The AuroraScience Project

F. S. Schifano[1]

[1]University of Ferrara and INFN-Ferrara

November 25-26, 2009

The AuroraScience Project

The AuroraScience project

- formally started in summer 2009
- has technological and scientific goals:
  - study of APE-like architectures based on off-the-shelf processors
  - development of scientific applications
  - study of programming methodologies for multi-core achitectures
- is divided in two main phases:
  - first phase develops a prototype of 20 Tflops and ends in Dec. 2010
  - second phase develops a prototype of $\approx$ 100 Tflops and ...
  - ... is scheduled for Jan. 2011 - Jun. 2012
  - start of second phase will be decided in summer 2010.

The AuroraScience Project Collaboration

- **ECT***
- **INFN** (Fe, Mi, Pr)
- Dipartimento di Fisica Università di Trento
- **DEI**: Dipartimento di Ingegneria dell'Informazione Università di Padova

and Trentino-located institutions:

- **IASMA**: Istituto Agrario S. Michele all'Adige
- **ATreP**: Agenzia Provinciale per la Protonterapia.

AuroraScience: Funding Structure

- The project has been formally approved by “Provincia Autonoma di Trento” (PAT)
- The funding is about 3 M€ (1.5 + 1.5)
- The funding includes:
  - delivering of two prototypes: 20Tflops + 100Tflops
  - several positions for hardware and software development and physics.

AuroraScience: Goals

Scientific goals:

- design and development of 3D network system a la APE (hw and sw)
- porting of scientific applications: mainly LQCD, LBE, ...
- ...but also genomic, medical physics, nuclear physics relevant for the "Trentino" institutions
- study of multi-core programming strategies.

Technological goals:

- use of latest generation of Intel CPUs
- assembly a $\approx$ 20 Tflops machine in 2010
- assembly $\approx$ 100 Tflops machine in 2011

AuroraScience vs Aurora

The hardware design of the machine has been done together with Eurotech before formal approval of the project:

- this allowed to have the machine available just a few months after the official start of the project ✔
- design details are not optimized only for LQCD and have a negative impact on costs ✘

However we think that:

- it is another relevant development in using non-custom CPUs in APE-like systems
- it is an important experience to be used as base for future projects.

The AuroraScience Machine: the processor

The choise of the processor is based on latest generations of Intel CPUs.

The project aim to use two/three versions of Intel processors:

- 4-core Nehalem, 50 Gflops peak double-precision
- 6-core Westmere, 75 Gflops peak double-precision
- 8-core Sandy Bridge $\sim$ 200 Gflops peak double-precision
  (2x cores + AVX 256-bit)

