



OPEN
Compute Project



OCP U.S. SUMMIT 2017

Santa Clara, CA



Facebook's Open Networking Hardware

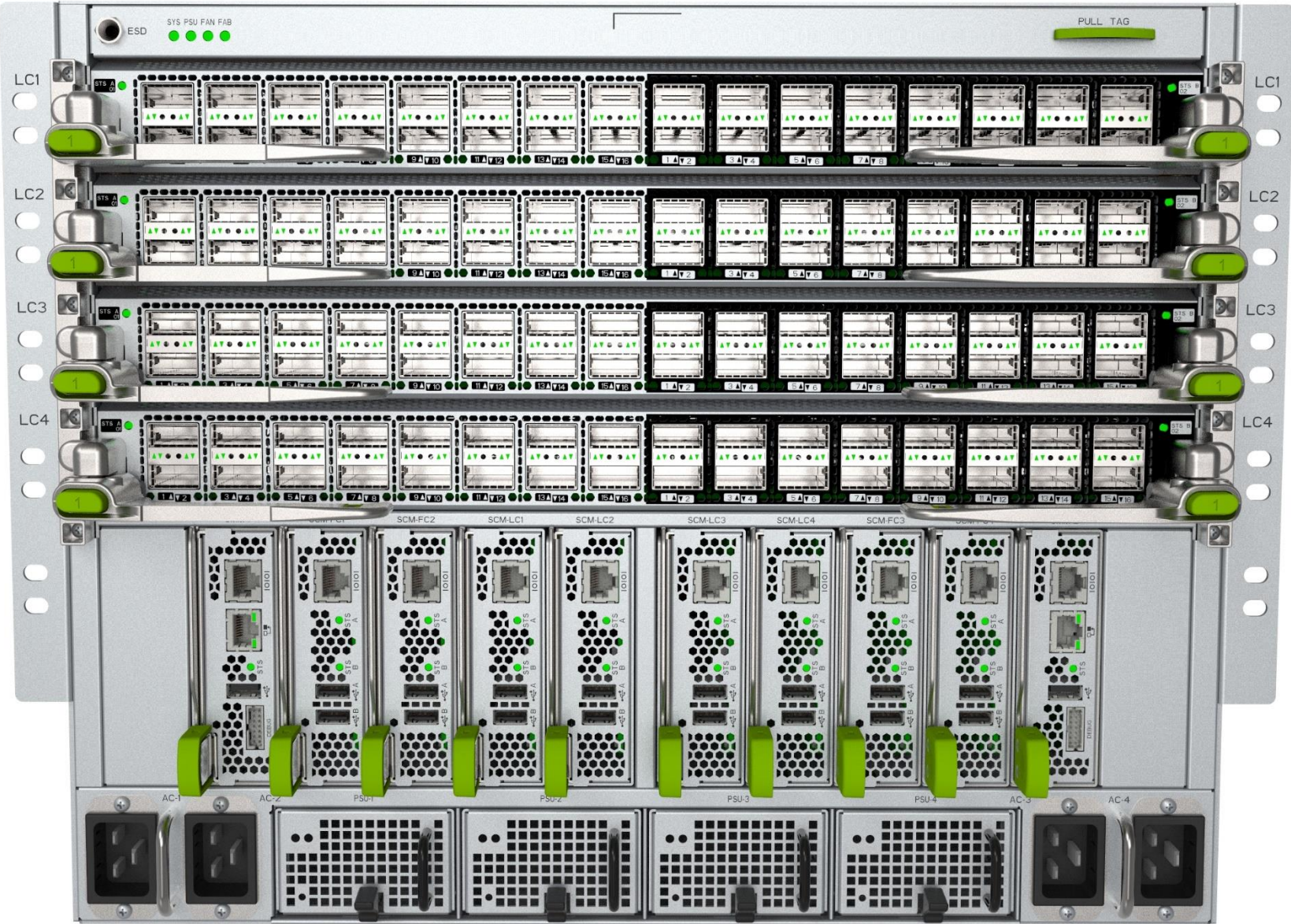
Zhiping Yao, Xu Wang, Facebook

OPEN HARDWARE.

OPEN SOFTWARE.

OPEN FUTURE.





From Wedge 100 to Backpack

Open

- Hardware
 - Contributed to OCP
- Software
 - FBOSS, OpenBMC

Disaggregated

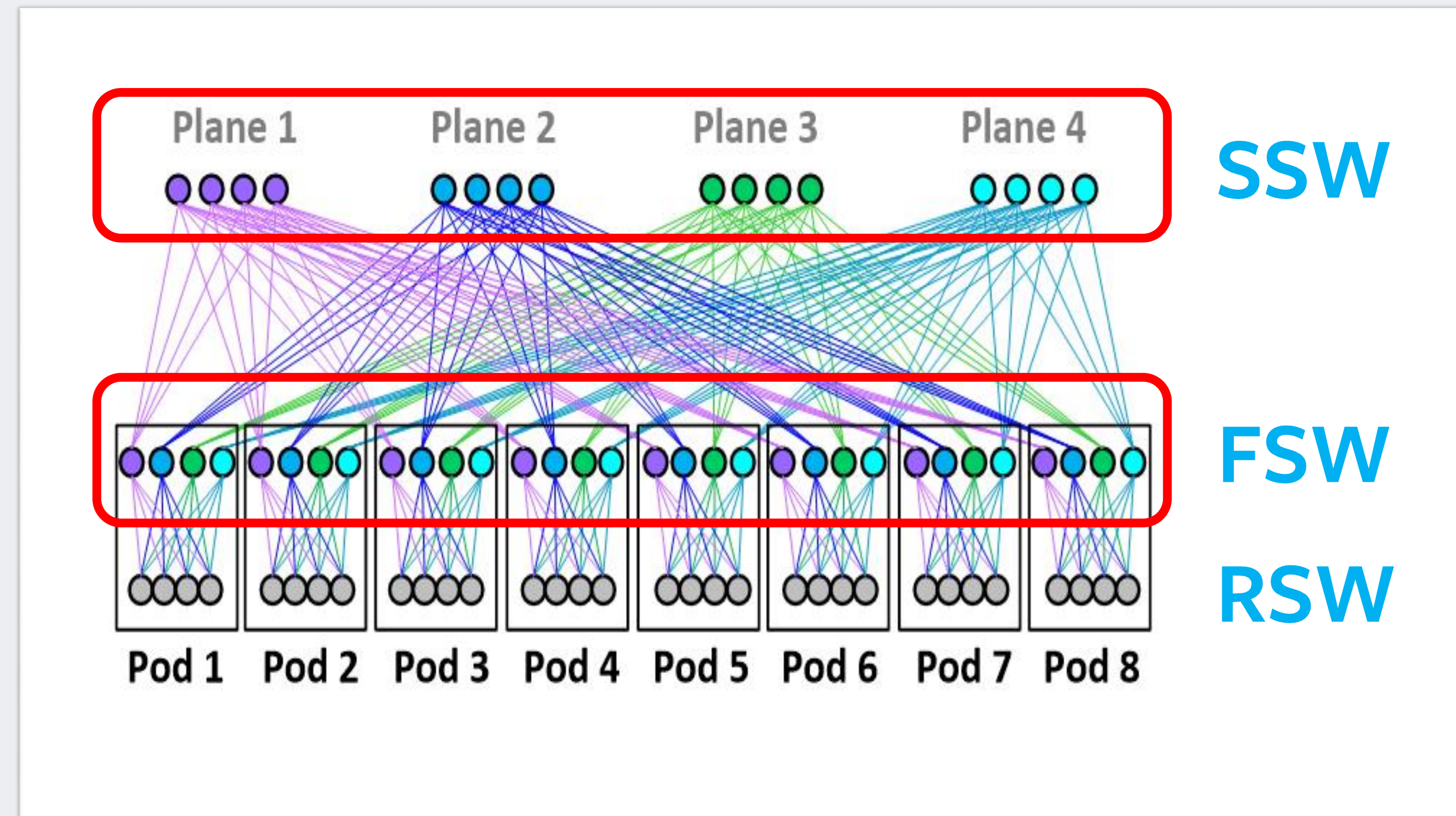
- Software and hardware disaggregation
- Based on a simple building block:
 - Switch Element (SWE)
 - Replicated 12 times
- Physically separated System Control Module and Switch ASIC
- Switch ASIC agnostic

Just a Bunch of Servers

- Managed by BMC
- Deploy and provision like server
- Analytics and monitoring
- Much faster software push for new features

What is Backpack?

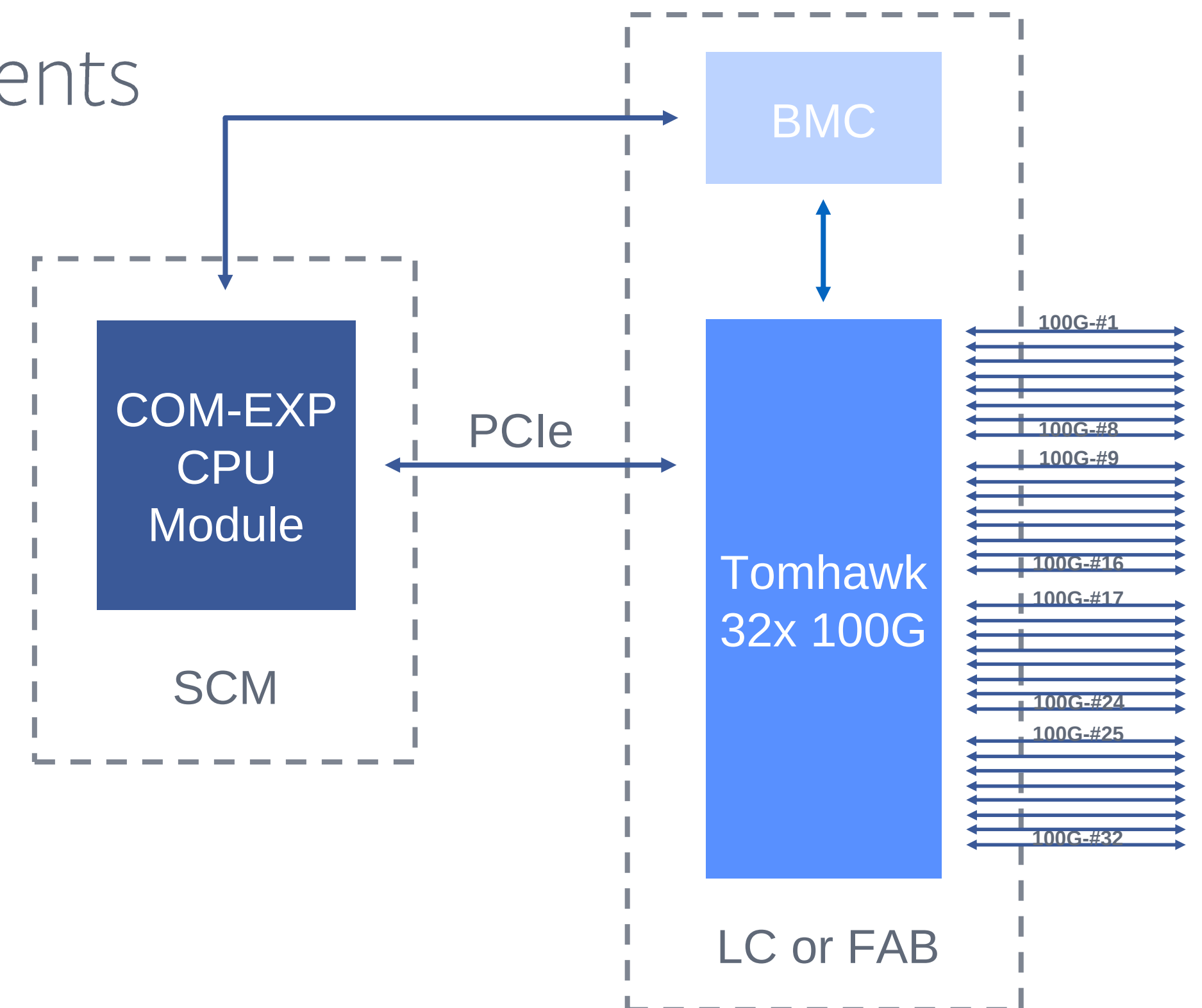
- FB 2nd Generation Modular Switch Platform
- 128 x 100G
- Used in Fabric Switch layer and Spine Switch layer in FB's data center network
- FBOSS and OpenBMC
- Tomahawk Switch ASIC



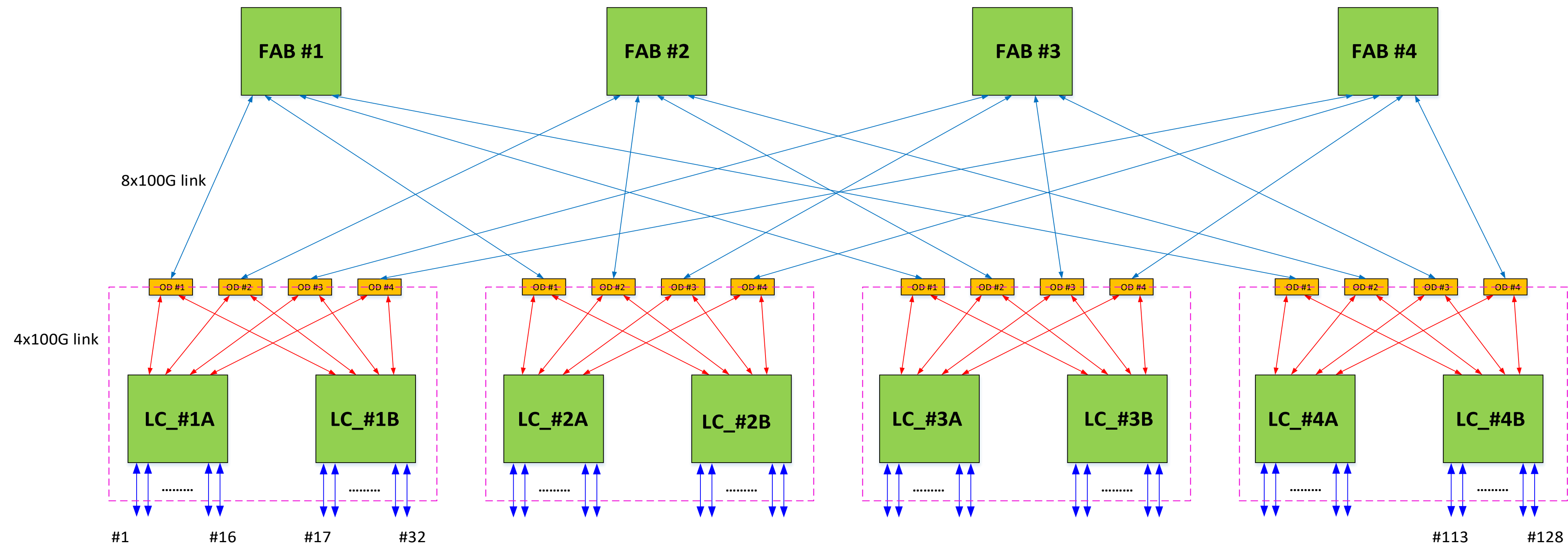
Backpack

Switch Element (SWE)

- A Switch Element consist of three components
 - Tomahawk switch ASIC
 - COM-e CPU module
 - BMC
- Wedge 100 is a single SWE
- 12 switch elements in Backpack
 - Each LC has 2 SWE
 - Each FAB has 1 SWE



