OPEN Compute Project®

OCP-driven Standardization to Deliver High Volume NPO Switch Systems for the Future

May 11, 2022

Xu Wang, Hardware Engineer
Meta Platforms, Inc.

Connect. Collaborate. Accelerate.

Acknowledgement

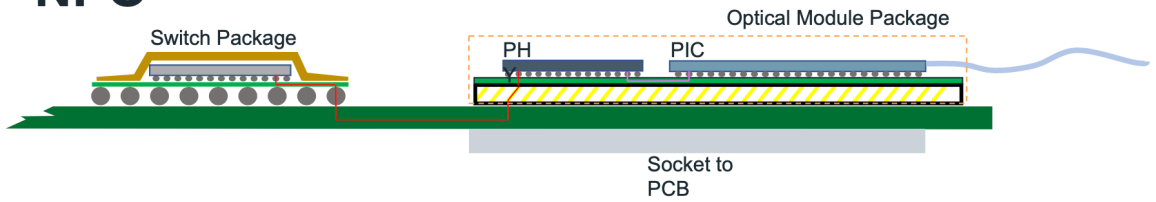
- Thanks to my Meta colleagues James Stewart (Director, Optical Technologies), Rob Stone (Technical Sourcing Manager), Srinivas Venkataraman (Signal Integrity Engineer), Ivy Wu (Mechanical Engineer), Melody Liu (Mechanical Engineer), Nhan Hoang (Mechanical Engineer), and Chris Berry (Optical Engineer) for the slides on CPO / NPO and the concept mezz card and system design!

Drivers for CPO/NPO in Data Centers

- CPO/NPO reduces switch power consumption by 10% to 30% relative to pluggable optics solutions depending on the implementation type and details.
- For 51.2 Tbps fabric switches, the system power limit in Meta DC can be achieved with pluggables using fly-over cables (VSR interface) or CPO/NPO.
- For 102.4 Tbps fabric switches, we do not currently see a path to achieve the system power limit for existing DCs with pluggable optics.
- Meta is developing both switch types (pluggable and NPO) for 51.2 Tbps fabric switches. This provides an opportunity to pipe-clean and prove-in CPO technology for broad deployment in 102.4 Tbps switches.

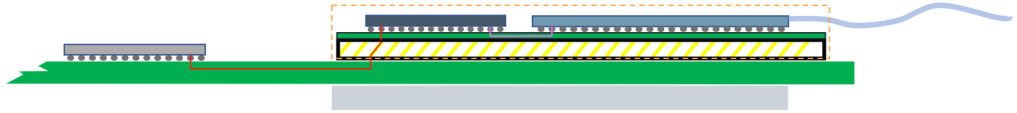
NPO and CPO Definitions

NPO



CPO

(packaged optical module)



