GPU TECHNOLOGY CONFERENCE
MARCH 24-27, 2014 | SAN JOSE, CA

The Brightest People
The Best Ideas
The Biggest Opportunities

FUSION-IO®

Robert Wipfel
David Atkisson
Vince Brisebois

RDMA GPU Direct for ioMemory

GPU Technology Conference: S4265

Abstract

- Learn how to eliminate I/O bottlenecks by integrating Fusion-io ioMemory Flash storage into your GPU apps

1 Technical overview of Fusion-io's PCIe attached ioMemory

2 Best practices and tuning for GPU applications using ioMemory based flash storage

- Topics will cover threading, pipe-lining, and data path acceleration via RDMA GPU Direct
- Sample code showing integration between RDMA GPU Direct and Fusion-io ioMemory will be presented

Outline

- Fusion-io ioMemory
- CUDA
  - Programming patterns
- GPU Direct
- RDMA GPU Direct for ioMemory
  - API
  - Sample code
  - Performance
- Conclusion

FUSION-IO®

ioMemory

Fusion-io Introduction

Fusion-io delivers the world's data faster. From workstations to datacenter servers, Fusion-io accelerates innovative leaders like:

and over 4000 other companies worldwide.

ioMemory Accelerates Applications

Relational Databases	Server Virtualization	VDI	Big Data	Search	Analytics
ORACLE, MySQL, SQL Server, IBM DB2 INFORMIX, SAP, SYBASE, INGRES, PostgreSQL	vmware, Windows Server Hyper-V	vmware, CITRIX	APACHE HBASE, Cassandra, mongoDB	fast, Autonomy, Lucene, ORACLE Text	AccessData, MarkLogic, LexisNexis

HPC	Messaging	Workstation	Collaboration	Caching	Security/Logging	Web
FLUENT, MagmaSoft, NX NASTRAN, lustre, IBM GPFS	IBM MQ, TIBCO The Power of Now, software AG	AUTODESK, Schlumberger, Adobe	Microsoft Exchange, Microsoft SharePoint 2010, IBM Lotus	Powered by Squid, VARNISH SOFTWARE	ArcSight, splunk>	LAMP, Microsoft .NET

Flash Results Are Dramatic

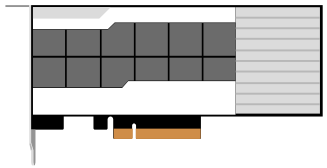
FINANCIALS	WEB	TECH	RETAIL	GOV / MFG
NYSE Euronext, STARKEY MORTGAGE, Callcredit Information Group, First Rate, TransUnion	facebook, LOCKERZ, HOST EUROPE, myspace, craigslist, NING, Kontera, HOST EUROPE, dwango, mixi, skyscanner	salesforce.com, V3, SevOne, PIPKINS, CLOUDMARK, PRIME FOCUS, zendesk, MCX, 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

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

ioMemory Purpose Built Products

ENTERPRISE

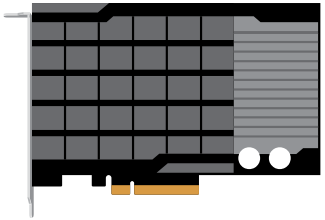
ioDrive® II



SCALE UP

HYPERSCALE

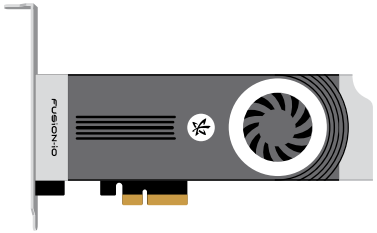
ioScale™



SCALE OUT

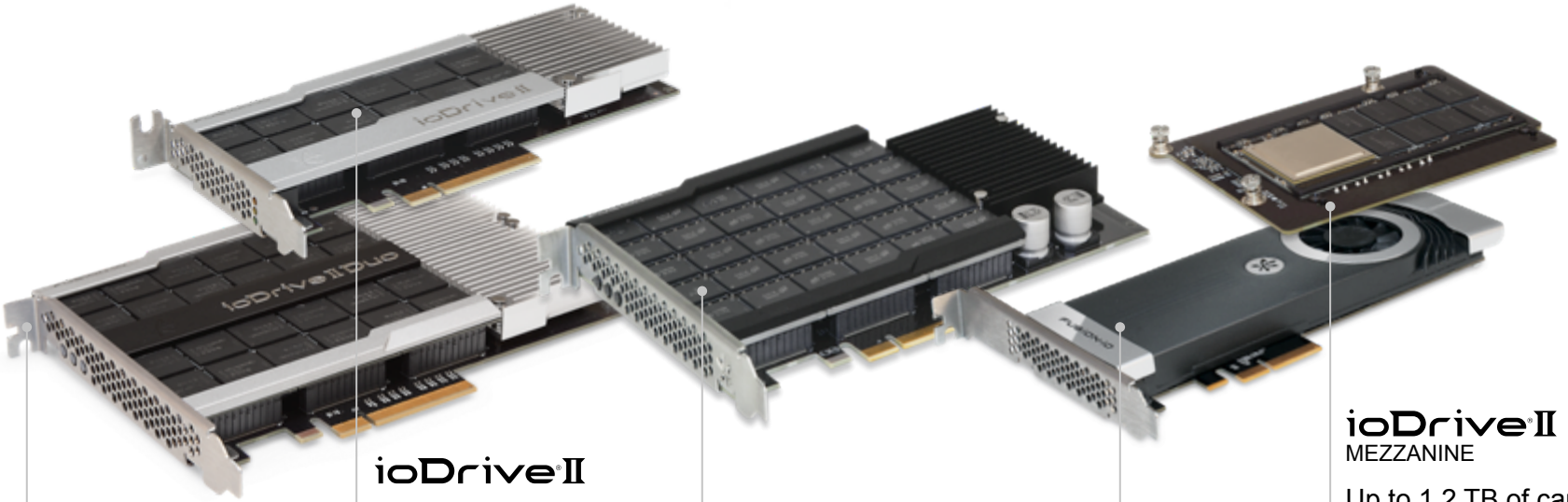
WORKSTATION

ioFX™



SINGLE USER

ioMemory®



ioDrive® II

Up to 3.0TB of capacity per PCIe slot

ioDrive® II MEZZANINE

Up to 1.2 TB of capacity per slot for maximum performance density

ioDrive® II Duo®

Up to 2.4TB of capacity per x8 PCIe slot

ioScale®

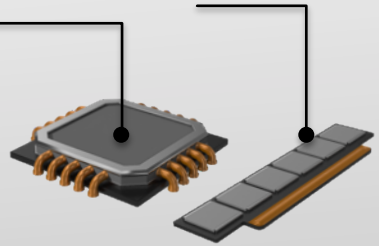
Up to 3.2TB of capacity per PCIe slot

ioFX®

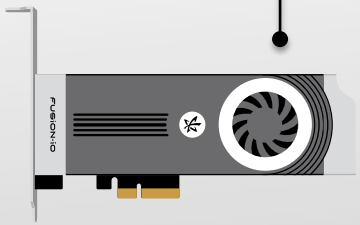
Up to 1.6TB of workstation acceleration

Access Delay Time

CPU Cache DRAM

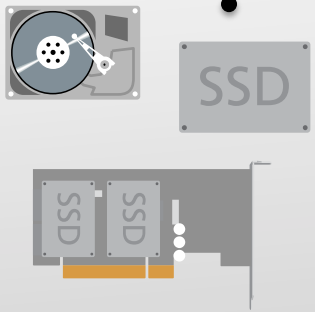


Server/Workstation Based Flash



- Significantly faster than disk
- More scalable than memory
- Less expensive than memory
- Data is persistent like on disk

