25-50-100G Ethernet Options and Experience in the Datacenter

Paul Zugnoni, LinkedIn

NANOG 68

25-50-100G is the new 10-40G

- Economically sensible
- Multiple router/switch/NIC products
- Multiple optics sources

- Pre-fab Twin-Ax cables
- MMF and SMF options
- Short range, Long range and Extended range options
- Divide by 4
- No speed change in optical-electrical operation

- IEEE standards for multiple transmission options

100G is the new 40G

100G QSFP28 Modules	40G QSFP+ Modules
100GBase-CR4 ● Copper Twin-ax up to 7m ● 4 wire lanes @ 25G each ● Breakout optional	40GBase-CR4 ● Copper Twin-ax up to 7m ● 4 wire lanes @ 10G each ● Breakout optional
100GBase-SR4 ● MMF/MTP up to 100m (OM4) ● 4 fiber lanes @ 25G each ● Breakout optional	40GBase-SR4 ● MMF/MTP up to 150m (OM4) ● 4 fiber lanes @ 10G each ● Breakout optional
100GBase-LR4 ● SMF/LC up to 10km ● 4 optical lanes @ 25G each	40GBase-LR4 ● SMF/LC up to 10km ● 4 optical lanes @ 10G each
100GBase-PSM4 ● SMF/MTP up to 500m or 2km ● 4 fiber lanes @ 25G each ● Breakout optional	Vendor Specific Parallel LR4 ● SMF/MTP (distances vary) ● 4 fiber lanes @ 10G each ● Breakout optional

100G - Additional datacenter interface options

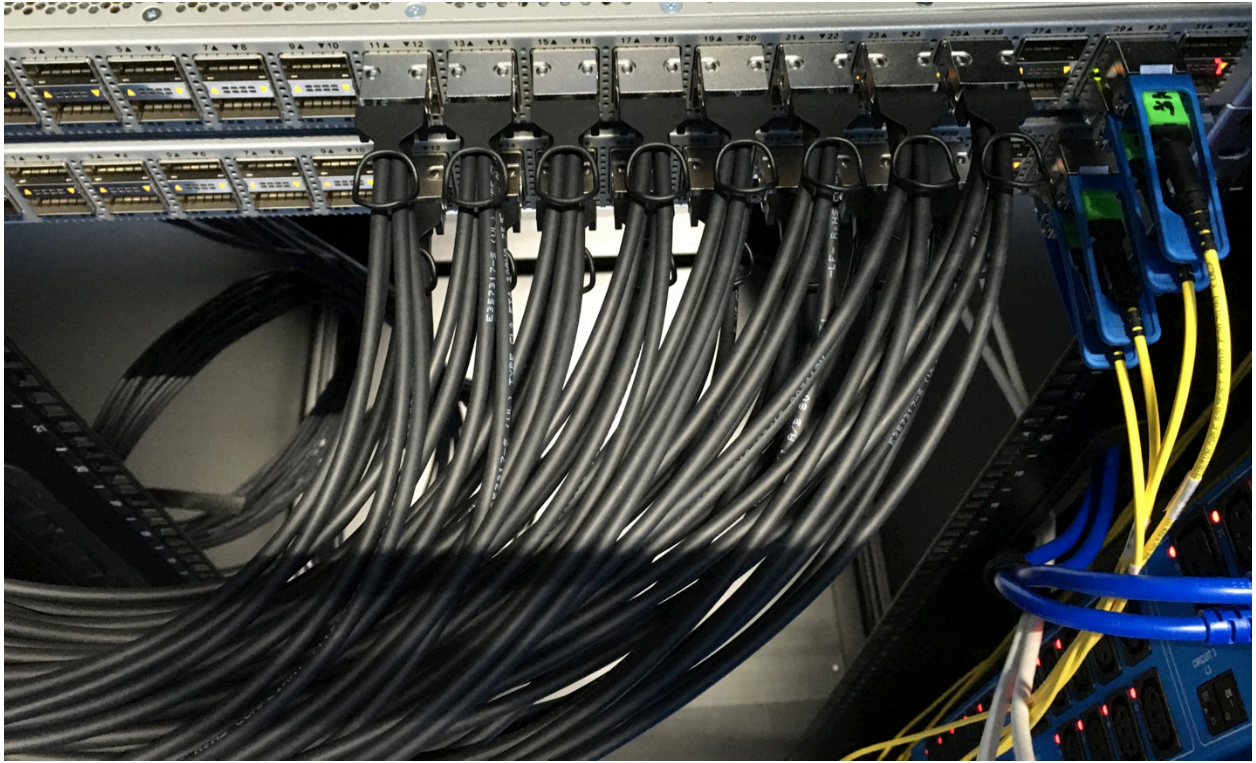
Other 100GBase Standards that might be useful in your datacenter:

- 100GBase-CR10
  - 10 x 10GBASE-CR (Twin Ax Breakout up to 7m)
- 100GBase-SR10
  - 10 x 10GBASE-SR (MMF/LC up to 100/150m)

Considerations for new pluggables: Environment

- Temperatures
  - Even in a climate controlled datacenter, check that the chassis of the switch you use can support the arrangement and quantity of optics you want.
- Density
  - Newer optics pack more power in same or smaller footprint
  - Pizza box switches are popular. Is your cable management ready?
  - Chassis based systems also supporting denser pluggables
- Distance to patch
  - Can you plan your patch panels around the model of network gear?
  - Or do you need to plan for more unknown?
- Existing gear, Existing cabling
  - Do you have existing core or edge gear that the new stuff needs to connect?
  - Are you connecting to NEW gear that doesn’t support the optimal line speeds or media type you want?
  - Do you have one cable type or another you MUST use?

Considerations for new pluggables: Density



Considerations for new pluggables: Interoperability

Interoperability -
Optic-to-Optic connecting two switches.

We tested modules from three suppliers.

Suppliers A, B, C

A == A: OK	A == B: OK
B == B: OK	A == C: OK
C == C: OK	B == C: FAIL

Issues

1. Whether FEC enabled by default vs. FEC cannot be disabled
2. 1550nm vs 1310nm, and finding that one side could not read Rx.

Both discovered through discussions with suppliers. Moved forward with “most compatible” two models.

Deployment experiences with PSM4

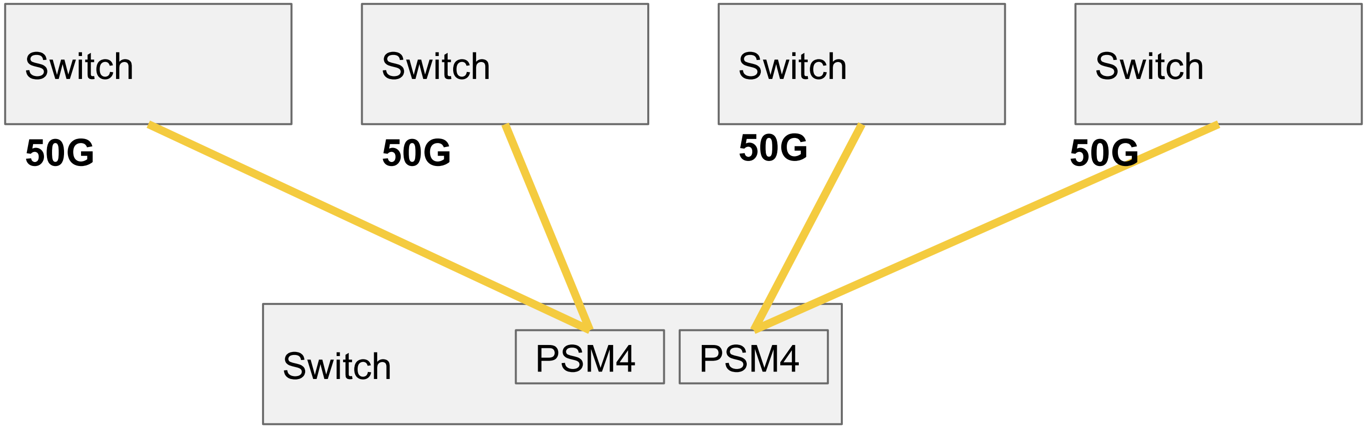
Our problem statement:

- Distance 150-1000m from “core network room” to ToRs in adjacent data hall or across the parking lot
- Leverage a 32-QSFP-port switch for more than 32 logical links
- Build with SMF now to leverage WDM-based 100G+ bandwidths in the future
- 160Gbps “Steady state” bandwidth from ToR upstream

Breaking out in the PSM4/100G world

What we liked about PSM4:

