INFN Istituto Nazionale di Fisica Nucleare

APE group

Many-core platforms and HEP experiments computing

davide.rossetti@roma1.infn.it

XVII SuperB Workshop and Kick-off Meeting

Elba, May 29-June 1, 2011

APE in a few words

APE group

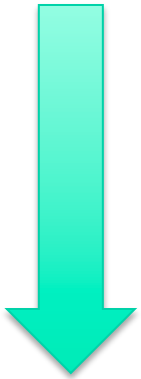
Our mission is providing solutions for theoretical numerical computing in INFN

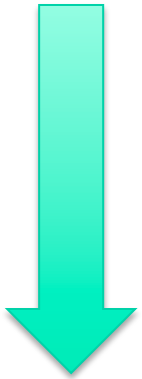
Our focus is in HPC architectures

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

Multi- vs Many- core processor architectures

Multi-core ⟷ Many-core





It's your laptop processor!

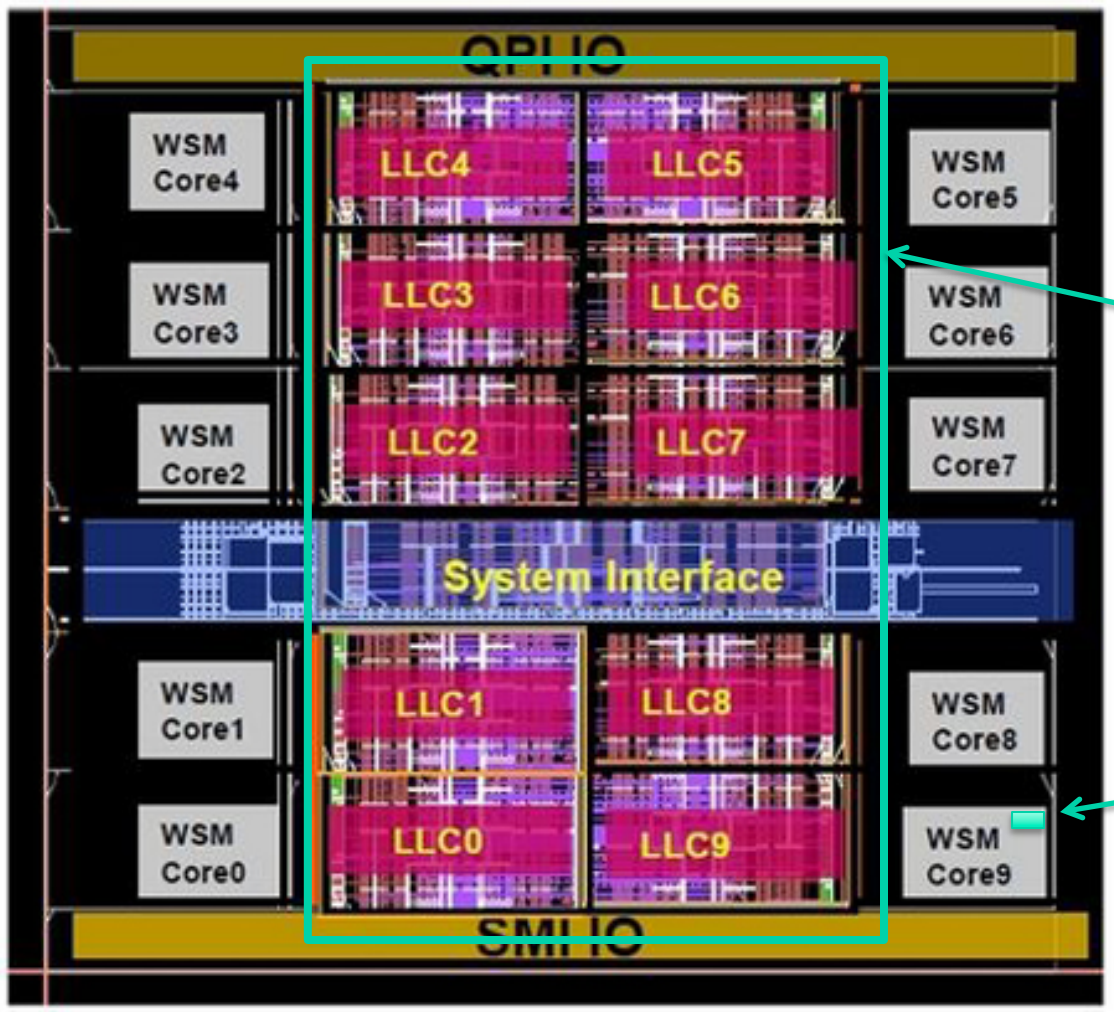
It's your laptop GPU!

So what ?

APE group

- What are the differences ?
- Why should we bother ?

