Meta’s Catalina Platform (NVL72)

Hotchips - August 24th, 2025

Datacenter Racks Tutorial

Meta

William Arnold

Technical Program Manager,
Infrastructure

Matt Bowman

Hardware Engineer,
Infrastructure

Agenda

AI at Meta

Meta’s Catalina Product

What motivated us to modify the standard NVL72 design

Resulting product

Supporting technologies

Hardware Details

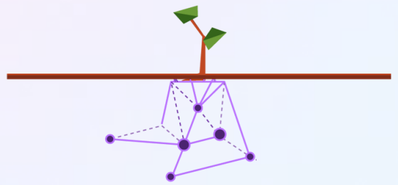
Server Architecture

Custom Boards

Supporting Technologies Detail

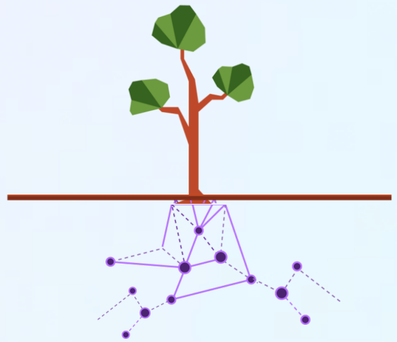
Since 2022, we've seen a major AI infrastructure change

2022

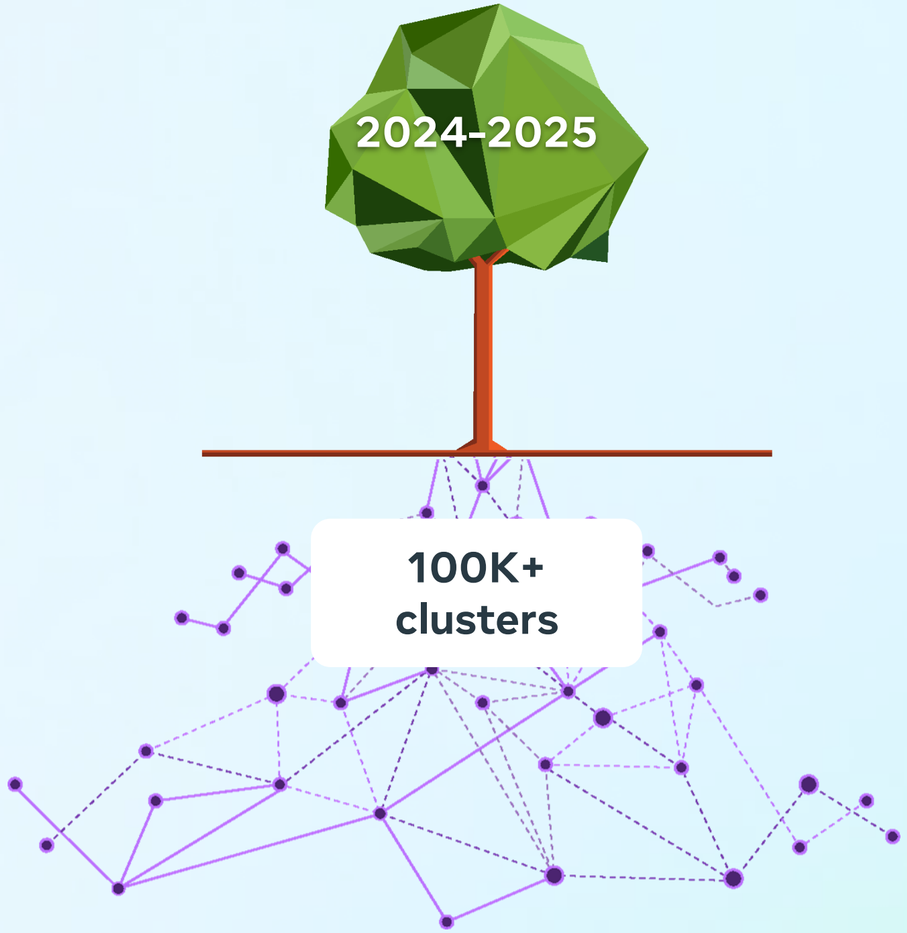


6K clusters
Job size: 128-512 GPUs

2023



16-24K clusters
Job size: 16K GPUs

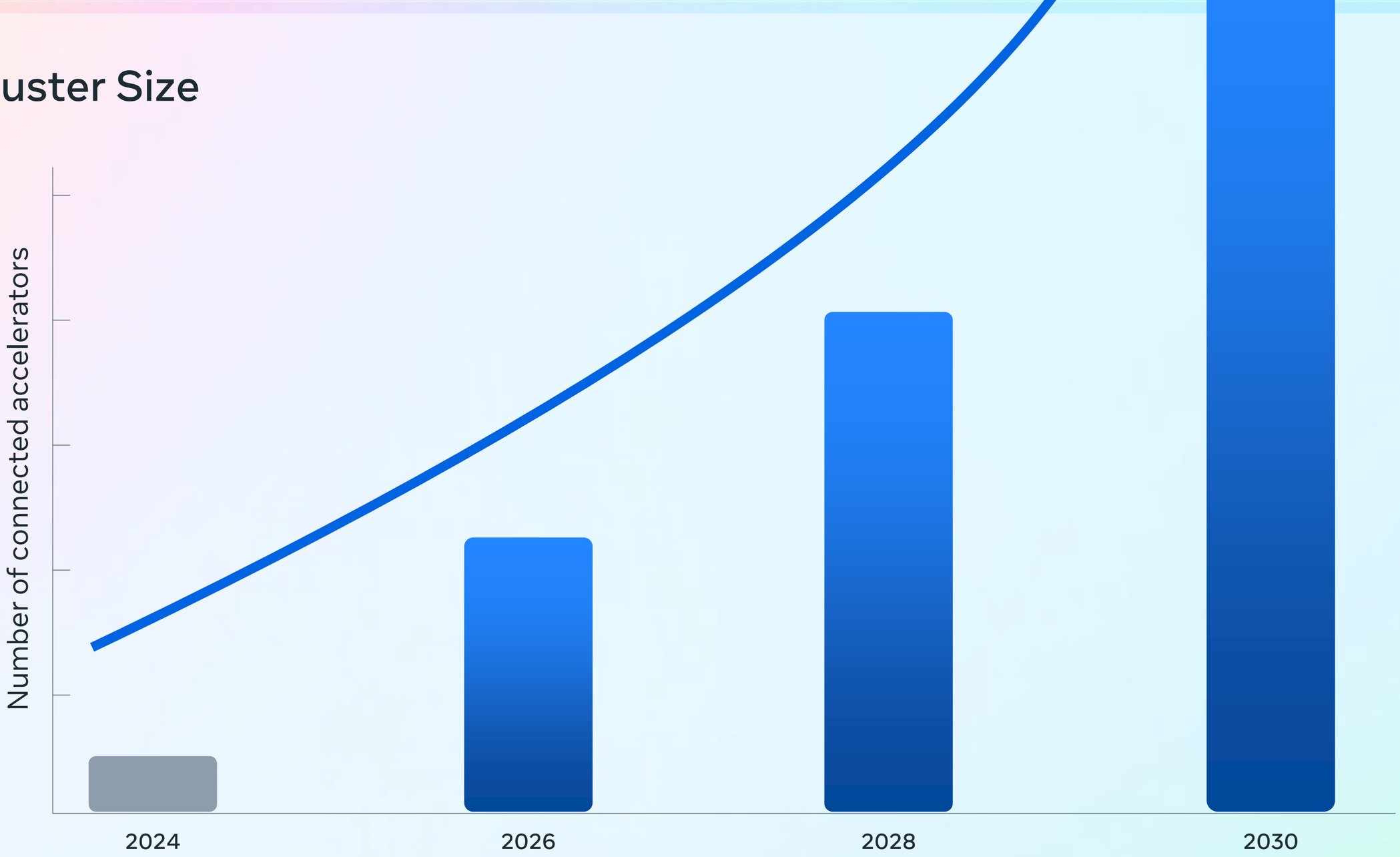


Models

Software Infra

Physical Infrastructure

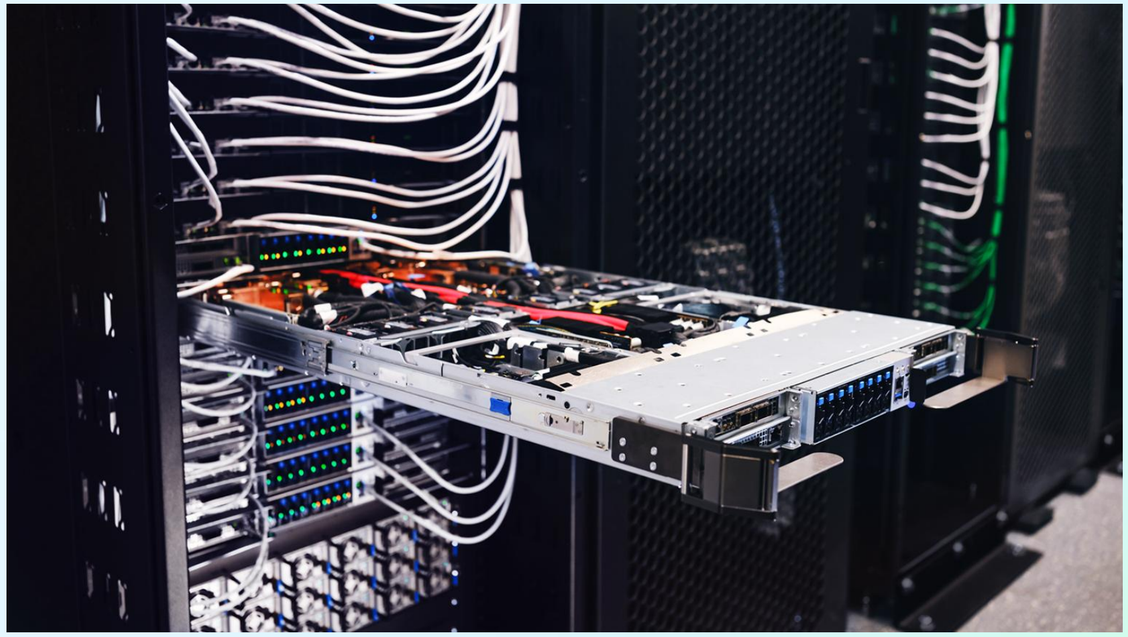
AI Cluster Size



Background

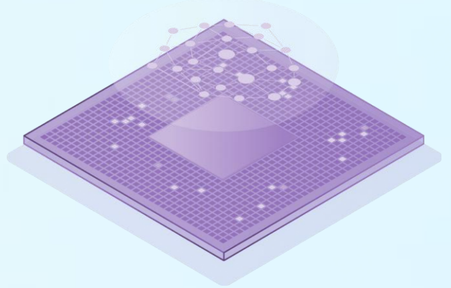
“NVIDIA has contributed MGX based GB200-NVL72 Rack and Compute and Switch Tray designs [to the Open Compute Project], while Meta is introducing Catalina AI Rack architecture for AI clusters.”

