OPEN. FOR BUSINESS.

OCP SUMMIT

Storage Track

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

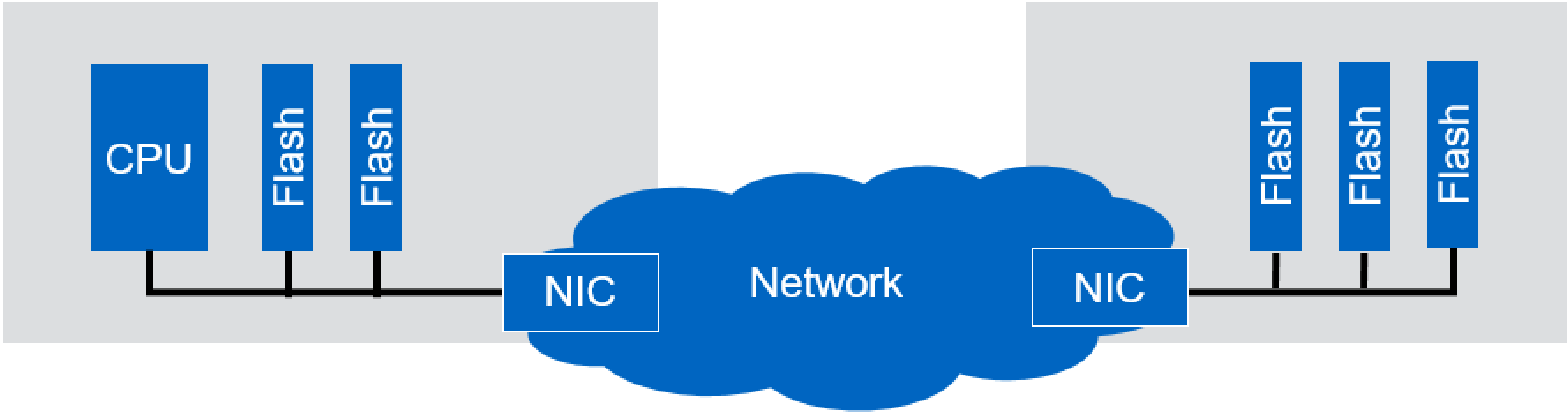
How Important is Flexibility?

