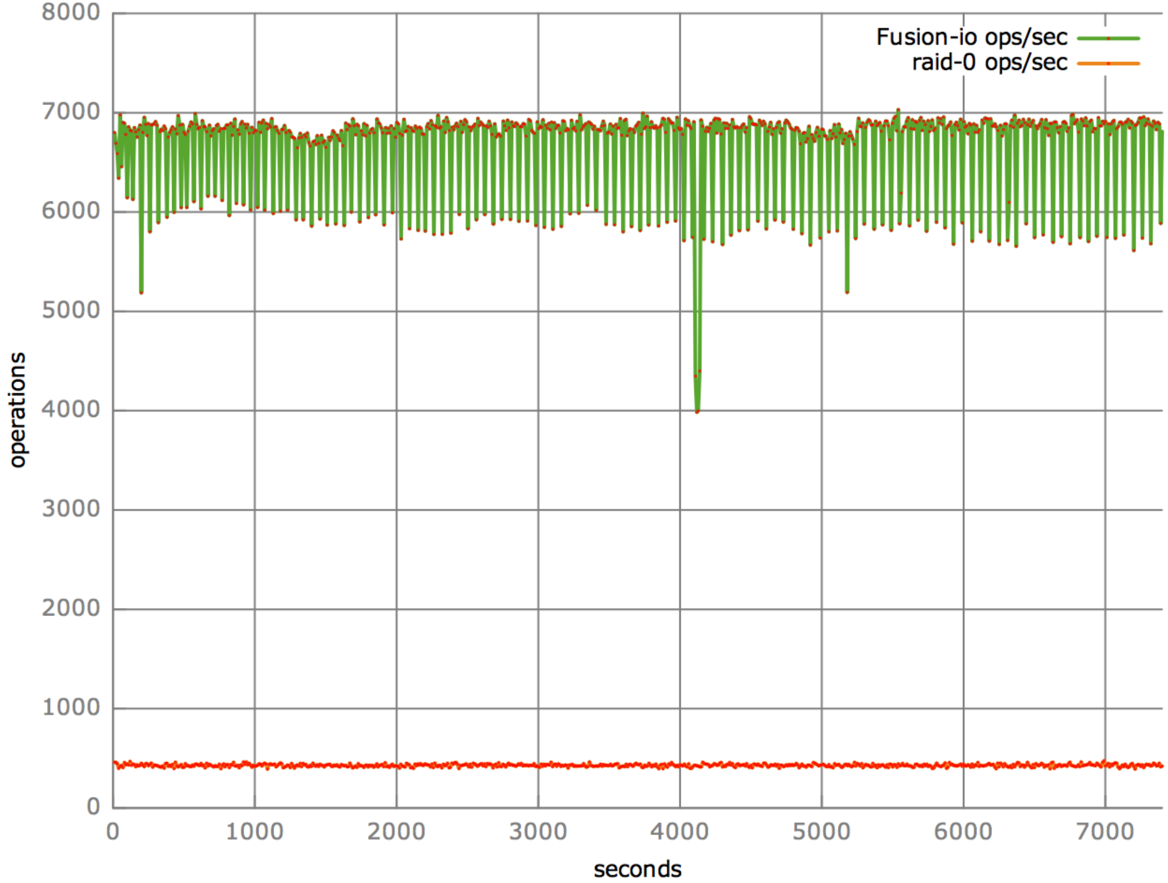
Test

- 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

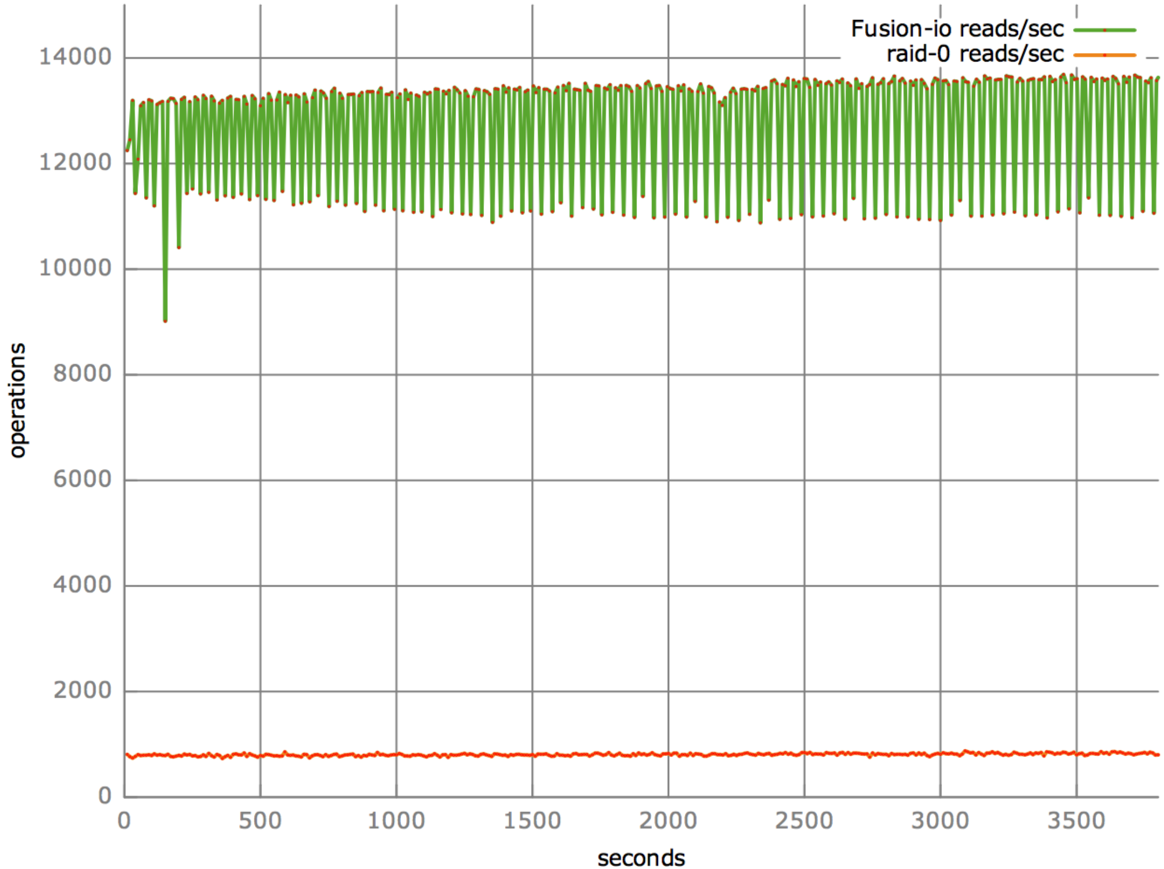