- 2024 Global OCP Summit

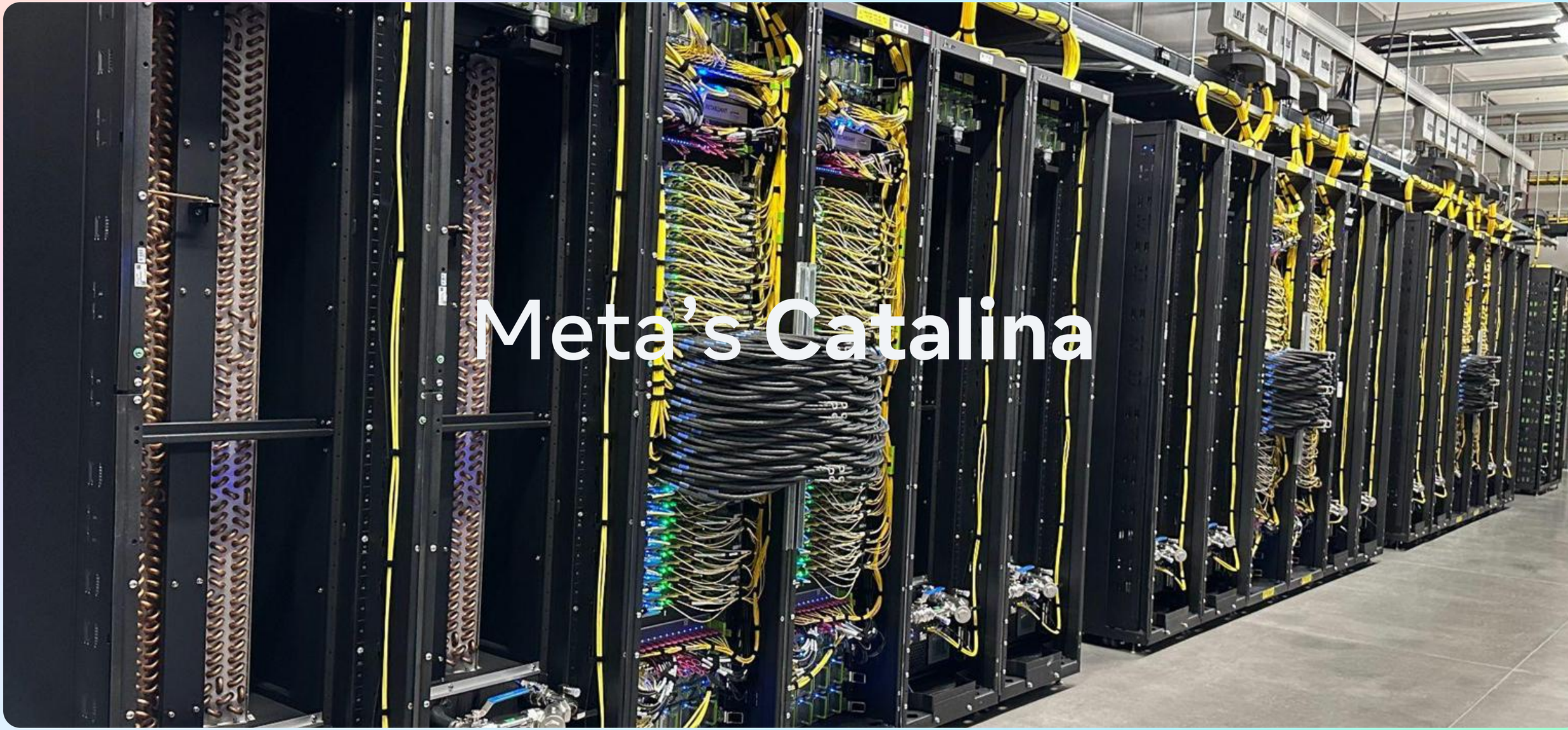


https://developer.nvidia.com/blog/nvidia-contributes-nvidia-gb200-nvl72-designs-to-open-compute-project/

