Tying GPUs together with APENet+

Roberto Ammendola

Istituto Nazionale di Fisica Nucleare, Sezione Roma Tor Vergata

Workshop CCR 2010, Napoli – 26 Jan 2010

APENet Motivations

The idea was to build a switch-less network characterized by:

- High bandwidth
- Low latency
- Natural fit with LQCD and numerical lattice algorithm.
- Multi hop hardware routing.
- Good performance scaling as a function of the number of processors.
- Very good cost scaling even for large number of processors.

APENet History

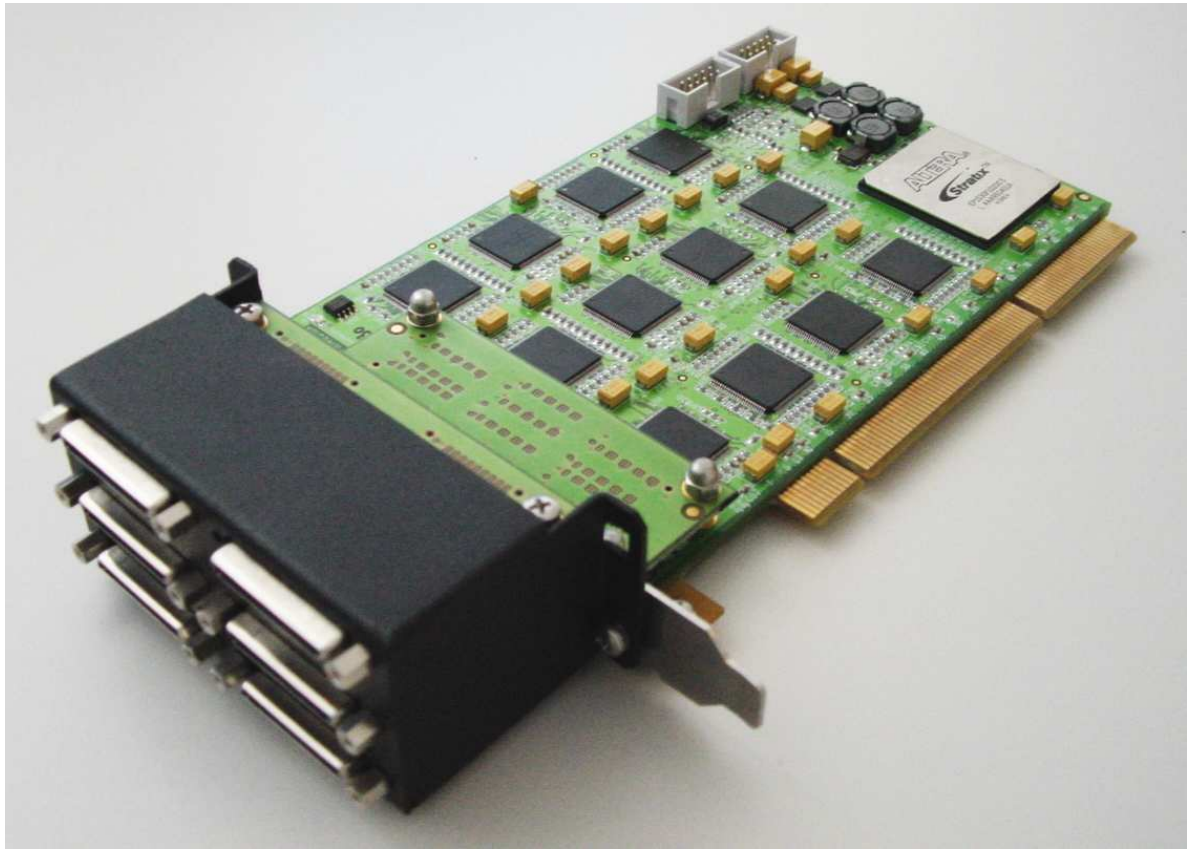
- March – Sep 2003: APENet HW development and production
- Sept 2004: 16 nodes APENet prototype cluster
- March – Nov 2005: 128 nodes APENet cluster
- 2006: hw/sw debugging
- 2007: Production starts

APENet Main Features

APENet is a 3D network of point-to-point links with toroidal topology.

