Note on Topology Processor Development

– Andi, Bruno, Eduard, Uli, Volker W. –

05/01/2012 14:27

Table of Contents

0 Prologue

Currently we do have neither a requirements document nor specifications. This document is meant to prepare the required paperwork, that's due very soon. It is important that some basics of the design are understood and agreed within the community – and in particular between the CMX- and the topo designers – before effort is invested in proper documentation.

1 Development line

Current plans:
GOLD – prototype1 (Q2 2012) – prototype2 (Q4 2012) – production modules (2013) – additional production modules for 2018.

Details of 2nd prototyping phase will depend on footprint compatibility of large FPGA devices. Additional 2018 modules are meant to process increased data volume from FEXes. They might be just further copies of 2013 modules, unless additional backplane bandwidth turns out to be required for routing data into final processing stage. Current baseline does not make use of backplane connectivity but rather relies on zero-latency forward duplication at upstream modules (CMX), should more than one topology processor module be required.

2 Requirements

Bandwidth requirement had been estimated at time of 2011 L1Calo Stockholm meeting (after correction for 16 octants):
"Preliminary data formats for jet and cluster processors: 96*(14*8+32) bit @40 MHz => 553Gb/s
For muon data assume data volume comparable to current MUCTPI input: 16*13*32 bit @ 40MHZ =>267Gb/s", i.e. 820 Gb/s total. That's under the assumption that all 96 data bits from each CP/jet slot are fully transmitted. No realistic muon data rate is known. Add some ε for energy sums.

The Stockholm presentation (June 2011) is found here:
http://www.staff.uni-mainz.de/uschaefe/browsable/Meeting/2011/2011-06-27-Stockholm/Uli-2011-06-27-corrected.pdf .
See slide 6 for the block diagram, slide 7 for bandwidth estimate.

3 Design principles – do it simple

The topo processor prototype tries to get closest possible to the final production modules. While GOLD is toying with connectors and mezzanines, allowing for maximum flexibility when testing data transmission, the prototype will be strictly purpose-built. While the initial prototype is being developed, the GOLD will continue to serve as a platform for on-going exploration of new technologies (opto links), and firmware development.

Due to recent concern about long high-speed lanes across connectors, PCBs, and backplanes, the topo processor prototype will be optimised for simplicity and integrity of electrical interconnects. In short: invest in active components, rather than in complex PCBs.

The processor module will, to the extent possible, make use of highest-density optical and optoelectronic components. Any duplication required will be done upstream. Any electrical high-speed traces will be routed on shortest possible links onto the closest FPGA. The current baseline scheme will even give up on the long-favoured scheme of electrical duplication on the outputs of the opto-electrical converters. When using highest density o/e converters, trace lengths would most likely need to be extended considerably, so as to fit fan-out chips. The module will make use of mid-board optoelectronic components only, interfaced to backplane and front panel via fibre ribbon pigtails.

So as to extract maximum data rate from the CMX, including data duplication, the CMX should provide ample output bandwidth, at both 6.4Gb/s and 10Gb/s line rate. The topo processor prototype is meant to be available for tests with 10Gb/s capable CMX technology demonstrators and/or prototypes from summer 2012. It is assumed that the CMX can easily, and at excellent signal integrity, communicate at both 6.4Gb/s and 10Gb/s to the topo processor, if the CMX shares the design principles described here.

The topo processor prototype is meant to be a full scale, full-function processor that can be used in conjunction with the CMX (and prototype) from day one. Therefore it seems appropriate to reduce the long discussed CMX input bandwidth to what's required for self-test diagnostics, and rather rely on the topo processor (and prototype) for any topology processing. This simplification of the CMX scheme should allow making additional bandwidth available on CMX optical outputs rather than inputs. For reason of signal integrity, on-FPGA signal duplication is preferred over electrical or passive-optical splitting.