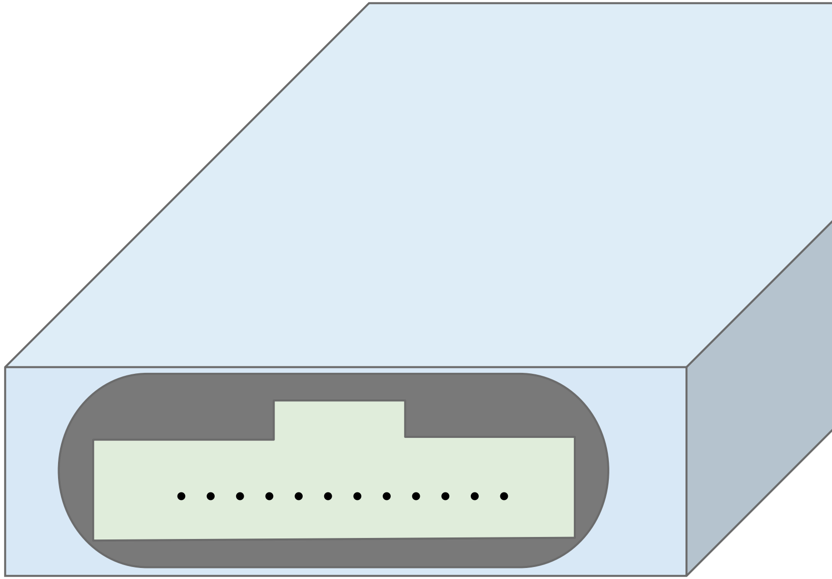
200G aggregate bandwidth cost less than 160G using 4 x 40G optics*



* Total port capacity on supported switches affected the formula, as did distance requirements.
* YMMV

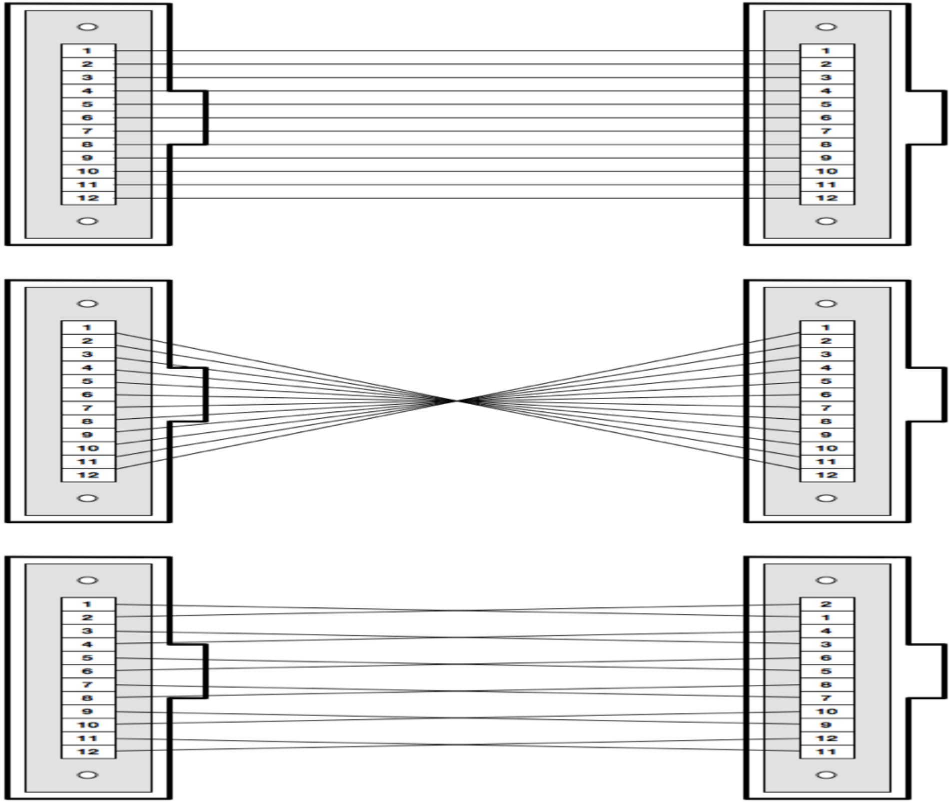
Tx/RX Positions of PSM4

- MPO Connector
  - When looking at module optical port:
  - Positions 1,2,3,4 TX lanes 1,2,3,4
  - Positions 9,10,11,12 RX lanes 4,3,2,1



Tx →	1	2	3	4					4	3	2	1	← Rx
	●	●	●	●	●	●	●	●	●	●	●	●	
Pos:	1	2	3	4	5	6	7	8	9	10	11	12	