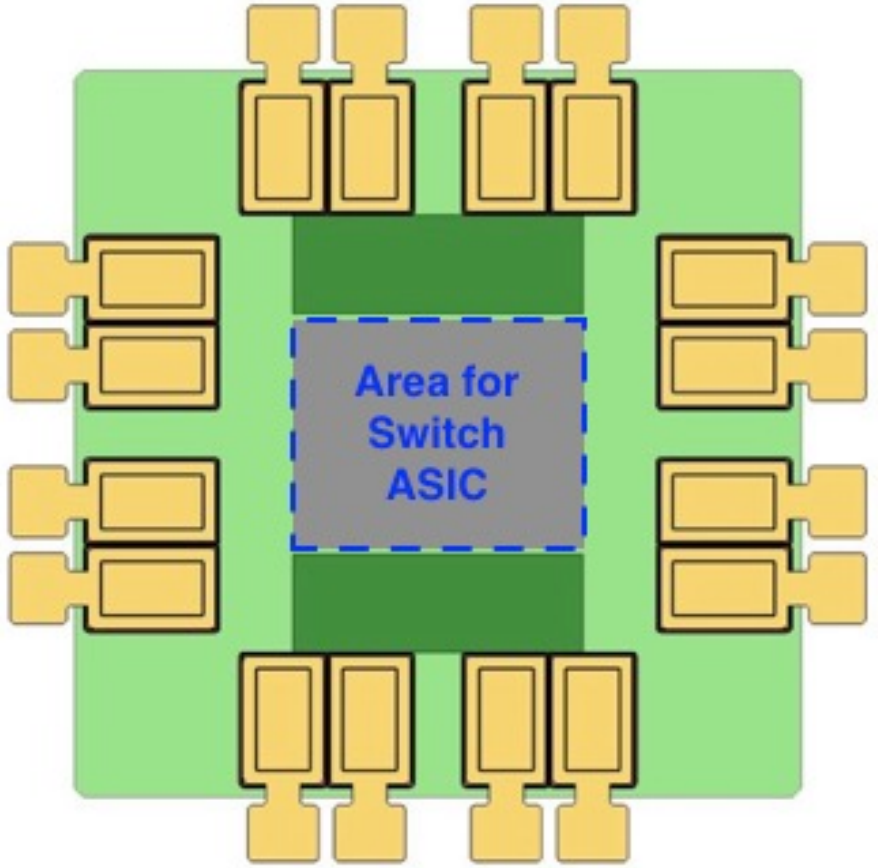
CPO

(bare die / solder attach optical module)

Switch	Optical Module	Switch - Optics Interface	Subsystem Package
Standard Package (BGA / LGA)	Packaged	Retimed, XSR / XSR+	HDI
Bare Die	Packaged	Retimed XSR / XSR+	Organic Build-Up
Bare Die	Bare Die	Retimed / Non-retimed ("Direct-Drive")	Organic Build-Up

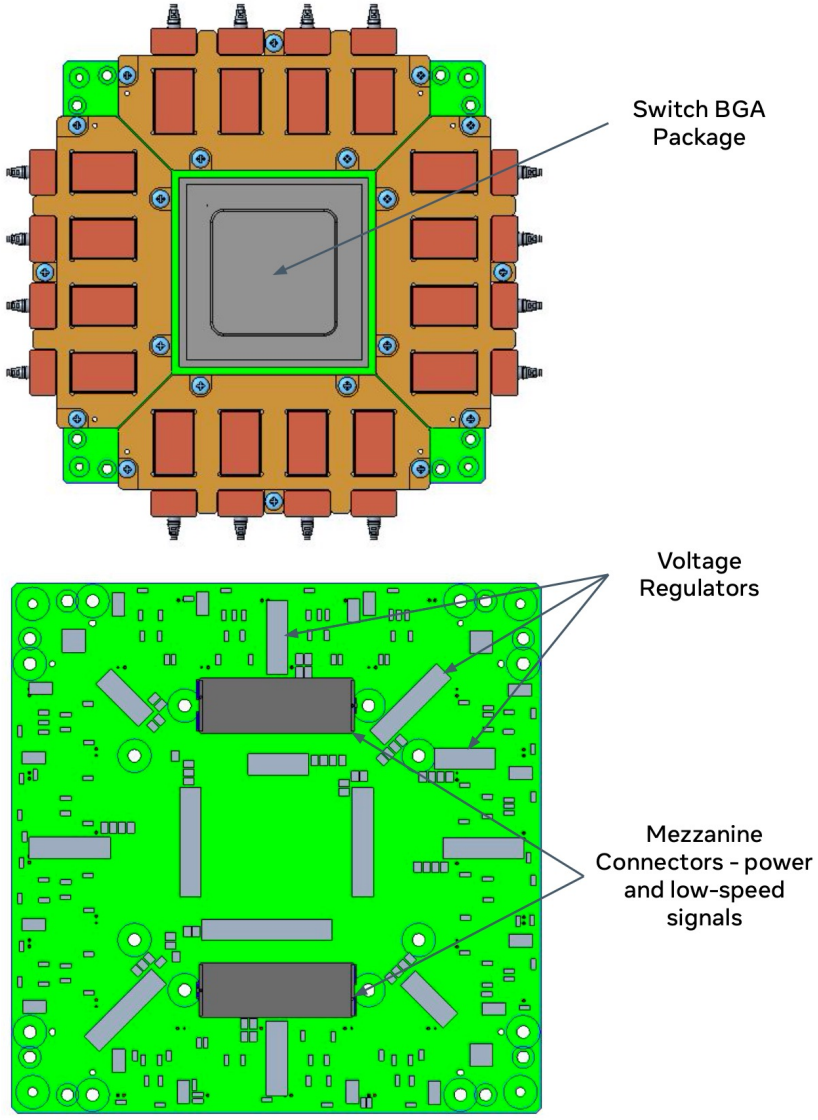
Near-term Implementation Considerations for CPO