The topo processor will employ two high-performance, high-bandwidth processor chips for real-time processing. They are indicated A and B (largest available Virtex-7) in Figure 1. With the devices expected to be available by spring 2012, for prototype1, the total MGT input connectivity would probably be 112 links, i.e. up to 0.9 Tb/s, if 10Gb/s links are fed in. With 6.4Gb/s links the total bandwidth would amount to 0.57Tb/s. The use of more advanced FPGAs on a second prototype and on production modules would boost the bandwidth by a factor of 1.4 or even 1.7 (80, as presented in Stockholm, or 96 links per chip). Control FPGA C (small Kintex-7) is mainly concerned with interfacing to the module controller via Ethernet.

The opto-electrical converters are assumed of miniPOD type. They are all mounted mid board and are connected to either front or back panel with octopus cables. MicroPOD devices might be used if they are available on the market in time, and if they turn out to be the baseline for FEX development. While the preferred interconnect is optical through-backplane connection, the mechanical properties of the blind-mate connectors have not yet been sufficiently explored. Dependent on future results from the GOLD, either back or front panel connectors will be chosen. The decision can be taken after PCB production. The use of mid-board transceivers requires an additional stage of fibre-optical connection, with the associated unavoidable signal loss.

The FPGA-to-FPGA interconnect will be done with parallel differential links. There is a penalty of approx. 1 tick of latency, if both FPGAs need to be joined together this way. Almost all bandwidth between FPGAs A and B will be made available for the real-time data path. It is anticipated to reserve one bank each for read and write control buses to C. Also, some rather narrow, parallel port should be reserved on both processor FPGAs for low latency connection to the CTP. Thus the possibly small number of trigger data on the critical path can be transmitted at minimum latency.

There is currently near-to-nothing known about the muon input links. It is anticipated that the muon signals will come in at 6.4 or 10 Gb/s optically.

For reason of latency optimisation, a symmetric dual-processor scheme (fig.1) is chosen for the topology processor.

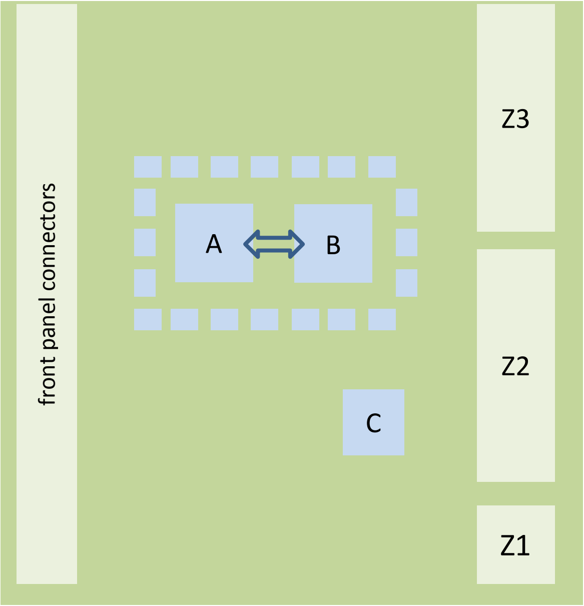

Figure 1: Floor plan topo module (not to scale)

Optical output bandwidth to the CTP (MGT links and low latency path) will be made available from both of the FPGAs. With the total input bandwidth provided, it is anticipated that most algorithms can be run in just one of the FPGAs, minimising overall latency. However, for more complex / high input volume algorithms it will be possible to cross chip boundaries at a latency penalty of about one clock tick.

It is anticipated to run the additional low latency, low bandwidth CTP links optically as well, so as to avoid any signal integrity issues. DC-balance coding would be implemented in the FPGA fabric.

For DAQ and ROI links MGT output lines from the two processors will be used. SFP / SNAP12 / miniPOD devices are considered for e/o conversion.

It should be noted that quite some ancillary circuitry and firmware will be required. It is anticipated that the ATCA specific I2C/power control scheme can be taken over from RAL. It is not currently decided whether L1Calo standard CAN monitoring will need to be implemented. Ethernet based control should be adapted from Bristol (IPBus). It should be noted that 7-Series FPGAs do not come with integrated MACs. Therefore the VHDL code would have to assemble Ethernet rather than IP packets. Due to availability of PCIe endpoints / root complex in 7-Series FPGAs the technically superior approach might be a standard PCIe-to-Ethernet device. However, it isn't currently known what level of service firmware would be required.