Workload A - 50/50% read/write mix



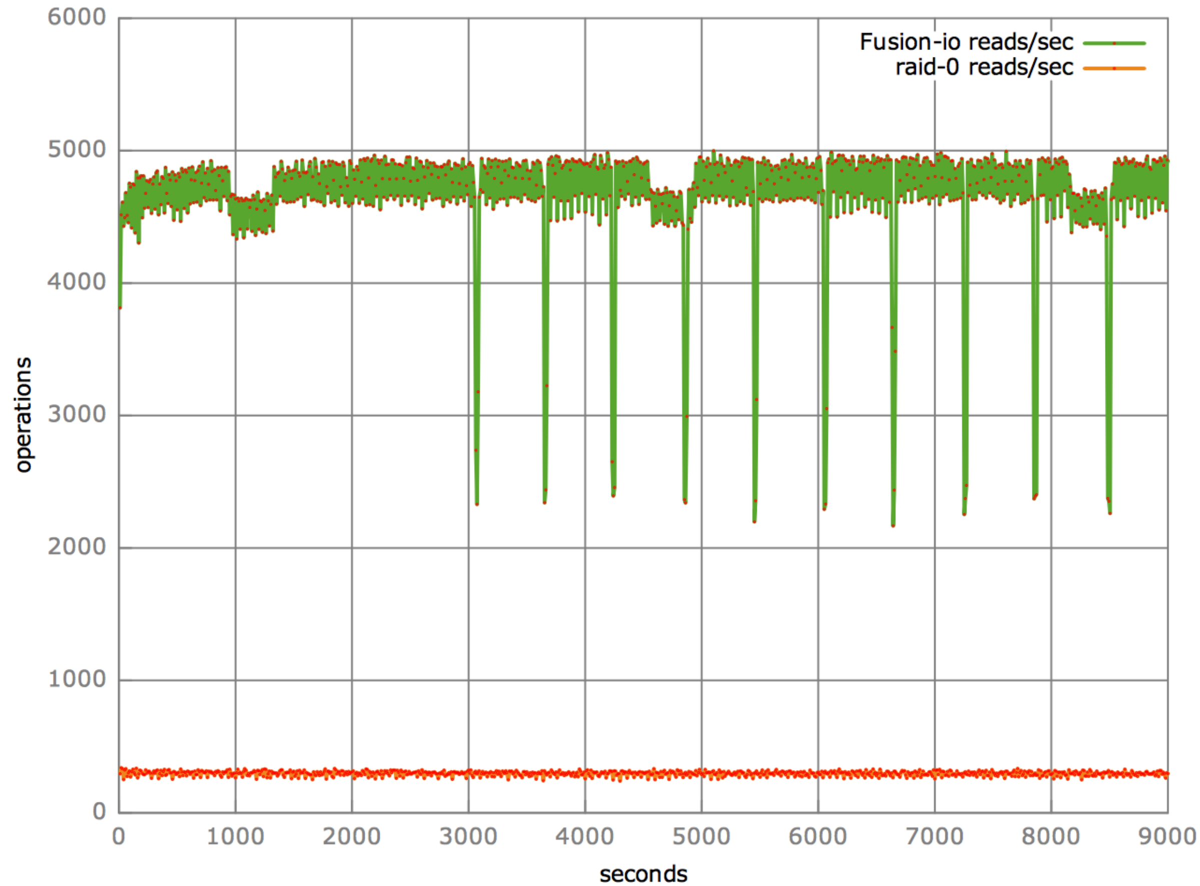
Workload B - 95/5% read/write mix



Workload C - 100% read