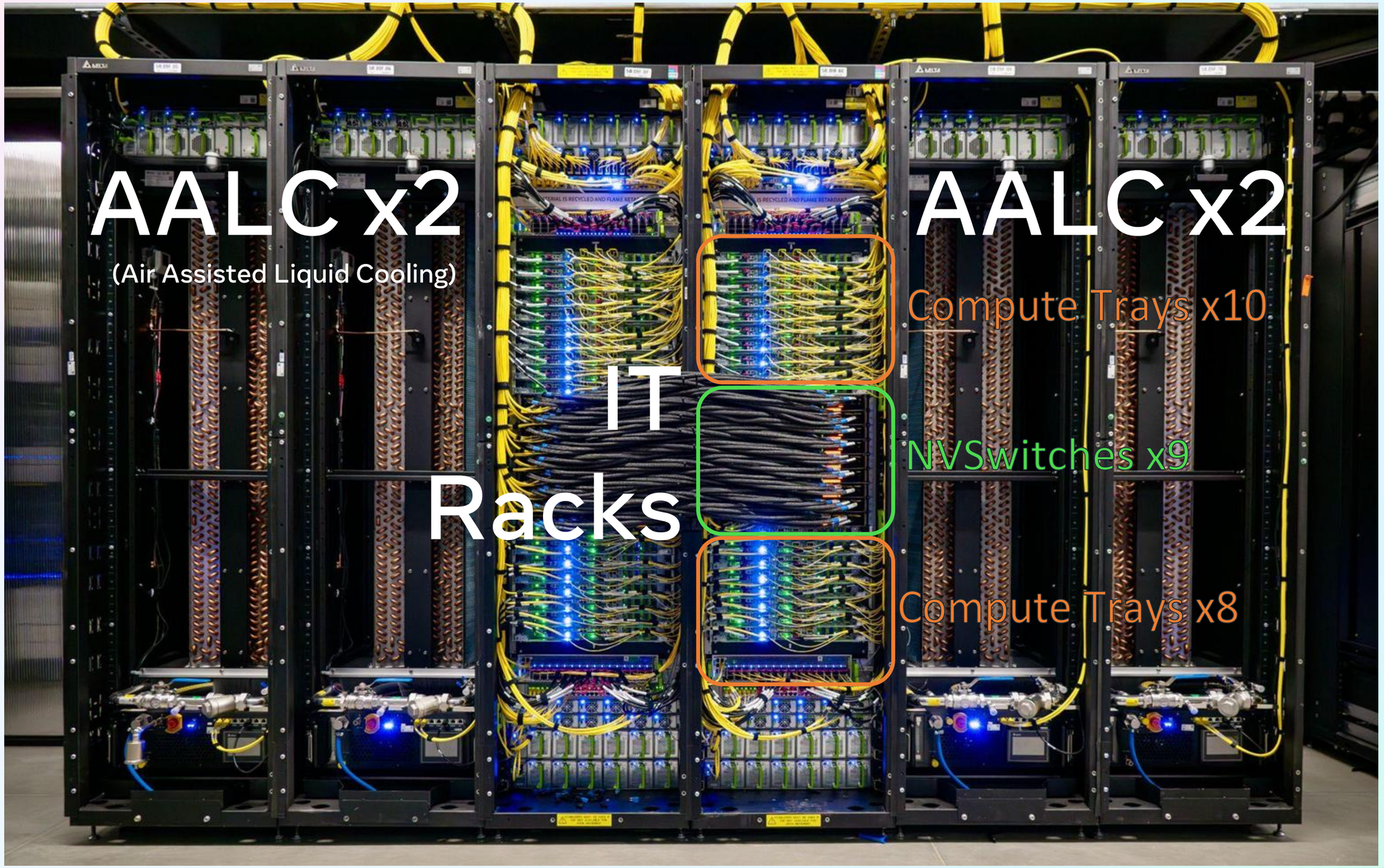
Models



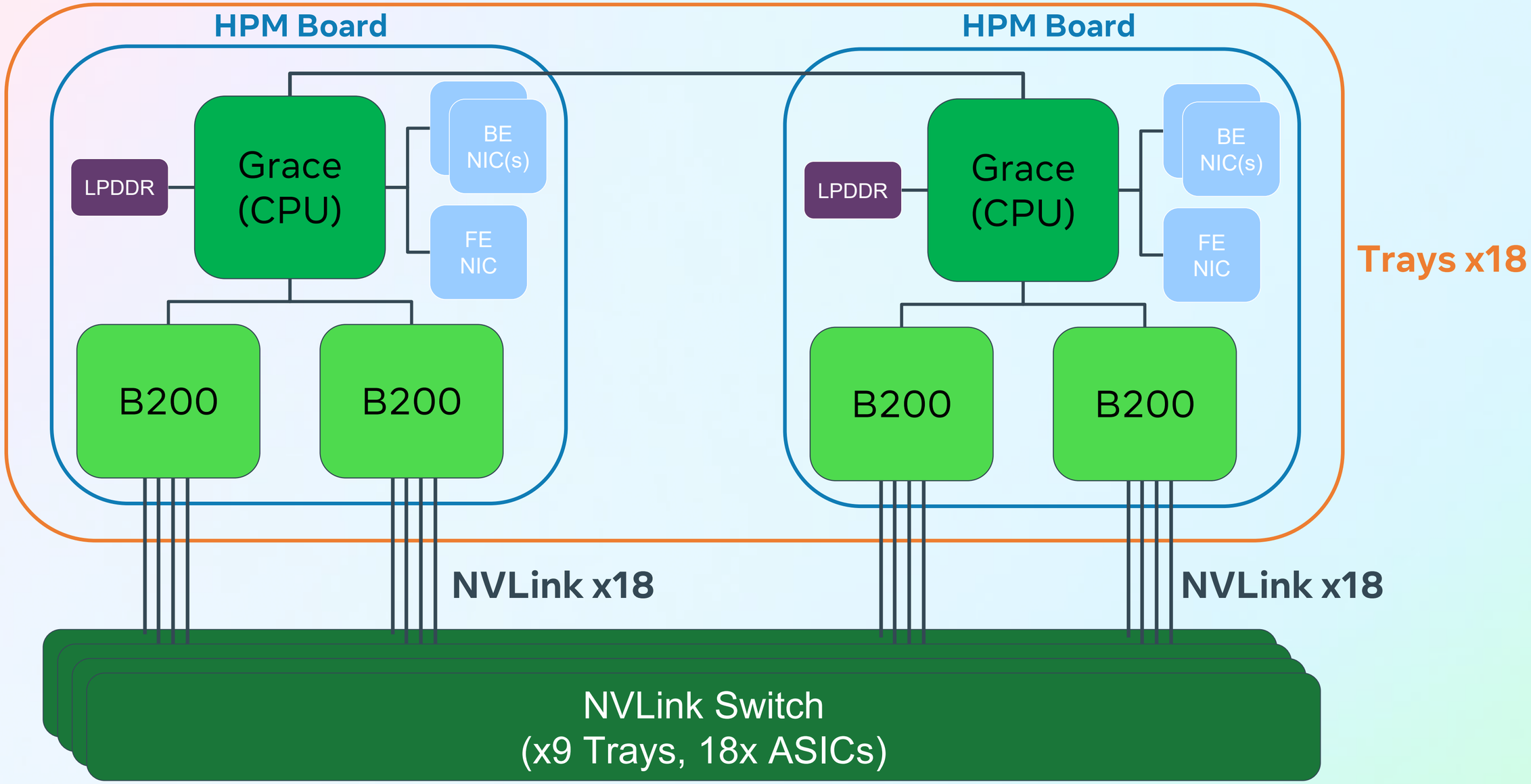
Physical Infrastructure


Meta’s Catalina

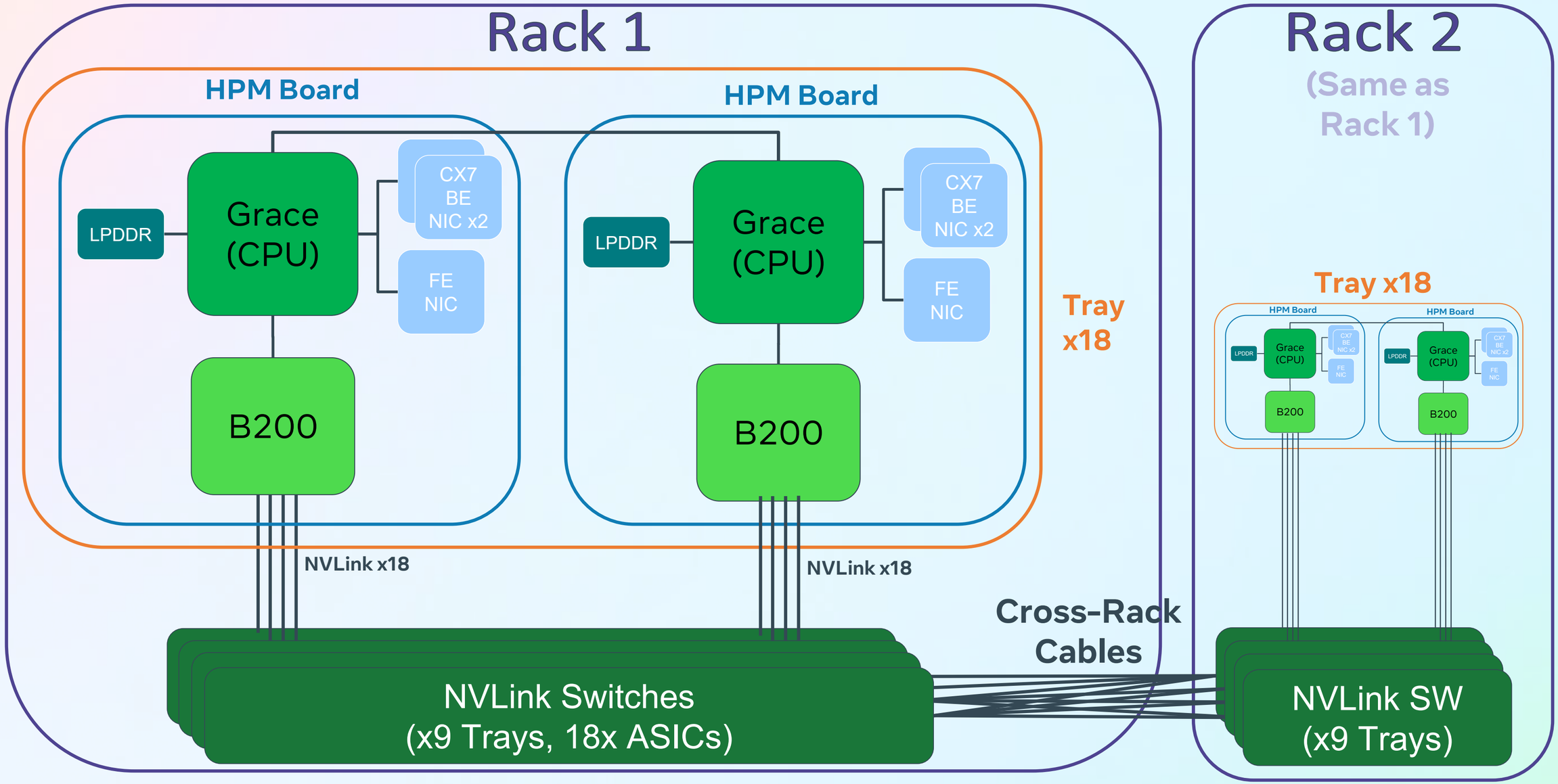
Catalina Pod



NVIDIA MGX GB200 Configuration



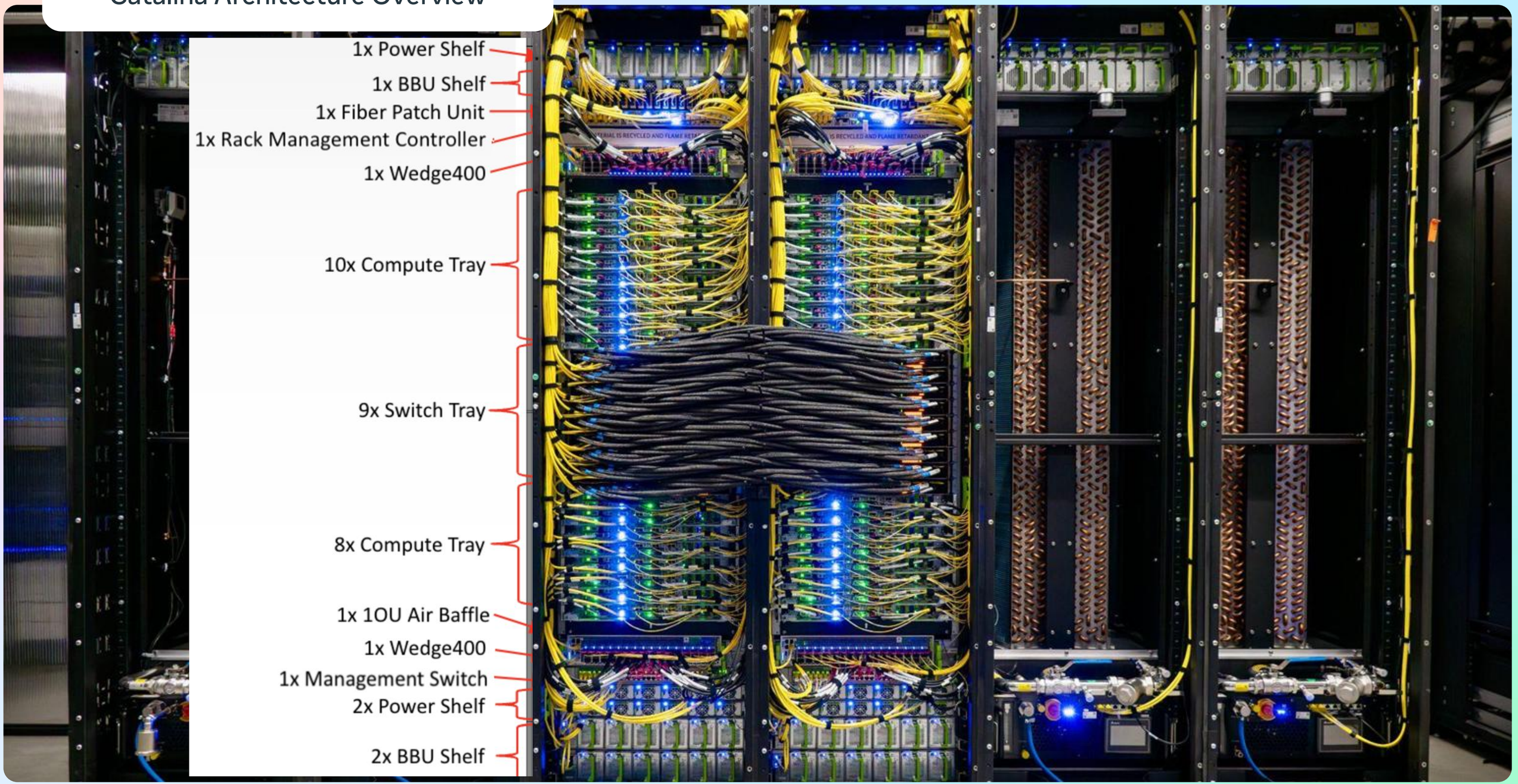
Meta Catalina GB200 Configuration



Scale-Up Domain - Key Resources

Per Pod (up to)	NVIDIA GB200 NVL72 Baseline	Meta GB200 NVL72
Blackwell GPUs	72	72
Aggregate GPU Memory	13.4 TB	13.4 TB
NVIDIA Grace CPUs	36	72
Aggregate LPDDR	17.3 TB	34.6 TB
Total cache-coherent memory	30 TB	48 TB
Physical IT Racks	1	2

Catalina Architecture Overview



There are a significant number of new technologies needed to support the NVL72 rack given the step function change in rack power density from previous generations. We used a blend of our own programs and 3rd party solutions.

Common to all NVL72

- NVSO72
  - NVLink Switches
  - Cable Cartridge
- Grace/ARM
- Universal Quick Disconnect (Valve for liquid cooling)

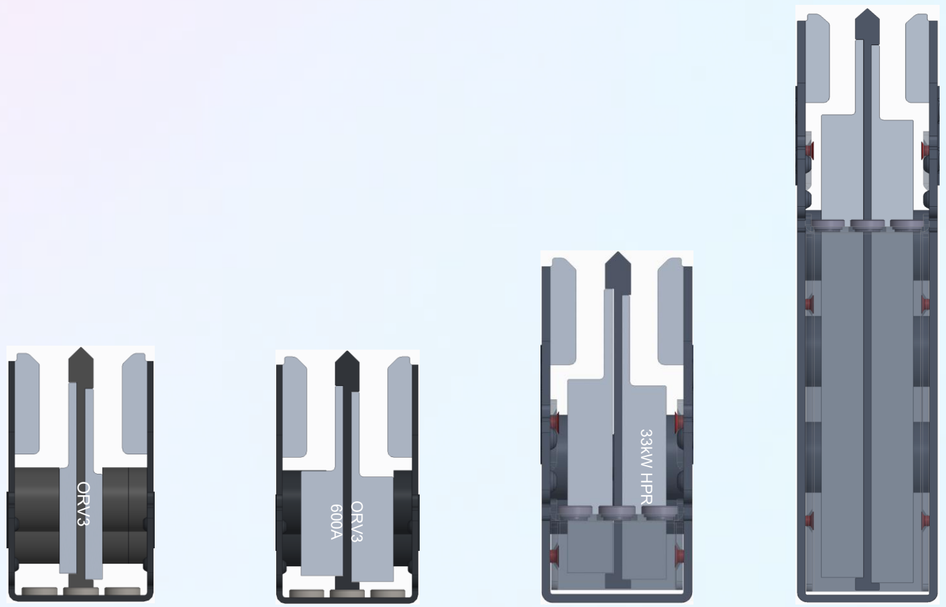
Unique to Meta's Version

- Open Rack v3 - High Power Rack
- Liquid Cooling
  - AALC
  - Rack Management Controller (RMC) - Liquid Cooling safety & orchestration device
- Disaggregated Scheduled Fabric (DSF) Network (Our program)
- Data Center Secure Control Module (DC-SCM)

High Power Rack

First deployment of HPR (High Power Rack) version of Open Rack v3

- Increased per-rack power to 93.5kW
- Enabled 5.5kW per PSU/BBU module; 33kW per power shelf
- Increased busbar power limit

