40G Ethernet Blog

Comments on new developments in 40G Ethernet

Archive for the ‘Technology’ Category

The XLPPI and using it with QSFP+ modules

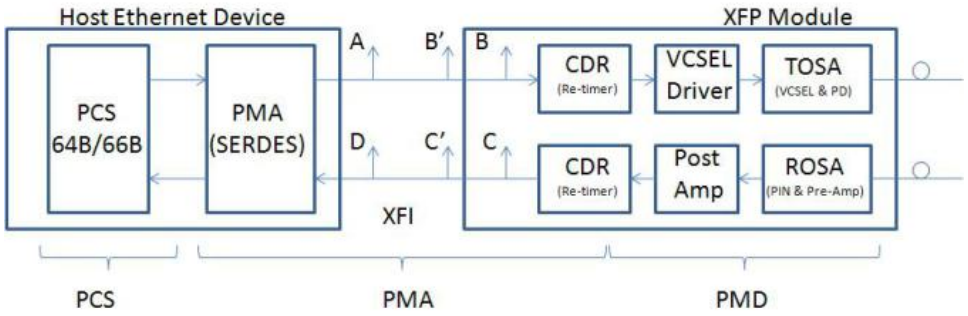
January 24, 2011
Overview of the XLPPI and using it with QSFP+ modules

In the initial development of the 100/40G Ethernet 802.3ba specification the intended interface to the optical module were the XLAUI and CAUI, which were derived from the 10Gbps XFI used on XFP transceivers. As with the XFI, both the XLAUI and CAUI require that the module perform retiming on both transmit and receive data. This retiming function is found in CFP modules but not in the popular QSFP+ module, which means that the XLAUI is not compatible with the quad 10Gbps interface of the QSFP+ module. To address the incompatibility between the XLAUI and the QSFP+ module the final version of the 802.3ba includes the description of the optional XLPPI (and CPPI) which allows direct connection to optics without the necessity of a re-timer function. This blog will look at the XLAUI and XLPPI with respect to their 10G predecessors (the XFI and SFI) and how the XLPPI is used with QSFP+ modules.

Background on the XFP/XFI and SFP+/SFI

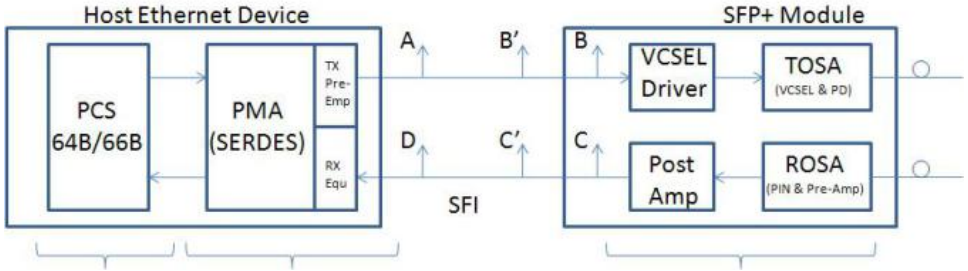
The XFP was the first 10Gbps pluggable module with a serial interface called the XFI and in order to enable the connecting ASIC’s to support the 10Gbps serial XFI the XFP module supported a PMA function for retiming of both the transmit and receive data. By introducing the retiming function into the XFP, the signal eye at the test points B’ /B and C/C’ were optimised for the benefit of the

host PMA SERDES at the expense of the cost and size of the XFP. The diagram below shows the functional blocks of the XFP module, the location of the test points and how it connects to a Ethernet host.



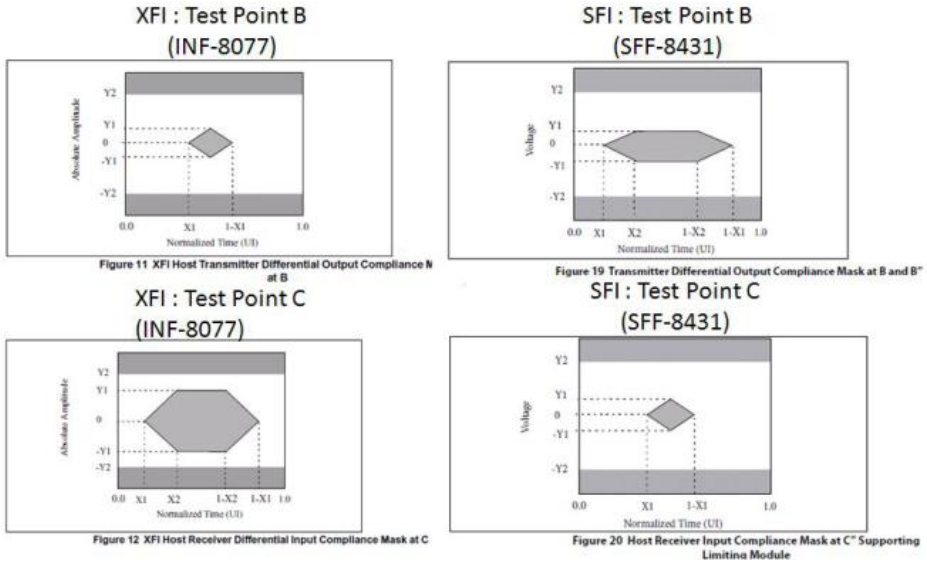
(http://40gethernet.files.wordpress.com/2011/01/xfp-and-xfi.jpg)

The next step in 10G module evolution was to enhance the SFP module (used for 1Gbps Ethernet and 1-4Gbps FC) to support 10Gbps thus creating the SFP+. The SFP+ module does not have internal CDR retiming which means the high speed electrical characteristics are very different to the XFI and as such a new interface was created, the SFI. The SFI only species the electrical requirements at point B'/B' and C/C', the PMA function in the Ethernet host to both operate with these signals and to compensate for any additional distortion created by the PCB traces between the ASIC and the SFP+ module.



(http://40gethernet.files.wordpress.com/2011/01/sfp-and-sfi.jpg)

In the transmit direction the SFI requirement for the signal eye at B' are much greater than those for the XFI, thus the host PMA must implement pre-emphasis to compensate for the signal distortion due to the PCB traces. On the receive side the SFI output eye at point C is not only smaller than with the XFI but the incoming jitter is higher, in addition a new element needs to be taken care of which is the pulse width shrinkage due to pattern dependency (DDPWS). In order to reliably recover the data from the SFI the host PMA needs to implement signal equalization and for some applications (like 10GBase-LRM) EDC is required.

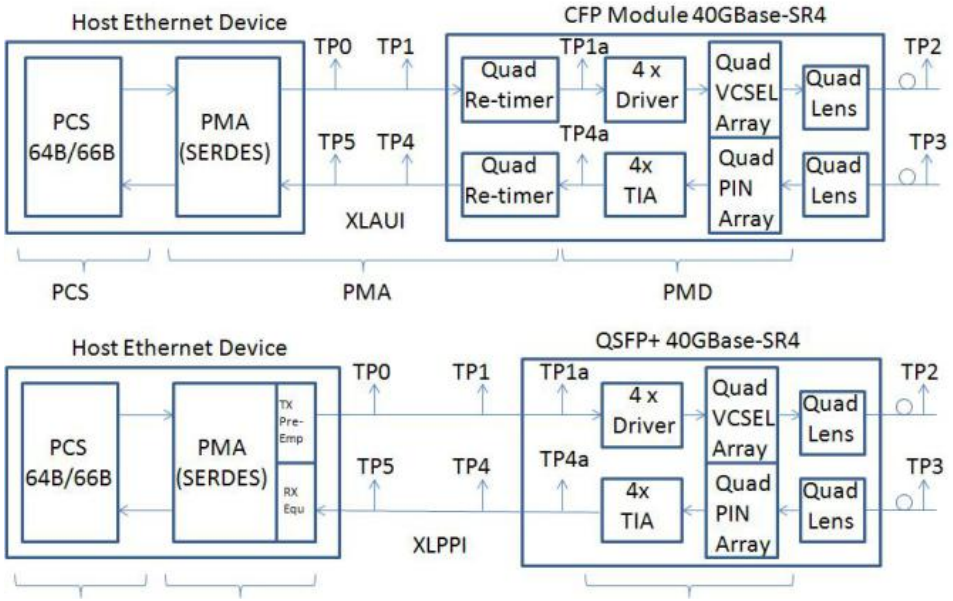


(http://40gethernet.files.wordpress.com/2011/01/xfi-and-sfi-eye.jpg)

The above eye diagrams give a visual indication of the difference between the XFI and SFI; the actual parameters can be found the table at the end of the blog.

