OPEN.

FOR BUSINESS.
Bryce Canyon - Facebook's Flexible Hard Drive Storage Architecture

Austin Cousineau
Storage Hardware Engineer – Facebook
Agenda

• Flexibility
• Disaggregation
• Overview of the Project
• Why We Built It
• Improvements
• Serviceability
• What’s Next
• Open Source Spirit
• Design Files
How Important is Flexibility?

- Technology is no longer scaling on the historical trends
  - new architectures and optimizations will be needed
- Designing a system for flexibility reduces the need for full re-designs
- Flexibility and modularity allows design re-use
- Utilizing OCP form factors and concepts allows faster adoption of new hardware and concepts
Disaggregation

- Storage capacity is scaling faster than storage throughput
- Increased network capacity and flexibility enable distributed systems
- Dense storage servers with integrated compute allows light weight local services to support disaggregated storage services and resources on separate compute clusters
- Robust network infrastructure combined with an optimized system configuration enables optimal resource utilization and performance on system, with shared resources on the network
What Is Bryce Canyon?

• Our latest disaggregated storage server and JBOD
• 4 OU storage server
  ━ Two distinct storage nodes, each with 36 drives
  ━ Leverages common 1P Compute Server (Mono Lake)
  ━ Uses OCP NICs (OCP Mezz)
• Modular and scalable to meet current and future challenges
Bryce Canyon – Major System Components

- Two storage nodes sharing a drive plane board
  - Logically separated front and back 36 drive sets
  - Separate compute cards, expander cards, and IO modules
Bryce Canyon – Mono Lake & Drive Plane Board

- 92 mm Fans x 4
- Power Entry – Cable Track
- Storage Controller Card
- HDD SAS Connector
- Rear DPB
- Air Baffle
- Front DPB
- Microserver 1P x2
- Input Output Module Connector

- DDR4 DIMMs (4 x 32 GB)
- M.2 SATA Boot Drive
- CPU – Broadwell DE
Bryce Canyon – Storage Controller Card

SCC With IOC & Expander

SAS Expander 12G
SAS IOC 12G

SCC With Expander Only
Bryce Canyon – Input Output Module (IOM)

Top View – M.2

Top View – IOC

Front Panel View

Front Panel View

TPM Module
BMC IC
SAS IO Controller
25 Gb/s OCP NIC
NVMe M.2 Drives x 2
Why Did We Build Bryce Canyon?

• Dense, modular design to accommodate different deployment configurations with a single chassis
• Enhanced serviceability of all major components
• Design reuse by leveraging existing micro server designs
• Improved system performance
• Efficient forced-air cooling for improved CFM/W and service time
• Maintain HDD performance over all operating conditions
How Does It Compare To Honey Badger?

<table>
<thead>
<tr>
<th></th>
<th>Warm Storage</th>
<th>Cold Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Previous Generation</td>
<td>Bryce Canyon</td>
</tr>
<tr>
<td>Compute</td>
<td>Avoton 8 core</td>
<td>Broadwell-DE 16 core</td>
</tr>
<tr>
<td>RAM per Compute</td>
<td>32GB DDR3</td>
<td>64GB DDR4</td>
</tr>
<tr>
<td>HDD per Compute</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>HDDs per Rack</td>
<td>450</td>
<td>576</td>
</tr>
<tr>
<td>SSD Slots (M.2) per Compute</td>
<td>1 x M.2 SATA</td>
<td>2 x M.2 NVMe</td>
</tr>
<tr>
<td>Max Network BW per Compute</td>
<td>10Gbps</td>
<td>50Gbps</td>
</tr>
</tbody>
</table>

The Warm Storage version of Bryce Canyon provides ~4x compute, 2x DRAM, consumes 30% less power / HDD, and helps achieve ~50% reduction in CFM/W.
# Hard Drive Performance Improvements