- 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

Disaggregated

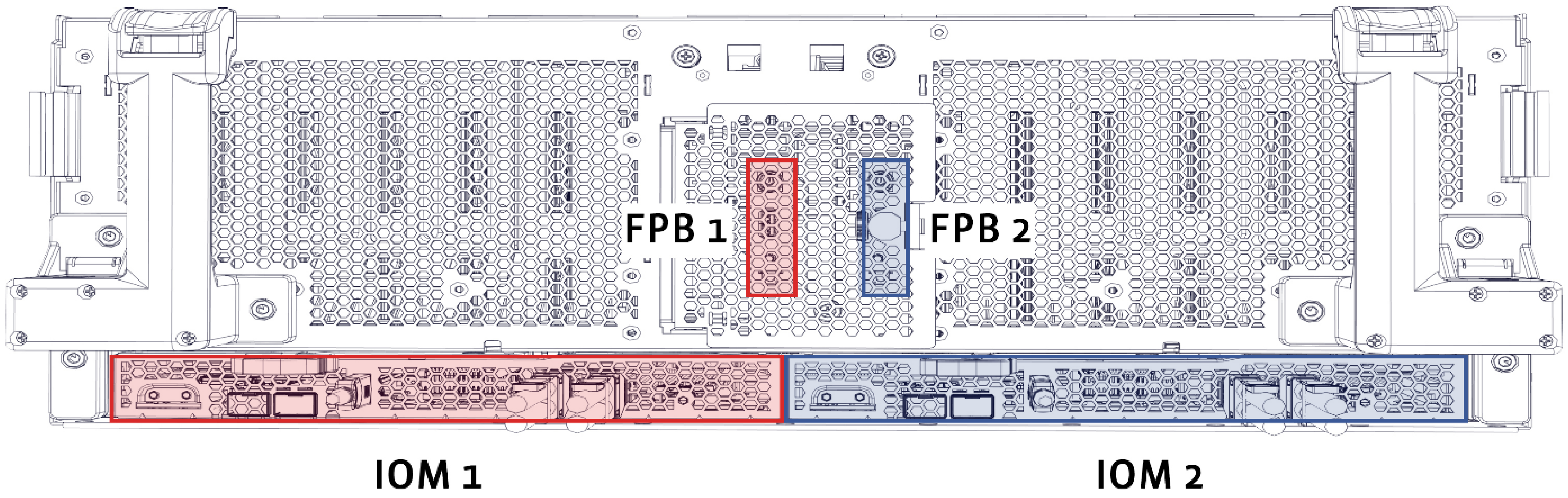


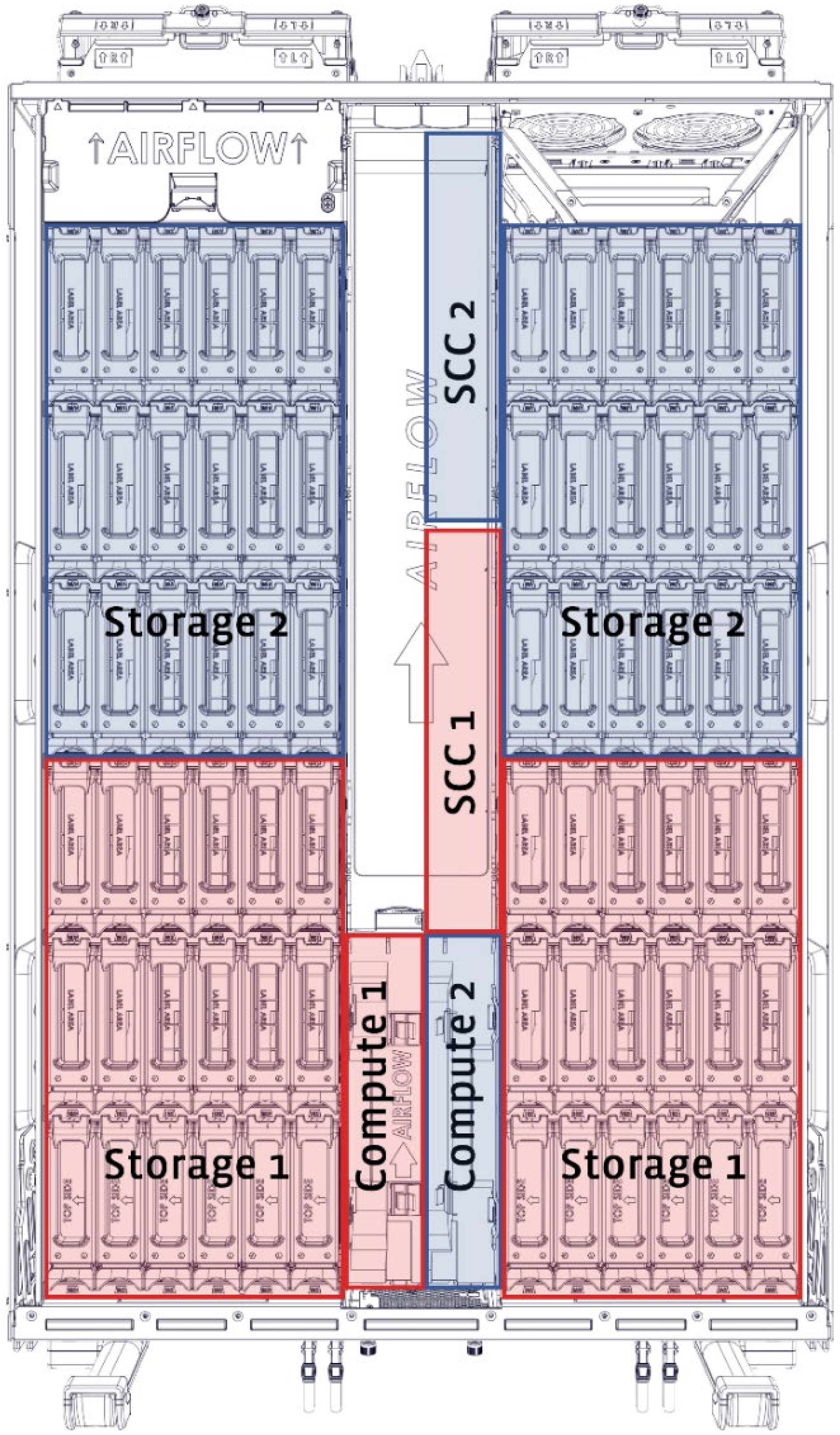
What Is Bryce Canyon?

