

Changing the Fundamentals of Data Storage for SuperComputers

An Overview

Michael H. Anderson, CTO

StreamScale Inc.

manderson@streamscale.com

10/21/09



Background

- “One definition of a SuperComputer used to be a computer that changed a compute bound problem to an I/O bound problem. Because the continuing increase in processor performance of roughly 60% per year has compounded much faster than the 20% per year of I/O rates, the problem has reversed so that I/O performance can be the bottleneck in solving complex physics problems. As these trends continue, the I/O bottleneck grows ever worse without some change in the fundamentals of data storage.”
 - Tyce McLarty, Lawrence Livermore National Laboratory

The StreamScale Vision of High Performance Computing

- Imagine an HPC system where:
 - Every disk rotation and every interconnect cycle is used to transfer data to or from an application
 - Every byte of data is validated, and if required, corrected or recreated, just prior to any computation
 - Every CPU cycle and every CPU resource is fully applied to computing a result
 - Every result is stored in a verifiable, redundant fashion
- *Every second of time and every watt of power was used effectively*
- You just had your first experience with ‘StreamScaling’, a fundamental change to HPC storage

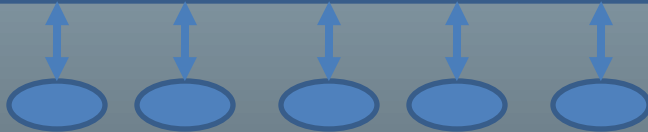
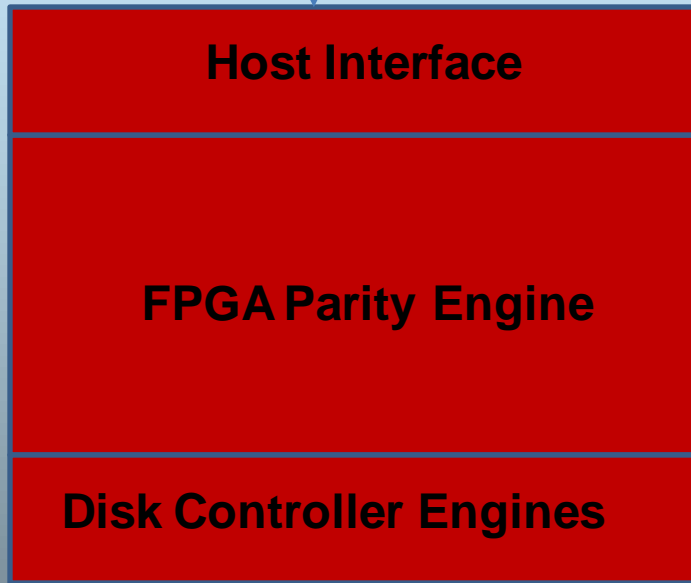
Presentation Overview

- Current fundamentals of HPC storage
- Changing the fundamentals of HPC storage
- Implementation and performance results
- Benefits of the change
- Real world examples and references
- Summary
- The Details

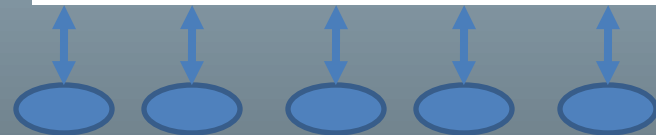
High Performance HPC Storage Architecture Comparison



