Datacenter Parallel Optics

Datacenter upgrade cycle impact on interconnects and server/switches
- 10GBASE-T and Direct Attach copper cables
- Active Optical Cables
- 40G/100G SR, LR4/LR10
- Notes from Interop Las Vegas and IEEE Silicon photonics events

Embedded Optical Modules (Mid-board optics)

Presented by Brad Smith
Brad@LightCounting.com
Datacenter Trends 1

- **New Intel/AMD servers**: offering 10-14 cores/MPU up from 2-4 cores
  - 2 MPUs per server to 2-4 MPU .... & Multiple GPU support

- **Server virtualization**: increasing the utilization of servers from 20% to 90%
  - Moving big computing tasks around the network stressing the interconnect infrastructure

- **Next generation I/O**: PCIe Gen3.0 doubling I/O bandwidth uplinks from servers
  - 5G/lane 8b/10b ..to.... 8G/lane 64/66 with QuickPath

- **Collapsing aggregation layers**: from 3 to 2 layers to reduce latency & complexity

- Increased use of video and increased bandwidth demand from iPhones, iPads etc.

- **All this adds up to more interconnect traffic and need for higher bandwidth!**
Datacenter Trends 2

- **Servers to Top-of-Rack switches uplinks upgrading** from 1G to 10G
  - With 10GBASE-T, SFP+ DAC and Optical transceivers

- **Top-of-Rack switch uplinks upgrading** from 10G SFP+ to 40G (4x10G)
  - To handle 24-48 ports @ 10G = need about 160G-360G

- Datacenters are getting bigger and longer reaches needed for 300m-600m reaches; LR4 transceivers and single mode fibers
New Intel Servers Start Shipping  
(Finally!)

• Intel announced availability of their next generation architecture / microprocessor called “Romley” in March (Intel’s Xeon)  
  – After 6-9 month delay  
  – IBM, HP, Dell, Cisco all announced availability of new servers (2H/2012)  
  – 80% performance improvement over earlier Xeon versions; large power savings  
  – 10 cores/MPU; lower power consumption, PCI Express Gen3 I/O bus faster I/O

• Delay put many hardware purchases “on hold” for 6+ months  
  – “Wait for new servers – don’t buy old gear” = WAIT!  
  – “Want older models but wait until new ones announced so can buy old ones at big discount” = WAIT!  
  – “Don’t need 10GbE SFP+ and 10GBASE-T for old servers” = WAIT!  
  – Net result: new purchases are delayed – until now

• Bottom line: More virtualization processes; more cores; more DRAM capabilities, will push more traffic in/out of servers and upwards into the switching infrastructure and drive interconnect migration to faster interconnects.  
  – 10GbE server uplinks with optical transceivers, DAC, and 10GBASE-T  
  – 40G and 100G switch uplinks for both SR and LR4

• Will kick the data center server and switch upgrade cycle into full gear for 2H/2012
Intel Server Microarchitecture Roadmap

LightCounting’s 10GBASE-T Report:  www.LightCounting.com/reports.cfm
Server and Switch Uplinks

• Upgradable Modular Interconnect Interfaces
  – Example: HP FexLOM, Dell Force10
  – Removable interconnect modules for servers and switches
  – Enables upgrading the uplink interfaces to higher data rates and different MSAs without having to change out the entire switch or server

  – Today: Quad 10GBASE-T; Quad SFP+; Dual QSFP;
  – Future path to CFP/2, CXP, CFP/2, for 100G and 25G and perhaps 40GBASE-T

  – Usually server/switch vendor specific and not traditional PCI Express bus line cards
Dell Force10 Modular Interconnects

- CFP/2? (100G)
- 2 QSFP (80G)
- 4 SFP+ (40G)
- 4 10GBASE-T (40G)
Direct Access Cables

