Reliability of NVM devices for I/O Acceleration on Supercomputing Systems

Hitoshi Sato*1, Shuichi Ihara2), Satoshi Matsuoka1)

*1 Tokyo Institute of Technology
*2 Data Direct Networks Japan
Requirements for Next-gen I/O Subsystem

• TSUBAME2
  – Pioneer in the use of local flash storage
    • 200TB of SSDs for productive operation since 2010
    • Tired and hybrid storage environments, combining local flash with external Lustre FSs

• Industry Status
  – Various flash devices emerging
    • Available at reasonable cost
  – New Approaches
    • Flat Buffers
    • Burst Buffers etc.,
Issues for Introducing Flash Devices on Next-gen Supercomputers

- Flash Devices have various performance characteristics
- How much reliability of flash devices is required for burst buffers for next-gen supercomputers?
  - Maximize
    - Throughput, IOPS, Capacity, Reliability, etc.
  - Minimize
    - Cost for introducing flash devices on large-scale systems
Today’s Talk

• Preliminary Evaluation of the endurance for SSD devices to support supercomputing I/O workloads

• Lustre I/O Monitoring for I/O workload analysis to evaluate the endurance
Towards Extreme-scale Supercomputing

• Memory/Storage Architecture
  – NVM (Non-Volatile Memory, Flash) is a key device for I/O subsystems
    • Capacity, Non-Volatility, Low energy consumption (vs. DRAM)
    • High throughput, Low latency, Low energy consumption (vs. HDD)
  – Various Flash Devices
    • Throughput: 200MB/s ~ over 1GB/s
    • IOPS: 10,000 ~ over 100,000
    • Interfaces: PCI-e attached, SATA3, mSATA, m.2, etc.
Reliability Evaluation of Flash Devices

- **Power on Hours**
  - Should support for operation duration
  - 4～5 years

- **Endurance**
  - Peta Byte Written (PBW), Tera Byte Written (TBW)
  - Deterioration Progress (%) = Total Write / PBW or TBW

- **Average Erase Count (AEC)**
  - Upper limit of erasure count (EC) is determined for a flash device
  - Deterioration Progress (%) = AEC / Upper Limit of EC

- **Incidence of Bad Blocks**
  - Impact for wear-leveling, GC
Reliability Evaluation of Flash Devices

• Power on Hours
  – Should support for operation duration
  – 4〜5 years

• Endurance
  – Peta Byte Written (PBW), Tera Byte Written (TBW)
  – Deterioration Progress (%) = Total Write / PBW or TBW

• Average Erase Count (AEC)
  – Upper limit of erasure count (EC) is determined for a flash device
  – Deterioration Progress (%) = AEC / Upper Limit of EC

• Incidence of Bad Blocks
  – Impact for wear-leveling, GC
Endurance of Existing Flash Devices

- A: 1.2TB (PCIe) - 17,000
- B: 2.4TB (PCIe) - 15,000
- C: 800GB (SATA3) - 450
- D: 512GB (SATA3.0) - 73
- E: 480GB (mSATA) - 72
- F: 480GB (m.2) - 72

Tera Byte
SSD Configuration on Supercomputers

Flat Buffers (Node Local)
- Client
- SSD
- Parallel File System

Burst Buffers (Intermediate Servers)
- Client
- BB Server
- SSD
- Parallel File System
• How many SSD devices can we use to support the endurance of the volumes during the operation duration?
• We analyze the required endurance based on Lustre I/O monitoring
Lustre I/O Monitoring

- Detailed I/O Analysis and Forecast
- Feedback for User/Job-based Stats for Users

Flexible Definition File

Virtualization tool

Data Analysis

Collectors runs on any type of machines

OpenTSDB
OpenTSDB

Scalable TimeSeries database

Hbase
Lustre I/O Monitoring

• Scalable and Flexible
  – Many Lustre stats, more than, 10M, 100M... stats
    • #OST * #client * stats * sampling
    • Future : #OST * #client * stats * JOBID(UID, etc) * sampling
  – Today, collected client’s exported stats from Lustre servers
    • Lustre-2.1 is still running on server, no jobstats
  – Store data into scalable backend database
    • OpenTSDB and HBase
  – Selectable OpenSource frontend
  – All Lustre version support
    • /proc/fs/lustre structure changes on several lustre versions
    • A shadow definition XML of each version’s /proc/fs/lustre
Endurance Evaluation

• Scheme
  – Estimate Total Write Size [PB] during operation years (1～5 years) based on Lustre Write Data Rate [GB/s] (Avg, Max)
  – Map the Total Write Size to aggregated SSD volumes
  – Evaluate the endurance of the aggregated SSD volumes based on the Total Write Size
TSUBAME2 System Overview

11PB (7PB HDD, 4PB Tape, 200TB SSD)

Computing Nodes: **17.1PFlops**(SFP), **5.76PFlops**(DFP), **224.69TFlops**(CPU), ~**100TB** MEM, ~**200TB** SSD

**Thin nodes**

1408nodes (32nodes x 44 Racks)

- HP Proliant SL390s G7 1408nodes
  - CPU: Intel Westmere-EP 2.93GHz
  - 6cores x 2 = 12cores/node
  - GPU: NVIDIA Tesla K20X, 3GPUs/node
  - Mem: 54GB (96GB)
  - SSD: 60GB x 2 = 120GB (120GB x 2 = 240GB)

**Medium nodes**

- HP Proliant DL580 G7 24nodes
  - CPU: Intel Nehalem-EX 2.0GHz
  - 8cores x 2 = 32cores/node
  - GPU: NVIDIA Tesla S1070, NextIO vCORE Express 2070
  - Mem: 128GB
  - SSD: 120GB x 4 = 480GB