Proprietary Components



Old Generation



New Generation

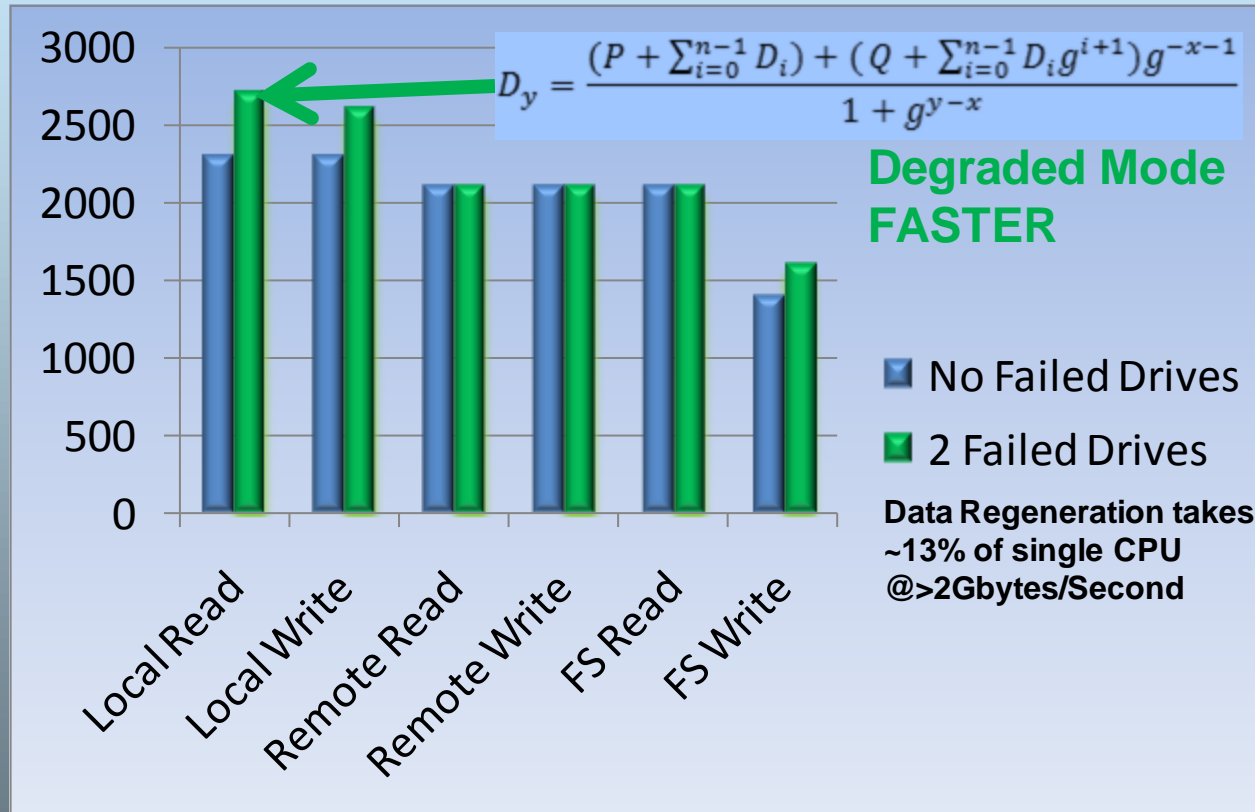
Standard Components

NUMARAID

NumaRAID Single 24 Drive Storage Node

Single IB Link, Single Stream DD Test

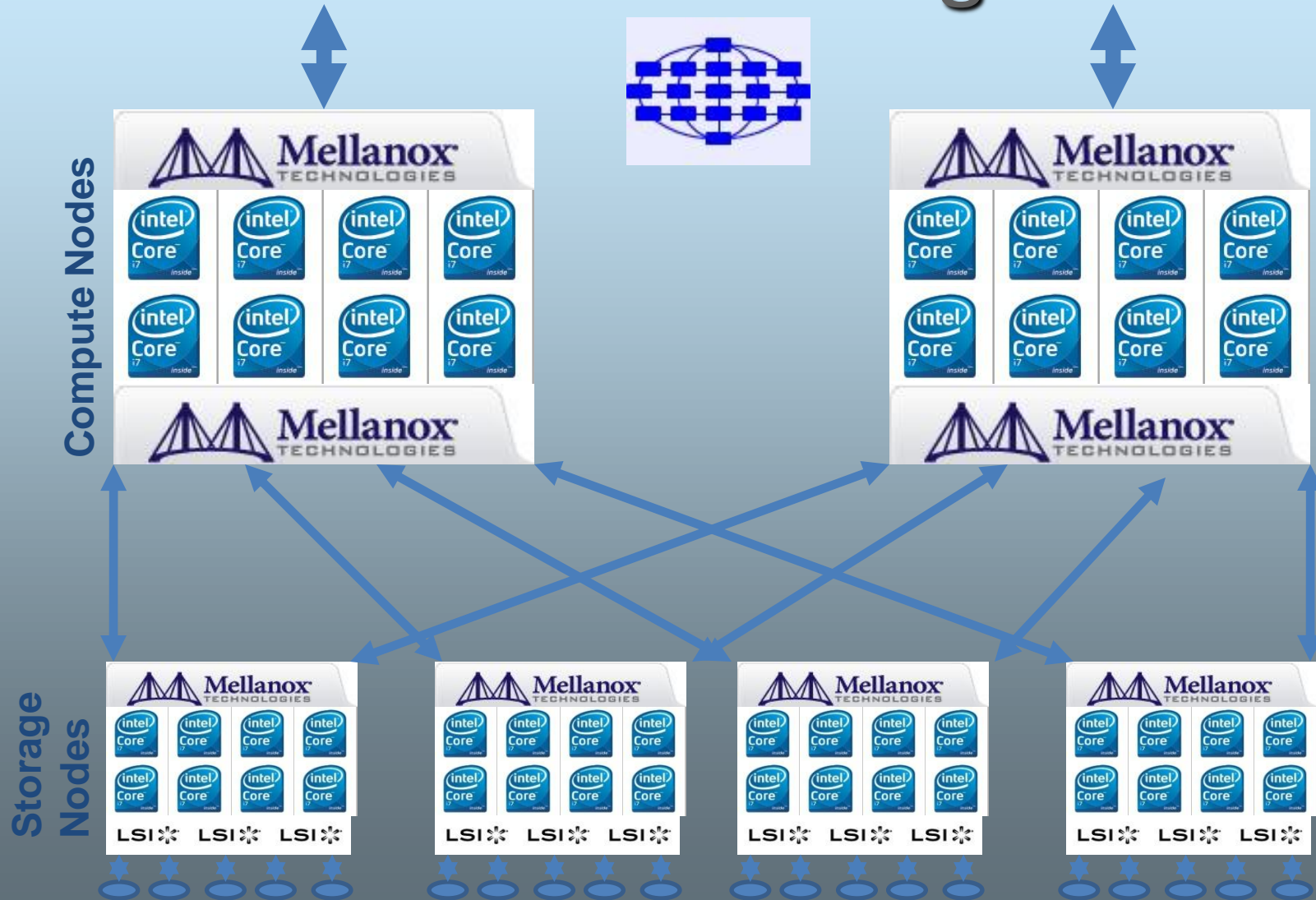
MB/Sec



Storage Node to Compute Node
Transfer Rates – 40Gbit IB

Measured on Mellanox Janus cluster

NumaRAID HPC Storage Natural Scaling



Benefits of the Change

- Decrease HPC storage hardware costs
 - High volume components benefit from economies of scale
 - Decrease power requirements by using latest generation components
 - Obsolescence Protection - Software licenses can be transferred as hardware platforms evolve – keep pace with current “best of breed”
- Increase HPC storage performance
 - Symmetrical read/write performance >2GBytes/sec per node today
 - Currently @ 13% single CPU load, efficiency increasing @ 60%/Year
 - 0% performance penalty in degraded mode
 - Performance often *increases* in the presence of failed drives
 - Lost data can be recomputed *faster* than existing drives can deliver it
- Increase HPC network reliability
 - Multi-node fault tolerance for both disk drives and arrays
 - High Speed detection and correction of Silent Data Corruption for both storage nodes and client nodes (full network coverage)
 - Validated with multiple file systems, including Lustre
 - Commercially deployed and independently verified
- Increase HPC application performance
 - Leverages years of application performance optimization
 - “Active Storage” architecture further decreases overall costs and simultaneously simplifies and accelerates HPC applications

NumaRAID Certifications



matrox

Blackmagicdesign



RORKE
DATA



POWERFUL SOLUTIONS FOR POST
DIGITAL FILM RESTORATION • DIGITAL WORKFLOW FOR DAILIES

BELLMICRO

IRIDAS.