The AuroraScience Machine: the node-card

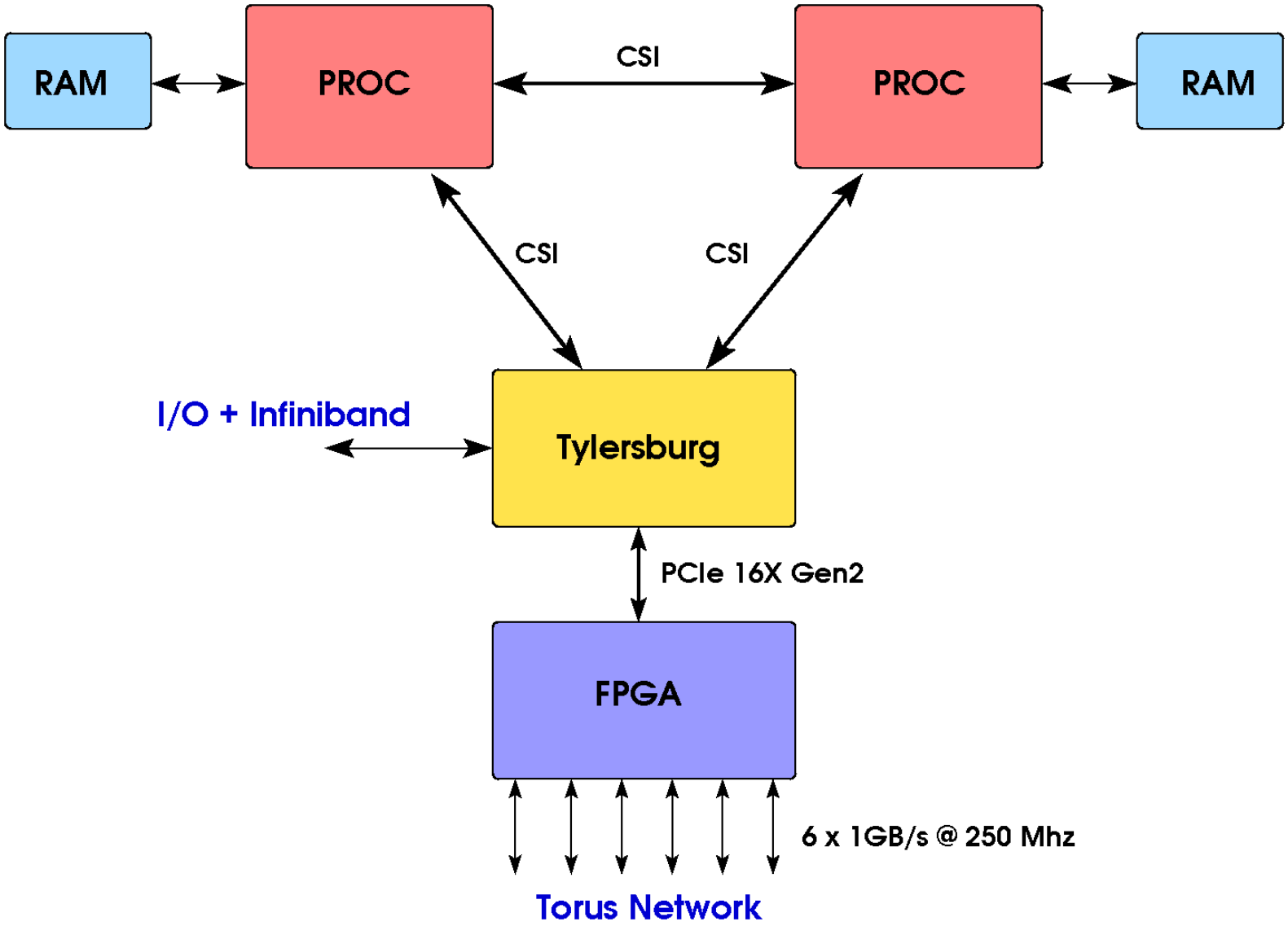
A node-card includes:

- 2 CPUs
- 6 GB of RAM per processor
- 1 Infiniband adapter
- 1 FPGA Altera Stratix IV GX230
- 6 PMC-Sierra quad-link PHYs

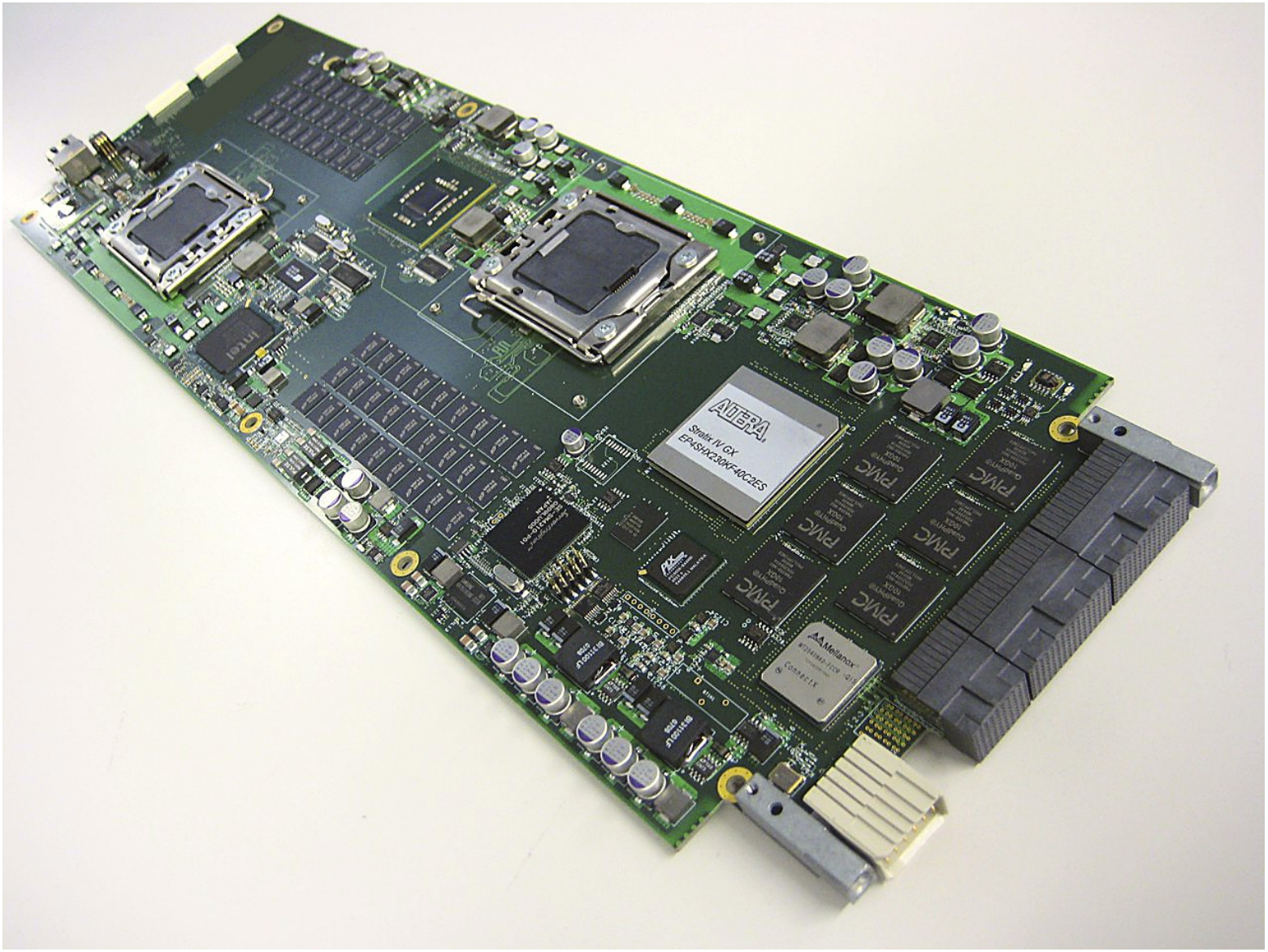
Peak perfomance:

- Nehalem version 100 Gflops
- Westmere version 150 Gflops
- Sandy Bridge version 400 Gflops

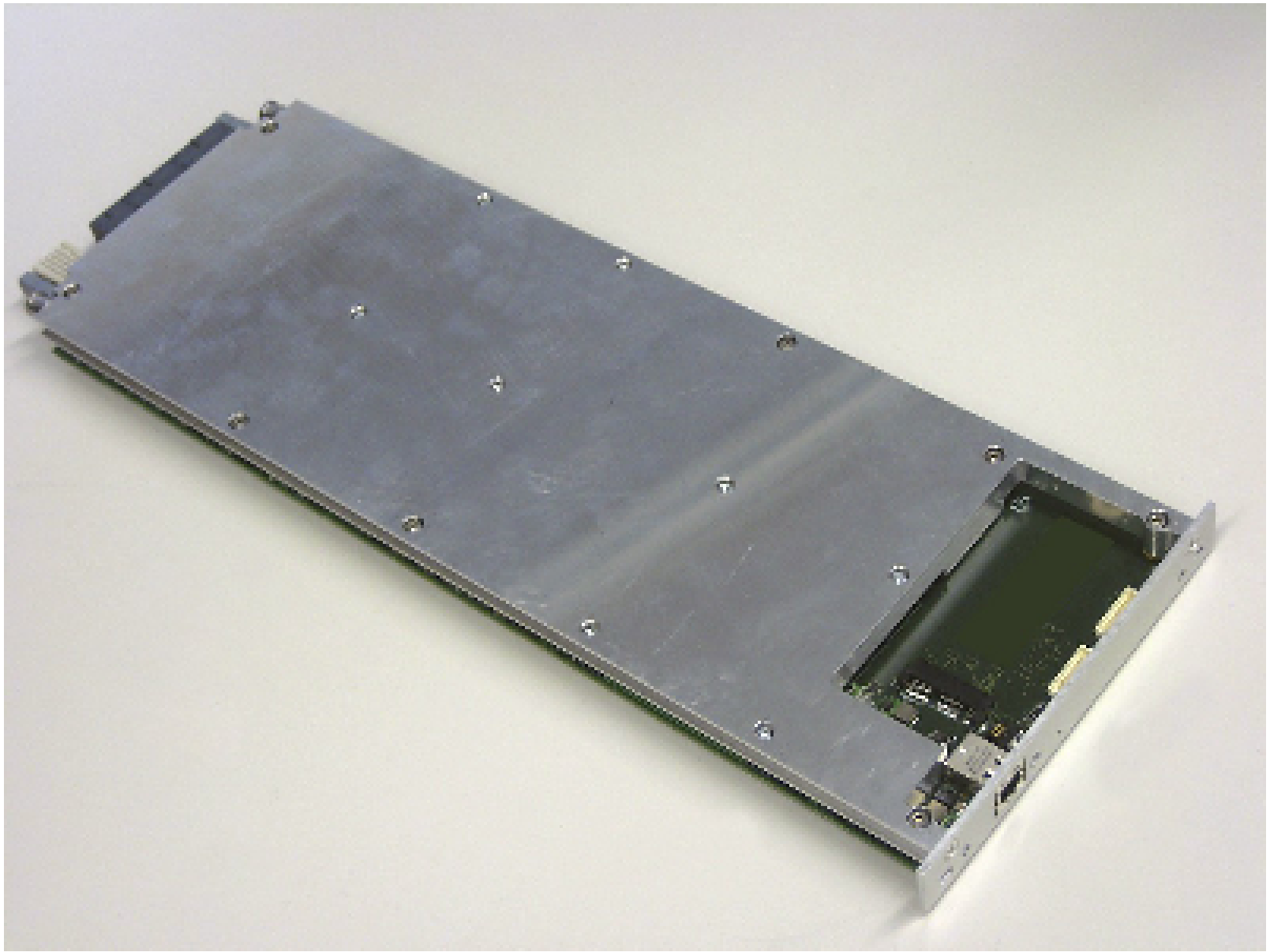
The AuroraScience Machine: the node-card block-diagram