The XLPPI and how it compares with the XLAUI and CFP modules

The XLPPI stands for XL(40Gbps) Parallel Physical Interface and is defined in 802.3ba Annex 86a as the interface between the PMA and PMD functions (where as the XLAUI dissects the PMA). The XLPPI is derived from the SFI interface and places higher signal integrity requirements on the host PMA than the XFI based XLAUI.



(http://40gethernet.files.wordpress.com/2011/01/cfp-and-qsfp.jpg)

The diagram above shows the differences between how a QSFP+ module uses the XLPPI versus a CFP module, which uses the XLAUI. The diagram also shows the 802.3ba test points which use a different naming convention than the SFF XFP and SFP+ specifications. As with the XFI and SFI the main difference is the increased signal integrity requirements for the host PMA and PCB signal traces.

Using the XLPPI with QSFP+ modules

The XLPPI interface is defined as being AC coupled on the host system, since the signals are running at 10Gbps (max 5GHz) it normal to route the traces as strip line on the top layer avoiding vias. When laying out the traces it is important to get both the impedance and differential trace length as close as possible to their target values. The impedance of the TXD and RXD pairs is 100Ω, with a termination mismatch of +/-5% at 1MHz; the closer your trace impedance is to 100Ω the more margin you'll have for the modules and ASIC's to vary their termination. The same applies for the length of the +ve and –ve traces, these need to be as close in length as possible since as any difference will eat away at the signal eye (note the effect of DDPWS in reducing the UI). However don't spend time trying to match the length of the four TXD+/- or RXD+/- pairs, this is not necessary as the MLD function takes care of this.

Currently QSFP+ modules support 40GBase-CR4 (both passive and active) and 40GBase-SR4 (either AOC or using the MPO/MTP®connector). However in June 2010 Colorchip announced their QSFP+ based 40GBase-LR4 module which is stated to be available in 2011, thus a system using the XLPPI/QSFP+ interface will soon be able to support the complete range of 40gBase-CR4, SR4 & LR4 – Excellent!

In addition to straight 40G interfaces there are also 4:1 hydra cables which have a QSFP+ at one end and four SFP+ cables at the other, these cables are typically direct attach copper and are used for aggregating servers to a top of rack switch. When being used for this application the high speed interface of the QSFP+ needs to conform to the SFI specification rather than the 802.3ba XLPPI.

Whilst the future trend is to adopt the XLPPI the XLAUI will probably still survive for special telecom orientated modules like SEI's 40Km 40GBase-LR4 CFP (more like a 40GBase-ER4), or any future 40GBase-FR module.

Short note on the CPPI and CXP

The CPPI is the equivalent interface to the XLPPI for 100G, currently it is only usable with the CXP module and it's not clear if the CXP module (defined initially for infiniband) will be used much for 100GigE due to concerns of reach and reliability of the MPO connector. The new 10x10MSA is currently defined as using CAUI though in the future this may move to CPPI for cost reduction. In the future 100G modules with a 4x28Gbps (CEI-28) interface will appear using a CFP2 format (similar in size to the X2 module).

Summary of electrical parameters

Interface comparison TX (MAC to Module)									
	MAC Interface (Output)				Module Interface (Input)			
	Interface	XFI	SFI	XLAUI	XLPPI	XFI	SFI	XLAUI	XLPPI
	Spec.	INF-8077	SFF-8431	802.3ba	802.3ba	INF-8077	SFF-8431	802.3ba	802.3ba
Parameters	X1	0.15UI	-	0.16UI	-	0.305UI	0.12UI	0.31UI	0.11UI
	X2	0.4UI	-	0.38UI	-	0.5UI	0.33UI	0.5UI	0.31UI
	Y1	180mV	-	200mV	-	60mV	95mV	42.5mV	95mV
	Y2	385mV	-	385mV	-	410mV	350mV	425mV	350mV
	Jitter	0.3UI	-	0.32UI	-	0.61UI	0.28UI	0.62UI	0.29UI

Interface comparison RX (Module to MAC)									
	MAC Interface (Input)				Module Interface (Output)			
	Interface	XFI	SFI	XLAUI	XLPPI	XFI	SFI	XLAUI	XLPPI
	Parameter	INF-8077	SFF-8431	802.3ba	802.3ba	INF-8077	SFF-8431	802.3ba	802.3ba
Parameters	X1	0.325UI	-	0.31UI	-	0.17UI	0.35UI	0.2UI	0.29UI
	X2	0.5UI	-	0.5UI	-	0.42UI	0.5UI	0.5UI	0.5UI
	Y1	55mV	-	42.5mV	-	170mV	150mV	136mV	150mV
	Y2	525mV	-	425mV	-	425mV	425mV	380mV	425mV
	Jitter	0.65UI	-	0.62UI	-	0.34UI	0.7UI	0.4UI	0.65UI
	DDPWS	-	-	-	-	-	0.34UI	-	0.34UI

(http://40gethernet.files.wordpress.com/2011/01/xfi-sfi-xlaui-xlpp-comparison.jpg)

The above table captures only a small portion of the high speed interface requirements, when designing a board it is important to get the complete specification. For SFI and XFI the relevant specifications can be downloaded free at http://www.sffcommittee.org (http://www.sffcommittee.org/) whilst the XLAUI and XLPPI specification is included in the (also free) 802.3ba specifications at http://standards.ieee.org/about/get/802/802.3.html (http://standards.ieee.org/about/get/802/802.3.html) .

The 40G Ethernet Resource Center
http://www.40GEthernet.com (http://www.40gethernet.com/)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

—-

Overview of TLA's

CAUI : 100Gbps (C) Attachment Unit Interface
CPPI : 100Gbps (C) Parallel Physical Interface
EDC : Electronic Dispersion Compensation
DDPWS : Data Dependant Pulse Width Shrinkage (in UI)
SFI : Small form Factor Interface

SFP : Small form Factor Pluggable
UI : Unit interval of a bit period, for all the interface above this is 96.97ps (10.3125Gbps)
XFI : 10Gbps(X) form Factor Interface
XFP : 10Gbps(X) form Factor Pluggable
XLAUI : 40Gbps (XL) Attachment Unit Interface
XLPPI : 40Gbps (XL) Parallel Physical Interface

Posted in Technology | 5 Comments »

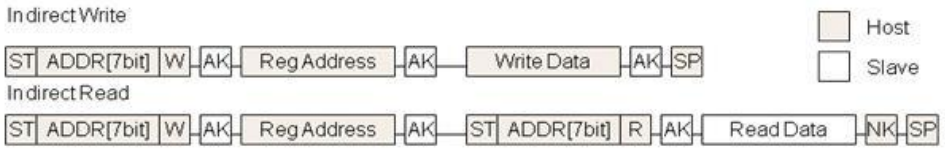
Control Interface : I2C and MDIO

January 6, 2011
For most pluggable optical modules the interface used for monitor and control is the I2C interface, also known as the 2-wire interface, however since the advent of the first 10GBase-LX4 Xenpack modules an alternative control interface has also been used, this being the MDIO interface. In this blog we'll look at how both control interfaces are used, their differences and how they will be used in the future.