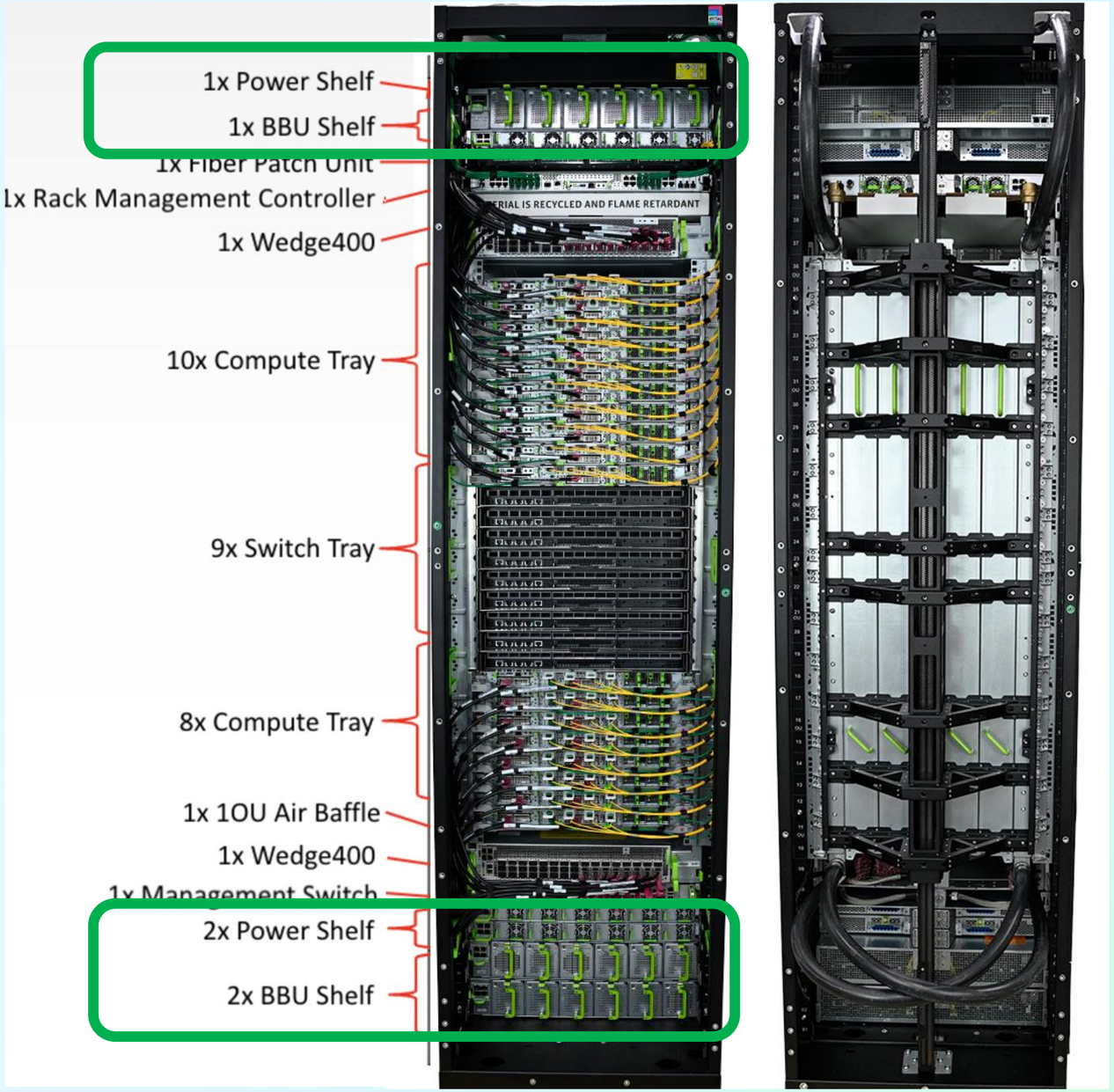


ORV3 “Vanilla”

ORV3 w/ new Busbar

ORV3 HPR 33kW

ORV3 HPR 94kW



Need to support traditional Data Center facilities built for air cooling, plus newly designed Facility Liquid Cooling (FLC) locations.

For deployments into traditional locations we utilize ‘Air Assisted Liquid Cooling’ (AALC) racks

Liquid flow is controlled by a device called the **Rack Management Controller (RMC)** that sits in each rack.

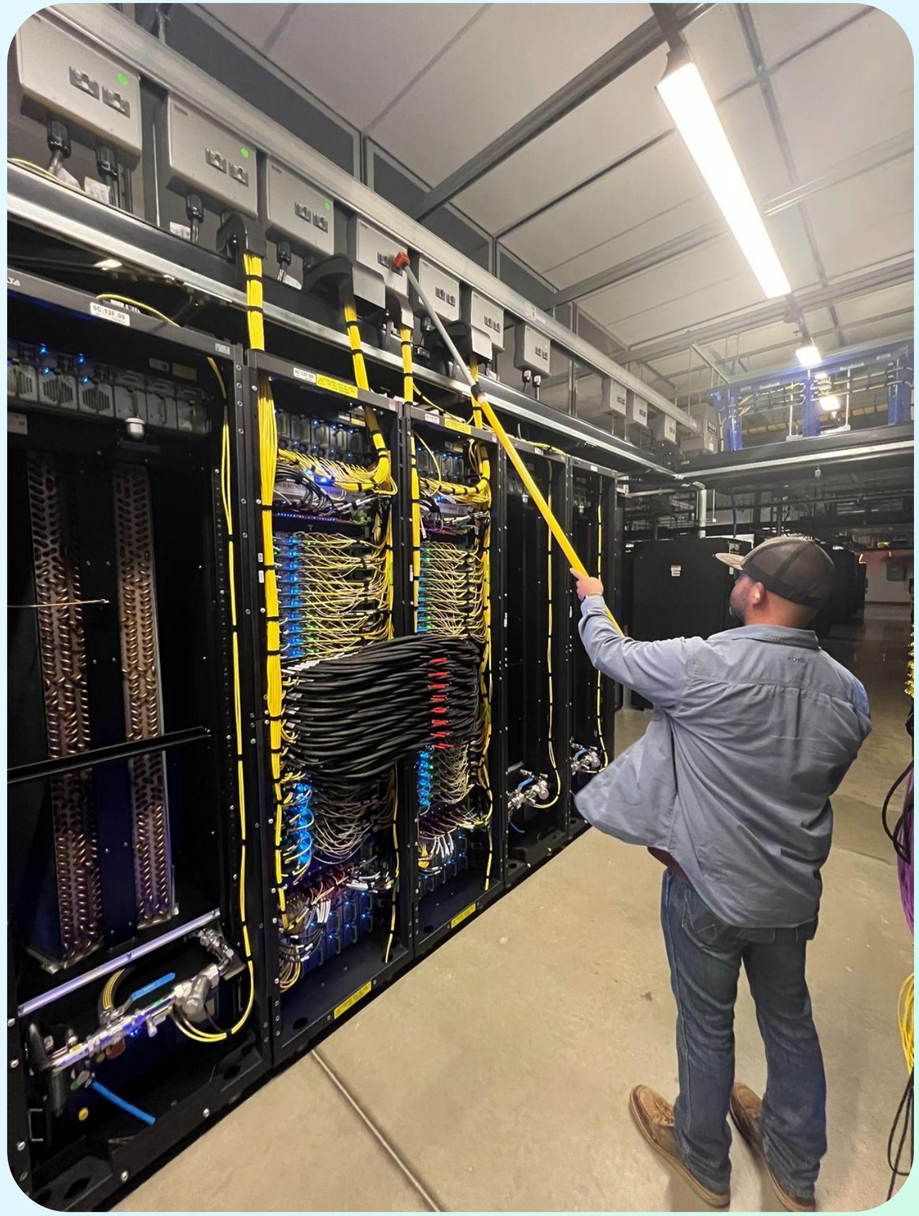
Leak Detection & Shutdown

Critical to respond to a leak anywhere in cooling path

- We have multiple sensors at key components, as well as the Tray, Rack, and Floor levels

Different flows happen depending on location and severity of the leak

The logic for leak control within the RMC



Network for Scale-out, AI Fabrics

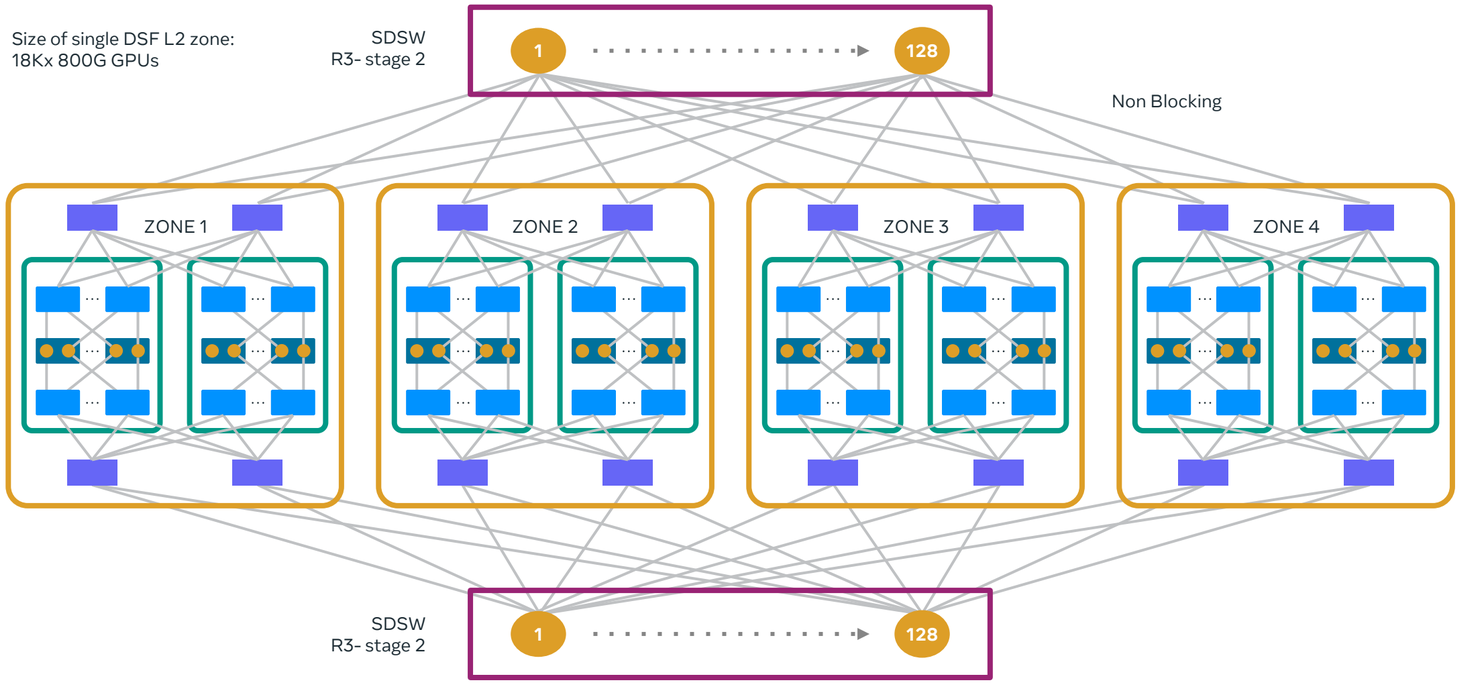
Disaggregated Scheduled Fabric (DSF): Lossless/reliable fabric of switches