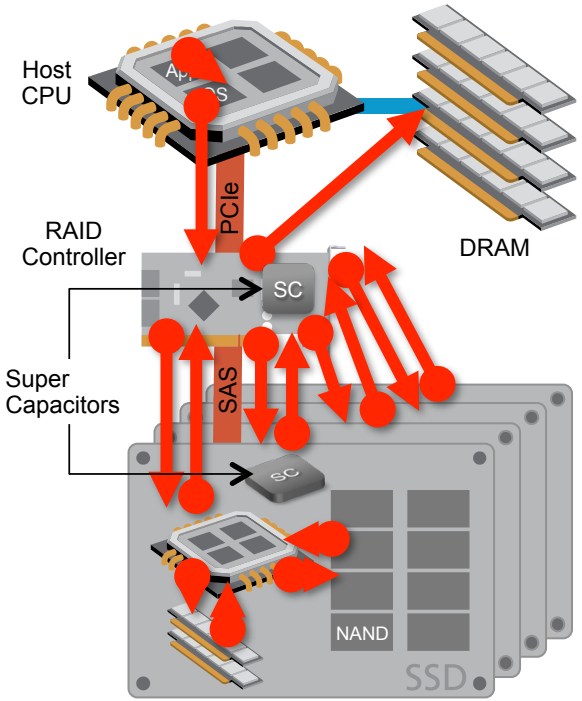
Disk Drives & SSDs



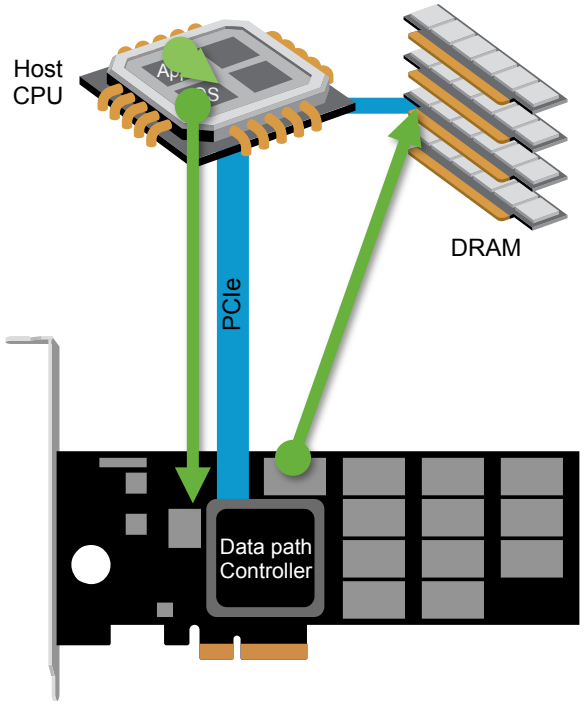
Nanoseconds 10^{-9} | Microseconds 10^{-6} | Milliseconds 10^{-3}

Cut-Through Architecture

Generic SSD Disk Approach



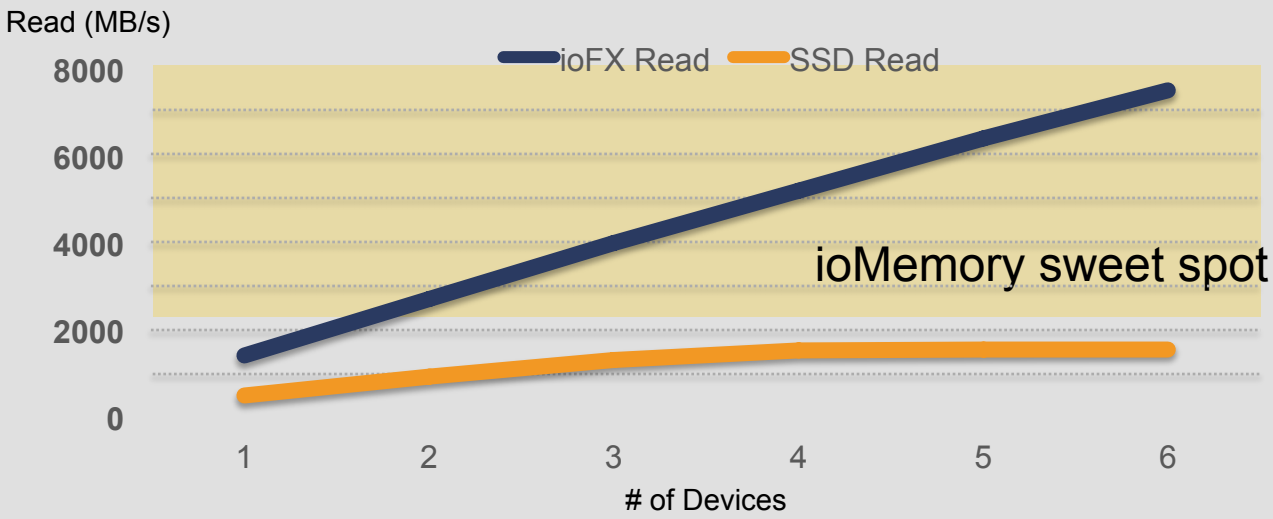
Fusion Cut-Through Approach



High Bandwidth Scalability

Bandwidth (MB/s)

ioMemory bandwidth **scales linearly** while SSDs and PCIe SSDs max out at around 2GB/s (when adding multiple devices)

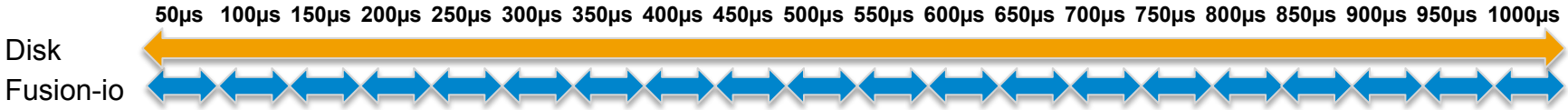


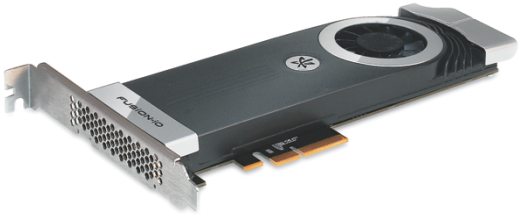
Example: Eight ioMemory devices in RAID0 provide a bandwidth of 12GB/s in a single system!

Latency

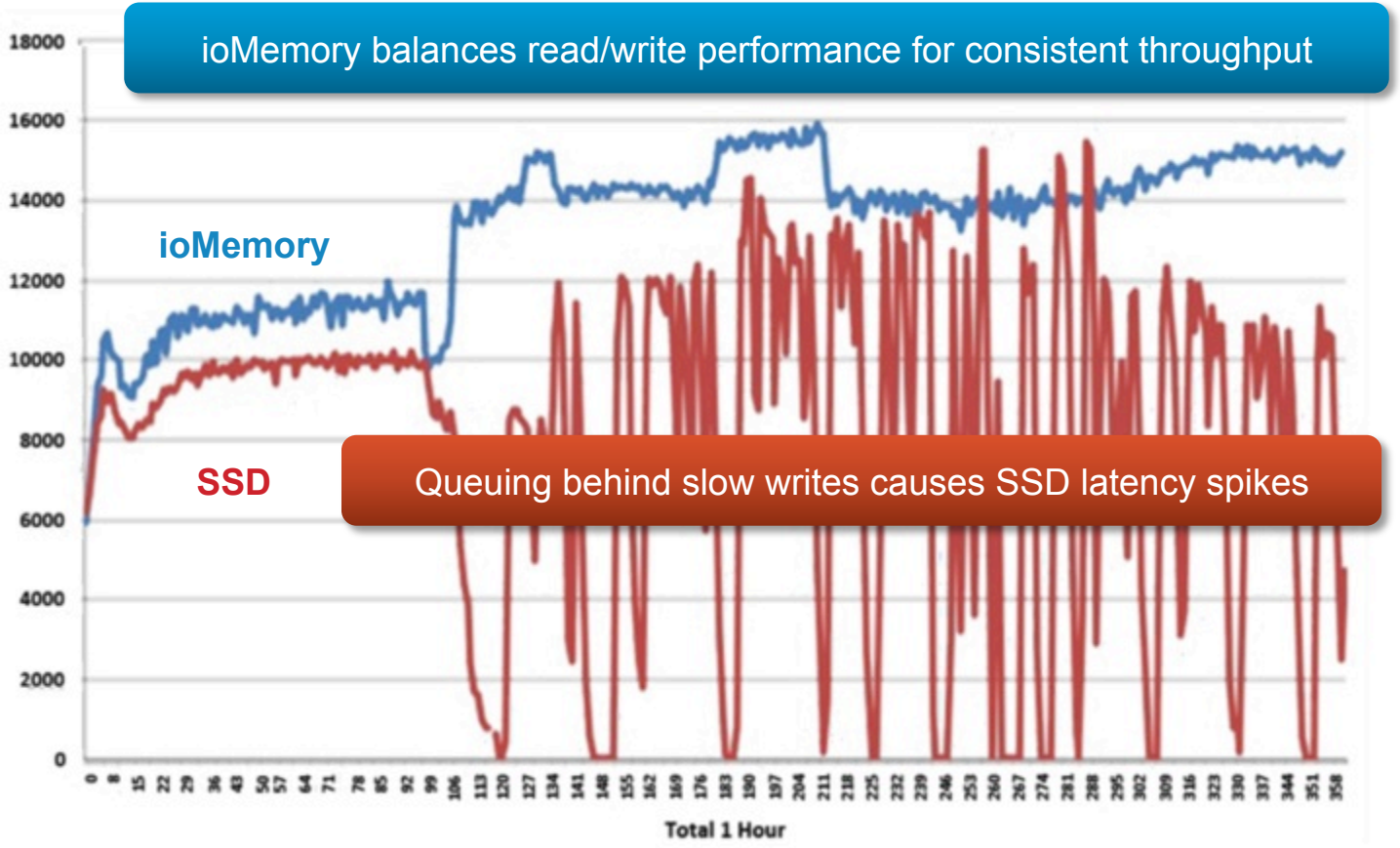
With an average response time of 50µs ioMemory will service **20** I/O's for every 1ms of disk access latency!