- 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



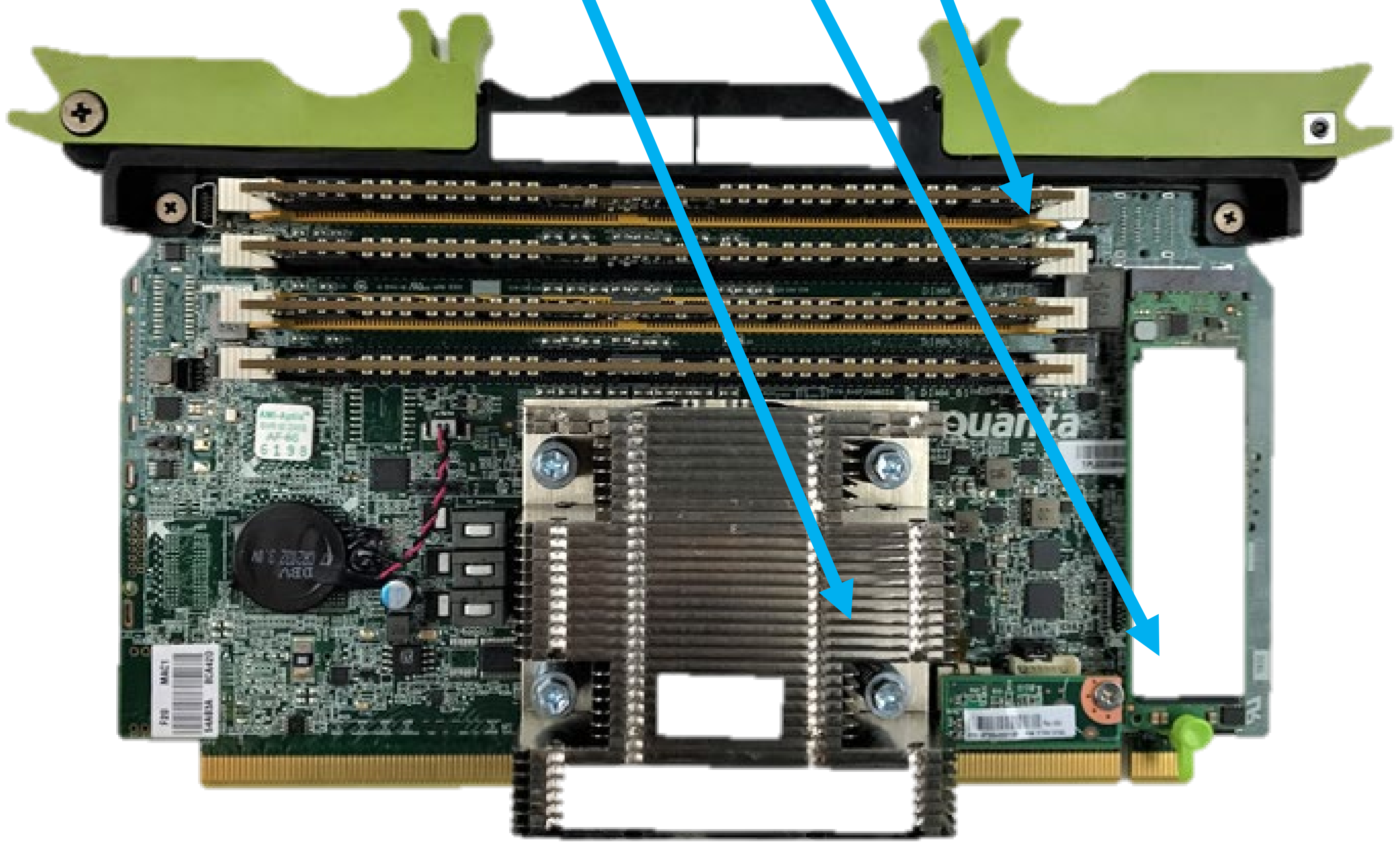


Bryce Canyon – Mono Lake & Drive Plane Board

DDR4 DIMMs (4 x 32 GB)

