Open. Together.

OCP
REGIONAL
SUMMIT

Progress Report on:

SERVER

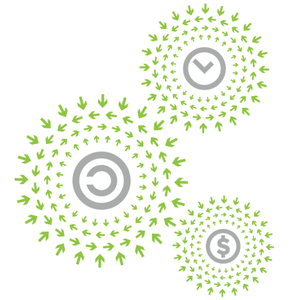
Datacenter-ready Secure Control Module and Interface

for Modular Building Block Architecture (**MBA**, *The Catalyst*)

Siamak Tavallaei, Principal Architect
Microsoft, Azure
OCP Server Project co-Lead

Sept 27, 2019

Specifications



OPEN PLATINUM™

Work-in-Progress

Microsoft & Google

Platform Teams Present

In preparation for an OCP specification, this slide deck is a progress report based on continued feedback received on **DC-SCM**, **DC-SCI**, and **PCIe Slot Cable Assembly.**
It is subject to change without notice.

Feedback from

Lenovo
Wiwynn
Intel
AMD
Dell
Inspur
Quanta
Inventec
Dell
Sanmina
Supermicro
…

Outline

SERVER

- Motivation, Background, and Review
- Progress Report on MBA and on DC-SCM & DC-SCI
- Received Feedback
- Open-source Activities

Motivation

- Open-source Modular approach for faster TTM
- Modeled after well-known interfaces such as PCIe
- Standardizing Common Blocks and Interfaces
- Target interoperability with ease!
  - High-speed Interconnect (PCIe Gen-4 and Gen-5)
  - Datacenter-ready Security, Control, and Management

For a successful Modular Building Block Architecture, we need:

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

Modular Building Block Architecture (**MBA**)

- Is based on small building blocks
  to allow flexible and agile system integration

- Clearly defines input/output ports
  for interoperability with CPU boards from various suppliers

- Riser-based & Cable-based IO Slots
  offer flexibility of choice-- ready for PCIe Gen-4 and Gen-5

MBA is a *Catalyst* for interoperable *Innovation!*

DC-SCM Facilitates MBA

A standards-ready secure control module, ***DC-SCM***, enables the design and deployment of CPU/Memory Complexes and Expansion Chassis to become simply a **routine exercise** based on guidelines from CPU and SoC suppliers!

Examples of

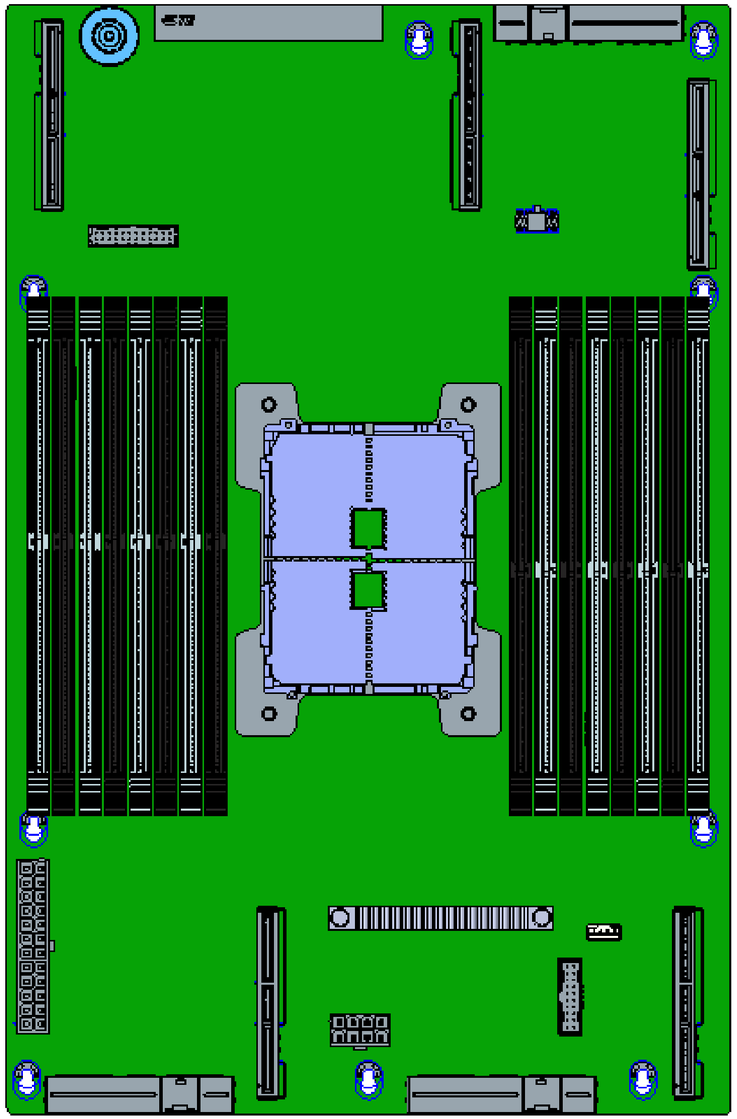
Modular Building Block Architecture

CPU/Mem/IO Module

Just the essential Central Compute Elements

High-speed Memory and
IO Connectors Close to the SoC

Get ready for *PCIe Gen-5!*

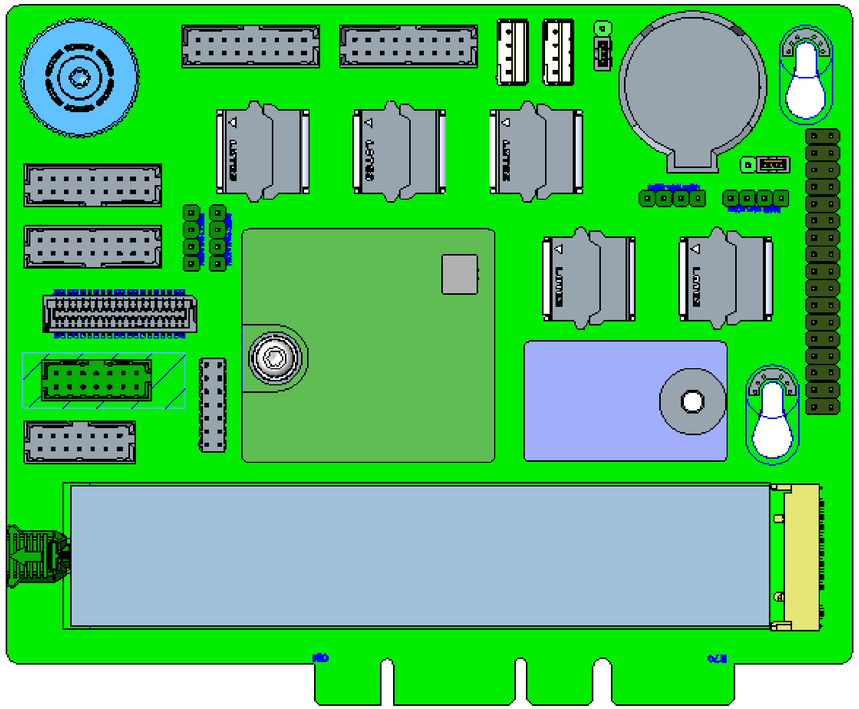


DC-SCM

Everything Else!

