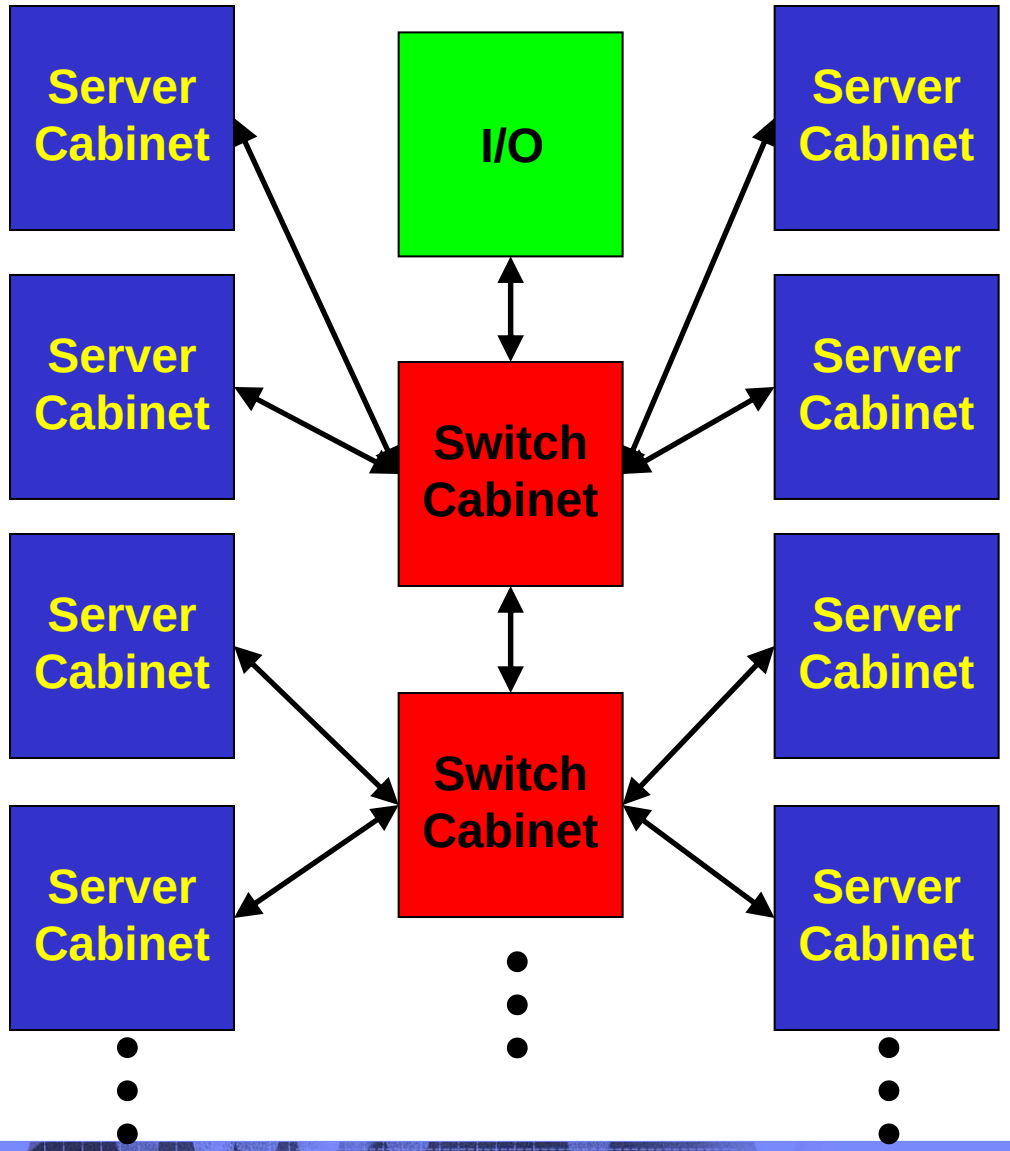
Market Potential and Technical Feasibility of 12 Channel Parallel Optical Interconnects for 802.3 HSSG

Petar Pepeljugoski, IBM
David Cunningham, Avago Technologies
Kenneth Jackson, Emcore