I2C / 2-wire Interface

The I2C interface was developed by Philips(now NXP) as a low speed, low pin-count control interface and is found in every type of application from consumer, industrial and communications and as such it is easy to find microcontrollers, control logic and EEPROM's with I2C interfaces. The I2C or 2-wire interface as it is also known as (due to trade mark issues) comprises of a bidirectional data line (SDA) and a clock signal (SCL) generated from the host, a serial byte based transfer protocol is used with each byte having an acknowledge bit (ACK) associated to it. In order to allow multiple bytes to be grouped together to form a message the first byte contains a Start bit (ST) and the last byte is followed by a Stop bit (SP). The smallest I2C message comprises of two byte transfers, the first byte contains a 7bit address with a R/W indicator and the second with the 8bit data, for small devices this is sufficient however most applications use the 7bit address to identify a particular physical device (on an I2C bus) and multiple data bytes to transfer 8 or 16bit address or data. For more details on how the I2C protocol works I'd advise looking at the "Using the I2C Bus" (http://www.robot-electronics.co.uk/acatalog/I2C_Tutorial.html) article, by Robot Electronics, browsing the excellent info at i2c-bus.org (http://www.i2c-bus.org/)and of course getting a copy of the I2C spec from NXP (http://www.nxp.com/documents/user_manual/UM10204.pdf).

A classic example is the method used to access an 8bit EEPROM via I2C using indirect addressing. Here the EEPROM has internal holding registers for both the address and the data, to write to a location three bytes are sent, an initial byte indicating that a write access(R/W=0) followed by a byte for the address and a byte for the data, note that all bytes have to be acknowledged else the transfer will be aborted. When performing a read access four byte transfers are used, the first two are used to write the address location to the EEPROM as above and the third byte contains a Read command (R/W=1), the fourth byte is sent by the EEPROM itself and contains the data of the location pointed to by the address register. In order to allow for the slave device to pause the transfer of bytes (so that it can read internal registers) it is possible to force the clock line low which prevents further cycles, this is called clock stretching.



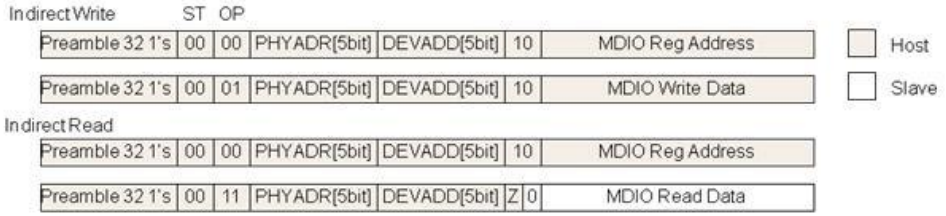
(http://40gethernet.files.wordpress.com/2011/01/i2c-frame1.jpg)

The EEPROM method of read/write access has been adopted by the SFF committee for communicating with SFP+(SFF-8431/8461), XFP(INF-8077i) & QSFP+(SFF-8436) modules as initially all that was contained in the modules was an EEPROM with identification/calibration information.

There are several variants of I2C specified with clock speeds upto 4MHz however most applications, including the SFF modules and 300pinMSA modules use either normal mode (100KHz) and fast mode (400KHz)

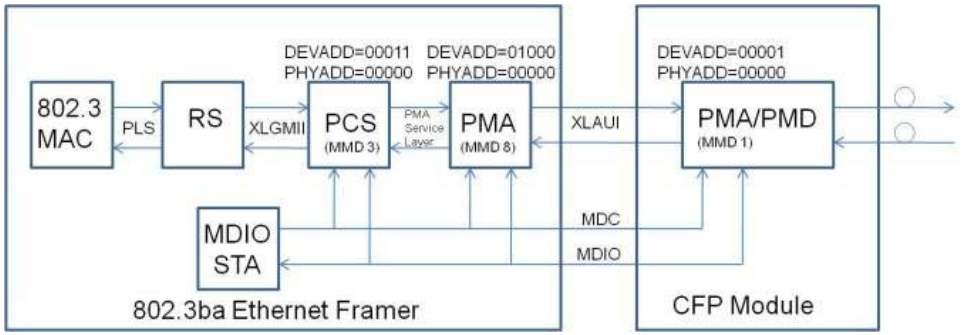
MDIO clause45 interface

The MDIO interface originates from the development of 100Mbps Ethernet and was part of the MII defined in 802.3 clause22, however for 10G and above application a new variant called clause45 is used, which uses a lower voltage (1.2V) and allows access to a 16bit address range. The MDIO is a two wire interface (clock and bidirectional data) with a transfer size of 64bits, the initial 32bit are preamble (set to 1) with the remaining 32bits containing a 16bit header and 16bit data. In the earlier clause 22 specification the header contained a 5bit address used to identify the registers in the individual physical interface device. However with the advent of 10G physical layer devices became more complex so a new variant, clause45, was created which along with a modified 16bit header uses a similar indirect addressing approach described above for the EEPROM. In clause45 an initial message is sent containing a 16bit address located in the data field and a second MDIO message is then sent to either read or write 16bits of data to the addressed register.



(http://40gethernet.files.wordpress.com/2011/01/mdio-frame1.jpg)

The 16bit header comprises a start indicator (ST), an operation command (OP) to indicate the transfer type (Address, Write, Read and Read Inc) and two 5bit address fields. The PHYADR field is used to select a particular physical channel/port, this is compared to external physical address lines used to identify the port. The DEVADD is used to select a particular device function used by the channel port, these functions are called MMD's (MDIO Manageble Devices) typically only three MMD functions exist. The example below shows a CFP module with a XLAUI interface, there are three MMD's per channel, the PCS and PMA located in the Ethernet Framer and the PMA/PMD located in the Module.



(http://40gethernet.files.wordpress.com/2011/01/mdio.jpg)

(http://40gethernet.files.wordpress.com/2011/01/mdio-connect.jpg)

Key differences between I2C and MDIO

Unlike I2C, the MDIO interface does not support an acknowledgment for the data transferred, thus when writing to a register the only way to check if the value has actually been accepted is to check by reading the register. When reading a device again there is no explicit acknowledgement, however the bit preceding the 16bit data should always be zero which can be used to verify that a data value of 0xFF is valid rather than a no-response.

A second difference with I2C is that the MDIO does not support any method for inserting delay's or wait states, this means that when a read or write command is sent the slave device must react immediately which has an impact on the implementation of the registers in a slave device. This makes the design of a MMD controller a little more complex as it is necessary to keep a shadow memory of registers in dual port memory to allow instant access via the MDIO whilst the local microcontroller in the module monitors the different electro-optic control circuitry.

Despite the above draw backs against I2C the main advantage that MDIO has is speed and transfer rate, the maximum MDIO clock is 4MHz and by not allowing the slave to insert wait states (and thus hold the bus) the message throughput of the MDIO is predictable. This makes MDIO a good choice for complex interfaces with a high number of registers such as longer reach 40G and 100G modules.

Summary

I2C is the control interface of choice for low cost pluggable module ranging from SFP+ & XFP for 10G and QSFP+ for 40G, the 100G CXP module defined by the IBTA also uses I2C and it is expected that any future QSFP module for 100G (4x28Gbps) will also use I2C.

MDIO clause45 is used for all 10G physical layer devices including 10GBaseT and 10GBaseKR and for specific 10GBase-LX4modules (XENPAK, X2 and XPACK). The MDIO clause45 is also the control interface for all 40/100G CFP modules (http://www.cfp-msa.org) and for OIF (http://www.oiforum.com) compliant 100G DWDM modules.

The 40G Ethernet Resource Center
http://www.40GEthernet.com (http://www.40GEthernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

Posted in Technology | 1 Comment »

OC-768 POS and 40G Infiniband

October 19, 2009
When it comes to 40G networking Ethernet is a relative newcomer as there are already two technologies being used these being OC-768 POS and Infiniband, below is a quick overview of these technologies and where they are used.

OC-768 POS

OC-768 and its European version STM-256 are the top rates for the venerable SONET/SDH transmission standards that over the last 20+ years have grown to be the backbone of the world telecommunication networks. Unlike its highly successful 10G predecessors, OC192/STM-64, the 40G OC768 never really made it into transport networks mainly due to cost, limitation with optical performance (due to dispersion) and the deployment of DWDM systems using OTN. A few companies tried it for replacing multiple 10G links between co-located systems but even that never really caught on due to the arrival of XFP's which removed any cost advantage.