Tuned for AI

Provides flexibility & speed across multiple generations and types of accelerators and NICs

Uses network racks at the end of each row to connect the Catalina racks to the Backend network

DSF Dual stage (Building, 4x AI ZONE)



Catalina Interconnects

Compute Tray

- Frontend Network [North-South]: 200G DAC to W400 per GPU
- Backend [East-West / Scale Out] Network: 2x400G per GPU via fiber

NVLink Switches

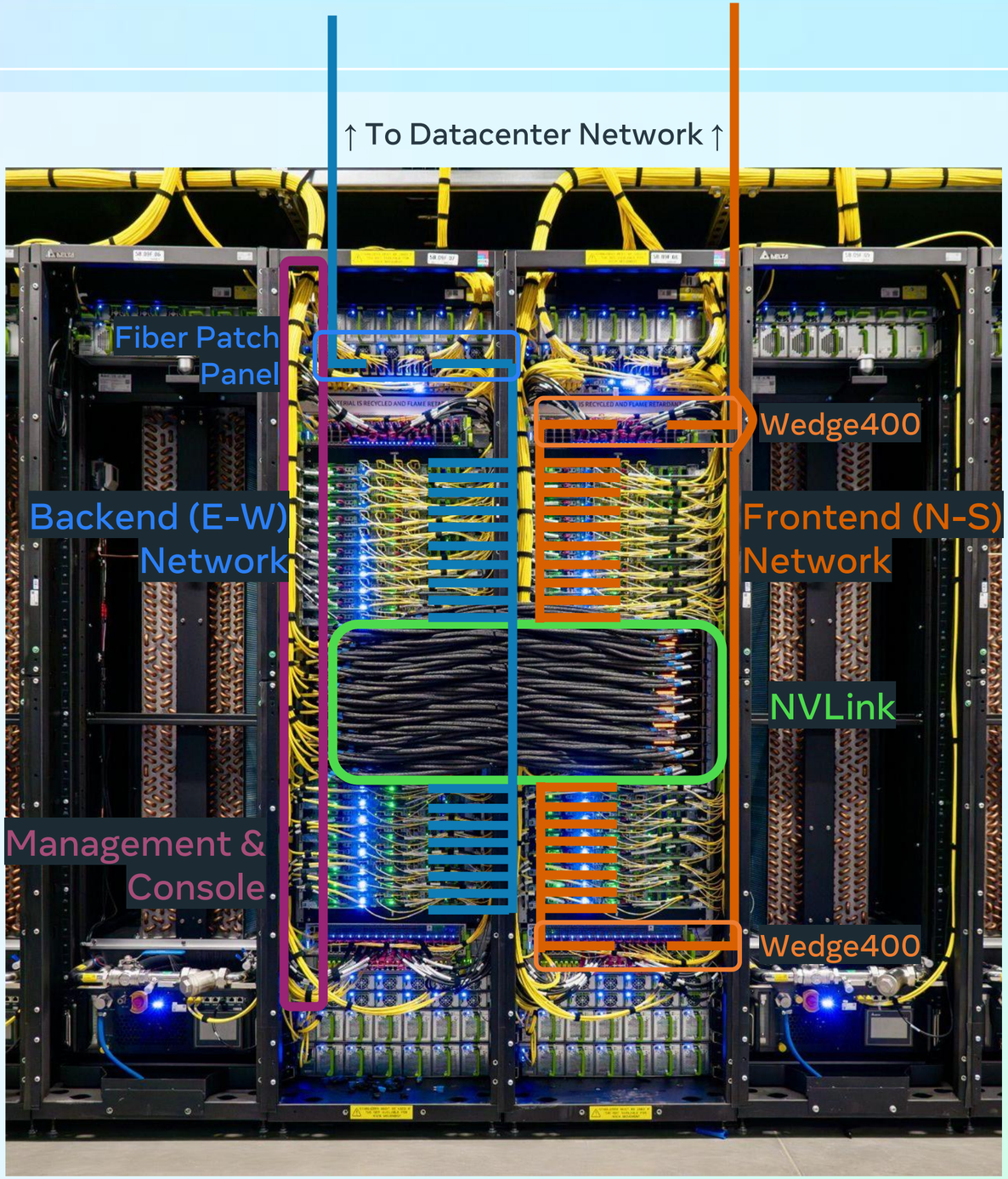
- Cross rack NVLink cables
- Backplane connection to GPU tray

Wedge400 x2 as Frontend Network switch

Fiber patch panel to connect to End of Row (EoR) Switch Rack

Management & Console Cables for switch configuration

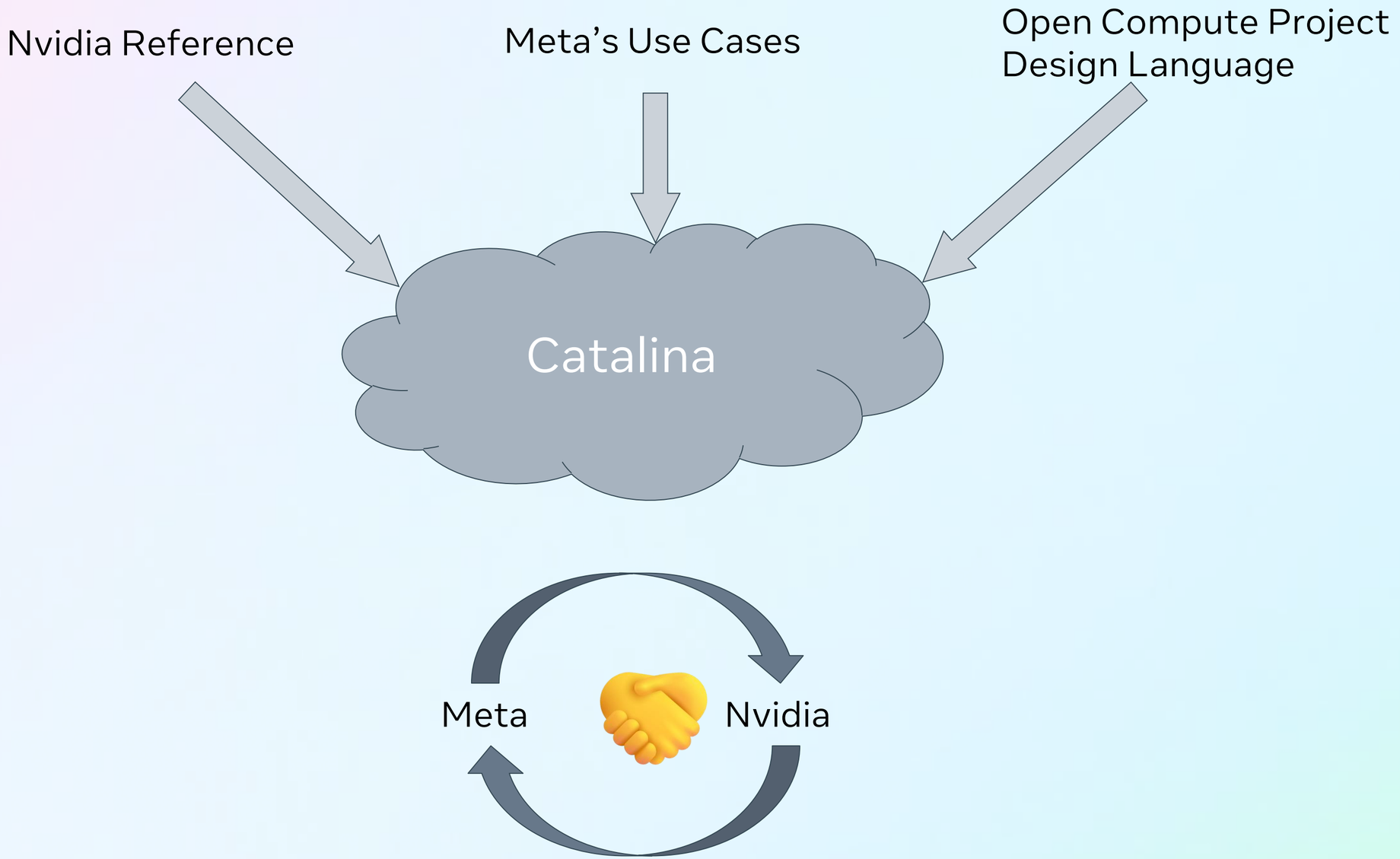
Note: All cables shown exist on both racks