System Topology



- Two Stage Spine-leaf architecture
- Fully non-blocking
- 12 Switch Elements

Backpack

Innovative Design

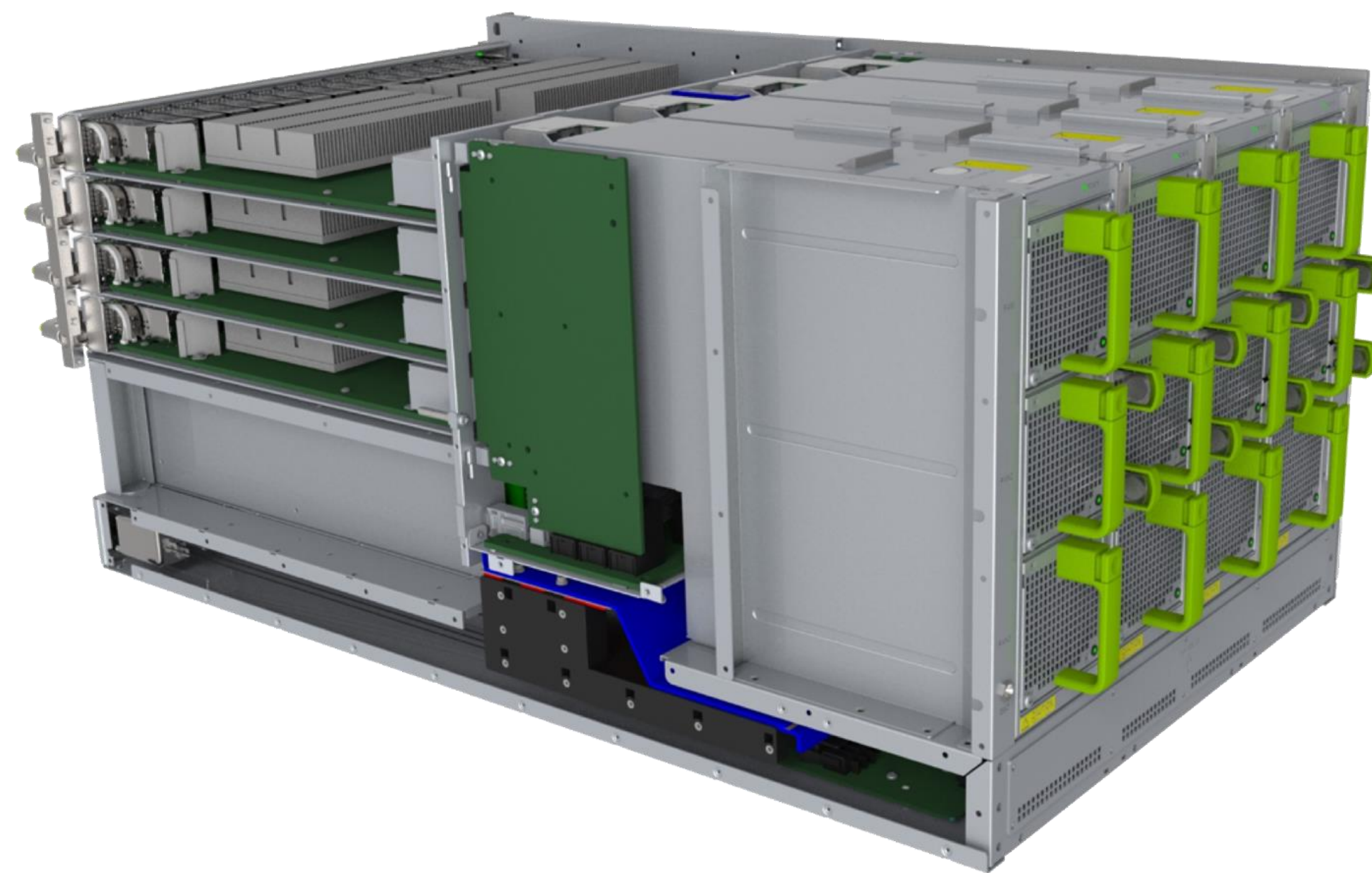
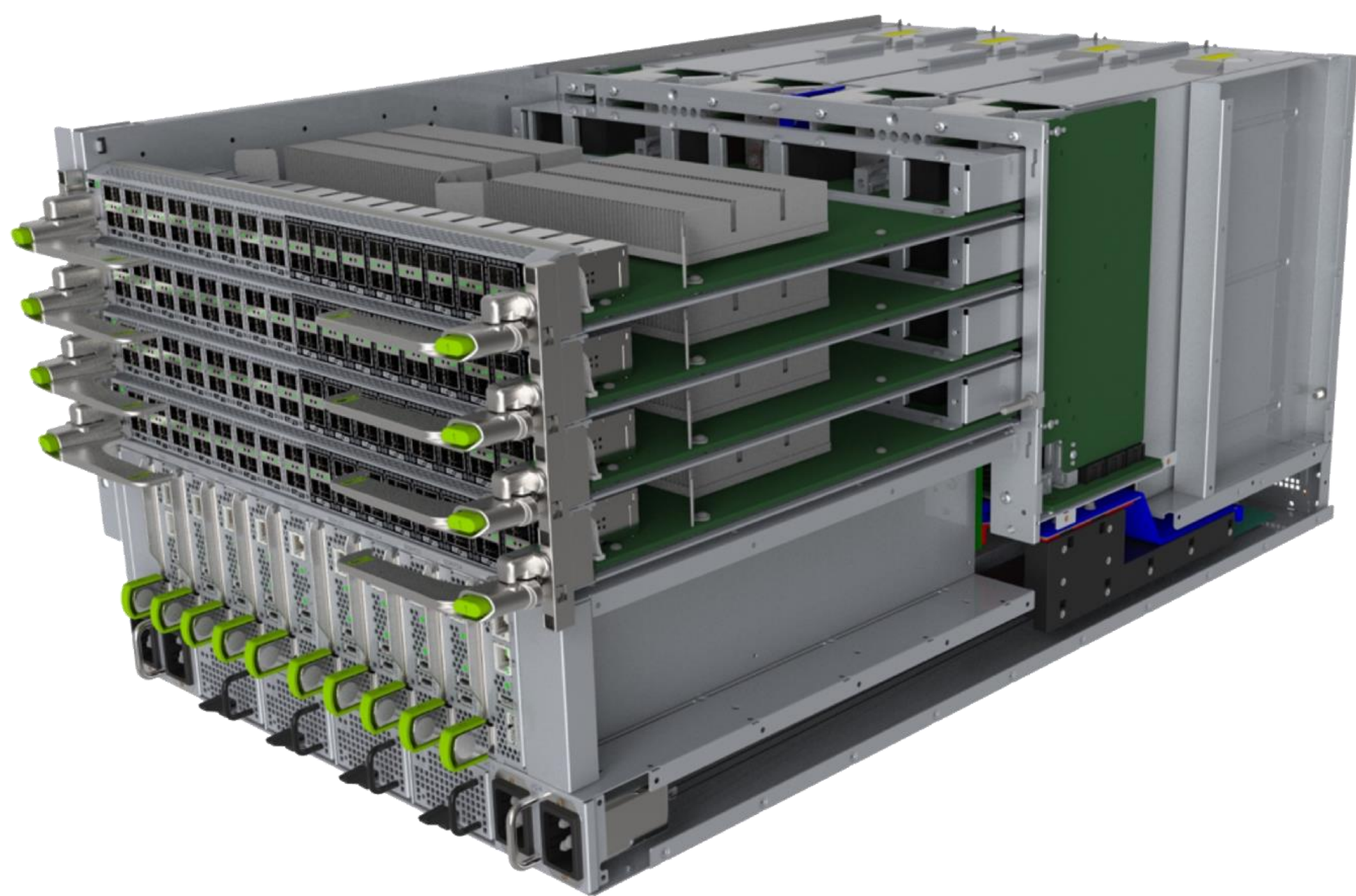
Fully Disaggregated Architecture

- 12 Switch Elements (SWE)
- Separate SCM module for Control CPU

An Orthogonal Direct Chassis Architecture

- All major module cards are designed to be mated orthogonally
- Open up more air channel for a better thermal performance
- Reduce the PCB trace length for better signal integrity

A sophisticated thermal design to support low cost 55C CWDM4 optics

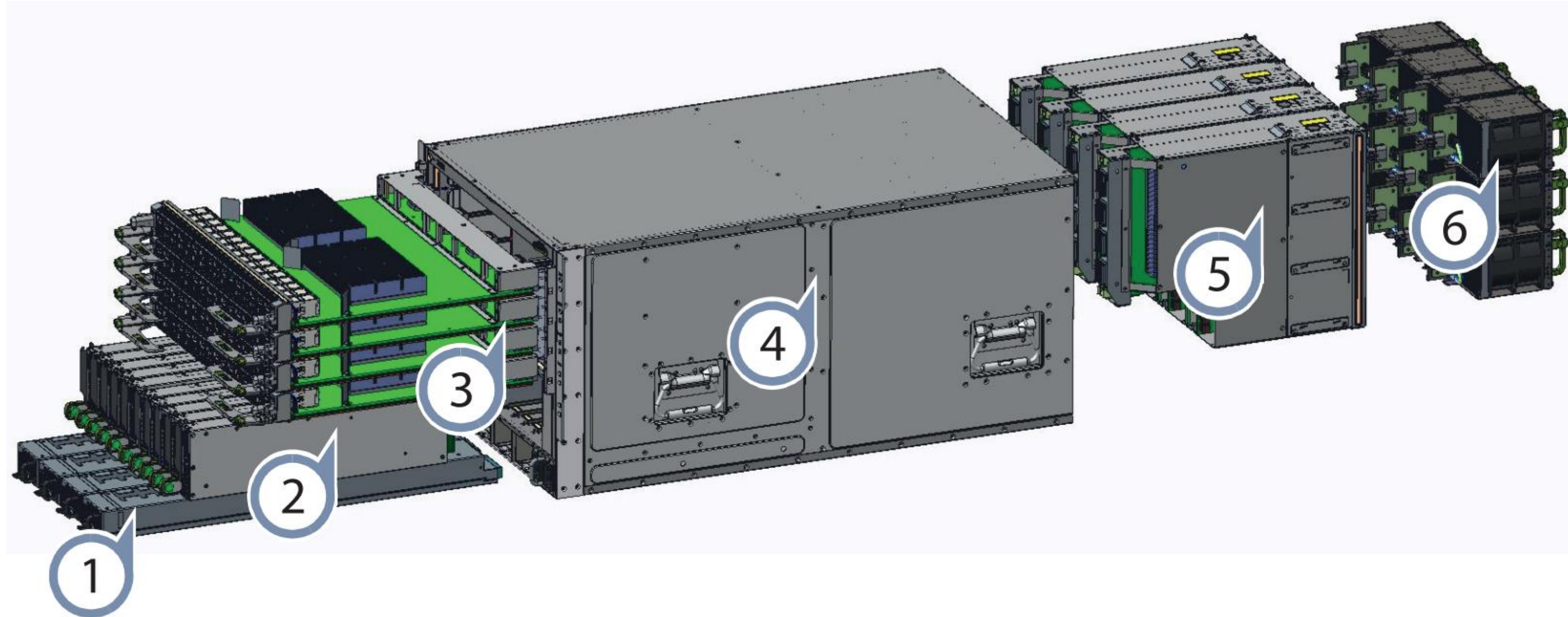


System Components

- Line Cards (LC)
- Fabric Card (FAB)
- System Controller Module (SCM)
- Chassis Management Module (CMM)
- Horizontal Control Plane (HCP)
- Vertical Control Plane (VCP-L, VCP-R)
- Bus Bar Assembly (BBA)
- Horizontal Bus Bar (HBAR)
- Vertical Bus Bar (VBAR)
- Horizontal Power Distribution Board (HPD)
- Fan Control Board (FCB)
- Power Distribution Board (PDB)

Backpack

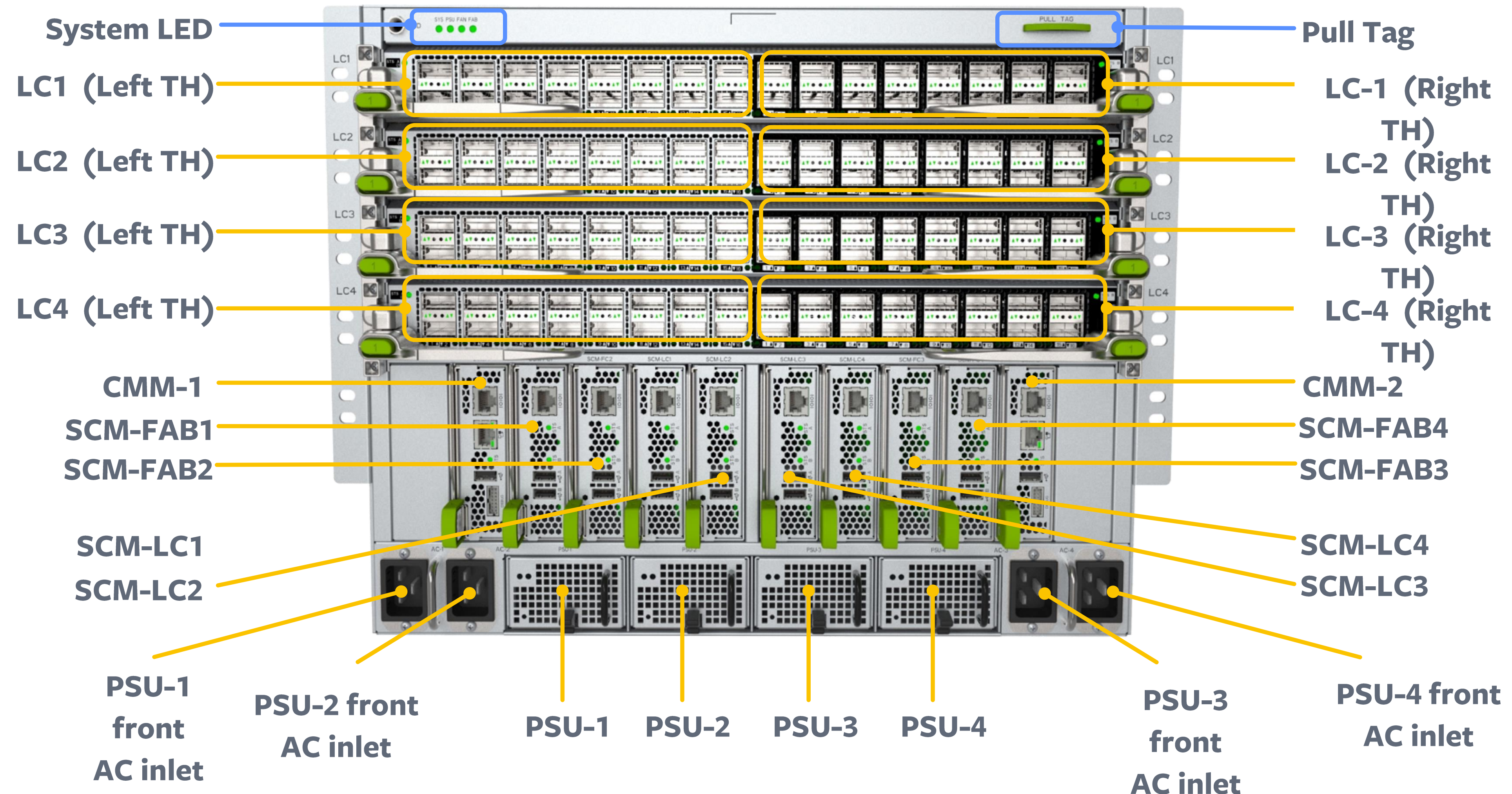
Chassis Assembly



- | | | |
|---------------|-------------------|---------------|
| ① PSU modules | ② SCM&CMM modules | ③ LC modules |
| ④ Chassis | ⑤ FAB modules | ⑥ FAN modules |