Where OC-768 did gain a foothold was as the interface for the top end Core Routers especially from Cisco who's GSR1200 and CSR1 kept the 40G market going from 2002 to 2008. When used as an interface for Routers the OC-768 is called a Packet over SONET or POS interface. POS interfaces have been around since the mid 1990's and were initially developed by Cisco to overcome the limitation of ATM and support data rates from OC-3 upto OC-768. The concept of POS is to map IP packets directly into the payload of the SONET frame using byte synchronous PPP (no TDM multiplexing), this allowed Routers to plug directly into the SONET network so they could be connected in a mesh configuration (no additional Layer 2 such as ATM).

The key components for a 40G OC-768 POS interface on a Core Router are:

- » 40G capable NPU with SPI-5 or Interlaken packet interfaces, such as the EZ-Chip (http://www.ezchip.com)NP-4 Network processor that support Interlaken as well as XAUI interfaces.
- » OC768 POS framer such as the Cortina Systems (http://www.cortina-systems.com) CS1999, which has a Interlaken packet interface and SFI-5 interface to connect to a 40G optical module
- » 40G optical module which has a SFI-5 system interface

The 40G optical modules used on Routers are often refered to as being 300pinMSA (http://www.300pinmsa.org) modules due to the 300pin system connector (for the OIF (http://www.oiforum.com) defined SFI-5 interface) and that they are designed to a common MSA specification allowing modules from several vendors to be used. The modules are also refered to as "VSR2000" modules which indicates that the module is "Very Short Reach" at just 2000 meter, so only really applicable for connecting co-located system. Initially the size of a 40G VSR2000 MSA compliant module was 5" x 9" but 2009 has seen several companies (Finisar (http://www.finisar.com), Opnext (http://www.opnext.com) and CoreOptics (http://www.coreoptics.com)) releas small form factor 40G VSR2000 modules which are 5" x 4" (following the 10G 300pinMSA form factor).

For connecting over long distance the OC-768 interface would normally connect to a DWDM system that would wrap the OC-768 into a OTU3 and then using a combination of special 43Gbps DWDM optical modules (supporting modulations such as DPSK) and fancy optical components like dispersion compensators and optical amplifies transmit it the signal a very long way indeed.

Cisco however is starting to promote IPoDWDM which is basically the integration of the DWDM optics into the 40G Line card as discussed in the video link from Cisco below

Cisco on 40G v. 100G and IPoDWDM at NXTcomm08 you tube video : http://www.youtube.com/watch?v=zSktgLwCmUk (http://www.youtube.com/watch?v=zSktgLwCmUk)

The change to Ethernet on the Core Router certainly won't happen overnight and OC-768 POS will be around for many years and we'll still need to make sure systems can interface to it (a key driver behind a Serial 40GBase-R standard to allow for multi-protocol support of POS and Ethernet) but in the future OC-768 POS will wear the mantle of "legacy interface".

40G Infiniband

Infiniband has had an interesting life; born in the internet bubble of 2000 as a technology that would allow the unification of computing, storage and local area networks (hey it was so hip it even used IPV6 addressing). As the bubble burst it could have disappeared but managed to find a niche in becoming the defacto solution for connecting together CPU Blades to form processing cluster for high performance computing applications.

Infinband is not a big market and there exists only one vendor supplying both the switching and adaptor ASICs, and that be Mellanox (http://www.mellanox.com). However what Infiniband has lacked in market size its made up adaption of high speed low cost links, from 10Gbps upto 40Gbps, mostly using QSFP modules, both optical and direct attach copper . Infact the adoption of 40G Ethernet is benefiting greatly from the fact that there is an already existing market for 40G Optical & Copper modules that it can tap into (just as Gigabit Ethernet initially benefited from 1G FiberChannel).

Infiniband has a very flexible physical layer implementation, the link data rate 2.5Gbps (SDR), 5Gbps (DDR) and 10Gbps (QDR) and there can be 1, 4 or 12 lanes (each lane must be the same speed). For 40Gbps Infiniband the physical layer is 4 lanes of 10GBps (QDR) and typically uses QSFP modules which come in three types;

- » Optical module with MPO connector – e.g. Avago (http://www.avagotech.com) & MergeOptics (http://www.mergeoptics.com)
- » Active Optical Cable – e.g. Finisar, Luxtera (http://www.luxtera.com) & MergeOptics
- » Active Copper Cable – e.g. Gore (http://www.gore.com), CINCH (http://www.cinch.com) & Quellan (http://www.quellan.com)

It should be noted that Infiniband uses the venerable 8B/10B encoding, so for a 40Gbps Infiniband link the actual data rate is 32Gbps, just like the line rate for 40G Ethernet is actually 41.25Gbps due to the 64/66B encoding.

To put Infiniband in context with Fiber Channel and Ethernet;

- » Ethernet : Primarily for Transport of TCP/IP
- » FiberChannel : For Transport of SCSI
- » Infiniband : Primarily used for Remote DMA access between CPU processors

Now as we know Ethernet is starting to eat in the Fiber Channel market, both with iSCSI (that runs on top of TCP/IP) and also with a combination of the new FCoE standard that makes use the Data Center Ethernet enhancements. Ethernet can also be used to network CPU clusters using the TCP/IP based iWARP standard, however this does not cut it for the key high performance due issues such as latency. Work is in progress on dedicated RDMAoE and the latest Ethernet adaptors from Mellanox are supporting RDMA over Ethernet.

In the end the spread of adoption of Ethernet as the layers 1&2 for storage and computing networks is going to be more to do with the strategies of the incumbant suppliers. They can either dig in and defend their turf but accept they will be a niche player or they take a gamble and make a play for Ethernet. Looking at what's happening to Brocade, QLogic and Mellanox its pretty clear they've decided to go for the later and take a bet on being a major player in Ethernet.

The 40G Ethernet Resource Center
http://www.40GEthernet.com (http://www.40gethernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

Posted in Technology | Leave a Comment »

40G Ethernet component and product update

October 12, 2009
September 2009 was great month for the advancement of 40GEthernet as it saw several significant product releases, both at the Intel Developer forum and at ECOC, Europes premier show for optical communications. Here's an overview of what we thought were the key product highlights :-

World first 40G Ethernet NIC card
It was no surprise that Mellanox (http://www.mellanox.com), the leader (and for a long time sole supplier) of 10/40G Infiniband solutions announced the availability of the worlds first 40G Ethernet NIC card, the ConnectX EN 40G.

http://mellanox.com/content/pages.php?pg=press_release_item&rec_id=350 (http://mellanox.com/content/pages.php?pg=press_release_item&rec_id=350)

What is surpsing is how compact the ConnectX EN 40G really is, as shown in the picture at the link below :-

http://mellanox.com/content/pages.php?pg=products_dyn&product_family=70&menu_section=28 (http://mellanox.com/content/pages.php?pg=products_dyn&product_family=70&menu_section=28)

The ConnextX EN 40G is based on its ConnectX 2 silicon and supports a single 40Gbps interface provided by a pluggable QSFP module for network connection and a PCIe2.0 system interface.

The card supports all the standard Ethernet features as well as priority based flow control which should be usefull in the data center and to help TCP/IP applications use all that bandwidth the ConnectX EN 40G suppoprts TCP/UDP/IP stateless offoad, like checksum calculation/check. The adapator is not just for the standard TCP/IP world and Mellanox have added support for both FCoE and RDMA (which is a bit like IBoE).

40G CFP modules from Finisar and SEI

At ECOC 09, Vienna, both Finisar (http://www.finisar.com) and SEI (http://www.global-sei.com) (Sumitomo Electric Inc) introduced 40GBase-LR4 modules using the CFP MSA format. The CFP MSA can be used for both 40G Ethernet and 100G Ethernet (that's the C in the CFP) and both companies were founder members of the CFP MSA organisation along with Opnext and Avago.

