

# Storage Is Changing Fast – Be Ready or Be Left Behind

April 16, 2008

2:00pm EDT, 11:00am PDT



Henry Newman, CTO of Instrumental Inc. and  
Enterprise Storage Forum columnist



## Housekeeping

- Submitting questions to speakers
  - Questions will be answered during 10 minute Q&A session at end of presentation
- Product names mentioned in this document may be trademarks and/or registered trademarks of their respective companies and are the property of these companies.
- Technical difficulties?



# Main Presentation

## The Future Will Be Different...

- The storage landscape is headed for dramatic change, thanks to new technologies like Fibre Channel over Ethernet (FCoE), NFS v4.1 (pNFS), object-based storage and SAS that will affect everything from storage interconnects and trickle down to RAID and the underlying devices.



## What We Will Cover

- New Interconnection technologies
- NFS v4.1 new data access method
- Future storage technologies
- Continuing storage challenges

Fibre channel over Ethernet is likely going to change the world

# NEW INTERCONNECTION TECHNOLOGIES



## Why FCOE Will Be a Disruptive Technology Shift

- Commodity technologies are driving the storage market
  - Take the case of enterprise disks (FC today) as compared to SATA
    - Estimates are a factor of 10x different in volume
- FCOE combines commodity technology Ethernet with enterprise technology FC protocol
  - The cost per Gbit/sec of Ethernet has been historically far lower than the cost of FC just given the volume
    - Who does not have Gbit Ethernet on the motherboard?

## What Is the Impact for the Enterprise?

- Today Ethernet networks are used for TCP/IP traffic and FC is used for data to RAID/disk and tape storage
  - A small percentage use Infiniband (IB) for data traffic and communication
- Having a single fabric to manage switches, NICs and cables is appealing given the cost
  - FCOE will likely require dedicated traffic to storage
- Currently 10 GbE is expensive but prices are dropping

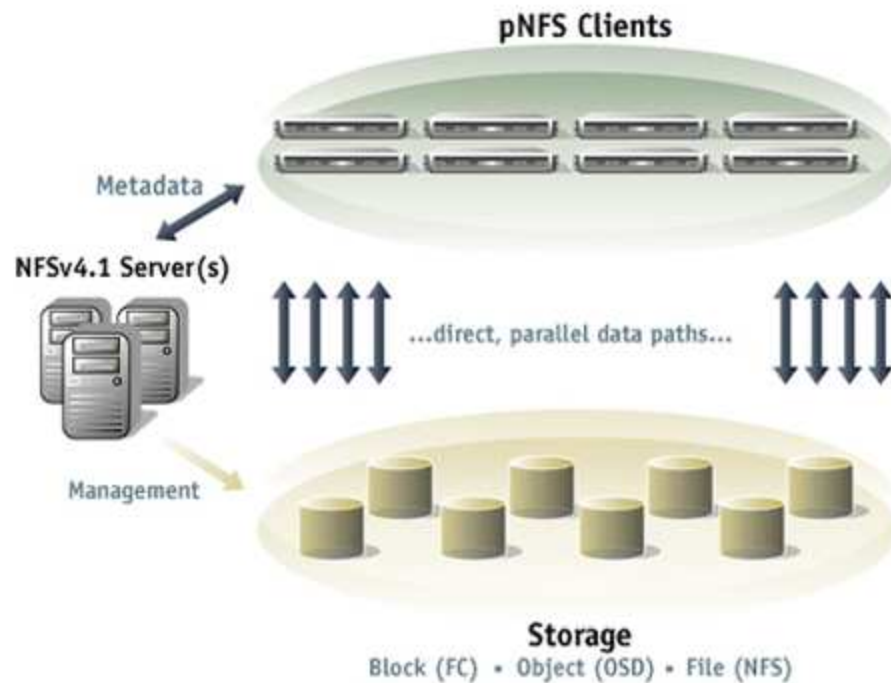
High performance data access via NFS will be a reality

# **NFS v4.1 NEW DATA ACCESS METHOD**

## Currently NFS Does Not Scale

- A draft for changes to NFS v4, called NFS v4.1, exists and includes pNFS
  - The draft can be found on the IETF web site and is called *draft-ietf-nfsv4-minorversion1-13.txt*
    - NFS v4.1 is often referred to as pNFS as that was the original name when first proposed to the IETF. Features needing to be added to the NFS v4 protocol include:
      - Sessions/RDMA
      - Directory delegations

# What Will NFS v4.1 Look Like?



Courtesy of Panasas

## Parallel access with DMA support and multiple access methods

## Who Is Working on NFS v4.1

- Three different access methods have been proposed
  - FILES: NFS/ONCRPC/TCP/IP/GE for files built on sub-files
    - NetApp, Sun, IBM, University of Michigan CITI
  - BLOCKS: SBC/FCP/FC or SBC/iSCSI for files built on blocks
    - EMC, IBM (<http://tools.ietf.org/html/draft-ietf-nfsv4-pnfs-blocks-03.txt>)
  - OBJECTS: OSD/iSCSI/TCP/IP/GE for files built on objects
    - Panasas, Sun (<http://tools.ietf.org/html/draft-ietf-nfsv4-pnfs-obj-03.txt>)