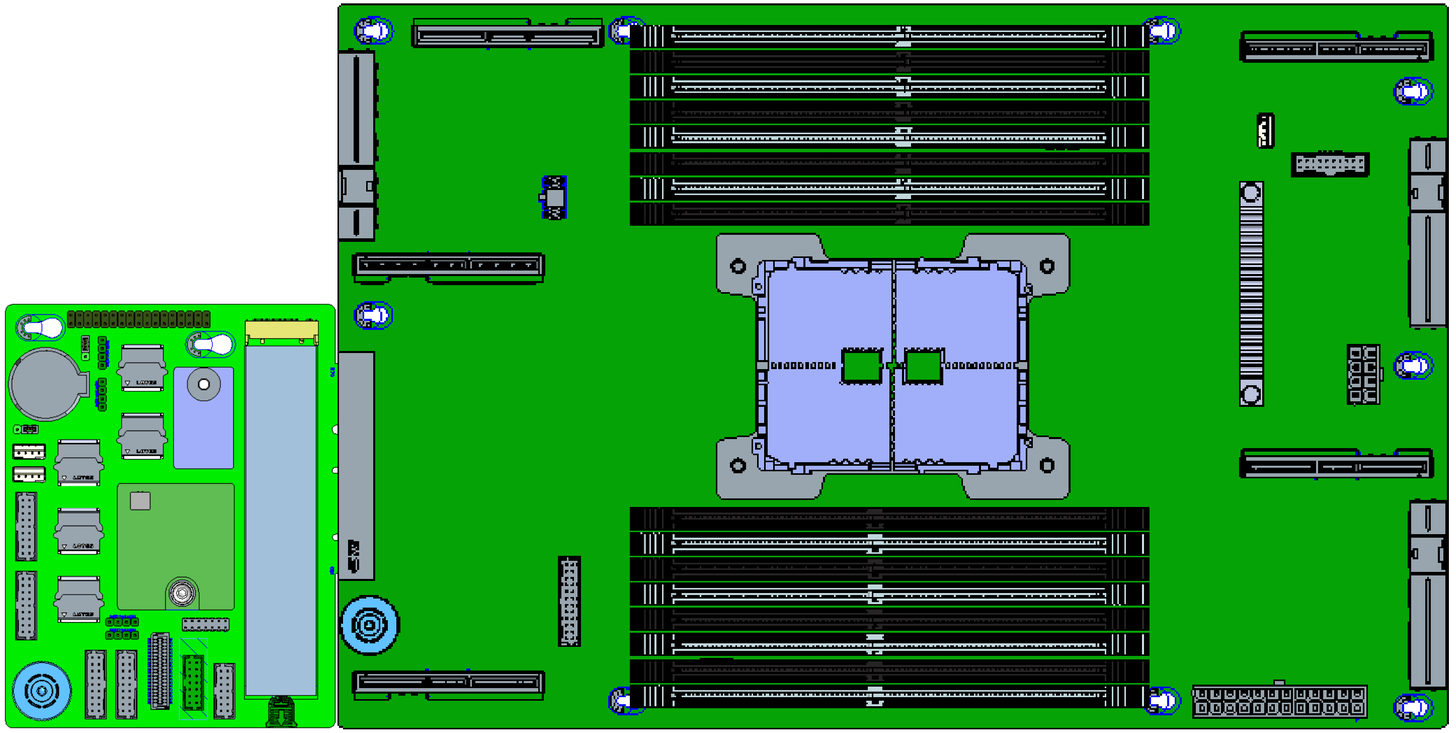
Security, Control, Management

DC-SCM

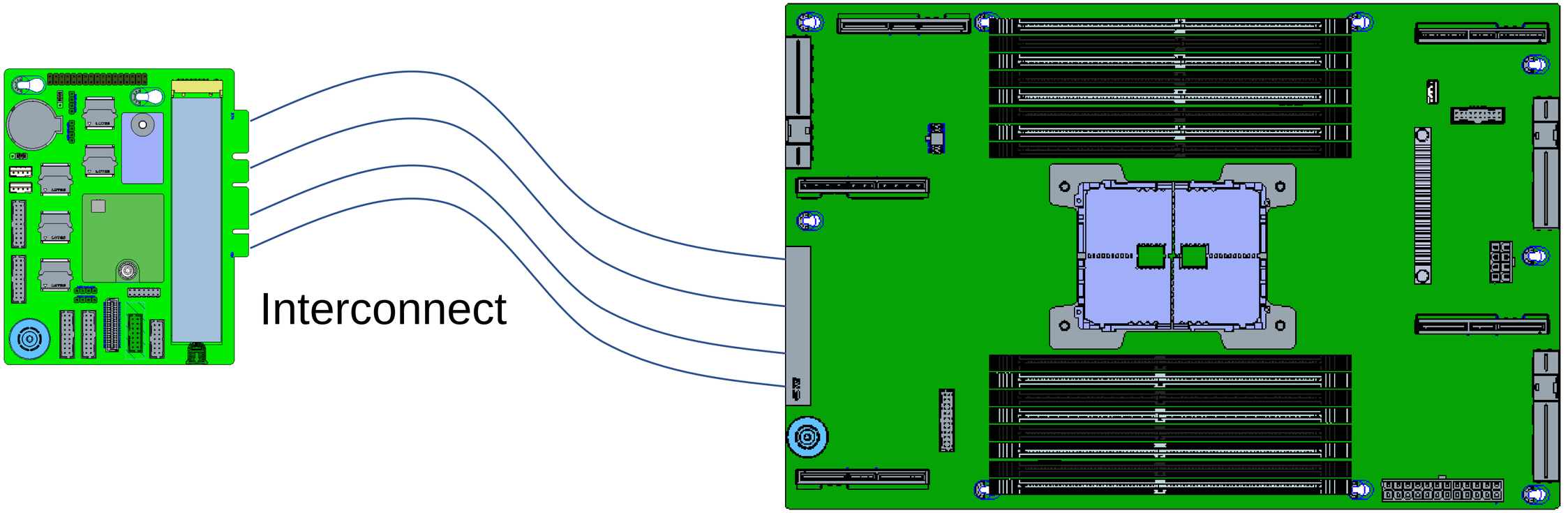


DC-SCI

DC-SCM + CPU/Mem/IO

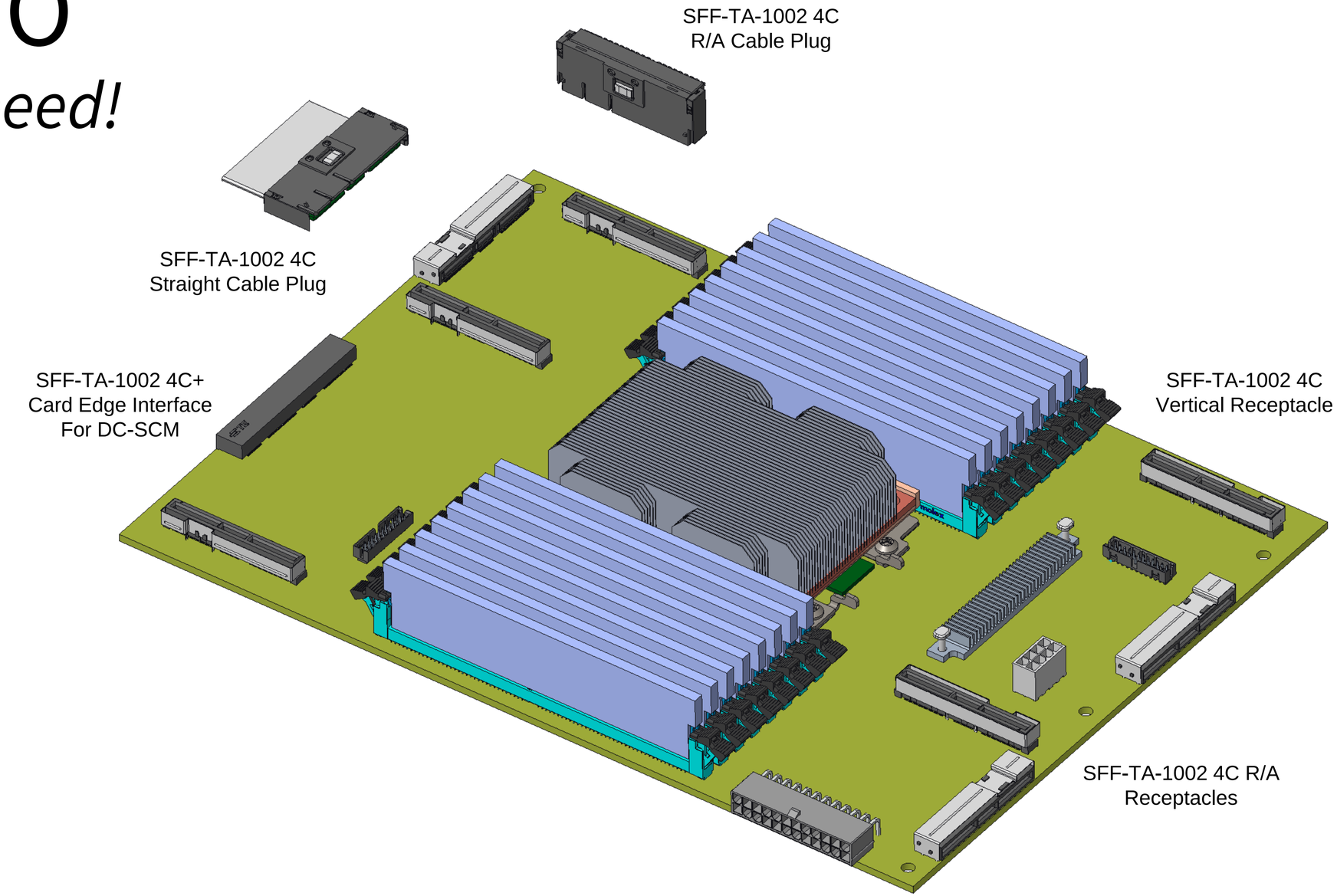


DC-SCM + CPU/Mem/IO



CPU/Mem/IO

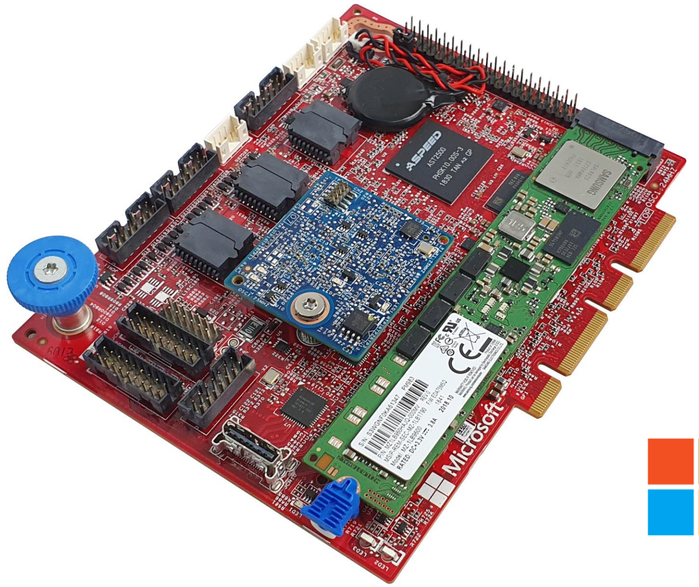
Ready for High-speed!



Started with concepts

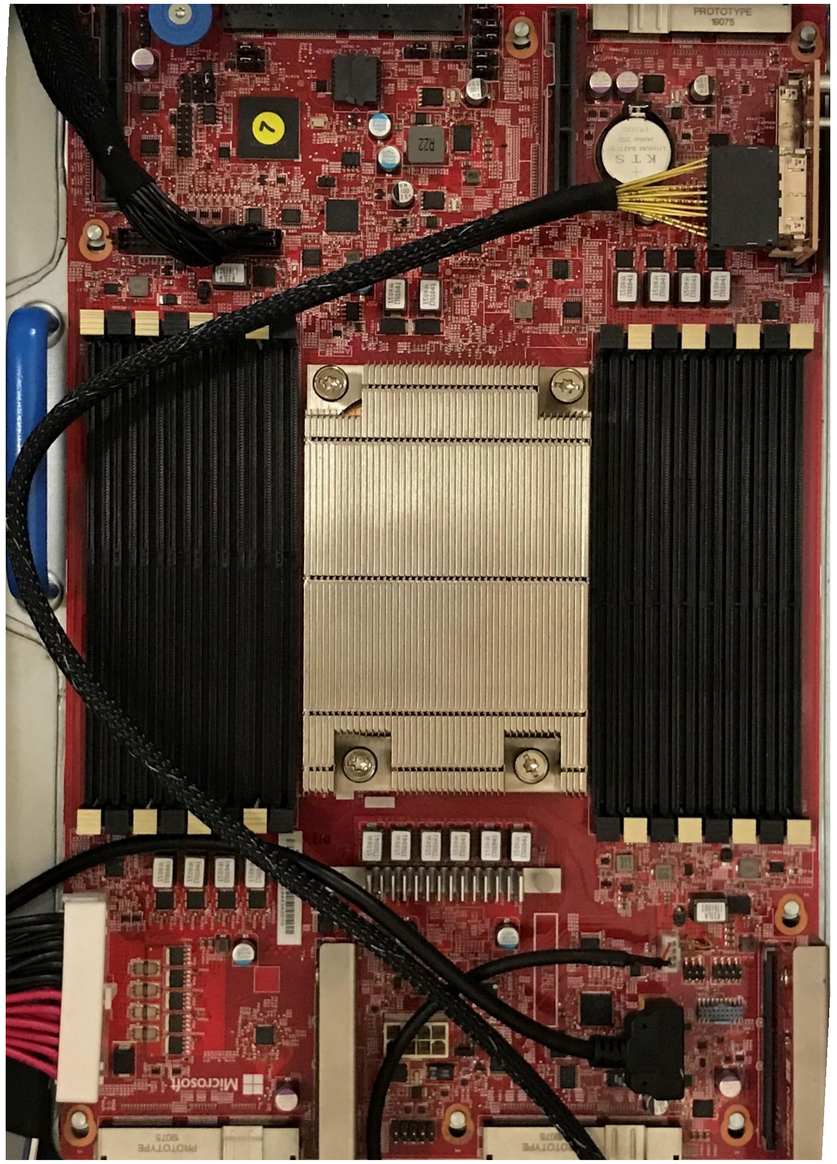
CPU/Mem/IO + DC-SCM

Realized the concept



wiwynn®

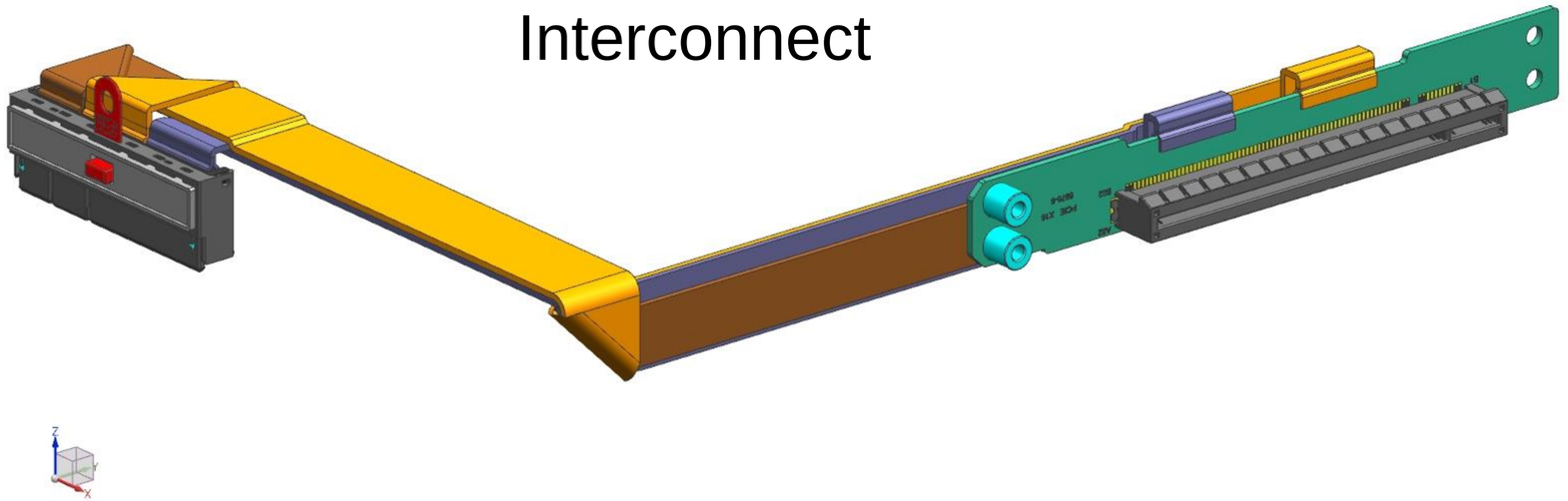
Microsoft



Add-in Card (AIC) Attachment

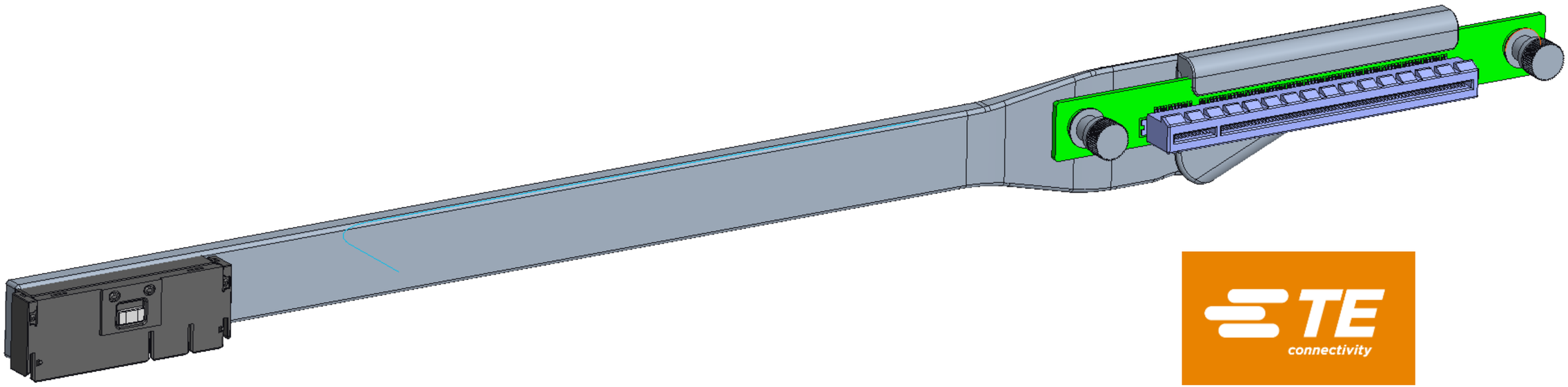
IO Slot to CPU Board Cable Harness

Ready for High-speed!