PSM4 Cabling: MTP Polarity Methods

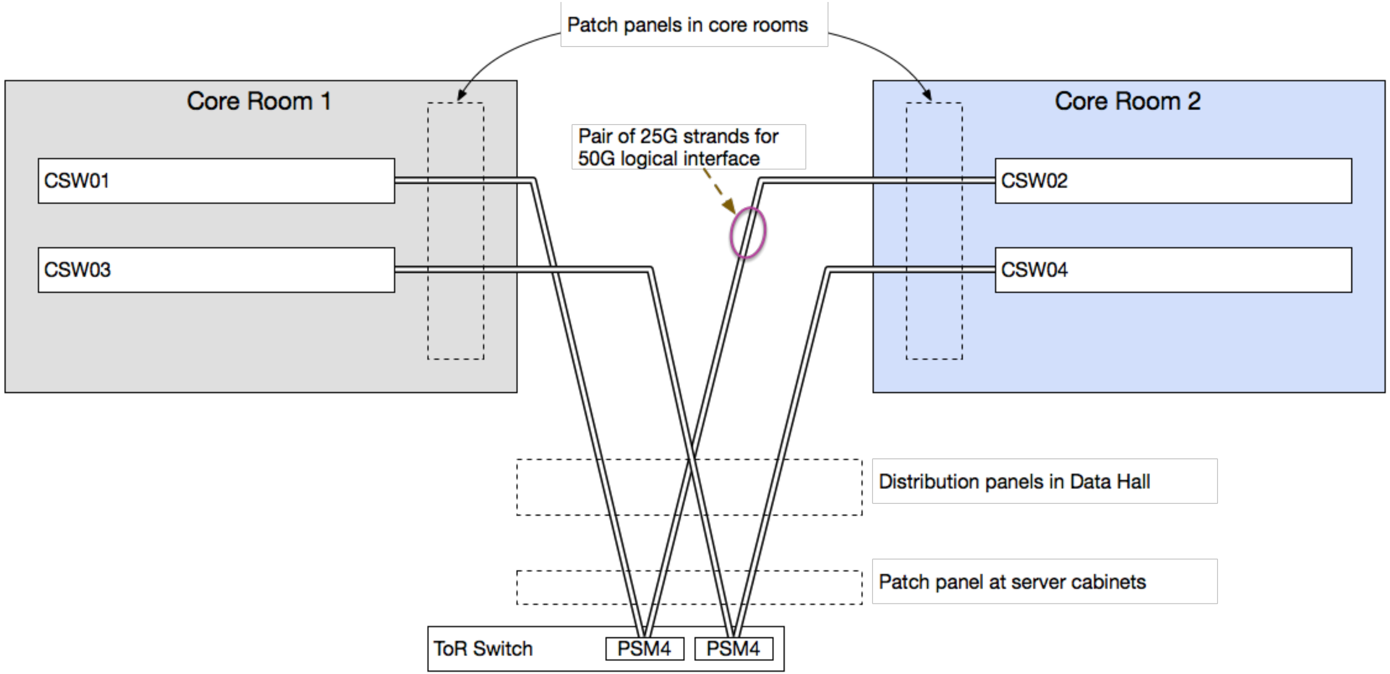


A: Straight Thru

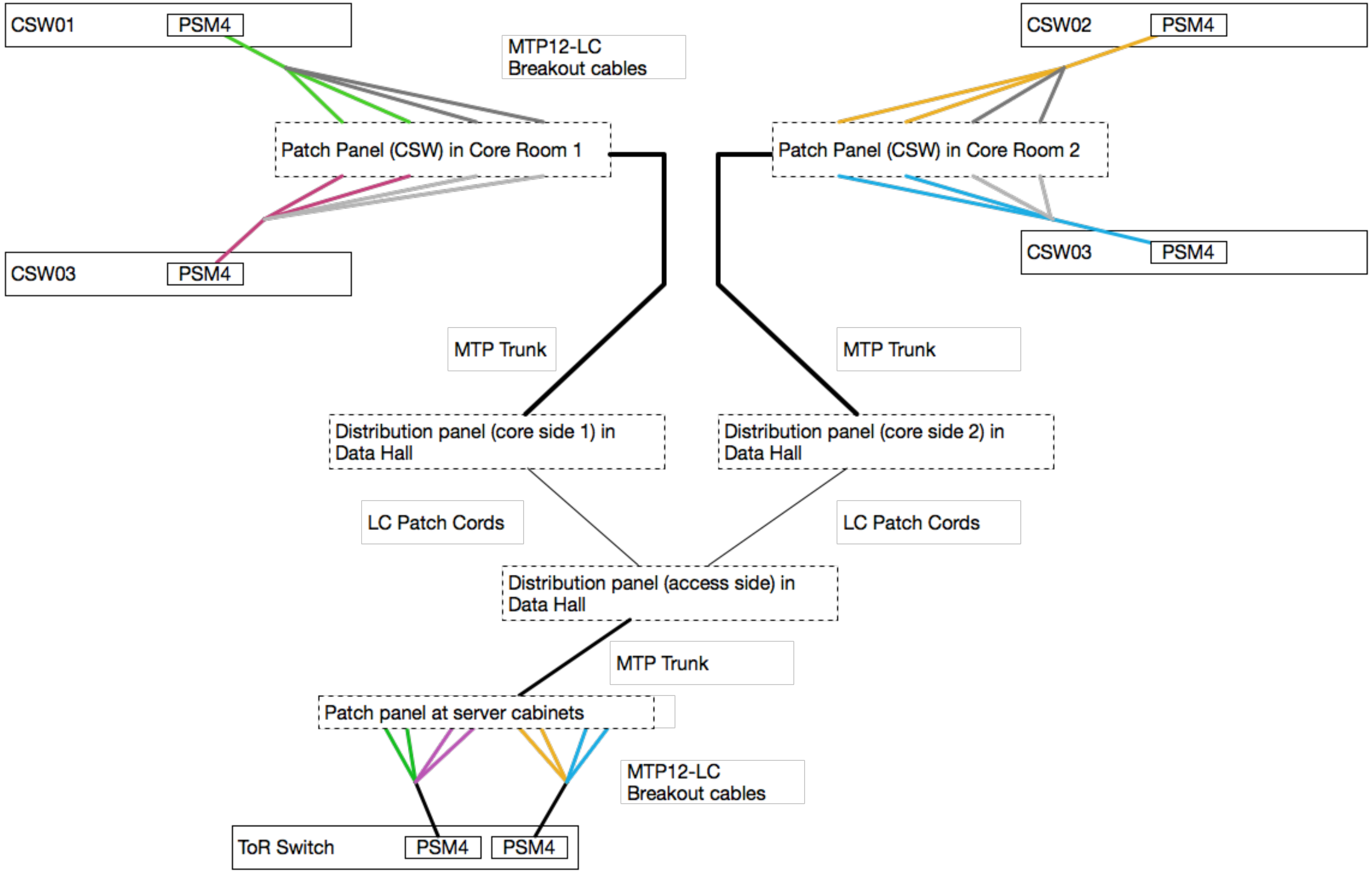
B: Rollover

C: Flipped Pairs

Cabling for 50G at LinkedIn - Patching Challenge

