Redefine

Isilon Performance

Name
Agenda

• Architecture Overview
• Next Generation Hardware
• Performance Caching
• Performance Streaming Reads
• Performance Tuning
OneFS
Architecture Overview
OneFS Network Architecture

- **Servers**
  - Windows
  - Linux
  - Mac

- **Client/Application Layer**

- **Standard 1GbE / 10GbE layer**

- **Isilon IQ Storage Layer**

- **IB for Intracluster Communication**

- **NFS, CIFS, FTP, HTTP, HDFS, RAN**

- **(optional 2nd switch)**
Isilon Scale-Out NAS Product Family

- **S200**: Purpose-built for transactional & IOPS-intensive workloads.
- **X200**: A slimmer solution cost-optimized for performance.
- **X400**: A flexible solution to accelerate high-concurrent and sequential throughput workloads.
- **NL400**: Purpose-built for cost-effective, high capacity storage.

Near-Linear Scaling of Performance and Capacity
Storage Workloads

- **X400**: A flexible solution to accelerate high-concurrent and sequential throughput workloads.
- **X200**: An optimized solution for high performance.
- **S200**: Purpose-built for transactional & IOPS-intensive workloads.
- **NL400**: Purpose-built for high capacity storage.
- **Hadoop / Analytics**
- **Home Directories**
- **Medical Imaging**
- **Electronic Design Automation**
- **Media & Entertainment**
- **Scientific HPC**
- **Active Archive**
- **Cold Archive**
- **Commercial HPC**

© Copyright 2014 EMC Corporation. All rights reserved.
Next Generation Hardware
S210 – Next Generation S Series

Specifications

• **CPU**: dual, 6-core Ivy Bridge Processors
• **RAM**: 32GB to 256GB
• **Drives**: 24 X 2.5-inch bays
  - **HDD**: 2.5-inch SAS, 300GB-1.2TB each
  - **SSD**: up to 6 SSDs, 200GB-800GB each
  - Self-encrypted options available
• **Front-end I/O**: 2x1GbE + 2x10GbE
• **Back-end I/O**: QDR Infiniband, 1m-100m cabling
• **Chassis**
  - Standard 2U enclosure
  - Dual redundant, hot swappable PSUs
Specifications

- **CPU**: dual, 8-Core Ivy Bridge Processors
- **RAM**: 32GB to 256GB
- **Drives**: 36 X 3.5-inch bays
  - **HDD**: 3.5-inch SATA, 1/2/3/4TB each
  - **SSD**: up to 6 SSDs, 400GB/800GB each
  - Self-encrypted options available
- **Front-end I/O**: 2x1GbE + 2x10GbE
- **Back-end I/O**: QDR Infiniband, 1m-100m cabling
- **Chassis**: Standard 4U enclosure
  - Dual redundant, hot swappable PSUs (high line only)
Flash Uses

Cost

Performance

Flash Uses

Data on Flash
File Data is stored on flash. Combined with either Metadata Read or Metadata Write policies.

Metadata Write
Stores all metadata on flash for faster writes and reads.

L3
(new in OneFS 7.1.1)
Intelligently chooses data and metadata to cache on flash.

GNA
Global Name Acceleration adds flash to existing clusters/nodes that otherwise wouldn’t have access.

© Copyright 2014 EMC Corporation. All rights reserved.
Caching

The importance of caching on OneFS

- OneFS uses extra RAM to cache frequently used file data and internal structures
- Blocks are cached at multiple layers
  - L1 - RAM
  - L2 - RAM
  - L3 - SSD - New in 7.1.1
Cache Layout

Node Memory

- Most node RAM dedicated to read cache
- L1 read cache
  - Cached File Data
  - Local to node
  - Removes internode latency, ~100 µs
- L2 read cache
  - Cached Block Data from local drives
  - Accessible by other nodes
  - Removes disk latency, ~5-7 ms (SATA)
Cache Layout with L3

Node Memory

Services

Protocols

Filesystem State

L3 Cache
- SSDs
- Stores active
  - Metadata
  - Data Blocks

Node SSDs

© Copyright 2014 EMC Corporation. All rights reserved.
L3 Cache: Functional Overview

- L3 cache on SSDs, not in RAM
  - Think of it as an extension of L2 cache (just not in RAM)
- Durable
  - Unlike L2, no cache loss on reboot
- Reduce expensive random read I/O from HDDs
- Caches both metadata and user data, unlike other metadata policies
L3 Cache: Performance

• SpecSFS
  – Op/s equivalent to metadata read acceleration
  – Metadata-write SSD strategy still provides best performance

• 10-20x latency improvements for repeated random read workflows
  – EDA/build, VMDK ...