Optics for High Performance Server Cluster Architectures

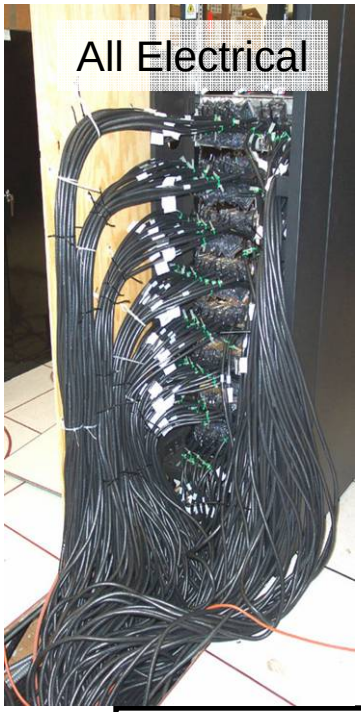


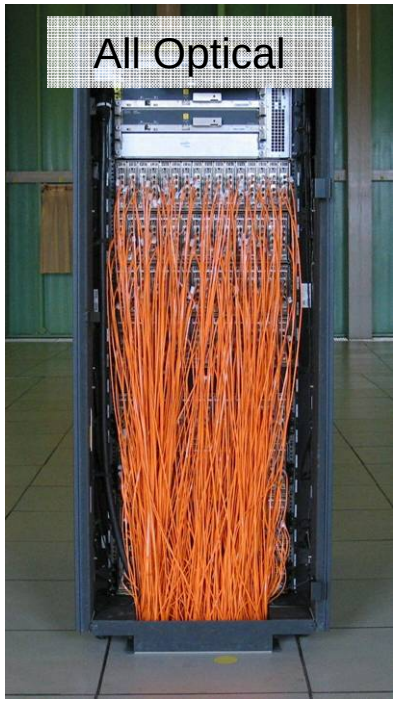
- Example is 8 cabinets and 2 switches
  - Real systems are 10-100s of cabinets and several racks of switches
- Rack-to-rack connections moving to parallel optics
  - **Distances <100m**
  - Optics will increasingly replace copper at shorter and shorter distances
- Trend will accelerate as bitrates (in the media) increase and costs come down
  - 2.5Gb/s → 5 Gb/s → 10Gb/s
  - Relative cost seen dropping from 10-12x to 1-1.2x* (full duplex link)

* Measure is cost per Gb/s relative to short wavelength 4 Gb/s SFP per Gb/s

Optics Today in Super Computers: 4X and 12X Infiniband

IBM ASCI Purple Server (LLNL)
75.7 TeraFLOP/s, ~3000 parallel links 12+12@2.5Gb/s/ch

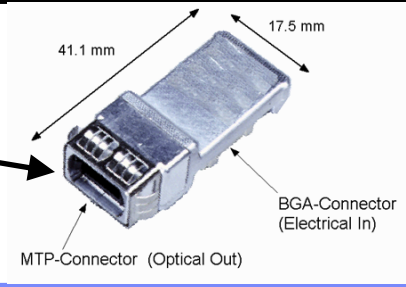




IBM Federation Switch
- Copper (bulk, bend and weight)
- Optical – very organized

- Next generation optical links will have 60, then 120 Gb/s
  - 12X DDR IB in 2-3 years, QDR 4-5 years

Snap 12 module
12 Tx or Rx at 2.5Gbd



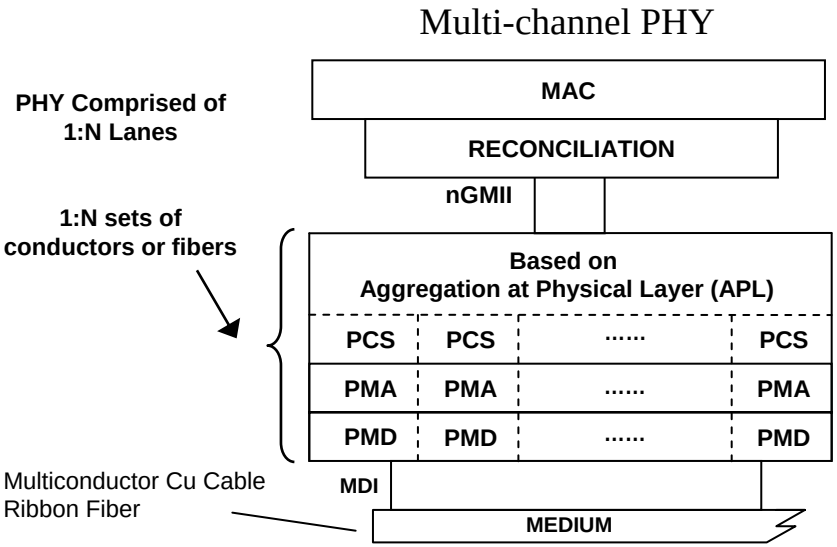
Parallel Optics in Super Computers

- #4, #5 fastest supercomputers use parallel optical interconnects.
- Future systems may have >10,000 processors & tens of thousands of fibers for each machine
- Many of IBM's other 235 machines in the top 500 list also use high-BW optics, for cluster and storage networks

