# Be Ready for CPO

## Integration and Enhancing CPO Switches with SONiC

OCP GLOBAL SUMMIT

OCT 15–17, 2024
SAN JOSE, CA

NETWORKING

# Be Ready for CPO

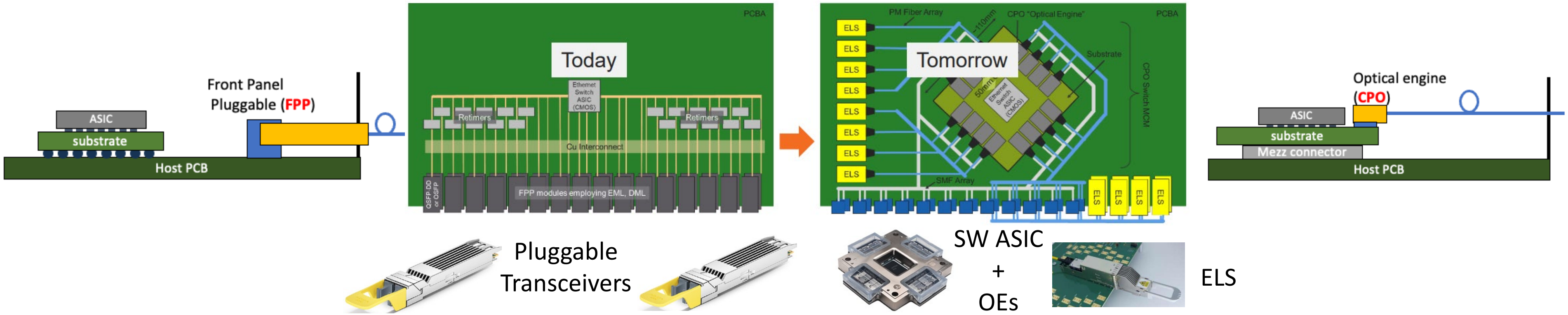
## Integration and Enhancing CPO Switches with SONiC

Wataru Ishida - NTT Devices

Yuki Arikawa - NTT

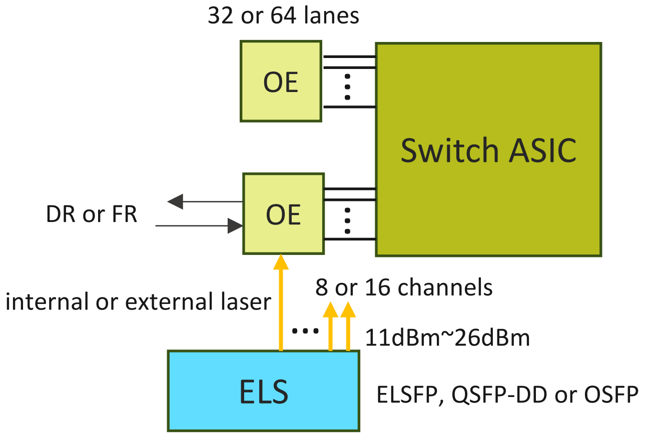
# Co-Packaged Optics(CPO) Switch

- No pluggable transceivers - optical engine(OE) and (optional) external laser source(ELS)
- Motivation
  - Power saving - shorter electrical trace between ASIC and Transceivers
  - Future-proof architecture for higher SerDes rate (> 200G/lane)
  - Potential higher reliability by redundant external laser source