-OR-





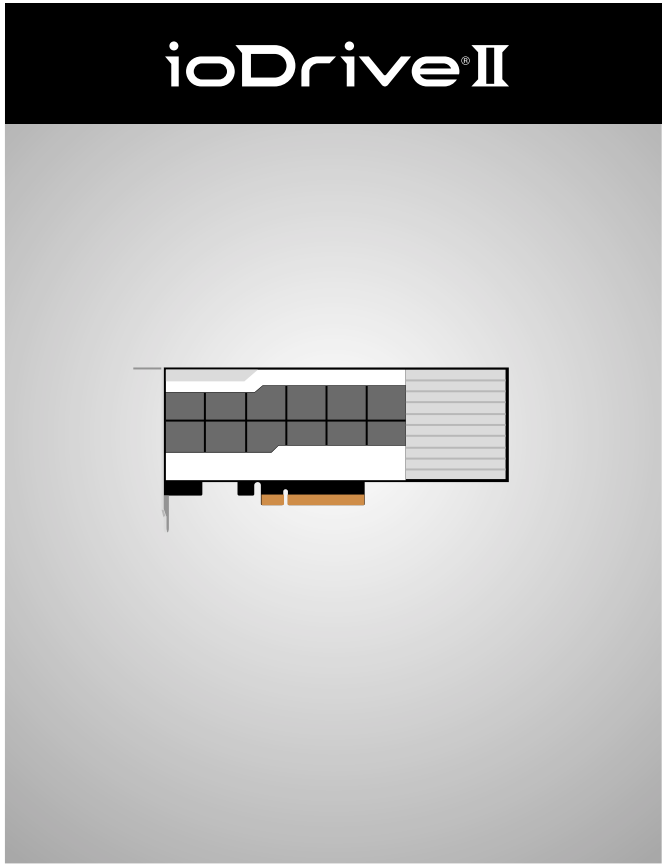
Consistent Application Performance



ioDrive2 – ioMemory for Enterprise

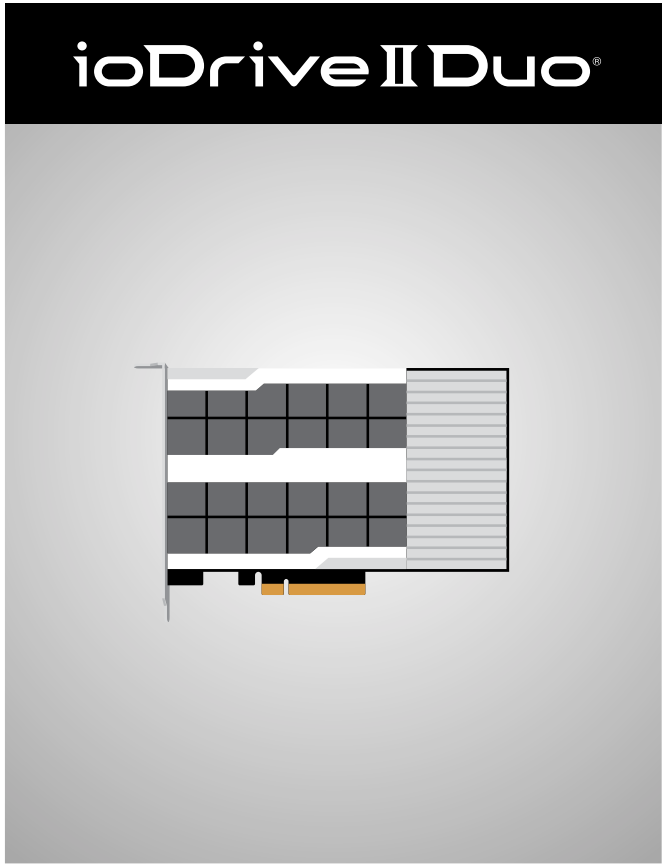
Pushing Performance Density Envelope with ioMemory Platform

- Combines VSL and ioMemory into an ioMemory platform that takes enterprise applications and databases to the next level
  - Consistent Low Latency Performance
  - Industry-leading Capacity
  - Lowest CPU Utilization Per Unit of Application Work
  - Unrivaled Endurance
  - Enterprise Reliability via Self-Healing, Wear Management, and Predictive Monitoring
  - Extensive OS Support and Flexibility



ioDrive2 Duo – Heaviest Workloads

- Based on the same ioMemory platform as the ioDrive2
- The ioDrive2 Duo provides 2.4TB of ioMemory per x4 and x8 PCI Express slot
- Organizations can:
  - Handle twice the transactions, queries and requests
  - Add capacity for future growth
  - Ensure applications seamlessly handle larger traffic spikes
  - Consolidate more infrastructure



Drivers for Hyperscale Data Centers

MARKET SEGMENTS

- Large-scale Transaction Processing
- Content Distribution On-Demand Streaming
- Big Data Management and Meta Data Indexing
- Cloud Computing and Multi-tenancy
- Business Intelligence: Data Warehousing

KEY DRIVERS

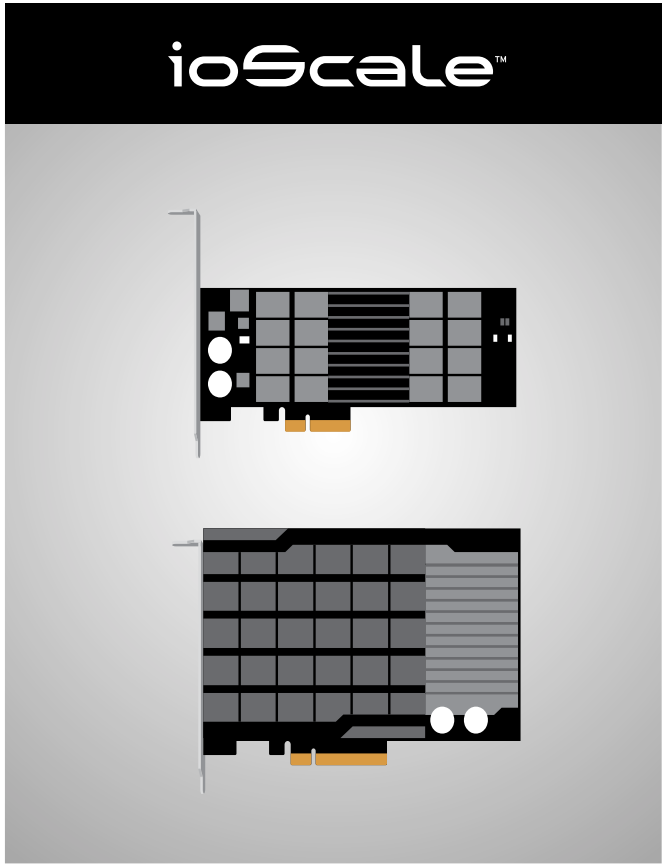
- Increase Transaction Throughput
- Lower Latency and Command Completion Times
- Lower Flash TCO