- Mainly used for North-South connections linking servers to Top-of-Rack switches
- DAC or “Twinax” cables continue to see solid market adoption
  - Started appearing in 2008; took off in 2010/2011
  - Primarily playing on market delays in 10GBASE-T implementations (lower power and costs)
- Active DA Cables enable 15 meters at the same cost as 7 meter cables
  - Thinner cables: Use 30 gauge copper instead of 20 gauge to offset chip cost
  - Enables some East-West connections but still not to end-of-row
- Number of problems surfacing:
  - “Linear” and Non-Limiting” port issues on servers and switches surfacing
    - They don’t mix well on the same cable
  - Flood of “marginal quality” cables fueled by Internet Data Centers in Asia
  - Industry not sure about active DAC yet
- Reach keeps shrinking with increasing data rates
  - 7 meters at 10G; 3 meters at 14G (InfiniBand FDR)
  - 25G will likely need active elements to achieve needed reaches
  - Eventually, a cross over pricing with optical occurs
- Bottom line: Likely to remain a popular choice in data centers for 10G server to switch uplinks until 10GBASE-T takes off late 2013/2014
10GBASE-T

- Adoption continues to pick up as chip vendors release newer 40-nm versions with improved feature sets; Still lackluster unit volumes
  - *Reach auto sense*: mechanism to reduce power for short links
  - *Energy Efficient Ethernet (EEE)*: to turn off power consuming circuits when not in use

- Power consumption is still an issue ~3-3.5W/port
  - Likely to drop to ~2W/port with 28-nm implementation
  - New daughter cards have PC like heat pipes for cooling!

- Interop – May- Las Vegas
  - Saw a lot of 10GBASE-T server daughter cards and top-of-rack switch announcements;
  - Still no server LAN-on-motherboard announcement (LOM)
10GBASE-T

• **IEEE considering 40GBASE-T** (More news at July meeting)
  – 4x10G in a single chip or package
  – Lots or rumors and discussions in the market
  – Could be of interest with next generation Intel “Romley+1” architecture with PCI Express Gen4 and perhaps 40G server uplinks

• **10GBASE-T supplier news:**
  – **PLX Technology** is in play to be sold to IDT (June 15th)
    • Cash and stock deal worth about $330 million
  – **Aquantia** raised $35M series-F VC funding totaling $132 M

• **Bottom line:**
  – Data centers are slowly moving to 10G at the server
  – At 28-nm, 10GBASE-T will likely show massive adoption ~2014
  – Likely to severely depress DAC sales
  – The only 100 meter alternative to optical transceivers .....but only 10G
40G Solutions: SR, LR4

**40G QSFP SR** (100m, MMF, 4x10G)
- Very important with new 10G/40G Top-of-Rack switches becoming popular (2H 2012)
- Will be popular for linking Top-of-Rack switches to End-of-Row switches (<100m)
- QSFP headed to be then next “SFP+” replacement in ToR switches
- Number of 40G QSFP suppliers is rapidly swelling – lots of recent announcements
  - Interop: saw a lot of switch systems with 40G uplink modules
- Barrier to building are now fairly low with transceiver assembly kits
- Lots of quality issues and junk transceivers surfacing
- Is headed for an ugly price war (ASPs range from $65 - $275)

**40G QSFP/CFP LR4** (10Km, SMF, 4x10G)
- Used to connect racks and rows to the far side of big datacenters (400-600m reach); for campuses and data center links to the metro loop
- QSFP will be the most popular in data centers; CFP in telecom
  - QSFP smallest size
- **Key feature**: SR and LR4 in the same QSFP socket (unlike CXP/CFP)
- **Upcoming report on 40G/100G for SR and LR4 in the data center (July)**
100G Solutions: LR4/LR10

Next generation MSAs for 100G with 25G signaling

- Power consumption of components and cooling issues and of course prices are the key issues
- Not clear what components will fit the power budget of the CFP/2 and CFP/4 or if alternative MSAs are needed
- New gearbox chips out based on CMOS boasting much lower power at 2.5Ws (Broadcom)
- Second generation lasers are out at lower power and costs (Finisar)

- IEEE meeting in July to resolve many open issues for 300-1,200 meter solutions
  - Reduced features for non-telecom applications
  - Decide best reach lengths
  - What components are in or on the system board, etc..
  - Retimers, CDRs, equalization, etc. issues
  - Packaging issues are still open
    - CFP/2, CFP/4, zCXP, zQSFP, QSFP28, CXP2, QSFP Extended, Enhanced QSFP++.........?