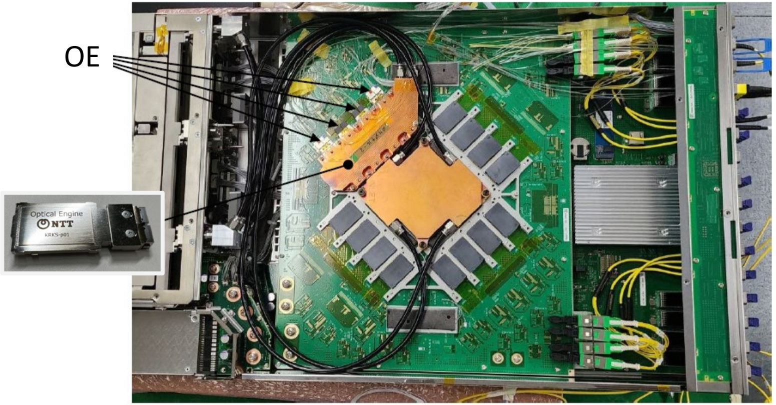
# CPO Switch - many design options

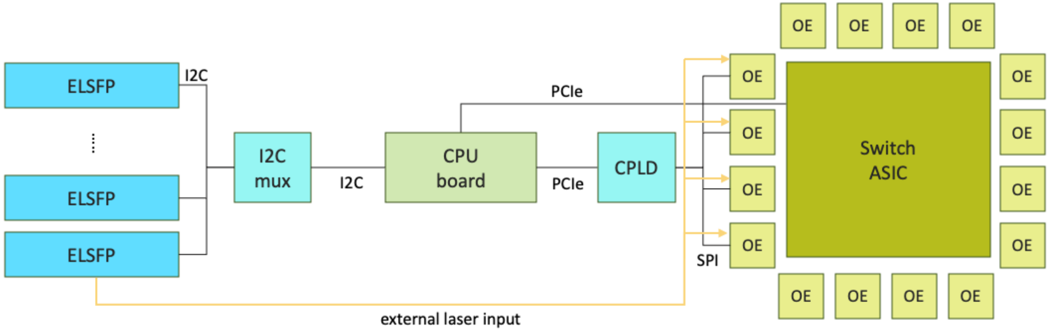
- no best-practice, no industry-accepted design yet
  - Standardization Effort: OIF IA for a 3.2Tb/s Co-Packaged(CPO) Module (March 2023)
- Optical Engine(OE) (non-pluggable):
  - # of lanes: 32, 64
  - PMD: DR, FR(CWDM), BiDi
  - laser: internal or external
- External Laser Source(ELS) (pluggable):
  - form factor: ELSFP(blind mate), QSFP-DD, OSFP
  - # of channels: 8, 16
  - output power: from Low Power 11dBm to Super High Power(26dBm)
- Mixed configuration is also possible: e.g. half-CPO, half-FPP



# NTT CPO project

- We developed an OIF IA-compliant CPO switch prototype
  - 51.2T switch ASIC
  - x16 3.2T OE (8x400G DR4)
  - x8 external lasers (8ch, 20dBm)