ASSIMILATE

AV

AVRORA



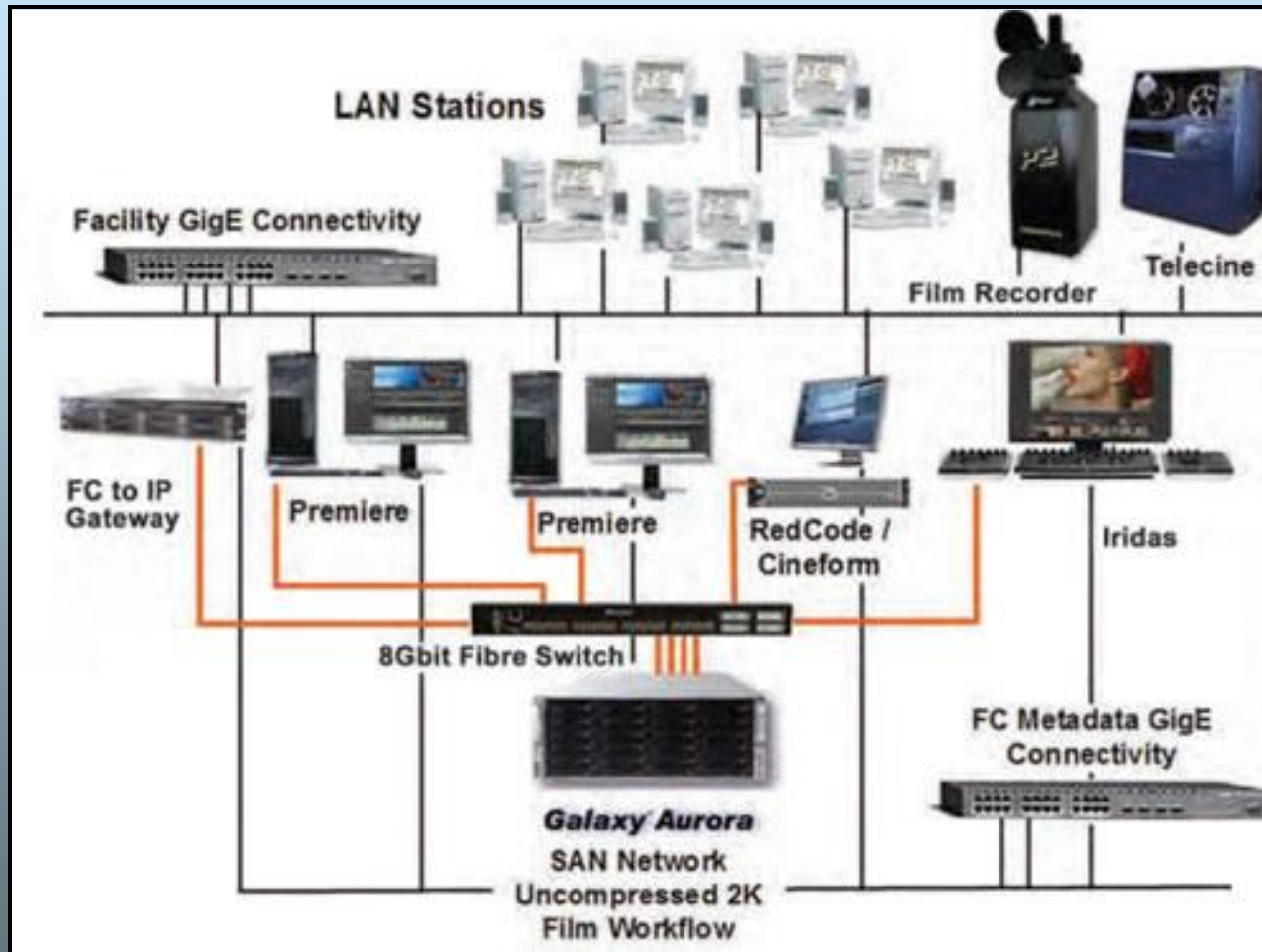
Russian Railways

MARIINSKY THEATRE



DrasticTech

NumaRAID Typical Deployment in Rich Media Applications



NumaRAID Storage Solution

An HPC Inspired RAID Stack

- Production level software, currently deployed worldwide
- Feature Rich
 - CLI and GUI for remote management
 - Centralized management console
 - Always at full speed, 0% degraded mode penalty
 - Silent Data Corruption detection and correction, partial reconstruction, fine grain cache controls, instant RAID6 initialize, embedded LUN security controls, WWN filtering and masking, port zoning, SNMP support, IPMI with power management
 - Extensive performance and reliability instrumentation with graphical display
- Supported by commercial business with active new feature development
 - N+3, N+4, N+M protection schemes
 - Comprehensive middleware solution to optimize HPC applications and maximize utilization
 - Virtualization specific storage acceleration
 - SSD Based High Transaction Rate Solutions
 - StreamScaling for Data Compression, Replication Recognition and Elimination
 - Rapid adoption of new technologies including
 - 6Gbit SAS, SSD, 10GigE, FCoE
 - Next generation AMD/Intel/nVidia Processors
 - High reliability enclosures

StreamScale

Product Offerings

- High performance storage solutions available now
 - Available on the Mellanox Cluster for remote evaluation
 - http://www.hpcadvisorycouncil.com/cluster_center_access.php
 - Fully configured, pre-tested hardware available from multiple integrators
 - Site licenses with full source code, documentation and training available from StreamScale
- Professional support and custom programming services available now
 - Highly qualified, experienced staff with deep parallel programming experience and extensive hardware infrastructure
 - ISO 9001 and CMMI Level 3 Certified
 - End-to-end project management
- Contacts
 - Mike Anderson, CTO - manderson@streamscale.com
 - Don McDonell, VP of Sales - dmcdonell@streamscale.com

Changing the Fundamentals of Data Storage for SuperComputers

The Detail



NumaRAID

An HPC Inspired RAID Stack

- NumaRAID is a set of Linux kernel modules, command line interfaces and graphical user interfaces
- NumaRAID is designed to increase the Performance, Reliability, Availability and Serviceability of compute and storage nodes in an HPC cluster
- By some definitions, NumaRAID transforms an HPC cluster into a SuperComputer
- NumaRAID is a commercial product, independently certified and deployed worldwide
- NumaRAID has about 25 man years of investment to date and an active development roadmap

Current Fundamentals of HPC Storage