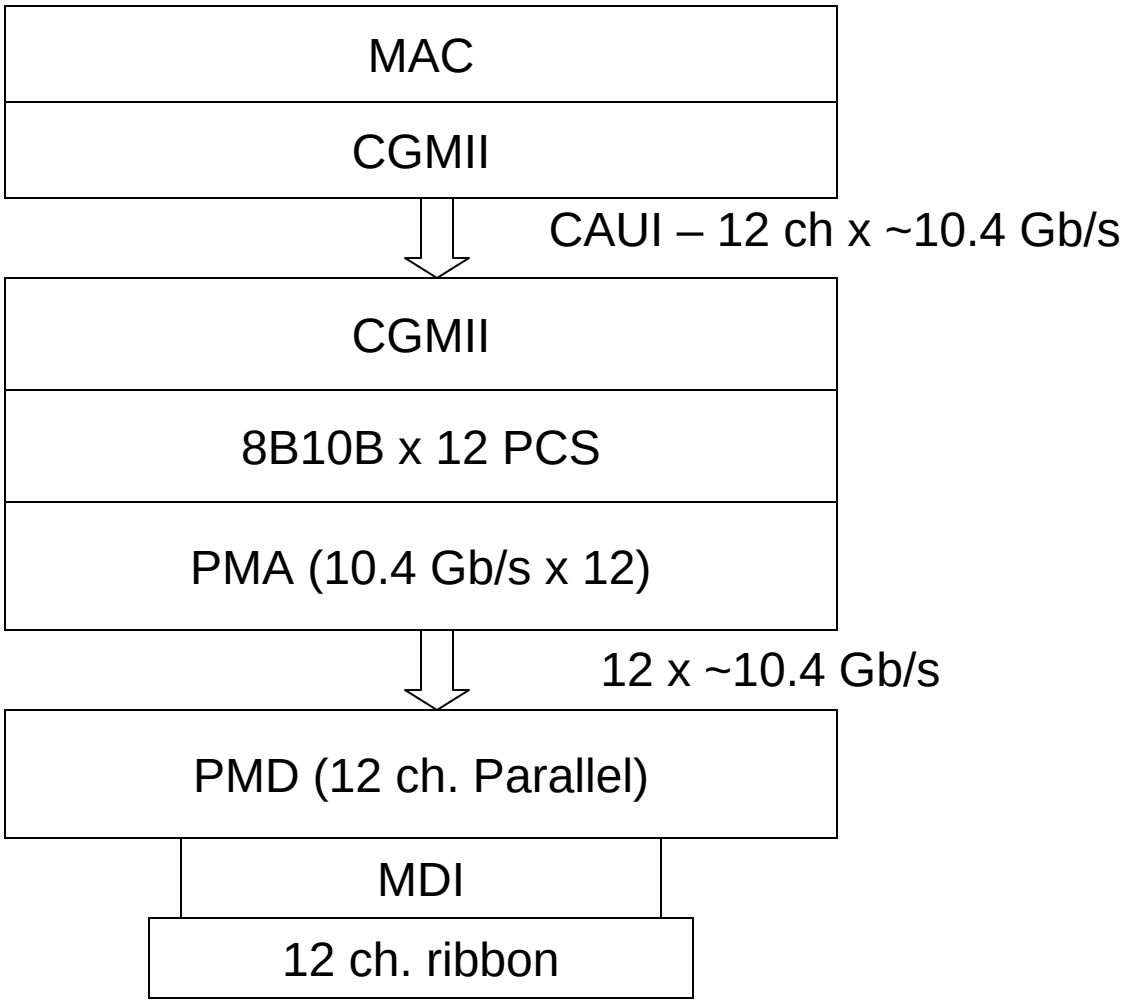
- Example: Mare Nostrum System (Barcelona Supercomputing Center) uses ~5000 Multimode Fiber Links
  - Combination of VCSEL/MMF links: serial for node/switch & 4x (POP-4) for inter-switch
- Future supercomputers will have much higher number of parallel optical interconnects