Slowly gaining market traction - but prices still too high
Datacenter needs reach of 400m-600m
Optical link budget is equivalent to 2Km but called “10Km”
Needs to move out of CFP to CFP/2

Upcoming report on 40G/100G for SR and LR4 in the data center (July)
Active Optical Cables (AOCs)

• **InfiniBand** is still the big market with supercomputers
  – 10G QDR transitioning to 14G FDR
  – EDR 26G likely 2015
  – 14G FDR Direct Attach copper now only 3 meters reach!

• **Ethernet AOCs** gaining significant traction in datacenters
  – Driven by low acquisition cost for short reach (<30-50m) where cabling access is easy
  – Driven by low operational cost - No optical-related field returns and connector cleaning
  • Optics are enclosed therefore no cleaning operating expense, re-installation errors, etc.
  • System vendors do not have field returns of optical line cards that when tested have return tag with: “No problem found-dirty optical connectors”

• **MSAs**: 4-channel QSFP dominates; 12-channel CXP is “still trying”
  – Enormous cost differential between 4x and 12x
  – CXP seeing strong interest from large switch and router companies especially in Asia

• **Starting to see Photoshop versions of SAS 12G AOC using Mini-SAS HD connector!**

• **First Thunderbolt AOCs** (Sumitomo) and USB 3.0 AOCs (Via and Hitachi Cable)

Embedded Optical Modules (EOMs)

- EOMs are parallel optic links that attach directly to PCBs
- Next generation SNAP12 and POP4 (circa 1999)
- Provide high port density; low power; better cooling; mid-board mounting
- Key feature:
  - Better signal integrity at 25G+
  - EMI/ESD immunity – especially important at 25G
- Used to link big chip “islands” to other chips -or- to another chassis 10-30m away!
  - Changes the system design game!
- Very early stages of an emerging transceiver segment
- Avago microPOD & miniPOD with IBM HPC blade shown at trade shows sparked interest from the high-speed, core switch and routers and server segments
- Different market from front panel, pluggable transceivers in many ways

LightCounting’s EOM Report:  http://www.lightcounting.com/reports.cfm
EOMs are Different From Transceivers

- Simplified optical transceiver but:
  - Without MPU & EPROM for system vendor lock-out codes
  - No standard electrical and optical connectors
  - No standard form factor design or MSAs
  - May snap in or screw mount to PCB
  - Much smaller than pluggables
  - Designed for 1-10m reaches not 100m as are front panel transceivers
  - Not designed for interoperability with other vendors EOMs

- New designs have faster data rates, smaller size, simplex and duplex, low power consumption

- *Business, sales and support dynamics are very different from selling pluggable front-panel transceivers*
Where Are EOMs Used?

EOMs are used in various applications such as ASICs, transceivers, and EOMs, which are essential components in Datacom, Telecom, CATV, FTTX, and Consumer markets. The following diagram illustrates the use of EOMs in these applications:

- **ASIC**: These are used as the primary components in electronic systems.
- **Transceiver**: They are used to facilitate the transmission of data between different devices.
- **EOM**: These are essential for the efficient transfer of data in complex electronic systems.

Key features mentioned in the diagram include:

- **Large number of parallel signal traces**: This indicates the complexity and volume of data that EOMs handle.
- **Switch ASIC**: This component is crucial for routing data within an electronic system.
- **Long PCB trace to transceiver**: This highlights the physical components and their connections in a circuit board.

These components and their roles are integral to the functionality of modern electronic systems, especially in high-speed interconnects.
40G-56G - Its Game Over for Copper!

- 10-12-inches at 10G ..........4-6-inches at 25G ............< 2-inches at 56G.

To illustrate this situation, the Optical Internetworking Forum (OIF) announced May 18th, 2012 that they are forming several 56G electrical interface projects:

- “Ultra Short Reach Electrical Interface project will define a link of less than 10 mm between an ASIC and an optical engine (OE) at data rates up to and including 56 Gbps. Such an interface will support the development of multi-chip modules (MCMs) and alternative packaging schemes optimized for minimum power consumption.”

- “Close Proximity Electrical Interface project will define a link with a reach of less than 50 mm from chip to discrete OE at data rates up to and including 56 Gbps. This effort aims to facilitate an efficient board mounted OE at low power.”