Add-in Card (AIC) Attachment

IO Slot to CPU Board Cable Harness



AIC Attachment

IO Slot to CPU Board Cable Harness

Gen-Z 4C Connector for attachment to CPU/Memory Module

SFF-TA-1002 4C Scalable Connector

PCIe Slot Connector for an Add-in Card in PCIe CEM form factor

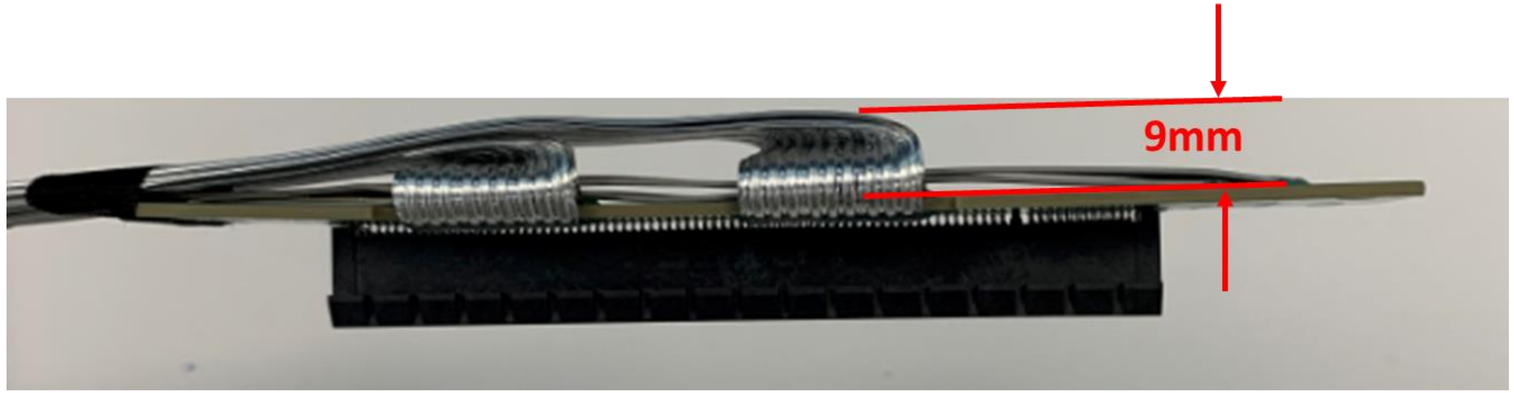
PCIe Slot to Gen-Z Pin Map (out for review– not final yet!)

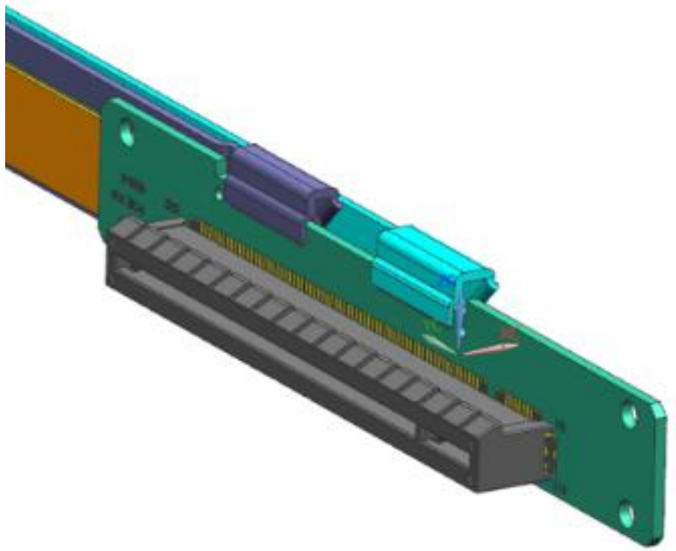
ASSEMBLY PINOUT TABLE		
PCIe Side-A		Gen-Z Side-A
P1	Description	P2
A1	PRSNT_1	A1
A2, A3	P12V	B1/B2/B3/B4/B5/B6
A4	GND	A5
A5	JTAG2	A42
A6	JTAG3	A2
A7	JTAG4	A3
A8	JTAG5	A4
A9, A10	P3.3V	A69/B68/B69
A11	PWRGD	B10
A12	GND	A6
A13	REFCLK_P	B15
A14	REFCLK_N	B14
A15	GND	A13
A16	HSIN_0 (RX)	A17
A17	HSIP_0 (RX)	A18
A18	GND	A16
A19	NC_RSVD_1	NC
A20	GND	A19
A21	HSIN_1(RX)	A20
A22	HSIP_1(RX)	A21
A23, A24	GND	A22
A25	HSIN_2(RX)	A23
A26	HSIP_2(RX)	A24
A27, A28	GND	A25
A29	HSIN_3(RX)	A26
A30	HSIP_3(RX)	A27
A31	GND	A28
A32	NC_RSVD_2	NC
A33	NC_RSVD_3	NC
A34	GND	A29
A35	HSIN_4(RX)	A30
A36	HSIP_4(RX)	A31
A37, A38	GND	A32
A39	HSIN_5(RX)	A33
A40	HSIP_5(RX)	A34
A41, A42	GND	A35
A43	HSIN_6(RX)	A36
A44	HSIP_6(RX)	A37
A45, A46	GND	A38
A47	HSIN_7(RX)	A39
A48	HSIP_7(RX)	A40
A49	GND	A41
A50	NC_RSVD_5	NC
A51	GND	A43
A52	HSIN_8(RX)	A44
A53	HSIP_8(RX)	A45
A54, A55	GND	A46
A56	HSIN_9(RX)	A47
A57	HSIP_9(RX)	A48
A58, A59	GND	A49
A60	HSIN_10(RX)	A50
A61	HSIP_10(RX)	A51
A62, A63	GND	A52
A64	HSIN_11(RX)	A53
A65	HSIP_11(RX)	A54
A66, A67	GND	A55
A68	HSIN_12(RX)	A56
A69	HSIP_12(RX)	A57
A70, A71	GND	A58
A72	HSIN_13(RX)	A59
A73	HSIP_13(RX)	A60
A74, A75	GND	A61
A76	HSIN_14(RX)	A62
A77	HSIP_14(RX)	A63
A78, A79	GND	A64
A80	HSIN_15(RX)	A65
A81	HSIP_15(RX)	A66
A82	GND	A67
NC	NC_MGMT_RST	A9
NC	NC_LED/ACTIVITY	A10
NC	NC_REFCLK1_P	A14
NC	NC_REFCLK1_N	A15

ASSEMBLY PINOUT TABLE		
PCIe Side-B		Gen-Z Side-B
P1	Description	P2
B1, B2, B3	P12V	B1/B2/B3/B4/B5/B6
B4	GND	GND
B5	SMCLK	A7
B6	SMDAT	A8
B7	GND	B13
B8	P3.3V	A69/B68/B69
B9	JTAG1	A68
B10	P3.3V_AUX	B11
B11	WAKE	A70
B12	CLKREQ	A11
B13	GND	B16
B14	HSON_0 (TX)	B17
B15	HSOP_0(TX)	B18
B16	GND	GND
B17	NC_PRSNT_2_B17	NC
B18	GND	B19
B19	HSON_1(TX)	B20
B20	HSOP_1(TX)	B21
B21, B22	GND	B22
B23	HSON_2(TX)	B23
B24	HSOP_2(TX)	B24
B25, B26	GND	B25
B27	HSON_3(TX)	B26
B28	HSOP_3(TX)	B27
B29	GND	B28
B30	PWRBRK	B8
B31	PRSNT_2_B31	A12
B32	GND	B29
B33	HSON_4(TX)	B30
B34	HSOP_4(TX)	B31
B35, B36	GND	B32
B37	HSON_5(TX)	B33
B38	HSOP_5(TX)	B34
B39, B40	GND	B35
B41	HSON_6(TX)	B36
B42	HSOP_6(TX)	B37
B43, B44	GND	B38
B45	HSON_7(TX)	B39
B46	HSOP_7(TX)	B40
B47	GND	B41
B48	PRSNT_2_B48	B42
B49	GND	B43
B50	HSON_8(TX)	B44
B51	HSOP_8(TX)	B45
B52, B53	GND	B46
B54	HSON_9(TX)	B47
B55	HSOP_9(TX)	B48
B56, B57	GND	B49
B58	HSON_10(TX)	B50
B59	HSOP_10(TX)	B51
B60, B61	GND	B52
B62	HSON_11(TX)	B53
B63	HSOP_11(TX)	B54
B64, B65	GND	B55
B66	HSON_12(TX)	B56
B67	HSOP_12(TX)	B57
B68, B69	GND	B58
B70	HSON_13(TX)	B59
B71	HSOP_13(TX)	B60
B72, B73	GND	B61
B74	HSON_14(TX)	B62
B75	HSOP_14(TX)	B63
B76, B77	GND	B64
B78	HSON_15(TX)	B65
B79	HSOP_15(TX)	B66
B80	GND	B67
B81	PRSNT_2_B81	B70
B82	GND	GND
NC	NC_MFG	B7
NC	NC_DUALPORTEN	B9
NC	NC_PWRDIS	B12

Ground pin	Zero volt reference, all tied together
Power pin	Supplies power to the card
High speed pin	High speed signals
Detect	Sense Pin
Other aux	May be pulled low or sensed by multiple cards
Reserved	Reserved for future use and no connect

3M™ Twin Ax Assembly





Cable compresses down to 7 mm without compromising performance
Insertion loss is not significantly impacted by folding

Adding Capacitors to the PCB

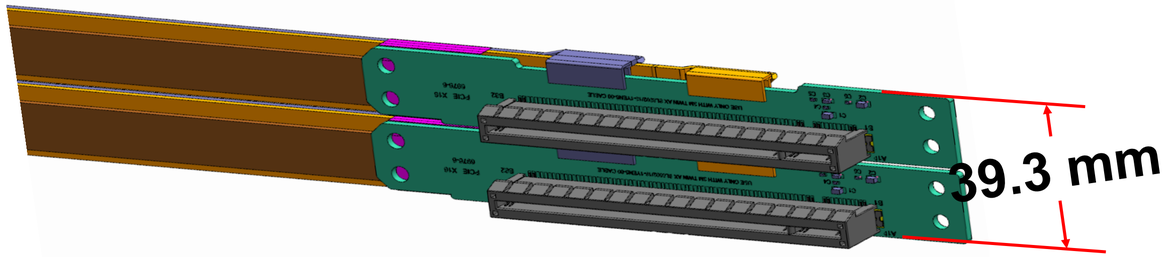
Added high-frequency by-pass capacitors to power rails

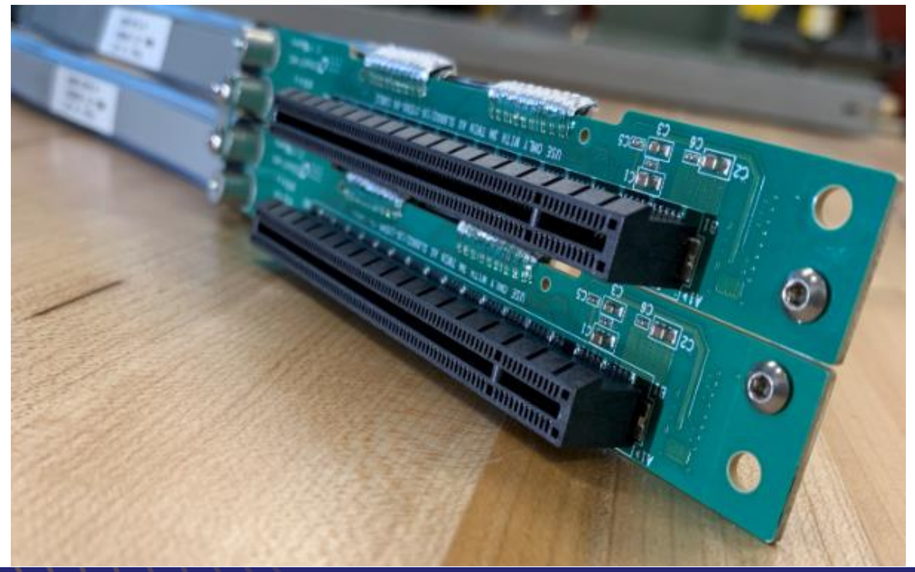
to improve Power Integrity and

to reduce Simultaneous Switching Noise (SSN) effects

Two Assemblies in 1U Height (44.5 mm)