- High bandwidth storage suitable for SuperComputer applications requires high drive-count RAID systems
- High drive-count RAID systems need multi-drive data protection (RAID6) for reliability
- The current high bandwidth RAID market is dominated by single purpose, hardware based RAID6 solutions
 - Proprietary hardware based Field Programmable Gate Arrays or ASICs perform the RAID6 (ECC) calculations
 - Require proprietary power, packaging, cooling and electronics
 - Single purpose, inflexible hardware
 - Relatively low volume drives high costs
- **Many similarities to *previous generation* SuperComputers**

Overview of 'StreamScaling' for RAID6 Storage

- Streaming increases Disk, DRAM and Interconnect Efficiency
 - Minimizes lost disk rotations, maximizes transfer rate
 - Minimizes command overhead of interconnect protocol stacks, maximizes transfer rate
 - Minimizes address cycles of DRAM, maximizes transfer rate
 - *Minimizes time and power to deliver data*
- Scaling increases Compute Efficiency
 - Optimized instruction selection, ordering and register usage maximizes CPU resources
 - Optimized scheduling, data placement and data flow maximizes multi-core and multi-CPU parallelism
 - *Minimizes time and power to compute result*
- 'StreamScaled' RAID6 solution outperforms hardware based solutions
 - >100:1 performance improvement over previous designs
- **Eliminates hardware/software dependency for HPC Storage**

Changing the Fundamentals of HPC Storage

- Separate the Hardware technology from the Software technology
 - Software and standard processors replaces high cost, proprietary ASICs and FPGAs
 - ‘StreamScaled’ RAID6 stack outperforms custom hardware solutions and fully leverages the 60% a year increase in processor performance
 - Transferrable licenses minimize the risk of obsolescence
- Leverage the economy of scale for hardware components and software licenses
 - Lower acquisition and service costs
 - Experienced and reliable Compute Node (commodity component) suppliers become High Performance Storage Node suppliers
- Keep pace with latest technology innovations
 - Select current “best of breed” as new hardware innovations are introduced
- Many similarities to *current generation* SuperComputers

NumaRAID

RAID6 Stack

- Written in C language, implemented in Linux
- Designed as a parallel processing RAID solution “from the ground up”
- Currently achieves >2Gbytes/second read/write data transfer under extremely heavy load (See Appendix 2)
 - Dual drive failure case requires solution to:

$$D_y = \frac{(P + \sum_{i=0}^{n-1} D_i) + (Q + \sum_{i=0}^{n-1} D_i g^{i+1})g^{-x-1}}{1 + g^{y-x}}$$