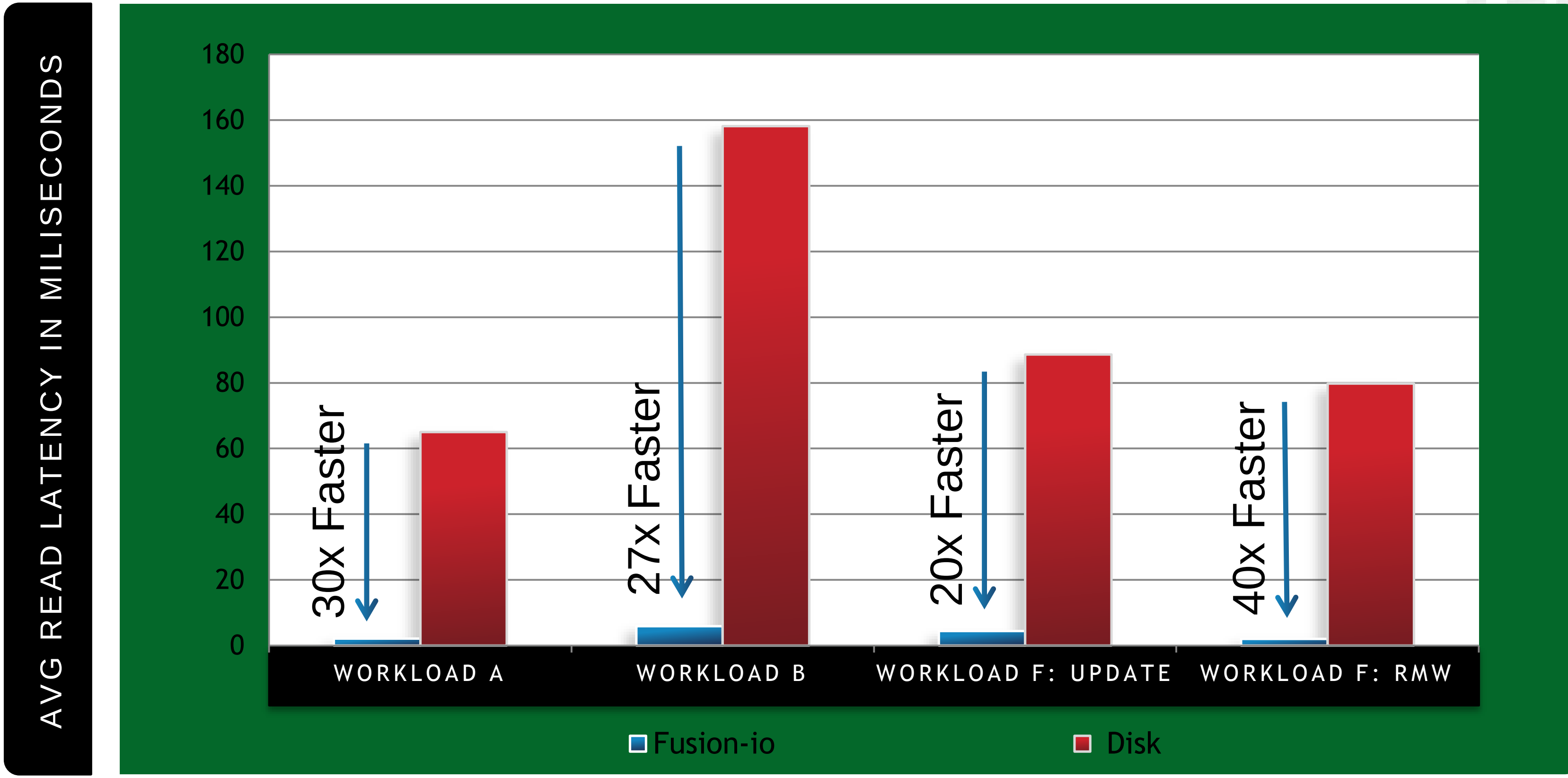


Workload F - 50/50% read/read+modify+write mix

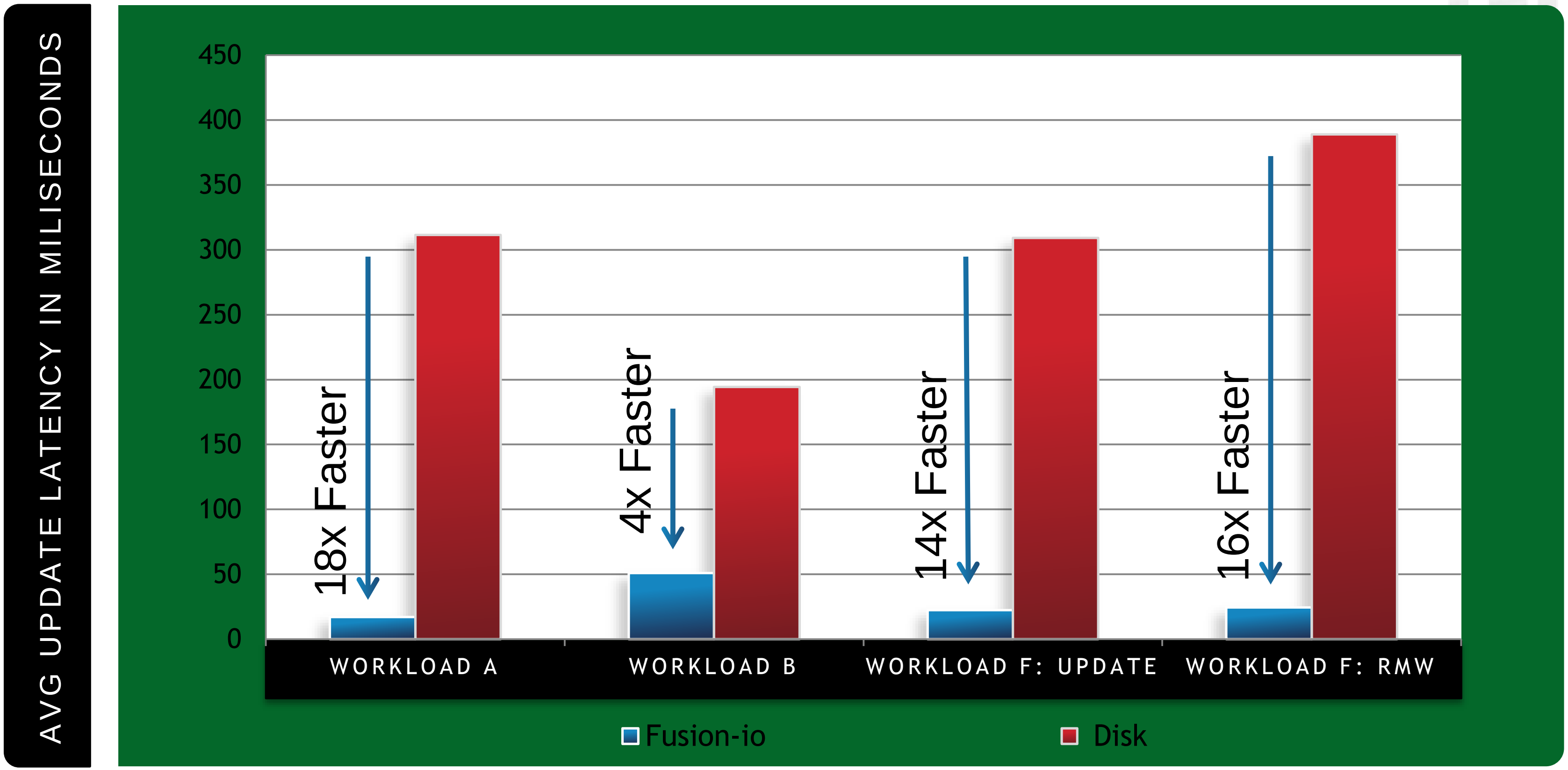


YCSB: Lowest read latency wins

FUSION-IO



YCSB: lowest Update latency wins



Architectural impact of 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 Amsterdam 2013

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
Advancing storage & information technology