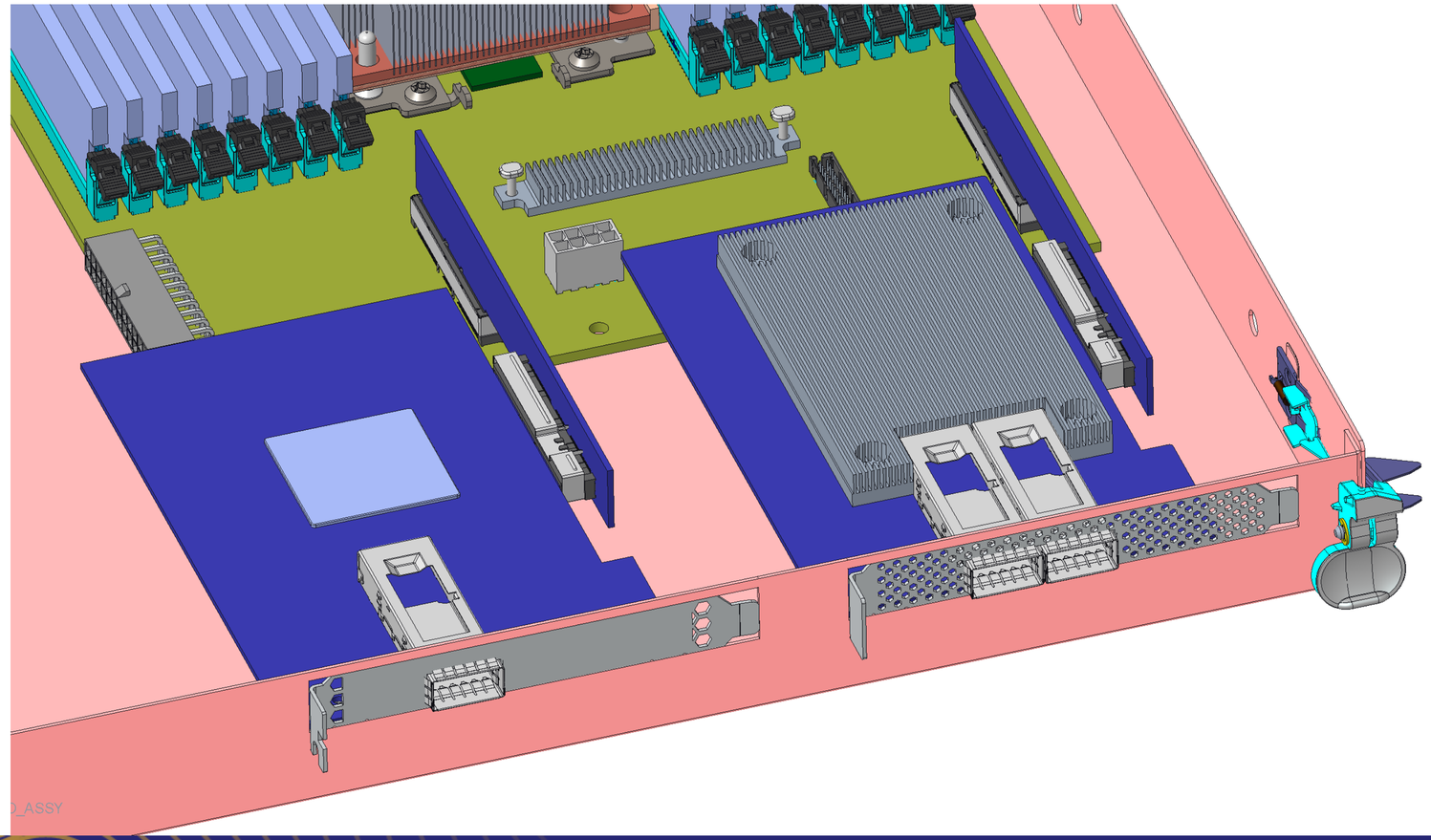
- Same PCB assembly for both assemblies to minimize number of SKUs
- Two PCIe Slots in 1U chassis
- Cable assembly folded onto itself





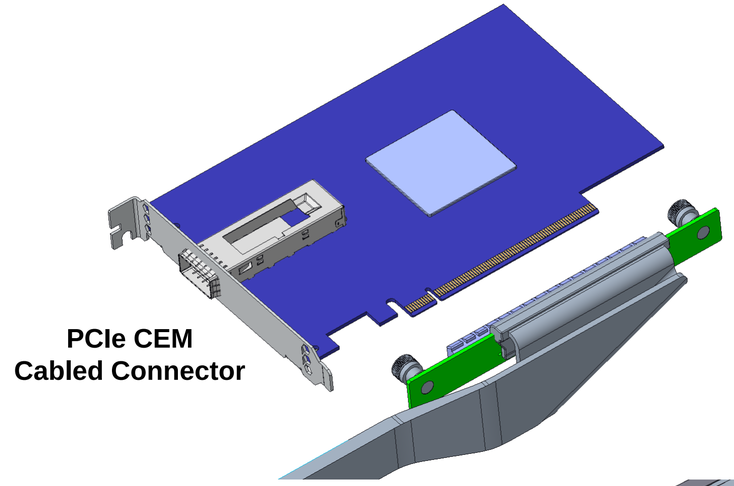
Enabled Options

Add-in Card (Riser Attached)

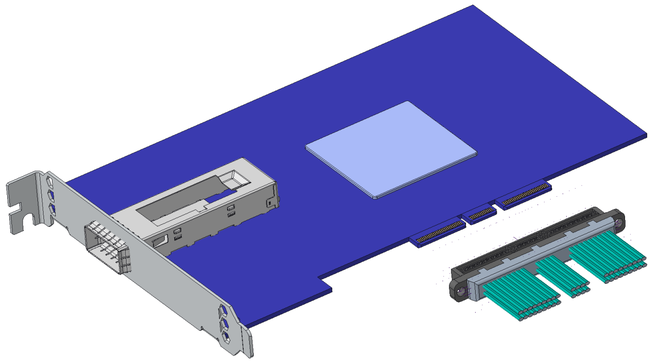


Enabled Options

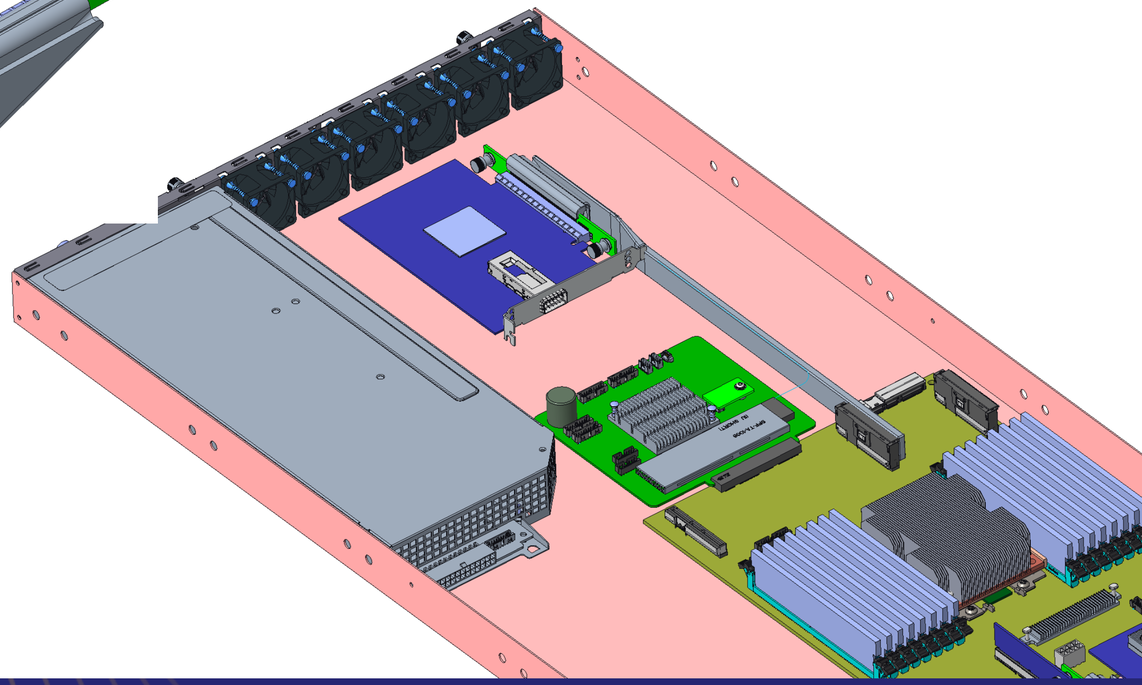
Add-in Card (Cable Attached) *Ready for High-speed!*