- » http://investor.finisar.com/releaseDetail.cfm?ReleaseID=410286 (http://investor.finisar.com/releaseDetail.cfm?ReleaseID=410286)
- » http://global-sei.com/news/press/09/09_13.html (http://global-sei.com/news/press/09/09_13.html)

We could try to write a few words about the features of the 40GBase-LR4 CFP modules but why bore you when you can watch the excellent video, made by Finisar at ECOC, showing their 40GBase-LR4 module being used with an EXFO 40G Ethernet tester to run error free over 10km :-

http://youtube.com/watch?v=Lir4g1SNRQ8&feature=related (http://youtube.com/watch?v=Lir4g1SNRQ8&feature=related)

In addition, Finisar have anounced new small form factor 40G VSR module based on the 300pin MSA which has the capabilty to operate at any rate from 39Gbps (SONET) upto 44Gbps (OTU3e) and also the 41.25Gbps rate required by serial 40G Ethernet!!

100G CFP modules by Opnext and Reflexphotonics

Whilst the blog is mainly about 40G Ethernet we can't help but mention that Opnext (http://www.opnext.com) showed a 100GBase-LR4 CFP and module and Reflexphotonics (http://www.reflexphotonics.com) showed their 100GBase-SR10 CFP module.

- » http://investor.opnext.com/releasedetail.cfm?ReleaseID=410405 (http://investor.opnext.com/releasedetail.cfm?ReleaseID=410405)
- » http://www.reflexphotonics.com/pressrelease203.htm (http://www.reflexphotonics.com/pressrelease203.htm)

QSFP and CXP modules from Merge Optics

MergeOptics (http://www.mergeoptics.com) have on offer a new range of 40G QSFP and 120G CXP modules. The 40G QSFP modules came in both the active cable variant and the MPO connector variant making them (we think) the first vendor to offer both. On the 120G CXP side the offered a 1:1 CXP active cable as well as unique 3x40G to 1x120G hydra active cable.

CFP Host Connector solution

Tyco Electronics (http://www.tycoelectronics.com) announced their complete solution for a CFP host connector (i.e. the stuff that goues on the line card to allow you to plug in a CFP module). The solution covers the 148pin Z axis connector, receptacle cover, guide rails, external bracket assembly, backer plate and heat sink.

Although still preliminary all parts come with the necessary drawings and 2/3D CAD files.

http://www.tycoelectronics.com/catalog/minf/en/714?BML=10576,17560,17553,17676
(http://www.tycoelectronics.com/catalog/minf/en/714?BML=10576,17560,17553,17676)

Best of luck to all the companies above and we hope your investement in 40G Ethernet pays off.

40G Ethernet Resource Center
http://www.40gethernet.com (http://www.40gethernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

Posted in Technology | 1 Comment »

Overview of the XAUI, XLAUI and CAUI: Part2

June 9, 2009

Part2 of the "Overview of the XAUI, XLAU and CAUI" blog focuses on the details of the XLAUI and CAUI interfaces and how the new MLD function works. Rather than re-inventing the wheel we suggest you grab copies of the top class presentations, from the IEEE P802.3ba Task Force Meeting Munich May 2008, listed in the text below.

The "hidden" XLGMII and CGMII interface
(XL/CGMII and RS proposal :
http://www.ieee802.org/3/ba/public/may08/gustlin_02_0508.pdf
(http://www.ieee802.org/3/ba/public/may08/gustlin_02_0508.pdf))

The XLGMII/CGMII is specified as a nominal 64bits wide data bus, separated into 8 lanes, with control/data signal per lane (not to be mixed up with the virtual lanes of the MLD function in the PCS). At 64bits the clock rate must be 625MHz for 40G and 1.5625GHz for 100G, though scaling this down to a typical 256width bus of 32 lanes would results in more reasonable values of 156MHz and 390Mhz.

The control lines, one per 8bit lane, indicate if the associated lanes contain data or control and help to identify the start/stop of the packet and indicate when IPG or Ordered Set control data is present. Unlike 10G, data is always assumed to be byte aligned which helps make the PCS a little less complicated (unless your build a 4×10G / 40G Ethernet MAC!)

XLAUI/CAUI – The PCS and its new function the MLD
(100/40Ge PCS proposal :
http://www.ieee802.org/3/ba/public/may08/gustlin_01_0508.pdf
(http://www.ieee802.org/3/ba/public/may08/gustlin_01_0508.pdf))

The XLGMII (and CGMII) provides the interface to the PCS block, sometimes referred to as the PCS/MLD block (technically the MLD is just a function inside the PCS). In the transmit direction the first stage is the 64b to 66b encoding where the information in the outband control/data lines is turned into a 2bit sync header and, if a control word is present, a block type field which resides in the first byte of the 64bit payload. The 64bit payload is then scrambled with the standard 1+ X^39+X^58 and the 2bit sync header added to make a 66bit block.

Upto this point it has been the same as serial 10G Ethernet and if the optical link were serial the data could have been passed straight away to the PMA function, however to pass over the multi 10G serial interface of the XLAUI and CAUI the data needs the help of the MLD block (Multi Lane Distribution) that maps the data into Virtual Lanes (VL's) so it can pass over the 4×10G XLAUI

and 10×10G CAUI. For the XLAUI, the number VL's is 4 and corresponds to both the number of 10G channels in the XLAUI and the 4 channels of the PMA for 40GBase-xR4. The CAUI however has 20 VL's! the reason for that will be explained later.

Assuming a 40G Ethernet link, the transmit path of the MLD is as follows; The data is transmitted by sending 64/66b code words on each lane in a round robin fashion, this basically means there are four 64/66B codes words being sent at the same time, one code word per phyiscal link (just like XAUI) . In order to keep alignment a 64/66b alignment word is sent simultaneously in each VL every 16384 66b blocks. Unlike the XAUI were the /A/ code is only sent in the IPG the MLD alignment 66b word can be sent in the middle of the payload so it needs to be removed by the MLD function in the receiver before the 64/66b decoding. The four VL's are then passed to four separate SERDES which each produce a 10.3125Gbps serial signal giving us the transmit XLAUI interface. The receive path is pretty much the reverse but with the need to perform 64/66b block alignment on each of the physical lanes of the XLAUI, to detect (and delete) the alignment word in each VL and to finally buffer the data on each VL so that the received 64/66b blocks are re-aligned so they can be decoded and passed to the XLGMII interface.

The process for the CAUI is identical apart from the fact that there are 20 VL's. This is due to the fact that whilst the CAUI is 10bits the PMA service layer can be either ten (100GBase-SR10) or four (100GBase-LR4). The lowest common dividable number for 10 and 4 is 20, hence 20VL's for 100G.

Available implementations from Sarance Technologies and MorethanIP
OK first point; this is not a competitive analysis and we think it's fantastic that these two companies have invested their time and resources in producing pre-standard building blocks to help 40/100G Ethernet hit the ground running.

Sarance Technologies is based in Ottowa,Canada and along with their 40/100G FPGA IP Core also provide FPGA IP for the 120G Interlaken interface as well as Core for Classification and Traffic Management.

The Sarance 40/100G Core is in three sub blocks; MAC, PCS and MLD and supports either a single channel 40 or 100G implementation. The data path of the Core can be selected to be either 256bits/200Mhz (for 40G) or 512bits/250MHz (for 100G). There is no internal XLGMII/CGMII interface so it seems it is intended to use all three block together. On the XLAUI/CAUI interface the core relies on the internal SERDES on the FPGA, whilst the diagram on the product brief implies that it will be a 10G SERDES but it can probably also support 2×5.12Gbps as they are working closley with Xilinx but its not stated in available info. The Sarance core is available for both Altera and Xilinx, but they seem to have more connection to Xilinx (http://www.xilinx.com), they were part of a demo with Xilinx, Netlogic (http://www.netlogicmicro.com), Avago (http://www.avagotech.com) and IXIA (http://www.ixiacom.com) at NFOEC in March 2009 and provide the 100G Ethernet core on the Xilinx Virtex5 TXT 100G Platform (http://www.xilinx.com/publications/prod_mktg/pn2094.pdf).

Visit the Sarance Technologies website at http://www.sarance.com (http://www.sarance.com)

MorethanIP is based in Munich,Germany and has been going for nearly 10 years. They have a good portfolio of Ethernet (GigE/10GE) and SONET FPGA IP Cores.