What other technology could change our world

# FUTURE STORAGE TECHNOLOGY

## T10 Object Storage Device

- OSD allocation of space, however, splits the name space and allocation between the host side and the storage which handles the allocation
- In the T10 OSD environment allocation is done by the storage controller and potentially by the disk drive
- Vendors are developing OSD hardware for the 2010 timeframe
  - OSD also provides an interface for device authentication and an encryption model. In addition, OSD has other significant advantages

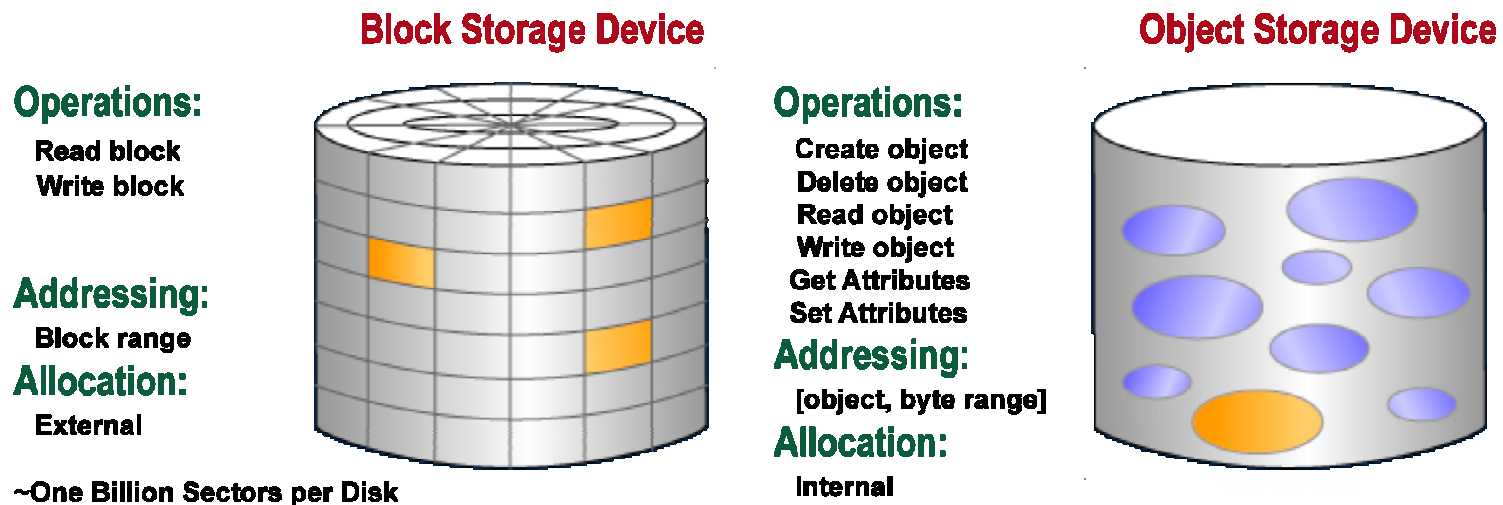
## It Will be a Disruptive Technology...

- In today's block based world, fragmentation is becoming a significant issue for long term sustained performance
  - With OSD, fragmentation is virtually eliminated
- Current RAID levels are fixed values with fixed allocations
  - The OSD object manager will be able to select the RAID level based on the object size



## What Will It Impact?

- OSD will change file and RAID controllers as well
  - The adoption of OSD is in question because it presents a chicken and egg scenario
- Without OSD drives available, most vendors will not develop an OSD file system or an OSD storage manager



Storage Needs to scale and the datapath needs to change

# CONTINUING STORAGE CHALLENGES

## Disk Drive Density Is Not Scaling

Year	Capacity (GBytes)	Overall Rate of Increase over last increase	Rate of Increase Per Year	Rate of Increase All Years
1990	0.5			
1991	1.0	2.00	2.00	2.00
1992	2.0	2.00	2.00	2.00
1994	4.0	2.00	1.00	2.00
1996	9.0	2.25	1.13	3.00
1998	18.0	2.00	1.00	4.50
1999	36.0	2.00	2.00	8.00
2000	72.0	2.00	2.00	14.40
2002	146.0	2.03	1.01	24.33
2005	300.0	2.05	0.68	40.00
2008	300.0	1.00	0.33	40.00