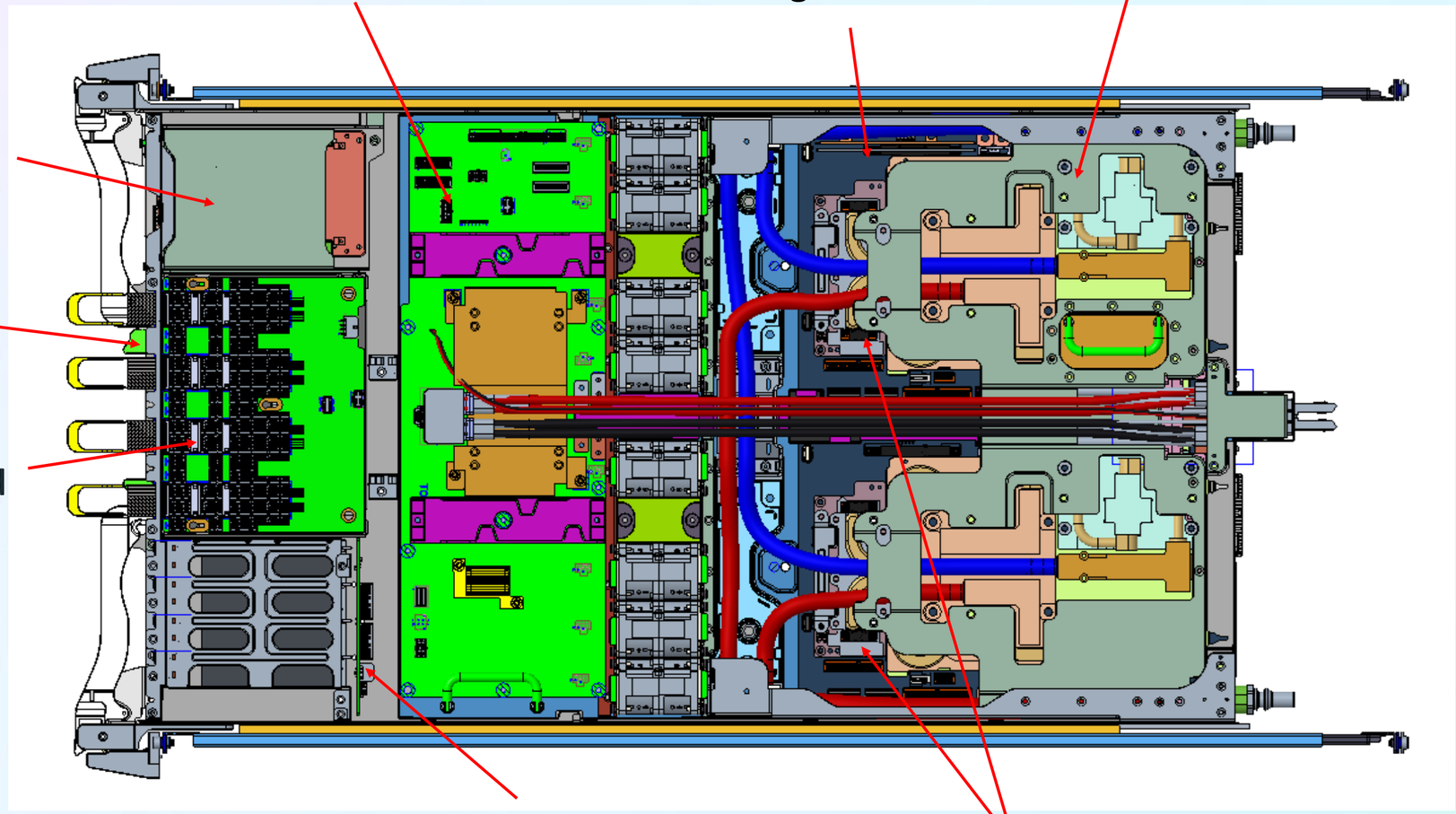
Hardware Architecture

Meta

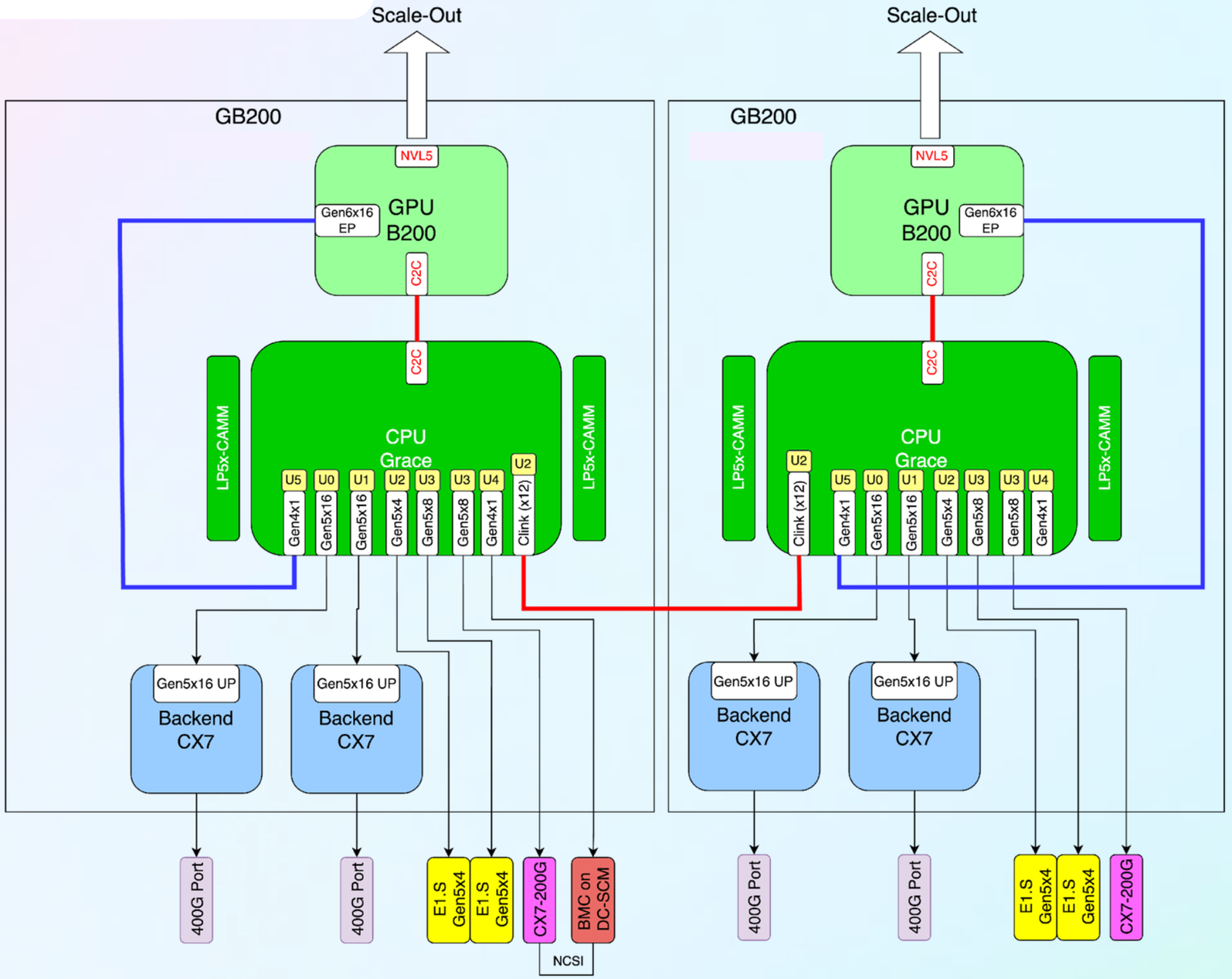
Design Philosophy & Collaboration



Compute Tray



Compute Tray Architecture



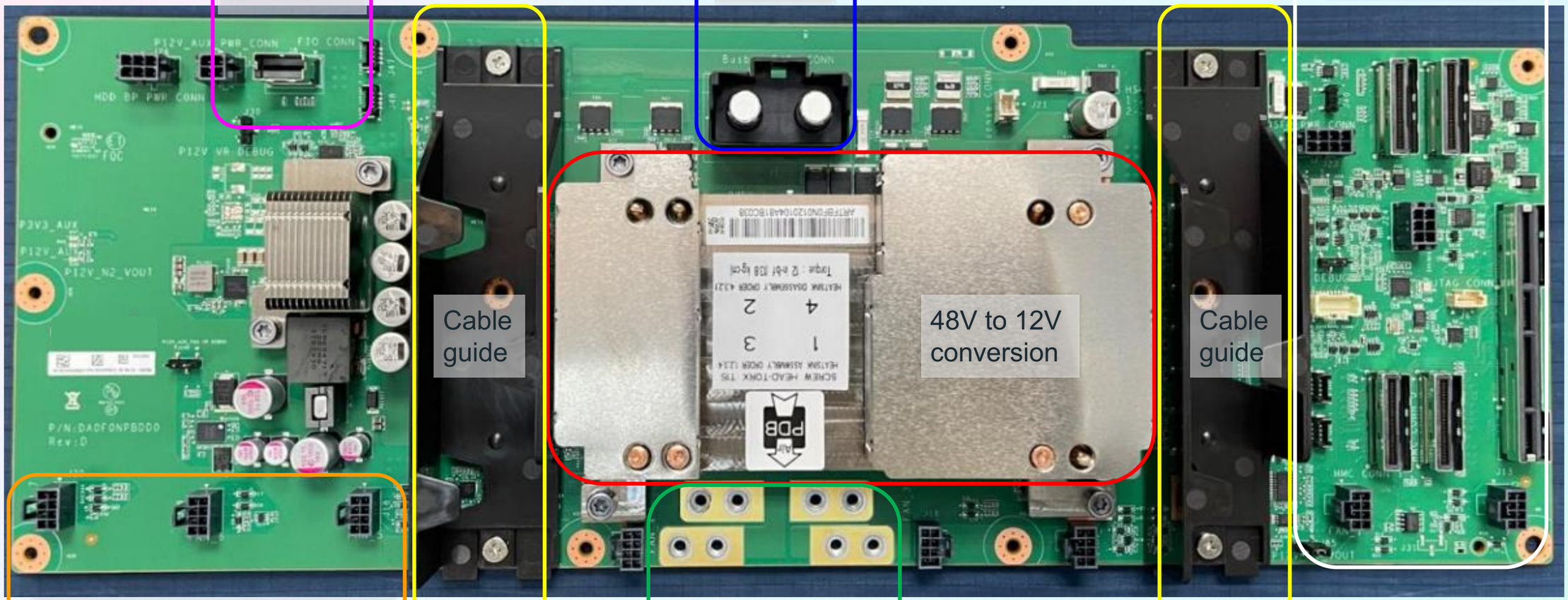
Meta Board Customization

	Meta's implementation compared to Nvidia's GB200 reference
Power distribution board	**New** design to support our management standby power
Management interposer board	Unused
BMC management board	Unused, BMC is on Meta's **DC-SCM**
E1.S midplane	Leveraged, but mechanical updates for 4x 25mm E1.S drives
PCIe riser	Unused, no CEM cards in Catalina. Using **OCP NICs**
OSFP board	Leveraged, but mechanical updates
IPEX board	Unused
Front panel IO board	Meta's FIO is used to connect to the **RMC**
M.2 riser	Unused
TPM board	Meta's TPM design used on the SCM

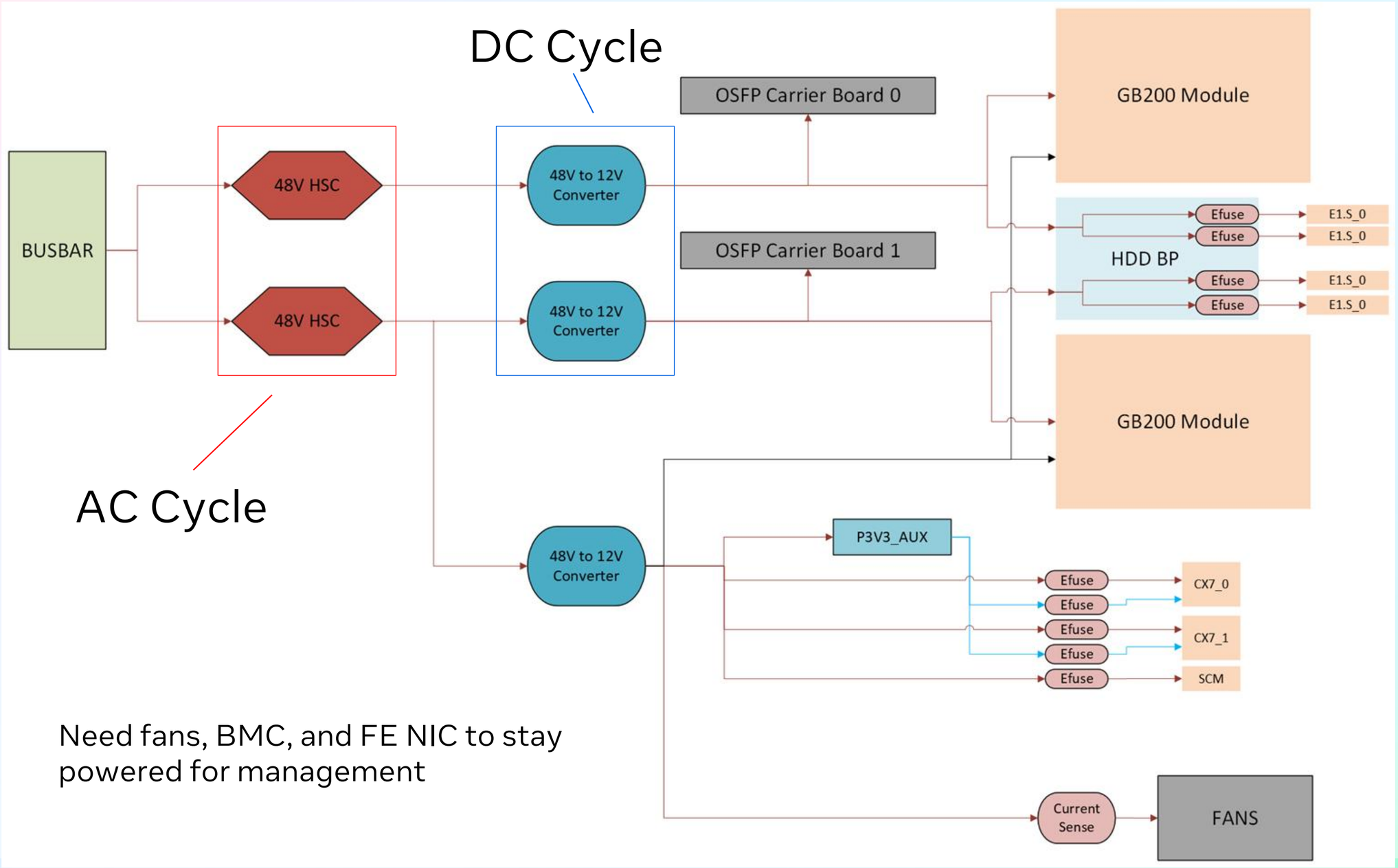
tl;dr:

Different standby power topology - Different leak detection methodology
OCP NICs instead of PCIe add in cards - BMC on a DC-SCM

PDB

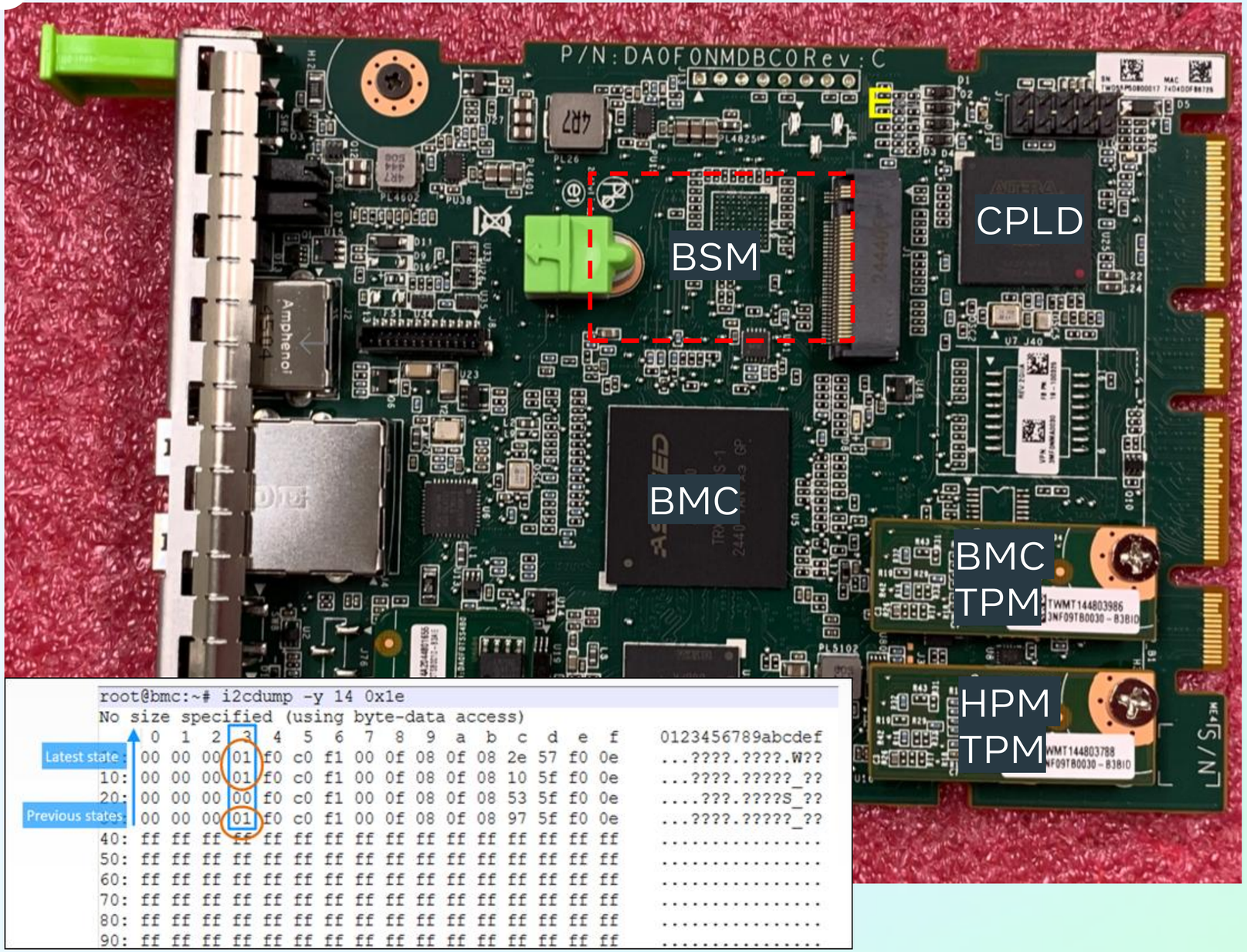

Front IO connection to RMC
Cable to busbar clip
IO Connections
Cable guide
48V to 12V conversion
Cable guide
Connections to leak sensors
Rigid busbar to HPMs

PDB



Need fans, BMC, and FE NIC to stay powered for management

DC-SCM



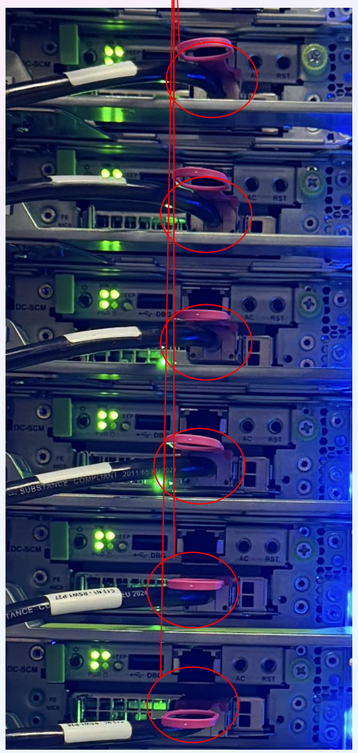

```
root@bmc:~# i2cdump -y 14 0x1e
No size specified (using byte-data access)
     0  1  2  3  4  5  6  7  8  9  a  b  c  d  e  f    0123456789abcdef
00: 00 00 00 01 f0 c0 f1 00 0f 08 0f 08 2e 57 f0 0e    ...????.????.W??
10: 00 00 00 01 f0 c0 f1 00 0f 08 0f 08 10 5f f0 0e    ...????.?????_??
20: 00 00 00 00 f0 c0 f1 00 0f 08 0f 08 53 5f f0 0e    ....???.????S_??
30: 00 00 00 01 f0 c0 f1 00 0f 08 0f 08 97 5f f0 0e    ...????.?????_??
40: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff    ................
50: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff    ................
60: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff    ................
70: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff    ................
80: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff    ................
90: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff    ................
```