The MorethanIP 40/100G IP Core comprises of a PCS Core and a MAC core, thus enabling then to be used separately. The interface between the MAC and PCS is a XLGMII or CGMII interface as described above. The MAC core supports the standard necessary features but they also added some support for per channel flow control and time stamping to help with applications such as FCoE (nice!). On the PCS side both 4 and 20VL's are supported with again the SERDES being provided by the FPGA, the SERDES can be either of the 10.3G or 5.15Gbps(XSBI) variety, the support for 5.15Gbps enable lower cost FPGA's to be used. The MorethanIP 40/100G Ethernet solution is available as both generic synthesizable cores, in both VHDL and Verilog, and a targeted version optimized for Altera (http://www.altera.com) StratixIV GT.

Visit the More than IP website at http://www.morethanip.com (http://www.morethanip.com)

Best of luck to both companies!

The 40G Ethernet Resource Center
http://www.40gethernet.com (http://www.40gethernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

Posted in Technology | 1 Comment »

Overview of the XAUI, XLAUI and CAUI: Part1

May 8, 2009
Every speed step made in the Ethernet standard has required a new front end interface to help the digital MAC framers interface to Physical layer devices and to preserve the continuity of the MAC functionality (It has to be noted that the MAC layer at 10Mbps is the same format as that used in 100Gbps). This blog tries to give a brief overview of how these interfaces have evolved and the features that have been added.

(To help with the following explanation a diagram of the Ethernet System interfaces and functional blocks (from 10Mbps to 100Gbps) can be downloaded from the 40G Ethernet Resource Center here (http://sites.google.com/site/40gethernet/documents)).

Background on 10/100/1000Mbps Ethernet system interfaces

At the very beginning we had the AUI interface for 10Mbps Ethernet that allowed a MAU to be plugged in (such a PHY that interfaced to Coax cable). The function that separated the MAC from the AUI was called the PLS which implemented the Manchester encoding along with other things. The PLS is long gone but even so the interface to the MAC itself is still referred to as the PLS service layer.

When 100Mbps arrived it was found the existing AUI was not suitable and it was replaced by the MII interface (and its derivatives RMII and SMII). The PLS function was dropped and replaced by the RS function which on one side has the MII interface and the other side the PLS service layer interface. The basic role of the RS layer is to incorporate any necessary data modifications to adapt future MII interfaces to the unchanged MAC interface. The MII would then interface to a Physical Layer device which used the 4B/5B coding scheme (taken from FDDI) plus other media specific functions, such as the MLT-3 encoding required for 100Base-T.

With the arrival of Gigabit Ethernet the GMII, and later RGMII / SGMII, were introduced along with the 8B/10B PCS coding, which was taken from Fibre Channel. At this point all GigE MAC's integrated the PCS layer and provided a TBI which allowed direct connection to available 10bit SERDES devices (which were used for Fibre Channel). Initially the TBI was more popular and many MAC devices did not support the GMII, this required the later 1000Base-T transceivers to support both GMII and TBI on the system interface.

10G Ethernet & the XAUI

For 10GEthernet the standard MAC/RS interface is the XGMII which interfaces to a Physical layer device that supports PCS/PMA/PMD function. For optical links the PCS coding is defined as using 64B/66B coding rather than 8B/10B in order to keep the line rate down to 10.3125Gbps (rather than 12.5Gbps required if 8B/10B was used). However it was also clear that whilst the XGMII interface would be rarely offered as an external interfaces on chips, due to the pin count, the PCS/PMA/PMD functions would initially reside in an external PHY layer SERDES device – thus some type of mid range interface would be required to allow the MAC to connect to the PHY and the solution was the XAUI.

The XAUI comprises of 4 x 3.125Gbps physical lanes with each use the 8B/10B coding scheme (rather than than 64B/66B) and in order to handle the potential deskew between the four lanes a new 8B/10B control word /A/ was introduced. The /A/ control word is periodically inserted, at exactly the same time, into each XAUI stream with the idle control word in the IPG being deleted to compensate for the added bandwidth. In the receive direction the /A/ are detected and used as alignment markers to re-synchronize the four XAUI lanes.

The XAUI interface not only found it's self as a chip to chip and backplane interface, but also as the interface to 10GBase-LX4 modules which use four WDM multiplexed lambda's to provide 10G Ethernet over OM3 cable, a concept that would be the basis for 40/100G Ethernet.

40/100G Ethernet and the XLAUI/CAUI

As with 10Gbps Ethernet the relevant MII interfaces, XLGMII(40G) and CGMII(100G), are no longer assumed to be external interfaces but just function boundaries inside devices, the role of providing an external interface is taken by the XLAUI and CAUI. The XLAUI and CAUI interfaces follow the concept of the XAUI interface but due to the bit rate required use the 64B/66B encoding, rather than 8B/10B, to keep the line rate down to 10.3125Gbps.

As with the XAUI interface it is necessary to compensate for mis-alignment of the various lanes, due to different delays in each path, which is solved by the insertion of a special 64B/66B alignment code word. To handle the various data path widths which range from 10, 5 and 4 the concept of virtual lanes is used and the management of these virtual lanes is handled in a new function called the MLD layer.

In part 2 of the blog we'll go into detail on the concept of virtual lanes and how the MLD function works plus take a look at recently available 40/100G Ethernet FPGA IP cores from Sarance (http://www.sarance.com) and MorethanIP (http://www.morethanip.com).

The 40G Ethernet Resource Center
http://www.40gethernet.com (http://www.40gethernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

——

Darn those TLA's (and FLA's – Four/Five Letter Abbreviations)

AUI : Attachment Unit Interface, originally connected to a MAU
MAU : Medium Attachment Unit, like a 10Base-2 transceiver.
MLT-3 : Multi Level Transmission 3; Three level line code (+1/0/-1) used by 100Base-T
XAUI : A 10G AUI, the X is the Roman numeral for 10; Data path is 4×3.125Gbps Lanes
XLAUI : A 40G AUI, XL being the Roman numeral for 40; Data path is 4×10.3125Gbps Lanes
CAUI: A 100G AUI, C (you guessed it) being the Roman numeral for 100; Data path is 10×10.3125Gbps Lanes

MII : Medium Independent Interface, 4bit wide data path.
RMII : Reduced MII, the MII but with less signals!
SMII : Serial MII, the data path is reduced to one bit.

GMII : Gigabit MII, 8bit wide data path.
RGMII : Reduced Gigabit MII.
SGMII : Serial Gigabit MII.
TBI : Ten Bit Interface

XGMII : 10G MII (this time the G made it in).
XGXS : XGMII eXtender Sublayer.
XLGMII : 40G MII.
CGMII : 100G MII.

MAC : Media Access Controller.
PLS : Physical Layer Signaling; for 10Mbps only, implemented the Manchester encoding.
RS : Reconciliation Sublayer.
PCS : Physical Coding Sublayer; e.g. 8B/10B.
MLD : Multi Lane Distribution
PMA : Physical Medium Attachment.
PMD : Physical Medium Dependant.

IPG : Inter Packet Gap; Code words sent between valid Ethernet Frames.

Posted in Technology | 4 Comments »

Using the 40/100G CFP MSA for nx10G Ethernet

April 27, 2009
When the 40/100G CFP MSA (http://www.cfp-msa.org (http://www.cfp-msa.org)) first came out in March I initially commented that it seemed optimized for 100G applications, though its size would make it easier to implement initial 40G modules. However after after taking a look at new systems from Extreme Newtorks (http://www.extremenetworks.com) and Arista Network

(http://www.aristanetworks.com), plus having some time to digest the finer details of what's written, it's now clear that the "size" of the CFP module open up the potential to use the format not only for 40/100G but also for high density 10G ports.

One of the concerns I've had about the adoption of 40/100G Ethernet is how easy it will be to build modular systems that allow people to switch from nx10G to 40G and finally on to 100G, since each line rates uses a different mechanical format (SFP+, QSFP and CFP). Having such a broad range of mechanical formats creates big problems in inventory management (you never have the right module) and was probably one of the factors that slowed the adoption of 10G Ethernet down (XENPACK, X2, XPACK, XFP and finally SFP+). It's interesting to note that the Fibre channel community avoided this as much as possible with 1Gbps upto 8Gbps FC all using the same SFP format.