Alternatively the use of an ARM based Zynq processor is considered. It contains Ethernet and CAN interfaces. This approach would require a local operating system to be deployed.

4 Component availability and timeline

It is difficult to precisely predict availability of new FPGA and high-density optoelectronic devices. However, it can be safely assumed that both engineering samples of moderately sized Virtex-7 devices and high-density o/e receivers will be available at a timescale appropriate for topology processor prototype1. These components will guarantee high-speed link operation at 6.4 Gb/s, 10Gb/s, and above. In case there were problems with FPGA supply, this would delay the prototype production. There is no fall-back option to Virtex-6. There might be considerable overlap of prototype1 tests and prototype2 design work, if V7 availability were severely delayed. Current assumption is that higher capacity FPGAs, expected to arrive in late 2012 or early 2013, will be footprint compatible to initial, smaller ES devices. In this case full capacity modules can be built and tested (to a large extent) before FPGAs for the production modules become available.

For the o/e components, time line seems to match the project. Though microPOD is not expected to be available in time, for miniPOD the situation looks promising. However, availability will need to be re-assessed in spring 2012. In case of severe delays, prototype1 would need to be designed with either conventional, standard footprint devices, as used on the GOLD and its mezzanines, or a mezzanine connector would have to be re-introduced. Either approach would reduce design density and increase high-speed trace lengths. Also, total optical bandwidth might be below the figures quoted above.

5 Module and system tests

The topo processor prototype will initially be tested in Mainz, along with the BLT. The GOLD will be adapted to serve as a generic source and sink module up to 6.4Gb/s line rate. For rates beyond the capabilities of BLT and GOLD, fibre loopback on the prototype will be the only test setup available initially.

It is anticipated that the test bench available in the L1Calo CERN test lab will be upgraded to allow for standalone tests, and tests including the CMX prototype. Availability of CTP prototypes will need to be discussed with the CERN group.

6 Cost and funding

Topo processor prototype1 is funded from available resources. Cost is dominated by FPGAs and PCB production. All multi-Gigabit input connectivity and a considerable fraction of output connectivity of the FPGA processors will be routed to transceiver sockets. Here cost will be incurred only for those transceivers and fibre bundles actually plugged into the module. We pay for what we need only. Component tests might be done with a partially equipped module. System tests will require a higher level of connectivity. Optic and optoelectronic components will be re-used in future test benches.

7 Technicalities

Data integrity is best preserved if minimising trace length and vias on the trace. Therefore the o/e converters will be mounted close to the FPGAs. However, there is competition for real estate between o/e converters and POL voltage converters.

Vias on MGT traces will be avoided by optionally connecting differential links in inverse polarity. Actual link polarity is programmable on the FPGAs and on the opto transceivers (check miniPOD).

Consider inverse polarity option also on parallel buses. This would make VHDL code slightly more cumbersome, but should be beneficial for signal integrity and board complexity.

Ideally one would run MGT links on top and bottom only. However, due to breakout density, that might be impossible. Micro or blind vias might have to be used to avoid stubs when connecting to inner layers. Options need to be discussed with PCB manufacturers and within the community. MGT clocks also deserve signal integrity focused attention. Check whether clock polarity matters. To avoid phase differences of recovered clocks, MGT reference clocks on all devices / all quads might possibly have to be of *same* polarity.

Though pin swapping would be possible across banks, there are limitations imposed on functionality of groups of pins. Therefore it is decided to connect full banks of processor A to full banks of processor B. No mixing of banks. All banks are 1.8V, LVDS only.

Due to the required dense packing, thermal considerations are required.

Standard FR4 laminate will be used, all micro strips facing a ground plane, all strips facing at least one ground plane and one continuous power plane.

Quote Xilinx (UG483):
"Substrates, such as Nelco, have lower dielectric loss and exhibit significantly less attenuation in the gigahertz range, thus increasing the maximum bandwidth of PCBs. At 3.125 Gb/s, the advantages of Nelco over FR4 are added voltage swing margin and longer trace lengths. At 10 Gb/s, a low-loss dielectric like Nelco is necessary unless high-speed traces are kept very short."