- Each computing node has 6 bi-directional full-duplex communication channels
- Computing nodes are arranged in a 3D cubic mesh
- Data is transmitted in packets (max size 4KB) which are routed to the destination node
- Lightweight low level protocol
- Wormhole routing
- Dimension ordered routing algorithm
- 2 Virtual Channels per receiving channel to prevent deadlocks

The interconnection card



- Altera Stratix EP1S30, 1020 pin package, fastest speed grade
- National Serializers/Deserializers DS90CR485/486, 48 bit 133 MHz

Usage of a programmable device allows possible logic redesign and quick on-field firmware upgrade.

APE128 (2006 $\longrightarrow$)

```
================================================================================
T/V                N    NB     P     Q               Time             Gflops
--------------------------------------------------------------------------------
WR01L2L2      105000    80     8    16             624.81          1.235e+03
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) =        0.0009871 ..... PASSED
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) =        0.0026707 ..... PASSED
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) =        0.0004767 ..... PASSED
================================================================================

Finished      1 tests with the following results:
              1 tests completed and passed residual checks,
              0 tests completed and failed residual checks,
              0 tests skipped because of illegal input values.
--------------------------------------------------------------------------------

End of Tests.
================================================================================
```

- 128 Dual Xeon "Nocona"
  - 3.4 *GHz*
  - 1 GB RAM DDR 333
- $4 \times 4 \times 8$ APENet network
- 1740.8 GFlops ($256 \times 2 \times 3.4$) Peak Performance
- 1235 GFlops (70.9%) Sustained

Submitting job

A job submitting environment has been developed, aware of allowed network topologies on a given machine. A configuration file describes allowed topologies:

```
APE_NET_TOPOLOGY        4 4 8

APE_NET_PARTITION       full            4 4 8   0 0 0
APE_NET_PARTITION       single          1 1 1   0 0 0

APE_NET_PARTITION       zline0          1 1 8   0 0 0
APE_NET_PARTITION       zline1          1 1 8   0 1 0
```

Jobs are submitted with an mpirun derived script:

```
aperun -topo zline0 cpi
```

APENet Limitations

As an hardware developed since 2003 APENet shows some limitations:

- Limited peak bandwidth: now we have Infiniband at 40 Gbps
- Offloading efficient RDMA operation needs
  - dedicated memory banks
  - programmable microcontroller
- Cabling is quite hard due to low wire gauge
- Time consuming on field firmware update

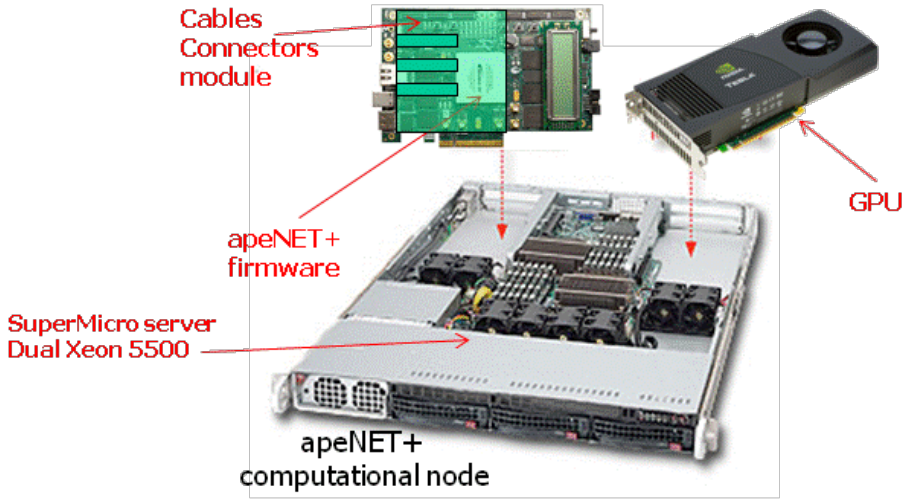
Additional requirements from GPUs adoption:

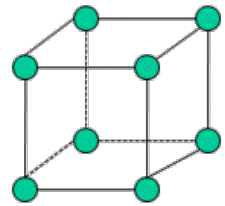
- Balancing increased computing resources with higher performance networks.
- Optimize interaction between interconnect and GPU card.

APENet+ in the GPU framework

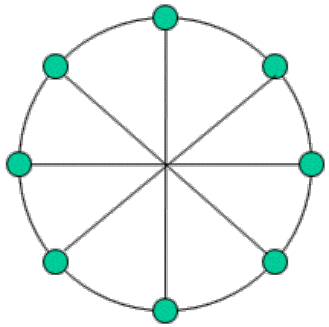
Leg I ($\longrightarrow$ 2Q 2010): evaluation of hybrid CPU+GPU architectures interconnected with commercial network (Infiniband) and custom network.