- Highly optimized design delivers ~100x performance increase
- “StreamScaling” design can be extended to solve other computational challenges (Active Storage)

NumaRAID

Performance and CPU Load

- RAID logic requires ~10% of a single CPU @ >2Gbytes/Second
 - Intel I7-920 2.6Ghz (See Appendix 3)
- ECC solution requires ~13% of a single CPU @>2Gbytes/Second
 - Intel I7-920 2.6Ghz (See Appendix 4)
- ECC can be applied to both data reconstruction and Silent Data Corruption detection *and* correction
- Driver can be deployed in storage node, and/or in compute node
 - Extends Silent Data Corruption solution to entire network
 - Extends multi-node fault tolerance across multiple RAIDs
 - Offers highly optimized software Compute Infrastructure

NumaRAID

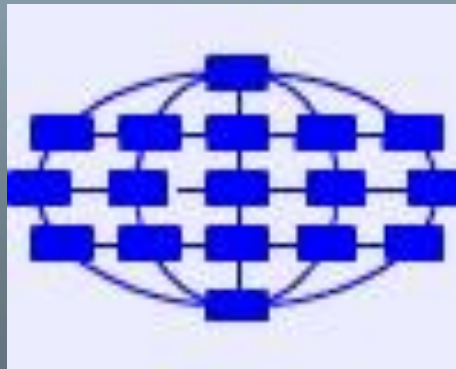
As a Compute Infrastructure

- Many HPC applications have very regular, but massive, data structures
- File systems often “get in the way” of high performance solutions
- Standard SAN topology rules (SRP) used to define single massive RAID with flexible device level interconnect hierarchy across nodes
 - Device level data flow eliminates file system bottlenecks
- Driver does “Heavy Lifting” of data path with StreamScale technology
 - Proven parallel infrastructure to maximize CPU utilization
 - Highly optimized methodology to deliver all data to/from CPU registers
 - Fine grain control to tune read/write caches to minimize latencies and maximize pipelining
- Well defined API to extend math processing capabilities
 - Completely isolated from device infrastructure
- Extensive embedded instrumentation
 - Simplified data flow and makes identification and elimination of performance bottlenecks straightforward

NumaRAID

Compute Infrastructure Example

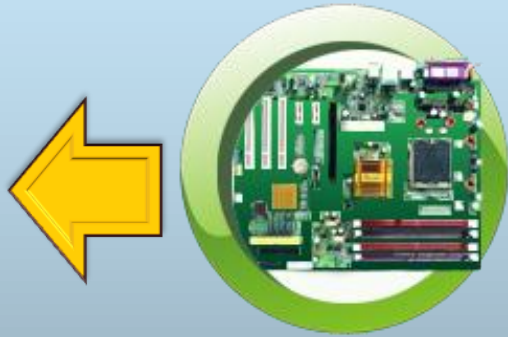
1. Load NumaRAID driver in compute nodes.
2. Configure NumaRAID interconnect topology using standard SAN LUN Masking rules (SRP) to match data flow interconnect to 'ideal' topology.
3. Configure NumaRAID math processing component and cache optimization rules in each compute node.
4. Perform 'dd' from existing file system to NumaRAID 'input device' to supply input.
5. Perform 'dd' from NumaRAID 'output device' to file system to record results.
6. Examine instrumentation results to optimize configuration for next run.





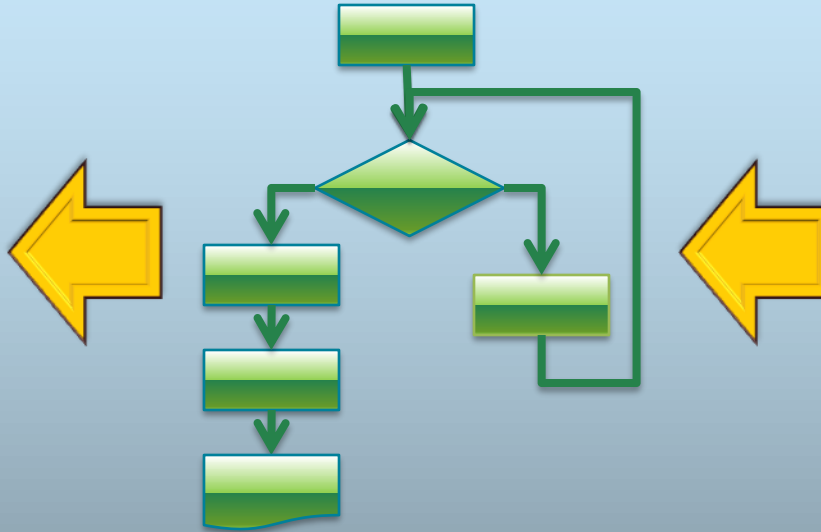
NumaRAID

What's inside the Box?