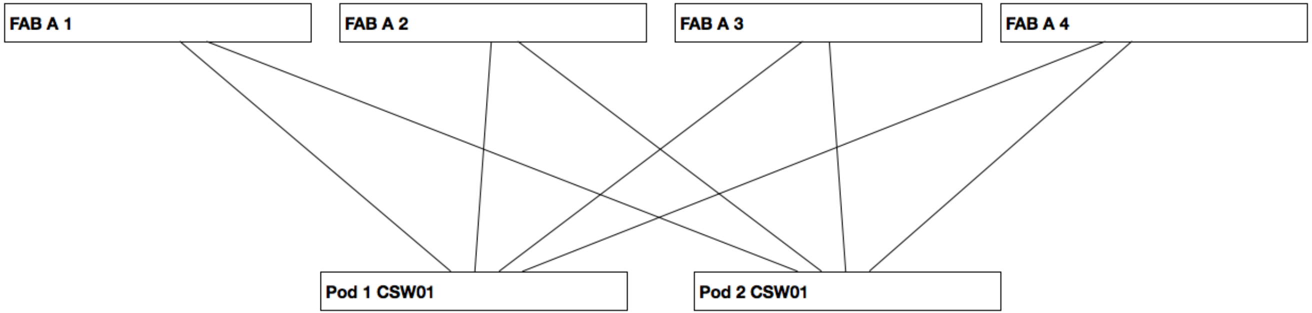

Patch panels in core rooms
Core Room 1
CSW01
CSW03
Pair of 25G strands for 50G logical interface
Core Room 2
CSW02
CSW04
Distribution panels in Data Hall
Patch panel at server cabinets
ToR Switch
PSM4
PSM4