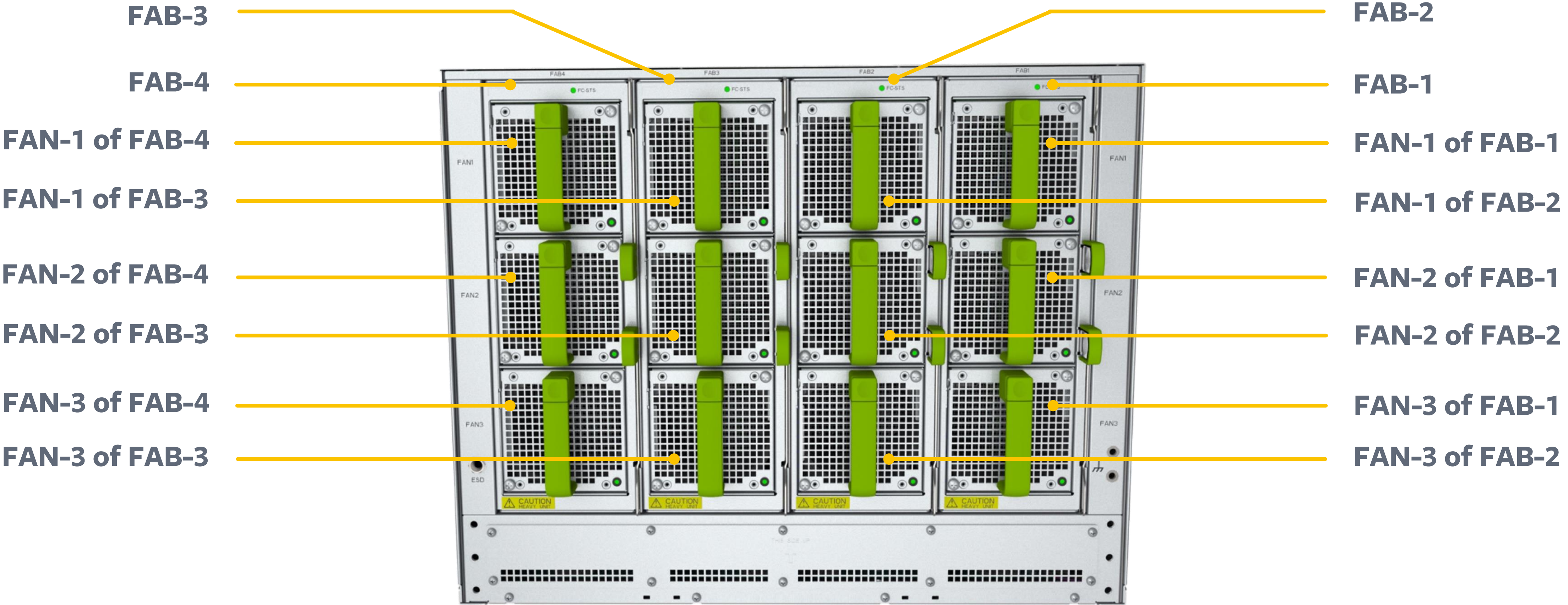
Backpack

Front View



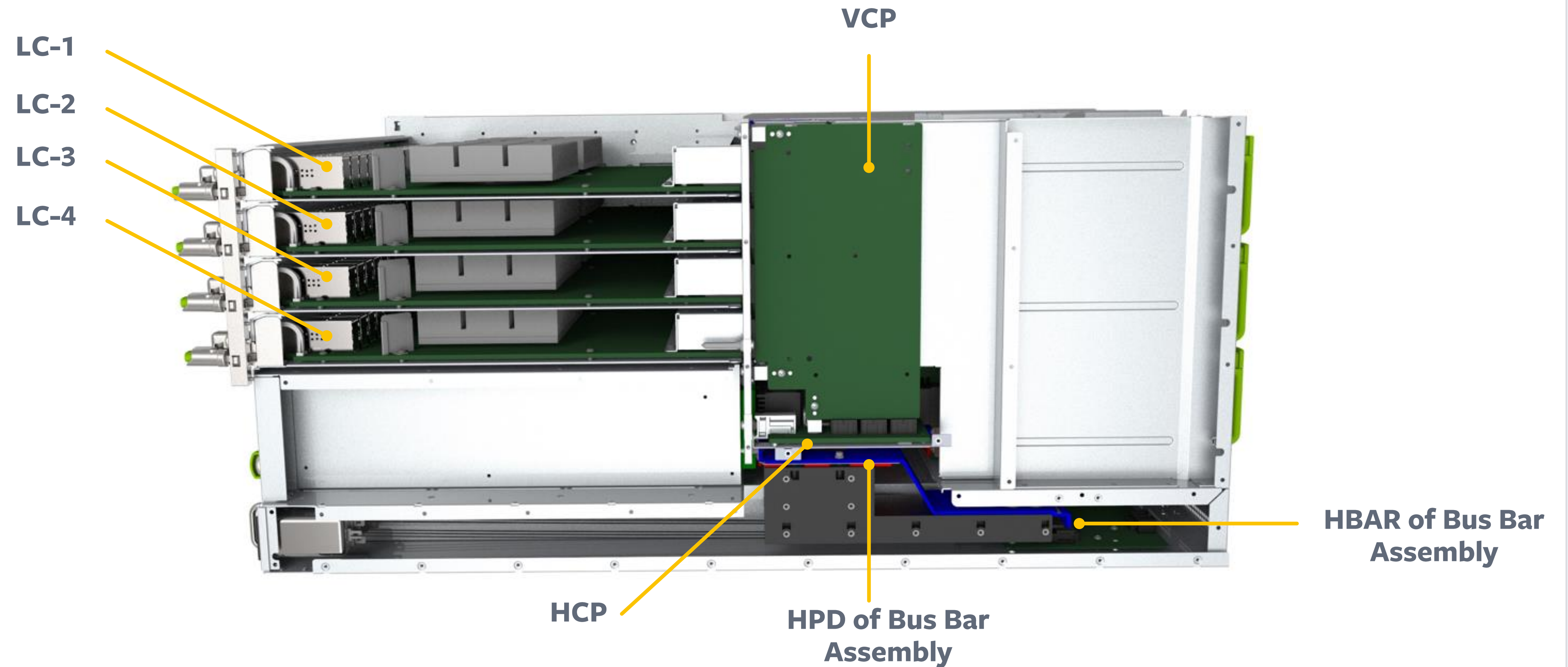
Backpack

Rear View



Backpack

Side View



Backpack

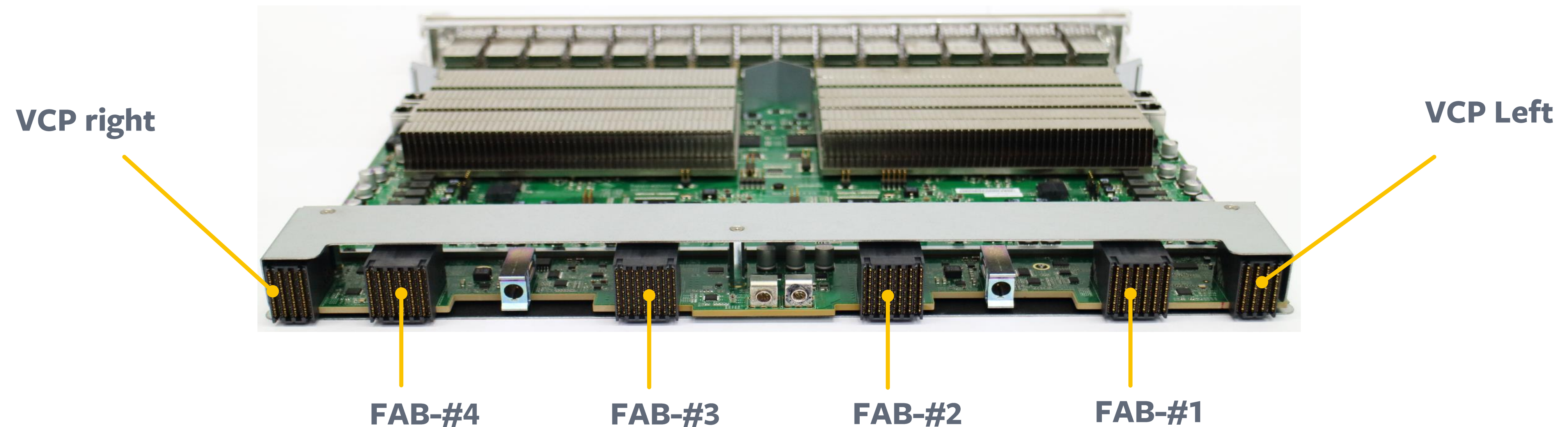
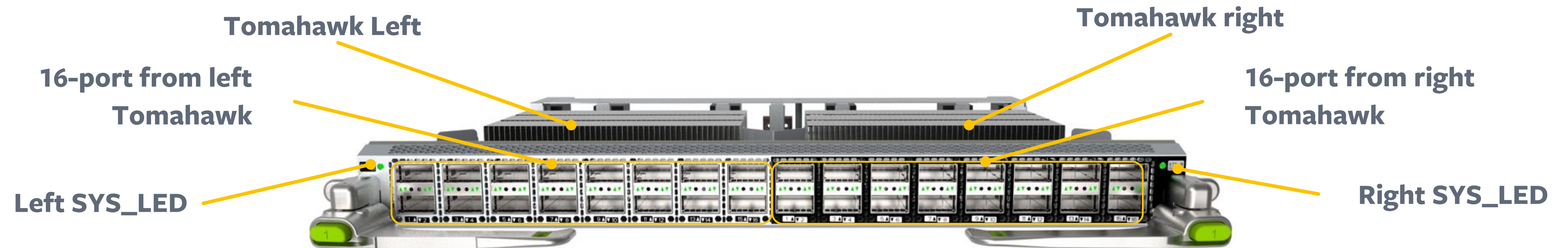
Line Card

- Two Tomahawk 3.2T ASIC
- 32 QSFP28 100G Ports
- Four DMO Connector to FAB for data plane signals
- Two DMO Connector to VCP for control plane signals



Backpack

Interface of Line Card



Backpack

Fabric Card

- One Tomahawk 3.2T ASIC
- Fan Control Board (FCB) is mezzanine card of FAB
- Four 6 x 12 DMO Connector to Four LC for data plane signals
- One 6x8 DMO Connector to HCP for control plane signals



Backpack

Components of a Fabric Card

Fan Control Board is mounted
on the bottom side of FAB

Tomahawk ASIC

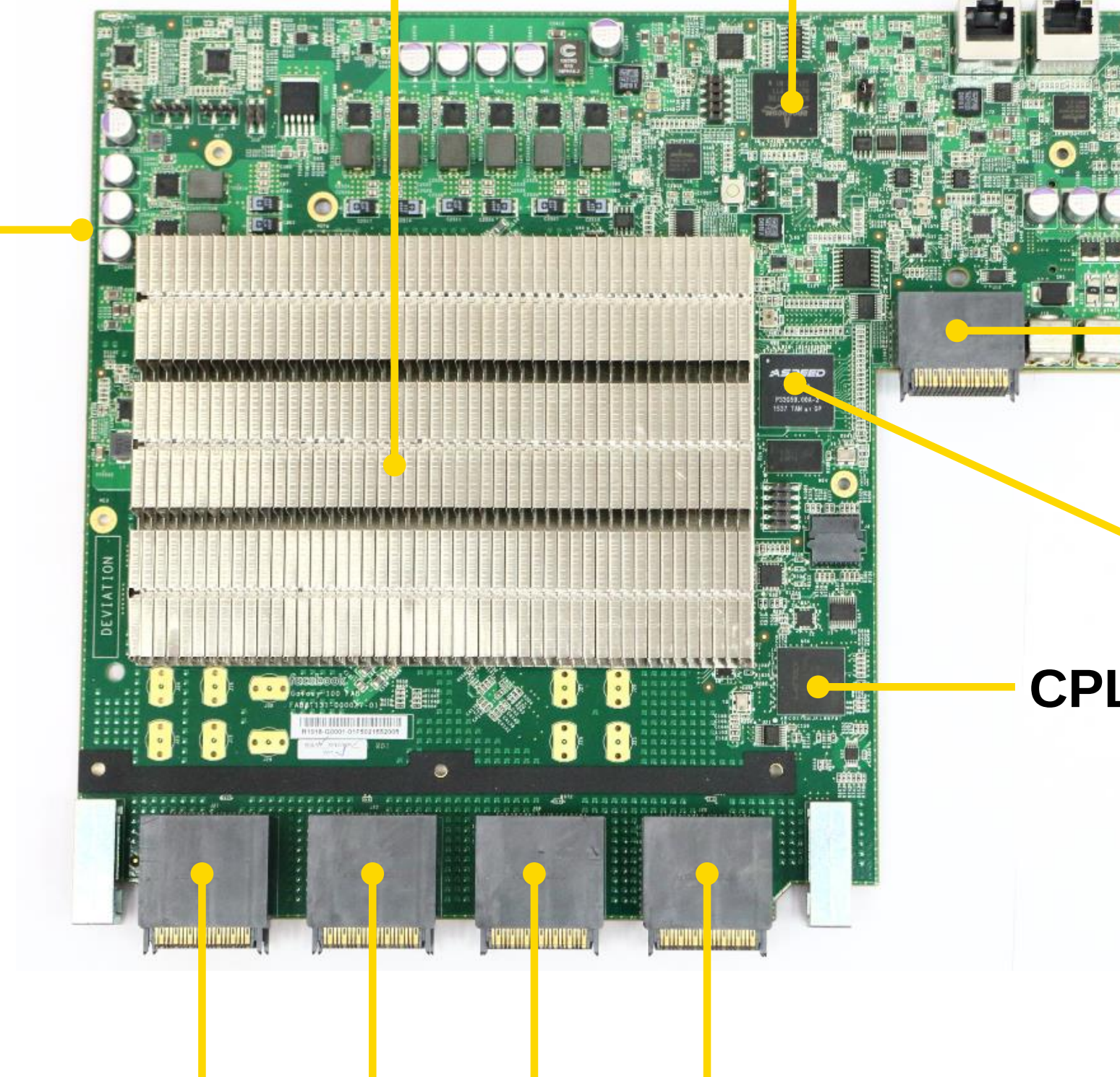
OOB 8-port GBE Switch

DMO Connector To SCM

BMC

CPLD

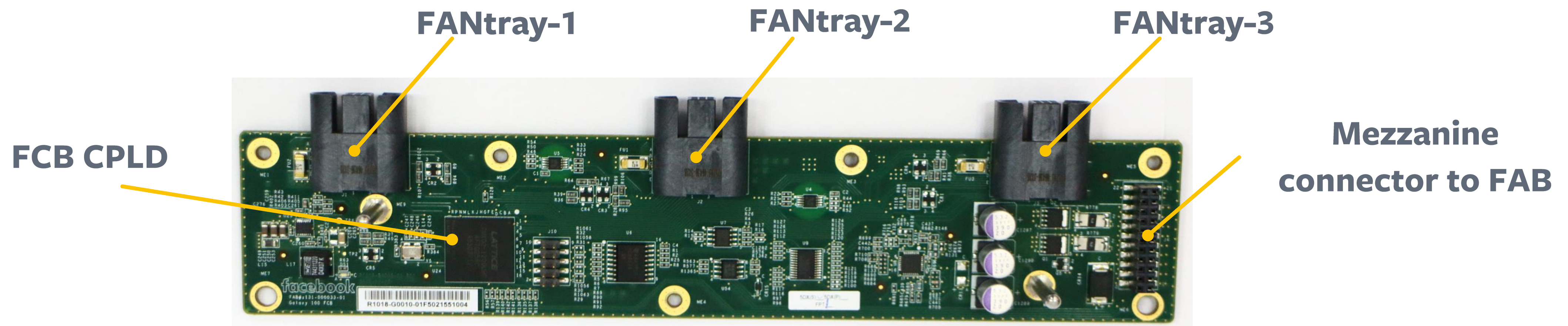
LC-#1 LC-#2 LC-#3 LC-#4



Backpack

Fan Control Board (FCB)

- Mezzanine Card of Fabric Card(FAB)
- Support 3 fan-tray
- FCB CPLD control fan speed, CMM can control FCB CPLD via system I2C bus



Backpack

Fabric Card (FAB) Assembly

- FAB Main Board + FCB Mezzanine Board
- Fan-tray plugged into FAB assembly and Mate to FCB



Backpack

Chassis Management Module (CMM)

- AS2540 BMC
- OOB 16 port Switch BCM5396
- Console UART MUX for all LC and FAB COM-e CPU and BMC CPU
- Chassis Management I2C bus



Backpack

System Control Module (SCM)

- Two COM-e CPU Module Cards
- COM-e CPU Module has one to one mapping to SWE
- 6x12 DMO connector to HCP
- SCM-LC has two COM-e mounted, SCM-FAB has only one COM-e mounted





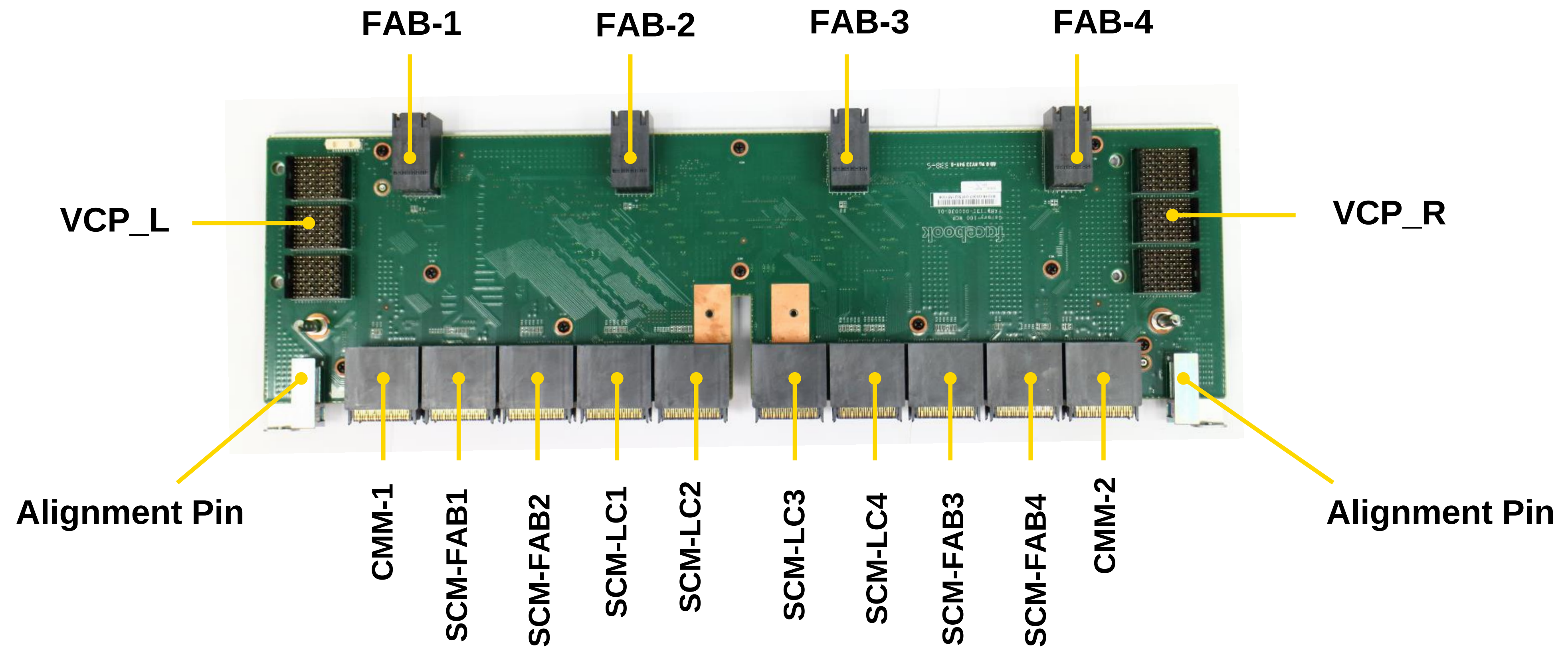
SCM-LC:
SCM for Line Card



SCM-FAB:
SCM for Fabric Card

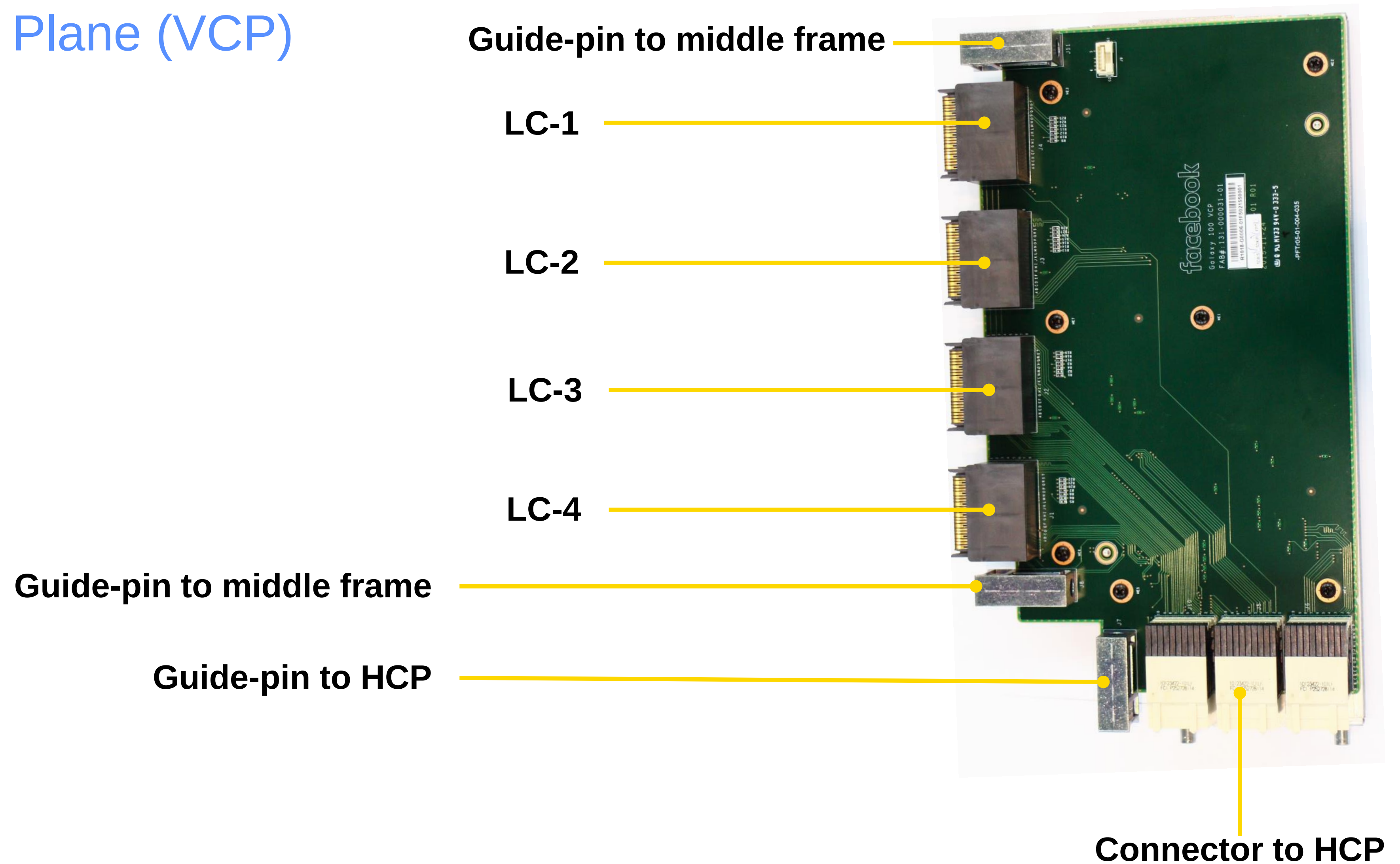
Backpack

Horizontal Control Plane (HCP)



