DC-MHS
Datacenter-ready Modular Hardware System

Siamak Tavallaei, Chief Systems Architect, Google Systems Infrastructure
Mark A. Shaw, Sr. Principal Architect, Azure Platform Architecture
DC-MHS
Datacenter-ready Modular Hardware System around DC-SCM and HPM for Hyperstack and DC-Stack

Siamak Tavallaei, Chief Systems Architect, Google
Mark A. Shaw, Sr. Principal Architect, Microsoft
2019: OCP Summit

2019
Articulated four elements for a Modular Building Block Architecture (MBA)

For a successful Modular Building Block Architecture, we need:

- Compute Modules (CPU/Memory/IO) (CMIOM)
- IO & Accelerator Add-in Card Modules (AIC)
- Security, Control, and Management (SCM)
- Data-plane Control
- An Interconnect

2021
The MBA has evolved to:
Open Accelerator Infrastructure (OAI) and Datacenter-ready Modular Hardware System (DC-MHS)
Datacenter-ready Modular Hardware System (DC-MHS)

OCP Server Project Monthly Call Presentation on DC-Stack (5/26/2021)
for Enterprise, Hyperscale, and Edge datacenter
Based on the current DC-SCM effort, our goal has been:

- Streamline the producer-to-consumer pathway
- **Win-win**: allow faster delivery of products into Hyperscaled, Enterprise, and Edge datacenters
- Reduce the complexity of providing a common management and security infrastructure into datacenters
- Increase the value-add and diversity of compute, storage, and IO elements that the suppliers may deliver into the products that Hyperscalers and Enterprise customers may consume
- While driving a standard for the interface to the HPM, limit the impact to the HPM; allow different instances of DC-SCMs for one or many HPM types (either directly or via an Interface Board)

Use the OCP legal framework for multi-party CLA based on OWFa to produce the **Base Specification**
Use appropriate framework for multi-party engagement for **Design Specifications** for various Modules

Each participant will contribute a portion:
- New Technologies
- Spec Chapters
- Program management
- PoC system
- Software, firmware, testbench, ...

**OPEN POSSIBILITIES.**
Alignment with other OCP Activities

The following figure depicts where Datacenter-ready Integrated System (DC-Stack) falls within the continuum from DC-SCM through the datacenter-level Cloud Service Model initiative within OCP.
Alignment with Other OCP Activities

We will align this system-level activity with the foundation we are building within OCP at the module level and deliver an integrated solution for others’ contribution at the datacenter level:

- **DC-SCM**: Starting with DC-SCM and DC-SCI specifications (an OCP subproject)
- **DC-XPI**: For interfacing Host Processor and Memory Module (HPM) to Modular IO (DC-MIO)
- **Modular Hardware System**: The system around DC-SCM, DC-XPI, and HPM and extend to Expansion Chassis such as storage and GPU/Accelerators
- **Datacenter-ready Integrated System (DC-Stack)** *(the effort outlined in this document)*: Add Software and Security apparatus to the Modular Hardware System
- **Open System Firmware** (OSF: an OCP Project)
  - Conforms to OSF 1.2 requirements to support owner control, circular economy
- **Security** (an OCP Project)
  - Implement “Gold” level Security as defined in the Composable Security Architectures
- **Test & Validation**: Accommodate Qualification and Certification (driving a standard diagnostics framework)
- **Benchmarking**: Allow standard benchmarking
- **Cloud Service Model** (an OCP Future Technology Initiative): Deliver the DC-Stack to the OCP Cloud Service Model (CSM) team for datacenter-level life-cycle management
**DC-Stack Vision:**
Streamline the producer-to-consumer pathways
Win-win: allow faster delivery of products into datacenters

**Open ecosystem**
Consumable by hyperscalers, testable by suppliers

Requirements expressed as:
- **Modular hardware**, enabling a vendor to build a base solution for multiple datacenters
- **Modular software**, with open-source reference implementation
- **Validation suite** certifying satisfaction of End Customers

![Diagram with sections for Validation, Reference SW, Reference HW, Continuous Integration, and Open Source SW]
Scaling to Handle Diversity

Hyperscaled DC
- Bare Machine Semantics
- Machine Management APIs
- Platform Security
  - Modular HW
  - Modular SW Stack
  - Test & Validation

Edge/Telco
- app-specific engineering
- Storage
- IO Mix
- Accels
- Power
- Thermal
- Chassis

open possibilities.