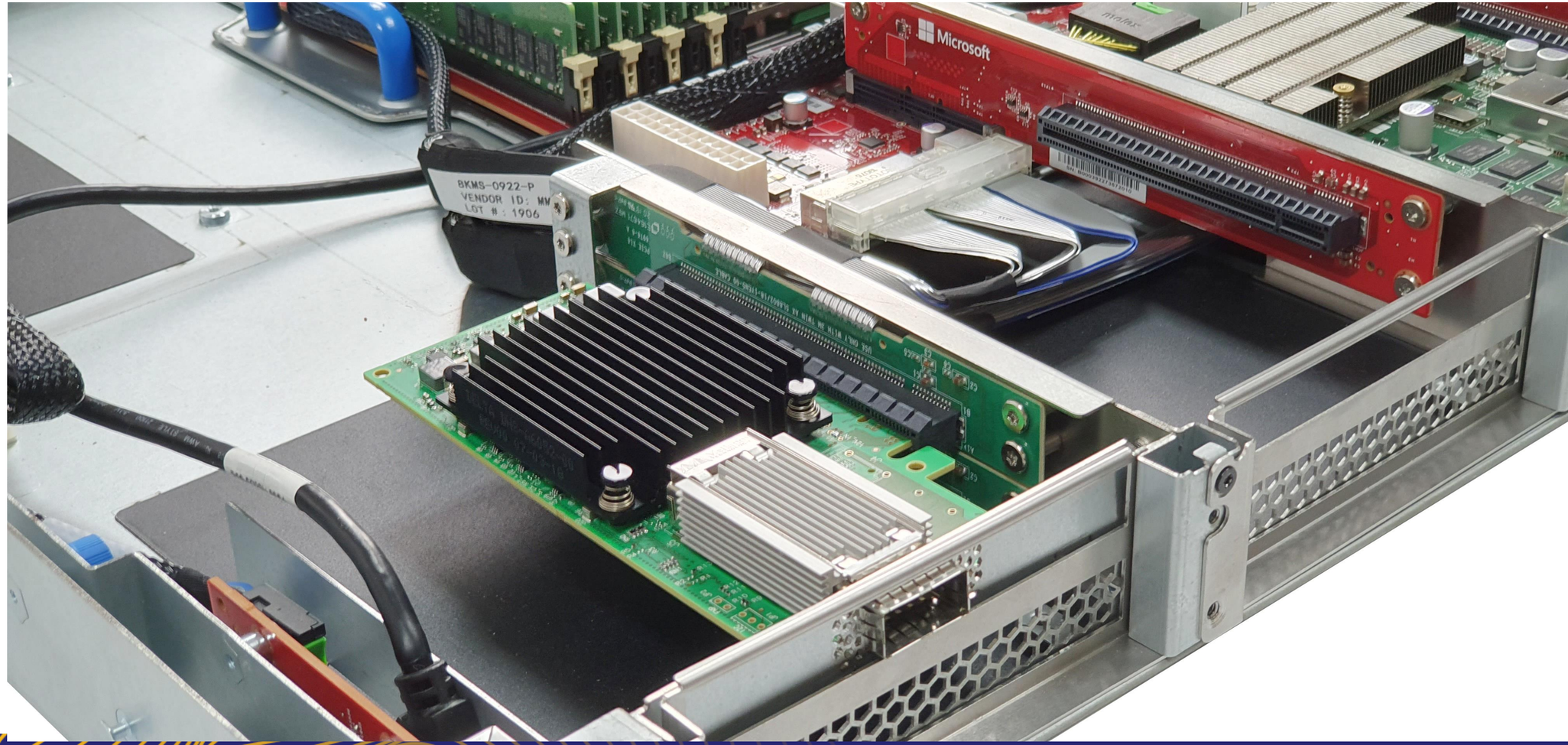
PCIe CEM
Cabled Connector



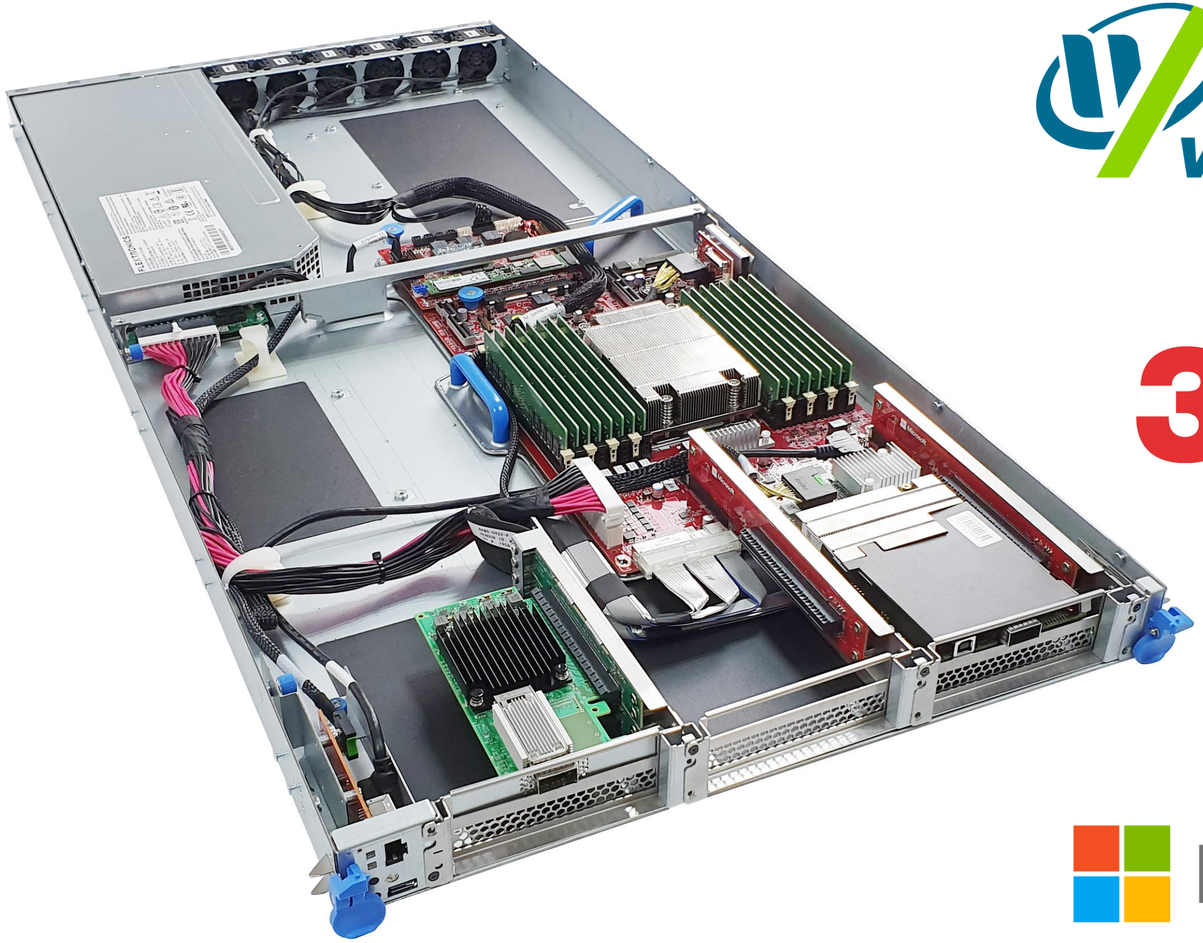
Native
SFF-TA-1002 4C
Cabled Receptacle



Realized Both Options (for Add-in Card attachment)

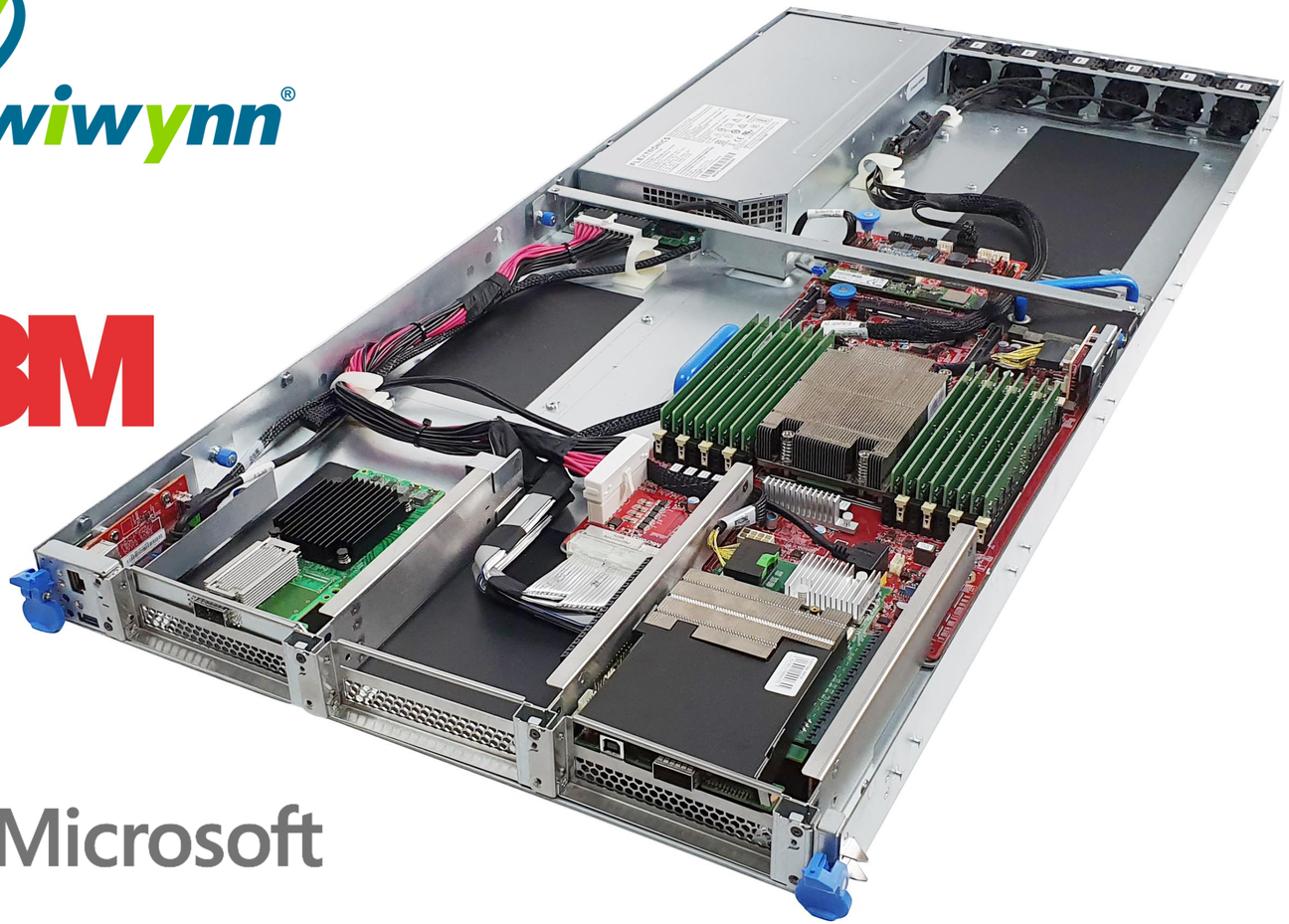


Examples of Modular Designs



wiwynn®

3M



Microsoft

wiwynn®

OCP REGIONAL SUMMIT

Open. Together.

The Datacenter-Ready Secure Control Module ***(DC-SCM)***

&

The Datacenter-Ready Secure Control Interface ***(DC-SCI)***

DC-SCM (in a nutshell)

- DC-SCM is “the heart of the motherboard” when we extract CPU(PCH), Memory, and IO Slots

- Given a traditional 1S, 2S, 4S, … Motherboard, extract CPU/PCH, DIMM Slots, IO Slots, and the associated VRs, Clock Drivers, and Reset Circuitry, and move them to a new Module

- The residual is the DC-SCM which will include everything else such as BMC, RoT, Flash, and PSU control along with optional Boot SSD and connectors for Fan control

DC-SCM (Motivation)

- Don’t reinvent the wheel with each new server design
- Unify the solution to support multiple architectures

“Same as before” with F/W, S/W, & Services-- maintaining the established tools and solutions experience with the same management, power sequencing, reset, FRU ID, VPD, …

A vehicle to drive a common Boot, Monitoring, Control, and Remote Debug procedures for Xeon, EPYC, ARM64, and Power Servers with the same firmware, diagnostic tools, manufacturing tools

Software Standardization

Collaborating with CPU suppliers, Open Computing Project community (**OCP**), Linux Foundation, and Open System Firmware (**OSF**) to standardize the hardware and software for **OpenBMC** with **RedFish** interface and for the system BIOS/UEFI based on **EDK-II**

DC-SCM (Ingredients)

- Most MBA building blocks are stateless
- The secure control module (DC-SCM) includes all system related components (other than CPU/Mem/IO) that are normally present on Motherboards
- Baseboard Management Controller (BMC), Realtime Clock (RTC), FAN/PSU Control, Root of Trust Chip (RoT: Cerberus/Other and the associated circuitry), BIOS & BMC Flash, and the Boot Device

- SCM holds control bits secured (no firmware on CPU/Mem Module)

DC-SCM (Form Factors)

The SCM is small enough to fit anywhere in the Chassis

Flexible as development vehicle or for Expansion Chassis:

1. Cabled to Chassis Edge for external connections such as RJ45, Serial Console, and cabled internally for Fans, PSUs, …

IO connectors at the Edge of the Chassis; DC-SCI for interfacing to CPU/Mem Module:

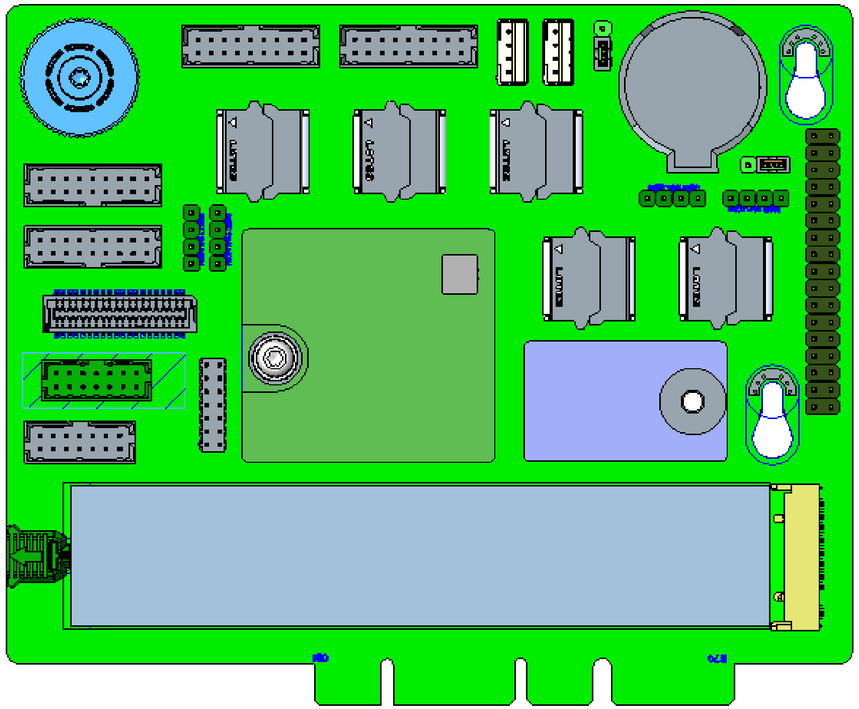
2. A plug-in module like OCP NIC-3: co-planar to CPU/Mem Module
3. A plug-in module like low-profile PCIe cards: plugs vertically into the CPU/Mem module

An example of DC-SCM

DC-SCM

- Receives Power
- Remote Control at Cloud Scale
  - CPU/Memory/IO Module (Xeon, EPYC, ARM64)
  - Expansion Chassis (JBOD, JBOG)
  - Fans, PSUs
- Includes
  - BMC and Rack Management Interface
  - Flash Devices (all Firmware)
  - RoT and TPM for Security
  - Optional Boot SSD
  - Remote, at-scale Debug