Backpack

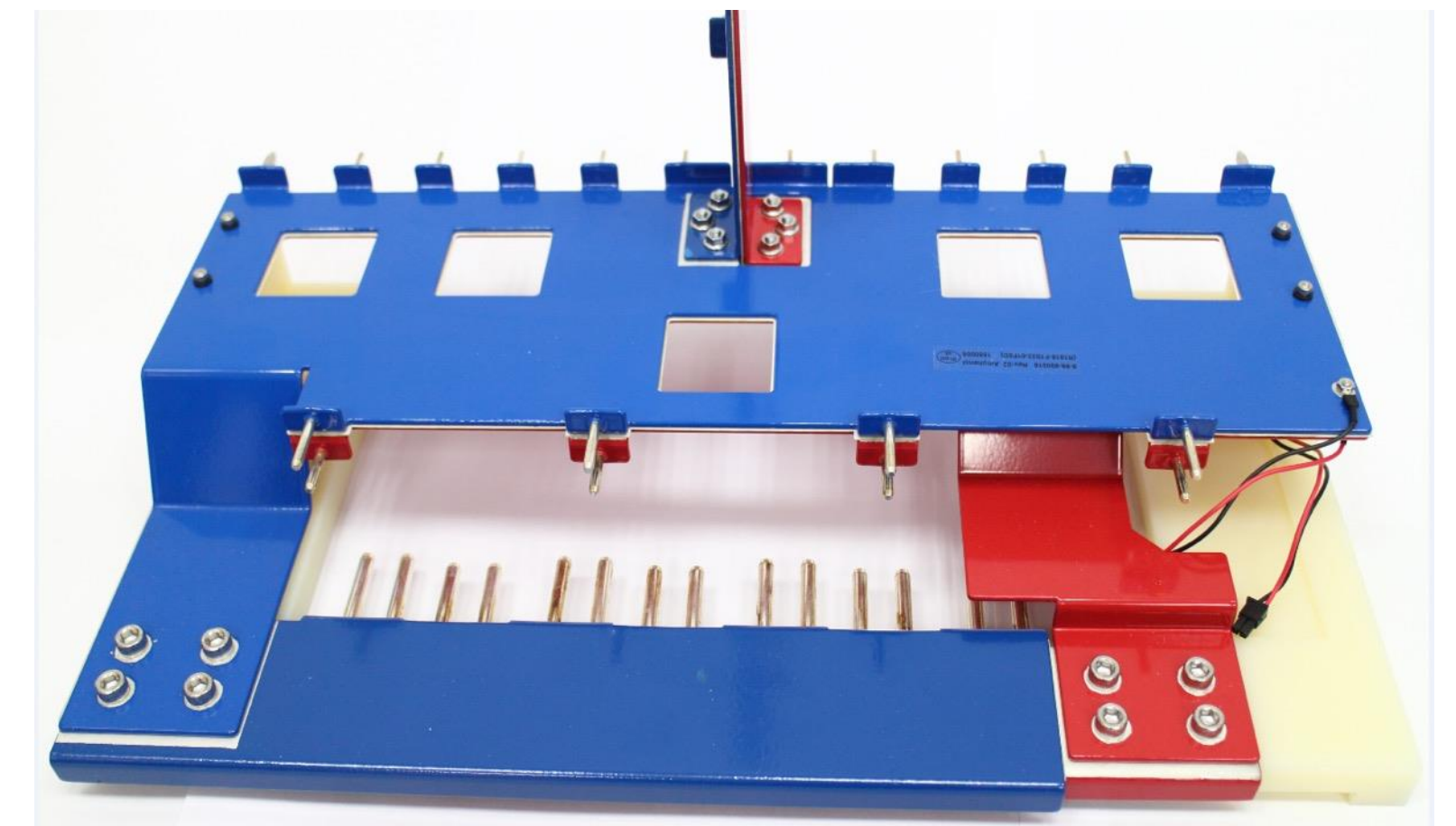
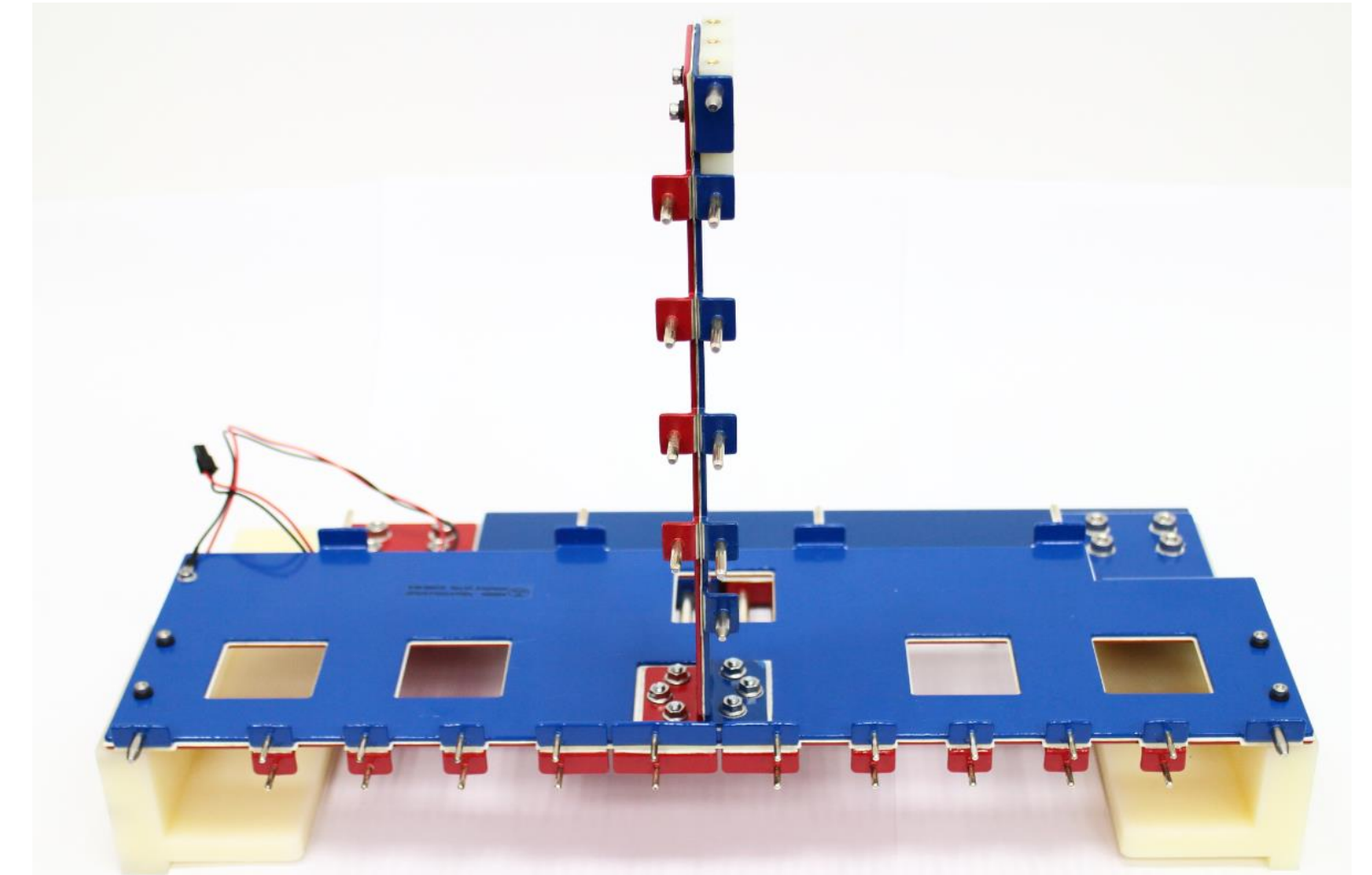
Vertical Control Plane (VCP)



Backpack

Bus BAR Assembly

- BBA(Bus Bar Assembly) consists of
 - HBAR: Horizontal Bar
 - HPD: Horizontal Power Distribution
 - VBAR: Vertical Bar
- HBAR Connect to PSU
- HPD for CMM, SCM and FAB
- VBAR for LC
- Sense wire to PDB



Backpack

Typical Power Consumption

PWM	20%	30%	40%	50%	60%	70%	80%	90%	100%
Normal condition	2670	2697	2726	2800	2896	3124	3290	3550	3729
One fan failure condition	2550	2627	2703	2768	2822	3056	3223	3428	3645

- Unit is watt
- Measured with 3.5W eload

Backpack

Thermal design

- Support 55C optic at ambient 35C environment
- 12 fan-tray on the rear panel
- Thermal Improved LC front panel design for better QSFP28 cooling
- Multiple on-board temperature sensor to monitor thermal healthy status of the system



Backpack

Optical transceivers

- QSFP28 100G optics
 - CWDM4
 - Can support other MSA, such as SR4, LR4, CLR4, etc.
- QSFP+ 40G optics
 - QSFP+ 40G SR4 optic (multi-mode fiber OM3/OM4)
 - QSFP+ 40G LR4 optic (single mode fiber)



Backpack

IN_USE!



Backpack

OCP Contribution

- Backpack OCP contribution
 - Specification
 - Design package
- Our ODM partner: Celestica
- Third-party switch software
 - Cumulus
 - SnapRoute
 - Broadcom



OPEN
Compute Project

<http://www.opencompute.org/wiki/Networking/SpecsAndDesigns>

Wedge 100S

Wedge 100S

An Improved Wedge 100

19-in SKU

For regular rack



21-in SKU

for Open Rack



Wedge 100S

Open 32x100GE TOR Switch

**Facebook's second generation
Open TOR Switch based on
Tomahawk Switch ASIC**

**Standard COM-Express
module as control plane
CPU**

BMC: Aspeed AST1250

**Supports Open BMC and
FBOSS**

Wedge 100S

Benefits and Applications



Connecting next generation servers

- Increased link speed with 50G and 25G NIC
- Increased compute density
- Lower oversubscription ratio



Improved system thermal design

- Enable adoption of 55°C 100G CWDM4 optic transceivers



Support for Open Rack V2

- Powered by Rack Bus Bar
- Rack Monitor



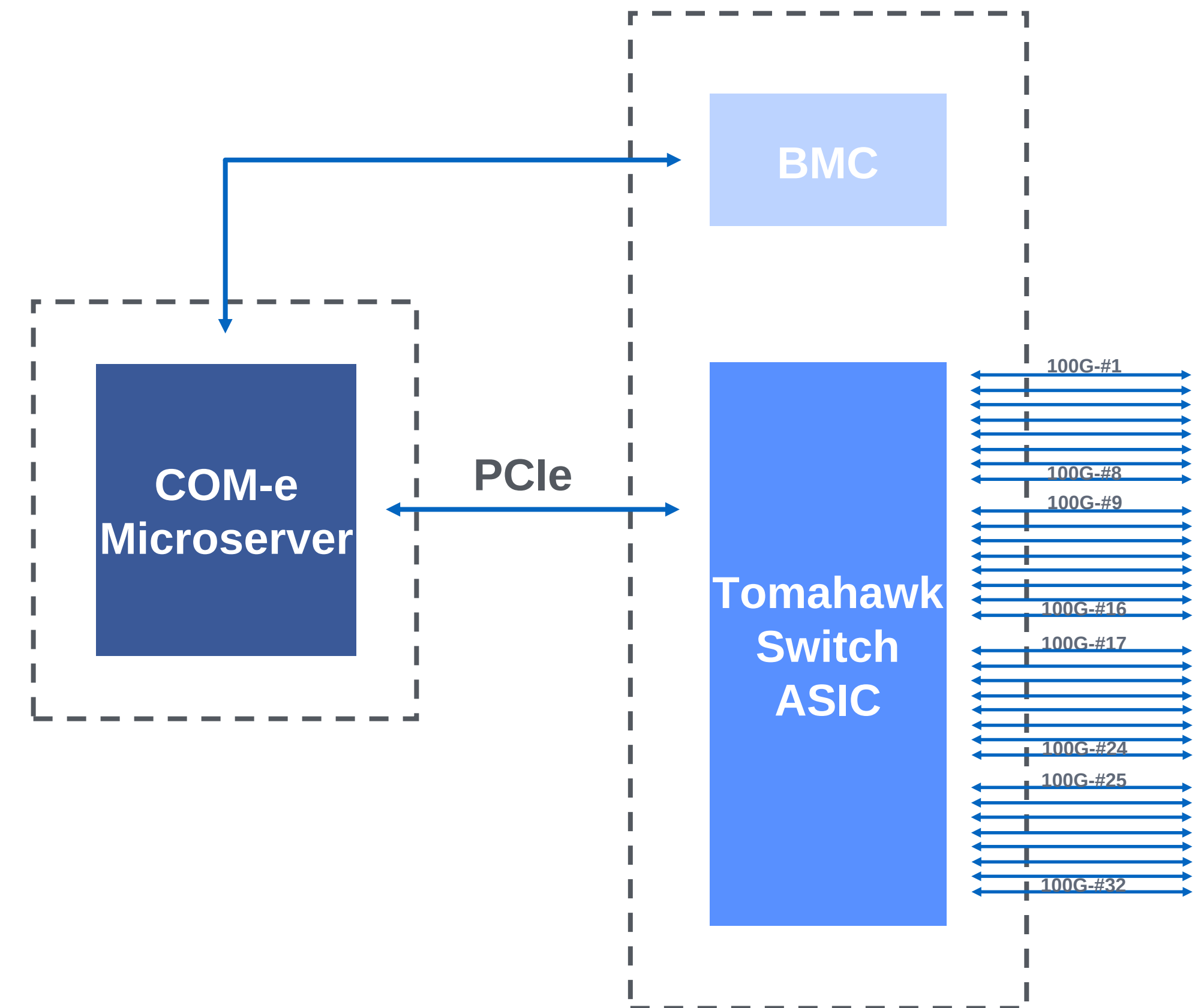
Broadwell-DE micro-server

- Higher performance
- Better RAS
- Intel Trusted Execution Technology (TXT)

Wedge 100S

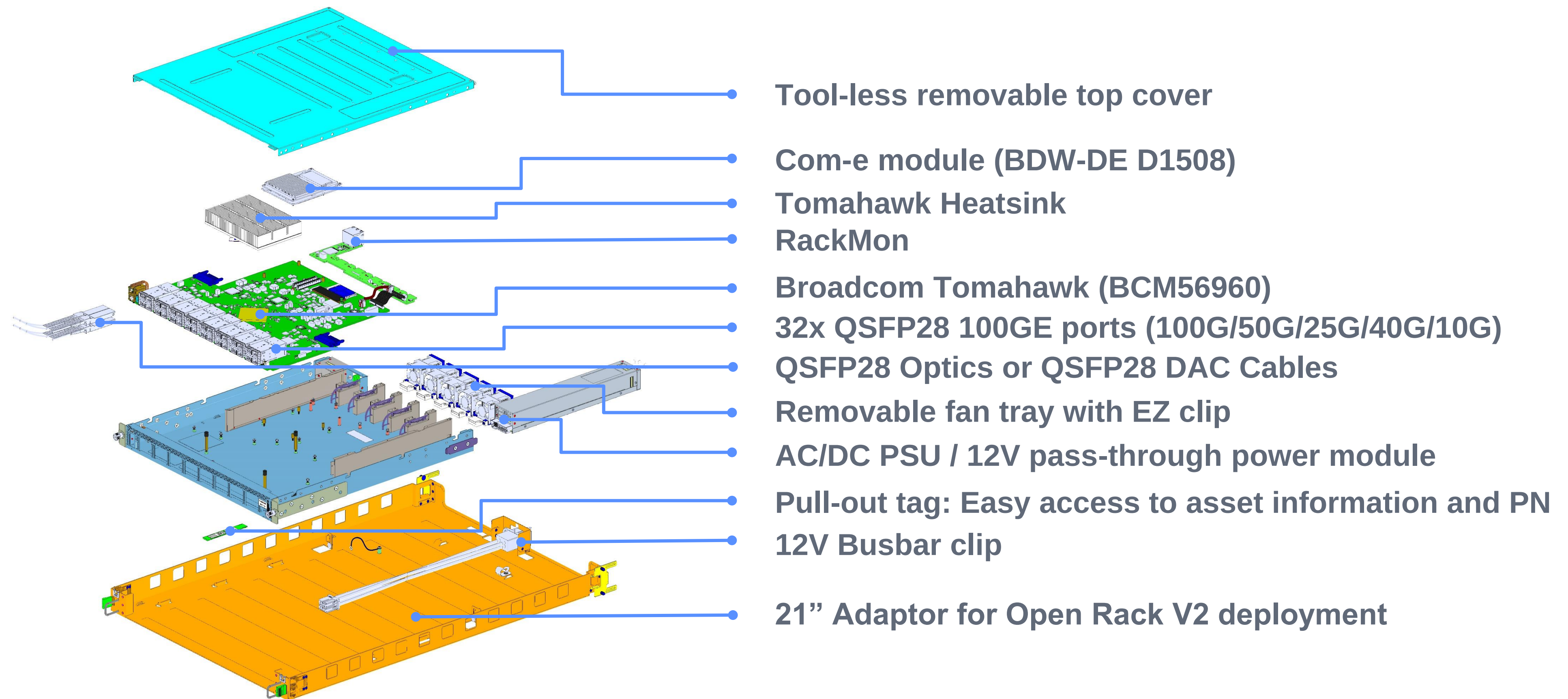
Open Hardware

- Hardware architecture: fully open
- Software: FBOSS and OpenBMC
- Manageability: operated just like a server
- Density: 32x QSFP28 100GE support
- Hardware building blocks
 - Broadcom Tomahawk ASIC, 32x 100GE, 3.2Tbps
 - COM-Express Microserver, Intel Broadwell-DE D1508
 - BMC, Aspeed AST1250