**Fat nodes**

- HP Proliant DL580 G7 10nodes
  - CPU: Intel Nehalem-EX 2.0GHz
  - 8cores x 2 = 32cores/node
  - GPU: NVIDIA Tesla S1070
  - Mem: 256GB (512GB)
  - SSD: 120GB x 4 = 480GB

**Interconnects:** Full-bisection Optical QDR Infiniband Network

- Core Switches
  - Voltaire Grid Director 4700 x 12
  - IB QDR: 324 ports

- Edge Switches
  - Voltaire Grid Director 4036 x 179
  - IB QDR: 36 ports

- Edge Switches /w 10GbE ports
  - Voltaire Grid Director 4036 x 6
  - IB QDR: 34 ports
  - 10GbE: 2port

**GPFS+Tape**

- 2.4 PB HDD + ~4PB Tape

**Lustre**

- Parallel File System Volumes
  - 3.6 PB

**Home**

- Home Volumes
  - 1.2PB

**Home**

- "cNFS/Clusterd Samba w/ GPFS"
- "NFS/CIFS/iSCSI by BlueARC"
Configuration of SSD Volumes

• Configuration
  – Same as TSUBAME2
  – 2 devices per node * 1408 nodes = 2816 devices

• Assumption
  – Written Data are equally distributed to the aggregated SSD volumes
  – The endurance (TBW/PBW) of the aggregated SSD volumes is equal to the aggregation of the endurance of each SSD device
Target Devices

- **A:** 1.2TB (PCIe) - R 1.5 GB/s, W 1.3 GB/s
- **B:** 2.4TB (PCIe) - R 3.2 GB/s, W 2.8 GB/s
- **C:** 800GB (SATA3) - R 500 MB/s, W 450 MB/s
- **D:** 512GB (SATA3.0) - R 540 MB/s, W 520 MB/s
- **E:** 480GB (mSATA) - R 500 MB/s, W 400 MB/s
- **F:** 480GB (m.2) - R 500 MB/s, W 400 MB/s

Tera Byte

<table>
<thead>
<tr>
<th>Tera Byte</th>
<th>A: 1.2TB (PCIe)</th>
<th>B: 2.4TB (PCIe)</th>
<th>C: 800GB (SATA3)</th>
<th>D: 512GB (SATA3.0)</th>
<th>E: 480GB (mSATA)</th>
<th>F: 480GB (m.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17000</td>
<td>15000</td>
<td>450</td>
<td>73</td>
<td>72</td>
<td>72</td>
</tr>
</tbody>
</table>
Aggregate Data Rate for TSUBAME2’s Lustre Volumes

AVG 2.07 GB/s
MAX 16.5 GB/s
Estimated Write Based on Lustre I/O Workload on TSUBAME2

Estimated Write [PB]

Average: 2.07 GB/s

311PB for 5 years
Endurance Based on TSUBAE2 SSD Configuration (Using 2816 devices)

Estimated Write [PB]

Operation Duration [year]

Avg: 2.07 GB/s

Aggregate Endurance of volumes using SSD devices (PBW/TBW)

Better
Endurance Based on TSUBAE2 SSD Configuration (Using 2816 devices)

Device A (1.7 PBW, PCIe)
Device B (1.5PBW, PCIe)
Device C (450 TBW, SATA3)
Device D (73TBW, SATA3)
Device E (72TBW, mSATA)
Device F (72TBW, m.2)

Aggregate Endurance of SSD volumes (PBW/TBW)

Operation Duration [years]
# of Required Devices to Support Estimated Total Write Size during Operation Duration (311PB for 5 years)

<table>
<thead>
<tr>
<th># of Devices</th>
<th>A: 1.2TB (PCIe)</th>
<th>B: 2.4TB (PCIe)</th>
<th>C: 800GB (SATA3)</th>
<th>D: 512GB (SATA3.0)</th>
<th>E: 480GB (mSATA)</th>
<th>F: 480GB (m.2)</th>
</tr>
</thead>
</table>
Aggregate Throughput

Throughput [GB/s]

Throughput

<table>
<thead>
<tr>
<th>Storage</th>
<th>READ</th>
<th>WRITE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 1.2TB (PCIe)</td>
<td>R 1.5 GB/s</td>
<td>W 1.3 GB/s</td>
</tr>
<tr>
<td>B: 2.4TB (PCIe)</td>
<td>R 3.2 GB/s</td>
<td>W 2.8 GB/s</td>
</tr>
<tr>
<td>C: 800GB (SATA3)</td>
<td>R 500 MB/s</td>
<td>W 450 MB/s</td>
</tr>
<tr>
<td>D: 512GB (SATA3.0)</td>
<td>R 540 MB/s</td>
<td>W 520 MB/s</td>
</tr>
<tr>
<td>E: 480GB (mSATA)</td>
<td>R 500 MB/s</td>
<td>W 400 MB/s</td>
</tr>
<tr>
<td>F: 480GB (m.2)</td>
<td>R 500 MB/s</td>
<td>W 400 MB/s</td>
</tr>
</tbody>
</table>
Discussion

• We can configure SSD-based buffer volumes using commodity-based SSD devices
  – Based on Lustre Write I/O workload
  – Several thousand of devices needed

• If we use high end SSD devices (PCI-e attached Flash), we can consolidate the SSD devices
  – Required Performance under the limited endurance of the SSD volumes

• Further evaluation by using more detailed I/O behavior is needed
  – Emulation of Burst Buffers etc.
Summary

• Preliminary Evaluation of Endurance for SSD devices to support supercomputing I/O workloads
  – Based on Lustre I/O monitoring

• High-performance Lustre I/O Monitoring for detailed I/O workload analysis