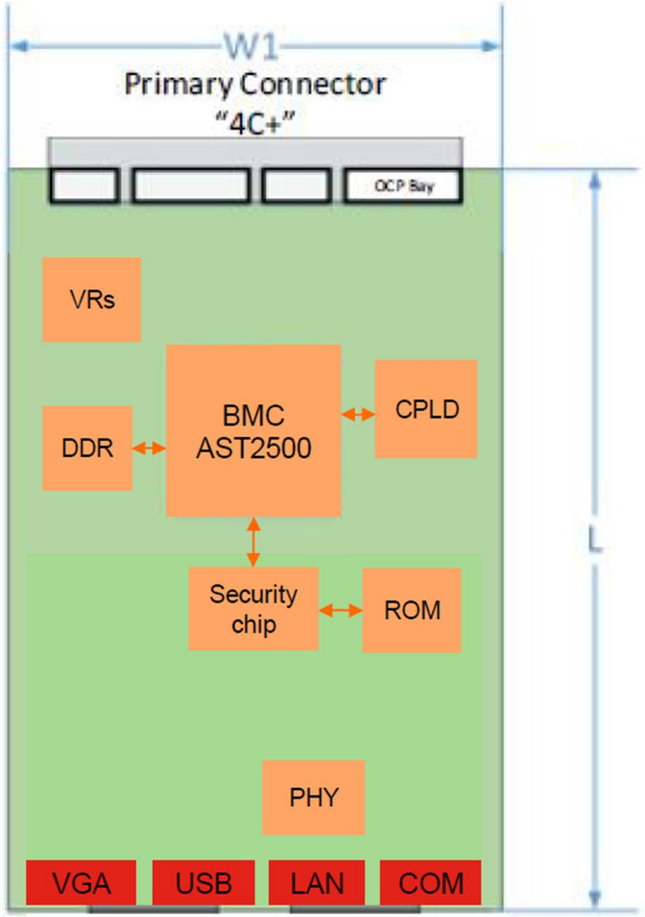
DC-SCM



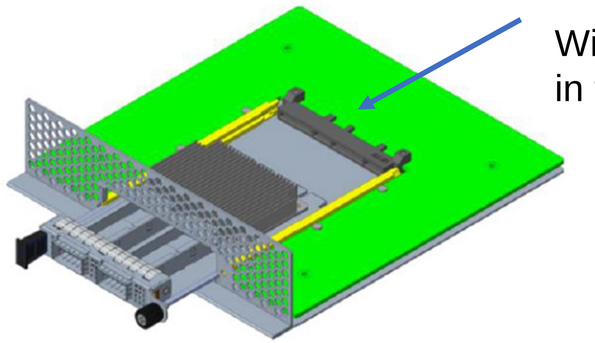
DC-SCI

SFF-TA-1002 4C+
168-pin, Scalable Connector

Another Example of DC-SCM (OCP NIC3 Form Factor)



SFF-TA-1002 4C+
168-pin, Scalable Connector



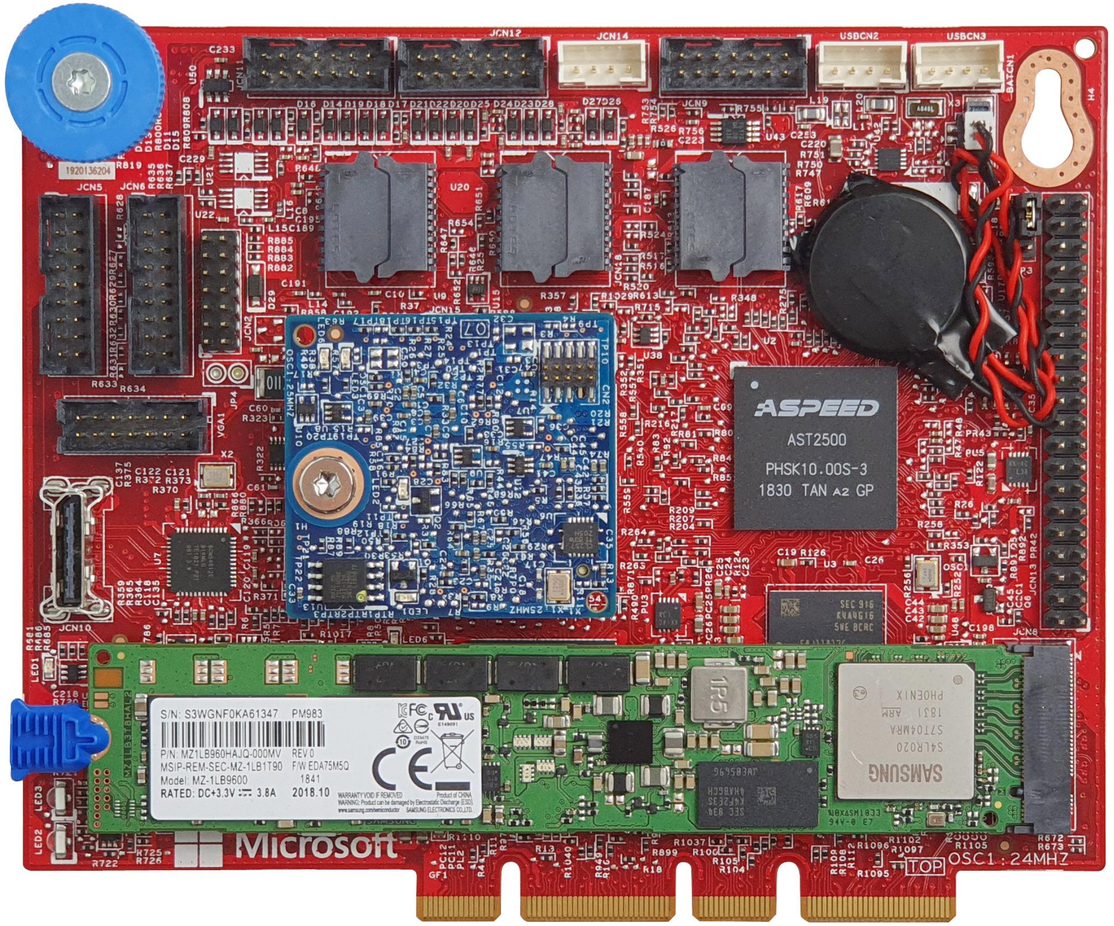
With a mechanical key to avoid plugging in the NIC 3.0 connector & vice-versa

Form Factor	Width	Depth	Primary Connector	Secondary Connector
SFF	W1 = 76 mm	L = 115 mm	"4C+" 168 pins	N/A

DC-SCM

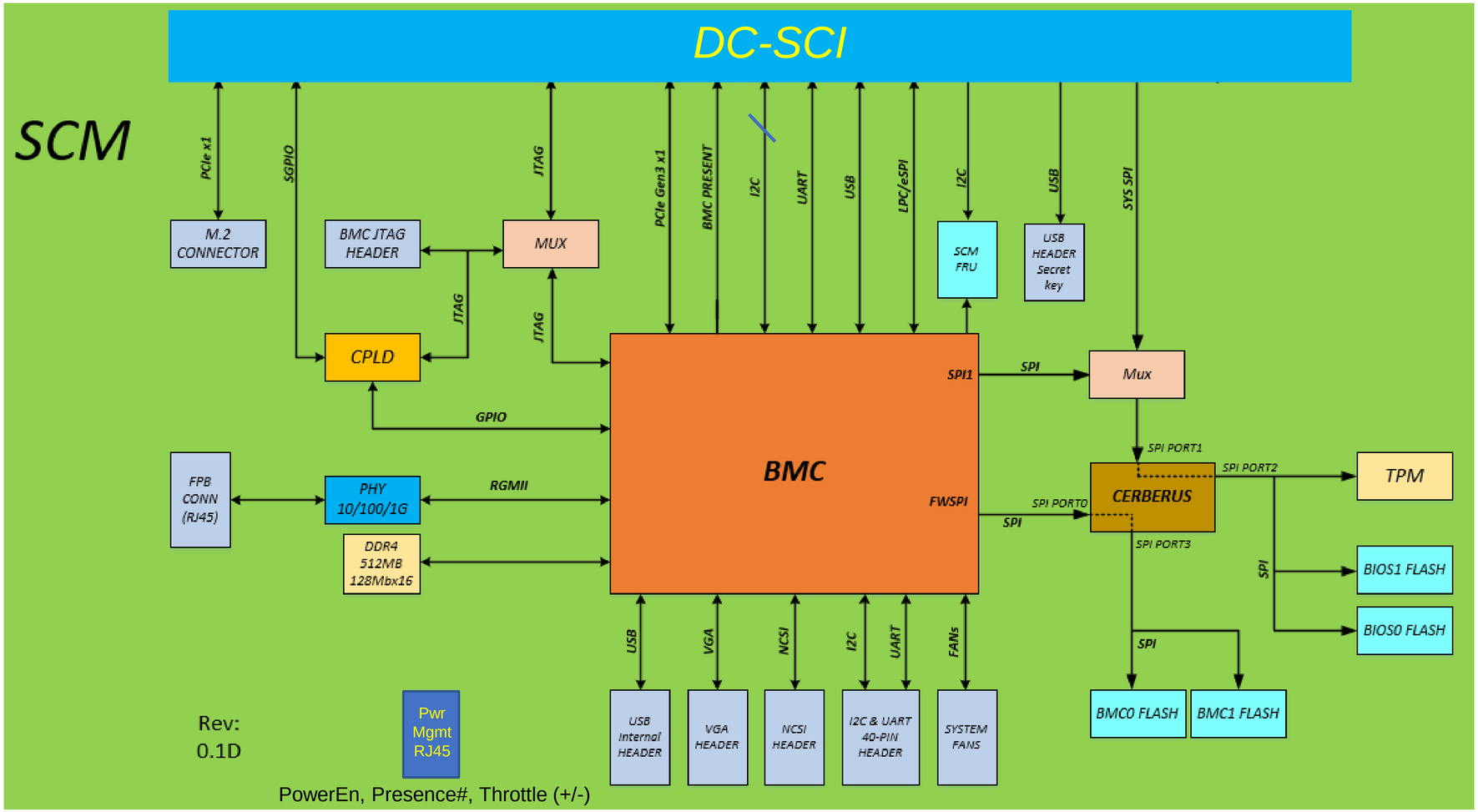
Realized

DC-SCM



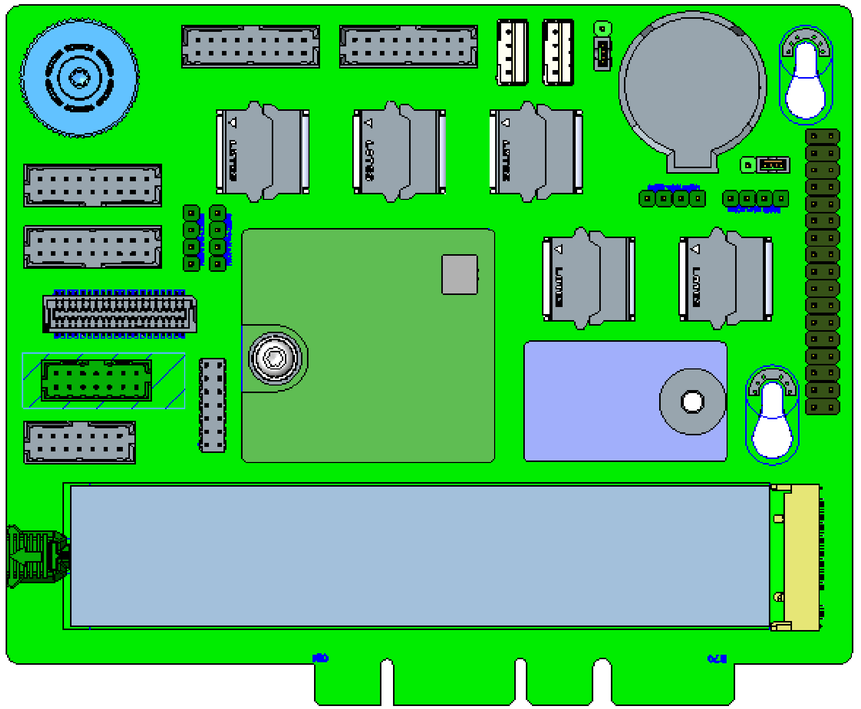
DC-SCI

DC-SCM (Block Diagram Example)



An Example of SCM Expander Connectors

Function	Qty
M2 socket	1
TPM connector	1
SPI socket	5
RoT connector	1
VGA cable connector	1
NCSI cable connector	1
Front panel cable connector	1
FAN cable connector	2
BMC debug UART Pin header	1
HOST UART Pin header	1
Auxiliary UART Pin header	1
Reserved UART Pin header	1
Pin header	2
PSU cable connector	2
JATG cable connector	1
Reserved USB cable header	1
ID LED header	1
Battery	1
I2C Header	1
Golden Finger	1