- CPU+GPU systems gathering
- Development of remote channel technology (QSFP) thanks to the Altera development board with a custom daughter card
- Development of first APENet+ prototypes (4/6) and deployment of a 4 nodes GPU platform for firmware and software validation





Topologia
rings connected



Topologia
1.5D

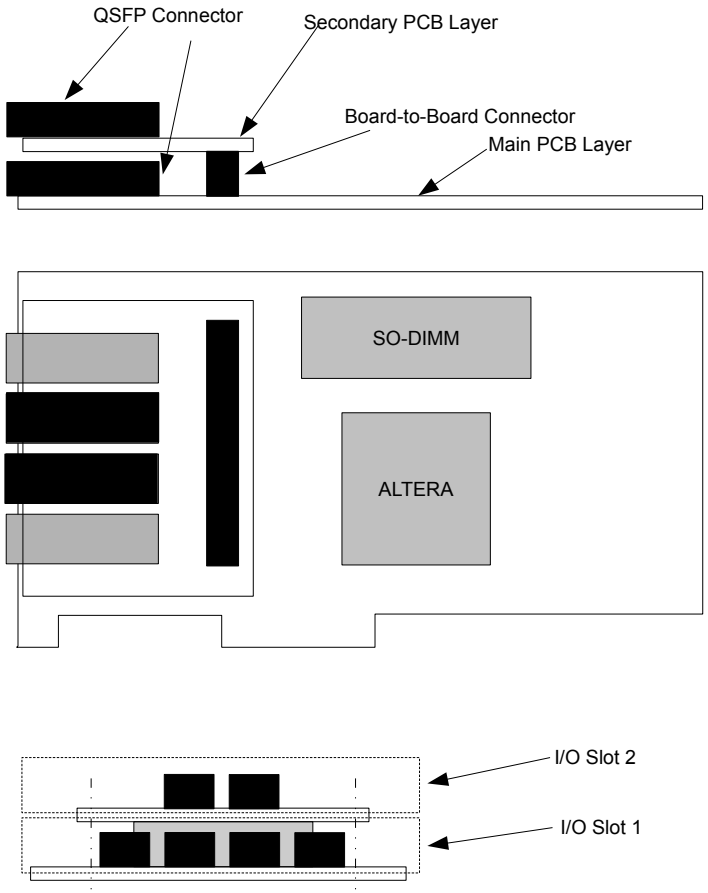
Leg II (3Q 2010 $\longrightarrow$): APENet+ integration

- Hardware test and firmware optimization
- API development for high efficient GPU-to-network communication
- Fine Tuning and application benchmarking