...
Enable diversity above the hyperscaled-optimized baseline
- Mech/thermal/power for Edge and Enterprise are different
- Different IO mix and flexibility: front vs back IO, disaggregation, etc..
- Machine size: mission-critical 8S/16S, DC 2S/1 (large) S, Edge 1 (small) S

Conquer with common baseline of requirements & reference implementation
- **Bare Machine** -- separate the customer from platform management
- **Platform Security** -- firmware integrity & control, physical protection of data confidentiality
- **Machine Management** -- telemetry & actuation for inventory and repairs

Historically focused on the hyperscaled datacenter optimizations for Vertical integration:
- Custom thermal, power, mech, security, machine management, SW stack
- Deliver maximal TCO/cycle at global DC scale, for internal and cloud customers
Datacenter-ready Integrated System (DC-Stack)

- Datacenter-ready Integrated System for Edge, Private Cloud, and Large Datacenters
  - HW, FW, SW, management, at-scale debug, security, and test & validation

- Built on successes within OCP efforts such as modular DC-SCM+HPM and OAM/OAI
  - Articulate one complete modular system (*The Base Specification*) for each solution category
  - Allow variations at each module (multiple *Design Specifications* based on the Base Spec)
  - Work with suppliers to build products (PCBA, Chassis, etc. based on Design Specs)

- Datacenter-ready Modular Hardware System (DC-MHS) (DC-SCM + HPM + DC-MIO + Modular Power)
  - DC-SCM (BMC, RoT, CPLD)
  - HPM (CPU/Memory/IO Slots)
  - Representative firmware for RoT and BMC (refer to the software strategy slide)
  - Modular IO (DC-XPI): Spec, cable/adapter prototypes

- Rack-level specifications (DC requirements: Mechanical, Power, Cooling, Weight, EMI, Acoustic, ...)
- Rack Manager Interface
- Contribute a reference design
  - Mechanicals (new enclosure which fits Open Rack and 1RU/2RU Blades)
  - Generic motherboard requirements (not secret sauce!)
  - Contribute the Base Specification to OCP (generic system)
  - Suppliers will contribute Design Specifications and build Products
Hyperstack Hardware Modules:
Logical Blocks overlaid on Physical Blocks for a Datacenter-ready Integrated System (DC-Stack)

DC Environmental Requirements
- Mechanical
- Power/Cooling
- EMI/Acoustic
- At-scale Debug
- Physical Security
- Management

DC-SCM
- OoB Control
- BMC
- RoT
- Partners go here
  - Internal
  - e.g., HPE’s iLo
  - Dell’s iDRAC
  - Lenovo
  - … others

HPM (baseboard)
- Compute
- Security & Control sidebands
- DC-SCI
- USB / I3C / 1xPCIE
- HPM goes here
- Partners go here
- OCP tracks go here
- 1S, 2S, 4S, 8S CPUs
- Xeon, EPYC, ARM64, …
- … others

DC-XPI (modular IO)
- Interconnect
- IO & Accelerators
- Form Factors (details here)
- NVMe requirements
- RoT requirements
- Cables & Interfaces
- OCP tracks go here
- SSD
- IB NICs
- Accelerators
- … others
- … others

SmartNIC
- Dataplane Control
- SmartNIC
- Partners go here
  - SmartNIC
  - e.g., Blue Field, Stingray,
    - AWS Nitro,
    - MSFT FPGA,
    - … others
  - … others
Hyperstack Software Components

Validation CUJs
- Upgrade
- Attest
- Inventory
- Load Test Fault Injection
-...
DC-Stack Compliance Suite

Upgrade
- Remote Firmware Upgrade/Downgrade (PLDM)
- Voltage/Current
- Memory Error Reporting
- Verify Locked/RO Firmware
- Redfish BMC Inventory (CIM)
- Kernel Panic/kexec/kdump
- Core Freq/Thermal
- Network Performance
- Power Stepping

Thermals/Power
- Fans/Liquid Cooling
- Soft Repair Support
- Disk Lock/Unlock Encryption
- FRU Reporting Compliance (CIM)
- Memory Bandwidth/Latency
- PCIe Feature Set

Get RAS
- ASPM State Test
- PCIe AER Support
- Remote Attestation Running vs. Static
- SMBIOS Device/CPU/Slot/Chassis/DIMM Checks
- PCIE Error Injection

Attest/Security
- Reset/Reboot Tests
- Platform MCE->BMC
- Remote Attestation Running vs. Static
- IPMI Power/PnP/CPU Interlink Checks
- Firmware Attestation Recovery
- CPU Interconnect Bandwidth/Latency
- Storage Performance

