Next Generation Datacenter Interfaces: Optics and Form Factors

Data Center Optical Interconnection Technology Summit
Optical Communication Technology Development Forum
5 September 2019
Shenzhen, China

Chris Cole

CIOE
中国国际光电博览会
CHINA INTERNATIONAL OPTOELECTRONIC EXPO

FINISAR®

Outline

- **Datacenter Optics Rates**
- Pluggable Form Factors
- Coherent in the Datacenter

Datacom (Ethernet) Gb/s Data Rates vs Time

Time	Datacom (Ethernet) Gb/s MAC Rates					Rate X
1990's - 2006	0.1	1	10			10

FINISAR

Datacom (Ethernet) Gb/s Data Rates vs Time

Time	Datacom (Ethernet) Gb/s MAC Rates					Rate X
1990's - 2006	0.1	1	10			10
2006 - 2007	0.1	1	10	100		10

FINISAR

40Gb/s vs. 100Gb/s IEEE Debate

- 100Gb/s pro arguments
  - 10x is the conventional rate step, minimizing deployment cost by minimizing number of rate steps
  - 25GBaud technology (100G = 4x25G NRZ) investment focus will lead to lower cost in the long-term
- 40Gb/s pro arguments
  - 10GBaud technology (40G = 4x10G NRZ) is mature, ready for low-cost, low-risk, high-volume deployment
  - 40G has nearly 3x radix vs. 100G for 1.28T switch ASIC
    - 100Gb/s: 12x
    - 40Gb/s: 32x
  - Server I/O step after 10Gb/s
- Both rates were adopted by the IEEE, after 40G was identified as important for Datacenter applications

Datacom (Ethernet) Gb/s Data Rates vs Time

<table>
<tr><th>Time</th><th colspan="6">Datacom (Ethernet) Gb/s MAC Rates</th><th>Rate
X</th></tr>
<tr><td>1990’s - 2006</td><td>0.1</td><td>1</td><td>10</td><td colspan="2"></td><td></td><td>10</td></tr>
<tr><td>2006 - 2007</td><td>0.1</td><td>1</td><td>10</td><td colspan="2">100</td><td></td><td>10</td></tr>
<tr><td>2008 - 2013</td><td colspan="2">1</td><td>10</td><td>40</td><td>100</td><td></td><td>4</td></tr>
</table>

Datacom (Ethernet) Gb/s Data Rates vs Time

<table>
<tr><th>Time</th><th colspan="7">Datacom (Ethernet) Gb/s MAC Rates</th><th>Rate X</th></tr>
<tr><td>1990’s - 2006</td><td>0.1</td><td>1</td><td colspan="2">10</td><td colspan="2"></td><td></td><td>10</td></tr>
<tr><td>2006 - 2007</td><td>0.1</td><td>1</td><td colspan="2">10</td><td colspan="2">100</td><td></td><td>10</td></tr>
<tr><td>2008 - 2013</td><td colspan="2">1</td><td colspan="2">10</td><td>40</td><td>100</td><td></td><td>4</td></tr>
<tr><td>2014 - 2015</td><td colspan="2">1</td><td>10</td><td>25</td><td>40</td><td>100</td><td>400</td><td>4</td></tr>
</table>

200Gb/s vs. 400Gb/s IEEE Debate

- 400Gb/s pro arguments
  - 4x is the new <u>conventional</u> rate step, minimizing deployment cost by minimizing number of rate steps
  - 50GBaud technology (400G = 4x100G PAM4) investment focus will lead to lower cost in the long-term
- 200Gb/s pro arguments
  - 25GBaud technology (200G = 4x50G PAM4) is mature, ready for low-cost, low-risk, high-volume deployment
  - 200G has 2x radix vs. 400G for 12.8T switch ASIC
    - 400Gb/s: 32x
    - 200Gb/s: 64x (or for 100Gb/s: 128x)
  - Server I/O step after 100Gb/s
- Both rates were adopted by the IEEE, after 200G was identified as important for Mobile applications in China