<table>
<thead>
<tr>
<th>Type of Operation</th>
<th>Specification</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous Operation Specification</td>
<td>&lt;5% Max</td>
<td>&lt;5% Max</td>
<td></td>
</tr>
<tr>
<td>Worst Case Continuous Operation</td>
<td>-5% Max</td>
<td>-1% Max</td>
<td></td>
</tr>
<tr>
<td>Non-Sustained Operation Specification</td>
<td>&lt;10% Max, &lt;7% Avg</td>
<td>&lt;10% Max, &lt;7% Avg</td>
<td></td>
</tr>
<tr>
<td>Worst Case Non-Sustained Operation</td>
<td>-99.7% Max</td>
<td>-2.3% Max</td>
<td></td>
</tr>
</tbody>
</table>
Hard Drive Performance Improvements

- Huge improvements in rear HDD degradation
  - More open fan guard
  - Improved fan blade shape
  - Metal honeycomb acoustic attenuator
System Thermal Improvements

- Optimized system mechanicals to increase service time and decrease CFM/W
  - Large improvement in rack level CFM/W (0.122 CFM/W @30 °C) over Honey Badger systems (0.3 CFM/W @30 °C)
  - Service time of 20 minutes in most conditions
  - Improved fans to reduce HDD RV/AV issues
# System Thermal Improvements – Service Time

<table>
<thead>
<tr>
<th>Thermal Solution</th>
<th>Original</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
</tr>
</thead>
<tbody>
<tr>
<td>without any solution</td>
<td>A hole on IOM chassis below NIC heatsink</td>
<td>A hole on IOM chassis below NIC heatsink</td>
<td>A hole on IOM chassis below NIC heatsink</td>
<td>[Reworked CNC DPB Cover] (with gasket to seal up gap between cover and IOM chassis) (Reduce the height of center ribs)</td>
</tr>
<tr>
<td>without any solution</td>
<td>[Reworked CNC DPB Cover] (with gasket to seal up gap between cover and IOM chassis) (Reduce the height of center ribs)</td>
<td>[CNC DPB Cover] (with gasket to seal up gap between cover and IOM chassis)</td>
<td>[Reworked CNC DPB Cover] (with gasket to seal up gap between cover and IOM chassis) (Reduce the height of center ribs)</td>
<td>[New HDD sponge] Remove the original U-shape sponge and add a new rectangular sponge on the bottom side of HDD latch + Mylar sheet under the sponge</td>
</tr>
<tr>
<td>without any solution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Service Time | 193s (3.2 minutes) | 309s (5.2 minutes) | 508s (8.5 minutes) | (>20 minutes) |
Mechanical Improvements

- Increased safety factor on extension rails
  - 198 lb chassis, 325 lb load on front handles
- Improved tool-less latches for most components
- Improved drive plane replacement
- More robust under shock and vibration

Tool Free HDD Replacement

Easy DPB Replacement
Serviceability

- Common service items can be swapped in under 3 minutes
- Less time servicing means lower down time
- Clearly marked modules and Open BMC allow quick diagnosis and replacement

NIC Latches  SSD Latches  SCC Latches
Deployment Flexibility

- Modular design enables easy changes to compute card, NICs, NVMe drives and expander card
- System is configurable for all in one storage server use, JBOD deployment and JBOD head node configuration
  - Supports both warm storage and cold storage
  - System flexibility allows designing of a rack configuration to meet specific application needs
- Modular sub-systems to scale for new media, technologies and performance requirements
What’s Next

- Updating design to accommodate future drive technology
- Updated compute modules
- Improved thermal and mechanical performance
- Increased electrical performance and efficiency
- Incorporate feedback and lessons from mass deployment
Open Source Spirit & Call To Action

- The Open Compute Project is a powerful community for collaboration and idea sharing
- Strong participation yields larger gains in technology and design enablement
- Large scale players can collaborate to steer the industry and suppliers
- Leveraging lessons from each other enables faster design and reduces duplicated efforts
- Join and participate in OCP collaboration calls
Contributions to OCP

- Specification v1.0 and Design Package
  - https://www.opencompute.org/contributions?menu%5Bspec_id%5D=S0147

- Where to Buy

- OpenBMC Github - https://github.com/facebook/openbmc
OPEN.

FOR BUSINESS.