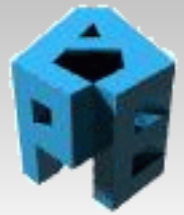
INFN Istituto Nazionale di Fisica Nucleare

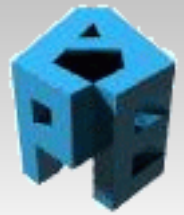


APE group

Present and future of APE projects

davide.rossetti@roma1.infn.it

APE in a few words



APE group

Our mission is providing solutions for theoretical numerical computing in INFN

Our focus is in HPC architectures

Lattice QCD is our main application, but historically other topics actively pursued (Glasses, CFD, Weather)

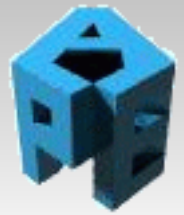
A long time ago…

APE group

As of three-four years ago, no doubts:

- buy IBM BlueGene, if you can ($ or ?)
- or choose your type of cluster (Intel or AMD, 1U or Blade)

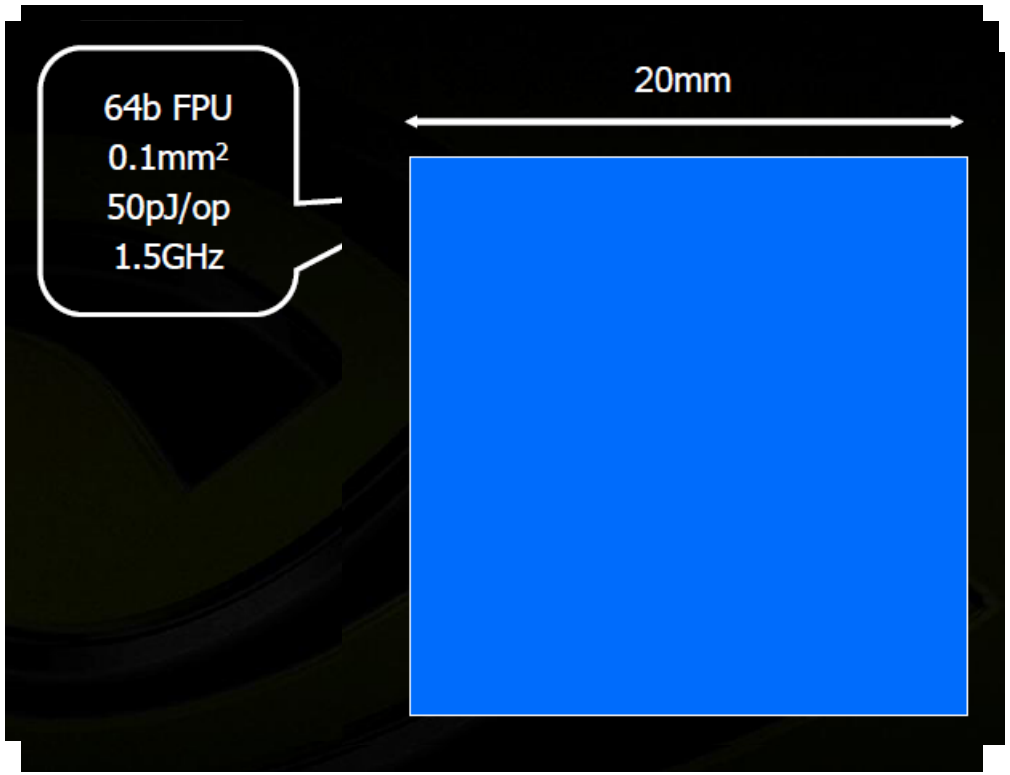
Where the power is spent



APE group

*"chips are power limited and most power is spent moving data around"**

- 4 cm^2 chip
- 4000 64bit FPU fit
- Moving 64bits on chip == 10FMAs
- Moving 64bits off chip == 20FMAs