M.2 SATA Boot Drive

CPU – Broadwell DE



92 mm Fans x 4

Power Entry – Cable Track

Storage Controller Card

HDD SAS Connector

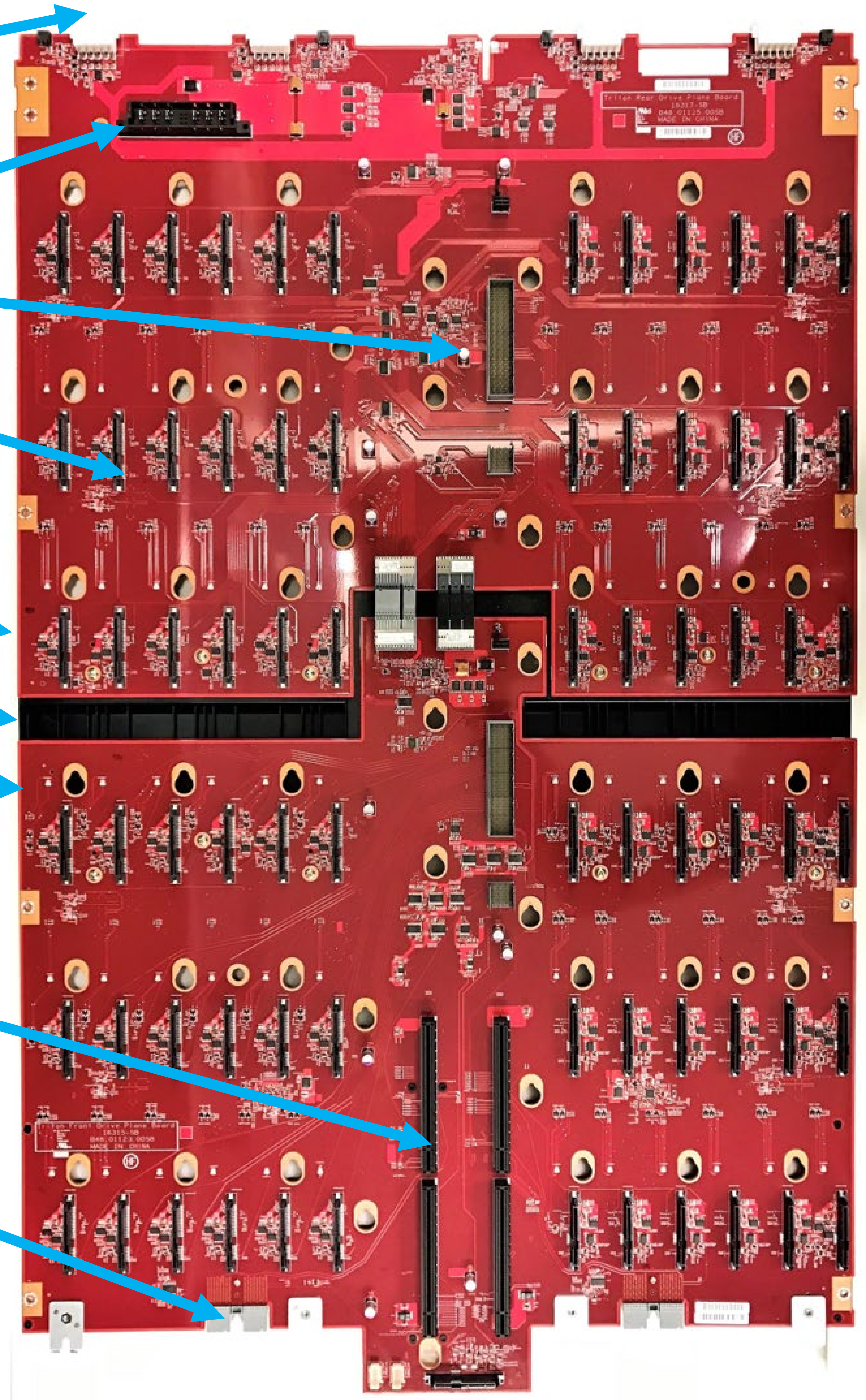
Rear DPB

Air Baffle

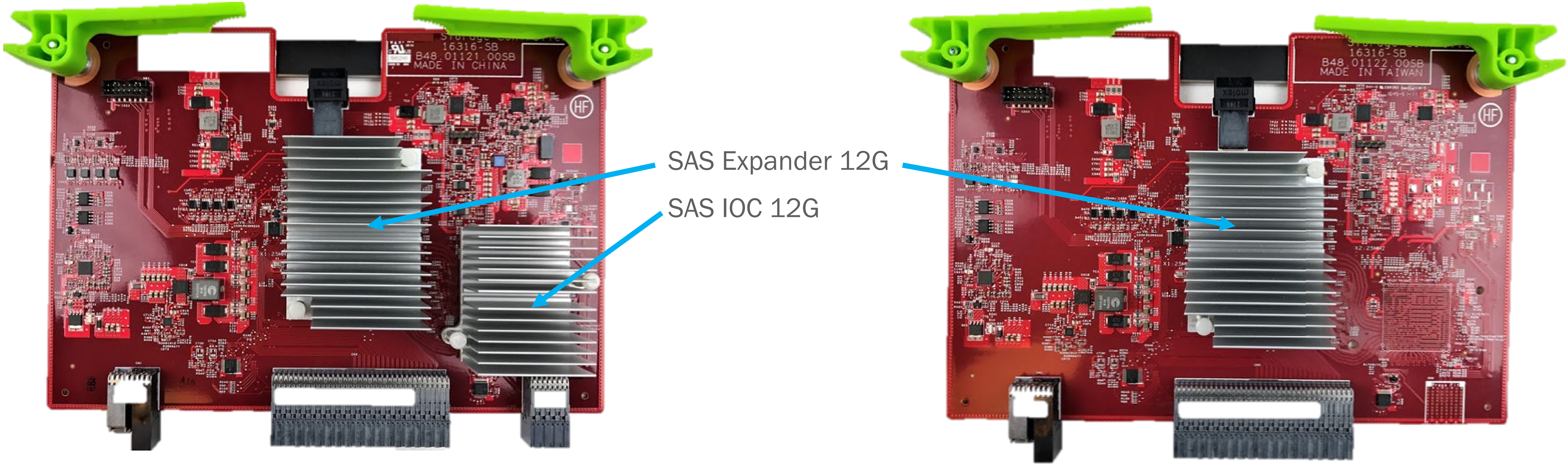
Front DPB

Microserver 1P x2

Input Output Module Connector



Bryce Canyon – Storage Controller Card

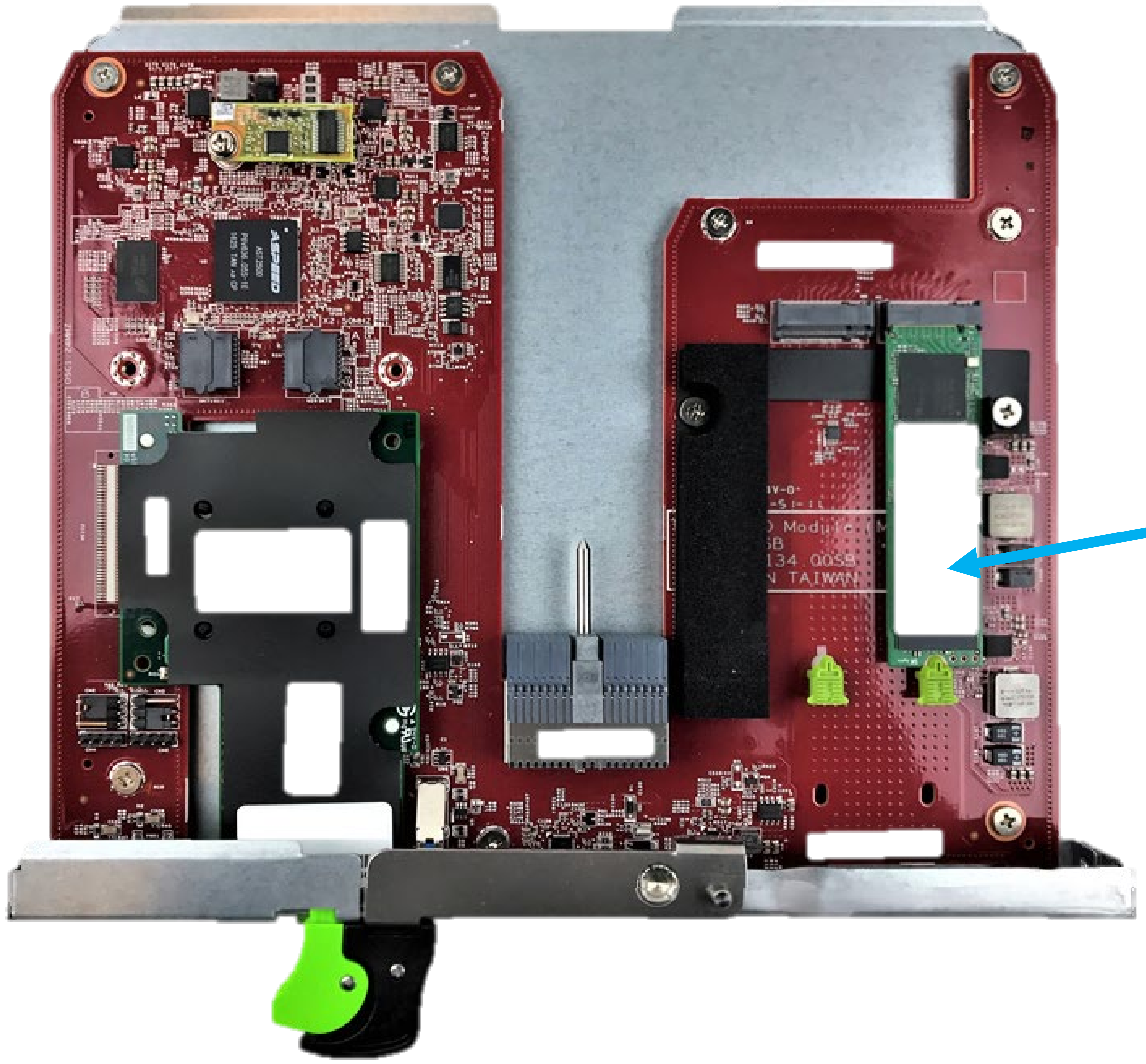


SCC With IOC & Expander

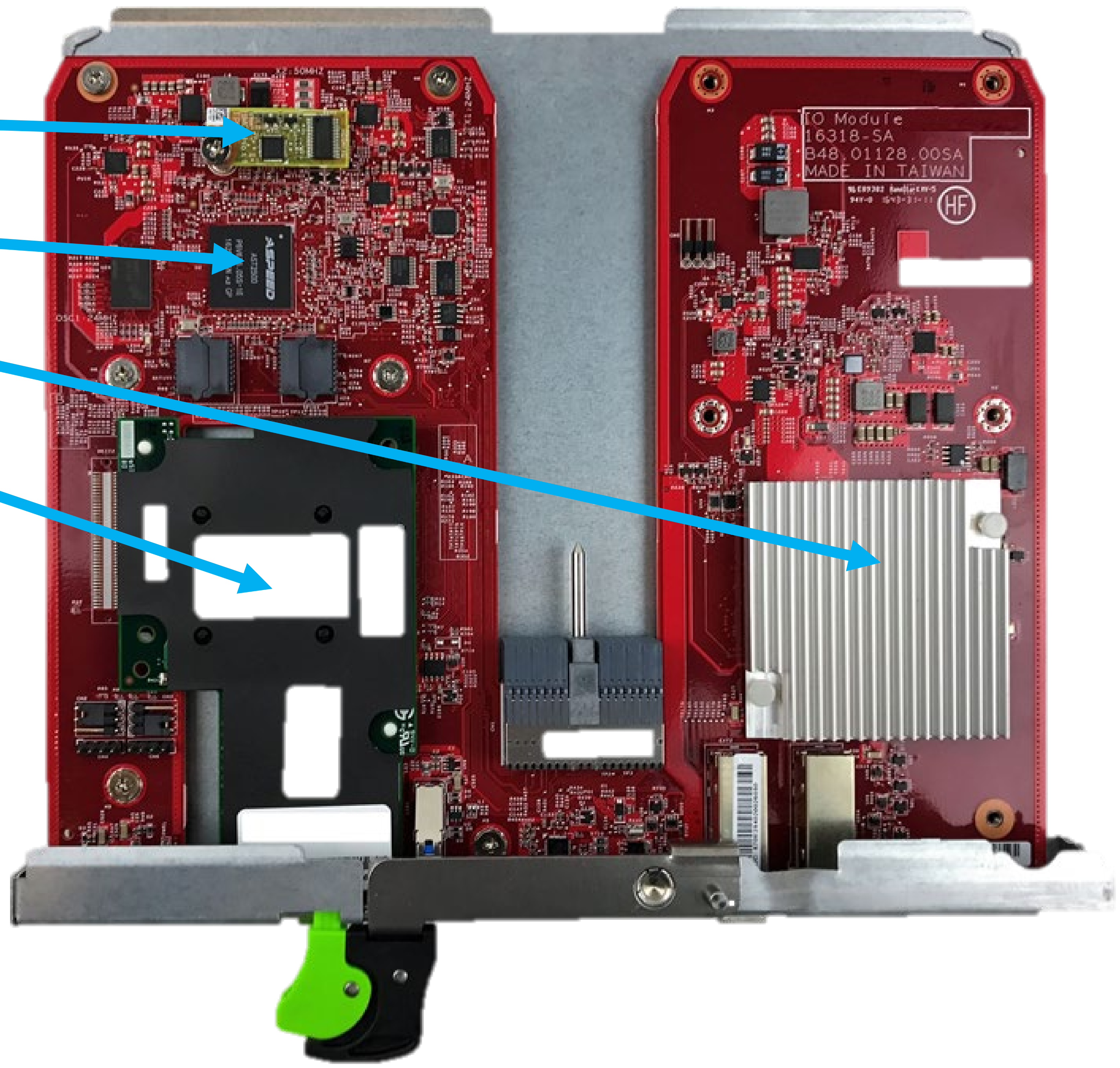
SCC With Expander Only

Bryce Canyon – Input Output Module (IOM)

Top View – M.2



Top View – IOC



TPM Module

BMC IC

SAS IO Controller

25 Gb/s OCP NIC

NVMe M.2 Drives x 2

Front Panel View

Front Panel View

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?

	Warm Storage		Cold Storage	
	Previous Generation	Bryce Canyon	Previous Generation	Bryce Canyon
Compute	Avoton 8 core	Broadwell-DE 16 core	Dual Haswell 12 core	Broadwell-DE 16 core
RAM per Compute	32GB DDR3	64GB DDR4	128GB DDR3	128GB DDR4
HDD per Compute	30	36	240	216
HDDs per Rack	450	576	480	648
SSD Slots (M.2) per Compute	1 x M.2 SATA	2 x M.2 NVMe	0	0
Max Network BW per Compute	10Gbps	50Gbps	10Gbps	50Gbps

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

	Performance Degradation (%)	
	Before	After
Continuous Operation Specification	<5% Max	<5% Max
Worst Case Continuous Operation	- 5% Max	-1% Max
Non-Sustained Operation Specification	<10% Max <7% Avg	<10% Max <7% Avg
Worst Case Non-Sustained Operation	-99.7% Max	-2.3% Max

