EURETILE: EUropean REference TILed architecture Experiment

Unified Networking Infrastructure for Embedded and HPC many-tile platforms

I. Bacivarov[1], I. Belaid[2], A. Biagioni[3], A. El Antably[2], N. Fournel[2], O. Frezza[3], W. Geurts[4], G. Goossens[4], J. Jovic[5], R. Leupers[5], F. Lo Cicero[3], A. Lonardo[3], L. Murillo[5], P. S. Paolucci[3], D. Rai[1], D. Rossetti[3], F. Rousseau[2], L. Schor[1], C. Schumacher[5], F. Simula[3], L. Thiele[1], L. Tosoratto[3], P. Vicini[3] and H. Yang[1]

Abstract

EURETILE is an FP7-funded project that aims to investigate and implement brain-inspired foundational innovations to massively parallel, tiled computer architectures and their corresponding programming paradigm. In this context, the project devises an HPC platform and an Embedded platform both exploiting the same custom interconnect IP which implements a 2D/3D toroidal direct network and provides RDMA capabilities to the system.
Inclusion of an ASIP to accelerate specific tasks has been investigated and systemic fault-tolerance features are scheduled for introduction during 2012.

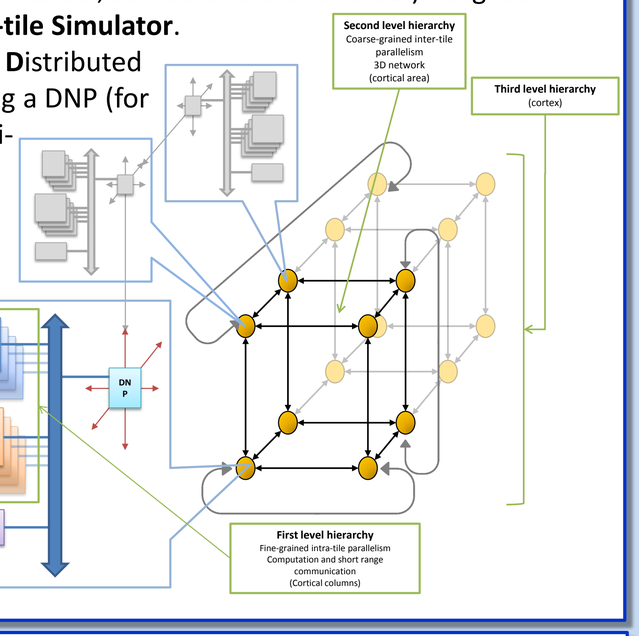
Contacts

1. CENL, ETH Zurich, Switzerland, email: firstname.lastname@tik.ee.ethz.ch – http://www.tik.ee.ethz.ch/~euretile
2. SLS Group, Laboratoire TIMA , France, email: firstname.lastname@imag.fr – http://tima.imag.fr
3. INFN Roma, Italy, email: firstname.lastname@roma1.infn.it – http://apegate.roma1.infn.it/APE
4. TARGET, Belgium, email: lastname@retarget.com – http://www.retarget.com
5. ICE, RWTH-Aachen University, Germany, email: firstname.lastname@ice.rwth-aachen.de – http://www.ice.rwth-aachen.de

EURETILE: http://euretile.roma1.infn.it

EURETILE platforms

EURETILE investigates and implements innovations for equipping an elementary HW tile with high-bandwidth, low-latency, brain-like inter-tile communication (mimicking 3 levels of connection hierarchy, namely *neural columns*, *cortical areas* and *cortex*). Target of EURETILE is a **fault-tolerant**, **many-tile** hardware platform, supplemented by a **many-tile Simulator**.
EURETILE builds upon a common network infrastructure and HW interconnect IP, the **D**istributed **N**etwork **P**rocessor; the elementary tile is an heterogeneous multi-processor including a DNP (for inter-tile comm.), a many-core Floating-Point engine (for numerical work) and a multi-core CPU (for control and OS). EURETILE develops two platforms:



HPC

✓The **HPC platform** consists in the QuOnG initiative by INFN, a cluster of Intel-based, off-the-shelf components networked by apeNET+ cards. These cards include the DNP and, according to the roadmap, its evolution will incorporate brain-inspired, fault-tolerance features and ASIPs on state-of-the-art FPGAs. Beyond applications coded in the EURETILE DAL-based software toolchain, this platform also runs standard C++/CUDA/MPI code.