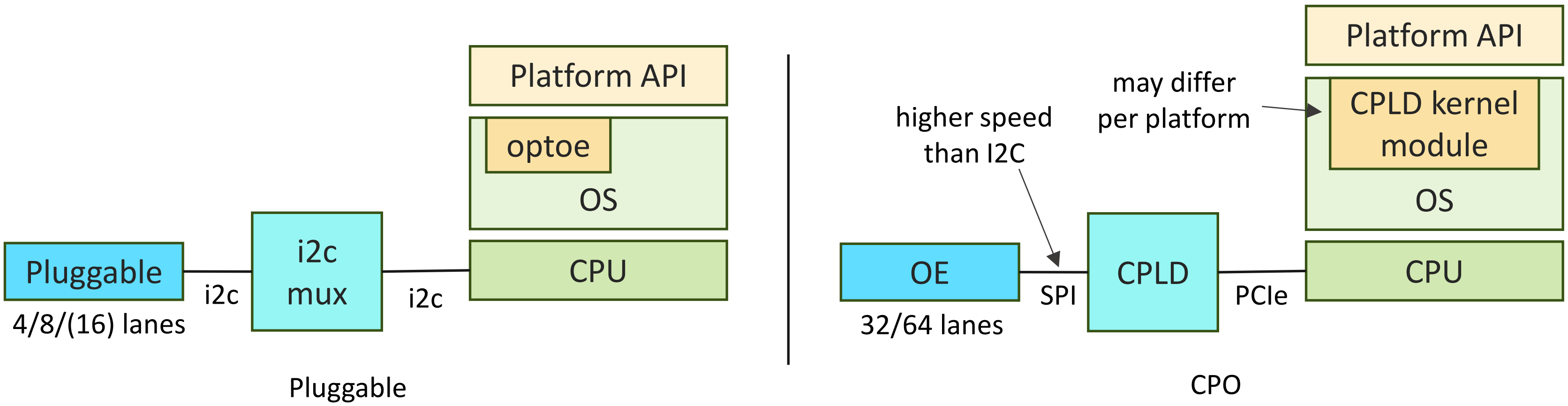




Control Bus Architecture

# OE management - SPI support

- Pluggable transceivers have been controlled via I2C
  - SONiC uses optoe for all platforms to provide EEPROM access
- This will change in CPO
  - SPI is chosen for the management communication interface(MCI) in OIF CPO IA
  - a kernel module for the CPLD that terminates the SPI bus will be needed



# OE management - CMIS 3D addressing (1/2)

- CMIS uses 3D addressing for the EEPROM memory space
  - [Bank, Page, Offset]
- Banked pages contain information for 8 lanes (see the example below)
- CPO OE typically has more than 32 lanes
  - OE management requires bank switching to cover all lanes
    - e.g. 4 banks to cover 32 lanes