Inventory
- SEL Persistence/BMC Resiliency
- BIOS/BMC “Fuzzer”
- CPU Feature Checks (BootGuard, AMD HVB, disable DCI/Tap, etc)
- DXE Inventory/Checks

Error Resilience/Recovery
- FRU Reporting
- DIMM Error Injection
- Firmware Attestation Recovery
- DIMM Error Injection
- CPU Interconnect Bandwidth/Latency
- Storage Performance

Performance & Load Test
- Platform MCE->BMC
- Remote Attestation Running vs. Static
- BIOS/BMC “Fuzzer”
- CPU Feature Checks (BootGuard, AMD HVB, disable DCI/Tap, etc)
- DXE Inventory/Checks

Platform Features
- Sel Persistence/BMC Resiliency
Datacenter-ready Integrated System *(DC-Stack)*
A convolution of many essential ingredients
Progress Status
Implementations
Example of a front I/O server using Modular I/O w/ vertical DC-XPI connectors (and DC-SCM).

HPM (mobo) PCB pulled back from front of chassis.

Front volume has been divided into four I/O “bays”.

DC-SCM (vertical style)
Implementations (cont’d)

1x16 CEM cabled I/O Adapter
2x8 CEM cabled I/O Adapter
Allows for riser-based I/O Adapters, as well

Multiple vertical DC-XPI connectors across front of HPM

Open Possibilities.
Implementations (cont’d)

Two 1x16 Cabled CEM I/O Adapters in an I/O Module
(top view)
DC-XPI Status

The DC-XPI 1.0 spec has been largely completed for productization in 2022.

Similar to DC-SCM 1.0, we hope to gather support and feedback from OCP members which could lead to a second iteration of the spec, i.e., DC-XPI 2.0.

We are targeting the **DC-XPI** 2.0 spec for use in 2023+ servers, coincident with the **DC-SCM** 2.0 and **DC-MHS** 1.0 specs for the datacenter-ready integrated system of **DC-Stack** 1.0.
DC-SCM 1.0 Designs

FPGA-based DC-SCM 1.0 prototypes for **LibreBMC**:  
Atmicro Blog: [https://antmicro.com/blog/2021/07/dc-scm-open-hardware-for-fpga-bmc/](https://antmicro.com/blog/2021/07/dc-scm-open-hardware-for-fpga-bmc/)

Designs can be found at;  
Based on Xilinx Artix-7 FPGA: [https://github.com/antmicro/artix-dc-scm](https://github.com/antmicro/artix-dc-scm)  
Based on Lattice ECP5 FPGA: [https://github.com/antmicro/ecp5-dc-scm](https://github.com/antmicro/ecp5-dc-scm)
Call to Action

- Adopt the Modular Building Block Architecture (MBA) using DC-SCM and DC-XPI specifications as the base. They are enabling high-volume designs going into production; take advantage of them in your new designs.
  - DC-XPI specification is available at: DC-XPI rev. 0.9 specification (1.0 soon to be released)
  - DC-SCM 1.0 specification is available at: DC-SCM 1.0 Specification_Released to OCP

- DC-SCM 2.0 specification is currently in revision 0.7; provide feedback to make it better for 2023+ products.
  - Find it at Hardware Management Module Subgroup: https://www.opencompute.org/wiki/Hardware_Management/Hardware_Management_Module

- Stay tuned for Datacenter-ready Modular Hardware System (DC-MHS) and the Datacenter-ready Integrated System (DC-Stack) specifications built around DC-SCM

OPEN POSSIBILITIES.
OPEN POSSIBILITIES.
DC-XPI Slides
DC-XPI
Datacenter-ready eXtended Peripheral Interface

Mike Branch, H/W Engineer, Google
Nilesh Dattani, H/W Engineer, Microsoft

OPEN POSSIBILITIES.
2019: OCP Summit

For a successful Modular Building Block Architecture, we need:

- Compute Modules (CPU/Memory/IO) (CM.IO.M)
- IO & Accelerator Add-in Card Modules (AIC)
- Security, Control, and Management (SCM)
- Data-plane Control
- An Interconnect

AIC Attachment
IO Slot to CPU Board Cable Harness

2021: The MBA has evolved to:
Open Accelerator Infrastructure (OAI) and Datacenter-ready Modular Hardware System (DC-MHS)
Datacenter-ready Modular Hardware System