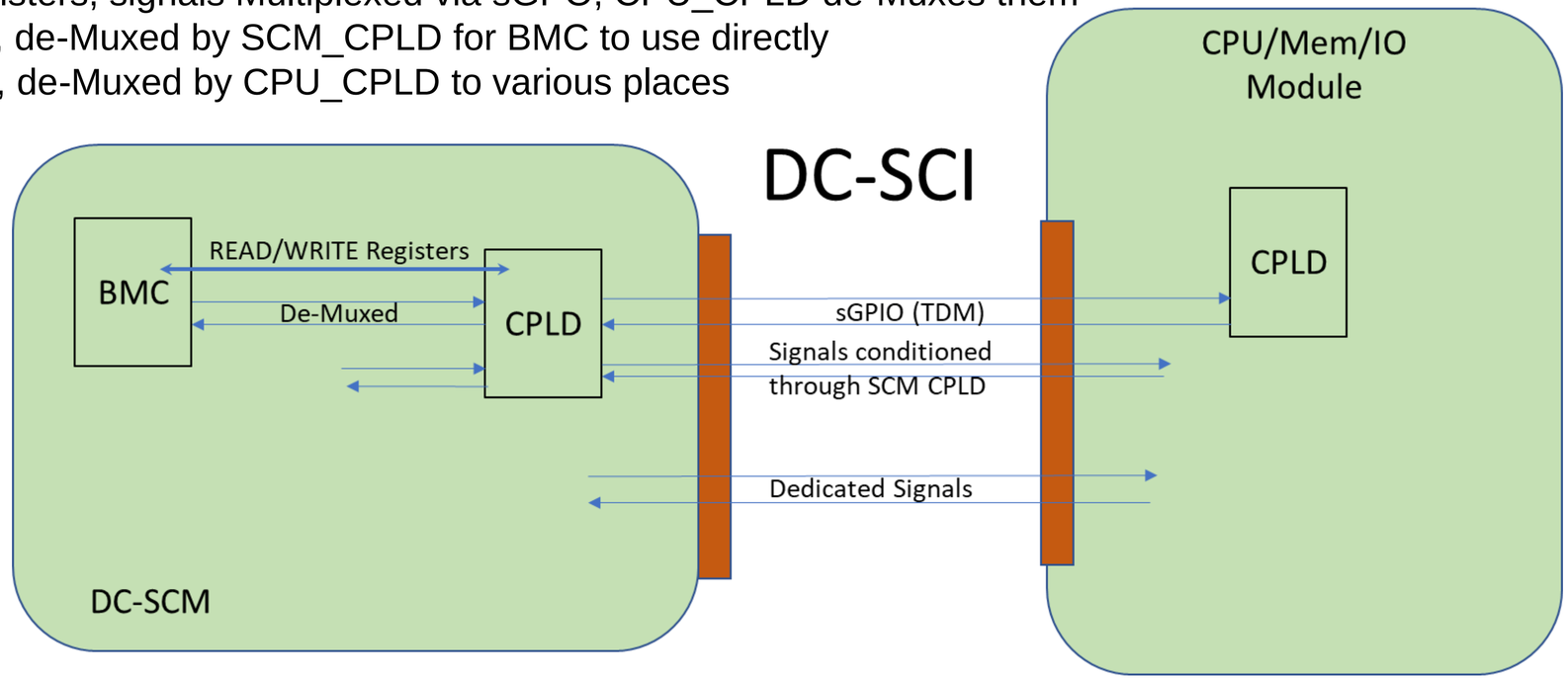
SFF-TA-1002 4C+
168-pin, Scalable Connector

DC-SCM to CPU/Mem Module Interface (DC-SCI)

Pinout and definition

Pin Reduction via SGPIO

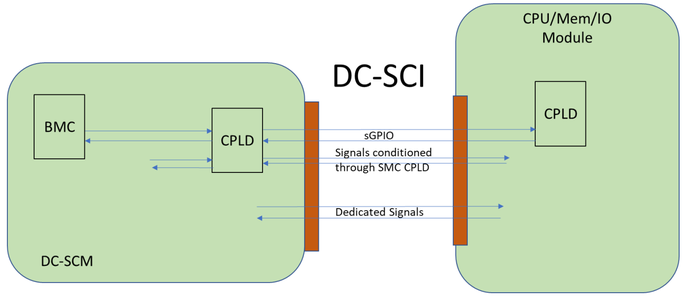
Dedicated Signals through DC-SCI
Signals conditioned through the SCM CPLD
rsGPI: Signals Multiplexed via CPU_CPLD, presented in Registers at SCM_CPLD, READ by BMC
rsGPO: BMC WRITEs into SCM_CPLD Registers, signals Multiplexed via sGPO, CPU_CPLD de-Muxes them
sGPI: Signals Multiplexed via CPU_CPLD, de-Muxed by SCM_CPLD for BMC to use directly
sGPO: Signals Multiplexed via BMC_CPLD, de-Muxed by CPU_CPLD to various places



SFF-TA-1002 4C+
168-pin, Scalable Connector
Gold-finger Plugin or Cabled

SFF-TA-1002 4C+
168-pin, Scalable Connector
Right-angled or Vertical

Pin Reduction Techniques



Various techniques for reducing required BMC and DC-SCI pins

CPU Module CPLD serially shifts GPIOs to/from SCM CPLD; SCM CPLD replicates them on pins connected to BMC GPIOs. (for latency-insensitive signals that need “exact” replication at BMC for firmware compatibility)

CPU Module CPLD serially shifts GPIs to SCM CPLD; In response to an Event/Interrupt, BMC READs GPIOs via SPI (or I2C) (for latency-insensitive signals and signals that don’t need “exact” replication at BMC for firmware compatibility)

BMC WRITEs GPOs via SPI (or I2C) into SCM CPLD. SCM CPLD serially shifts them onto CPU Module CPLD. CPU Module CPLD replicates them as parallel signals to go to various places on the CPU/Memory Module. (for latency-insensitive signals and signals that don’t need “exact” replication at BMC for firmware compatibility)

DC-SCI (pinout implemented in PoC)

Connector Type:

SFF-TA-1002 4C+

168-pin, Scalable Connector

DC-SCM Connector:

Gold-finger

CPU/Memory/IO Module Connector:

Right-angled or Vertical Receptacle

Function	168	Comment
BIOS SPI	6	Serboard SPI
BMC_PCIE	7	PCIE to BMC
CPLD_SGPIO	9	CPLC comunication interface between MB/SCM
Critical GPIO	30	The critical event dedicate pin
GND	23	GND
I2C ALERT	13	I2C ALERT signal
JTAG	5	JTAG to MB
LPC & ESPI	12	LPC or ESPI
M2_CLK	2	PCIE clock to M.2 connector
M2_PCIE	4	PCIE to M.2 connector
PECI	2	PECI for Intel platform
POWER	5	3*P12V_STBY, 1* for LPC/ESPI power difference
Sequence	3	For sequence control reference
I2C	26	I2C BUS
SPI	0	SOC/PCH SPI to SCM board
UART	2	SOC/PCH UART to BMC
USB	4	SOC/PCH USB to SCM
Remote Debug	4	For Intel/AMD remote debug
TPM	2	TPM IRQ & chip select
FPGA SPI	4	Reserve for the FPGA FW
RSVD	5	Reserved
PSU Connector	0	PSU Connector signals

Win-win!

DC-SCM accelerates deploying servers from various suppliers into the datacenter

Standardizing **DC-SCI** for ease of integration into various datacenters

Flexibility to use **BMC** and **RoT** chips of choice on any platform

<u>From a Datacenter point of view:</u>

with one DC-SCM, a datacenter may support multiple variants of servers (AMD-, Xeon-, ARM64-, Power-based 1S, 2S, 4S, …) and expansion chassis, JBODs, JBOG, JBOFs, …

<u>From OEM/ODMs' point of view:</u>

A product will fit datacenters of various CSPs or Hyperscalers

If we are smart, one DC-SCM may enable supplier products into different DC types; otherwise, each Datacenter Provider may have its own version of DC-SCM

Call to Action

Design your Servers, Expansion Chassis, JBODs, JBOGs, JBOFs, multi-Server Chassis, etc. with

DC-SCI connector in mind.

Make your solution Datacenter-Ready!

Join the effort to enhance DC-SCM and DC-SCI

https://www.opencompute.org/projects/server

OCP REGIONAL SUMMIT

Open. Together.