Cabling for 50G at LinkedIn - Patching Solution

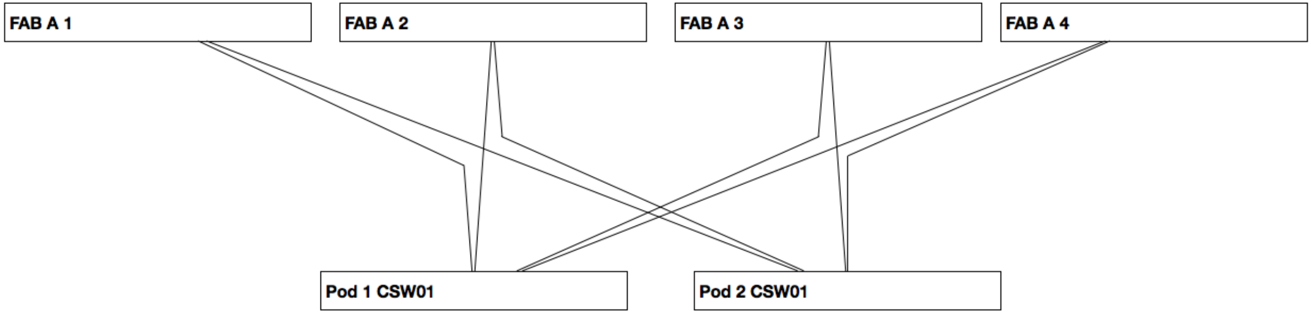

CSW01
PSM4
MTP12-LC
Breakout cables
CSW02
PSM4
Patch Panel (CSW) in Core Room 1
Patch Panel (CSW) in Core Room 2
CSW03
PSM4
CSW03
PSM4
MTP Trunk
MTP Trunk
Distribution panel (core side 1) in Data Hall
Distribution panel (core side 2) in Data Hall
LC Patch Cords
LC Patch Cords
Distribution panel (access side) in Data Hall
MTP Trunk
Patch panel at server cabinets
MTP12-LC
Breakout cables
ToR Switch
PSM4
PSM4

Cabling for 50G at LinkedIn - Fabric Cabling Challenge

Sample of FAB (Fabric) switches to Core (CSW) switches - **Logical**



Look! Wiring can be achieved by Y cables placed in this pattern:

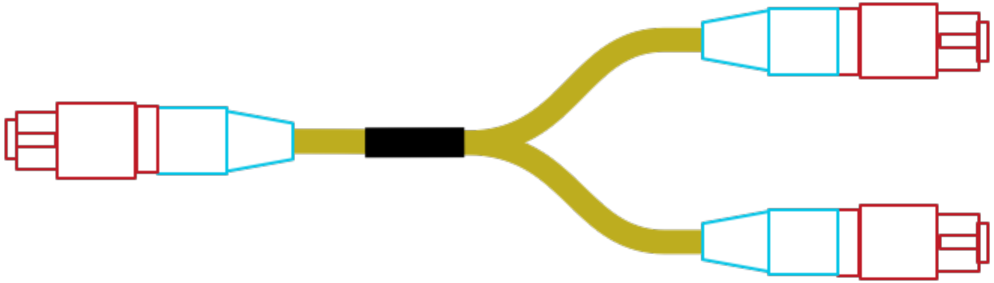


PSM4 Cabling for 50G at LinkedIn - The Y cable

We contracted the design and manufacture of a Y cable to split MTP to 2x50G

PSM4 Optic Side

POSITION	COLOR
1.	BLUE
2.	ORANGE
3.	GREEN
4.	BROWN
5.	GRAY
6.	WHITE
7.	RED
8.	BLACK
9.	YELLOW
10.	VIOLET
11.	ROSE
12.	AQUA



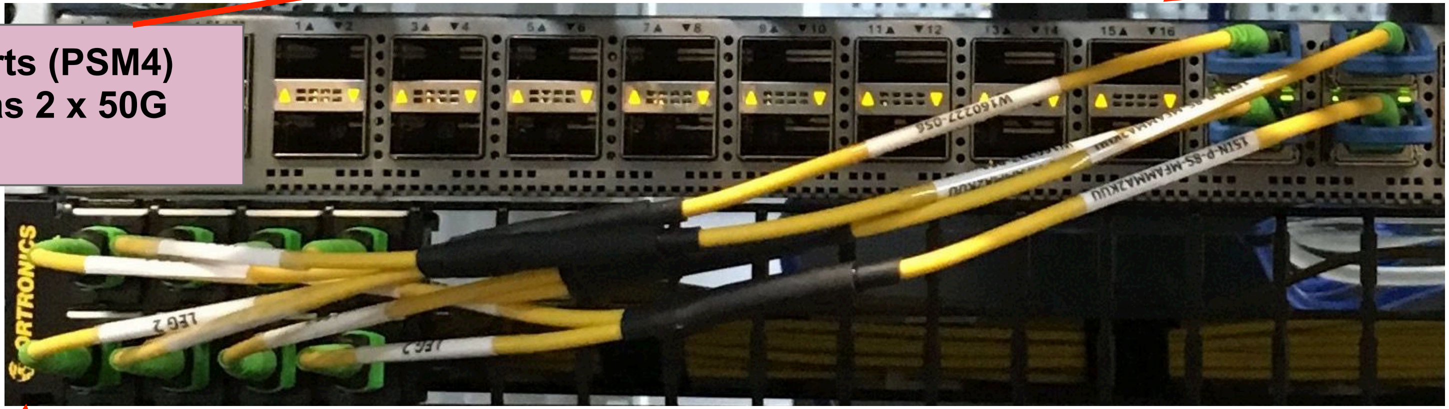
Patch Ports Side

POSITION	COLOR
1.	GREEN
2.	VIOLET
3.	BROWN
4.	YELLOW

POSITION	COLOR
1.	BLUE
2.	AQUA
3.	ORANGE
4.	ROSE

PSM4 Cabling for 50G at LinkedIn - The Y cable

Here’s how it turned out.