The AuroraScience Machine: the node-card



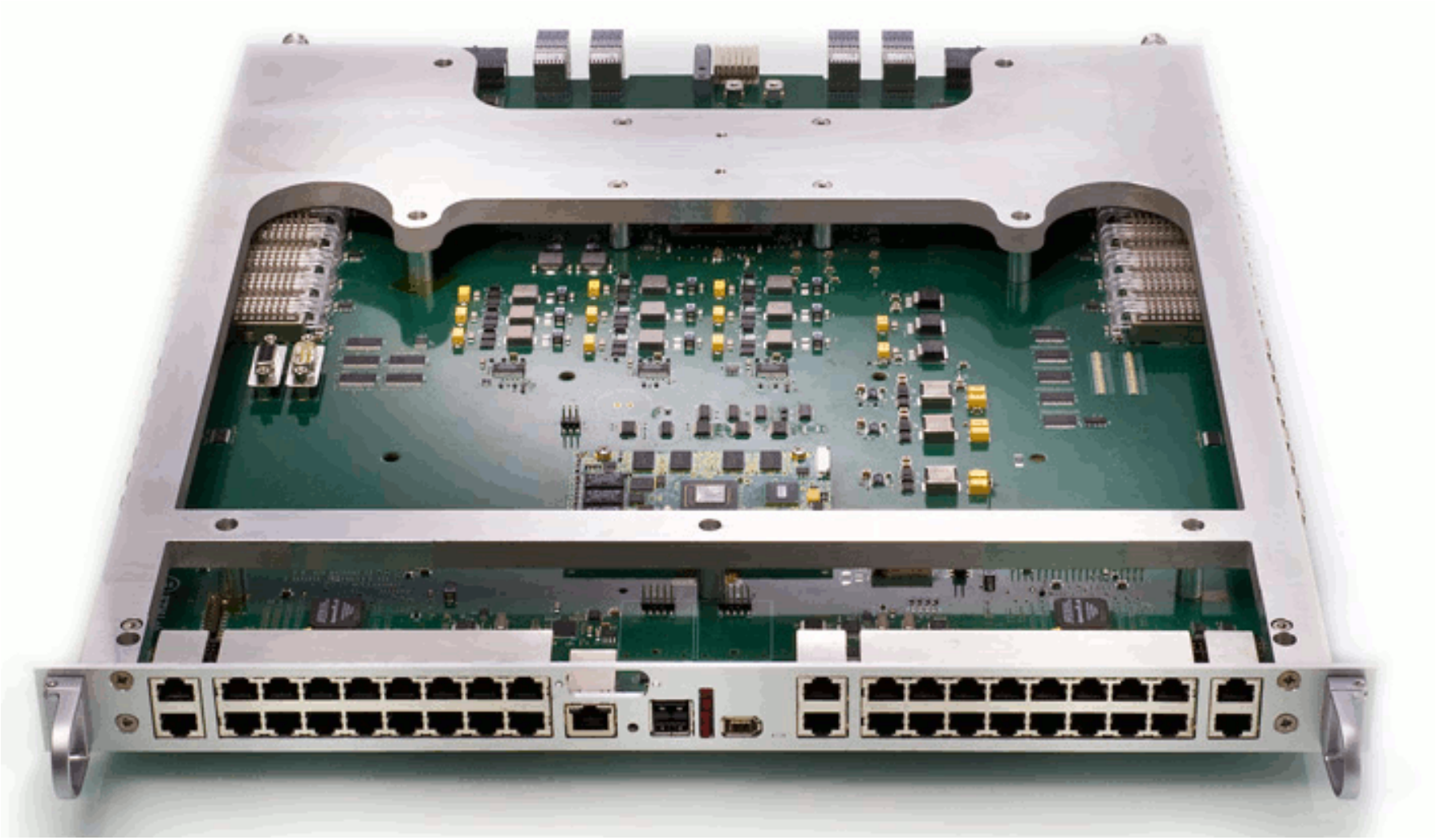
The Aurora Machine: the node-card with cold plate



