Cray Gemini Interconnect

Technical University of Munich
Parallel Programming Class of SS14
Denys Sobchyshak
1. Introduction
2. Overview
3. Architecture
4. Gemini Blocks
5. FMA & BTA
6. Fault tolerance
7. Software stack
8. Interesting to know
9. References
Introduction

- Torus interconnect is a network topology for connecting processing nodes in parallel computer system.
- Gemini is a network for Cray's supercomputer systems first used in 2010
- Successor of Gemini is an Aries interconnect network first used around 2013
Introduction

Most known example of Gemini interconnect based supercomputer currently occupying 2\textsuperscript{nd} place in TOP500:

*Titan - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x*
Overview

- Develops the highly scalable Seastar design used to deliver the 225,000 core Jaguar system, improving network functionality, latency and issue rate
- Novel system-on-chip (SoC) design to construct direct 3D torus networks that can scale to in excess of 100,000 multi-core nodes
- Capable of tens of millions of MPI messages per second
- Each hybrid compute node is interfaced to the Gemini interconnect through HyperTransport™ 3.0 technology. This direct connect architecture bypasses the PCI bottlenecks inherent in commodity networks and provides a peak of over 20 GB/s of injection bandwidth per node
- The proven 3-D torus topology provides powerful bisection and global bandwidth characteristics as well as support for dynamic routing of messages
Architecture

- Cray XT 3D Torus Network

- Each SeaStar router provides one node of 3D Torus, Gemini provides two.
Gemini Blocks

Each Gemini ASIC provides two network interface controllers (NICs), and a 48-port “Yarc” router. Each of the NICs has its own HyperTransport 3 host interface, enabling Gemini to connect two Opteron nodes to the network. This 2 node building block provides 10 torus links, 4 each in two of the dimension (‘x’ and ‘z’) and 2 in the third dimension (‘y’), as shown in previous slide. Traffic between the two nodes connected to a single Gemini is routed internally. The router uses a tiled design, with 8 tiles dedicated to the NICs and 40 (10 groups of 4) dedicated to the network.
**FMA & BTA**

- **Fast Memory Access (FMA)** is a mechanism whereby user processes generate network transactions, puts, gets and atomic memory operations (AMO), by storing directly to the NIC. The FMA block translates stores by the processor into fully qualified network requests. FMA provides both low latency and high issue rate on small transfers.

- **The Block Transfer Engine (BTE)** supports asynchronous transfer between local and remote memory. Software writes block transfer descriptors to a queue and the Gemini hardware performs the transfers asynchronously. The BTE supports memory operations (put/get) where the user specifies a local address, a network address and a transfer size.

- **To put it simply:** FMA is used for small transfers and BTE for large. FMA transfers are lower latency. BTE transfers take longer to start, but once running can transfer large amounts of data (upto 4GB) without CPU involvement.
Fault tolerance

- Gemini is designed for large systems in which failures are to be expected and applications must run on in the presence of errors. Each torus link comprises 4 groups of 3 lanes. Packet CRCs are checked by each device with automatic link level retry on error. In the event of the failure of a link, the router will select an alternate path of adaptively routed traffic. Gemini uses ECC to protect major memories and data paths within the device.

- Error-correcting code memory (Error Checking & Correction, ECC memory) is a type of computer data storage that can detect and correct the most common kinds of internal data corruption.

- A cyclic redundancy check (CRC) is an error-detecting code commonly used in digital networks and storage devices to detect accidental changes to raw data. Blocks of data entering these systems get a short check value attached, based on the remainder of a polynomial division of their contents; on retrieval the calculation is repeated, and corrective action can be taken against presumed data corruption if the check values do not match.

- Note: ECC memory is a common thing even in less complex systems, for example a simple rack DIMM (Dual In-line Memory Module, think of it as a set of RAM slots) might work with ECC memory circuits.
Software stack

Gemini supports both kernel communication, through a Linux device driver and direct user space communication, where the driver is used to establish communication domains and handle errors, but can be bypassed for data transfer. Parallel applications typically use a library such as Cray MPI or SHMEM in which the programmer makes explicit communication calls.
Interesting to know

- **Performance metrics:** on a quiet network, the end-point latency is 1.0μs or less for a remote put, 1.5μs or less for a small MPI message. The per hop latency is 105ns on a quiet network.

- **Updating SeaStar to Gemini:** The Gemini mezzanine card is pin compatible with the SeaStar network card used on Cray XT5 and XT6 systems allowing them to be upgraded to Gemini. The upgrade is straightforward, each blade is removed and its network card is replaced. There are no changes to the chassis, cabinets or cabling.

- **Future of Cray Interconnects:** Cray sold patterns (10% of whole pattern portfolio), technology itself (starting from Aries interconnect) and key people to Intel in 2012. The agreement excludes Intel from reselling standalone Aries chips to anyone but Cray till 2016.

  “It’s been clear that the important technology trends are starting to intersect, enabling more powerful and energy-efficient systems interconnects by more closely integrating this technology with the processor.” (Cray CEO Peter Ungaro)

  “What will be changing in 2017, and beyond, is that we won’t continue to build system interconnect hardware.” (Cray CEO Peter Ungaro)
References

- http://www.theregister.co.uk/2010/05/25/cray_xe6_baker_gemini/?page=2
- http://www.hpcwire.com/2012/04/25/intel_makes_a_deal_for_cray_s_interconnect_technology/
- http://www.top500.org/system/177975#.U44G6nWSwl9
- http://livingthewiccanlife.blogspot.de/2012/05/aries-march-21-april-20-aries-is.html
- http://sjastro.com/gemini-horoscope/