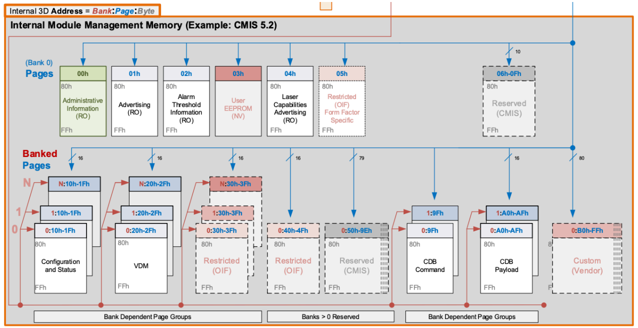


**Figure 8-1 CMIS Module Memory Map (Conceptual View)**

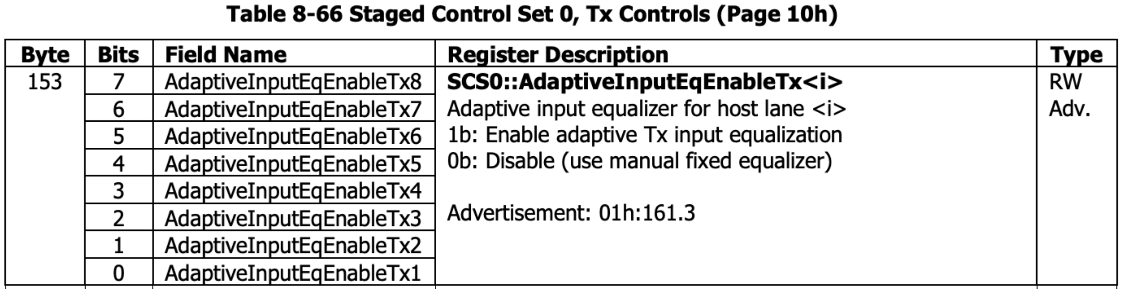


**Table 8-66 Staged Control Set 0, Tx Controls (Page 10h)**

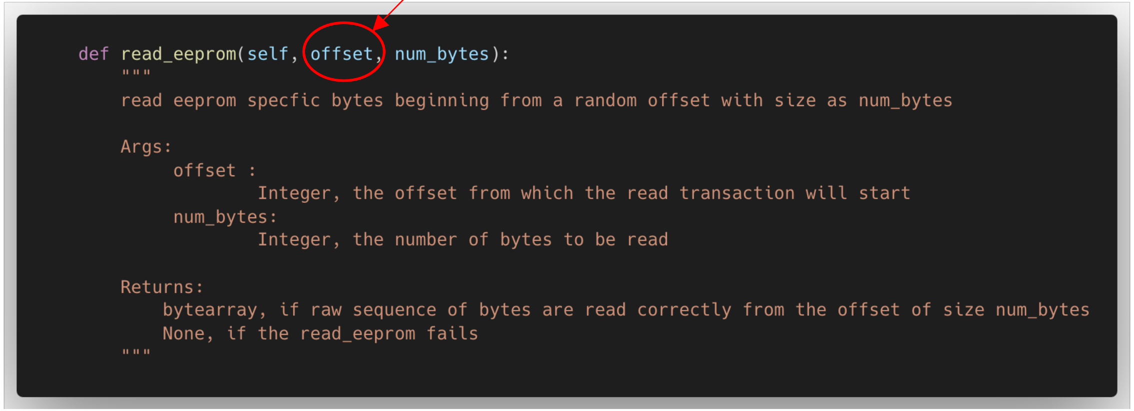
| Byte | Bits | Field Name | Register Description | Type |
|---|---|---|---|---|
| 153 | 7 | AdaptiveInputEqEnableTx8 | **SCS0::AdaptiveInputEqEnableTx<i>**<br>Adaptive input equalizer for host lane <i><br>1b: Enable adaptive Tx input equalization<br>0b: Disable (use manual fixed equalizer)<br>Advertisement: 01h:161.3 | RW<br>Adv. |
| | 6 | AdaptiveInputEqEnableTx7 | | |
| | 5 | AdaptiveInputEqEnableTx6 | | |
| | 4 | AdaptiveInputEqEnableTx5 | | |
| | 3 | AdaptiveInputEqEnableTx4 | | |
| | 2 | AdaptiveInputEqEnableTx3 | | |
| | 1 | AdaptiveInputEqEnableTx2 | | |
| | 0 | AdaptiveInputEqEnableTx1 | | |

Example of Banked CMIS register

# OE management - CMIS 3D addressing (2/2)

- SONiC (and optoe) doesn't support banked pages (Bank is always 0)
- optoe maps CMIS 2D address space to a linear address space
- SONiC xcvr API needs to be updated to support Bank

optoe linear address

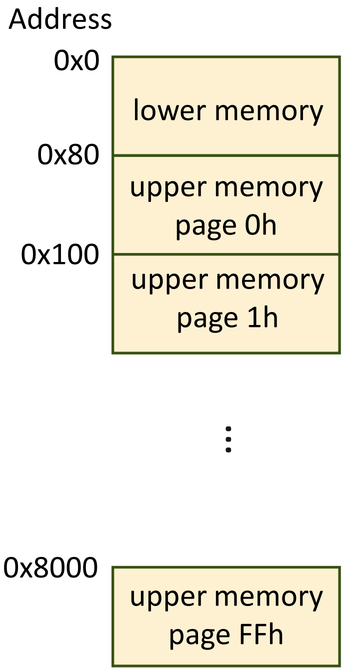


```
def read_eeprom(self, offset, num_bytes):
    """
    read eeprom specfic bytes beginning from a random offset with size as num_bytes

    Args:
         offset :
                 Integer, the offset from which the read transaction will start
         num_bytes:
                 Integer, the number of bytes to be read

    Returns:
        bytearray, if raw sequence of bytes are read correctly from the offset of size num_bytes
        None, if the read_eeprom fails
    """
```