The AuroraScience Machine: the Crate/Chassis (front)

The AuroraScience Machine: the Crate/Chassis (rear)

The AuroraScience Machine: the root-card



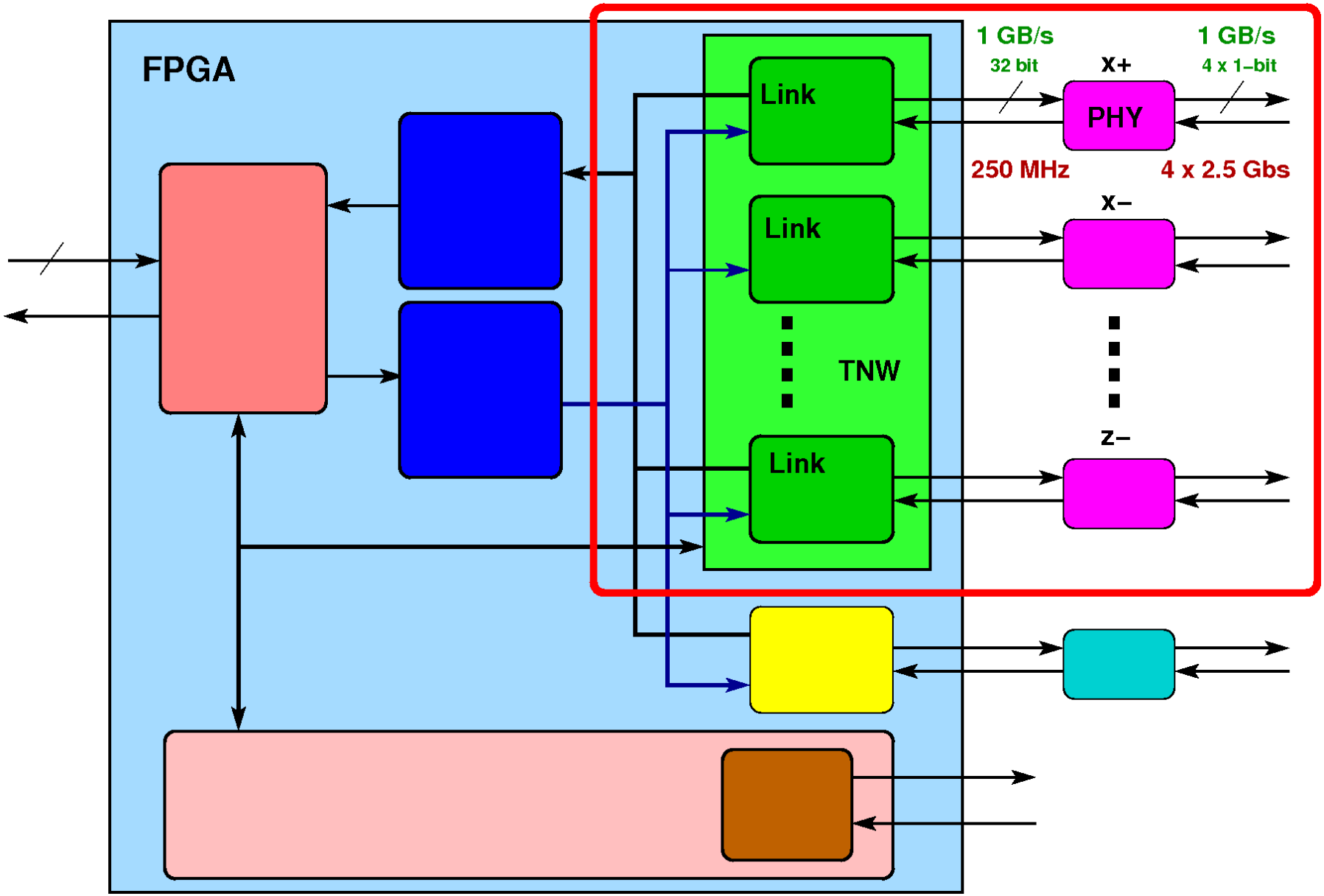
The AuroraScience (3D-torus) network

Aurora nodes are interconnected through a switched-network and a nearest neighbor 3D-torus network a la APE . . . QPACE.

