I/O: Toward The Exascale Era
Cray User Group

Keith Miller
Technical Director WW HPC and LS, DDN
DDN is the world leader in massively scalable storage and processing technology for unstructured & big data applications.

Established: 1998
Revenue: $250M+ Per Year, Profitable & Growing
Headquarters: Chatsworth, California USA
Employees: 500+ Worldwide
Worldwide Presence: 4 Continents, 15 Countries
Installed Base: 1,000+ Customers; 50 Countries
Go to Market: Global Partners, VARs, Resellers
World-Renowned & Award Winning:

DDN | The Scalability Leader
DDN | 2012 Update

INFRASTRUCTURE
Organizational L1/2/3, ProServe
Systems
Oracle

PEOPLE
Organizational
• 140: Data Intensive Computing Team
• World’s Largest Lustre Front-Line

R&D
Storage Fusion
• 1M+ of State-Machine Code
• 100s of Engineers

– The World’s Largest Independent Storage Company –
– Heavily Invested In Lustre and The Lustre Ecosystem –
– Driven By HPC –
Current Leadership

Real-Time, HPC State Machine
- 1M Lines of Zero-Interrupt Storage Engine Code
- Highly-Parallel Storage Processing Architecture
- Adaptive RAM Cache for Mixed Workloads, Journals
- Embedded Virtualization For ExaScaler™ Appliances

Quality Of Service
- Critical For Strided Writes & Reads
  - SFA Technology Maximizes Cluster Productivity
  - Performance Degredation Less Than 10%
- Real-Time Latency Management

Autonomous, Self-Healing Technology
- Automatic Drive Power Cycling
- Predicts and Prevents Drive Failures
- Minimizes Failure Instances by 80%
### DDN & WC | Partners in HPC

<table>
<thead>
<tr>
<th>T3S: Tech. Sales, Support &amp; Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The world’s largest team of HPC storage solution support and delivery experts</td>
</tr>
<tr>
<td>• 1 interface; 1 solution</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Whamcloud Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The de facto developers of core Lustre</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DataDirect Networks Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 10 years of developing Lustre storage solutions</td>
</tr>
<tr>
<td>• Real-Time, State Machine Leaders</td>
</tr>
</tbody>
</table>

One Tree For WW Lustre Supportability
A Look Toward Exascale
Exascale Systems will be big…

Billion-Way Parallelism Will Be A Reality

<table>
<thead>
<tr>
<th>System attributes</th>
<th>2010</th>
<th>“2018”</th>
</tr>
</thead>
<tbody>
<tr>
<td>System peak</td>
<td>2 Peta</td>
<td>1 Exaflop/sec</td>
</tr>
<tr>
<td>Power</td>
<td>6 MW</td>
<td>20 MW</td>
</tr>
<tr>
<td>System memory</td>
<td>0.3 PB</td>
<td>32-64 PB</td>
</tr>
<tr>
<td>Node performance</td>
<td>125 GF</td>
<td>1 TF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 TF</td>
</tr>
<tr>
<td>Node memory BW</td>
<td>25 GB/s</td>
<td>0.4 TB/sec</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 TB/sec</td>
</tr>
<tr>
<td>Node concurrency</td>
<td>12</td>
<td>O(1,000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>O(10,000)</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>18,700</td>
<td>1,000,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100,000</td>
</tr>
<tr>
<td>Total Node Interconnect BW</td>
<td>1.5 GB/s</td>
<td>200 GB/sec</td>
</tr>
<tr>
<td>MTTI</td>
<td>days</td>
<td>O(1 day)</td>
</tr>
</tbody>
</table>

New Configurations Are Outfitting 1000s of Cores Per Data Center Rack

The impact on Storage:

- **B-Way** Parallelism will challenge Exascale file system locking and data integrity mechanisms
- New **memory-class storage** will require new approaches for tiering and data locality
- Exascale consortiums are seeking new methods to **reduce cluster and network workload**
Hyper-Concurrency is a Result of Exascale CPU Evolution

Hyper-Linear Challenges for Locking Engines

<table>
<thead>
<tr>
<th>Year</th>
<th>Performance (TF)</th>
<th>Concurrency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>2000</td>
<td>200000</td>
</tr>
<tr>
<td>2015</td>
<td>20000</td>
<td>50000000</td>
</tr>
<tr>
<td>2018</td>
<td>10000000</td>
<td>10000000000</td>
</tr>
</tbody>
</table>
“It took 100 months for us to create our first Petabyte of data. It took us only one month to create our 18th Petabyte.”

Jeffrey Nichols, Associate Lab Director
October, 2011

Big Data Challenges @ Exascale:
Moving PB Objects, Datasets
Managing Diverse Data Models

Mainframe Era
1980s

Client:Server Era
1990s

Mobility/Web Era
2000s

Big Data Era
2010s

Convergence @ HyperScale

100Ms: Concurrent Users
The Internet of Everything

100Ms: Parallel Threads
Exascale Computing

Hyperscale

Requirements

CLOUD

HPC
Today’s IO Models: Too Much Clutter

Simplify the layers of Application Stack
- Collapse 5 separate data structures into one
  Each layer has its own “Database”
- Whether its named POSIX or SQL

Each layer has failover and redundancy
Each layer has complex recovery
Each layer has garbage collecting and maintenance

Data Transparency will address these issues
Object Storage Is Enabling A Re-Think On Data Layouts

Science – NetCDF, HDF5, SciDB, etc.
- Self-describing data
- Metadata built into the file itself

Web – NoSQL stores
- Flexible data structures (schemas)
- Easy
- Speed, scale

Objects (Cloud)
- Flat, global, arbitrary naming
- Integral sector management
- User-defined metadata
The Evolution of Object Storage

“Hundreds of millions of people use object storage every day – and don’t even know it.”