Embedded

✓The **Embedded platform** is an homogeneous mesh of RISC-like CPUs (IP by RWTH-Aachen) interconnected by DNPs; this is fully representative of an embedded multi/many-core system. The actual implementation of this platform is the Virtual Platform, a simulated platform containing a SystemC fast model of the DNP and a cycle accurate model of the RISC.

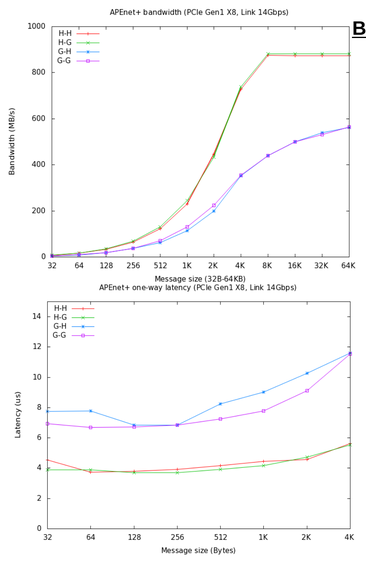
QuOnG: the HPC platform

QuOnG is an INFN initiative targeted at developing a high performance computing system dedicated to **Lattice QCD** computations; it is a massively parallel computing platform leveraging on commodity multi-core processors coupled with last generation GPUs as computing nodes interconnected by a point-to-point, high performance, low-latency 3D torus network. This network mesh is particularly suited to the transmission patterns of the set of algorithms LQCD belongs to.

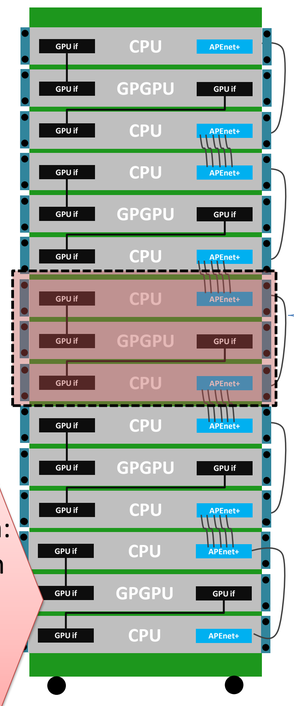
The network is built upon the **apeNET+** card:

- ✓PCIe board with signaling capabilities for up to X8 Gen2 (**4+4 GB/s** peak bi-dir with the host PC)
- ✓6 full bi-dir links on 4 bonded lanes over **QSFP+** cables
- ✓raw bandwidth of **34Gb/s** per dir
- ✓expected power envelope of **80W**
- ✓transfers are **RDMA** – CPU is not involved
- ✓custom-designed **network-to-GPU** interface on top of PCIe P2P transactions available on Fermi-class NVIDIA GPUs → significant reductions in access latency for inter-node data transfers.

The graphs below show transmission latency and bandwidth vs. message packet size for the apeNET+ card. Acronyms and colours in the legend distinguish if in- and out-bound packets are on **H**ost or **G**PU memory.



Below: a pictorial view of a QuOnG rack.



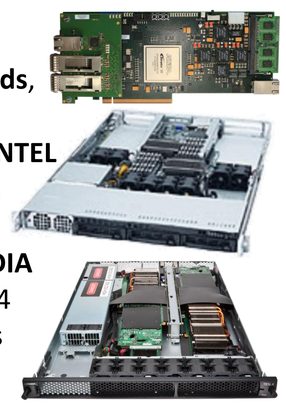
Complete QuOnG installation: a fully populated **42U** rack with
✓**60 TFlops/rack peak**
✓**25 kW/rack** (0.4 kW/Tflops)
✓**300 k€/rack** (5 K€/TFlops).

Besides: red box encloses an elementary mechanical assembly, a **3U stack** consisting of:

2x ✓**apeNET+ boards**, equipping

2x ✓**1U multicore INTEL rackable system**, "sandwiching"

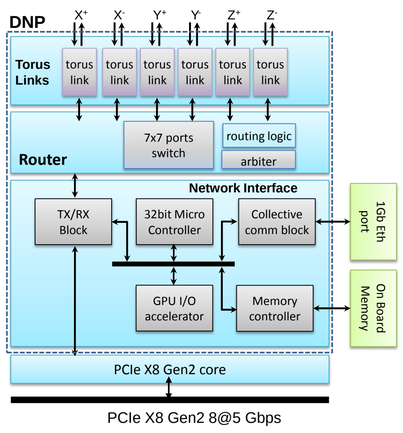
✓**1U S2090 NVIDIA system**, packing 4 Fermi-class GPUs and >4 TFlops



Distributed Network Processor (DNP)

The **DNP** is an INFN custom-designed IP at the core of a packet-based, direct network fabric on top of a 2D/3D toroidal mesh topology; its versatility allows it to be retargeted from being an HPC cluster interconnect to an embedded one. The DNP takes onto itself all the communication tasks like deadlock-free routing, flow control and integrity check; data transfer is managed through zero-copy RDMA primitives.