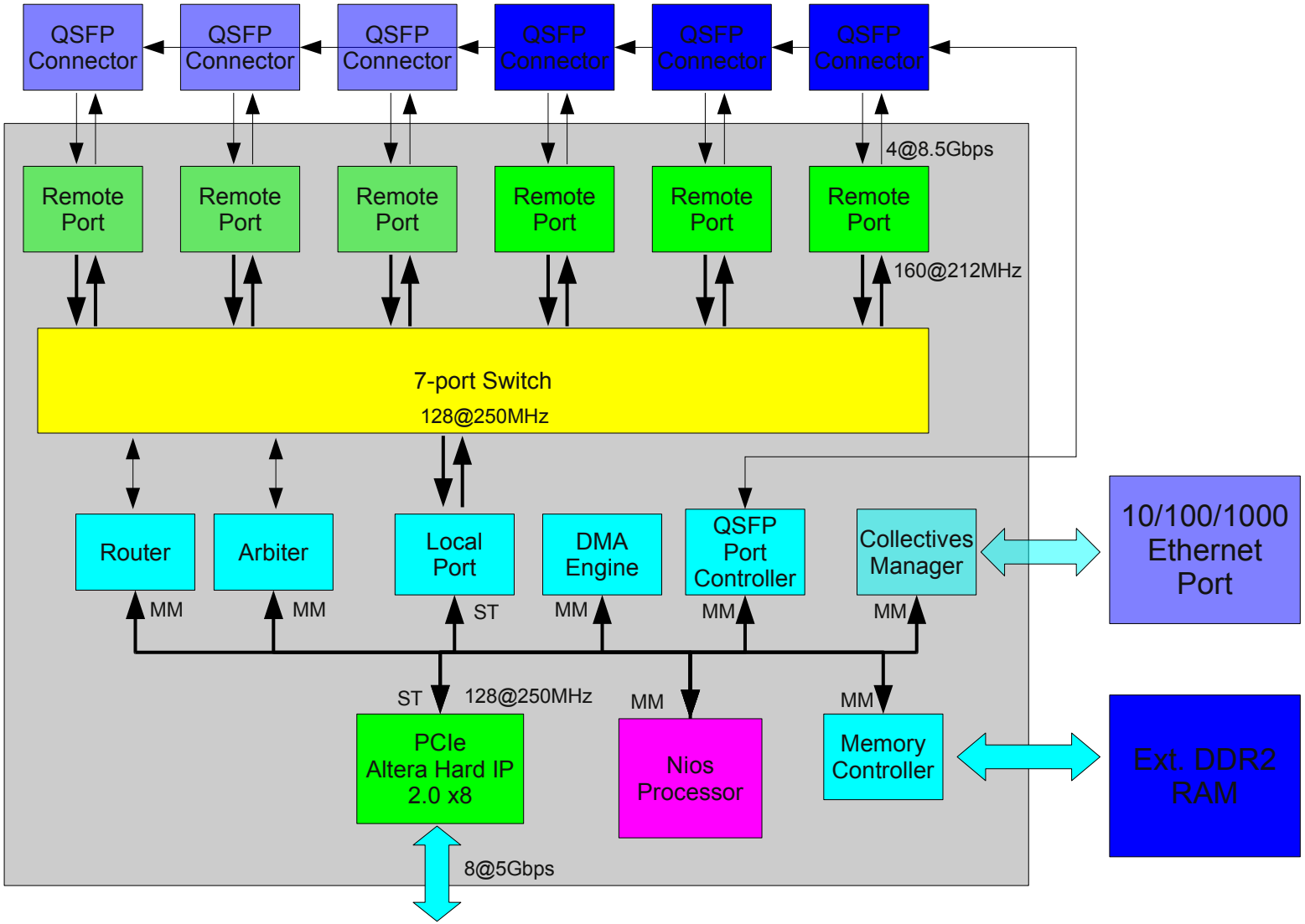
Leg III (2011 $\longrightarrow$): medium/large system deployment, production starts.