Hard Drive Performance Improvements

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

Fan Guard

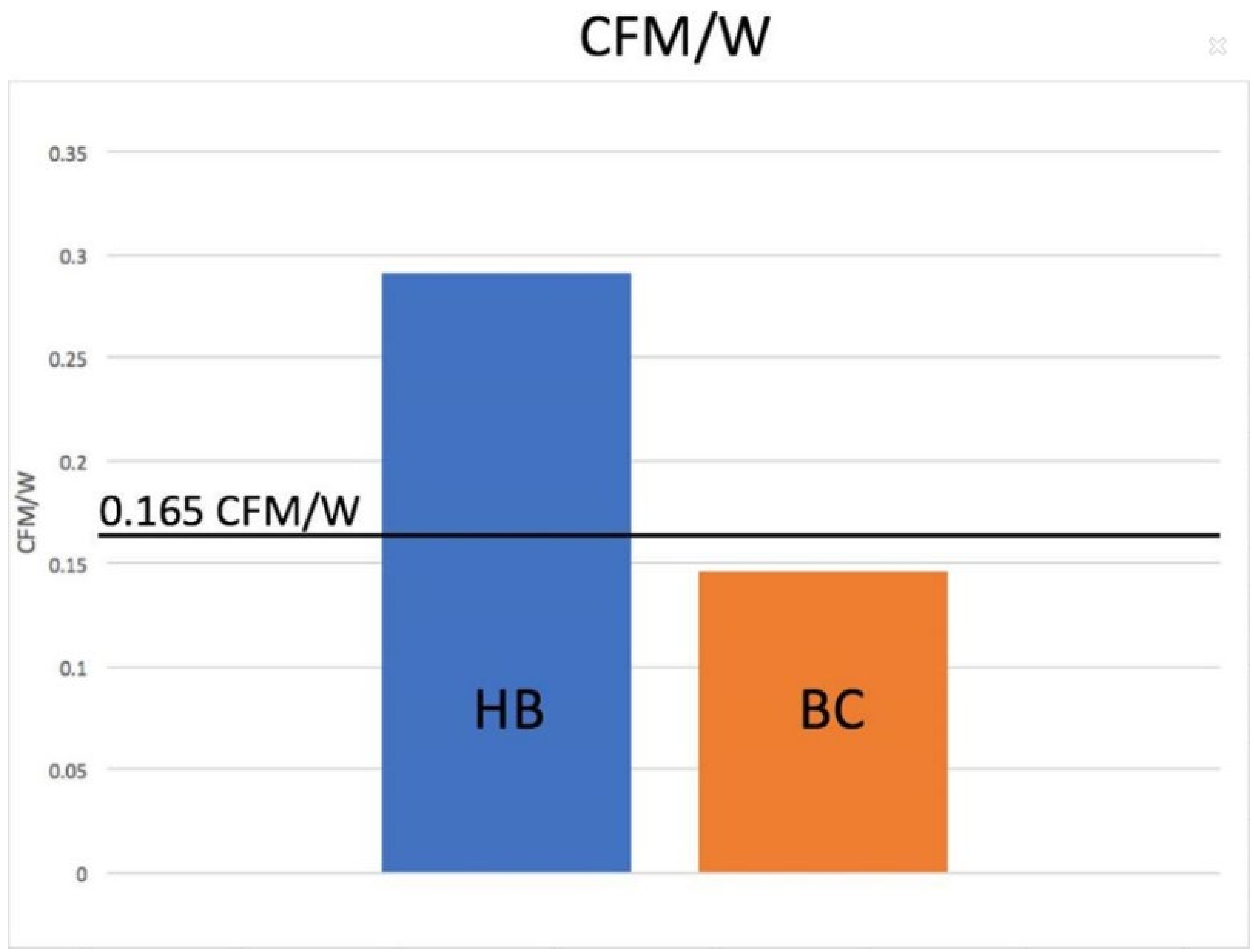
Old Blade Shape

New Blade Shape

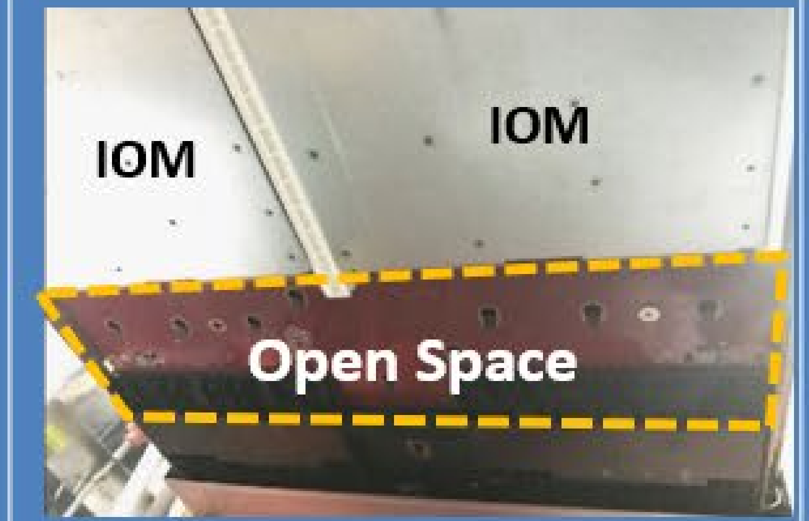
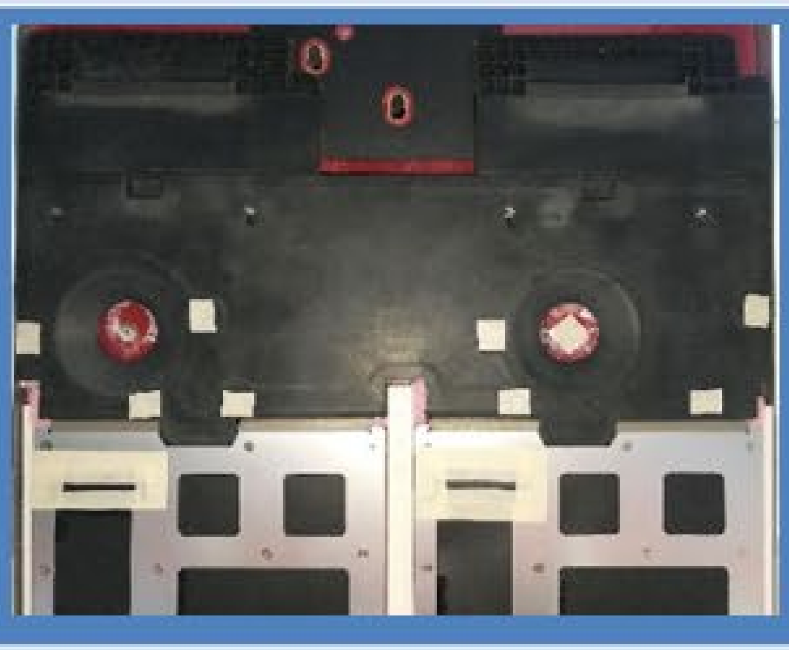
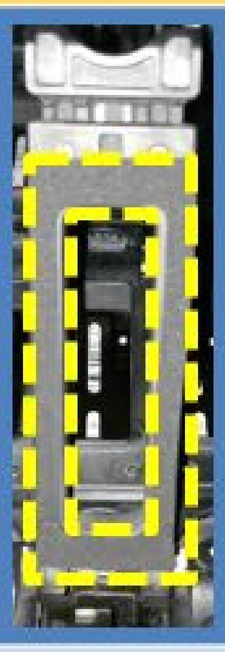
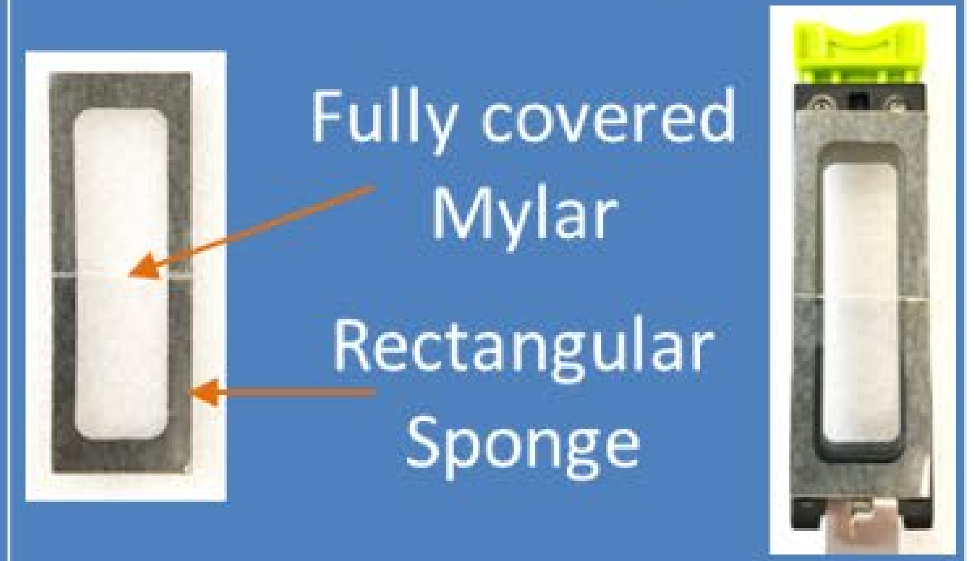
Attenuator

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>
<tr><th rowspan="5">Thermal Solution</th><th>Original</th><th>#1</th><th>#2</th><th>#3</th></tr>
<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></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></tr>
<tr><td>without any solution</td><td></td><td>[New HDD sponge] Remove the original U-shape sponge and add a new rectangular sponge on the bottom side of HDD latch</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>
</td><td></td><td></td><td>
</td></tr>
<tr><th>Service Time</th><td>193s (3.2 minutes)</td><td>309s (5.2 minutes)</td><td>508s (8.5 minutes)</td><td>(>20 minutes)</td></tr>
</table>