Source: HGST Flash Memory Summit Aug 2012

ioScale – ioMemory for Hyperscale

Purpose-built for the Largest Web Scale and Cloud Environments

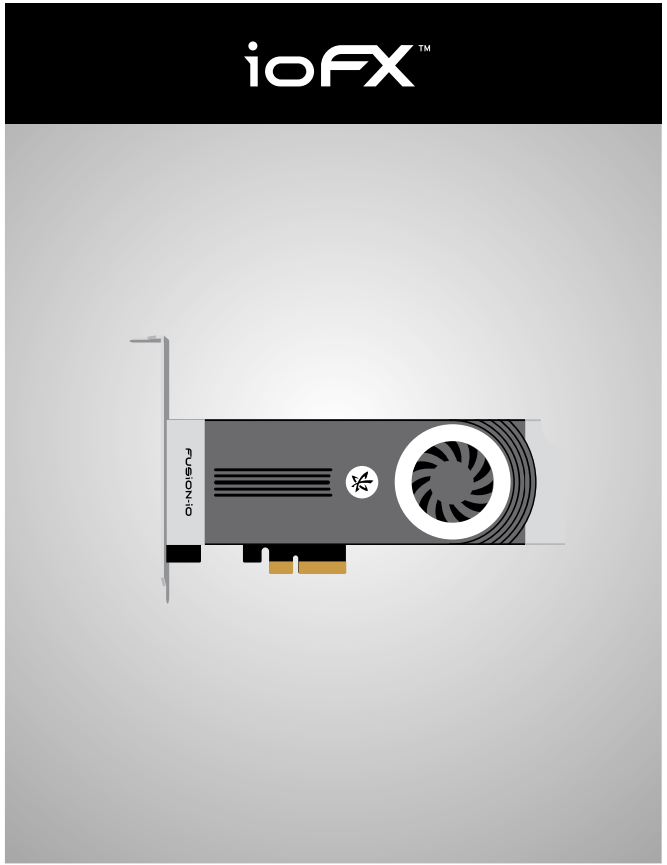
- Consistent low latency and high performance
- Reliability through Simplicity
- Unrivaled endurance
- Tools for Flash-Smart Applications
- Industry-Leading performance density
  - 410GB, 825GB, 1650GB and 3.2TB



ioFX – ioMemory for Workstations

Bridging the gap between creative potential and application performance

- Tuned for sustained performance in multithreaded applications
- Work on 2K, 4K and 5K content interactively, in full resolution
- Manipulate stereoscopic content in real-time
- Accelerate video and image editing and compositing
- Speed video playback
- Powerful throughput to maximize GPU processing
- Simplify and accelerate encoding and transcoding as well as other data-intensive activities in contemporary digital production



ioDrive2 Specifications

ioDrive2 Capacity	400GB SLC	600GB SLC	365GB MLC	785GB MLC*	1.2TB MLC*	3.0TB MLC
Read Bandwidth - 1MB	1.4 GB/s	1.5 GB/s	910 MB/s	1.5 GB/s	1.5 GB/s	1.5 GB/s
Write Bandwidth - 1MB	1.3 GB/s	1.3 GB/s	590 MB/s	1.1 GB/s	1.3 GB/s	1.3 GB/s
Ran. Read IOPS - 512B	360,000	365,000	137,000	270,000	275,000	143,000
Ran. Write IOPS - 512B	800,000	800,000	535,000	800,000	800,000	535,000
Ran. Read IOPS - 4K	270,000	290,000	110,000	215,000	245,000	136,000
Ran. Write IOPS - 4K	270,000	270,000	140,000	230,000	250,000	242,000
Read Access Latency	47µs	47µs	68µs	68µs	68µs	68µs
Write Access Latency	15µs	15µs	15µs	15µs	15µs	15µs
Bus Interface	PCI Express 2.0					
Weight	6.6 ounces					9.5 ounces
Form Factor	HH x HL					FH x HL
Warranty	5 years or maximum endurance used					

Supported Operating Systems	
Microsoft Windows	64-bit Windows Server 2012, Windows Server 2008 R2, Windows Server 2008, Windows Server 2003
Linux	RHEL 5/6; SLES 10/11; OEL 5/6; CentOS 5/6; Debian Squeeze; Fedora 16/17; openSUSE 12; Ubuntu 10/11/12
UNIX	Solaris 10/11 x64; OpenSolaris 2009.06 x64; OSX 10.6/10.7/10.8
Hypervisors	VMware ESX 4.0/4.1/ESXi 4.1/5.0/5.1, Windows 2008 R2 with Hyper-V, Hyper-V Server 2008 R2

www.fusionio.com August 2013

ioDrive2 Duo Specifications

ioDrive2 Duo Capacity	2.4TB MLC*	1.2TB SLC*
Read Bandwidth (1 MB)	3.0 GB/s	3.0 GB/s
Write Bandwidth (1 MB)	2.5 GB/s	2.5 GB/s
Ran. Read IOPS (512B)	540,000	700,000
Ran. Write IOPS (512B)	1,100,000	1,100,000
Ran. Read IOPS (4K)	480,000	580,000
Ran. Write IOPS (4K)	490,000	535,000
Read Access Latency	68µs	47µs
Write Access Latency	15µs	15µs
Bus Interface	PCI Express 2.0 x8 electrical x8 physical	
Weight	Less than 11 ounces	
Form Factor	Full-height, half-length	
Warranty	5 years or maximum endurance used	

Supported Operating Systems	
Microsoft Windows	64-bit Windows Server 2012, Windows Server 2008 R2, Windows Server 2008, Windows Server 2003
Linux	RHEL 5/6; SLES 10/11; OEL 5/6; CentOS 5/6; Debian Squeeze; Fedora 16/17; openSUSE 12; Ubuntu 10/11/12
UNIX	Solaris 10/11 x64; OpenSolaris 2009.06 x64; OSX 10.6/10.7/10.8
Hypervisors	VMware ESX 4.0/4.1/ESXi 4.1/5.0/5.1, Windows 2008 R2 with Hyper-V, Hyper-V Server 2008 R2

www.fusionio.com August 2013

ioScale Specifications

	ioScale 410GB	ioScale 825GB	ioScale 1650GB	ioScale 3.2TB
Form Factor	HHxHL	HHxHL	HHxHL	FHxHL
Read Bandwidth (1MB)	1.4 GB/s	1.4 GB/s	1.4 GB/s	1.5 GB/s
Write Bandwidth (1MB)	700 MB/s	1.1 GB/s	1.1 GB/s	1.3 GB/s
Random Read IOPS (512 bytes)	78K	135K	144K	115K
Random Write IOPS (512 bytes)	535K	535K	535K	535K
Read Access Latency	77µs	77µs	77µs	92µs
Write Access Latency	19µs	19µs	19µs	19µs
Power Profile	< 25W	< 25W	< 25W	< 25W
PBW (typical)	2	4	8	20