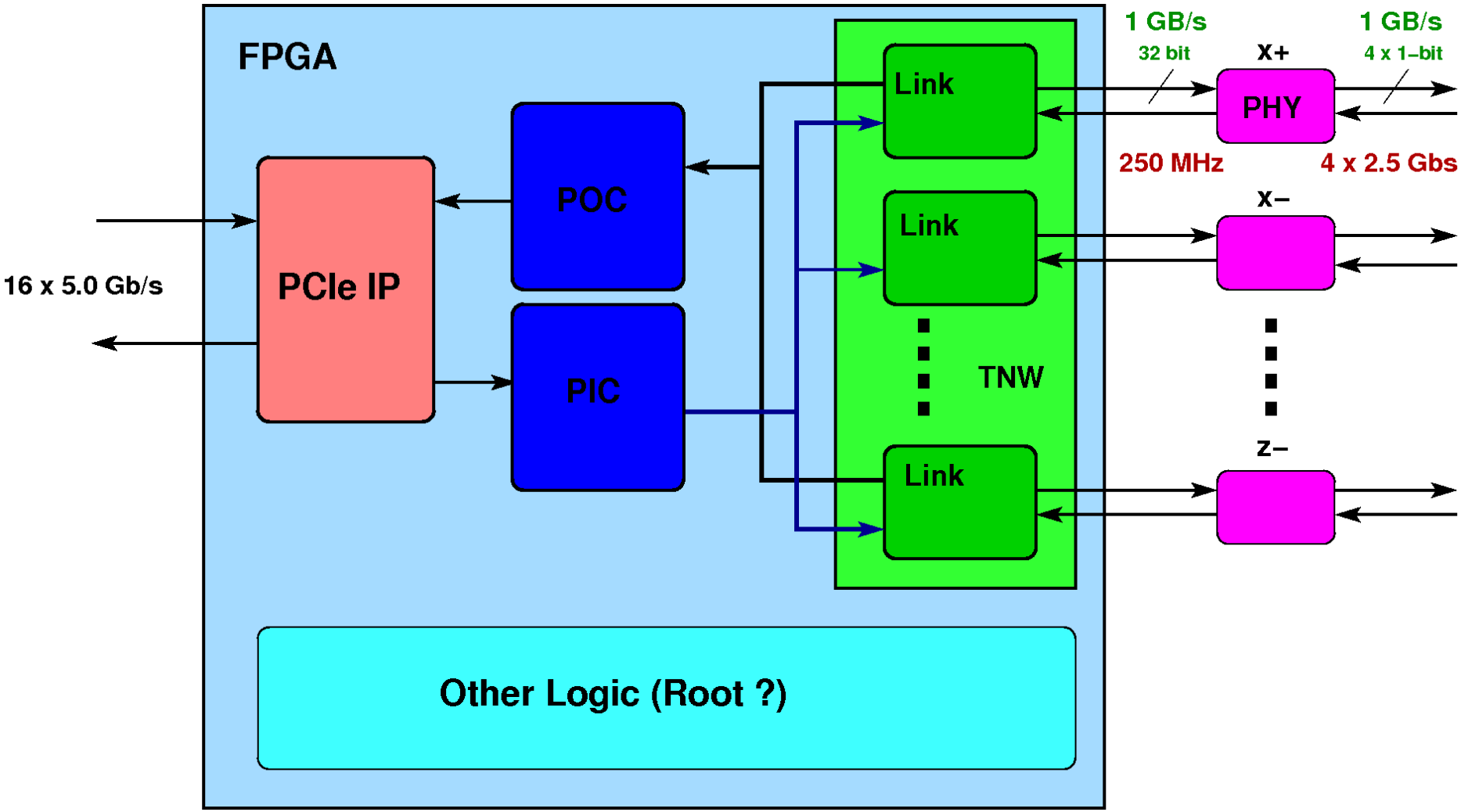
3D-torus:

- processor interface based on standard PCIe Gen2 16x
- network processor implemented on FPGA (Altera Stratix IV GX230)
- routing logic supporting nearest-neighbor communications plus . . . under study
- 6 data-links:
  - physical level based on 10Gbit PMC-Sierra quadPHY
  - bandwidth: 1 GByte / link / direction, latency: $\sim 2 - 3\ \mu$ sec