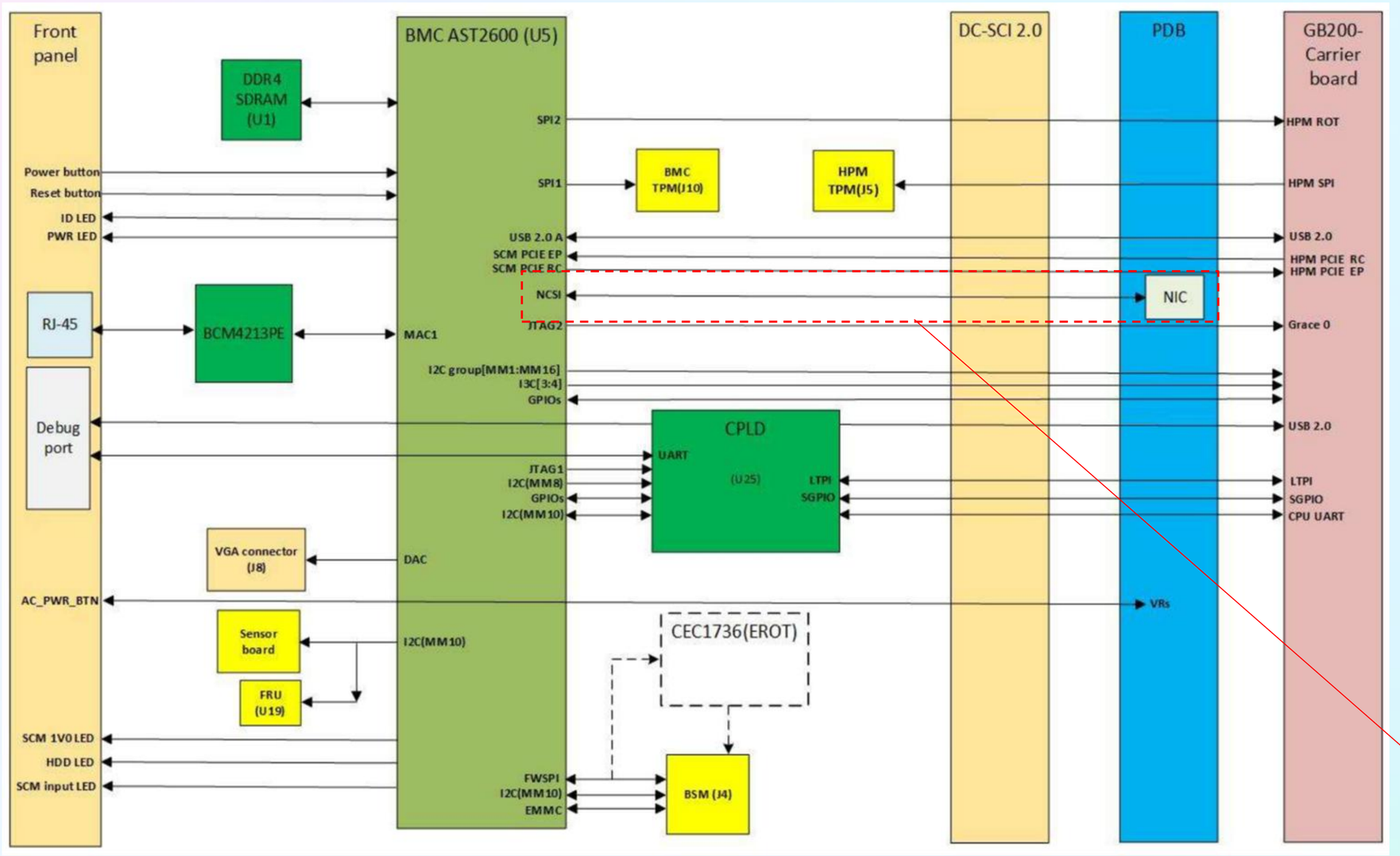
Latest state

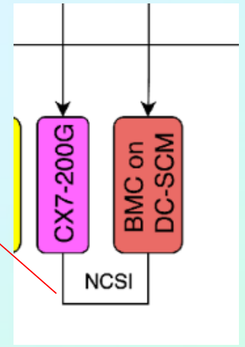
Previous states

SCM

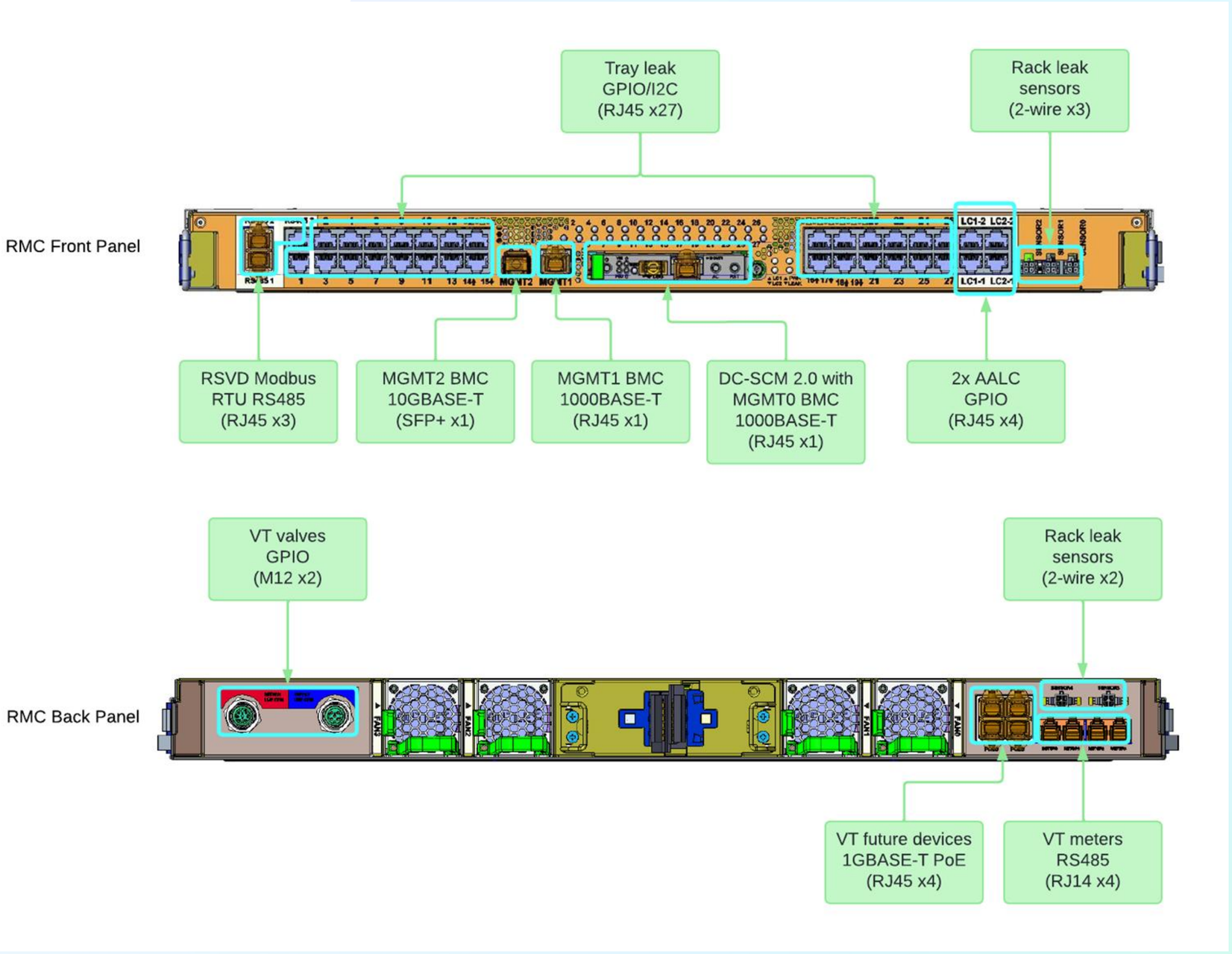
OOB
&
Data




Front panel
DDR4 SDRAM (U1)
BMC AST2600 (U5)
DC-SCI 2.0
PDB
GB200-Carrier board
Power button
Reset button
ID LED
PWR LED
SPI2
SPI1
USB 2.0 A
SCM PCIE EP
SCM PCIE RC
NCSI
JTAG2
I2C group[MM1:MM16]
I3C[3:4]
GPIOs
BMC TPM(J10)
HPM TPM(J5)
HPM ROT
HPM SPI
USB 2.0
HPM PCIE RC
HPM PCIE EP
Grace 0
NIC
RJ-45
BCM4213PE
MAC1
Debug port
CPLD
UART
(U25)
JTAG1
I2C(MM8)
GPIOs
I2C(MM10)
LTPI
SGPIO
USB 2.0
LTPI
SGPIO
CPU UART
VGA connector (J8)
DAC
AC_PWR_BTN
VRs
Sensor board
FRU (U19)
I2C(MM10)
CEC1736(EROT)
SCM 1V0 LED
HDD LED
SCM input LED
FWSPI
I2C(MM10)
EMMC
BSM (J4)

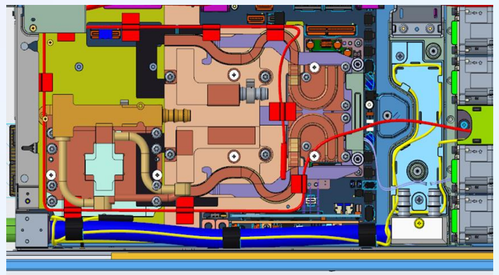

CX7-200G
BMC on DC-SCM
NCSI

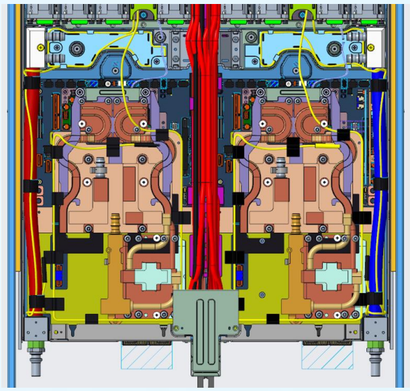
RMC Design/Connectivity

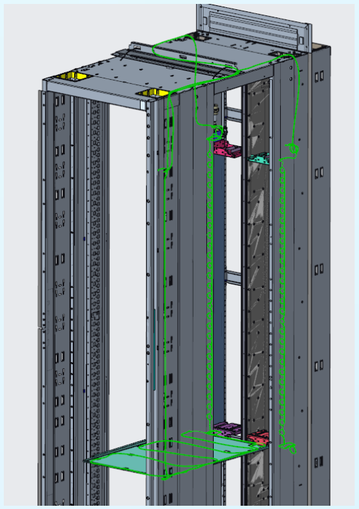


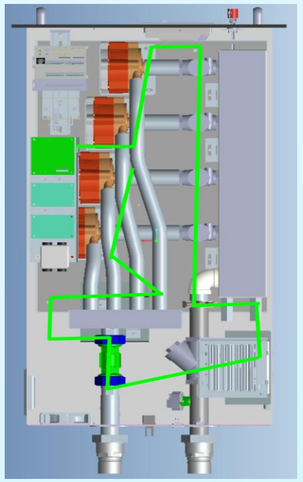
Leakage Sensing

Leakage Level	Small Leak	Tray Leak	Rack Leak	AALC Leak
Areas of Coverage	• Cold plate and quick disconnect joints.	• HPMs and bottom of tray	• Rack Manifold and Drip Pan	• Pump and heat exchanger

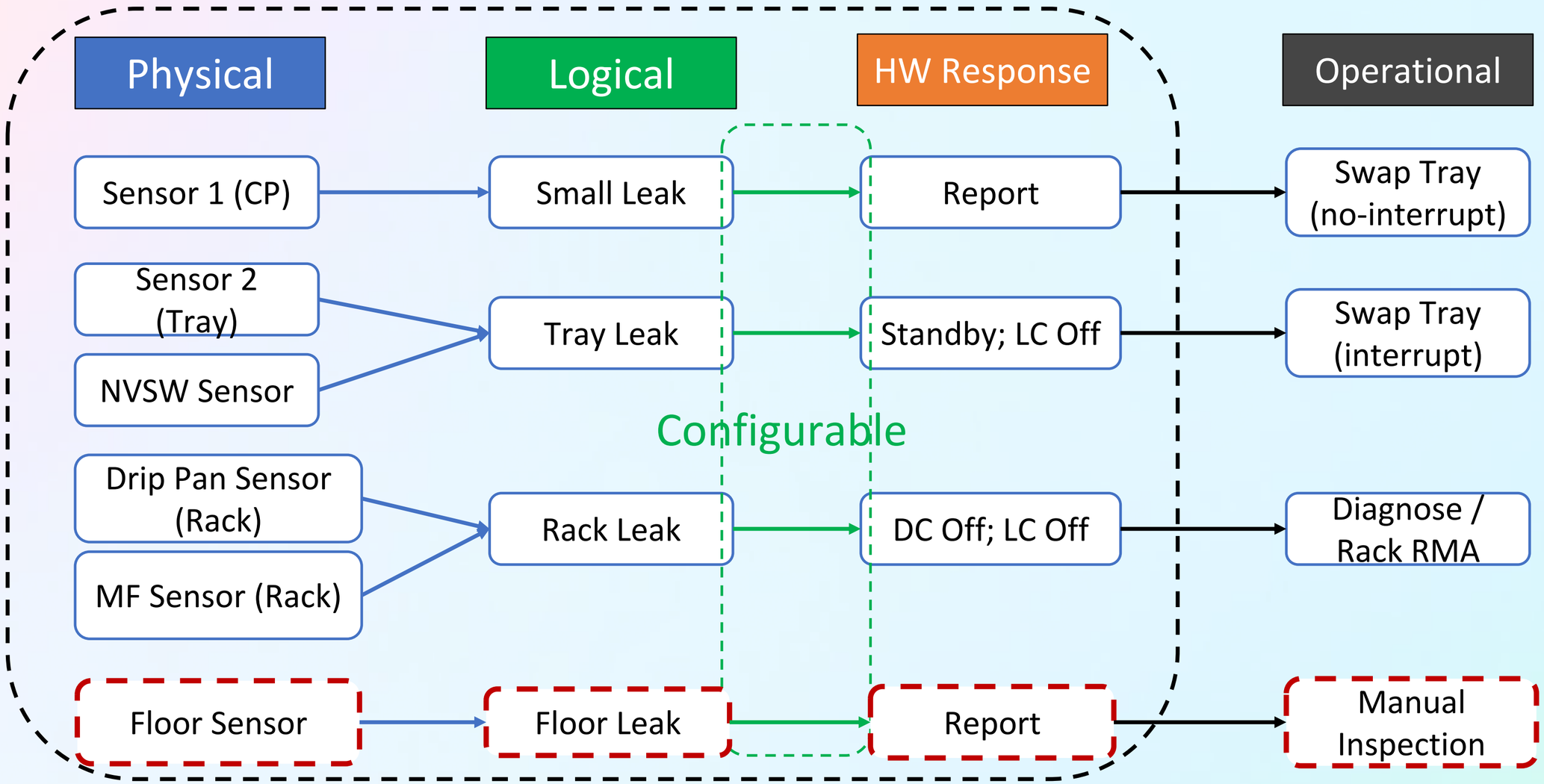








Tier Mapping



Power

Cooling

DC Infra

Large complex AI systems require a huge XFN effort to land, bring up, and scale!

Network

Hardware

Software


Meta