Supported Operating Systems	
Microsoft Windows	64-bit Windows Server 2012, Windows Server 2008 R2, Windows Server 2008, Windows Server 2003
Linux	RHEL 5/6; SLES 10/11; OEL 5/6; CentOS 5/6; Debian Squeeze; Fedora 16/17; openSUSE 12; Ubuntu 10/11/12
UNIX	Solaris 10/11 x64; OpenSolaris 2009.06 x64; OSX 10.6/10.7/10.8
Hypervisors	VMware ESX 4.0/4.1/ESXi 4.1/5.0/5.1, Windows 2008 R2 with Hyper-V, Hyper-V Server 2008 R2

ioFX Specifications

ioFX Capacity	420GB	1650GB
NAND Type	MLC (Multi Level Cell)	
Read Bandwidth (1MB)	1.4 GB/s	1.4GB/s
Write Bandwidth (1MB)	700MB/s	1.1GB/s
Read Access Latency (4K)	77µs	77µs
Write Access Latency (4K)	19µs	19µs
Bus Interface	PCI Express 2.0 x4 (x8 physical)	PCI Express 2.0 x4 (x4 physical)
Bundled Mgmt Software	ioSphere	
Operating Systems	***Microsoft Windows***: 64-bit Windows Server 2012, Windows Server 2008 R2, Windows 8, Windows 7 ***Linux:*** RHEL 5/6, SLES 10/11, CentOS 5/6, Debian Squeeze, Fedora 16/17, Ubuntu 10/11/12, OELv5/6, OpenSUSE 12 ***OSX:*** MAC OSX 10.6.7/10.7/10.8 ***Hypervisors:*** VMware ESX 4.0/4.1, VMware ESXi 4.1/5.0/5.1, Windows 2008 R2 with Hyper-V, Hyper-V Server 2008 R2	
Warranty	3 years or maximum endurance used	

www.fusionio.com August 2013

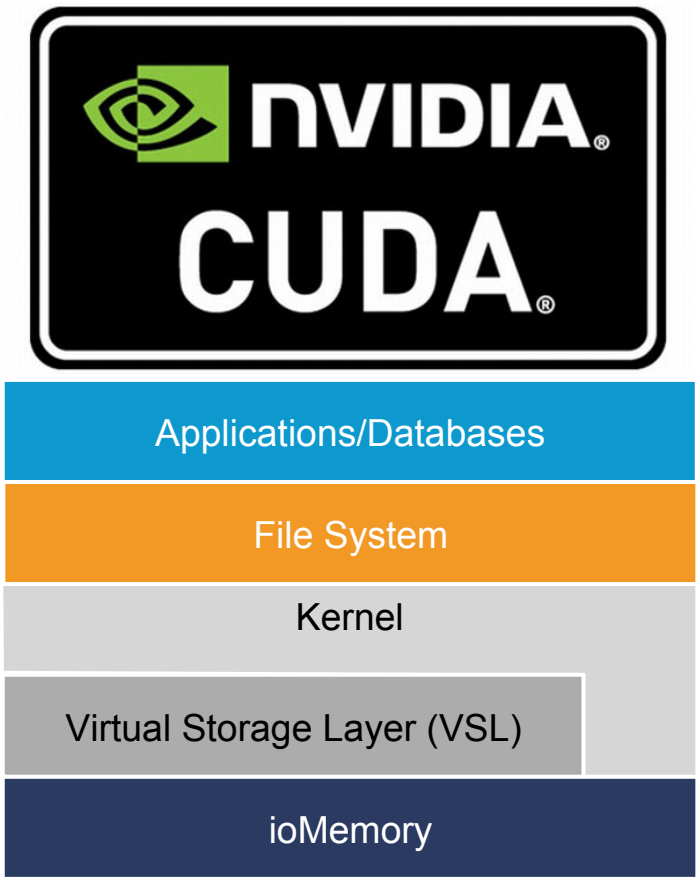
Datasheets

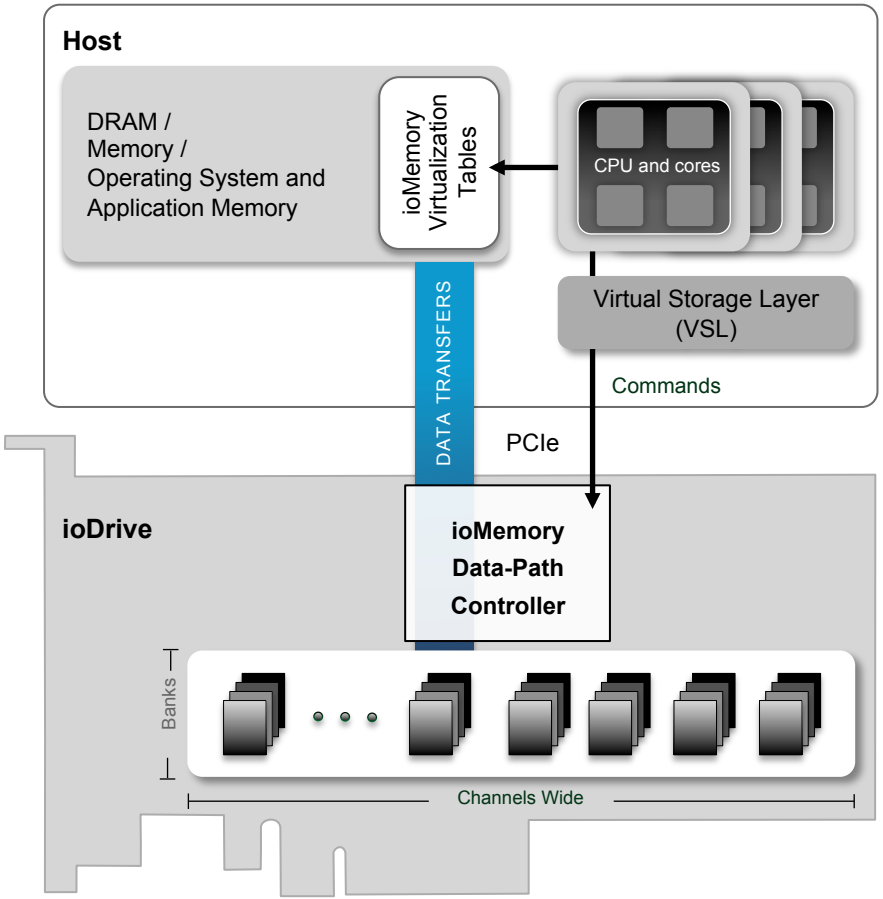
http://www.fusionio.com/

FUSION-IO®

CUDA

CUDA and the ioMemory VSL





ioMemory -> GPU

- OS-pinned CUDA buffer

```
// Alloc OS-pinned memory
cudaHostAlloc((void**)&h_odata, memSize,
              (wc) ? cudaHostAllocWriteCombined : 0);
```