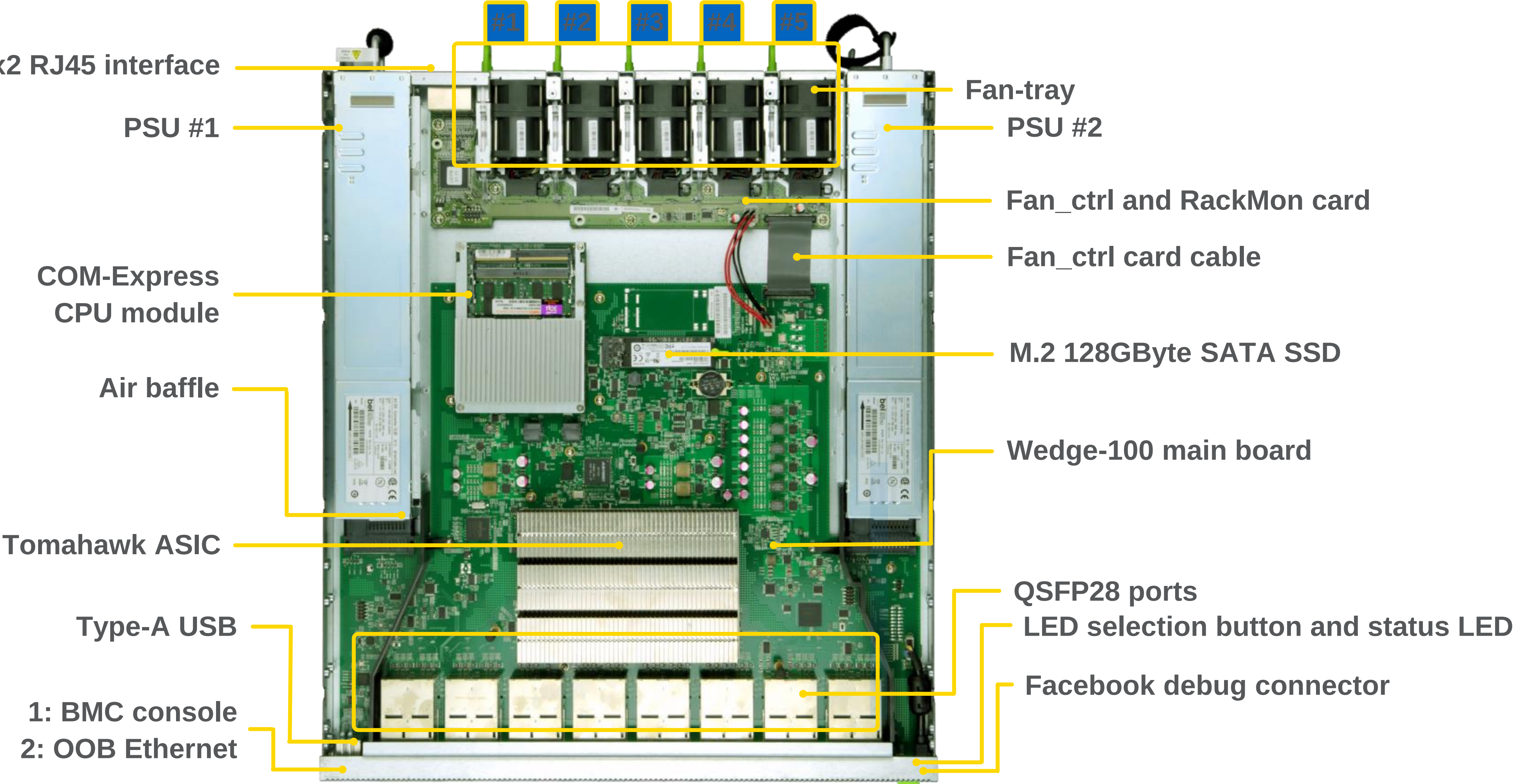
Wedge 100S

Component View



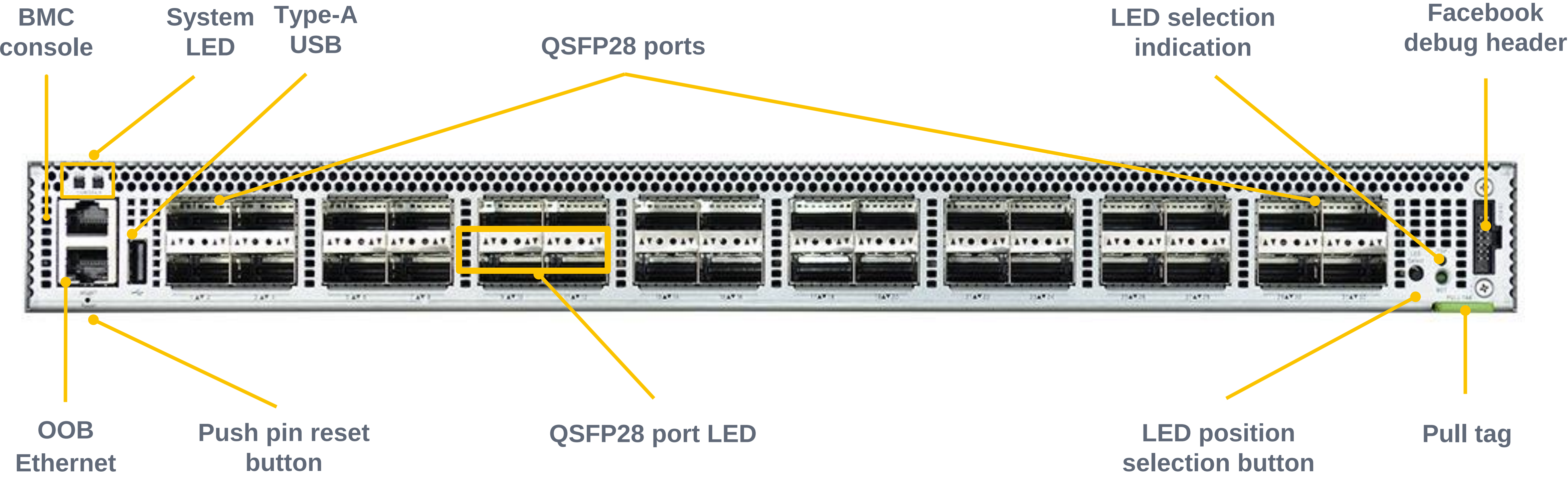
Wedge 100S

Top View



Wedge 100S

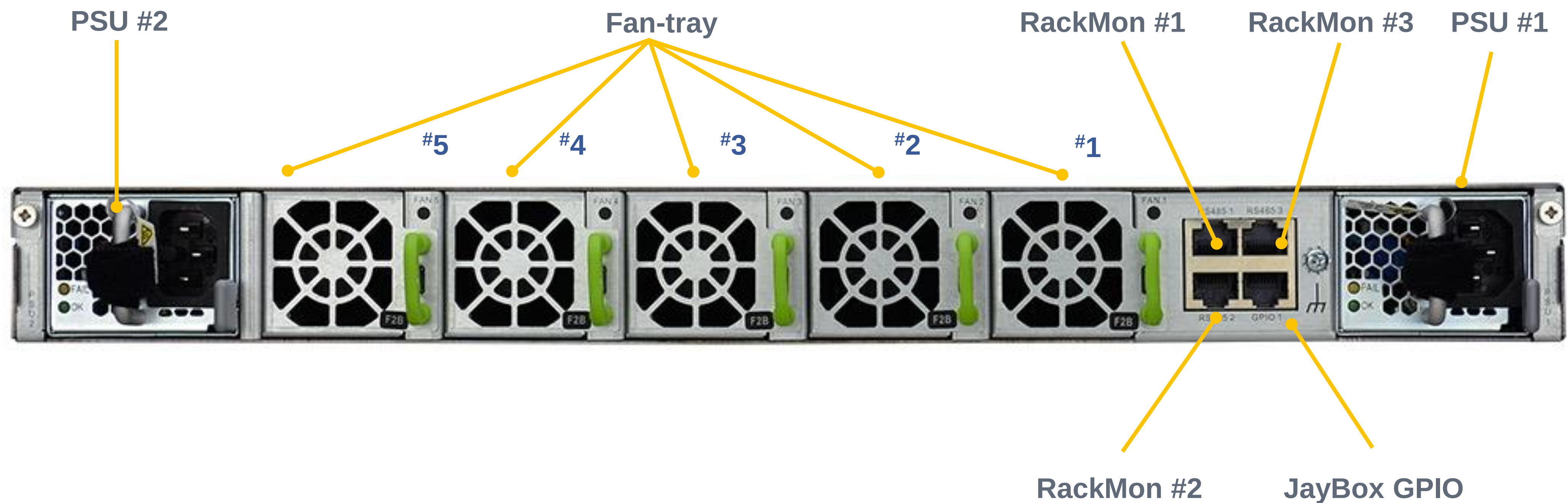
Front View



Wedge 100S

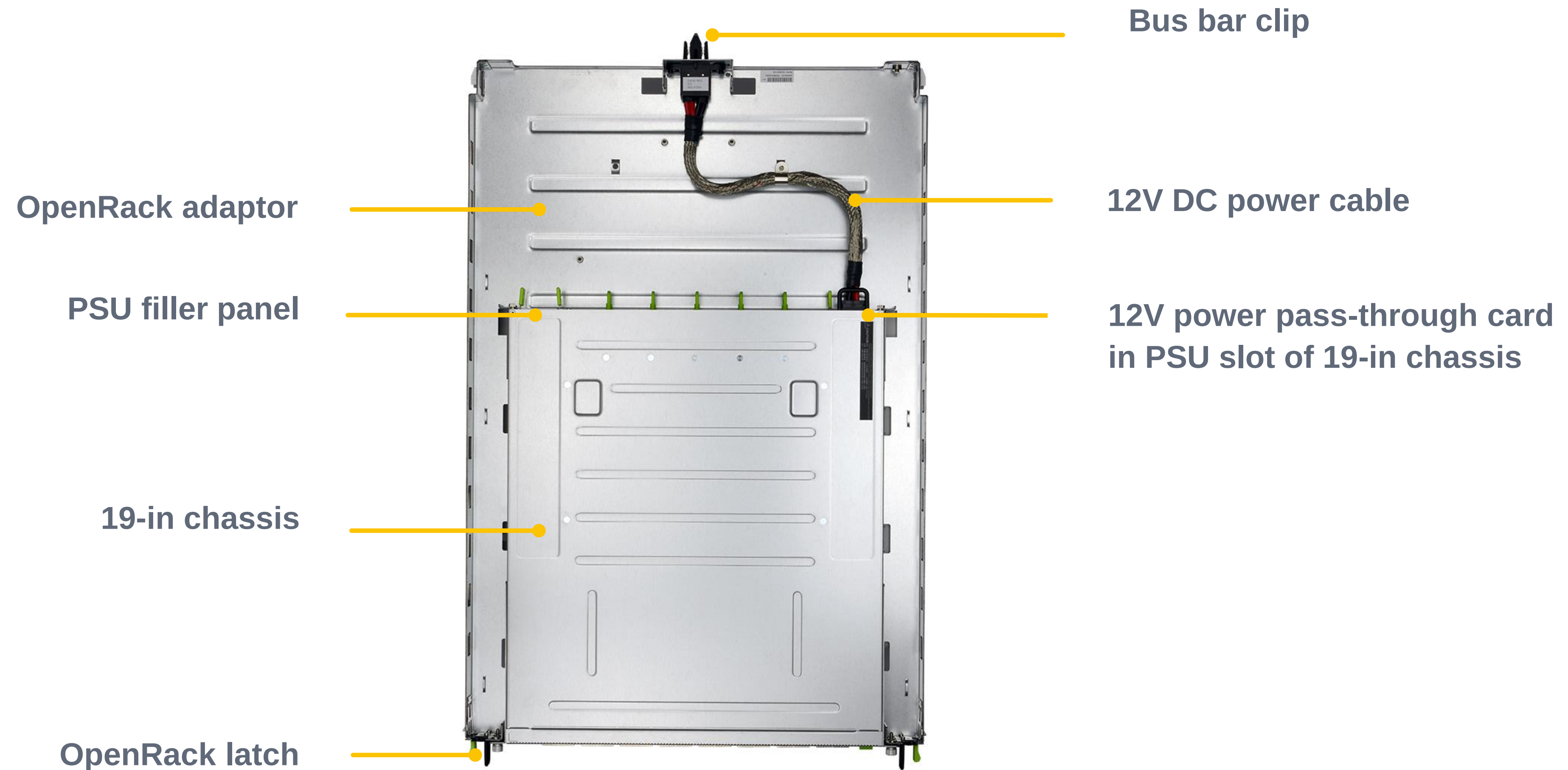
Rear View

- Fan Tray changed to latch design for tool less insertion or removal
- Fan LED now can be displayed on rear panel to indicate fan tray failure.



Wedge 100S

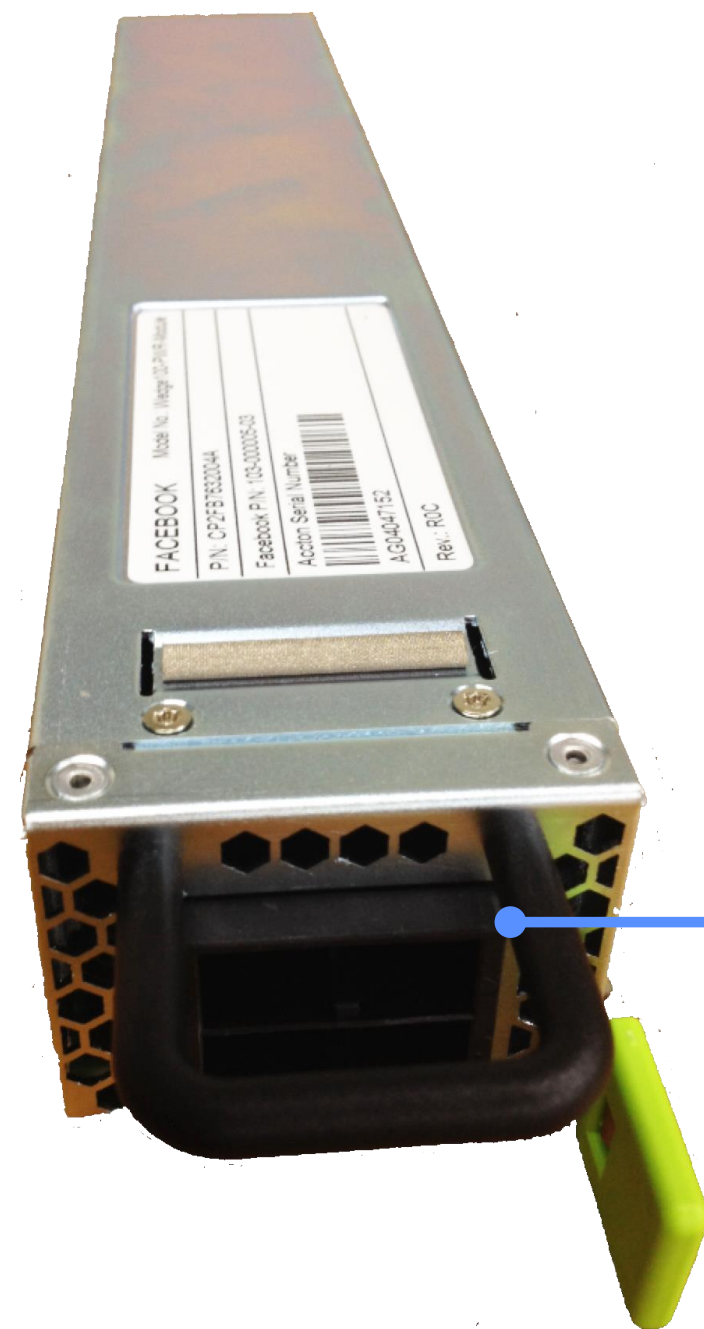
21-in Adaptor Tray



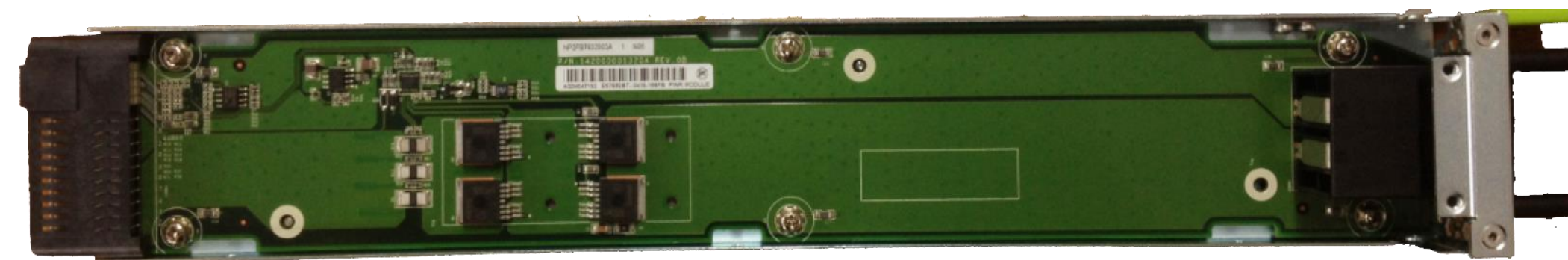
Wedge 100S

OCP Power Pass-through Card

- Power pass-through card with the same form factor as the AC/DC PSU



12V bus bar power
input connector



Connector to
Wedge-100 main board



Wedge 100S

QSFP28 Port Modes

1x100G: use CWDM4 QSFP28
optic to connect to 100G
aggregation switch

2x50G: use QSFP28-
2xQSFP28 DAC cable to
connect to 50G NIC

4x25G: use QSFP28-4xSFP28
fan-out cable to connect to
25G NIC

1x40G: support QSFP+ 40G
SR4 optic and QSFP+ 40G
LR4 optic

4x10G: use QSFP-4xSFP fan-
out cable to connect to 10G
SFP+ NIC



Wedge 100S

Optical Transceivers

- **QSFP28 100G optics**
 - CWDM4
 - Can support other MSA, such as SR4, LR4, CLR4, etc.
- **QSFP+ 40G optics**
 - QSFP+ 40G SR4 optic (multi-mode fiber OM3/OM4)
 - QSFP+ 40G LR4 optic (single mode fiber)



Wedge 100S

DAC Cables

- **25Gbps/lane class cables**
 - QSFP28 100G to QSFP28 100G cable, 1M, 2M, and 3M
 - QSFP28 100G to two QSFP28 50G cable, 1M, 2M, and 3M
 - QSFP28 100G to four SFP28 25G cable, 1M, 2M, and 3M
- **10Gbps/lane class cables**
 - QSFP+ to 4xSFP+ 40G fan-out cable