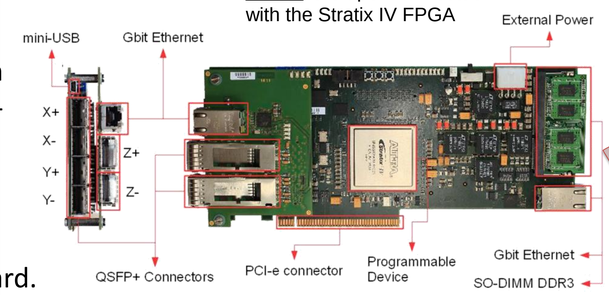
In its HPC incarnation, the DNP is a VHDL firmware synthesized onto a Stratix® FPGA at the core of the **apeNET+** card. The schematic below shows its block structure, which is split into:



- ✓**torus links** – bi-dir DC-balanced Ser/Des with auto-retransmission capability thanks to a *word-stuffing* CRC-protected low-level packet protocol;
- ✓**router** – for packet arbitration and dimension-ordered routing, guaranteed deadlock-free by using virtual channels (60ns routing latency);
- ✓**network interface** – for packet injection and processing logic comprising host interface, TX/RX logic and two auxiliary blocks:
  - **micro controller** – part of the FPGA, relieves the DNP core from some chores of RDMA implementation (for fast LUT management on its on-board memory)
  - **GPU/IO accelerator** – custom block for acceleration of GPU-initiated network operations.

The board can support 4 or 6 channels for 2D or 3D mesh supplemented by a piggy-back card.

Below: the apeNET+ board with the Stratix IV FPGA



Virtual Embedded Platform

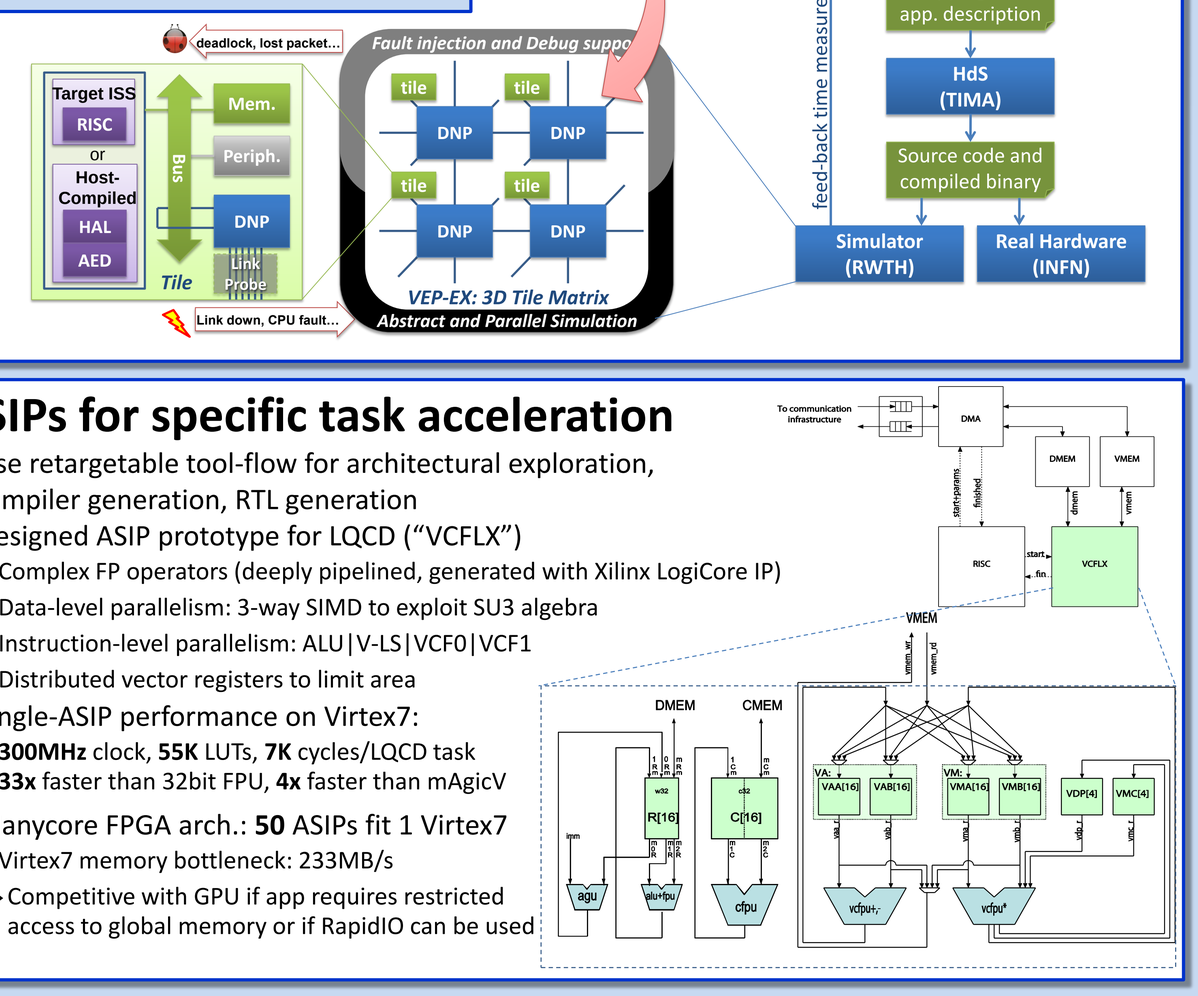
The Virtual Embedded Platform – Experimental (VEP-EX) is an embedded system simulator tailored to the specific demands of the EURETILE project consortium:

- ✓ Fast and scalable simulation to enable experimentation with a possibly large number of tiles
- ✓ Fault injection support
- ✓ Debugging and profiling for massively-tiled systems

The simulation framework will fulfil multiple roles in the course of the project:

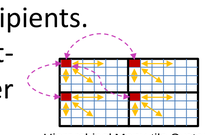
- ✓ Serve as a customizable platform to evaluate the overall EURETILE methodology without the need to wait for actual hardware to become available
- ✓ Aid in operating system, application and driver development by providing tracing and inspection capabilities beyond those typically available in hardware
- ✓ Answer performance queries from the application task mapping framework

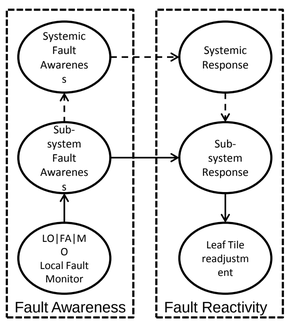
Further on, the simulation framework will offer acceleration technologies like **abstract** and **parallel simulation** to ensure simulation as a productivity tool will stay viable.



Fault-tolerance features in EURETILE

When scaling to peta/exa-scale in HPC, usage of techniques that aim to maintain a low Failure In Time (FIT) ratio is mandatory. To this purpose, in EURETILE we use a systemic approach based on the idea of splitting the **fault-tolerance** feature into **fault awareness** and **fault reactivity**. A hierarchical structure of the HW/SW components allows to make a systemic decision on the basis of information gathered from the lower levels and to push the countermeasure to the relevant recipients.
At the moment (Jan 2012) a specification draft of this fault-tolerant architecture is complete and will be further investigated and implemented during 2012.


Hierarchical Many-tile System



Hierarchical Fault Tolerance

ASIPs for specific task acceleration

- ✓Use retargetable tool-flow for architectural exploration, compiler generation, RTL generation
- ✓Designed ASIP prototype for LQCD ("VCFLX")
  - Complex FP operators (deeply pipelined, generated with Xilinx LogiCore IP)
  - Data-level parallelism: 3-way SIMD to exploit SU3 algebra
  - Instruction-level parallelism: ALU|V-LS|VCF0|VCF1
  - Distributed vector registers to limit area
- ✓Single-ASIP performance on Virtex7:
  - **300MHz** clock, **55K** LUTs, **7K** cycles/LQCD task
    33x faster than 32bit FPU, **4x** faster than mAgicV
- ✓Manycore FPGA arch.: **50** ASIPs fit 1 Virtex7
  - Virtex7 memory bottleneck: 233MB/s

→Competitive with GPU if app requires restricted access to global memory or if RapidIO can be used

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

INFN Istituto Nazionale di Fisica Nucleare

RWTH AACHEN UNIVERSITY

iima GRENOBLE 1 UNIVERSITE JOSEPH FOURIER SCIENCES.TECHNOLOGIE.SANTÉ

TARGET™ THE ASIP COMPANY