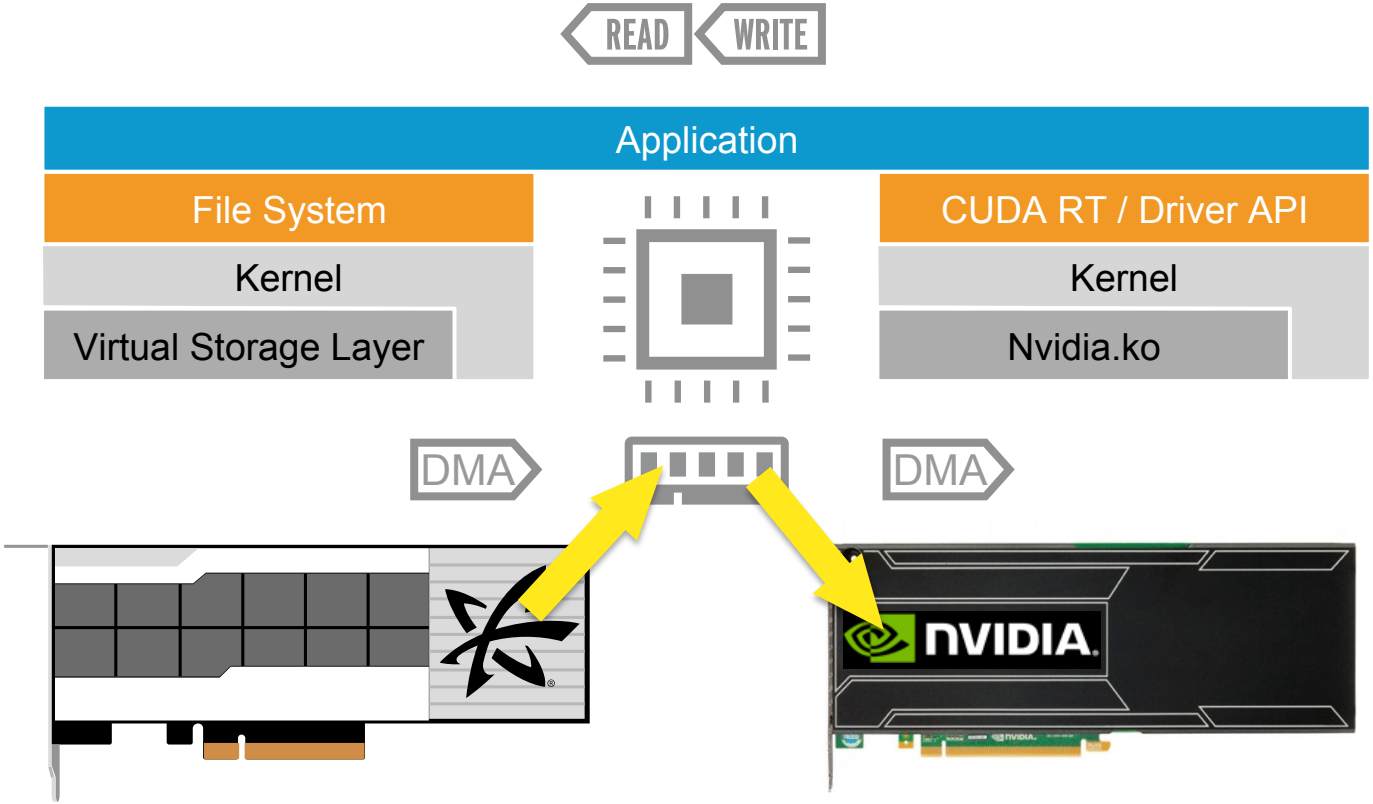
- Read from ioMemory

```
fd = open("/mnt/ioMemoryFile", O_RDWR | O_DIRECT);
rc = read(fd, h_odata, memSize);
```

- Copy (DMA) to GPU

```
cudaMemcpyAsync(d_idata, h_odata, memSize,
                cudaMemcpyHostToDevice, stream);
```

ioMemory -> GPU



GPU -> ioMemory

- OS-pinned CUDA buffer

```
// Alloc OS-pinned memory
cudaHostAlloc((void**)&h_odata, memSize,
              (wc) ? cudaHostAllocWriteCombined : 0);
```

- Copy (DMA) from GPU

```
cudaMemcpyAsync(h_odata, d_idata, memSize,
                cudaMemcpyDeviceToHost, stream);
```

- Write to ioMemory

```
fd = open("/mnt/ioMemoryFile", O_RDWR | O_DIRECT);
rc = write(fd, h_odata, memSize);
```

Programming Patterns

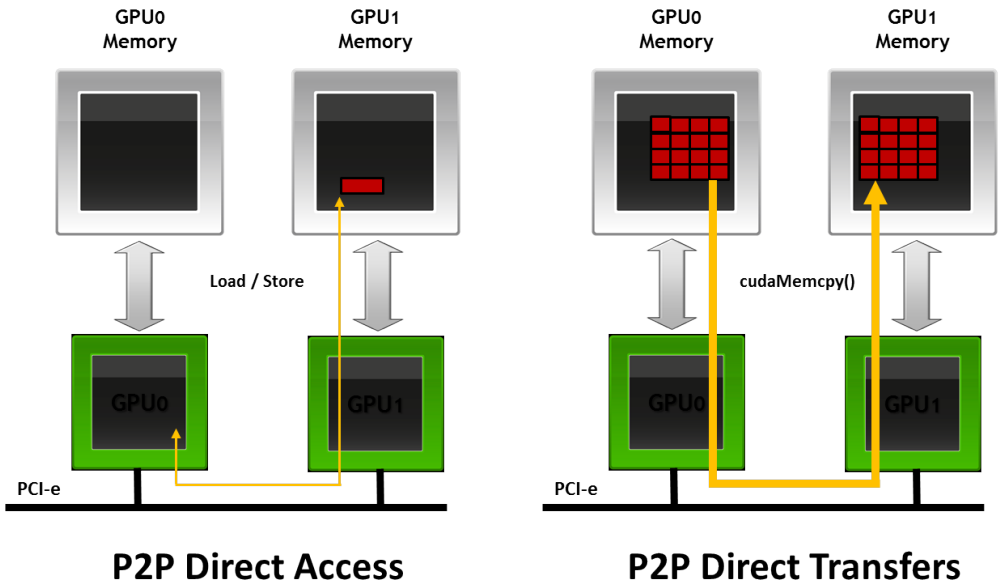
- Pipelines
  - CPU threads
  - CUDA streams
- Ring buffers
  - Parallel DMA
- Direct I/O

FUSION-IO®

GPU Direct

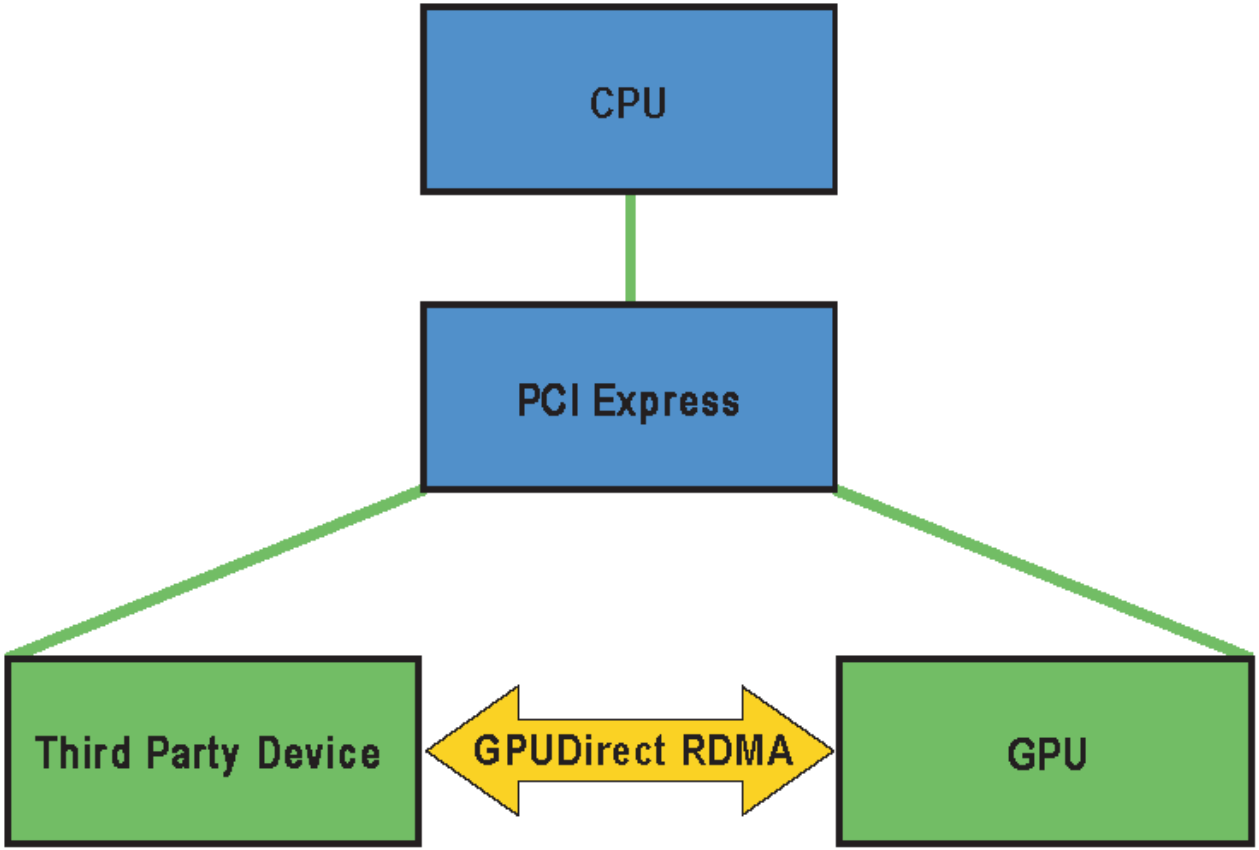
GPU Direct
https://developer.nvidia.com/gpudirect