APENet+ Tech Specs

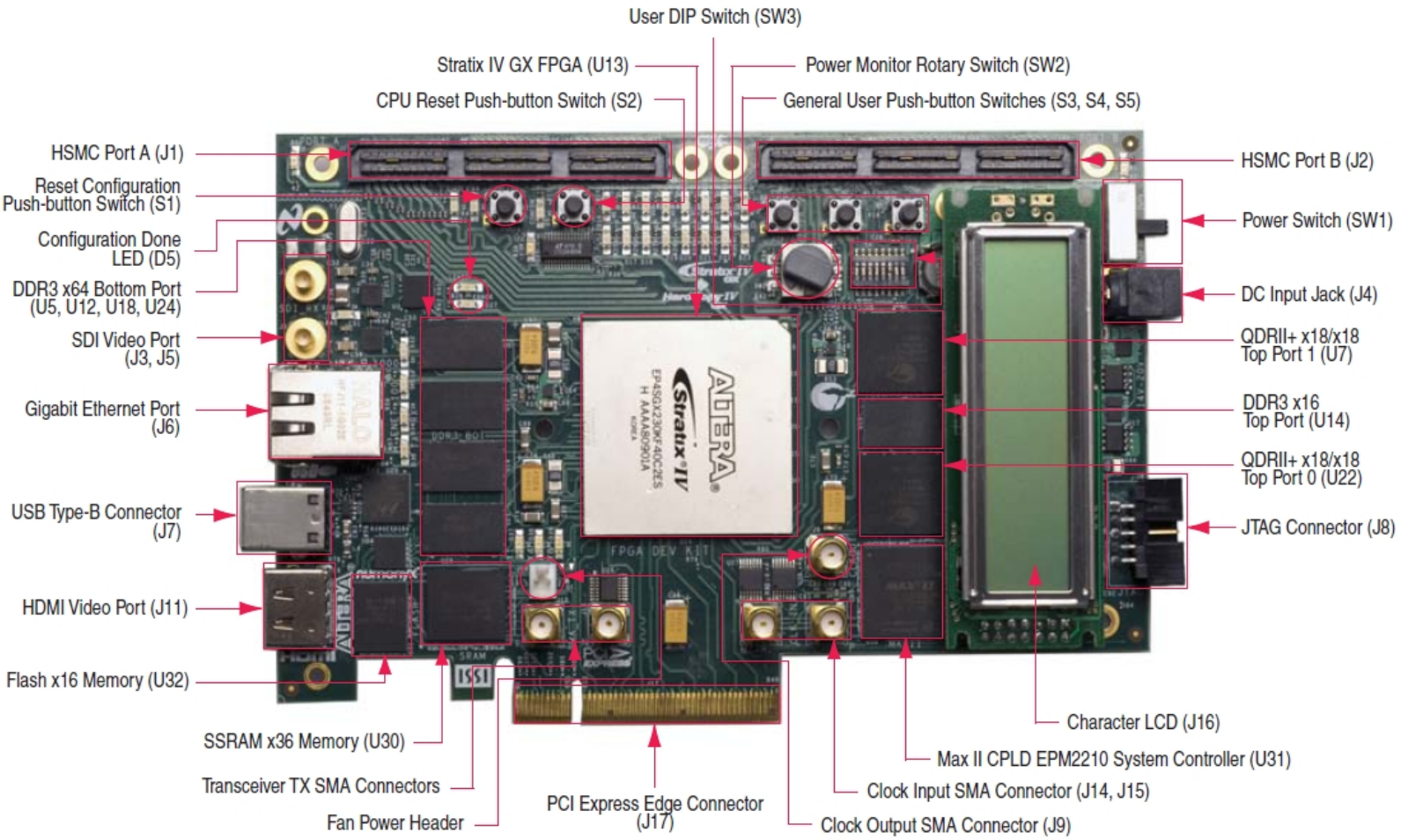
- 6 Remote Channels based on QSFP technology (up to 34 Gbps with 4 bonded Altera embedded transceivers)
- Host Connection based on PCIe x8 v2.0 (4 GB/s)
- SO-DIMM DDR3 socket (512 MB – 2 GB)
- USB/PCIe re-programming
- 1U Chassis Compliance