OCP Regional Summit

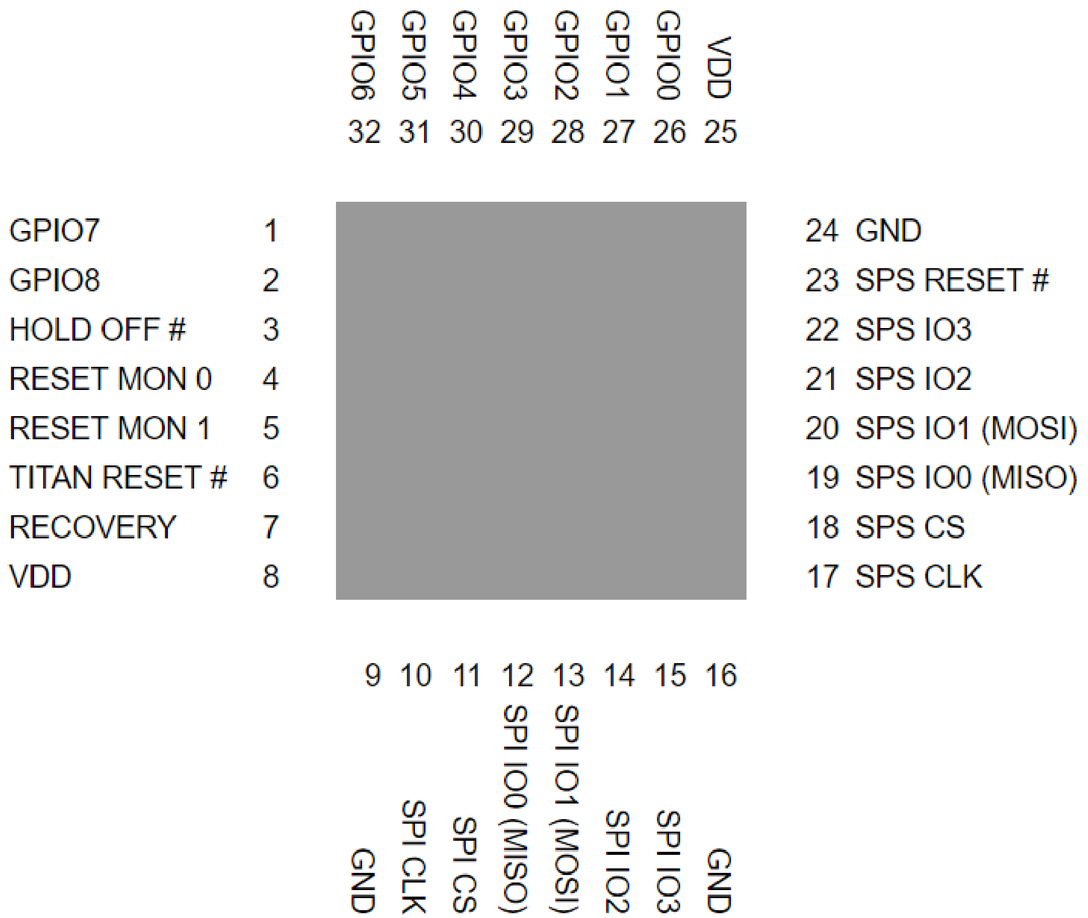
26–27, September, 2019

Presenter

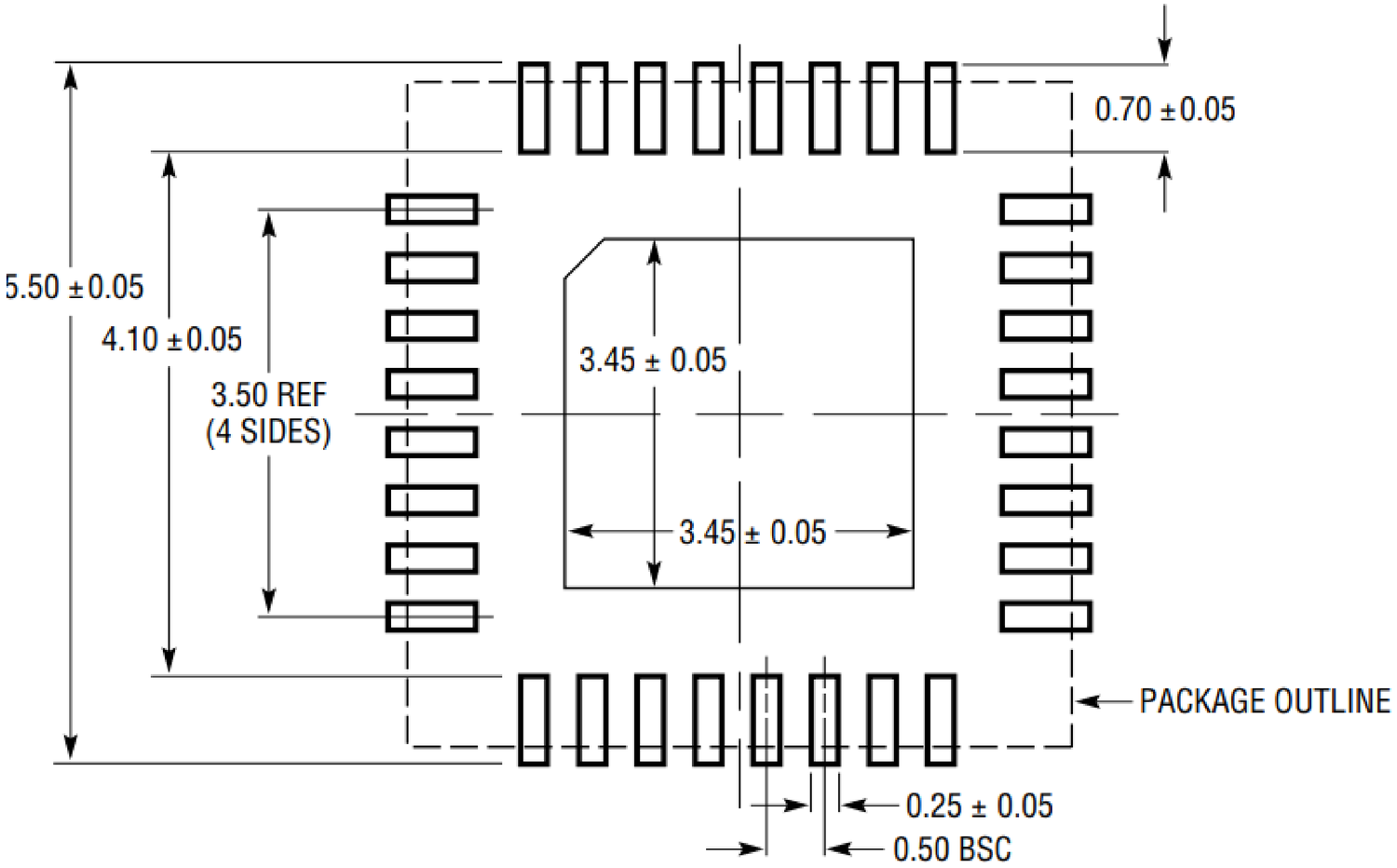
Siamak Tavallaei is a Principal Architect at Microsoft Azure and co-chair of OCP Server Project. Collaborating with industry partners, he drives several initiatives in research, design, and deployment of hardware for Microsoft's cloud-scale services at Azure. He is interested in Big Compute, Big Data, and Artificial Intelligence solutions based on distributed, heterogeneous, accelerated, and energy-efficient computing. His current focus is the optimization of large-scale, mega-datacenters for general-purpose computing and accelerated, tightly-connected, problem-solving machines built on collaborative designs of hardware, software, and management.

Backup Slides

An Example of an OCP RoT Chip



An Example of an OCP RoT Chip



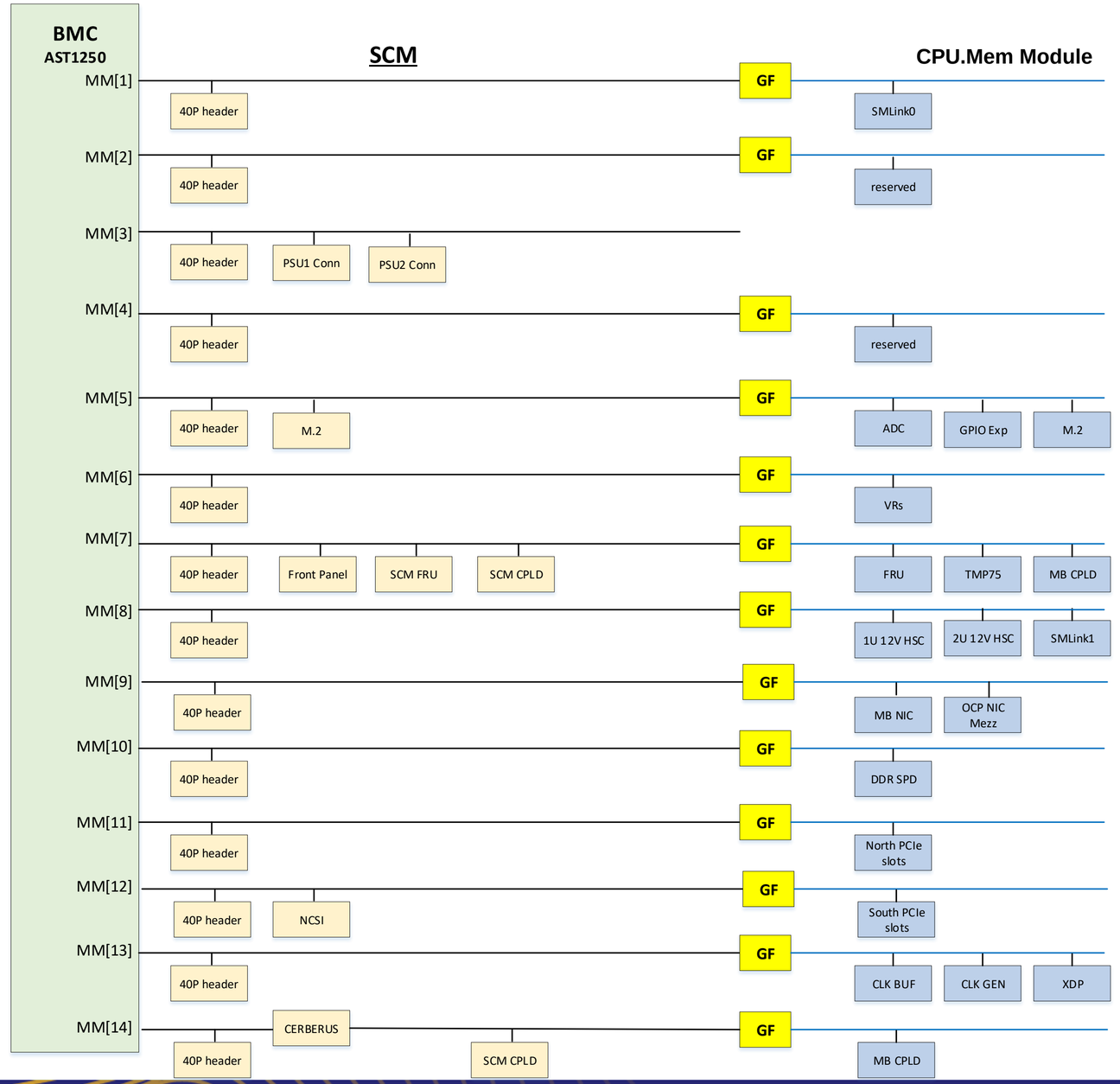
RoT Pin Description

GPIO	Function 1	Function 2
0	SMBus0 Alert#	JTAG TDO
1	SMBus0 SCL	JTAG TDI
2	SMBus0 SDA	JTAG TMS
3	Flash Reset #	JTAG TCK
4	UART TX	
5	UART RX	
6	SMBus1 Alert#	
7	SMBus1 SCL	USB N
8	SMBus1 SDA	USB P

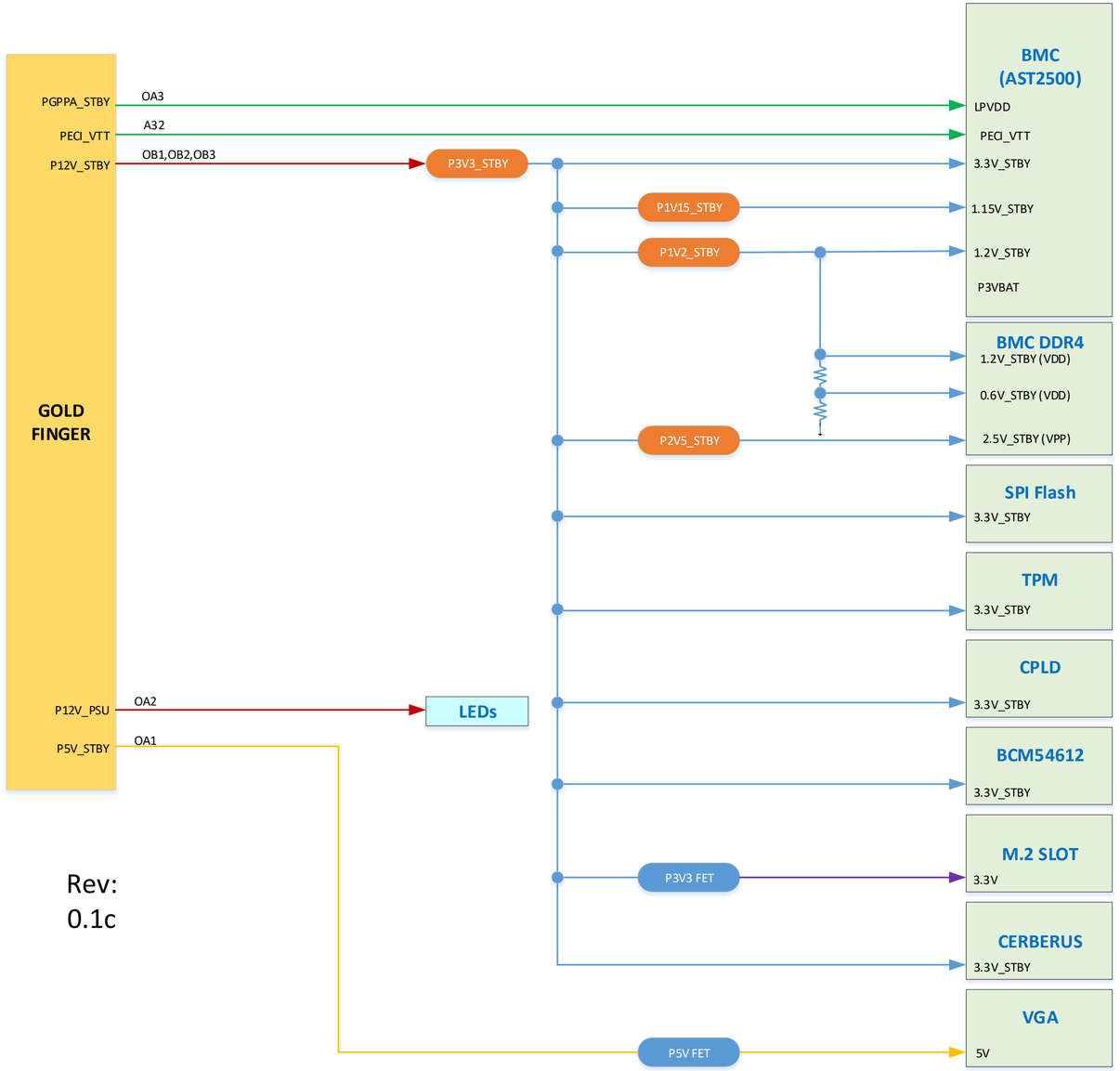
Pin	Signal	When Unused	I/O	Description
3	HOLD OFF #	PU	O	Active low output, holds the boot device in reset while the flash is being verified
4	RESET MON 0#	PU	I	Active low input, used to monitor the reset line of the boot device
5	RESET MON 1#	PU	I	Active low input, used to monitor the reset line of the boot device
6	RESET#	-	I	Active low chip reset. Must be held low for at least Xms after Vdd goes high
7	RECOVERY	-	I	Hold to VDD at reset to trigger new firmware image to be read in on the SPI slave interface. This pin should only be driven by a header. See the recommended circuit application note.
10	SPI CLK		O	SPI master clock
11	SPI CS #	PU	O	SPI master chip select
12	SPI IO0 (MISO)		I/O	SPI master I/O 0; MISO in single mode, IO0 in dual/quad mode
13	SPI IO1 (MOSI)	PU	I/O	SPI master I/O 1; MOSI in single mode, IO1 in dual/quad mode
14	SPI IO2	NC	I/O	SPI master I/O 2 in quad mode
15	SPI IO3	NC	I/O	SPI master I/O 3 in quad mode
17	SPS CLK	NC	I	SPI slave clock
18	SPS CS #	PU	I	SPI slave chip select
19	SPS IO0 (MISO)	NC	I/O	SPI slave I/O 0; MISO in single mode, IO0 in dual/quad mode
20	SPS IO1 (MOSI) eSPI Alert#	NC	I/O	SPI slave I/O 1; MOSI in single mode, IO1 in dual/quad mode; eSPI Alert#
21	SPS IO2	NC	I/O	SPI slave I/O 2 in quad mode
22	SPS IO3	NC	I/O	SPI slave I/O 3 in quad mode
23	SPS RESET #	PU	I	eSPI slave reset

I²C Tree

Rev:
0.1D



SCM Power Distribution



Rev:
0.1c