- Copy data directly to/from CUDA pinned host memory
  - Avoid one copy but not a host memory bypass
  - Currently supported by ioDrive (as described previously)
- Peer to peer transfers
  - P2P DMA
    - between PCIe GPUs
- Peer to peer memory access between GPUs
  - NUMA
    - between CUDA kernels



RDMA GPU Direct

http://docs.nvidia.com/cuda/gpudirect-rdma/index.html



RDMA GPU Direct

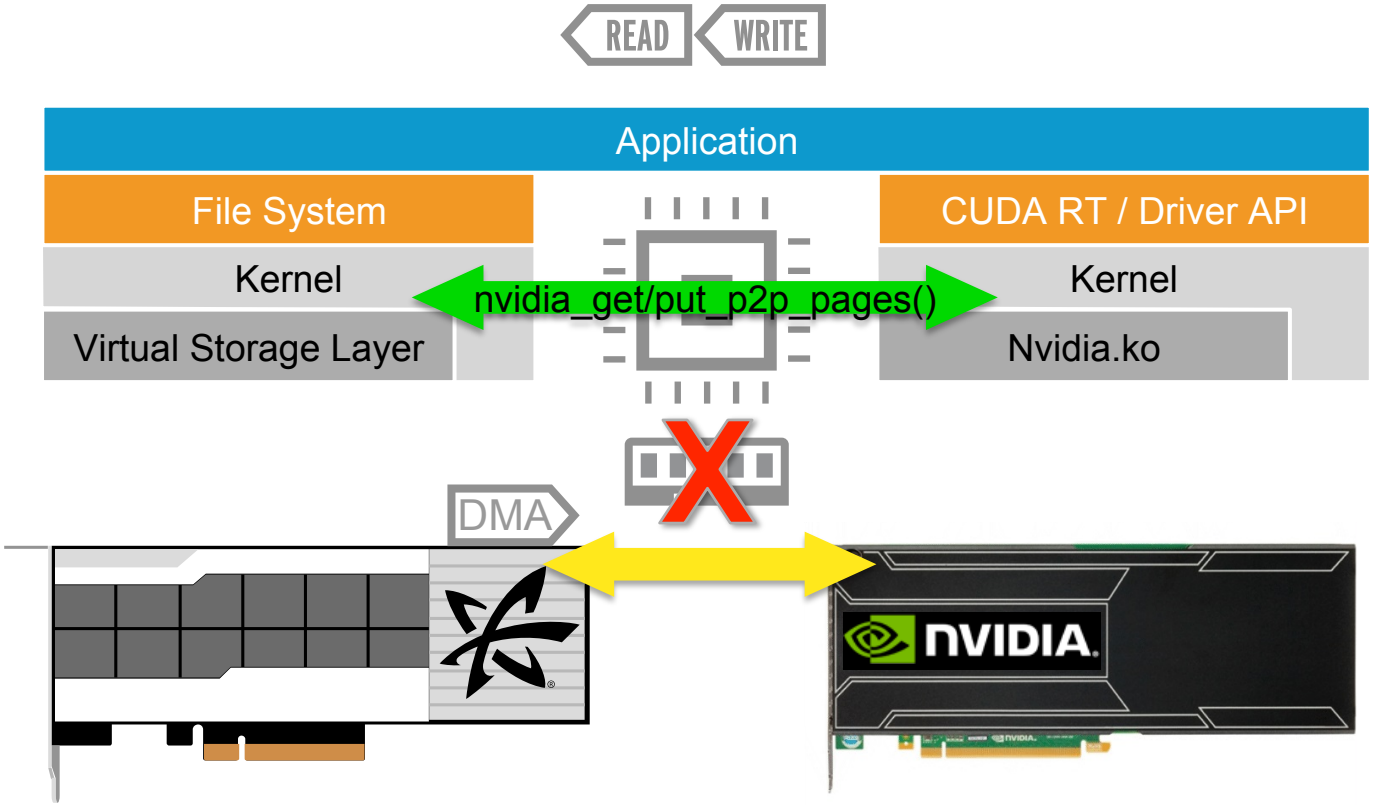
- Available in Kepler-class GPUs and CUDA 5.0+
- Fastest possible inter-device communication
  - Utilizes PCIe peer-to-peer DMA transfer
- Avoids host memory altogether
  - No bounce buffer nor copy-in/copy-out
- Third party device can directly access GPU memory
  - Tech preview (proof of concept) for ioMemory

TECH PREVIEW

FUSION-IO®

RDMA GPU Direct for ioMemory

ioMemory <-> GPU



Design Considerations

- Supported systems
  - Kepler-class GPUs
  - Intel Sandy/Ivy Bridge
- User space tokens
  - Query CUDA and pass to VSL
- Lazy unpinning
- Kernel driver linkage
  - VSL uses nvidia_get/put_pages() APIs
  - Adopt the Mellanox GPU Direct approach

CUDA API

- Data structure

```
typedef struct CUDA_POINTER_ATTRIBUTE_P2P_TOKENS_st {
  unsigned long long p2pToken;
  unsigned int vaSpaceToken;
} CUDA_POINTER_ATTRIBUTE_P2P_TOKENS;
```

- Function

```
CUresult CUDAAPI cuPointerGetAttribute(void *data,
  CUpointer_attribute attribute, CUdeviceptr pointer);
```

- Usage

```
ptr = cudaMalloc();
cuPointerGetAttribute(&tokens,
  CU_POINTER_ATTRIBUTE_P2P_TOKENS, ptr);
```

VSL user space API

```
typedef struct fusion_iovec {
    fio_block_range_t iov_range; ///< Range of sectors
    uint32_t  iov_op;            ///< FIOV_TRIM/WRITE/READ
    uint32_t  iov_flags;         ///< FIOV_FLAGS_* (GPU)
    uint64_t  iov_base;          ///< Mem addr (CPU or GPU)
                  ///< Special tokens for p2p GPU transfers
    unsigned long long iov_p2p_token;
    unsigned int iov_va_space_token;
    unsigned int iov_padding_reserved;
} fusion_iovec_t;

int ufct_iovec_transfer(int fd,
       const struct fusion_iovec *iov,
       int iovcnt, uint32_t *sectors_transferred);
```