IB – 40 Gb/sec
FC – 8 Gb/sec

Galois Field
Based ECC
Data Validation
And Recovery



SATA – 120 MB/sec
SAS – 200 MB/sec
SSD - 250+ MB/sec

$$D_y = \frac{(P + \sum_{i=0}^{n-1} D_i) + (Q + \sum_{i=0}^{n-1} D_i g^{i+1}) g^{-x-1}}{1 + g^{y-x}}$$

*3rd party applications (Active Storage)
Majority of compute power available*

NumaRAID Deployments in Commercial Applications

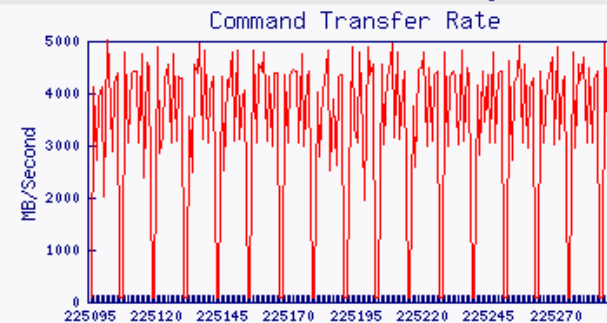
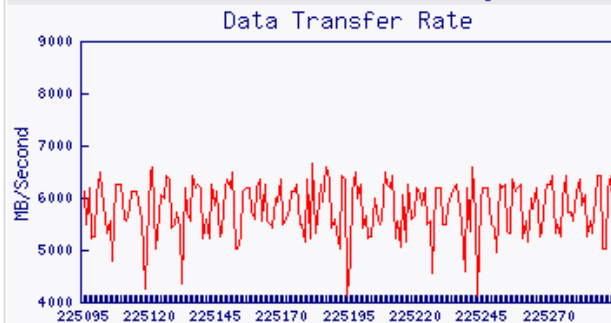
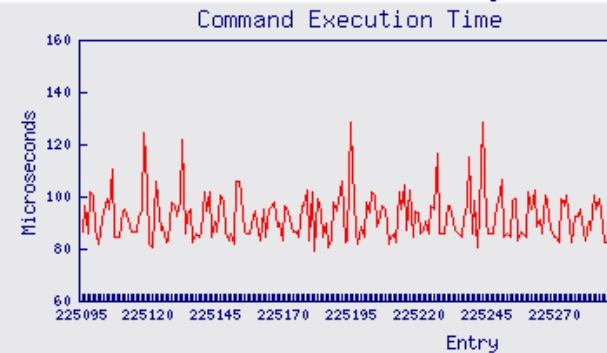
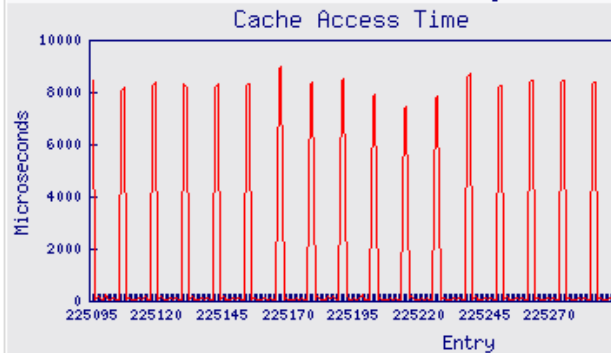
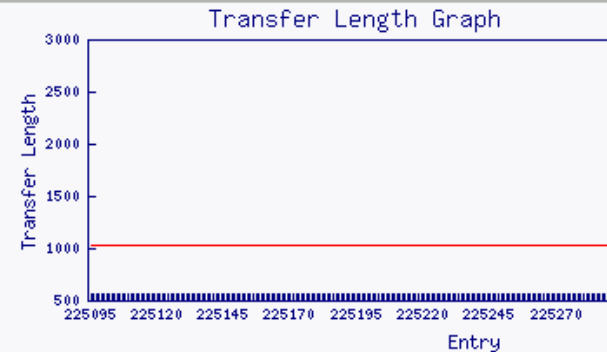
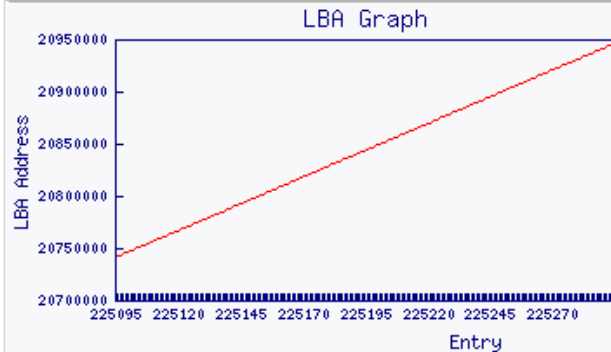
- Audio Video Storage Market – In Production
 - Rorke Data, a subsidiary of Bell Microproducts
 - Product branded “Galaxy Aurora”
- Enterprise Market – In Production
 - AvroRAID, St. Petersburg, Russia
 - Backup product supplied to Russian Railroads
- Digital Film Companies – Official Certifications
 - AJA, Assimilate, Black Magic, Drastic Tech , Iridas, Matrox, MTI
- Digital Video Supplier – Tested Compatibility
 - Autodesk, Avid, Adobe, Apple, Thomson, Quantel
- Client OS Support – Tested Compatibility
 - AIX, Apple OSX, Linux, Solaris, Windows