Datacom (Ethernet) Gb/s Data Rates vs Time

Time	Datacom (Ethernet) Gb/s MAC Rates								Rate X
1990's - 2006	0.1	1	10						10
2006 - 2007	0.1	1	10		100				10
2008 - 2013	1		10		40	100			4
2014 - 2015	1		10	25	40	100	400		4
2016 to today	2.5	5	10	25	40 50	100	200	400	2

FINISAR

The Big Four Plans - 2019

- AWS
  400G-DR4 broken out to four 100G-DR
- Google
  Shifting from 100G to 200G in the form of 2x200G modules. 2x400G will be their next step.
- Facebook
  New high-density 100G switch fabric for 4X capacity.
  Next step 200G.
- Microsoft
  Will deploy 400G inside data centers after 400ZR available to interconnect regional data centers

No clear plans to deploy true 400GbE for some time!

LightCounting High-Speed Ethernet Optics Report – April 2019 – page 12

Next High Volume Ethernet Data Rates

- Huge industry investment to support 400GbE as the next high volume Datacom rate will not see ROI for many years
- 1st Gen 400GbE optics will have small volume, primarily in telecom applications
- 200GbE is the next high volume Datacom rate
- Commonly used characterization of 200GbE as an "interim" step to 400GbE is meaningless
  - 200GbE is an "interim" step to 400GbE, just like 40GbE was an "interim" step to 100GbE
- 400GbE will be high volume when following is mature:
  - 100Gb/s lane SerDes
  - 7nm CMOS PHYs
  - Sufficient bandwidth TX to generate open PAM4 eyes

What's After 400Gb/s Ethernet?

- Future rates prediction based on 2x rate increases:

10	25	40/50	100	200	400	800	1600

- Broad industry consensus that 800G is the next step
- Are we falling into the same conventional thinking trap?
- Could there be finer Ethernet rate increments than 2x?
- Transport no longer follows conventional fixed rate steps:

  Transport per λ rates:

  100 → 200 → 300 → 400 → 500 → 600 → 800
- Ethernet not likely to follow; the overhead is not worth it
- However, FlexEthernet could start to introduce sub-rating into the Datacenter

Outline

- Datacenter Optics Rates
- **Pluggable Form Factors**
- Coherent in the Datacenter