- Reduce risk by socketing (w/ LGA socket) Optical Modules (OMs).
  - OMs do not need to survive soldering process.
  - OMs can be replaced after assembly is complete.
- Socketed OMs drive the need for large package substrate (>120mm per side) which will not be available/scalable in timeframe required for 51.2 Tbps switch applications
- Switch ASICs are currently delivered fully tested in BGA or LGA packages. For a CPO switch design this cannot be the case. Either the bare switch die needs to be tested or the switch IC on a large CPO substrate needs to be tested. Switch IC vendors reluctant to change test infrastructure for initial CPO use case.



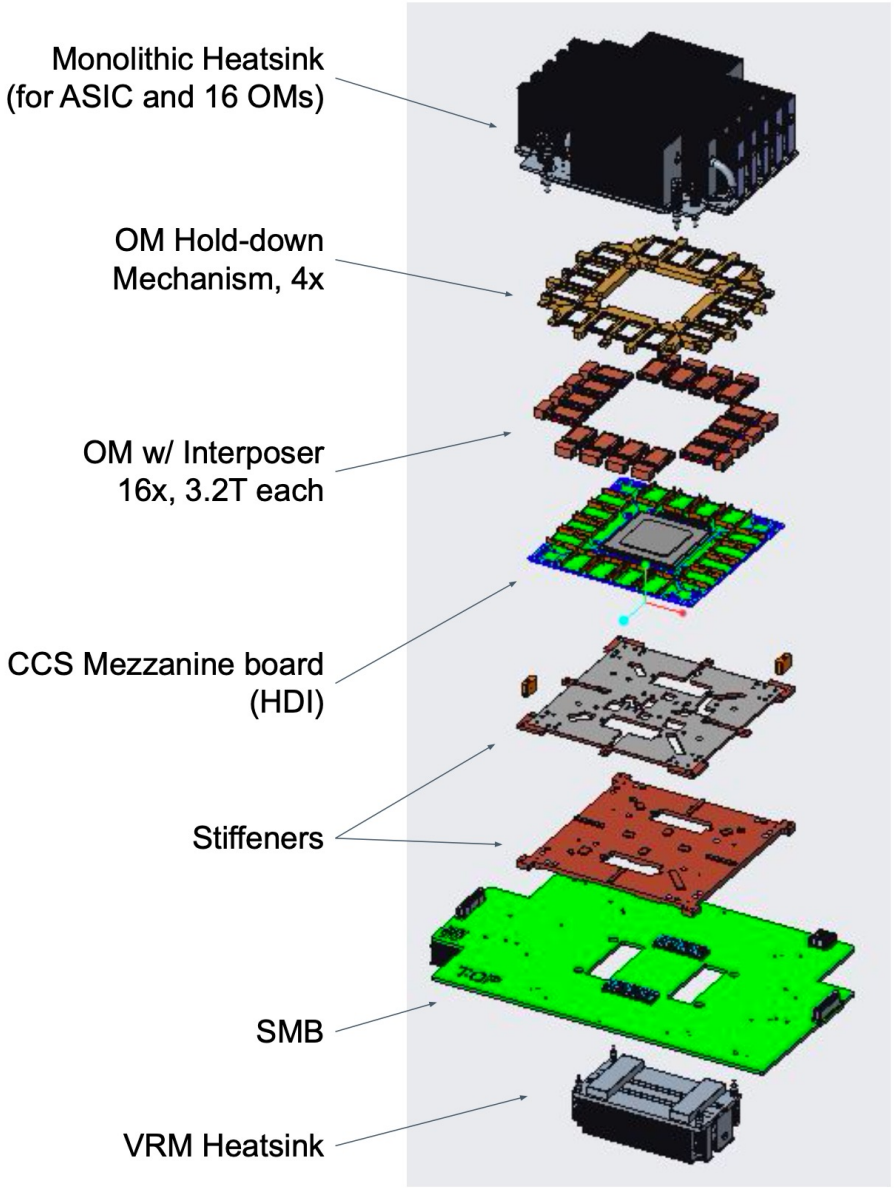
NPO Mezzanine Card

- High-density, socketed, On-Board Optic (OBO) mounted near the switch IC aligned with OIF CPO framework
- Standard BGA package for switch IC
- High Density Interconnect (HDI) mezzanine PCB
  - Mezzanine connectors attach mezz card to Switch Base Board (SBB)
  - Voltage regulators on back side used for high-current, low-voltage supplies (e.g. 0.7V). Reduces the current through the mezz connector
  - 190mm Long x 190mm Wide x 2mm Thick (Subject to change)