NumaRAID

Extensive Embedded Instrumentation

TRACE Details

Trace Results 'all' Total Entries=225320 Display=200 Current entry=225095



Appendix 1

Local, Remote and File System Tests: No failed drives

```
• =====  
• REMOTE TESTS  
• =====  
• [root@compute-01-22 ~]# mount -t xfs /dev/sdb2 /ib-storage/  
• [root@compute-01-22 ~]# dd if=/dev/zero of=/ib-storage/zzz bs=8M oflag=direct  
• dd: writing `/ib-storage/zzz': No space left on device  
• 58816+0 records in  
• 58815+0 records out  
• 493375979520 bytes (493 GB) copied, 340.982 seconds, 1.4 GB/s  
• [root@compute-01-22 ~]# dd of=/dev/null if=/ib-storage/zzz bs=8M iflag=direct  
• 58815+0 records in  
• 58815+0 records out  
• 493375979520 bytes (493 GB) copied, 233.833 seconds, 2.1 GB/s  
• [root@compute-01-22 ~]# umount /ib-storage/  
• [root@compute-01-22 ~]# dd of=/dev/null if=/dev/sdb2 bs=8M iflag=direct  
• 58831+1 records in  
• 58831+1 records out  
• 493516800000 bytes (494 GB) copied, 232.209 seconds, 2.1 GB/s  
• [root@compute-01-22 ~]# dd if=/dev/zero of=/dev/sdb2 bs=8M oflag=direct  
• dd: writing `/dev/sdb2': No space left on device  
• 58832+0 records in  
• 58831+0 records out  
• 493516800000 bytes (494 GB) copied, 233.749 seconds, 2.1 GB/s  
• =====  
• LOCAL TESTS  
• =====  
• [root@aurora NumaRAID]# dd if=/dev/sdab2 of=/dev/null bs=8M iflag=direct  
• 58831+1 records in  
• 58831+1 records out  
• 493516800000 bytes (494 GB) copied, 212.329 seconds, 2.3 GB/s  
• [root@aurora NumaRAID]# dd of=/dev/sdab2 if=/dev/zero bs=8M oflag=direct  
• dd: writing `/dev/sdab2': No space left on device  
• 58832+0 records in  
• 58831+0 records out  
• 493516800000 bytes (494 GB) copied, 217.843 seconds, 2.3 GB/s
```

Appendix 2

Local, Remote and File System Tests: 2 failed drives

```
• =====  
• REMOTE TESTS  
• =====  
• [root@compute-01-22 ~]# dd if=/dev/zero of=/ib-storage/zzz bs=8M oflag=direct  
• dd: writing `/ib-storage/zzz': No space left on device  
• 58816+0 records in  
• 58815+0 records out  
• 493375979520 bytes (493 GB) copied, 306.402 seconds, 1.6 GB/s  
• [root@compute-01-22 ~]# dd of=/dev/null if=/ib-storage/zzz bs=8M iflag=direct  
• 58815+0 records in  
• 58815+0 records out  
• 493375979520 bytes (493 GB) copied, 233.476 seconds, 2.1 GB/s  
• [root@compute-01-22 ~]# umount /dev/sdb2  
• [root@compute-01-22 ~]# dd if=/dev/zero of=/dev/sdb2 bs=8M oflag=direct  
• dd: writing `/dev/sdb2': No space left on device  
• 58832+0 records in  
• 58831+0 records out  
• 493516800000 bytes (494 GB) copied, 237.077 seconds, 2.1 GB/s  
• [root@compute-01-22 ~]# dd of=/dev/null if=/dev/sdb2 bs=8M iflag=direct  
• 58831+1 records in  
• 58831+1 records out  
• 493516800000 bytes (494 GB) copied, 231.788 seconds, 2.1 GB/s  
• =====  
• LOCAL TESTS  
• =====  
• [root@aurora NumaRAID]# dd of=/dev/sdab2 if=/dev/zero bs=8M oflag=direct  
• dd: writing `/dev/sdab2': No space left on device  
• 58832+0 records in  
• 58831+0 records out  
• 493516800000 bytes (494 GB) copied, 190.611 seconds, 2.6 GB/s  
• [root@aurora NumaRAID]# dd if=/dev/sdab2 of=/dev/null bs=8M iflag=direct  
• 58831+1 records in  
• 58831+1 records out  
• 493516800000 bytes (494 GB) copied, 185.895 seconds, 2.7 GB/s
```