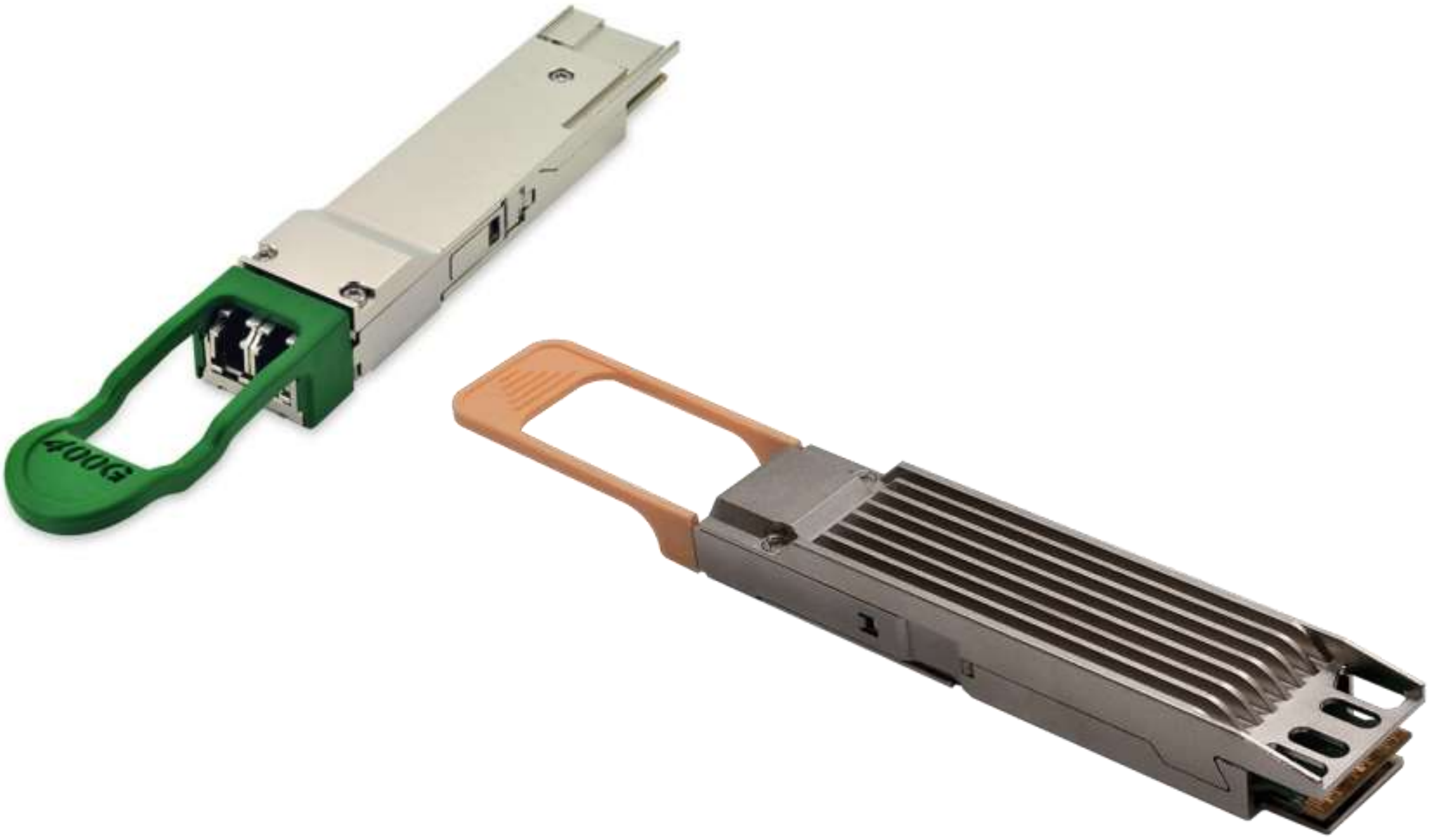
FINISAR

Mainstream Pluggable Form Factor Evolution

I/O Count	10G I/O	25G I/O	50G I/O	100G I/O
Single Dual	SFP+	SFP28	SFP56	SFP112
			SFP-DD56	SFP-DD112
			DSFP	DSFP
Quad Octal	QSFP+	QSFP28	QSFP56	QSFP112
			QSFP-DD56	QSFP-DD112
			OSFP	OSFP
Ten to Hex	CFP	CFP2	CFP8	CFP8