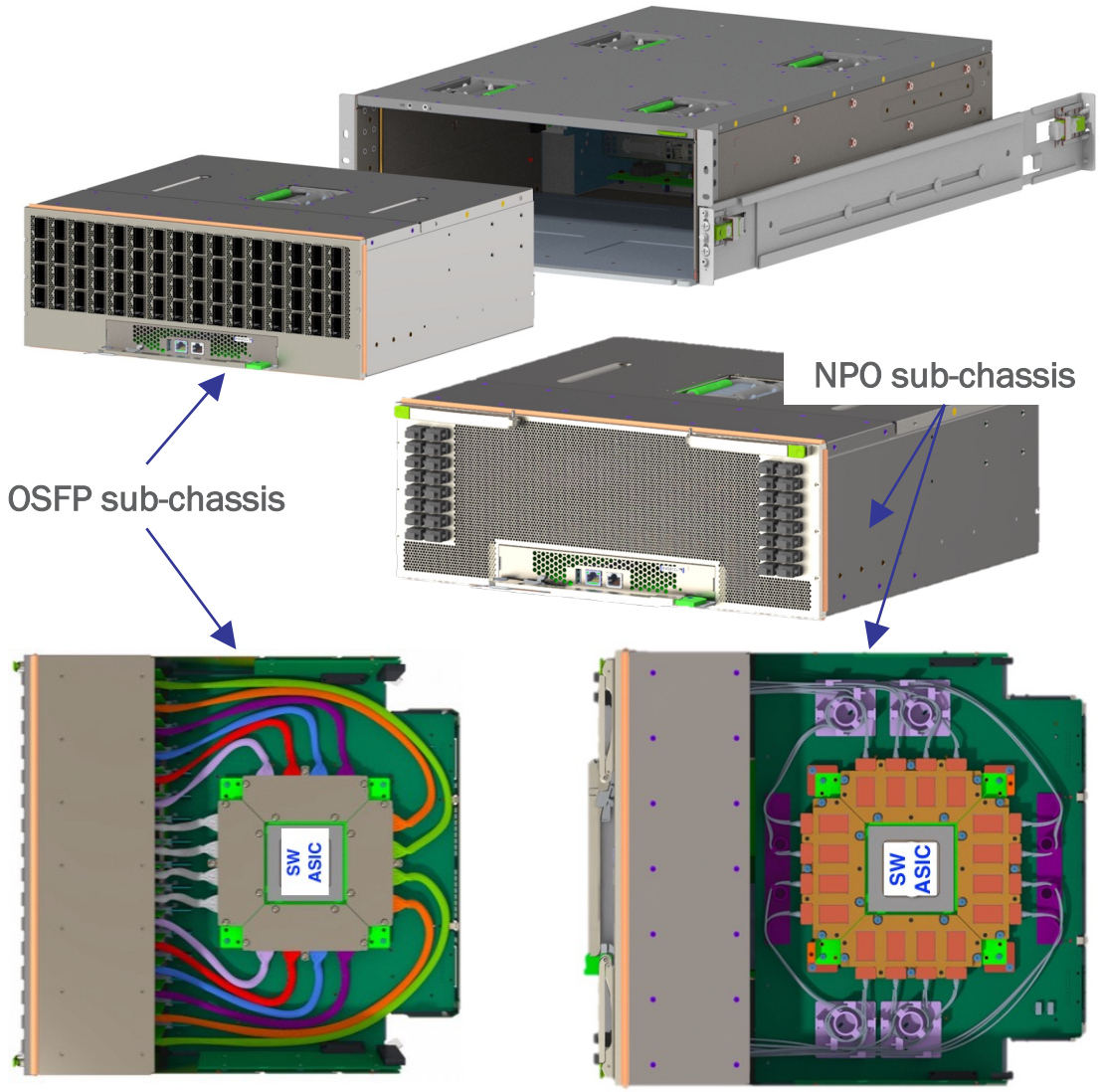
NPO Mezzanine Card Details

- Mezz card attached to stiffener for rigidity and mechanical attachment.
- Mounting plates (4x, 1/side) compress 4 LGA OMs to mezz card to make electrical attachment.
- Mezz card electrically attaches to SMB through mezz connector. Power and control signals only through mezz connector.
- Heat sink for voltage regulators on the back side of the mezz card is applied from bottom-side of SMB through openings.