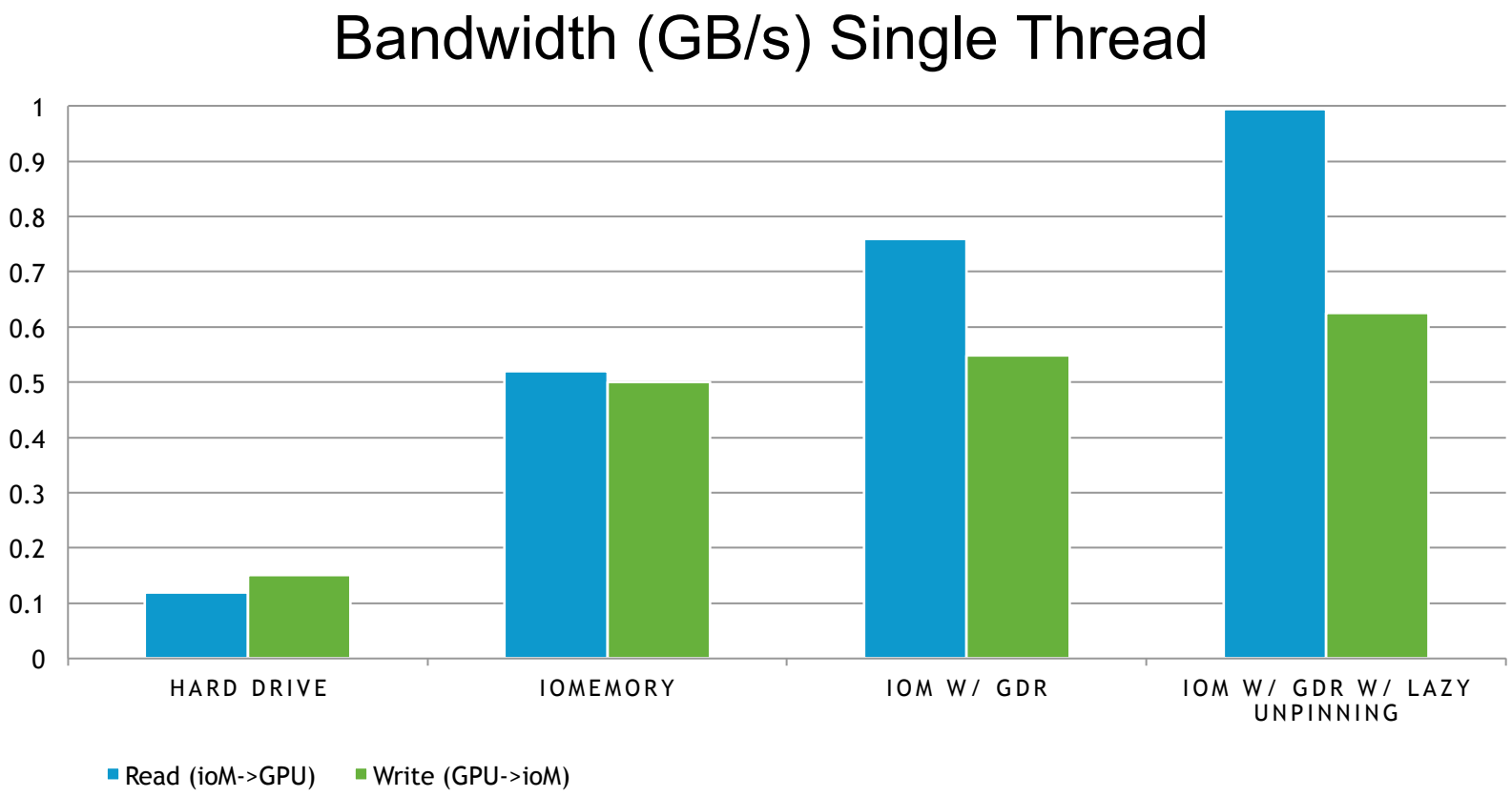
Sample Code

```
cuda_err = cudaMalloc((void**)&buf, allocate_size);
cu_err = cuPointerGetAttribute(&cuda_tokens,
                               CU_POINTER_ATTRIBUTE_P2P_TOKENS,
                               (Cudeviceptr)buf);

iov[0].iov_range.base      = sector;  // ioMemory block addr
iov[0].iov_range.length    = individual_buf_size;
iov[0].iov_op     = is_write ? FIOV_WRITE : FIOV_READ;
iov[0].iov_flags = is_gpu    ? FIOV_FLAGS_GPU : 0;
iov[0].iov_base            = buf;     // GPU memory addr
iov[0].iov_p2p_token       = cuda_tokens.p2pToken;
iov[0].iov_va_space_token = cuda_tokens.vaSpaceToken;

ufct_iovec_transfer(fd, iov, iovcnt, &sectors_transferred);
```

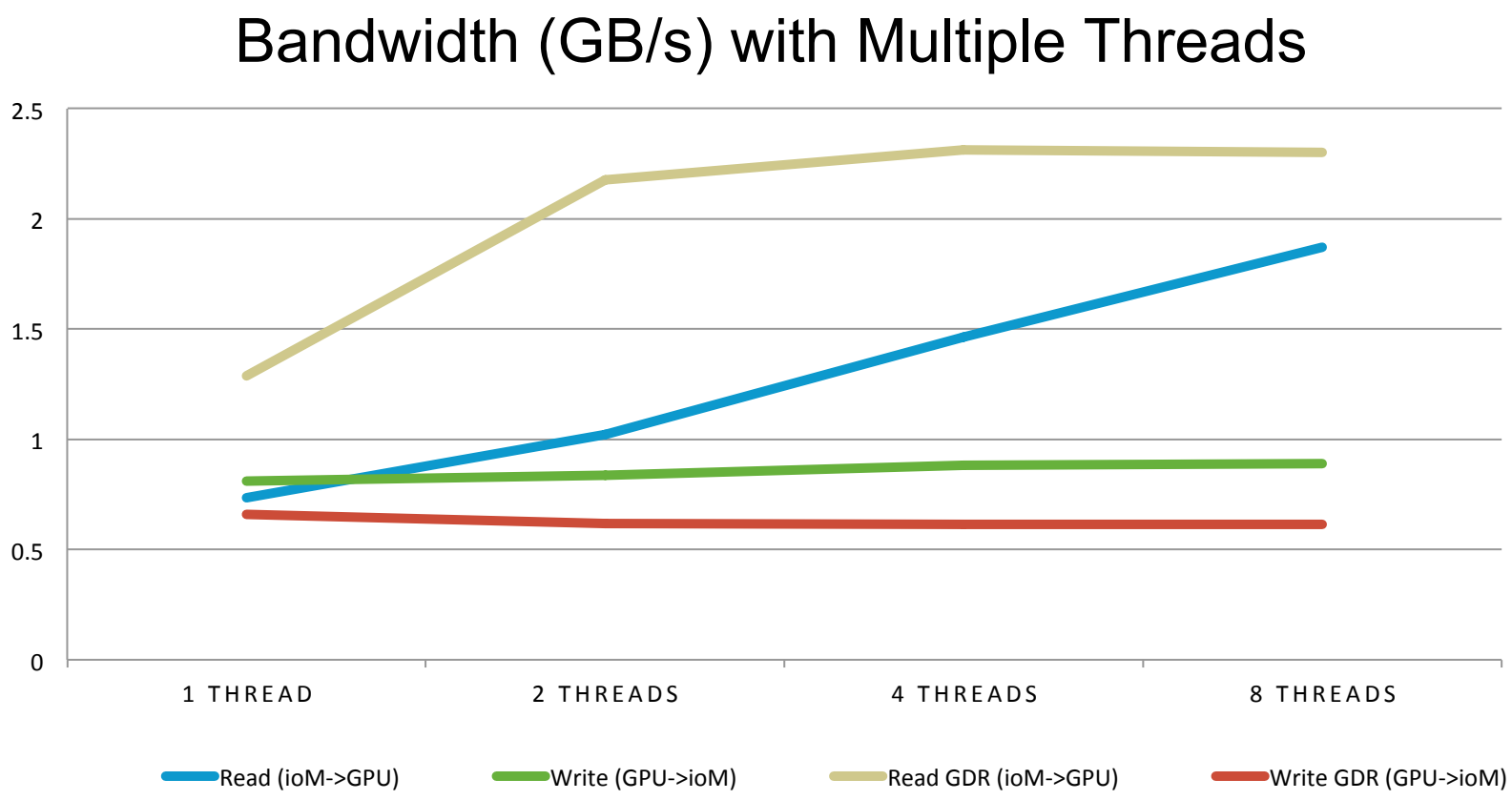
Benchmark

- System details
  - Intel Sandy Bridge
  - NVIDIA Tesla K20
- Modified CUDA bandwidth test
- Modified fio-iovec-transfer test

Modified CUDA bandwidth test



Modified fio-iovec-transfer test



FUSION-IO®

Conclusion

Conclusion

- ioMemory for high performance GPU I/O
  - Available now!
- Tech preview of RDMA GPU Direct for ioMemory
  - Bypass host memory to further increase I/O performance
- ioMemory can accelerate your GPU applications
- Invite feedback
  - vbrisebois@fusionio.com

Thank You

fusionio.com | DELIVERING THE WORLD'S DATA. FASTER.