An overview from: OCP Server Project Monthly Call Presentation on DC-Stack (5/26/2021) for Enterprise, Hyperscale, and Edge datacenter

Hyperstack Hardware Modules:
Logical Blocks overlaid on Physical Blocks for a Datacenter-ready Integrated System (DC-Stack)

SERVER

DC Environmental Requirements
- Mechanical
- Power/Cooling
- EMI/Acoustic
- At-scale Debug
- Physical Security
- Management

DC-SCM (OoS Control)
- BMC
- RoT

HPM (baseboard) Compute
- Security & Control sidebands (DC-SCI)
- USB / I2C / 1xPCIe

CPUs or GPUs, TPUs, xPUs

DIMMs
- OCP tracks go here
- 1S, 2S, 4S, 8S CPUs
  - Xeon, EPYC, ARM64, ...
  - xPU Expansion Chassis
  - ... others

DC-MIO (modular IO)
- Interconnect
- IO & Accelerators
  - NVMe requirements
  - RoT requirements

SmartNIC (Dataplane Control)
- Cables & Interfaces
  - DC-XPI (eXtended Peripheral Interconnect)
- Partners go here
  - SmartNIC
  - e.g., Blue Field, Stingray
  - AWS Nitro
  - MSFT FPGA, ...
  - ... others
Why I/O Modularity?

● Interface speeds have been increasing
  ○ Increasing mobo material costs and/or
  ○ Increasing need for re-timers
● Higher power peripherals (requiring additional cabling)
● Increasing # of peripheral shapes to support (CEM, U.2, EDSFF, custom, …)
● Desire for “pay-as-you-go” addition of peripherals
● Increasing # of server platforms; desire to reduce validation time & effort
Datacenter-ready Modular I/O (DC-MIO)

- Packaging approach that separates the motherboard (HPM\(^1\)) from the I/O peripherals
- Allows high-speed I/O connector(s) near the CPU(s)
- Uses I/O Adapters to connect peripherals to the HPM
- System cost reduction opportunities:
  - Reduces motherboard size & cost
  - Allows for cabled and riser-style I/O Adapters
  - Cabled I/O adapters may eliminate need for retimers
- Accommodates multiple peripheral form factors
- I/O Adapters can be installed as-needed based on tray config

\(^1\)Host Processor/Memory Module
Implementation Goals for DC-XPI 1.0

How should this modular interface be implemented?

Goals:
- A high-speed (up to PCIe Gen6), high-density connector
- A high-volume connector with multiple sources
- Cable and riser-card support
- Support for x16 (not too concerned with optimizing for smaller width connectors)
- Support (12V) higher-power peripherals without additional cables
- Support a flexible set of sideband interfaces, supporting a wide range of standard peripherals
- Re-use existing high-volume connector and pinout if possible
- Support flexible mounting orientations: vertical/horizontal/coplanar (1U/2U/...)

OPEN POSSIBILITIES.
An Implementation

Datacenter-ready eXtended Peripheral Interface (DC-XPI 1.0)

- SFF-TA-1002 4C+ connector provided the desired speed, density and pin count
  - PCIe Gen6, 0.6mm/<3” length, x16 + sidebands
- Connector already has volumes being driven by OCP NIC & DC-SCM
- Allows for cabled and riser-style I/O adapters
- Created a pinout that supports high power (150W) peripheral(s)
  - Supports 2x 75W CEM cards
- Optional (separate) auxiliary power block to support up to 400W peripheral(s)
- Rich set of sideband interfaces including USB2, USB3, UART, I2C
- Supports individual Presence Detect for I/O Adapter and Peripheral
A New Pinout for 4C+?

Several existing pinout/connector options, including:

- EDSFF / PECFF (4C)
- PECFF-HP-12V (4C)
- OCP NIC 3.0 / PECFF (4C+)

4C+ connector meets most goals, but existing pinouts don’t support:

- High power (150W) peripherals without additional power cables
- A rich set of sideband interfaces including USB2, USB3, and UART
Implementations
Example of a front I/O server using Modular I/O w/ vertical DC-XPI connectors (and DC-SCM).

- Front volume has been divided into four I/O “bays”.
- HPM (mobo) PCB pulled back from front of chassis.

DC-SCM (vertical style)
Implementations (cont’d)

1x16 CEM cabled I/O Adapter
2x8 CEM cabled I/O Adapter
Allows for riser-based I/O Adapters, as well

Multiple vertical DC-XPI connectors across front of HPM

(top view)
Implementations (cont’d)

Two 1x16 Cabled CEM I/O Adapters in an I/O Module
(top view)
Progress Status

**DC-SCM 1.0** specification is available. It is enabling high-volume designs going into production; take advantage of it in your new designs.