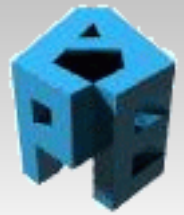
*Bill Dally, Nvidia Corp. talk at SC09

One year ago in Naples

APE group

- GPUs are rising
- A lot research being done on GPUs (compbio, moldyn, CFD, astrophys, …), also in INFN (D'Elia, Bosi, …)
- Some LQCD groups are adopting GPUs (US, TW, Japan, …)
- Still a few GPUs in a single host
- Next big step is **multi-GPU**

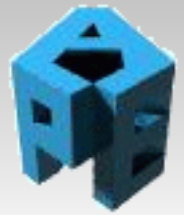
Last year the unexpected



APE group

- GPUs are super-cool, scaling from the desktop to the supercomputer
- NVidia fixed many design problems in Tesla and went out with Fermi (caches, DP, …)
- AMD struggling to merge CPUs and GPUs (Fusion)
- Intel canceled its own GPU (Larrabee)
- IBM BlueGene no more so cool

An exciting period



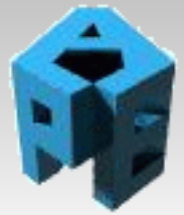
APE group

What is confusing today:

- NVidia has a great GPGPU, moves well in HPC, now wants to integrate a low-power CPU (Denver)
- Intel does great mid-to-high power GPUs, and now promises a GPGPU without Graphics (MIC, morning talk today)
- AMD still struggling

Everybody wants to step on each other toes :)

What about us?



APE group

Recently we sat, calmed down and did our planning.

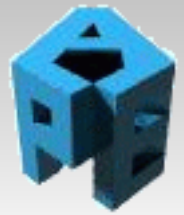
Two time scales:

- short term, next 3-4 years, do our best, working with industrial partners, serving our community, funded *partly* by INFN
- long term, in 5-10 years, preparing for the next big thing (?), funded 100% by EU FET

As in financial M&A, point is leveraging synergies:

- INFN € can buy HW, no people
- EU € can pay people, no HW (laptops are ok :)

Short term plan



APE group

- An hybrid cluster
- Multicore nodes in **your** choice of cha$$is (1U, blades)
- Accelerated with 1 or 2 GPUs per node (M20x0, S20x0)
- Communicating with optimized custom interconnect (APEnet+)
- With a standard software stack (MPI, OpenMP, ...)
- Optionally an augmented programming model (cuOS)
- With a community of researchers sharing codes (see talk D'Elia,…)

GPU by NVidia:

- solid HW
- good SW
- open to collaboration with customers (wow!)

APEnet+ interconnect

APE group

A network *à la APE*:

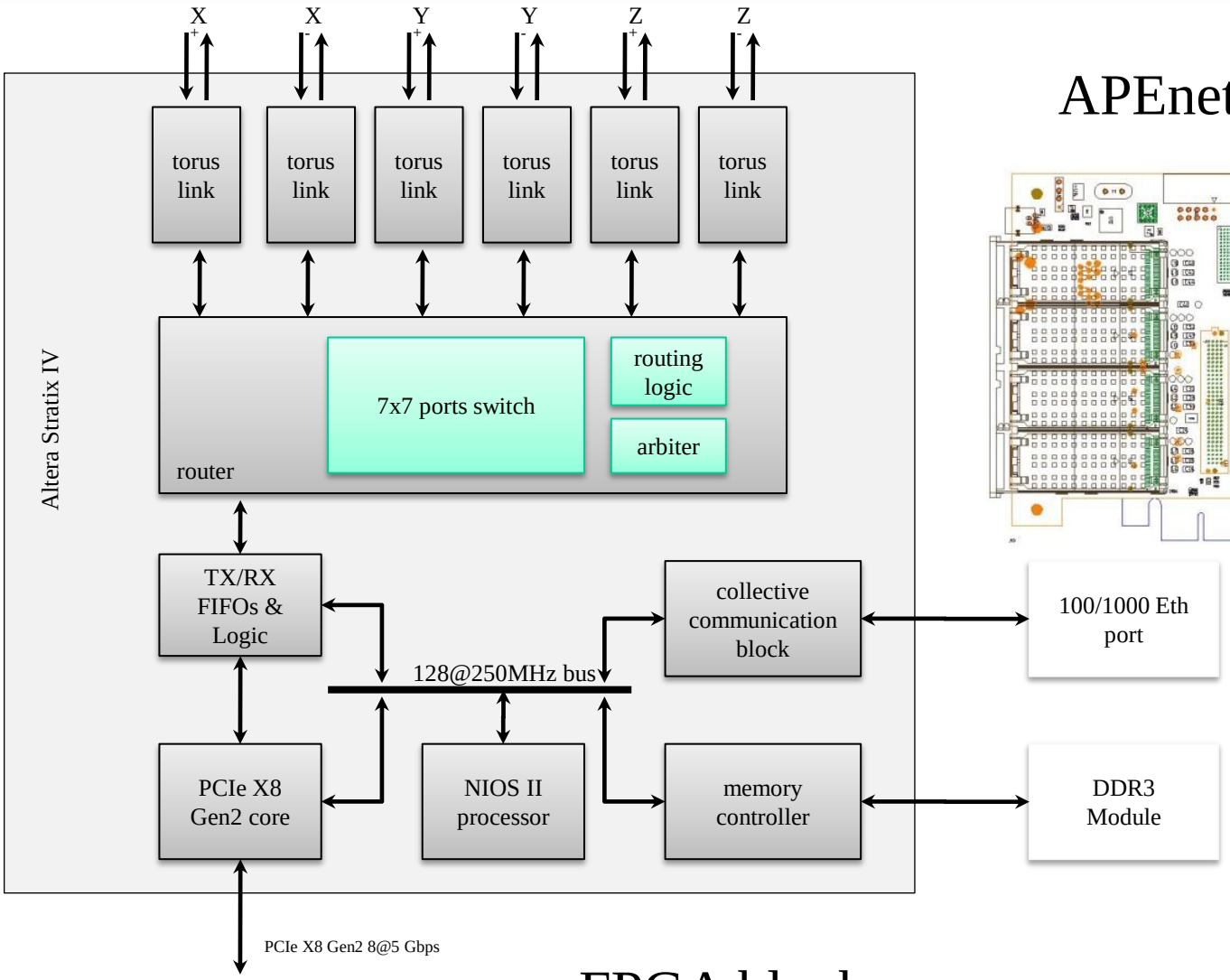
- 3D torus, packet based *Network Processor*
- scaling up to thousands of nodes
- packet auto-routing
- top bandwidth (6 x 30+30Gbps links)
- PCIe X8 gen2
- evoluted zero-copy RDMA CPU interface
- SW: MPI (high-level), RDMA API (low-level)

Working on:

- fault tolerance
- experimental direct GPU interface
- on-board processing (embedded processor)