# Tape Drive Density Is Not Scaling

Vendor	Drive	Media	Introduced	Capacity GB	Overall Rate of Increase over last increase	Rate of Increase Per Year	Rate of Increase All Years
IBM	3490E	3480	1991	0.4			
IBM	3490E	3490E	1992	0.8	2.00	2.00	2.00
IBM	3590	3590	1995	10.0	12.80	4.27	6.40
STK	SD-3	SD-3	1995	50.0	5.00	21.33	32.00
IBM	3590E	3590E	1999	20.0	0.40	0.10	6.40
IBM	3950E	3590E	2000	40.0	2.00	2.00	11.38
STK	T9940A	9940	2000	60.0	1.50	3.00	17.07
LTO	LTO	LTO	2000	100.0	1.67	5.00	28.44
STK	T9940B	9940	2002	200.0	2.00	1.00	46.55
LTO	LTO-II	LTO	2003	200.0	1.00	1.00	42.67
IBM	3592	3592	2004	300.0	1.50	1.50	59.08
LTO	LTO-III	LTO	2005	400.0	1.33	1.33	73.14
IBM	3592J	3592J	2005	500.0	1.25	1.67	91.43
STK	T10000	Titanium	2005	500.0	1.00	1.67	91.43
IBM	TS1120	3592 JA/JW	2006	700.0	1.40	1.40	119.47
LTO	LTO-IV	LTO	2007	800.0	1.14	1.14	128.00
T10000B	T10000	Titanium	2008	1000.0	1.25	1.25	150.59



## Performance Is Not Scaling for HDD or Tape

Disk	Average performance MB/sec	Inprovment since 1990
FC/SAS	100	25
SATA	70	25

Tape	Compressed MB/sec	Uncompressed MB/sec	Compressed Inprovment since 1990	Uncompressed Inprovment since 1990
LTO-4	240	120	192	96
T10000	360	120	288	96

## Latency Scaling Is Worse than Performance

Year	Approx. Seek +latency in milliseconds	Improvement
1991	20.8	
2008 3.5 inch	6	3.5
2008 2.5 inch	5	4.2

**This trend will not change**

## Latency Tolerance for CPUs (last 35 years)

- Hardware evolution based on memory component latencies have caused dramatic shifts in design
  - Vectors started it all because of memory performance limitations
  - Multiple levels of memory (L1, L2, L3, NUMA)
  - Multi-threaded CPUs
- All of these changes are based on the need to hide latency when accessing memory as latency has increased as a function of CPU performance
  - This latency trend is not going to change

## Latency Tolerance for the Data Path (last 35 years)

- Application data path has not changed in a similar way to address latency
  - Storage latency has changed a small amount compared to CPU performance
  - File systems do not pass topology to block devices to impact latency
    - This has a major impact on RAID storage and read-ahead
    - You cannot read-ahead if you do not know where the data is

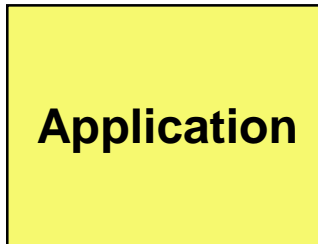
## It's All About Latency...

- Application changes such as multithreading are the same techniques used to address latency issues in the late 70s and 80s with vector based computers
  - Systems today are efficient if they hide memory latency
  - I/O is efficient today if it can hide latency by multi-threading
  - I/O can be efficient by making large I/O requests
- Latency in the data path is growing as a function of computation
  - True for memory
  - True for I/O

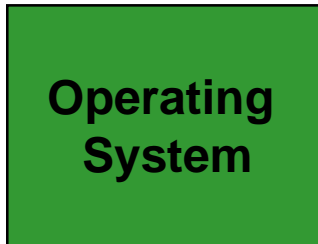
# Data Path Research Needed and Updates Required

## Current

Current POSIX system calls open/read/write/aio. Limited communication with OS layer



POSIX Atomic operations open/read/write/aio. No communication with physical layer



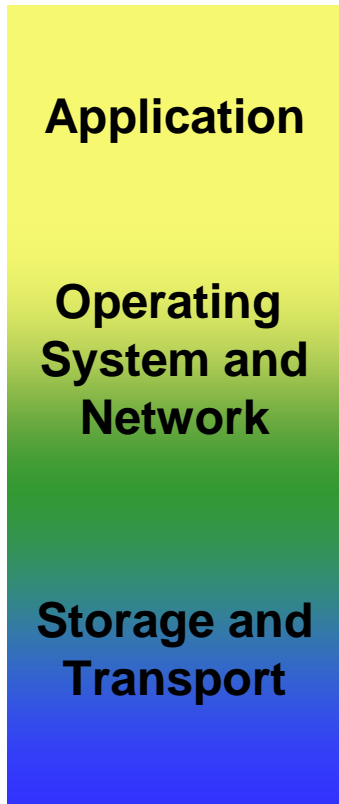
Block based storage and limitations of 30+ year old technology



**Data path today is the same as the data path 20 years ago**

## Future

Changes to support new constructs for different types of latencies for data



OSD combined with networking constructs could address different latencies

Given physical limitations of storage, optimizations must be done at a higher level to impact the technology

**Need benchmarks that focus on an end-to-end view of I/O**



internet.com<sup>®</sup>

Questions?

**Thank you for attending**

**If you have any further questions, e-mail**  
**[webcasts@jupitermedia.com](mailto:webcasts@jupitermedia.com)**

**For future internet.com Webcasts, visit**  
**[www.internet.com/webcasts](http://www.internet.com/webcasts)**