The QPACE 3D-torus network processor



The AuroraScience 3D-torus network processor



The AuroraScience 3D-torus network processor

Status:

- the link components (fifo, memory, etc) has been ported from Xilinx (QPACE) to Altera environment
- the not-open modules of the QPACE NWP design has been removed
- synthesys and test of the torus with all 6 links under Altera environment has been done
- a 8-lane GEN1 pci-express interface has been implemented and tested
- a basic linux-driver and user low-level library for communications is available
- preliminary communication tests are running

Atnw2 Resouces Occupation

Fitter Summary	
Fitter Status	Successful - Mon Nov 9 10:42:28 2009
Quartus II 64-Bit Version	9.0 Build 132 02/25/2009 SJ Full Version
Revision Name	topHw
Top-level Entity Name	topHw
Family	Stratix IV
Device	EP4SGX230KF40C2ES
Timing Models	Preliminary
Logic utilization	**17 %**
Combinational ALUTs	22,194 / 182,400 (12 %)
Memory ALUTs	120 / 91,200 (< 1 %)
Dedicated logic registers	24,417 / 182,400 (13 %)
Total registers	24918
Total pins	562 / 888 (63 %)
Total virtual pins	0
Total block memory bits	**1,106,865 / 14,625,792 (8 %)**
DSP block 18-bit elements	0 / 1,288 (0 %)
Total GXB Receiver Channel PCS	8 / 24 (33 %)
Total GXB Receiver Channel PMA	8 / 36 (22 %)
Total GXB Transmitter Channel PCS	8 / 24 (33 %)
Total GXB Transmitter Channel PMA	8 / 36 (22 %)
Total PLLs	7 / 8 (88 %)
Total DLLs	0 / 4 (0 %)

Torus box requires 20911 logic-elements (11% of FPGA) and
$\approx$ 1 Mbit of memory (5% of FPGA)

ATNW2 Processor Interface

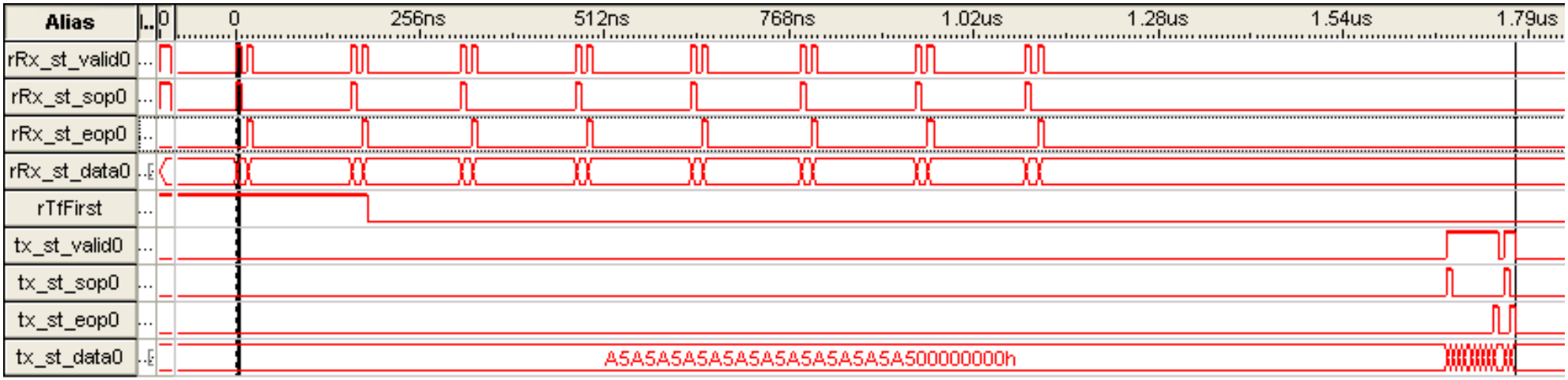
Configuration:

- based on single 8-lane GEN1 IP, 128-bit @ 125 MHz, 2 GB/s
- 1 64-bit (prefetchable) BAR mapping all send fifos
- 1 32-bit BAR mapping all status/monitor/debug/config registers
- CPU writes data to FPGA send-fifos
- include a DMA engine to move data from FPGA to CPU (memWrite)

Improvements:

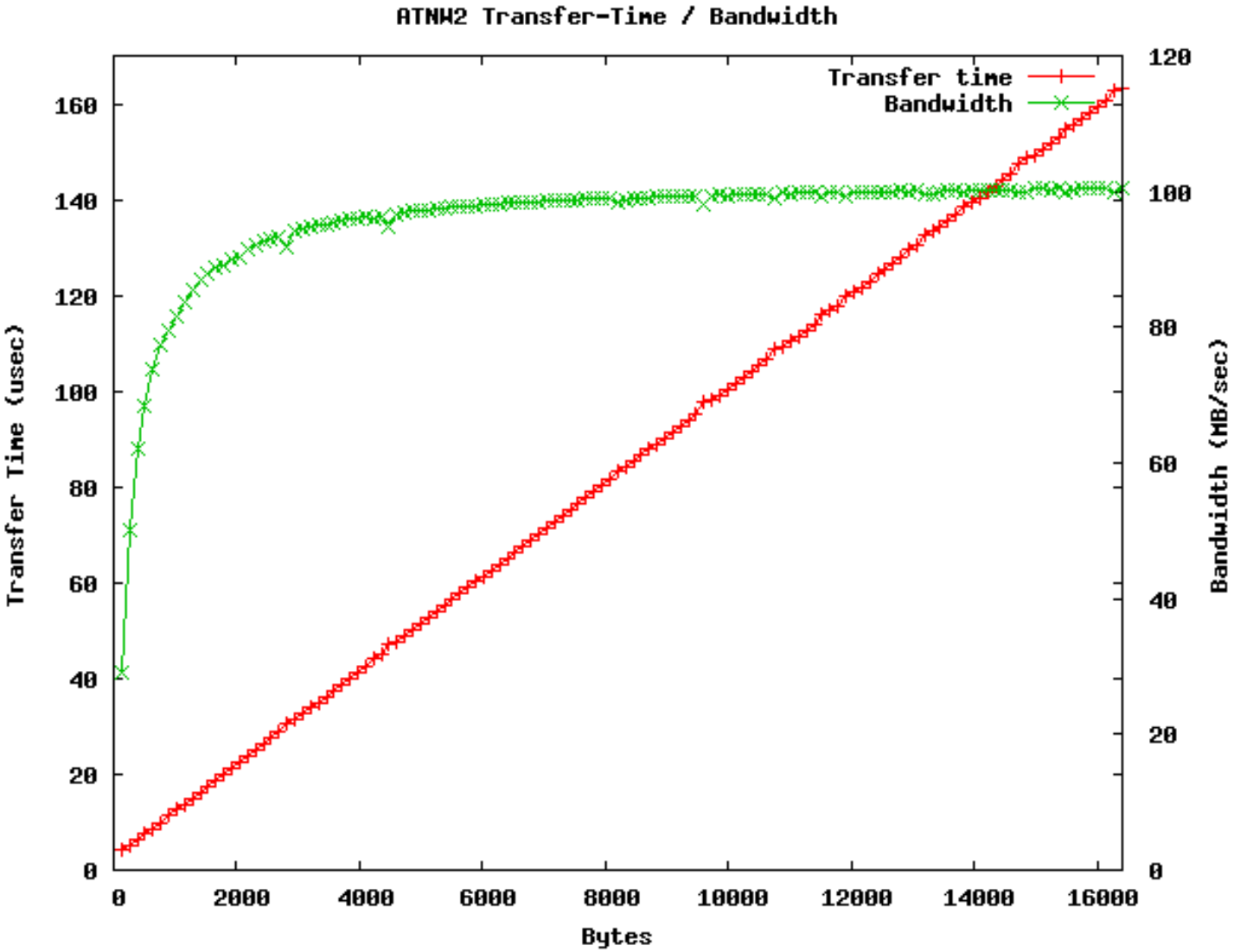
- 2 8-lane GEN2 IP, 128-bit @ 250 MHz, 8 GB/s
- implement DMA engine to move data from CPU to FPGA (memRead)

ATNW2 Throughput Details



- one item (16B) every 152 ns (19 τ @ 8 ns)
- one packet every 1.2 μsec
- fly-time inside TNW, including cable, is 630 ns

ATNW2 Transfer Time



red line is fit by $T(x) = 2.44 + 0.0098 * x$.

The AuroraScience System

Processor	Nehalem/Westmere	Sandy Bridge
Node Card	2 Processors	
	100-150 Gflops*	≈400 Gflops*
	270 W	
Chassis	16 Node Card	
	1.6-2.4 Tflops*	≈6.4 Tflops*
	4.3 kW	
half-Rack	8 Chassis	
	12.8-19.2 Tflops*	≈50 Tflops*
	34.4 kW	
Rack	16 Chassis	
	25.6-38.4 Tflops*	≈100 Tflops*
	70 kW	

* double-precision peak

Relevant installation but not leading-edge for the LQCD community.