Other pluggable form factors: CXP, uQSFP, DSFP-DD

QSFP-DD & OSFP Form Factors

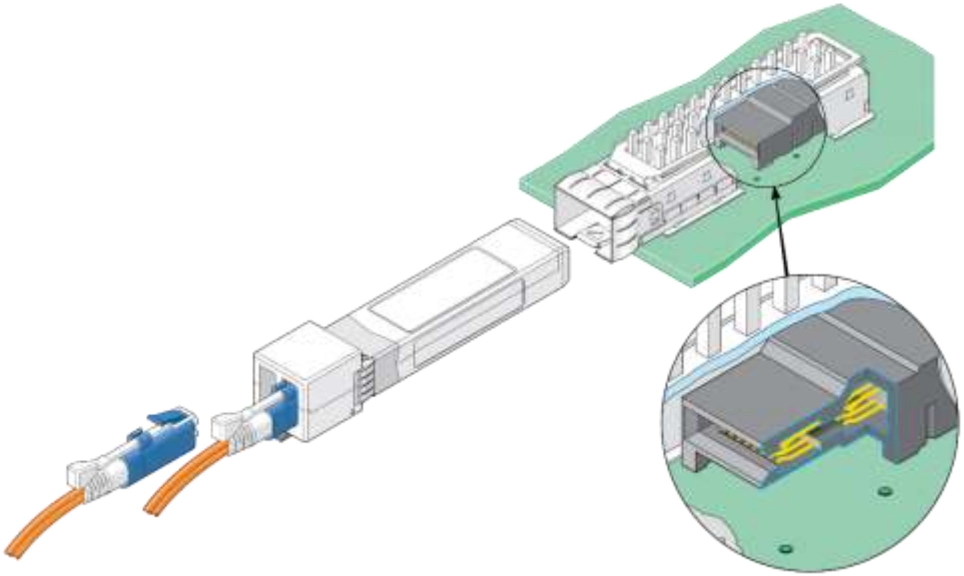


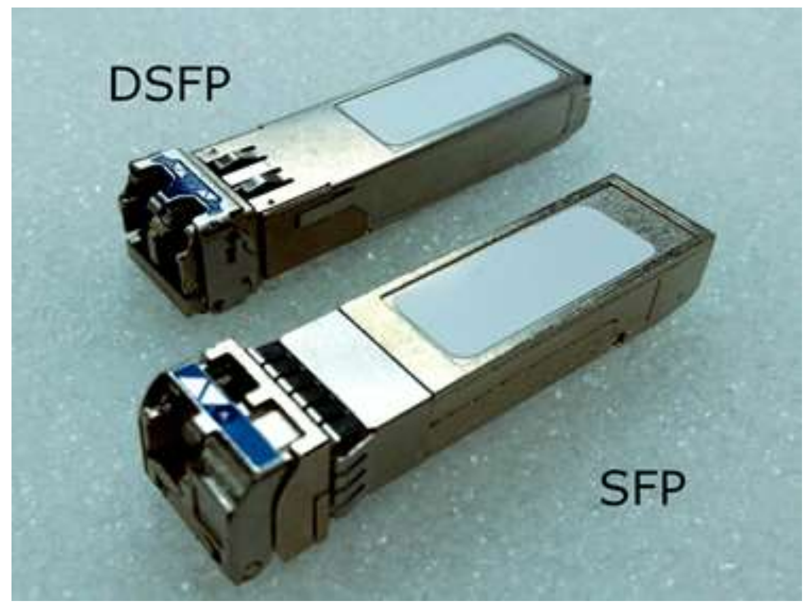
FINISAR

QSFP-DD vs. OSFP Comparison

Category	QSFP-DD	OSFP	Comments
Compatibility	**QSFP+, QSFP28**	none	
1 RU Front Ports	**36x**	32x	36x OSFP is marginally possible
Connector	Double row (76 contacts)	**Single row (60 contacts)**	OSFP connector is QSFP28 style
Signal Integrity (worst host lines)	28GBaud (56 PAM4)	**56 GBaud (112 PAM4)**	DD overfly leads degrade S.I.
Thermal interface power density	2x	**1x**	DD top surface roughness, flatness specs. are ~2x harder
Heat dissipation	35mm outside	**inside**	DD has similar thermal management issues as CXP
Hear Sink Configs.	ridding	ridding, **integral**	Integral sink has no temp. drop at module at interface
Internal volume	1x	**2x**	OSFP enables larger components
Cost	>1x	**<1x**	Connector, top surface, and internal volume drive cost

SFP-DD & DSFP Form Factors





SFP-DD vs. DSFP Comparison

Category	SFP-DD	DSFP	Comments
SW Compatibility	**SFP+, SFP28**	OSFP	
HW Compatibility	SFP+, SFP28	SFP+, SFP28	DSFP requires additional host circuits to support SFP+, SFP28
Control I/O	SFP+, SFP28	OSFP	see above
1 RU Front Ports	48x	48x	
Connector	Double row (40 contacts)	**Single row (22 contacts)**	DSFP connector is SFP28 style
Signal Integrity (worst host lines)	28GBaud (56 PAM4)	**56 GBaud (112 PAM4)**	DD overfly leads degrade S.I.
Host card depth	>>SFP+	**SFP+**	DD has double row connector (mobile and NIC issue)
Hear Sink Configs.	ridding	ridding	
Cost	>1x	**<1x**	Connector drives cost