Compliance Era:
Document Vaults
WORM Storage
Eg. EMC Centera

Parallel I/O Era:
Scale-Out
Eg. Lustre, GPFS

Cloud Era
Peer:Peer, Hyperscale
Eg. Amazon S3
A Hyperscale Case Study

The Cloud Scales: Amazon S3 Growth

Peak Requests:
650,000+
per second

Total Number of Objects Stored in Amazon S3

Q4 2006: 2.9 Billion
Q4 2007: 14 Billion
Q4 2008: 40 Billion
Q4 2009: 102 Billion
Q4 2010: 262 Billion
Q4 2011: 762 Billion
Q1 2012: 905 Billion

Source: Amazon S3 Blog
Storage: Sources of Latency

Hardware Chain
- Disk drive servo operation
- Multiple SCSI layers
- Multiple bus transitions
- Memory bandwidth limitations
- Network service latencies

Software Chain
- Memory copies
- Kernel operations
- Layers of consecutive operations including the service of V-nodes, I-nodes and FAT
- Serial data transport processes

WOS is designed to reduce latency in all phases of data capture and retrieval
Understand the data usage model in a collaborative environment where immutable data is shared and studied

- A simplified data access system
- Eliminates the concept of FAT, extent lists to maximize efficiency
- Reduce the instruction set to only PUT, GET, & DELETE
- Add the concept of locality based on latency to data and load balance

Abandons storage convention entirely
WOS | Overview

GeoDistributed, Scale-out Object Storage System
► Hyper-Scalable Cloud Storage Foundation

TRUE End:End Object Storage
► Maximum Performance From Every Media

Easy to Manage at Hyperscale
► Single namespace, Single Global Cluster Interface

Autonomous, Self-Healing Big Data Infrastructure
► Intelligent, Fail-In-Place Architecture

Flexible Cloud Storage Service Platform
► Multimodal Access Featuring Billing & Multi-Tenancy
WOS Exascale Advantages

WOS is an end to end object placement file system.

- WOS has no concept of fragmentation
- Objects 1MB or less are stored in contiguous space minimizing actuator usage in rotating media and simplifying internal maps in solid state media

WOS is efficient

- Objects are immutable so there is no concept of “File open for xWRITE”
- Locking is completely eliminated
- Like size objects are always stored together in Object Resource Groups (ORG) so that there is no concept of “garbage collection” on a block level basis
- Operations of PUT and GET are accomplished in one concise internal transaction layer
WOS addresses data corruption in multiple dimensions

- All objects are written with a checksum
- The checksum is evaluated for every GET and every bus transition
- All objects can be written with erasure codes distributed on multiple storage devices
- Nodes automatically recover data in the case of silent data corruption with a rapid, object aware, rebuild operation
- Nodes automatically recover data in the case of media failure

WOS was designed for large scale data operations

- Test limits are 256 billion objects utilizing 256 nodes

WOS is a peer to peer data distribution system which could be utilized to enable collaboration
The Power of Object Storage

<table>
<thead>
<tr>
<th></th>
<th>EMC Atmos</th>
<th>HPCS</th>
<th>Megastore</th>
<th>Amazon S3</th>
<th>WOS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reads/Day</strong></td>
<td>500,000,000</td>
<td>2,764,800,000</td>
<td>20,000,000,000</td>
<td>32,000,000,000</td>
<td>55,472,947,200</td>
</tr>
<tr>
<td><strong>Writes/Day</strong></td>
<td>500,000,000</td>
<td>2,764,800,000</td>
<td>3,000,000,000</td>
<td>-</td>
<td>23,113,728,000</td>
</tr>
</tbody>
</table>

Amazon: [http://aws.typepad.com/aws](http://aws.typepad.com/aws)
EMC: [http://reg.cx/1P1E](http://reg.cx/1P1E)
HPCS: [http://www.spscicomp.org/ScicomP13/Presentations/IBM/GPFSGunda.pdf](http://www.spscicomp.org/ScicomP13/Presentations/IBM/GPFSGunda.pdf)
Intelligent Object Storage Lays Foundation For Exascale Efficiency

Object Stores
► Versatile data model that works with distributed block management, security and data reliability/recovery
► Tiered, Scalable to B’s of reads/writes per second

Converged Storage & Processing
► Pre and post processing functions and achieving integrated ILM services while making applications more data aware

Integrated Analytics
► Integrated Map-Reduce and Analytics services designed to turn PBs of data into knowledge
Even More Integrated, Comprehensive Appliance Management Tools

Exascale Engineering
Objects, Data Models, Storage-Class Memory

Stay Tuned....

www.ddn.com
Thank You