Appendix 3

CPU Measurements: No ECC Load, 2.1Gbyte/Second over IB

- top - 23:55:56 up 7 days, 11:46, 1 user, load average: 1.94, 0.99, 0.42
- Tasks: 195 total, 7 running, 188 sleeping, 0 stopped, 0 zombie
- Cpu(s): 0.0%us, 9.6%sy, 0.0%ni, 89.7%id, 0.0%wa, 0.6%hi, 0.1%si, 0.0%st
- Mem: 5967248k total, 4881880k used, 1085368k free, 168984k buffers
- Swap: 786424k total, 0k used, 786424k free, 365288k cached

```
•
• PID USER PR NI VIRT RES SHR S %CPU %MEM TIME+ COMMAND
• 8249 root 10 -5 0 0 0R 9 0.0 7:21.10 NR T11 0 [3]
• 8246 root 10 -5 0 0 0S 9 0.0 7:22.69 NR T8 0 [0]
• 8248 root 10 -5 0 0 0R 9 0.0 7:23.16 NR T10 0 [2]
• 8250 root 10 -5 0 0 0S 8 0.0 7:22.71 NR T12 0 [4]
• 8247 root 10 -5 0 0 0S 8 0.0 7:19.62 NR T9 0 [1]
• 8252 root 10 -5 0 0 0R 8 0.0 7:23.04 NR T14 0 [6]
• 8253 root 10 -5 0 0 0R 8 0.0 7:21.08 NR T15 0 [7]
• 8251 root 10 -5 0 0 0R 8 0.0 7:20.34 NR T13 0 [5]
• 8256 root 10 -5 0 0 0S 2 0.0 1:33.25 NR T18 0 [10]
• 8257 root 10 -5 0 0 0S 2 0.0 1:34.35 NR T19 0 [11]
• 8260 root 10 -5 0 0 0S 2 0.0 1:33.58 NR T22 0 [14]
• 8261 root 10 -5 0 0 0S 2 0.0 1:33.61 NR T23 0 [15]
• 8254 root 10 -5 0 0 0R 1 0.0 1:33.52 NR T16 0 [8]
• 8255 root 10 -5 0 0 0S 1 0.0 1:33.49 NR T17 0 [9]
• 8259 root 10 -5 0 0 0S 1 0.0 1:33.52 NR T21 0 [13]
• 30694 root 15 0 12716 1156 816 R 0 0.0 0:00.42 top
```

Appendix 4

CPU Measurements: Full ECC Load, 2.1GByte/Second over IB

- top - 07:42:38 up 6 days, 19:33, 1 user, load average: 1.26, 2.43, 2.18
- Tasks: 194 total, 6 running, 188 sleeping, 0 stopped, 0 zombie
- Cpu(s): 0.0%us, 23.3%sy, 0.0%ni, 75.7%id, 0.0%wa, 0.7%hi, 0.2%si, 0.0%st
- Mem: 5967248k total, 4870892k used, 1096356k free, 168112k buffers
- Swap: 786424k total, 0k used, 786424k free, 364200k cached

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
8188	root	10	-5	0	0	0R	32	0.0	18:03.80	NR T0	GF [0]
8190	root	11	-5	0	0	0R	30	0.0	18:05.76	NR T2	GF [2]
8191	root	10	-5	0	0	0R	29	0.0	18:08.18	NR T3	GF [3]
8189	root	10	-5	0	0	0R	27	0.0	18:06.13	NR T1	GF [1]
8251	root	10	-5	0	0	0S	8	0.0	5:42.26	NR T13	0 [5]
8247	root	10	-5	0	0	0S	8	0.0	5:41.47	NR T9	0 [1]
8246	root	10	-5	0	0	0S	7	0.0	5:42.60	NR T8	0 [0]
8252	root	10	-5	0	0	0S	7	0.0	5:42.79	NR T14	0 [6]
8248	root	10	-5	0	0	0S	7	0.0	5:42.89	NR T10	0 [2]
8249	root	10	-5	0	0	0R	7	0.0	5:43.03	NR T11	0 [3]
8250	root	10	-5	0	0	0S	7	0.0	5:42.59	NR T12	0 [4]
8253	root	10	-5	0	0	0S	7	0.0	5:42.73	NR T15	0 [7]
8255	root	10	-5	0	0	0S	2	0.0	1:28.08	NR T17	0 [9]
8259	root	10	-5	0	0	0S	2	0.0	1:28.31	NR T21	0 [13]
8256	root	10	-5	0	0	0S	2	0.0	1:28.01	NR T18	0 [10]
8258	root	10	-5	0	0	0S	2	0.0	1:27.93	NR T20	0 [12]
8254	root	10	-5	0	0	0S	1	0.0	1:27.89	NR T16	0 [8]
8257	root	10	-5	0	0	0S	1	0.0	1:29.12	NR T19	0 [11]
8260	root	10	-5	0	0	0S	1	0.0	1:28.36	NR T22	0 [14]
8261	root	10	-5	0	0	0S	1	0.0	1:28.18	NR T23	0 [15]

About the Author

Michael H. Anderson

Mr. Anderson has been designing high performance storage systems for over 30 years. He currently holds 10 US patents related to storage.

Number	Title	Issue Date
6971042	Media server with single chip storage controller	11.29.2005
6608966	VCR-type controls for video server system	08.19.2003
6473875	Error correction for network delivery of video streams using packet resequencing	10.29.2002
6442649	Dynamic expansion of storage device array	08.27.2002
6430118	Data storage utilizing parity data to enhance performance	6.8.2002
5805919	Method and system for interleaving the distribution of data segments from different logical volumes on a single physical drive	8.9.1998
6640235	Expandable mass disk drive storage system	10.28.2003
6148142	Multi-user, on-demand video server system including independent, concurrently operating remote data retrieval controllers	11.14.2000
5519435	Multi-user, on-demand video storage and retrieval system including video signature computation for preventing excessive instantaneous server data rate	05.21.1996
5191584	Mass storage array with efficient parity calculation	2.3.1993

Thank you

“The RAID *is* the SuperComputer.”