After discussing this concern with a friend I was pointed to Extreme Networks and Arista Networks to take a look at how these two companies were approaching the problem. The latest system from Extreme (X650) has an uplink module of eight 10GBase-T links using link aggregation, it's easy to envisage the step to a plug in module using 2 x 40G especially as the internal system interface "could" be identical (8xXFI could also work as 2xXLAUI). This lead me to consider how the CFP format could be used not just as a "plug in" Optical module but as a network interface module covering both Optical and Electrical interfaces, here's what I mean…

- » 1x40GBase-SR4/CR4
  A single QSFP cage with quad SFI+ retimer, such as those from Gennum, interfacing to the CFP XLAUI system interface. The available QSFP pluggable modules seem to come in three types, optical modules with an MTP/MPO connector (e.g. Avago (http://www.avagotech.com)) , "Active" optical cable (e.g. Finisar (http://www.finisar.com)) and direct attach which is for copper (e.g. Quellan (http://www.quellan.com)).

- » 4x10GBase-SR/LRM SFP+ module based on CFP format
  Using a 4xSFP+ cage (e.g Pulse Engineering (http://www.pulseeng.com)), which just about fits into the width and height of the CFP format, along with a Quad SFI+ retimer from Gennum to interface to the XLAUI system interface. This would allow the use of 10G SFP+ modules such as those from Opnext (http://www.opnext.com), Excelight (http://www.excelight.com), Finisar and Avago.

- » 4x10GBase-T module based on CFP format
  Using a 4xRJ45 ganged connector (e.g Pulse Engineering or Belfuse (http://www.belfuse.com)) with two dual 10GBase-T transceivers from Teranetics (http://www.teranetics.com) which support XFI for the system interface. The Teranetics devices are 6W per channel so the module would consume in the region of 25W (including on board controller) which is just on the border line between the CFP class 3 and class4 rating.

All the above assumes that the MAC can be switched from 4x10G to 1x40G, this is probably not unreasonable and the announced Prestera CX from Marvell (http://www.marvell.com) is described as having this capability, we're eagerly waiting further detailed public info on this new switch family to find out more.

Initially I thought that the maximum number of 10G ports one could fit on a CFP would be 4, but then I checked out what Arista were doing. What really caught my eye here was how their "Pass Through" cards (used with IBM BaldeCenterH) use QSFP to 4xSFP+ direct attach cables to consolidate the cabling – now this got me really thinking (or dreaming)….

- » 3x40GBase-SR4/CR4 or 12x10GBase-SR/CR1
  Use the 3 x QSFP option in the CFP MSA and either direct attach QSP:4xSFP+ modules or standard QSFP Optical modules with MTP/MPO to 8xLC connector fiber optic cables from Molex (http://www.molex.com).

To take advantage of the possibility to aggregate 3x40G and 12x10G onto a single CFP would require a very flexible 10/40/100G MAC but just imagine the possibilities : a 24+2 switch (by that I mean 24x10G + 2x100G) on a 1U platform! It will take a while as we'll need an Ethernet switch chip that is capable of 0.5 Tbyte capacity that supports Data Center Ethernet (DCE) but definitely something to look forward to.

40G Ethernet Resource Center
http://www.40gethernet.com (http://www.40gethernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

Posted in Technology | 1 Comment »

40G Ethernet over DWDM : Part2

April 19, 2009
With the technical barriers of how to map a 40G Ethernet signal into an OTU3 pretty much resolved (see part 1), early adopters will face the problem that current 40G DWDM systems support only OC-768/STM-256 or OTU3 client interfaces! Since the DWDM market moves slower in releasing new technology than in the enterprise (due to component complexity and longer qualification times) this problem will be with use for some time.

First it's best to recap on the two types of interfaces that are currently provided, OC768 and OTU3. The OC768 interface run at 39.8Gbps and is used mainly for connection to Routers for OC768 POS. The second client mode offered by most systems is an 43.02Gbps OTU3 client, more correctly described as an OTM0.3 interface, which uses OTU3 framing with G709 RS-FEC and is used mainly as an internetwork interface between DWDM systems (IaDr or IrDr). The optical interface for both OC768 and OTM0.3 is 40G serial and currently uses 300pin MSA VSR modules which are available from several vendors. Currently OC768 and OTU3 framing are incompatible with the XLAUI system interface, due to the 4x10Gbps data path, however there is talk in ITU about a proposal to map OC768 and OTU3 over XLAUI which we'll touch on later.

To allow a 40G Ethernet signal to be connected to a current DWDM system with an OTM0.3 client (sorry, but no OC768) a Media Converter would need to provide the PCS transcoding function, bidirectional 40GBase-R MAC monitoring along with an OTU3 framer with G709 RS FEC, this could be realized using one of the latest FPGA's, especially since the standardization for the 1024B/1027B transcoding is still pending in the ITU, or a combination of FPGA and 40G FEC device. On the optical interfaces the 40G Ethernet would be supported via a 40GBase-xR4 and a XLAUI interface whilst on the OTU3 side a SFI-5.1 or SFI5.2 interface would connect to 300pin MSA 40G VSR modules to provide the OTM0.3 signal.

Eventually the functions of transcoding and mapping the 40G Ethernet into the OTU3 would get absorbed into new DWDM line cards, however they would also need to resolve the issue of having a common optical interface which would support OC768, OTM0.3 and 40GBase-xR4. As mentioned above one proposal would be to standardize a mapping of OC768 and OTU3 framing over XLAUI using a form of byte interleaving, this however would not be backward compatible with existing DWDM systems! Perhaps a better solution would be the standardization of serial 40G Ethernet (e.g. 40GBase-VSR from an earlier blog) with both a XLAUI and SFI-5.2 system interface, as supported on the new 40G/100G CFP MSA. We think serial 40G Ethernet would provide the best solution for the LAN-WAN interconnect and it is great to hear that the IEEE are going to restart the discussion at their next meeting in April.

Whilst the role for Media Converters to interface to DWDM system may be just stop gap solution a similar function would also be necessary to allow 40GBase-CR4/SR4 interfaces to be converted to future 40GBase-LR4. This would allow large campus networks to take advantage of the 10KM reach and any possible inclusion of 64B/66B inband FEC – let's see.

40G Ethernet Resource Center
http://www.40gethernet.com (http://www.40gethernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

——-

In both part 1 & 2 of this blog we've assumed some (if not too much) knowledge of the G709 standards. Rather than try to go into detail on this subject we think it best to refer people to the excellent OTN Tutorial by Tim Walker which is available "free" from

the ITU website.
http://www.itu.int/ITU-T/studygroups/com15/otn/OTNtutorial.pdf (http://www.itu.int/ITU-T/studygroups/com15/otn/OTNtutorial.pdf) or just search for "ITU OTN Tutorial".

Oh yes and some more of those dreaded TLA's :

DWDM : Dense wave Division Multiplex
FEC : Forward Error Correction
IaDr : Intra Domain interface (a type of interface between DWDM systems)
IrDr : Inter Domain interface (another type of interface between DWDM systems)
POS : Packet over SONET (a method of mapping TCP/IP into SONET)
SFI-5.1 : SERDES to Framer Interface Level 5, interface standard from the OIF for 40G using 17bit datapath
SFI-5.2 : SERDES to Framer Interface Level 5, interface standard from the OIF for 40G using 5bit datapath
VSR : Very Short Reach (VSR2000 means 2KM reach on SMF)

Posted in Technology | Leave a Comment »

40G Ethernet over DWDM : Part1

April 12, 2009
Even though 40G Ethernet is classed as a technology that will be found in the LAN or SAN, at some point in time it will be necessary to transport 40G Ethernet over DWDM links in order to connect remote sites. Typically for links of more than 80km this will mean interfacing to DWDM systems built around the ITU's G709 standard for OTN, which specifies a digital wrapper with FEC. I had planned to write this as one blog but due to the torrid history of aligning the 10Gbps Ethernet data rates with the associated 10Gbps Telecom data rates I decided to split it into two parts, the first covering how 40G Ethernet is mapped and the second on the challenges faced by early adopters.