Options to Create a Fatter Pipe Using Aggregation at Physical Layer (APL)

- Multi-channel PHY
  - Multi-core ribbon fiber – how wide
- Multi-wavelength (WDM) PHY
  - N wavelengths on single fiber pair
- Multi-wavelength (DWDM) system
  - Single wavelength per module
  - External optical MUX/DEMUX



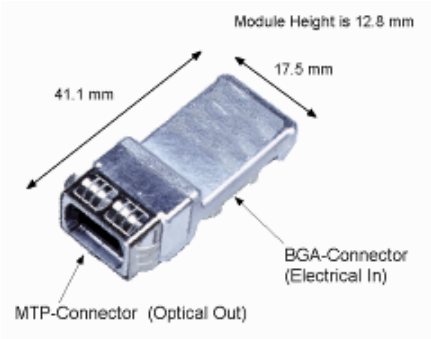
Architecture based on 12 channel PMD



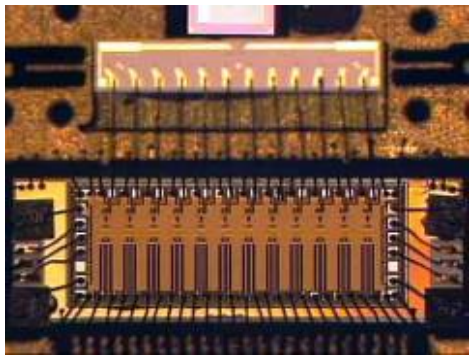
Why 12 Channel 8b/10b for <100m links?

- Straightforward extension of XAUI to CAUI (3x wider, 3.333x faster) at 100 Gbps
- Use 12-lane wide connectors and cables based on standardized MT ferrule
  - Manufacturers have experience with 12 lane drivers, lasers, detectors, amps
  - 10 channel interface would waste 2 lanes in many existing OM3 parallel cables
- 8B/10B gives low overhead de-skew, prior experience with XAUI implementation
- 8B/10B provides guaranteed run-length
  - Better signal integrity at the same data rate than 64B/66B
  - More robust error and simpler detection and correction
- Commonality with 12x-QDR InfiniBand
  - Increased volumes for better commercial viability

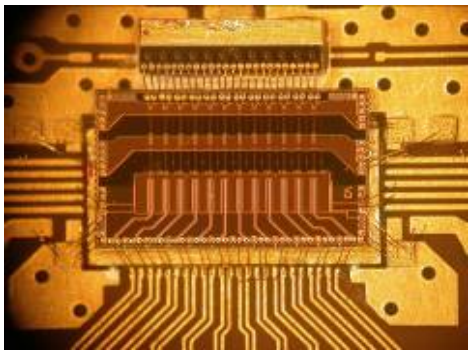
12 Channel Parallel Link for 100 GbE



- Extension to 12x10 Gb/s demonstrated in 2003*
  - Based on Picolight commercial 3Gbit/sec module
  - IBM Res designed SiGe LDD, Infineon SiGe Rx
  - 10Gbit/sec lasers from Picolight
  - IBM redesigned electrical flex packaging and mega-array connector
- 12 channels at BER $<10^{-12}$, link length 316m using OM3 fiber
  - The most cost effective solution is for lengths shorter than 100m
- Inexpensive 12 channels, 9.9-11.0 Gb/s/channel tester designed and built
- Compatibility with QDR Infiniband – avoid form-factor/module type proliferation

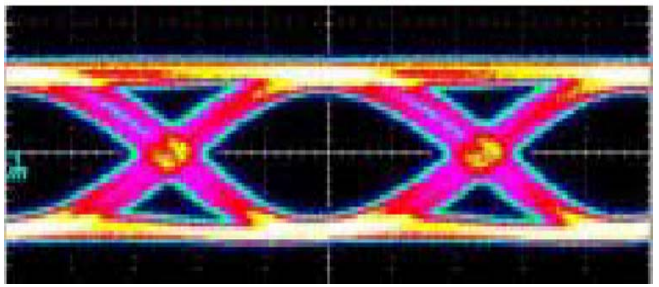


12 ch. TX OE chips



12 ch. RX OE chips

DC-wander minimized for 8B/10B



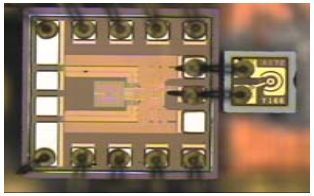
Eye diagram at 10Gbit/sec

*Kuchta et al., "120-Gb/s VCSEL-Based Parallel-Optical Interconnect and Custom 120-Gb/s Testing Station, IEEE Journal of Lightwave Technology, vol.22, no.4, 2004