The **DC-XPI 1.0** spec has been largely completed for productization in 2022.

**DC-SCM 2.0** specification is currently in revision 0.7; provide feedback to make it better for 2023 products. Find it on the [Hardware Management Module Subgroup Wiki Page](#).

Similar to DC-SCM 1.0, we hope to gather support and feedback from OCP members which could lead to a second iteration of the spec, i.e., DC-XPI 2.0.

We are targeting the **DC-XPI 2.0** spec for use in 2023+ servers, coincident with the **DC-SCM 2.0** and **DC-MHS 1.0** specs for **DC-Stack 1.0**.
Call to Action

• Adopt the Modular Building Block Architecture using **DC-SCM** and **DC-XPI** as the base. They are enabling high-volume designs going into production; take advantage of them in your new designs.
  ○ DC-XPI specification is available at: [DC-XPI rev. 0.9 specification](#) (1.0 soon to be released)
  ○ DC-SCM 1.0 specification is available at: [DC-SCM 1.0 Specification Released to OCP](#)

• DC-SCM 2.0 specification is currently in revision 0.7; provide feedback to make it better for 2023 products.
  ○ Find it at Hardware Management Module Subgroup: [https://www.opencompute.org/wiki/Hardware_Management/Hardware_Management_Module](https://www.opencompute.org/wiki/Hardware_Management/Hardware_Management_Module)

• Stay tuned for Datacenter-ready Modular Hardware System (**DC-MHS**) and the Datacenter-ready Integrated System (**DC-Stack**) built around DC-SCM

**OPEN POSSIBILITIES.**
OPEN POSSIBILITIES.
DC-SCM 1.0 Update

Priya Raghu, Sr. HW Engineer, Microsoft
prraghu@microsoft.com

Nathan Folkner, HW Engineer, Google
folkinator@google.com
DC-SCM 1.0 Recap

Top Level Block Diagram

- Modularizes management and Security functionality.
- CPU and BMC vendor agnostic
- Scalable 1S, 2S, 4S…GPU, AI
- Standardized connector interface
- Standardized form factors
- Future proof
DC-SCM 1.0 Form Factors

Vertical Form Factor

Horizontal Form Factor
What’s New Since OCP 2020?

• Released OCP DC-SCM Spec to v1.0 (Link)

• Incorporated feedback received over previous iterations of the spec (Thank you for the great feedback!)

• Some major changes
  • Added two additional I2C busses
  • Added a x4 PCIe Gen5 interface for future expansion
What worked well?

- It has enabled us to build smaller/less expensive HPMs by moving the management circuit onto a board with lower cost/area.

- It has decoupled the BMC and RoT implementation from the server, allowing them to innovate and iterate at different rates.

- It has provided us a line-of-sight on having DC-SCM designs across multiple server programs, saving design and validation time.
Challenges

• Pinout and form-factor covers vast majority of use-cases. Some small number of corner cases not supported in DC-SCM v1.0.

• Requires up-front work (Hardware and Firmware) to make DC-SCM design work across multiple HPM architectures. "Plug and Program" still involves work for each server.

• Requires up-front work to enable standard CPLD implementation and Serial GPIO mappings.
Looking Ahead

- **Google**: We see it filling the needs of several upcoming server programs and will continue to use it until OCP DC-SCM 2.0 is finalized and needed to support our designs.

- **MSFT**: Common OCP DC-SCM 1.0 hardware across several of our current generation programs, and current line of sight indicates that we will continue that trend in the future. Actively involved in DC-SCM 2.0 definition at OCP and evaluating it for future designs.
Call to Action

- Adopt the Modular Building Block Architecture using DC-SCM as the base.
  - DC-SCM 1.0 specification is available. It is enabling high-volume designs going into production; take advantage of it in your new designs.
  - DC-SCM 2.0 specification is currently in revision 0.7; provide feedback to make it better for 2023 products. Find it on the Hardware Management Module Subgroup Wiki Page.

- Stay tuned for Datacenter-ready Modular Hardware System (DC-MHS) and the Datacenter-ready Integrated System (DC-Stack) built around DC-SCM 2.0.

- Get involved: OCP-HWMgt-Module@OCP-All.groups.io
OPEN POSSIBILITIES.