Looking at how 10G Ethernet/Fiber Channel is transported over a 10.7Gbps OTU2 DWDM link the first problem we hit is that the OTU2 was initially specified to carry SONET/SDH traffic (9.96Gbps) and does not have the payload capacity to transport 10G Ethernet (10.3Gbps) and 10G FiberChannel (10.5Gbps). The Telcom folks had proposed a 10GBase-W which was basically a rate shaped 10G Ethernet mapped into a OC192 SONET signal, however the response from the data community to rate shaping could be politely summed up as "you want us to do what…..!"

So regardless of the 10GBase-W, the initial market solution was to "overclock" the OTU2 frame (OTU2e) so that it had the extra capacity (something that can be done for OTN but not for SONET).This however has spawned four different overclocked datarates, two for 10G Ethernet and two for 10G Fiberchannel depending on how the data is mapped.

Although the use of OTU2e overclocking is very wide spread it is acknowledged now as an undesirable solution for the WAN as it makes multiplexing these signals into a higher rate 40G signal very difficult (needs an overclocked OTU3) and pretty much prevents any reasonable form of OTN crossconnect.

For 10G Ethernet, a method has been devised by AMCC (http://www.amcc.com) to map the data into a standard rate OTU2, this is done by making the OTU2 payload slightly bigger (using some unused overhead bytes) and deleting Idle 64B/66B characters in the IPG whilst still being able to transport certain ordered set messages likes local/remote defect. This proposal, which is an extension of the ITU G.7041 GFP based mapping, is now part of the ITU's G.Sup43 and is pretty much accepted as the industries preferred solution (though not optimal). However the solution does not work for the higher datarate of 10.5Gbps Fiberchannel.

Despite the problems lessons were learned, and in the definition of the ITU's 100G OTN standard a landmark was reached where, rather than a telecom signal like SONET being assumed payload, the proposed OTU4 being specified by ITU is based on the 100G Ethernet data rate (103.125Gbps incl. PCS) – so no more need to overclock – hopefully.

However what about 40G Ethernet (41.25Gbps line rate), since 43Gbps OTU3 is optimized for the SONET 39.8Gbps OC-768, do we need to overclock the OTU3(e.g 45Gbps)? Well the good news is that it's not necessary and that is thanks again to the work done in the IEEE (kudos to Steve Trowbridge at Alcatel-Lucent) where a method of transcoding the 40G Ethernet 64B/66B signal into a 1024B/1027B signal (original proposal was for 512B/513B) has been devised which "just" fits inside the payload of the OTU3 – hurray!

The proposed transcoding method does not support full IPG transparency but as long as vendors stick to the 40G PCS coding rules it will work. After all that said it is still possible to map a 40G Ethernet into an overclocked OTU3 using the same method as used in 10G overclocking – but lets hope not.

For the moment the 40G Ethernet transcoding is still just a proposal with the IEEE passing to initial work done over to the ITU's standards sub committee Study Group 15 . The most likely next step is that it will be included in the next update of G.sup43.

So, even though nearly all of the technical details have been sorted out it will be a while before native 40G Ethernet ports appear on DWDM system which will cause extra challenges to early adopters. Part 2 of the blog discusses these challenges and some possible solutions.

40G Ethernet Resource Center
http://www.40gethernet.com (http://www.40gethernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

—-

In both part 1 & 2 of this blog we've assumed some (if not too much) knowledge of the G709 standards. Rather than try to go into detail on this subject we think it best to refer people to the excellent OTN Tutorial by Tim Walker which is available "free" from the ITU website.
http://www.itu.int/ITU-T/studygroups/com15/otn/OTNtutorial.pdf (http://www.itu.int/ITU-T/studygroups/com15/otn/OTNtutorial.pdf) or just search for "ITU OTN Tutorial".

(http://www.40gethernet.com)

(http://www.40gethernet.com)

Posted in Technology | 4 Comments »

40G Ethernet Network Interface Cards

April 4, 2009
Back in early 2007 when 100G Ethernet was being discussed several companies including SUN (http://www.sun.com), Intel (http://www.intel.com) and Broadcom (http://www.broadcom.com) raised the need for 40G Ethernet, one of the main drivers for a mid way point between 10G and 100G was the networking requirements for high end server NIC cards. The advocates for 40G Ethernet pointed out that the bandwidth gap between future 10G and 100G was too great and if a 40G Ethernet was not standardized other technologies such as Infiniband would fill the gap.

http://www.ieee802.org/3/hssg/public/apr07/hays_01_0407.pdf (http://www.ieee802.org/3/hssg/public/apr07/hays_01_0407.pdf)

Fast forward two years and we find the market for 10G Ethernet NIC cards, and associated CNA's, growing quickly with around 10 suppliers based on chips from Intel & Broadcom as well as several smaller niche suppliers. Virtually all offer 2x10GBE with PCI-E and many are now offering SFP+ and 10GBase-T (Cat6e!!).

One of the big drivers for the shift to 10G on the server has been the rapid adoption of FCoE in the Data Center. The reason why you need 10GBE is because your combining both the bandwidth of the Ethernet NIC and the Fiber Channel HBA (1,2,4 or 8Gbs), a NIC card that supports both TCP/IP and FCoE is called a Converged Network Adaptor. If, as seems to be the case, that FCoE and the associated Data Center Ethernet (DCE) is here to stay then its only a matter of time before the combined traffic of TCP/IP and FCoE on servers exceed 10G and force a move to 40G Ethernet.

If the competition in the 10G Ethernet market was not enough Mellenox (http://www.mellanox.com), a leader in Infiniband, released both ConnectX silicon and adapter cards with 2x40Gbps data rate for Infiniband (along with associated switch silicon) in late 2008. Mellanox also offers CNA's supporting 10G Ethernet (TCP/IP, FCoE) and 10/40G Infiniband which can auto sense the type of interface they are connected to, this they call Virtual Connect Protocol

According to David Gross's blog in SeekingAlpha, the percentage of 40G revenue for Mellanox reached over 30% of their total in 4Q2008, so the demand seems really there!

http://seekingalpha.com/article/117329-mellanox-sees-rapid-increase-in-40g-revenue (http://seekingalpha.com/article/117329-mellanox-sees-rapid-increase-in-40g-revenue)

So, not only has the market for 10G NIC/CNA cards grown faster than expected so did the potential technology alternative, 40G Infinband. It seems a long time since the market has outpaced predictions in the communication market segment and shows that even in todays current downturn there are still bright spots.

In the past realizing such high end adapter cards, such as 40GBE CNA, would require an ASIC but with new high end FPGA's from Altera (http://www.altera.com) and Xilinx (http://www.xilinx.com), which both support 10G serial links it should be possible to implement a 40G Ethernet PCI-E to XLAUI adaptor using off the shelf IP cores for both the PCI-E and the 40G MAC/PCS (see the 40G Ethernet Resource Center website for list of 40G Ethernet and related FPGA&IP vendors).

Infact, given that there are no 40G Ethernet switch ports out there (yet!) a multi-function CNA that could operate in either 40G Infinband mode or 40G Ethernet mode would seem to make most sense as both markets can be addressed. Let's see (but on this one I doubt we will have to wait too long!)

40G Ethernet Resource Center
http://www.40gethernet.com (http://www.40gethernet.com)
http://twitter.com/40GEthernet (http://twitter.com/40GEthernet)

——-

There's allot of TLA's above so here's a break down:

CNA : Converged Network Adapter, basically an Ethernet NIC that supports both the transport of standard TCP/IP and FCoE.
DCE : Data Center Ethernet, an Ethernet network/switch that can support "lossless" protocols ,such as Fiber Channel or Infiniband, as well as the existing TCP/IP with can tolerate the odd dropped packet.
HBA : Host Bus Adapter, basically a Fiber Channel NIC.
HCA : Host Channel Adaptor, an Infiniband NIC.
NIC : Network Interface card, typically Ethernet.
TLA : Three Letter Abreviation!

Posted in Technology | 5 Comments »