- Single heat sink mounted on top for switch IC and OMs.



Switch System Design

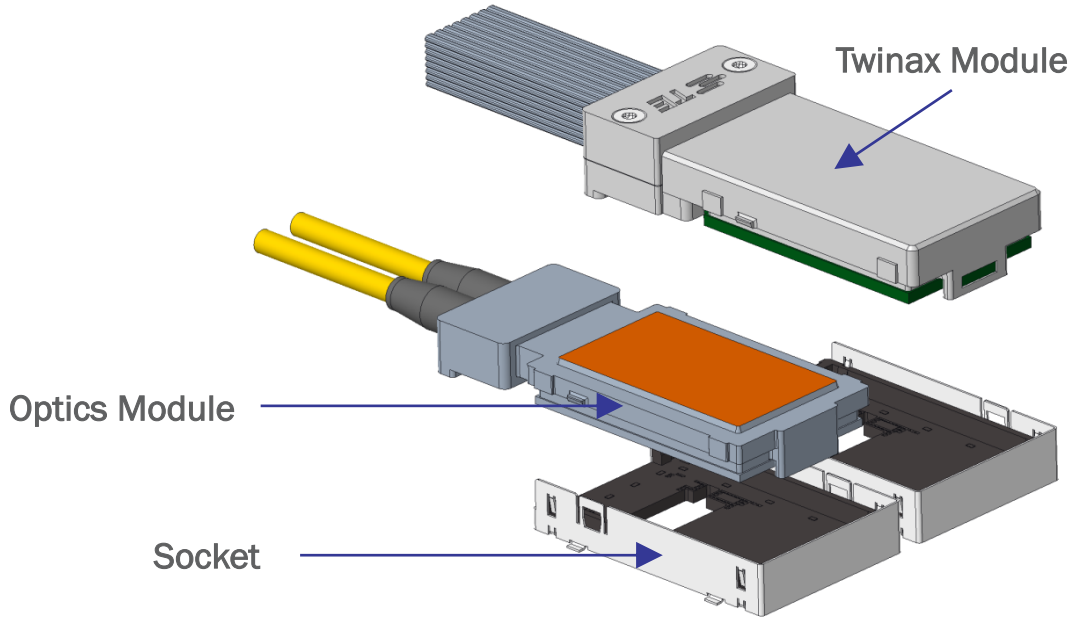
- Single switch ASIC-based system
- SBB and NPO mezz card are contained in a sub-chassis that can be inserted/removed from the front side of the switch.
- Chassis height of 4RU driven by Meta requirement to minimize fan power.
- Common platform
  - NPO: Internal fiber routing from the NPO OM to the front panel
  - Pluggables: Flyover cable routing from the OM footprints to the front panel OSFP