Wedge 100S

Thermal Design

1

Supports 55°C optic
in 35°C environments

2

Five fan-trays on
the rear panel

3

PSUs have a separate air
channel to avoid air recycling

4

Multiple on-board temperature
sensors to monitor thermal
status of the system



Wedge 100S

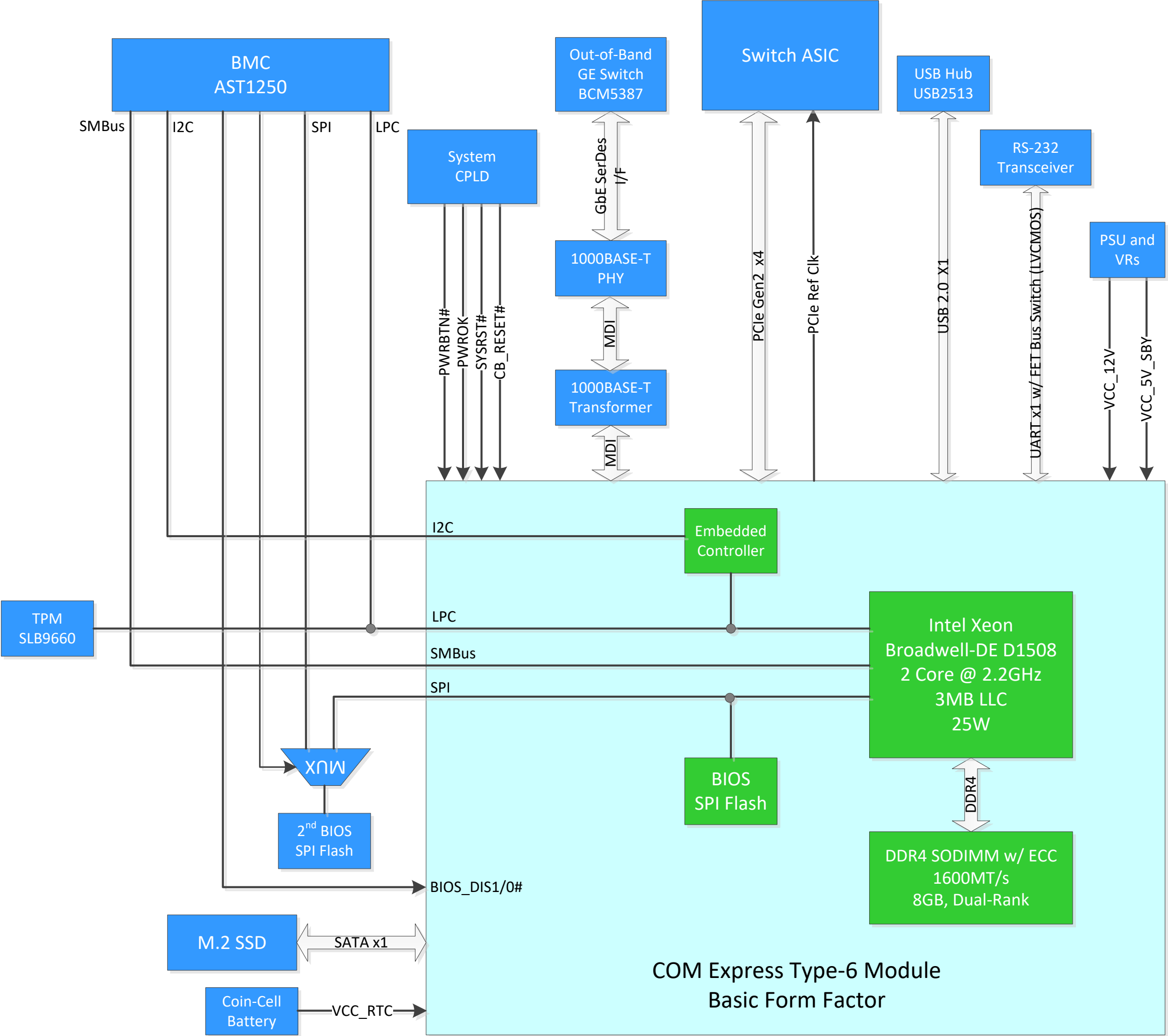
Fan Tray

- Screwless latch design for easy maintenance
- Powerful Counter-Rotating fans
- LED on the rear panel



Wedge 100S

COM-Express Module Interfaces



Wedge 100S

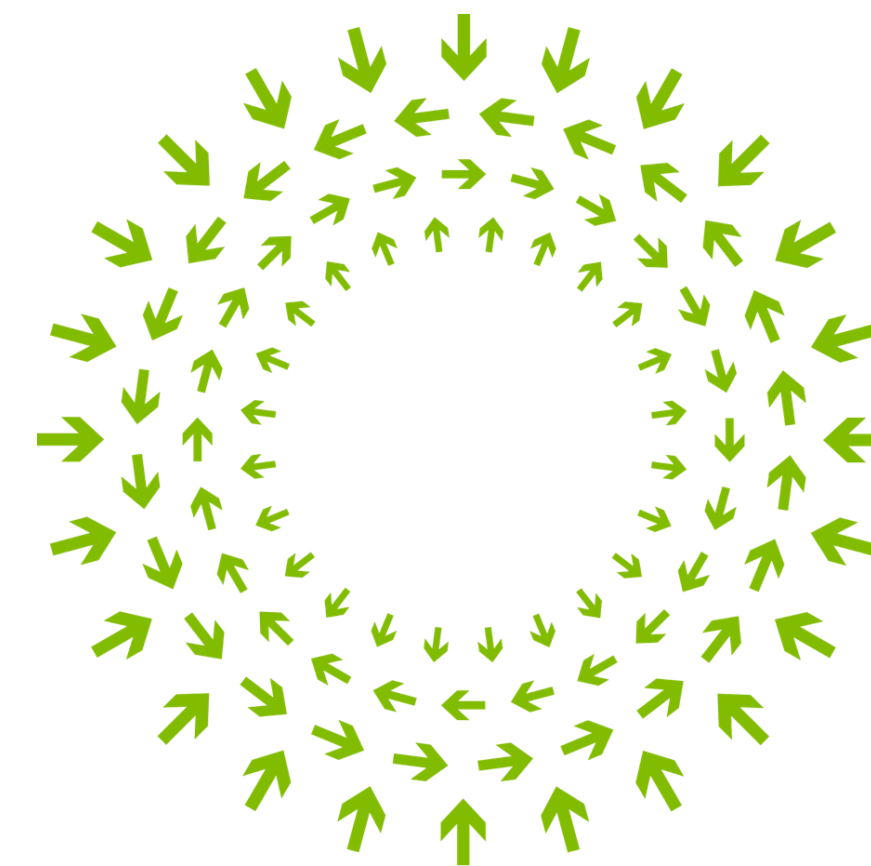
CPU SKU Selection Criteria and Process

- ToR power efficiency – important for data center applications
- Minimize increase in power consumption while migrating to Broadwell-DE
- No performance degradation
- Better performance if needed

Wedge 100S

OCP Contribution

- Wedge100S specification
- Wedge100S design package



OPEN
Compute Project

<http://www.opencompute.org/wiki/Networking/SpecsAndDesigns>

Wedge 100B



Wedge 100B Series

Akin Koyuncuoglu, Barefoot Networks

OPEN HARDWARE.

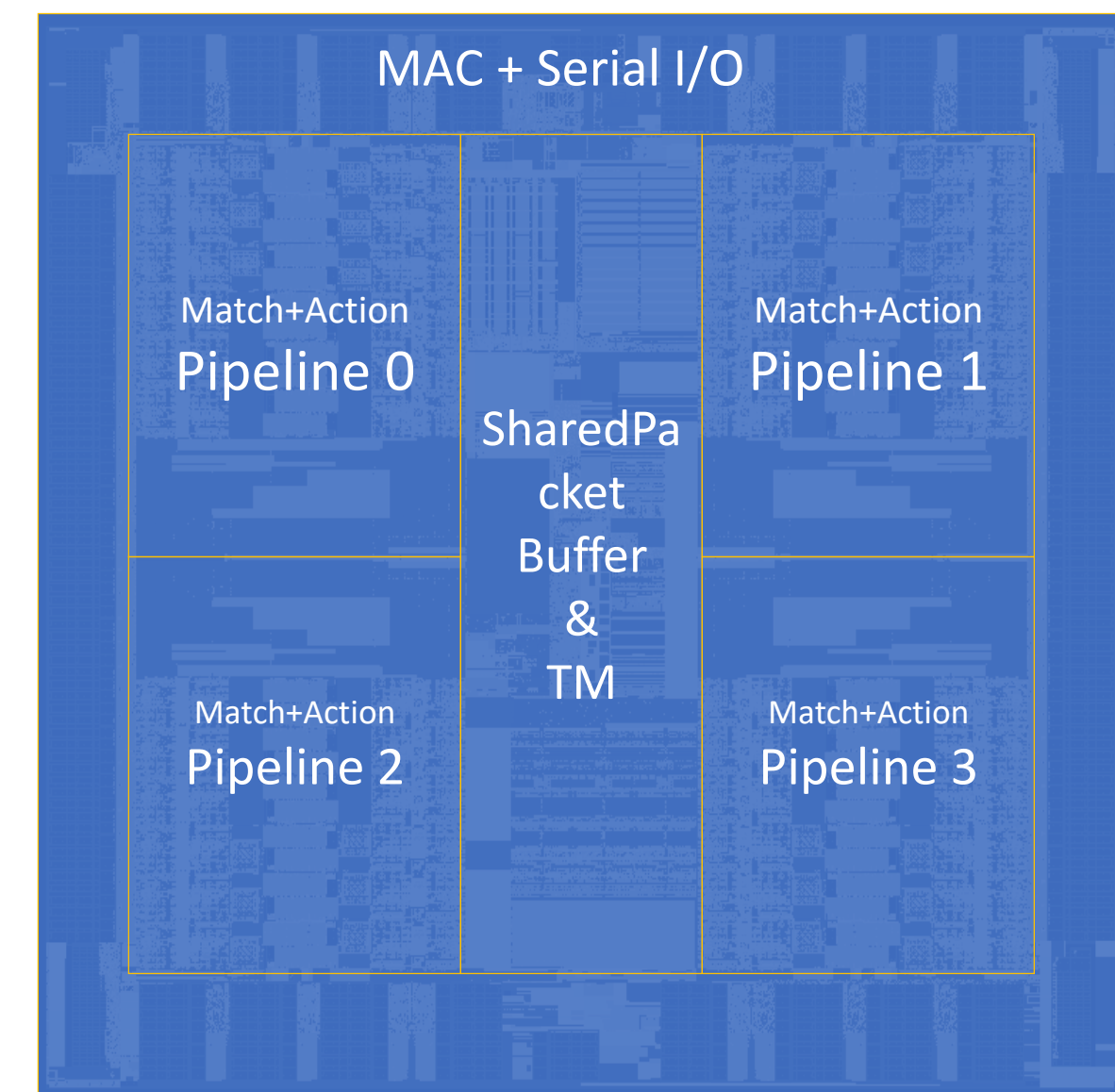
OPEN SOFTWARE.

OPEN FUTURE.

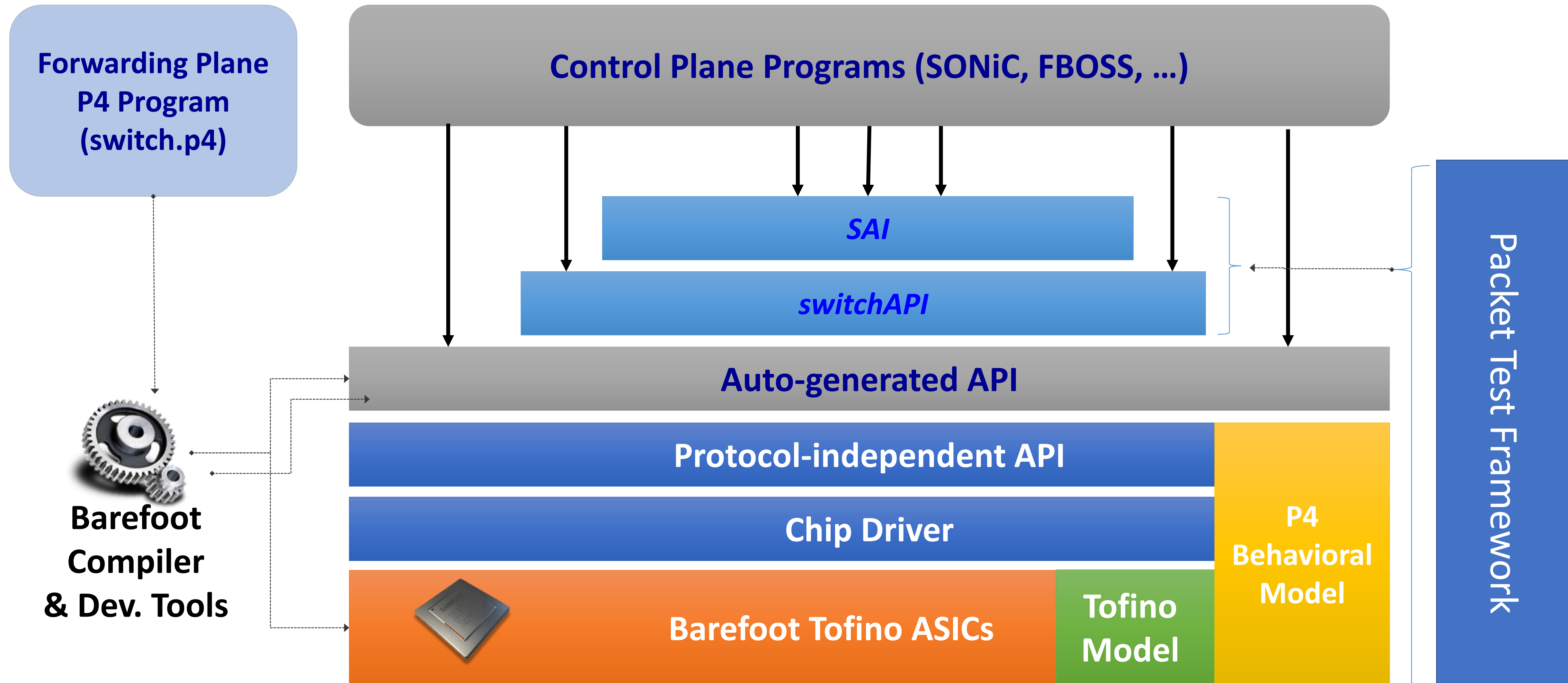


Barefoot Tofino™

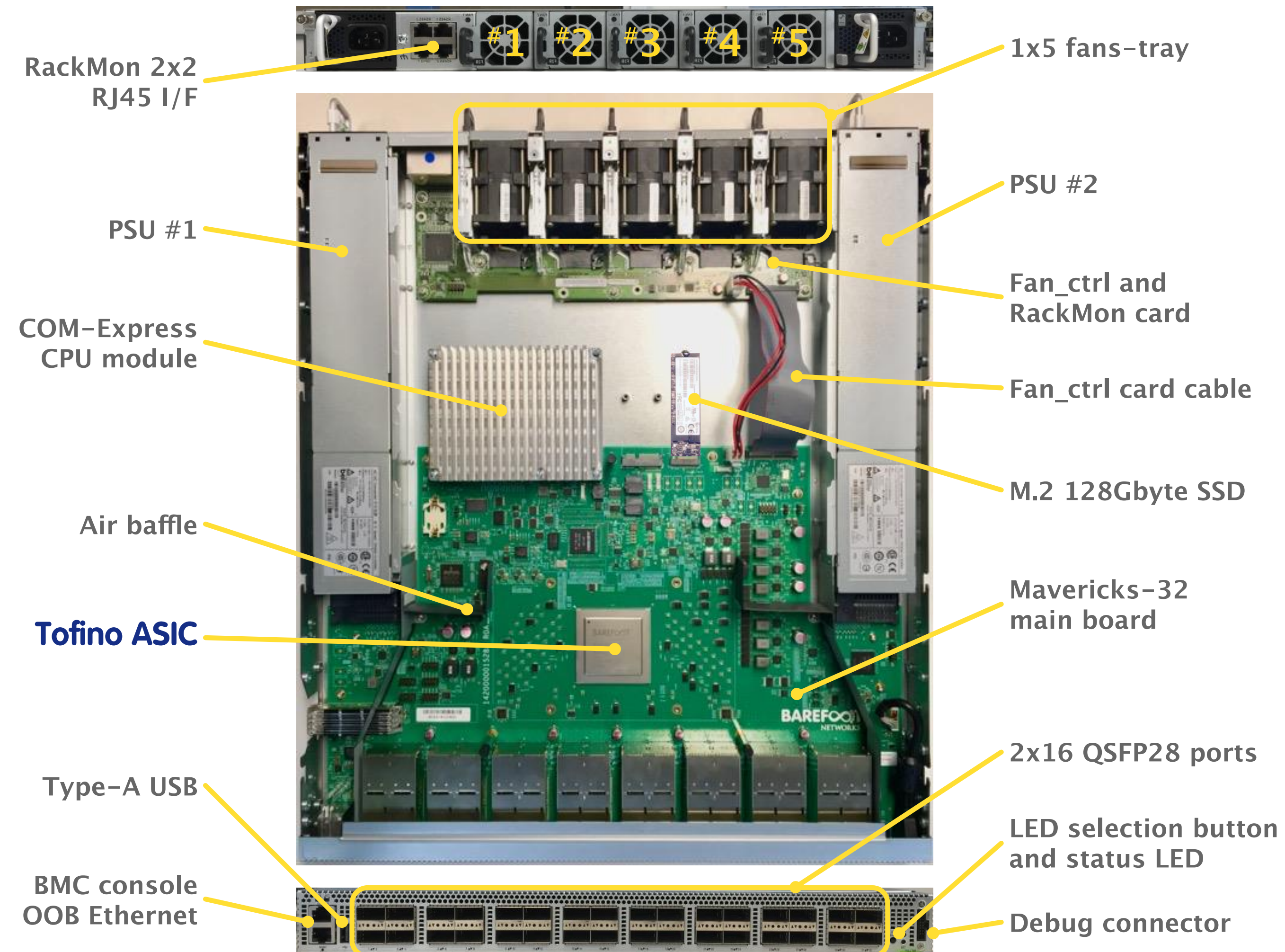
- 6.5 Tb/s P4-programmable Ethernet Switch
- Port Configurations
 - 65x100GE/40GE
 - 130x50GE
 - 260x10/25GE
- Programmable Packet Processing Pipeline
 - switch.p4 forwarding plane program
- CPU Interface
 - PCIe Gen3 x4