- This clearly shows where things are headed longer term
Where Are EOMs Used

Demo board from US Conec and Avago
Three Generations of EOMs

1ST Generation EOM (SNAP12)

2nd Generation EOMs (microPODs)
- Avago
- UC Conec
- US Conec Prism LightTurn Connector

3rd Generation EOMs
- Avago
- NEC, Reflex, Silicon-Line, Samtec, TE Connectivity, Finisar
- Emcore/Sumitomo
Products, Prototypes, Trade Show Demos and Photoshop Product Concepts
Sony AOC/EOM

CXN2006, CXN2007 and CXN2300 optical interconnection modules

- Designed for 4K x2K video for HDTVs and digital cameras
- 4x10G based on VCSELs with a lens-less internal design
- 10mm x 16mm

EOM Summary

• 3rd generation EOM market is just beginning; Lots of unknowns
• Late 2013: 25G VCSELs in will be a key enabler for EOMs
  – Most protocols and interfaces converge on ~100G with ~25G signaling
• ~2015: EOMs will play an important role at 25Gs but will compete strongly against traditional electronic solutions with new PCB materials and compensating electronics
  – Retimers, CDRs, Pre-emphasis, Regeneration electronics
• ~2017: At 40G and 56G, EOMs will play a dominant role and include silicon photonics solutions and integrated ASICs and transceivers in multi-chip modules and/or integrated as 3DICs

• Over time, 25G and EOM technologies will ripple down the data center infrastructure as each protocol moves to 16G-25G and higher

• Compensating electronics will always compete head on but...at what total cost and power consumption?

LightCounting’s EOM Report:  http://www.lightcounting.com/reports.cfm
Xilinx/Luxterta Announcement

3D-IC Co-Packaged FPGA and Optical Transceiver

• Leading FPGA vendor announced integrating Luxtera’s 4x28G silicon photonics chip into the same chip package as the FPGA
• Uses copper pillars and flip-chip technologies to connect the 28-nm FPGA to the Luxtera transceiver chip
• Enables a programmable single chip with:
  – 72 electrical transceivers at 13.1G (pins)
  – 8 simplex optical interconnects at 28G each (fibers)
• Targeted towards telecom and core switch and routers with 100G and 400G internal core systems

• **Bottom line:** Now at the top of the food chain but shows what is in store for the technology as prices declines and the technology improves
Silicon Photonics

• Interest in silicon photonics is rapidly escalating due to:
  – Cisco purchasing Lightwire $271M – Why so much?
  – Very few 25G VCSEL demos or announcements at OFC/2012 March – What’s next at 40G-56G? Is it the end of VCSELs?
  – Embedded Optical Modules and Xilinx/Luxtera announcement

• IEEE Santa Fe (May) event had about 3-4 times number attendees and presenters as last year
  – High-speed modulation (40G+) and integration are key interests
  – How are we as an industry going to deal with the faster line rates needed to support the exploding traffic demands?

• OPSIS – University of Delaware
  – Silicon photonics fab for companies to run test wafers on using Luxtera tool kit, libraries

• Silicon photonics may find an entry point at 25G modulation (100G/400G)
LightCounting Q&A

Submit a question in the Q/A window to the right of your screen

Research Reports

Storage Networking Report (September 2014): $5,000 (USD)
With v1 Storage, users only accessed the entire LUN (LARGE): Dead or alive: platform storage on the horizon. As storage is changing so fast, it’s hard to keep up with the latest developments. However, a new and innovative approach to storage, known as “active storage”, is emerging. This approach aims to make storage more efficient and scalable, allowing organizations to store and manage data more effectively. To find out more, refer to the Storage Networking Report.

Market Update Report (September 2015): $5,000 (USD)
Sales of optical components and modules continue to increase, with the market growing at a rate of 20%. The demand for optical components and modules is driven by the growing demand for high-speed interconnects in data centers and telecom networks. According to LightCounting, the market for optical components and modules is expected to grow at a CAGR of 25% over the next five years.

Research Databases

LightTrends Newsletters

http://www.lightcounting.com/reports.cfm
Contact: Renee@LightCounting.com at 408.962.4851