• Improves many workflows
  – The primary exception are those where Metadata Write is required today
L3 Benefits

• Greatly simplify sizing

• Allows nodes to be sold with fewer SSDs
  – 2% rule for SSD capacity is not required when using L3

• Less degradation if your metadata exceeds your SSD capacity

• 7.1.1 will default to L3
Performance Characteristics
Streaming Read

- OneFS spreads out a single file across multiple disks
  - Streaming performance isn’t limited to a single disk
- A single large read (>128KB) can be serviced by multiple disks at once
- Prefetch - OneFS issues read requests to drives ahead of the client requests to hide latency
Streaming read

Visual Example

**Key**

- **In L2**
- **On Disk**
- **In L1**
- **Read**

© Copyright 2014 EMC Corporation. All rights reserved.
Streaming Read changes in 7.1.1

Prefetch Improvements

• Prefetch can be offloaded to a separate thread
• Up to 64MB of data from disk to L2
• Up to 8MB of data into L1
• Previous versions are less aggressive and less efficient
Beginning Of SMB 3.0 - MultiChannel

- 7.1.1 brings the first SMB 3.0 feature to OneFS
- SMB 3.0 client can establish multiple TCP connections to file servers
  - High performance with bandwidth aggregation
  - Fault tolerant with multiple network paths
- 3 configurations can benefit:
  - Single RSS-capable NIC
  - Multiple NICs
  - Teamed NICs (LACP)
SMB 3.0 Multi-Channel

• Requirements
  – Windows 2012/Windows 8 or later
  – OneFS 7.1.1 or later

• Limited to a single node

• Interfaces must be the same type

• Still requires the client and application to be smart about requesting and processing data
Biggest Performance Increase Is Using Multiple NICs w/MultiChannel

1 session, without Multichannel
• No automatic failover
• Can’t use full bandwidth
  • Only one NIC engaged
  • Only one CPU core engaged

1 session, with Multichannel
• Automatic NIC failover
• Combined NIC bandwidth
  • Multiple NICs engaged
  • Multiple CPU cores engaged
Peak Aggregate Throughput
SMB2 Read, 5-node clusters

Total MiB/s

- 7.0.2.3
- 7.1.0.1
- 7.1.1.rc
Performance Tuning
OneFS Performance Tuning

• Tuning should be done via File Pool Policies
  – Easy to configure and verify
  – Access to some tuning parameters

• Can be done in the Web Interface File System Explorer
  – Doesn’t require SmartPools license
  – Not easy to verify configuration
  – Access to some tuning parameters

• Can be done in the CLI with the *isi set* command
  – Doesn’t require SmartPools license
  – Not easy to verify configuration
  – Access to all tuning parameters
## File Policies
### Choosing The Correct SSD Strategy

<table>
<thead>
<tr>
<th>SSD Usage</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3 Cache (Default in 7.1.1)</td>
<td>Use for all except Write centric workflows</td>
</tr>
<tr>
<td>Metadata Read Acceleration</td>
<td>Read centric, low SSD:HDD ratio</td>
</tr>
<tr>
<td>Metadata Write Acceleration</td>
<td>Write centric, moderate SSD:HDD ratio</td>
</tr>
<tr>
<td>Data On SSD</td>
<td>Performance critical, small data set, high SSD:HDD ratio</td>
</tr>
<tr>
<td>Avoid SSD</td>
<td>Archived, cold data</td>
</tr>
</tbody>
</table>
### File Policies

Choosing the Correct Access Policy

<table>
<thead>
<tr>
<th>Policy</th>
<th>Prefetch</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimize for Random Access</td>
<td>Disabled</td>
<td>Virtual Machines</td>
</tr>
<tr>
<td>Optimize for concurrent Access</td>
<td>Minimum</td>
<td>Home directories</td>
</tr>
<tr>
<td>Optimize for streaming Access</td>
<td>Maximum</td>
<td>Large files</td>
</tr>
</tbody>
</table>

#### SmartCache

<table>
<thead>
<tr>
<th></th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>On</td>
<td>Random</td>
</tr>
<tr>
<td>Off</td>
<td>Streaming</td>
</tr>
</tbody>
</table>
Spec Total Ops/s
Home Directory Mix, Per-Node

Total Ops/sec

- 7.0.2.3
- 7.1.0.1
- 7.1.1.rc

© Copyright 2014 EMC Corporation. All rights reserved.
Takeaways
Get excited about 7.1.1 & Performance

Where We Really Deliver With This Release

• Customers with lots of namespace operations
  – Home directories, EDA, etc
  – X400 nodes get a ~10% improvement in SpecSFS tests
  – X410 nodes get 30-60% more performance than X400 nodes

• Aggregate read performance is 15-30% better

• Applications that need great single-stream performance with Win8/Windows2012 – up to 1.4GB/s

• This software update will allow customers to unlock performance on nodes they already own!