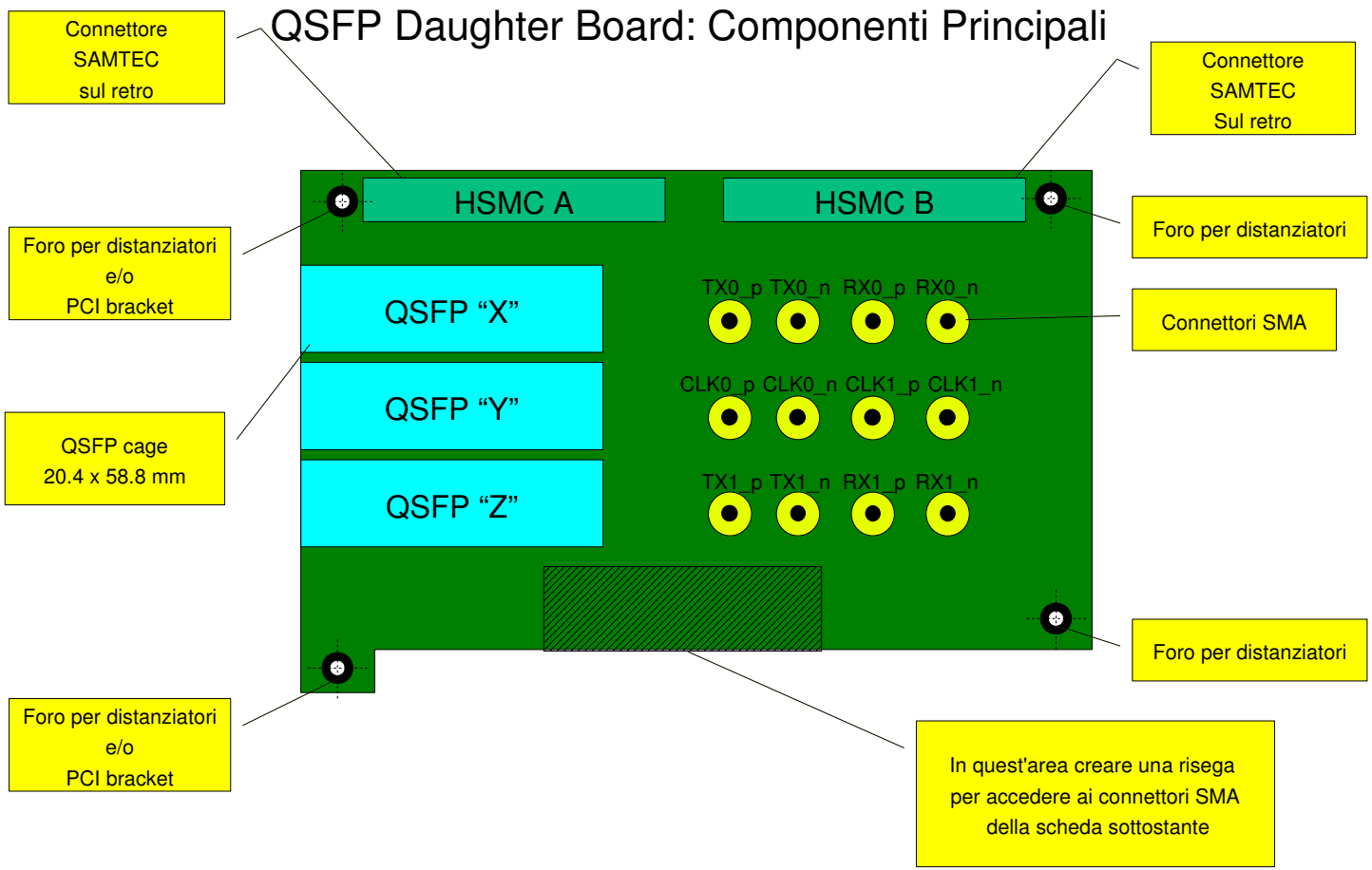
APENet+ Logic Blocks



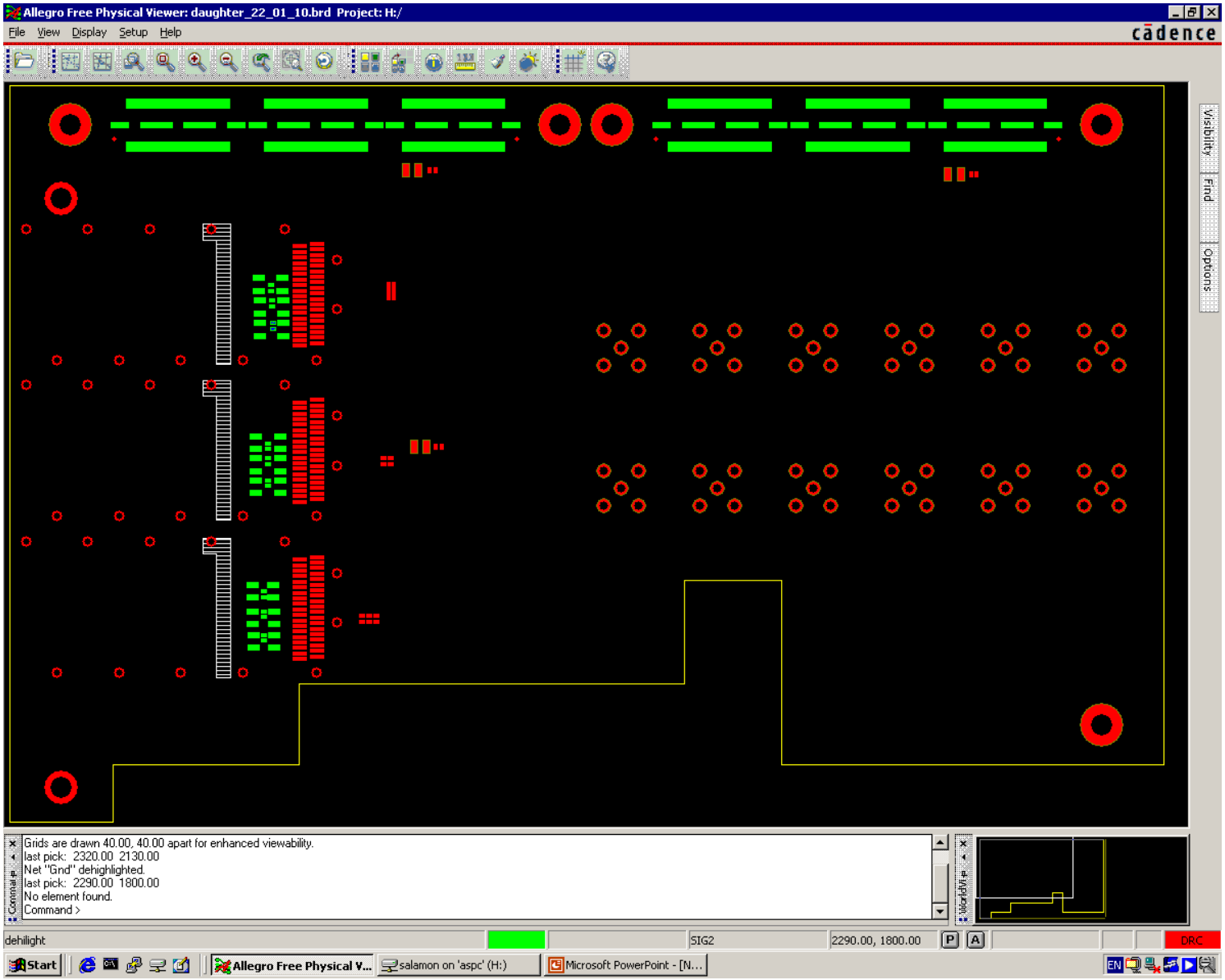
Altera Stratix4 Development Board



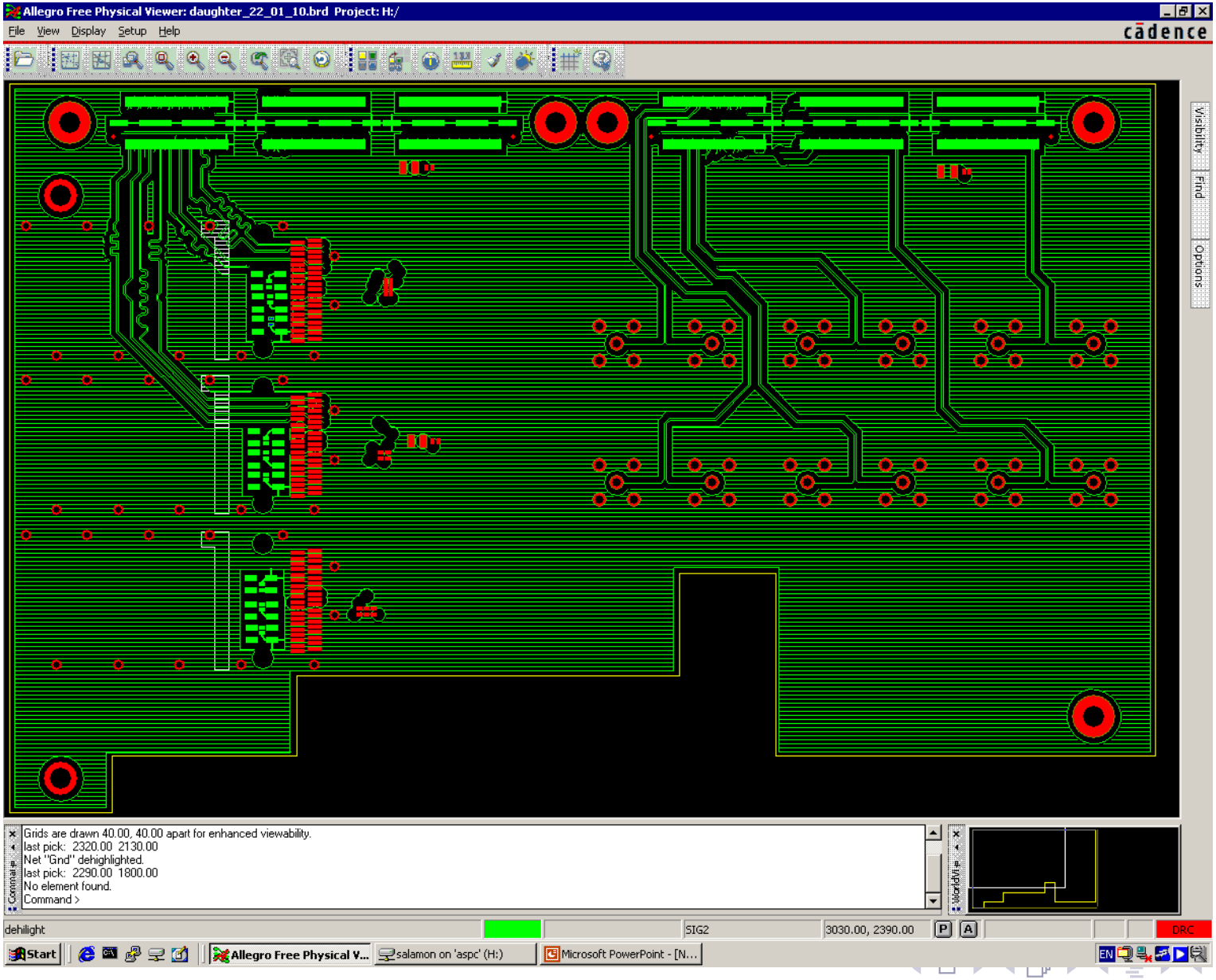
APE QSFP Daughter Card



APE QSFP Daughter Card II



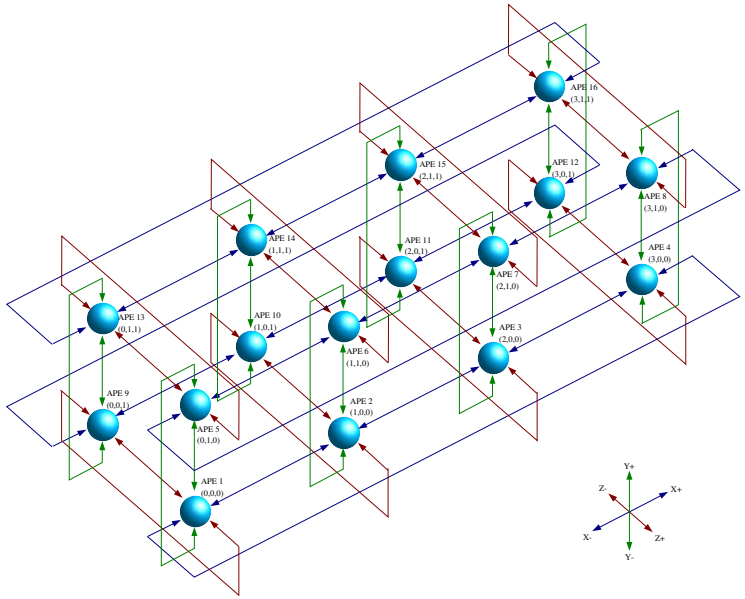
APE QSFP Daughter Card III



Tying GPUs together

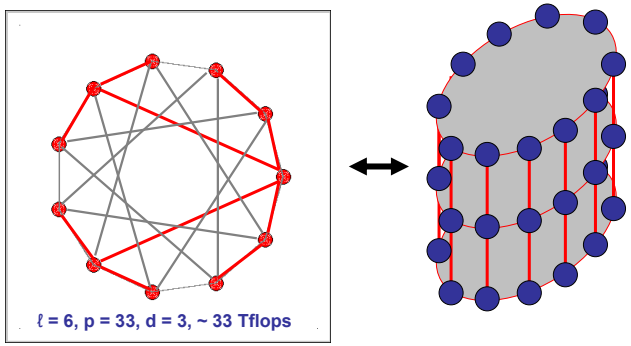
Depending on applications, arranging with the 6 links different network topologies can exploit to higher code efficiency

- 3D torus can be a natural fit for n-Dim load balanced first neighbour codes (LQCD)
- higher connectivity can be reached for "unbalanced" codes (MD) [Rossi, Salina et al.]



Connecting network arrangement, # 3

- 4 of the available $\ell = 6$ links per node are used to connect $p = 11$ processors in a plane (no full connectivity)
- The remaining 2 links are used to connect each processor in a plane to its upper and lower lying processor (in a periodic configuration)
- Minimal connecting distance between any two processors is $d = 2$ in each plane, but $d = 3$ if the whole drum is looked at



Conclusions

- To satisfy current and short term computational requirements we need to look at new computing architectures.
- GPUs are a good chance to lower overall costs and gain in performances more than Moore's Law for commodity processors allows.
- For not embarassingly parallel problems which are not going to fit into 1 GPU, network is again a critical issue.
- APENet+ is a low risk and short term solution for building PC Cluster with GPUs for fully parallel applications.
- Porting on CUDA/OpenCL of applications is anyway needed.

People involved

- Roma 1
  - O. Frezza
  - A. Lonardo
  - D. Rossetti
  - P. Vicini
  - more Juniors. . .
- Roma 2
  - R. Ammendola
  - R. Petronzio
  - A. Salamon
  - N. Tantalo
- LABE Roma 1