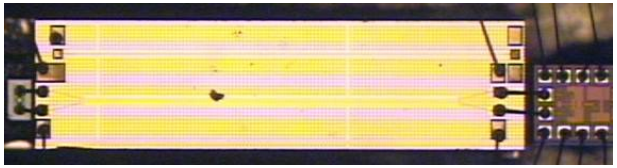
Experimental 17 Gb/s Link

- CMOS driver and CMOS receiver integrated circuit demonstrated
  - 0.13 μm CMOS 8RF-LM technology, 250x350μm
  - Power consumption 100mW for TX, 120 mW for RX at 17 Gb/s
- Performance limited by RX bandwidth

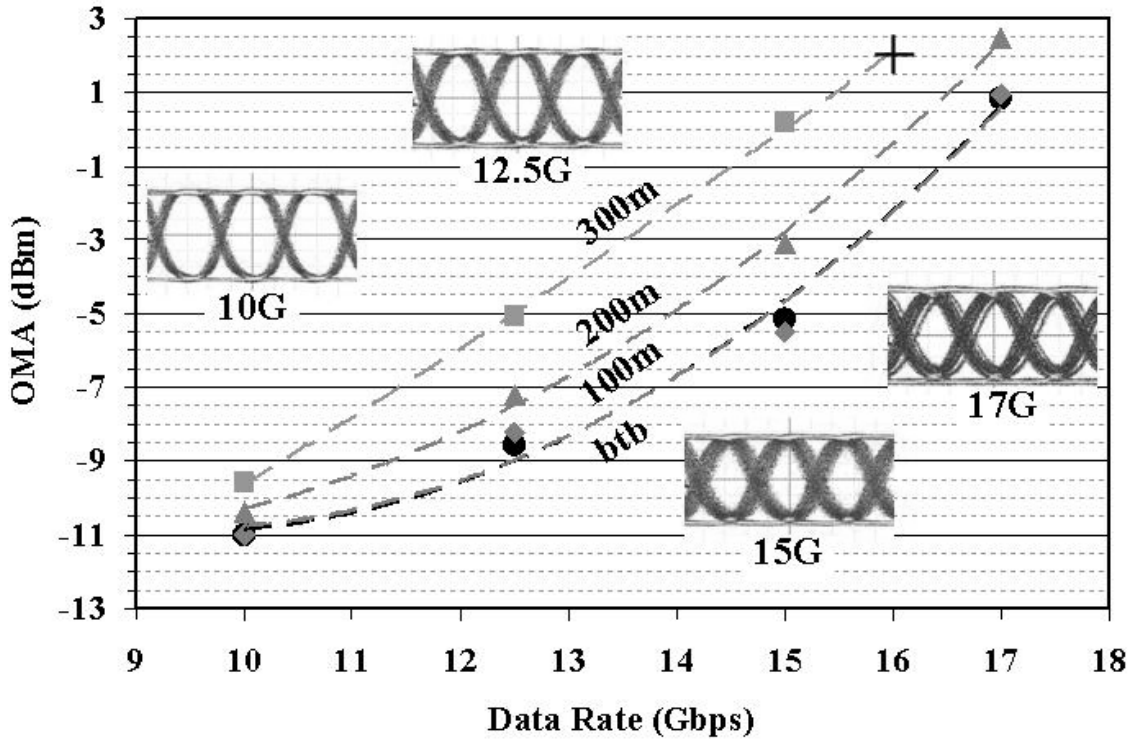
TX



RX



O.Liboiron-Ladouceur et al., LEOS 2006 Annual meeting



Conclusions

- There is a broad market potential for short optical interconnects based on parallel optics in HPC environment
- Technical feasibility demonstrated to implement cost effective 12 channel wide PMD over OM3 for 100 Gb Ethernet
  - >10Gb/s per lane is possible and will offer power and density advantages, but at higher cost
- Parallel optical link compatible with QDR Infiniband is the most cost effective option for DataCenter/HPC environment
  - avoids form-factor/module type proliferation