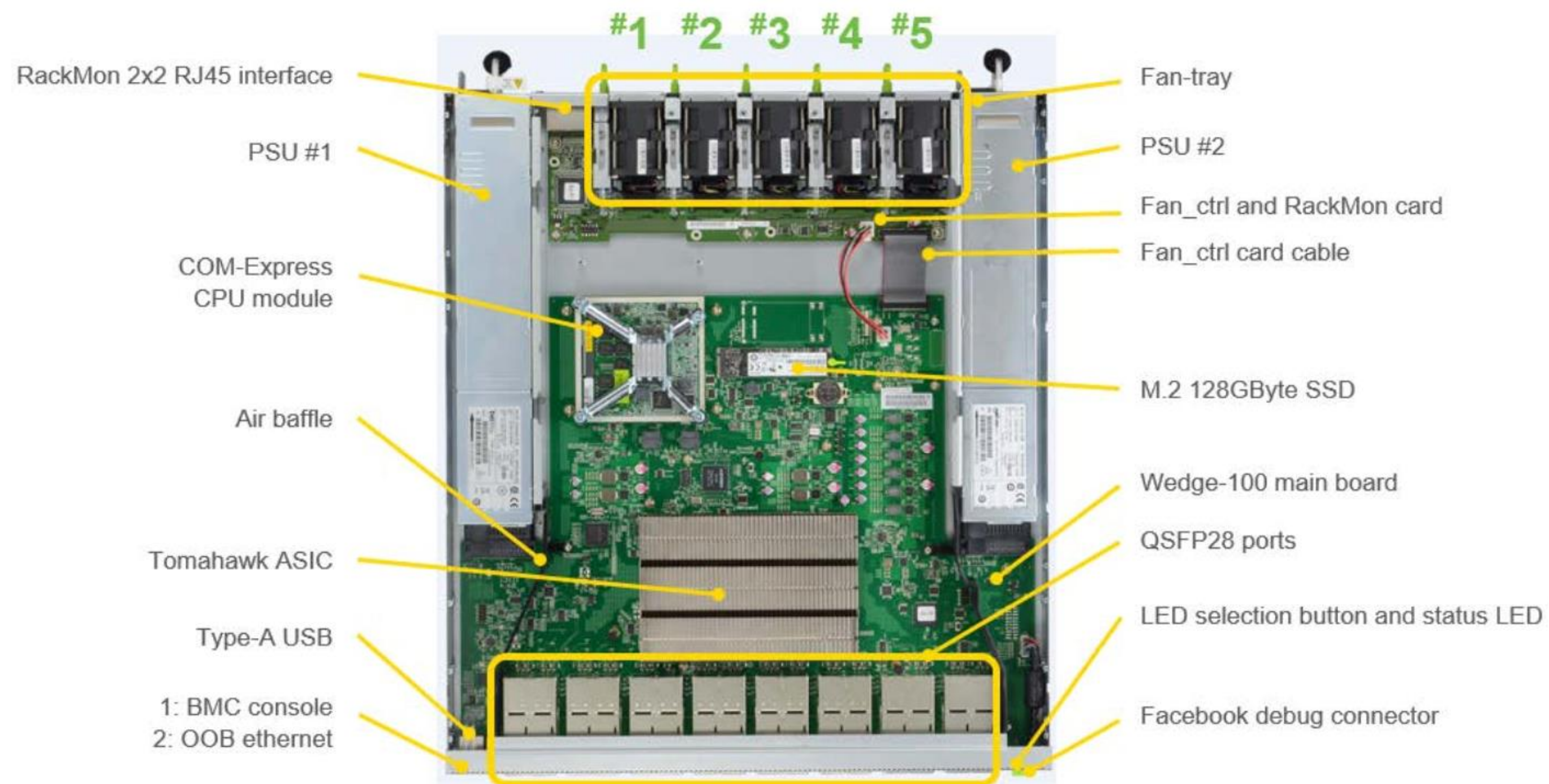
Barefoot Capilano™ SDE



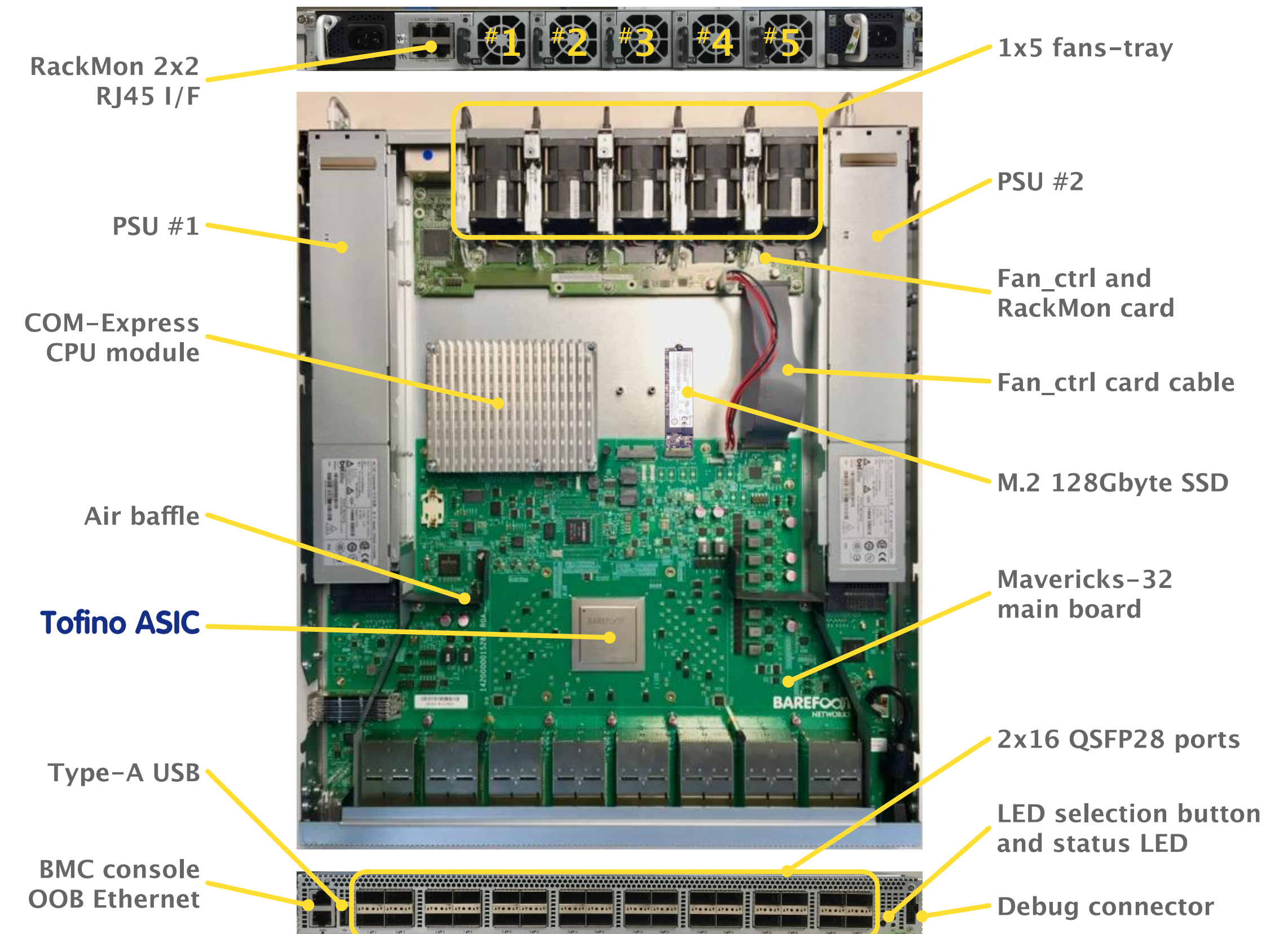
Wedge 100BF-32X



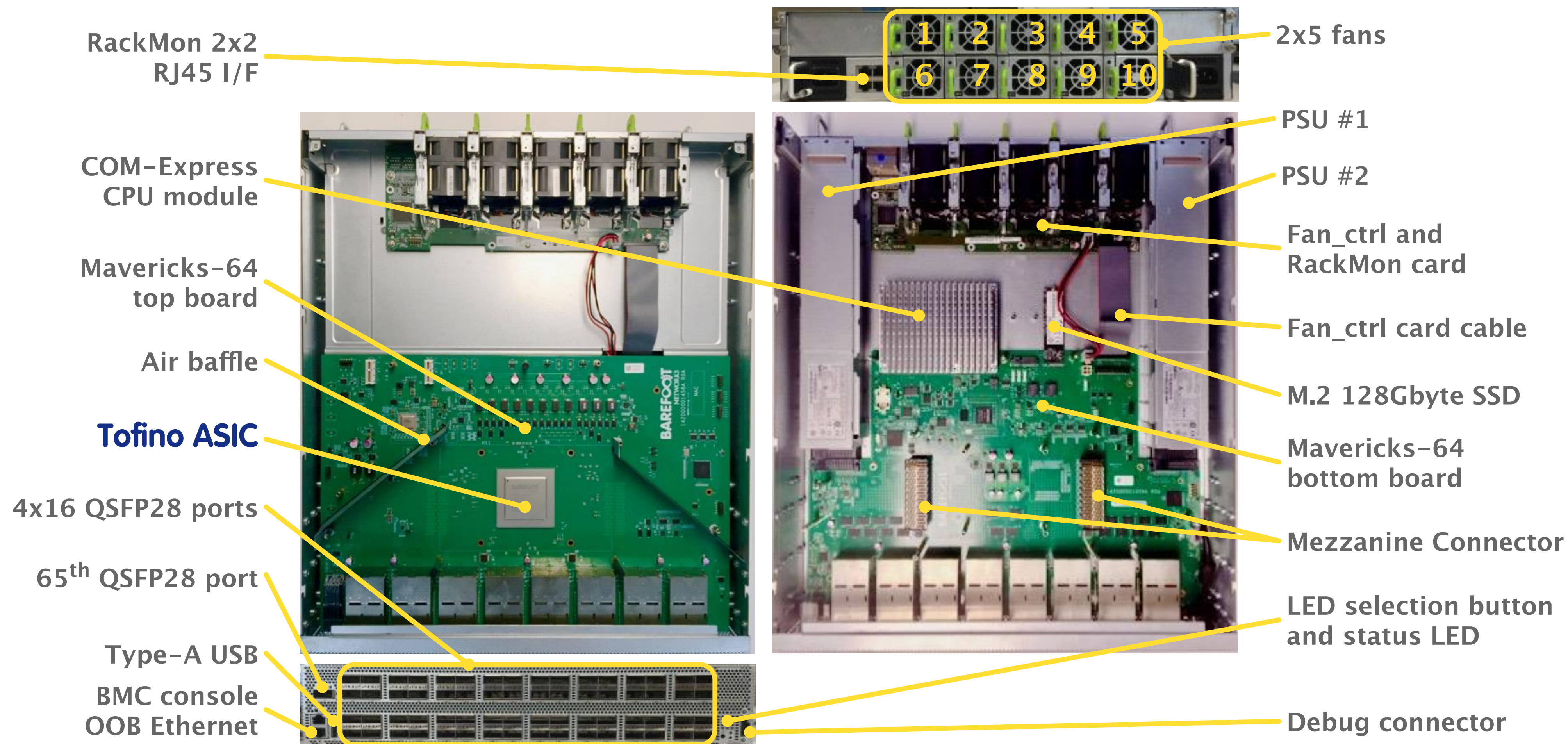
Wedge 100



Wedge 100BF-32X



Wedge 100BF-65X



CPU, BMC, Platform Software

- CPU
 - Modular COM Express Type 6
 - PCIe Gen2/3 x4
 - Intel Xeon D-1517
 - 4 Core@2.2GHz
 - 8GB DDR4 SDRAM
 - 128GB M.2 SSD
- BMC
 - ASPEED AST1250A1-GP
 - 400MHz
- Platform Software
 - OpenBMC
 - ONIE
 - ONL
 - Debian



Wedge 100B Series – A Summary

- Two platforms
 - **Wedge 100BF-32X** - 1RU 32x100GE
 - **Wedge 100BF-65X** - 2RU 65x100GE
- Derived from Wedge 100
- Includes several cost and quality enhancements
- User-programmable forwarding plane using P4, Tofino & Capilano
- Includes the World's Fastest and Fully Programmable Ethernet Switch ASIC.



Wedge 100C



Wedge 100C

Alik Fishman, Cavium

OPEN HARDWARE.

OPEN SOFTWARE.

OPEN FUTURE.



Wedge 100C

Highlights

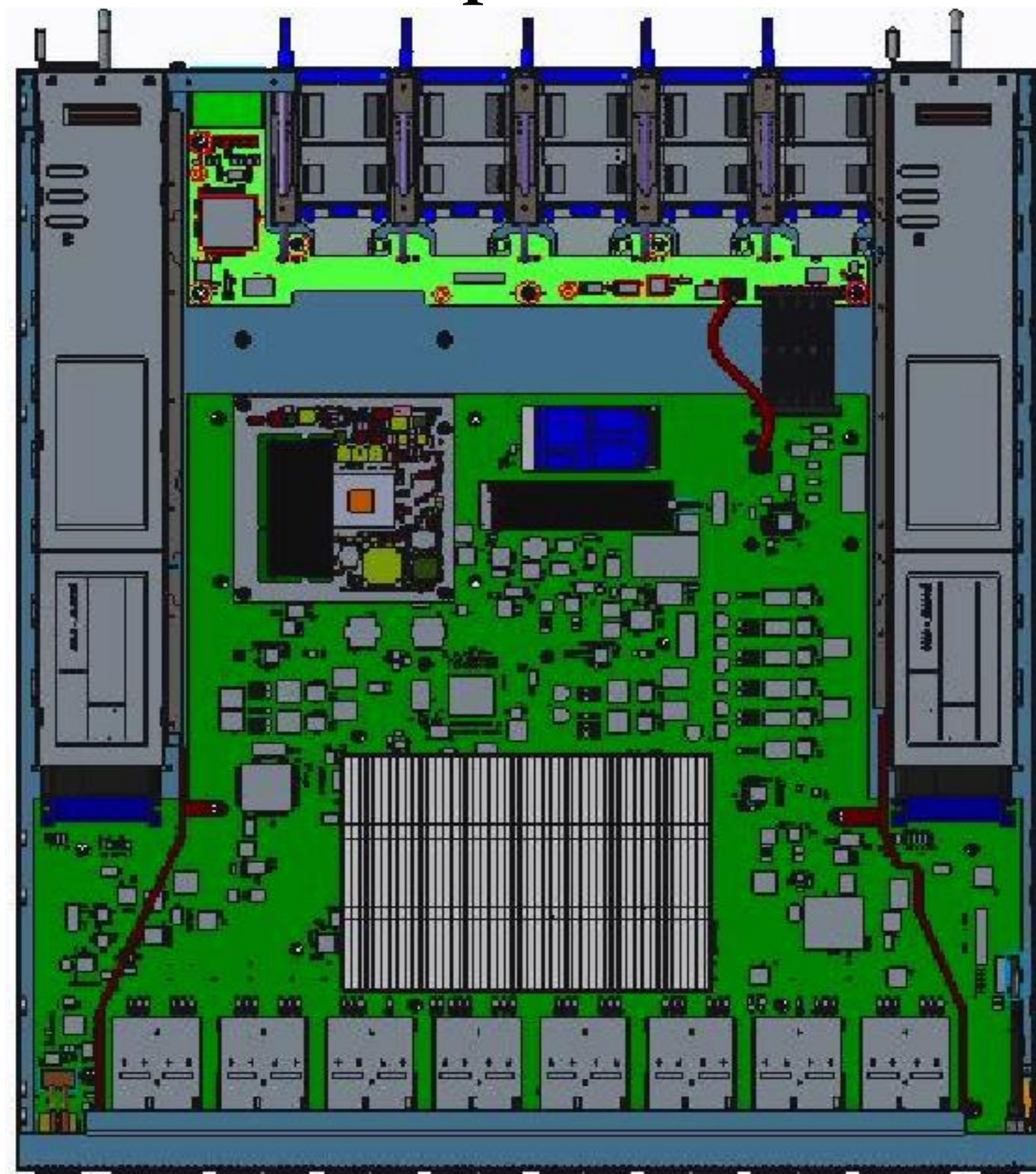
- 1RU 3.2Tb Switching platform with 32 x QSFP28 ports
 - Supports any combination of 10/25/40/50/100GbE interfaces
- Based on the OCP-ACCEPTED™ Wedge 100 contributed by Facebook
 - Wedge 100 platform hardware characteristics preserved, including OpenBMC
- Uses Cavium-XPliant CNX88091 programmable field-proven ASIC
 - The first and the only programmable ASIC available in production today



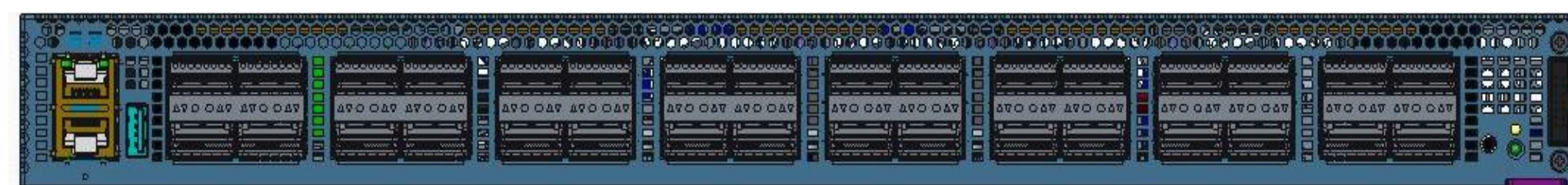
Wedge 100C

Placement & Layout

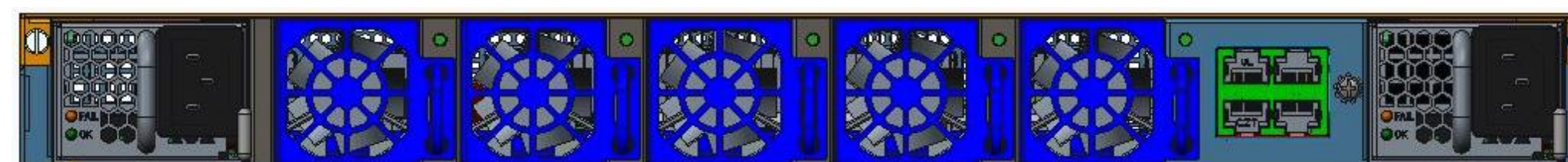
Top View



Front Panel View

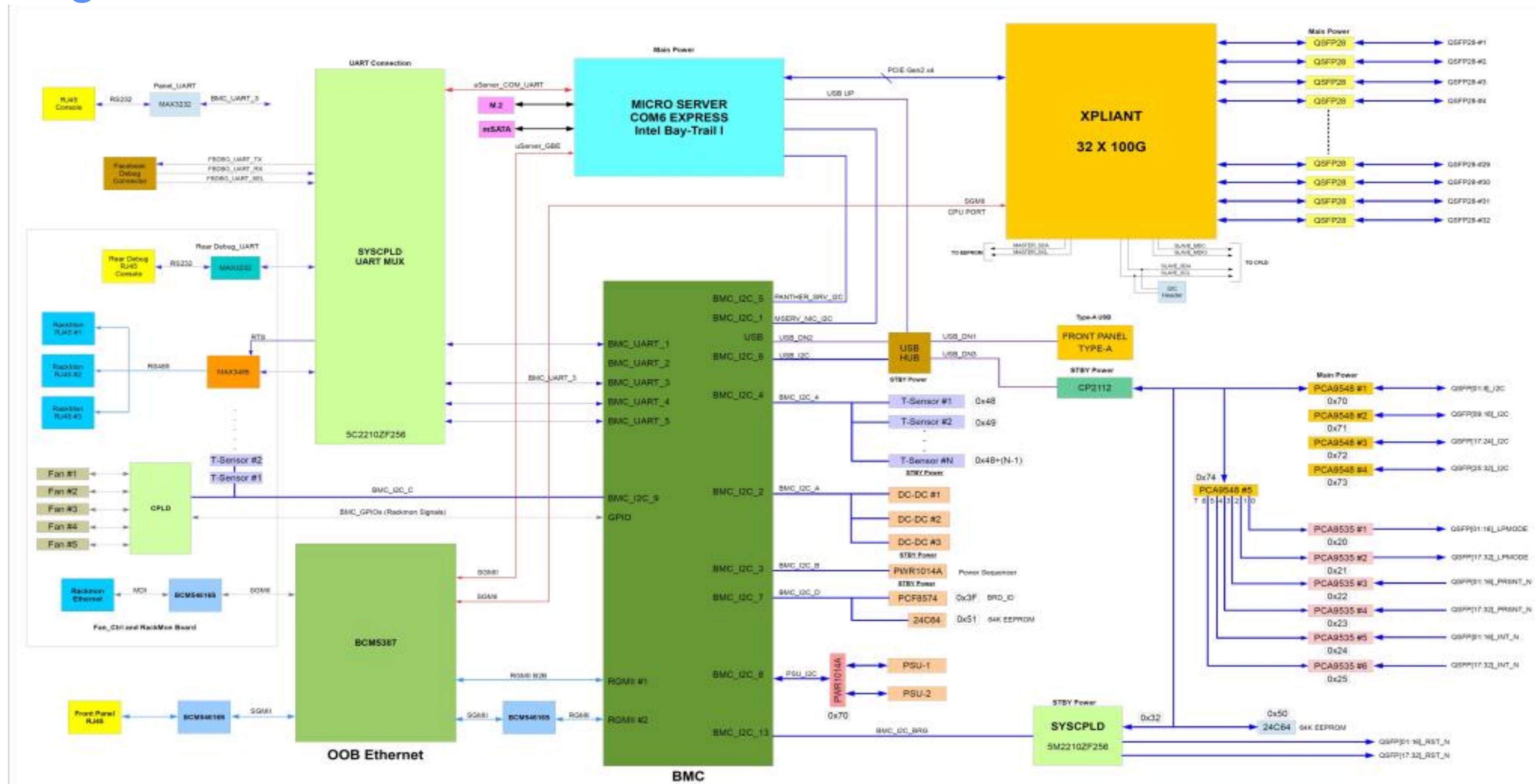


Rear Panel View



Wedge 100C

Block Diagram



19-in Standard & 21-in OpenRack SKUs are available: share the same main board PCB