A Message from SNIA Technical Council

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

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**

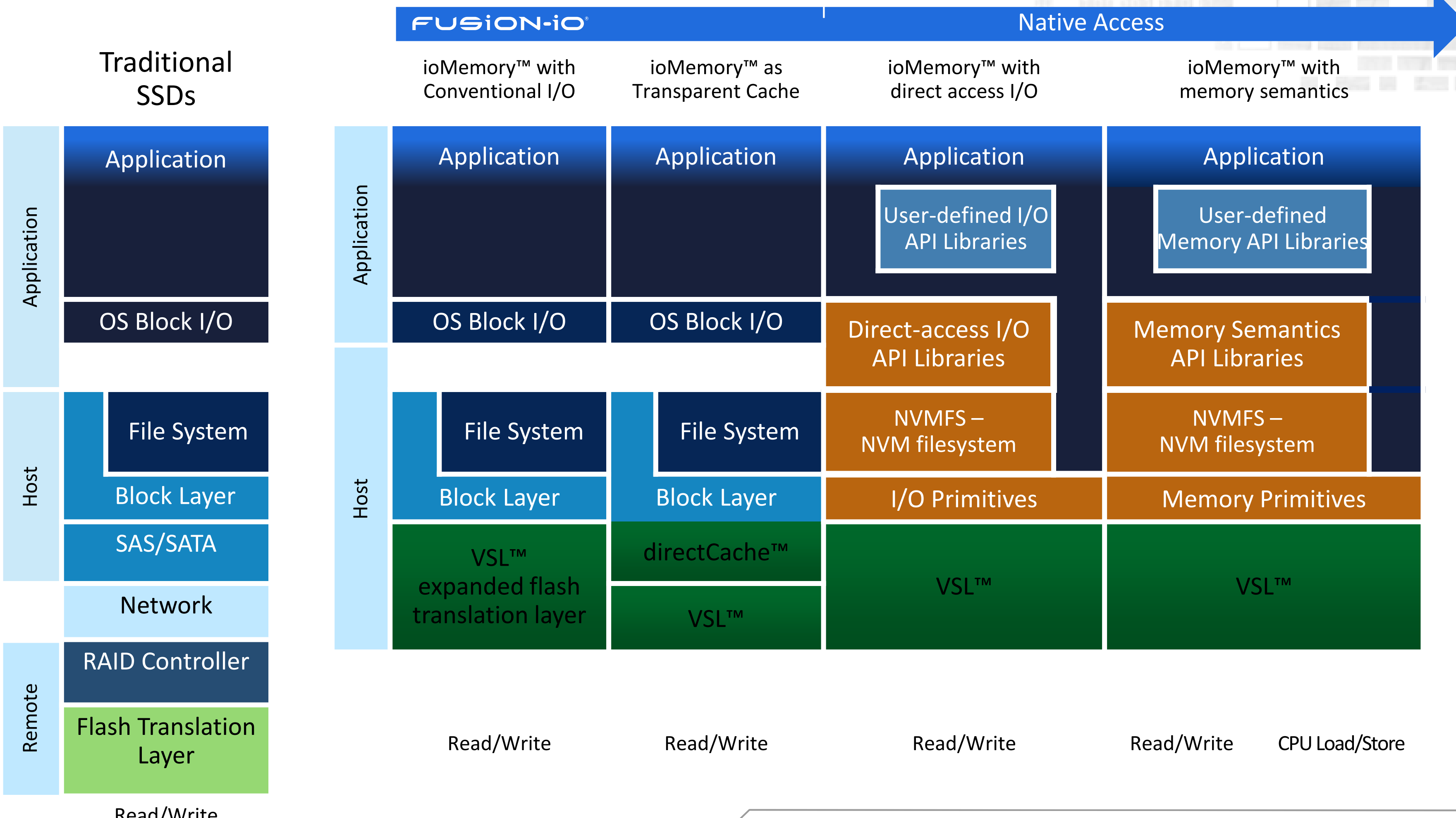
Posted in Blocks and Files, 20th April 2012 07:29 GMT

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



Comprehensive Customer Success

FUSION-IO

FINANCIALS	WEB	TECHNOLOGY	RETAIL	MANUFACTURING/ GOVERNMENT
NYSE Euronext, Starkey Mortgage, Callcredit Information Group, First Rate, TransUnion	facebook, LOCKERZ, HOST EUROPE, myspace, craigslist, NING, Kontera, HOST EUROPE, dwango, mixi, skyscanner	salesforce.com, V3, SevOne, PIPKINS, CLOUDMARK, PRIME FOCUS, MCX, ak, FRONTSAFE, PANDORA, BrainPad	alibris, Rakuten, Zappos.com, wine.com, XBOX LIVE, CyberAgent, BETONSOFT, DRECOM, Jointown	LLOG, Lawrence Livermore National Laboratory, Revere Electric Supply Co., Syntricity, Esbjerg Kommune
CREDIT SUISSE	Answers.com — The world's leading Q&A site	datalogix	EQUINOX	tabulex
5x FASTER DATA ANALYSIS	30x FASTER DATABASE REPLICATION	40x FASTER DATA WAREHOUSE QUERIES	15x QUERY PROCESSING THROUGHPUT	15x FASTER QUERIES

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

THANK YOU!

fusionio.com | REDEFINE WHAT’S POSSIBLE