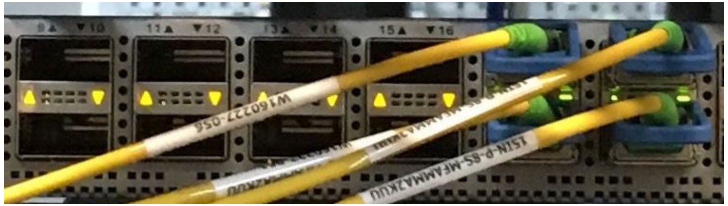
4 x 100G ports (PSM4) configured as 2 x 50G each

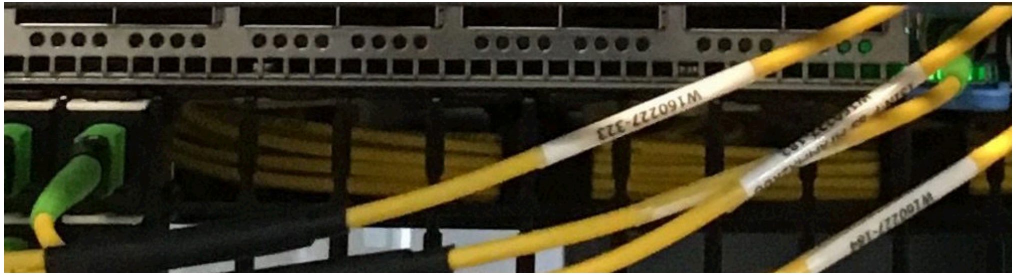
8 x MTP-12 ports carrying 2 x 25 lanes each

One status light per 100G port. Do you see a problem with this configuration?

Other onsite twist for PSM4: Status Lights

How many lights do you see?





I see FOUR LIGHTS!

Considerations for PSM4: Platform support

```
# show interface Ethernet1/2/1
Ethernet1/2/1 is up
admin state is up, Dedicated Interface
  Hardware: 50000 Ethernet, address: 00f2.8b9f.35db
  MTU 9100 bytes, BW 50000000 Kbit, DLY 10 usec
  reliability 255/255, txload 1/255, rxload 1/255
  Encapsulation ARPA, medium is broadcast
  full-duplex, 50 Gb/s, media type is 100G
```

Does your switch see all the lanes?

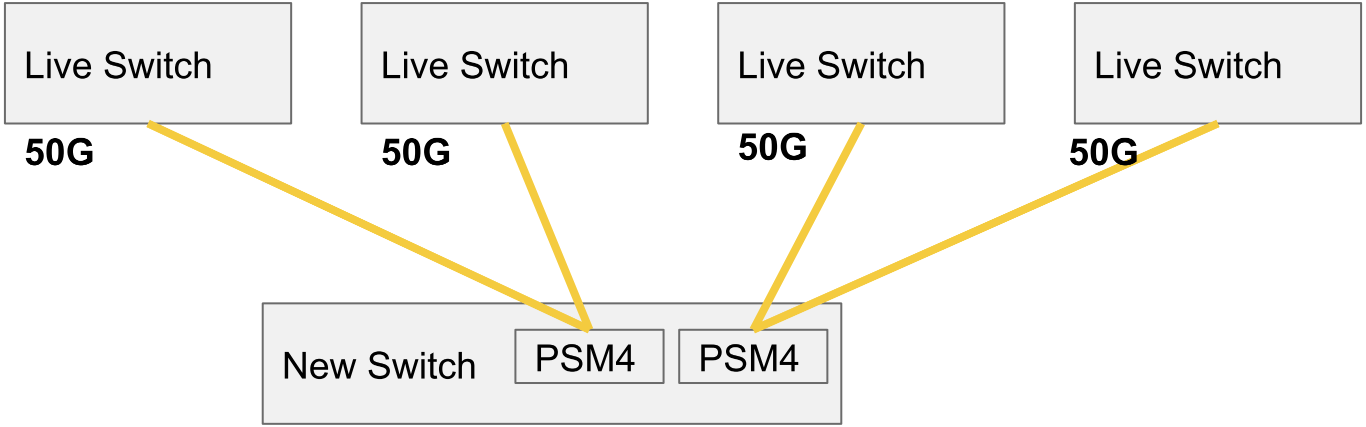
```
# show interface Ethernet1/2/1 transceiver de
Ethernet1/2/1
    transceiver is present
    type is QSFP-100G-PSM4
    ...
    nominal bitrate is 25500 MBit/sec per channel
    Link length supported for 9/125um fiber is 2 km

Lane Number:1 Network Lane
           SFP Detail Diagnostics Information (internal calibration)
  ----------------------------------------------------------------------------
                Current              Alarms                  Warnings
                Measurement     High         Low         High          Low
  ----------------------------------------------------------------------------
  Temperature    0.17 C        75.00 C     -5.00 C     73.00 C       -3.00 C
  Voltage        0.01 V   --    3.63 V      2.97 V      3.46 V        3.13 V
  Current       24.72 mA       65.00 mA    15.00 mA    60.00 mA      20.00 mA
  Tx Power       6.01 dBm ++    2.99 dBm  -11.42 dBm    1.99 dBm     -9.43 dBm
  Rx Power     -25.22 dBm --    2.99 dBm  -14.68 dBm    1.99 dBm    -12.67 dBm
  Transmit Fault Count = 0
  ----------------------------------------------------------------------------
  Note: ++  high-alarm; +  high-warning; --  low-alarm; -  low-warning

  *** SFP diagnostics data may be invalid!  ***
```