Wedge 100C

Complete HW Design Contribution

- Available at:
 - <http://www.opencompute.org/wiki/Networking/SpecsAndDesigns>
- The contribution includes all the files required for production:
 - README.txt, submission presentation slides
 - /Electrical/BOM
 - /Electrical/Gerber
 - /Electrical/PCB Layout
 - /Electrical/Schematic
 - /Mechanical/
 - /Programming_Files
 - /Specification

Wedge 100C

Value Proposition

- Wedge 100 hardware platform benefits:
 - Field-proven cost optimized design
 - OpenBMC baseboard management & COM-E for control plane CPU
 - 19-in Standard & 21-in OpenRack SKUs
- Cavium-XPliant CNX88091 programmable field-proven ASIC benefits:
 - Programmable Packet Pipeline – Flexible Networking Profiles
 - Embedded Resources Elasticity – Scale and Efficiency
 - Dynamic Buffering – 24MB Fully Shared
 - The 1st and the only Programmable ASIC Shipping & Field Deployed Today
- Variety of Disaggregated Software Choice:
 - FBOSS
 - OpenSwitch, Pica8, SONIC
 - SAI, OpenFlow





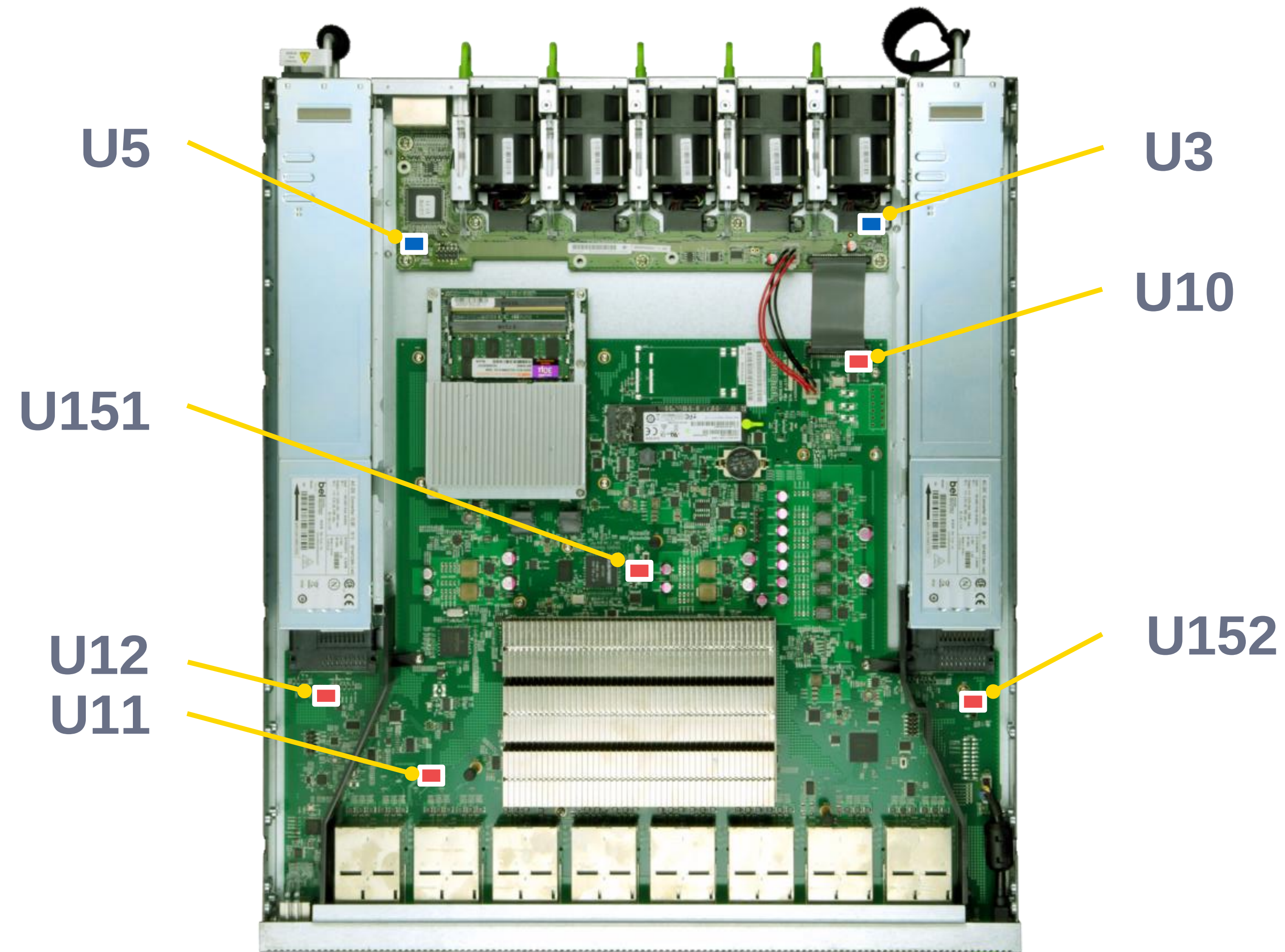
OPEN

Compute Project



Backup

Placement of Temperature Sensors in Wedge 100S



QSFP28 100G to QSFP28 100G Cable

Four lanes, QSFP28 at both ends
25Gbps per lane

QSFP28 100G
All 4 lanes used

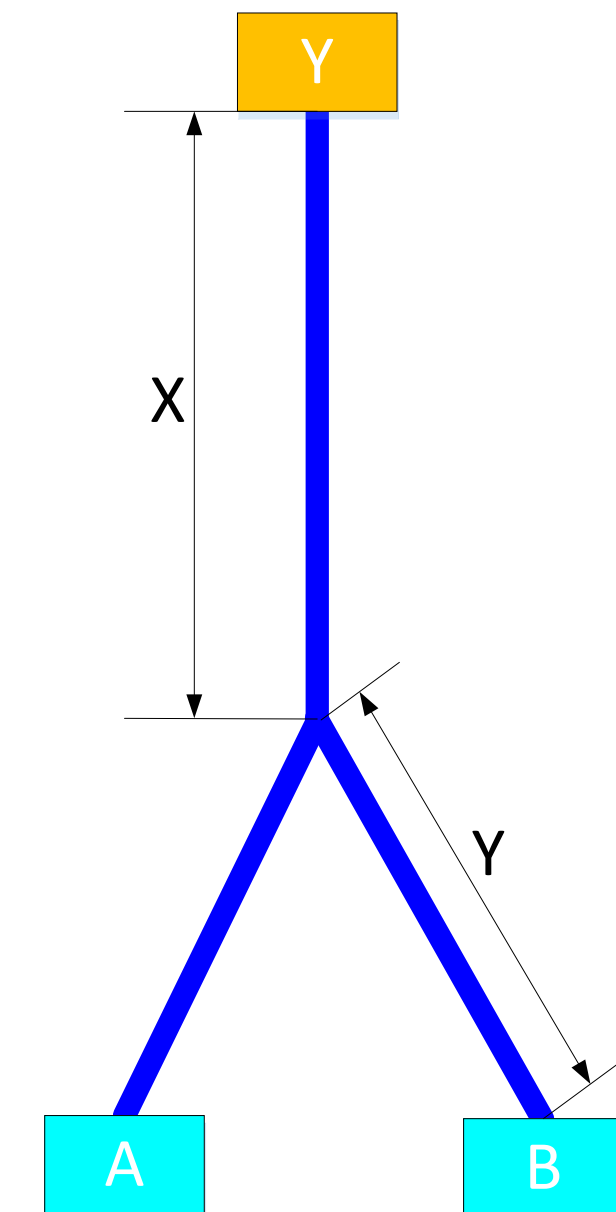
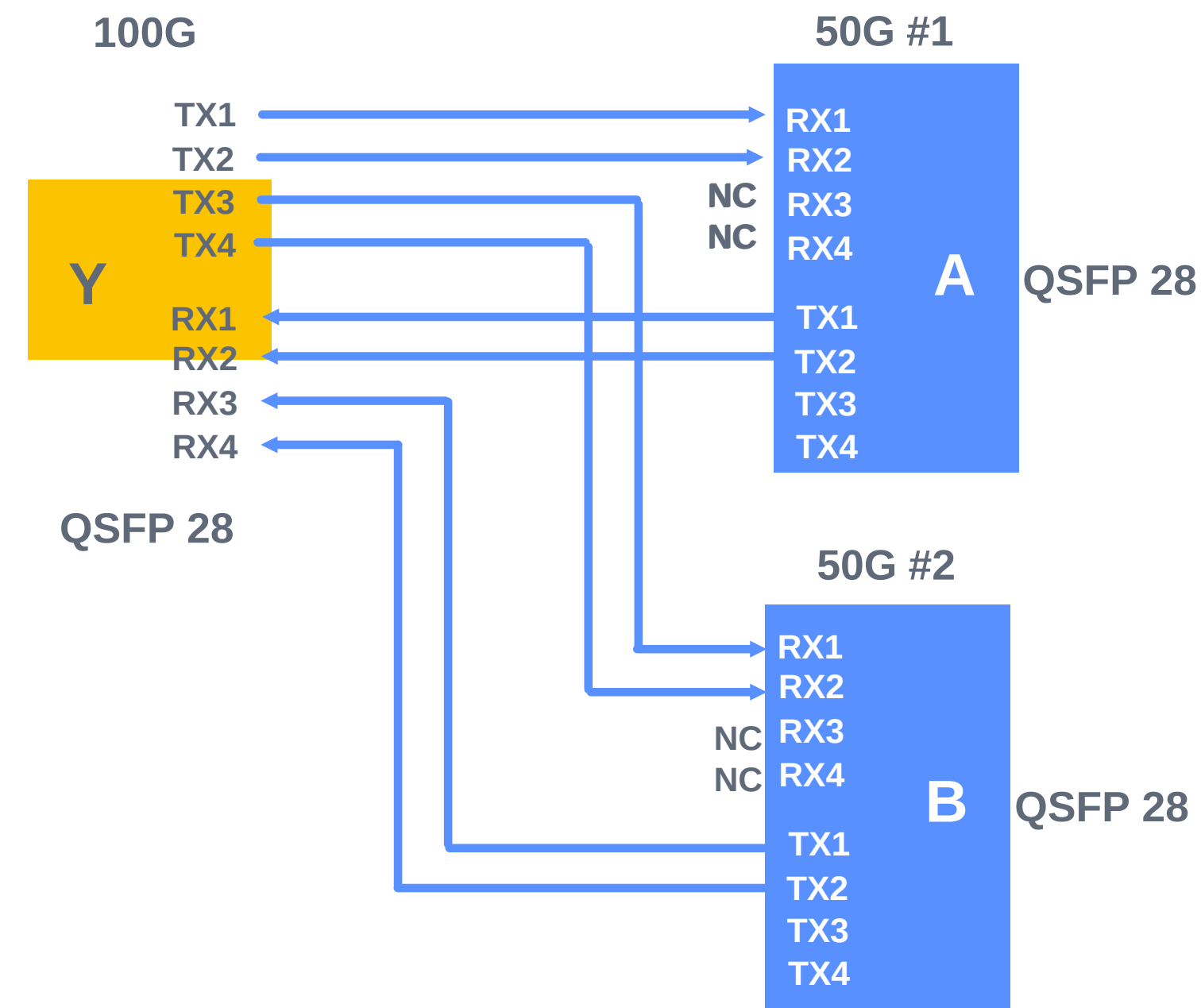
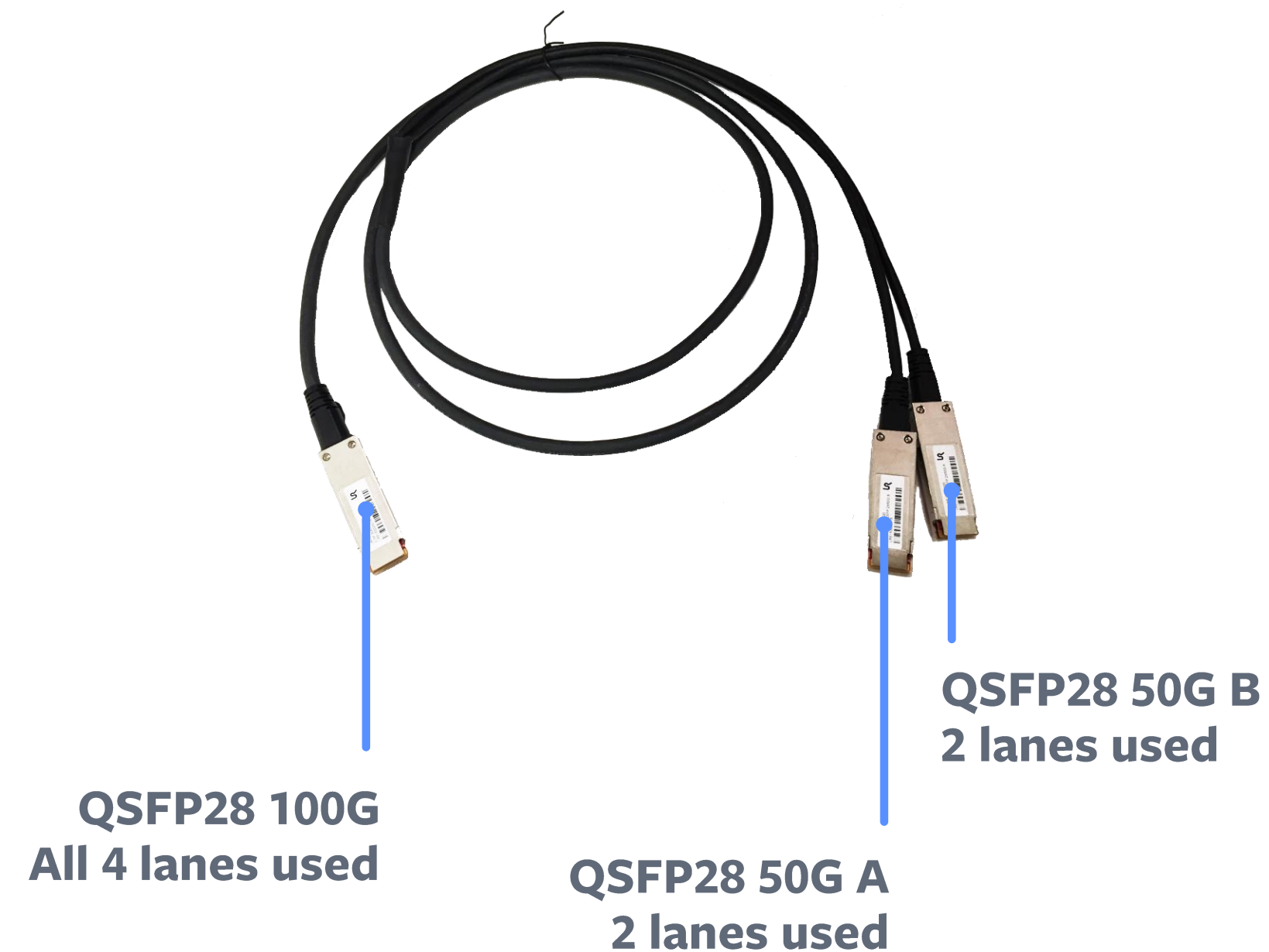


QSFP28 100G
All 4 lanes used

QSFP28 100G to 2x QSFP28 50G Cable

- a.k.a Y-Cable, this was for Wedge-100 to inter-operate with Yosemite 50G NIC.
- Yosemite 50G CX4 NIC uses QSFP28 form factor, but only 2 lanes used.
- Header Y use all 4 lanes, header A and B use 2 lanes.
- All lanes are 25Gbps.

Facebook 100G to 2x50G Y Cable Pinout Diagram



QSFP28 100G to 4x SFP28 25G Cable

- Four lanes in QSFP28 side, 25Gbps per lane
- One lane at SFP28 side, 25Gbps
- 1M, 2M, 3M

**Four SFP28 25G
header, each SFP28
supports
25Gbps single lane**

**QSFP28 100G
All 4 lanes used**