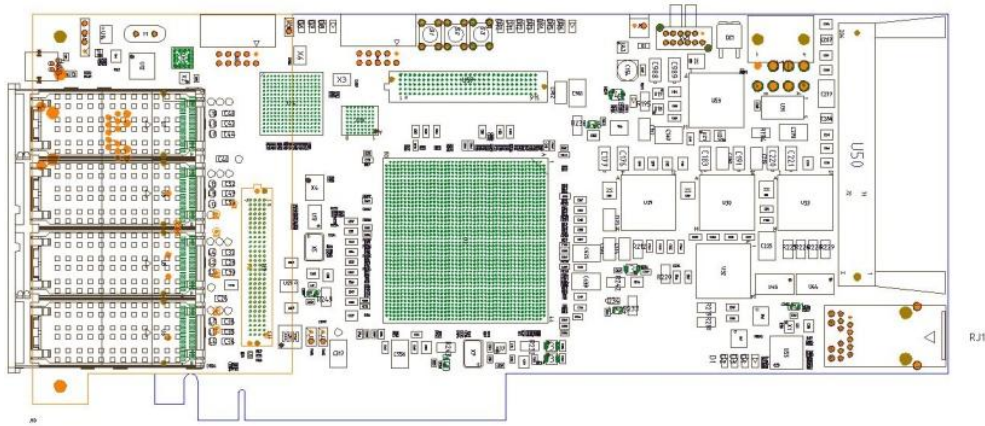
Some eye candies …

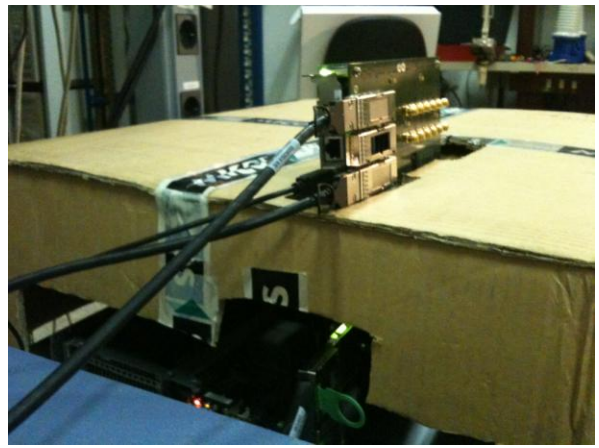
APE group



FPGA blocks

APEnet+ final board, 4+2 links





3 link test board

Experimental direct GPU interface

APE group

Sneak peek preview

HW optimizations (w coll. NVidia):

1. TX/RX: APEnet+ can read/write GPU memory buffers with PCIe peer-to-peer transfers
2. TX: GPU can directly send pkts thru APEnet+

SW improvements:

- Integrate (1) in OpenMPI, i.e. MPI_Send(gpu_addr)
- Use MPI directly in CUDA kernels, run-time switch between 2 *paths*:
  - using (1) send small & fast pkts, no CPU
  - using cuOS, pass MPI call to CPU

Augmented programming model

APE group

leveraging our cuOS (CUDA Off-loaded System services) :

- cuMPI: MPI APIs ...
- cuSTDIO: file read/write ...

... in CUDA kernels!

Features:

- program large GPU kernels
- without explicit CPU code
- hidden use of direct GPU interface
- need resident blocks (global sync)

cuOS is developed by APE and is open source
http://code.google.com/p/cuos

Overlapping computation and networking in CUDA kernels

APE group

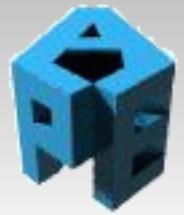
traditional CUDA:

```
//GPU
__global__ void compute_borders(){}
__global__ void compute_bulk(){}
__global__ void reduce(){}
//CPU
main() {
    do {
        compute_bulk<<<,1>>>();
        cudaMemcpyAsync(frames, 0);
        cudaStreamSynchronize(0);
        MPI_Sendrecv(frames);
        cudaMemcpyAsync(frames, 0);
        compute_borders<<<,0>>>();
        cudaStreamSynchronize(0);
        cudaStreamSynchronize(1);
        local_residue<<<,1>>>();
        cudaMemcpyAsync(lres, 1);
        cudaStreamSynchronize(1);
        MPI_Reduce(gres, lres);
   } while(gres > eps);
}
```

using in-kernel MPI (cuOS):

```
//GPU
__global__ void solver() {
    do {
        cuMPI_Isendrecv(frames);
        compute_bulk();
        cuMPI_Wait();
        compute_borders();
        local_residue(lres);
        cuMPI_Reduce(gres, lres);
   } while(gres > eps);
}
// CPU
main() {
    ...
    solver<<<nblocks,nthreads>>>();
    cuos->HandleSystemServices();
    ...
}
```

Status and deliverables



APE group

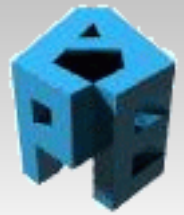
Today

- 7 GPU nodes with Infiniband for applications development: 2 C1060 + 2 M2050 + S2050
- 2 nodes HW devel: C2050 + 3 links APEnet+

Next steps:

- APEnet+ early prototypes: May ’11
- Debugging and fixes: 3-4 months
- Procurement and mass production: Q4 ’11
- € for 30-40 TFlops development cluster with APEnet+: early 2012
- More Tflops if more €

Game over…



APE group

Thank you for your patience!