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

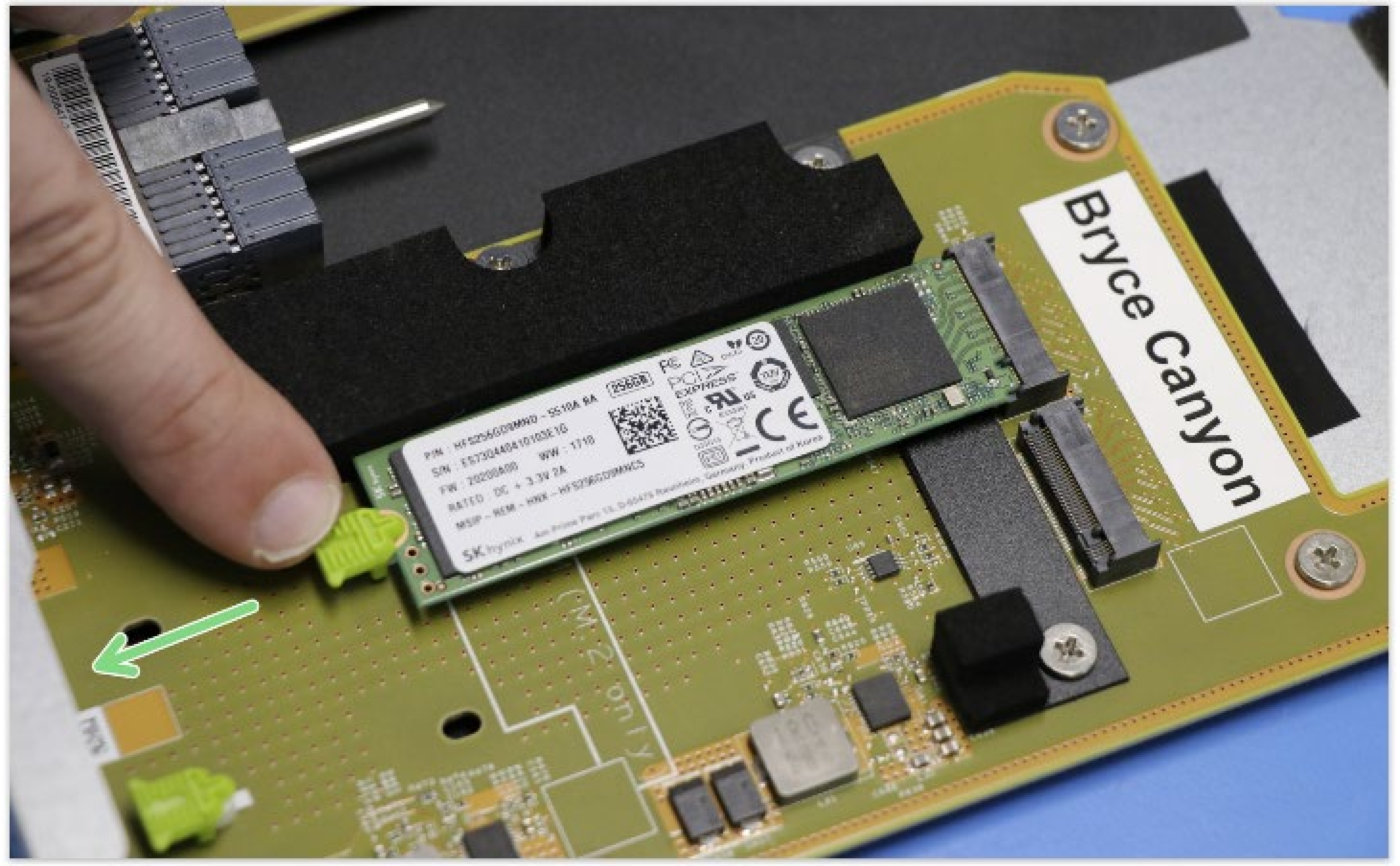
Tool Free HDD Replacement

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

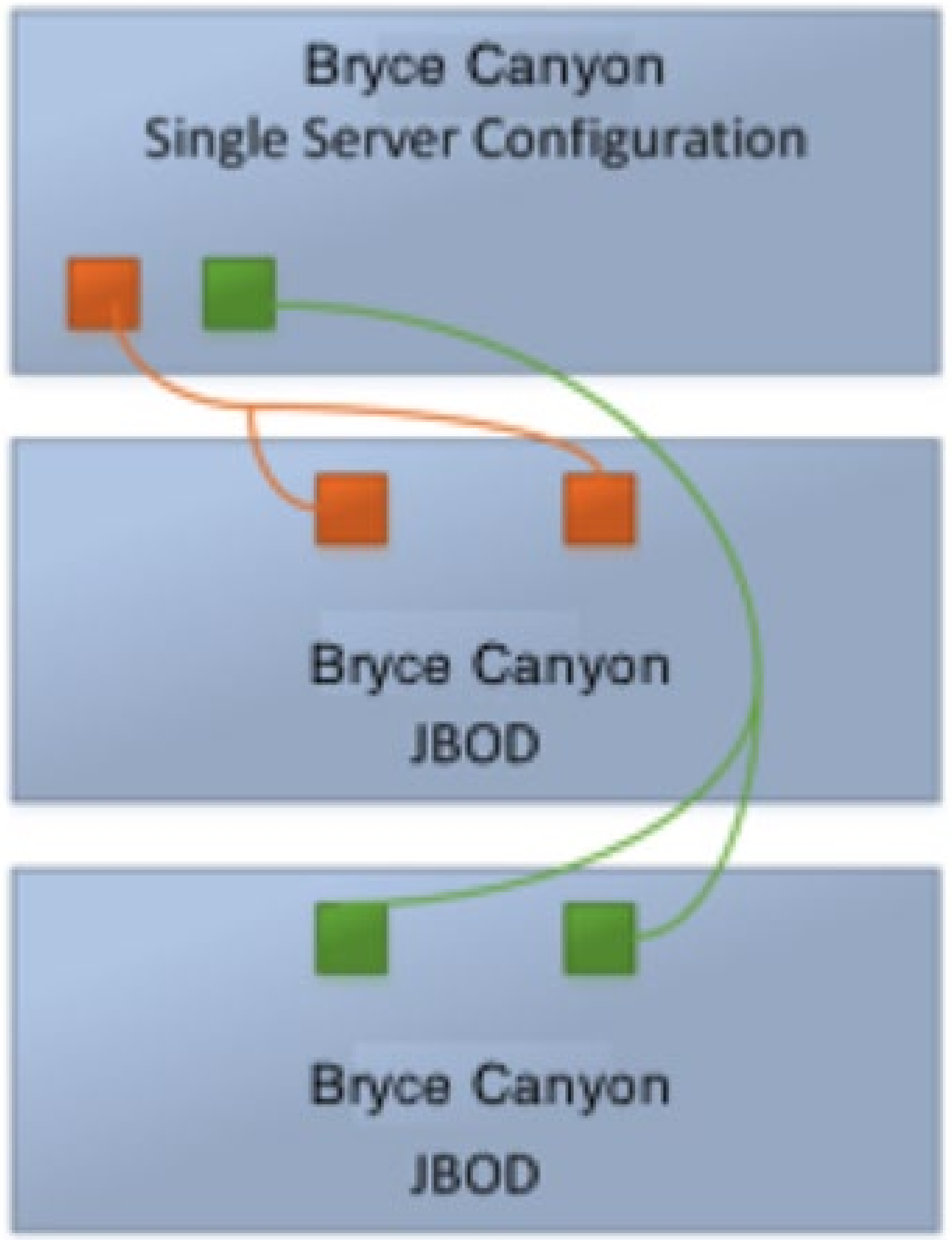


SSD Latches

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

OPEN
Compute Project

Contributions to OCP

- Specification v1.0 and Design Package
  - https://www.opencompute.org/contributions?menu%5Bspec_id%5D=S0147
- Where to Buy
  - https://www.opencompute.org/products/214/wiwynn-bryce-canyon-sas12g-storage-server-comprising-2-server-cards-with-up-to-36-hot-pluggable-hdds-each
- OpenBMC Github - https://github.com/facebook/openbmc

OCP SUMMIT

OPEN.

FOR BUSINESS.