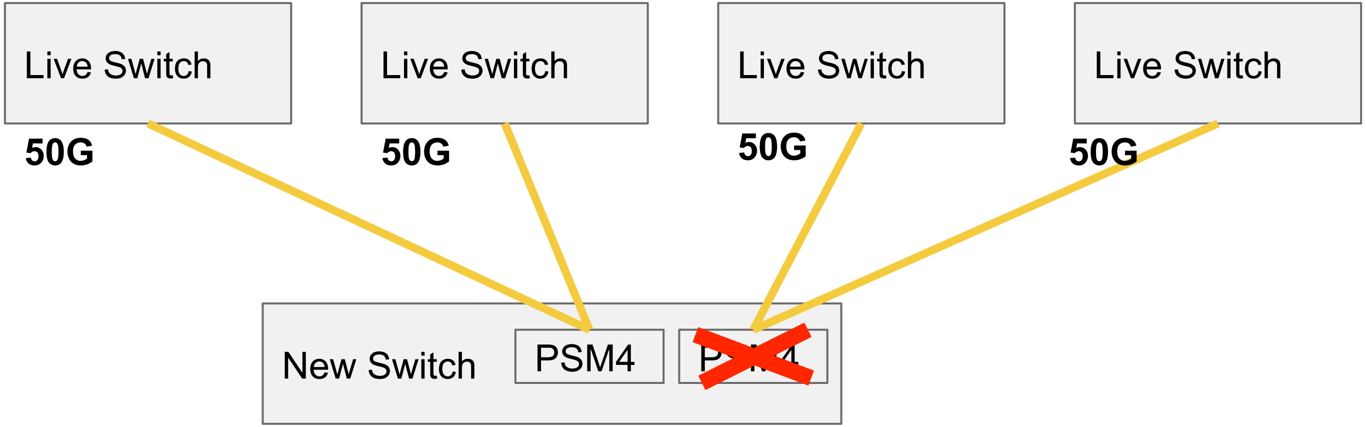
!

Considerations for Breakout: ZTP on the front panel



If the front panel ports (forwarding interfaces) are being used by the new switch for initial provisioning, how does it know to configure itself for 50G ports?

Considerations for Breakout: Capacity Impact



When using a breakout on a single optic for capacity-critical interfaces, consider a single fault's impact to aggregate bandwidth.

Summary

* Consider 25-50-100G Ethernet if you’re doing 40G now

* PSM4 offers an inexpensive 25-50-100G interface for a SMF environment

* Check for platform support for selected optic and interoperability

Lessons for us and for all:

* If this is your first dive into SMF intra-datacenter cabling, review best practices

* Budget time for design, testing, redesign, finalizing network design with datacenter team, finalizing again, (redesign again if needed), testing one more time, and patience with onsite datacenter staff during the initial build

* Test your optics, test your cables

* Plan your cable routing, plan your failure domains

Additional References

25G Ethernet Consortium: http://25gethernet.org/

PSM4 Consortium, Specifications: www.psm4.org

Latest Trends in Optical Interconnects by Christian Urricariet, Finisar at NANOG66: https://www.nanog.org/meetings/abstract?id=2754

Detailed PHYs of 100G optics (2013) including 10x10 and 4x25: http://www.ieee802.org/3/bm/public/jan13/petrilla_03a_0113_optx.pdf

LinkedIn datacenter network: https://engineering.linkedin.com/blog/topic/100G-50G

Questions?

Thank You!

www.linkedin.com/in/paulzugnoni

More from LinkedIn Engineering:
https://engineering.linkedin.com/blog/topic/datacenter

Yuval Bachar

Brad Peterson

Shawn Zandi

Jacob Rose