SONiC xcvr API read_eeprom function signature



Address

0x0

lower memory

0x80

upper memory page 0h

0x100

upper memory page 1h

⋮

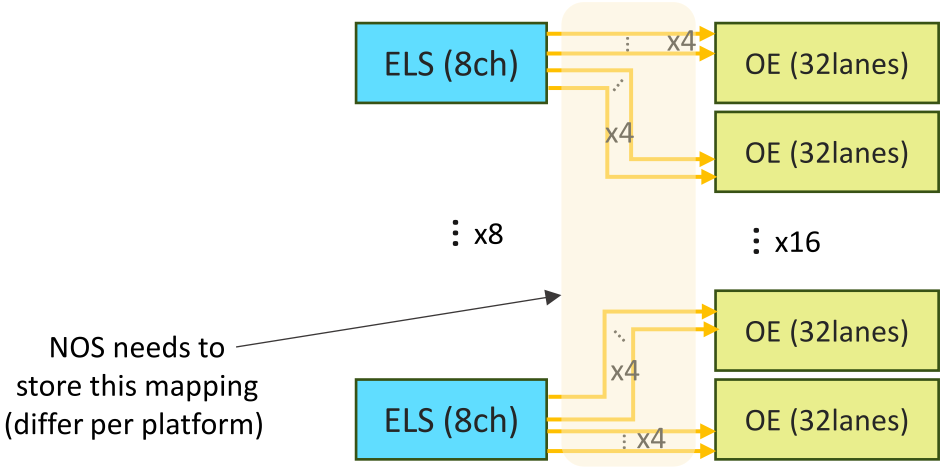
0x8000

upper memory page FFh

optoe linear address space

# OE - ELSFP coordination (1/2)

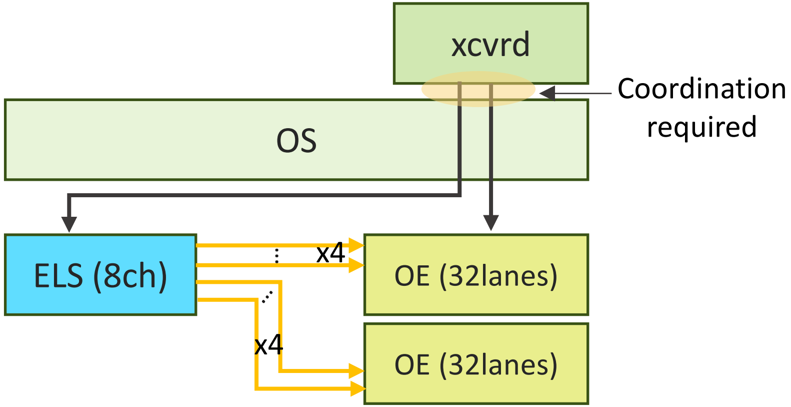
- NOS needs to understand which ELSFP channel is used for which OE lane
  - physical fiber mapping between ELSFP and OE
- In the case of our prototype, 1 laser channel is used for 8 lanes of an OE.
  - disabling one laser channel turns down multiple lanes (and possibly ports)



NOS needs to store this mapping (differ per platform)

# OE - ELSFP coordination (2/2)

- Before initializing an OE, NOS needs to turn on the connected laser channel
  - additional coordination is required between ELSFP and OE
- In the case of SONiC, xcvrd is the component that manages transceivers
  - To support CPO, xcvrd needs to manage ELSFP as well as OE



laser channel needs to be turned on here

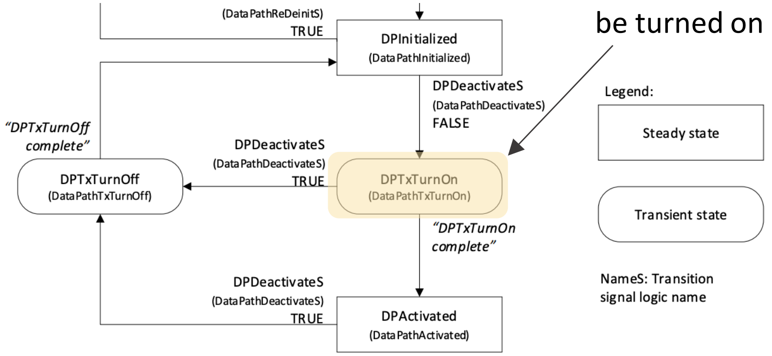


**Figure 6-5 Data Path State Machine (DPSM) State Transition Diagram**

OE state machine (partial)

# Enhancement of SONiC Virtual Switch Environment

- CPO switch has many design options

- capability to experiment/develop new mechanisms to support various CPO switches in a virtualized environment without a real platform is important

- In SONiC, xcvrd will be the main component that needs to be updated to support the CPO switch
  - However, currently, xcvrd is disabled in the SONiC Virtual Switch Environment
  - We have enabled xcvrd in the SONiC Virtual Switch Environment
  - Details will be covered in a presentation at the SONiC Workshop Part2 on Oct. 18

# Call to Action

- The development of CPO switches is not just a hardware issue
  - support in the NOS is also required
- CPO switches come with various design options
  - software architecture that can handle them uniformly is required
- Deepen discussions within the OCP and SONiC community
  - Promote and build awareness of the CPO system

# Thank you!