Pluggable Available Technology Configurations

Switch BW Tb/s	Optical Rate Gb/s	Port Count	Port rows	Ports/ row	I/O Rate Gb/s	I/O Pin Count
1.28	40	32	2	16	10	512
3.2	100	32	2	16	25	512
12.8	100 200	128 64	2 2	64 32	50	1024
25.6	200	128	4	32	50	2048
25.6	200 400	128 64	2 2	64 32	100	1024
51.2	400	128	4	32	100	2048

Pluggable Form Factors Discussion

- Pluggable paradigm is viable for 12.8T, 25.6T and 51.2T Switch nodes using available technology
- This is at the cost of increasing SerDes power
- Possible new technologies that could extend the pluggable paradigm to 102.4T Switch node:
  - Low-cost flyover miniature copper cables
  - High-density Hex pluggable connector
  - Low-power 200G/lane SerDes
- For when the pluggable paradigm finally runs out of gas, optics industry is investigating new paradigms:
  - High-density on-board optics
  - Co-packaged optics w/ promise of 20-30% power savings
- There is no consensus on how and when this will happen

Outline

- Datacenter Optics Rates
- Pluggable Form Factors
- **Coherent in the Datacenter**

IMDD vs. Coherent in the Datacenter

- 10G/λ Transport: IMDD
  (Intensity Modulation Direct Detection)
- 40G/λ Transport: IMDD and Coherent
- 100G/λ and above Transport, ≥80km links: Coherent
- 200G/λ ≥40km links: Coherent
- 400G/λ ≥25km links: Coherent
- Coherent advantages over IMDD:
  - Chromatic Dispersion (CD) and Polarization Mode Dispersion (PMD) compensation because of signal amplitude and phase recovery followed by DSP
  - Higher SNR because of RX front end LO mixing
- Conventional thinking is that Coherent will soon replace IMDD for links inside the datacenter

Datacenter Link Limits

- Longest internal link distance: 1km
- Example CWDM4 λs 1km SMF Spec Limits
- L0 λ: 1271nm (1264.5 to 1277.5nm span)
  λ_{min} = 1264.5nm and $\lambda_{zero\_dispersion\_max}$ = 1324nm:
  - CD = -6 ps/nm
  - PMD = 0.5 ps
  - Loss = 0.47dB
- L3 λ: 1331nm (1324.5 to 1337.5nm span)
  λ_{max} = 1337.5nm and $\lambda_{zero\_dispersion\_min}$ = 1304nm:
  - CD = 3 ps/nm
  - PMD = 0.5 ps
  - Loss = 0.43dB
- These values do not require compensation for IMDD links

SNR Comparison of IMDD vs. Coherent

Application	Direct Detection SNR NRZ, PAM4		SNR Compare	Coherent SNR QPSK, QAM16	
	Implementation			Implementation	
	TX	RX		TX	RX
4dB typical datacenter link budget	EML, DML single λ or TFF, PLC WDM	PIN single λ or TFF, PLC WDM	>>	SiP	SiP
Laser AOP constrained	single λ SiP	single λ SiP	>>	SiP	SiP
	WDM SiP	WDM SiP	≈	SiP	SiP
Transport	Any	Any	<<	SiP	SiP

C. Cole, "Direct Detection vs. Coherent SNR Inside the Datacenter," Will Coherent Optics Become a Reality for Intra-data Center Applications? Workshop, OFC 2019, San Diego, CA, 3 March 2019.

IMDD & Coherent in the Datacenter Discussion

- Conventional IMDD has better SNR than SiPIC Coherent for typical datacenter links
- Conventional thinking is not based on link analysis
- Coherent dispersion compensation processing is unnecessary and offers no advantages for these links
- Choice of IMDD or Coherent for the datacenter should only be based on specific implementation trade-offs
- Coherent maybe required inside the datacenter for high loss links (>10dB), for example those that include passive components like optical switches
- Outside the datacenter, for reaches >20km, Coherent advantages dominate

Next Generation Datacenter Interfaces

Thank You