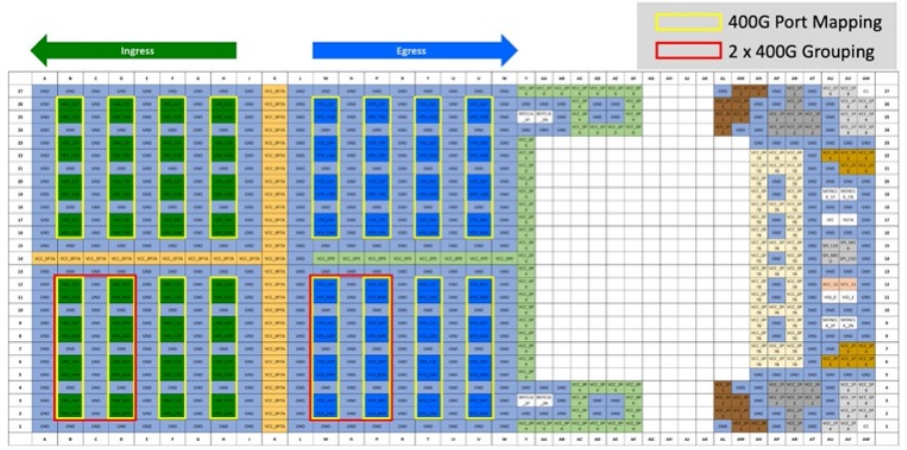
Common Connector Socket (CCS)

- A twin-ax connector can be used on the same LGA footprint used for the NPO OM.

- Common Connector Socket (CCS-twinax) on one side breaks out to twin-ax cables and connecting to an OSFP cage (or other pluggable form factors) on the other side



Artwork from TE contribution to OIF: oif2022.039.02



OM socket pin definition being standardized at OIF

Mezzanine Connector Pin Lists

- Power delivery
  - High voltage power pins for high power rails, such as 48V or similar intermediate bus voltages
  - Low voltage power pins for low power rails, such as 5V, 3.3V, etc.
- System control
  - Power sequencing control signals
  - PCIe interface for switch ASIC control, including PCIe reference clock
  - Other control interfaces for the switch ASIC, such as SPI, I2C, and LED serial interfaces
  - Control interface for the OMs, such as SPI
  - Control interface for the voltage regulators, such as PMBus/I2C
- Clocking
  - SerDes reference clocks for the switch ASIC and the OMs
  - PTP signals

System Advantages of NPO Mezzanine Card

- Modularity
  - Separates the ASIC + OM subsystem from the base system
  - Common platform for optics variants: NPO or pluggables
  - Common platform for different switch ASICs
  - Base system components
    - Micro-server
    - BMC
    - Power delivery
    - Cooling and mechanical support
- Velocity of development and deployment
  - Eliminates SI and power delivery challenges from the common platform
- High-speed PCB cost reduction

Challenges for NPO Mezzanine Card

- Pros and cons with HDI PCB technology
  - Smaller geometry compared to traditional PCBs
    - Thinner dielectrics
    - Finer trace width and spacing
    - Smaller via diameters
  - More flexible routing options and generally better SI
  - Limited in layer count
  - Limited in board thickness
- Alternatives: Sequential lamination with conventional PCB geometries

Challenges for NPO Mezzanine Card (2)

- Power delivery
  - Power conversion on the mezz card and vertical power delivery reduce DC power loss.
  - Very challenging to implement the power conversion on the mezz card due to limited space
  - Due to the layer count limit, difficult to achieve low DC power loss when distributing power horizontally
  - If adding dedicated power connectors for large current from the SBB, alignment across multiple connectors can be challenging.

Call for Action

The OM form factor and its footprint, including CCS-twinax, is being standardized at OIF.

OCP CPO Project

The OCP community can make an impact by driving commonality of form factor, power envelop and mezz connectors to ensure we have the most efficient outcome for all players in the ecosystem to leverage on.

We propose to standardize the NPO mezzanine card form factor in the OCP CPO project.

- *Mechanical spec: Card profile, etc.*
- *Thermal spec: Air cooling and liquid cooling*
- *Mezzanine card connectors and signal definitions*

Thank You

Connect. Collaborate. Accelerate.