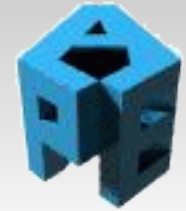
Intel Westmere-EX



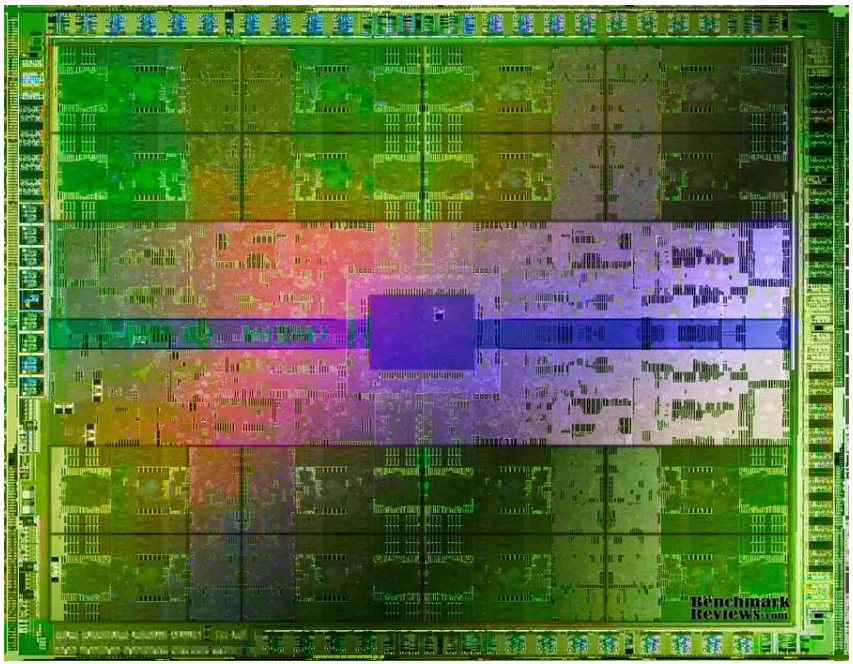
Lot's of caches!!!

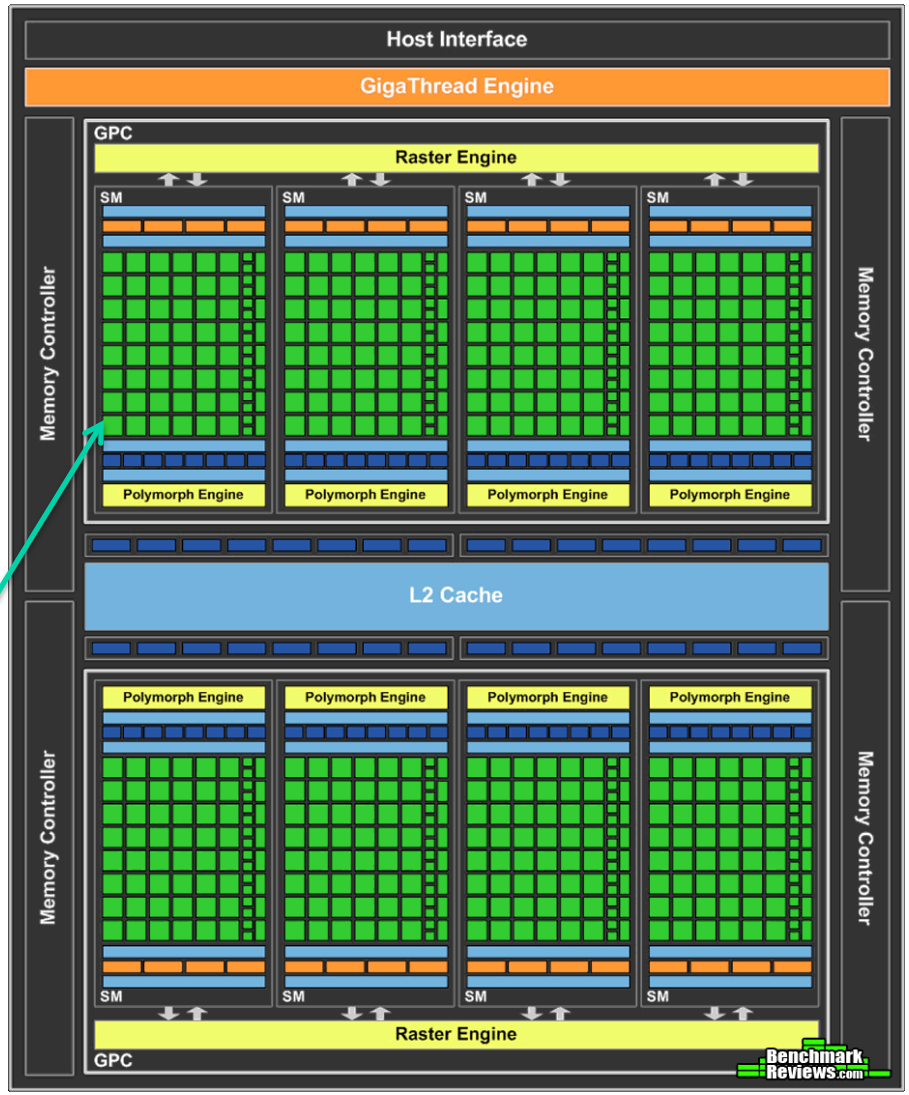
Few processing:
4 FP units are probably
1 pixel wide !!!

NVidia GPGPU



APE group





Lot's of computing units !!!

So what ?

APE group

- What are the differences ?
- Why should we bother ?

They show different trade-offs !!

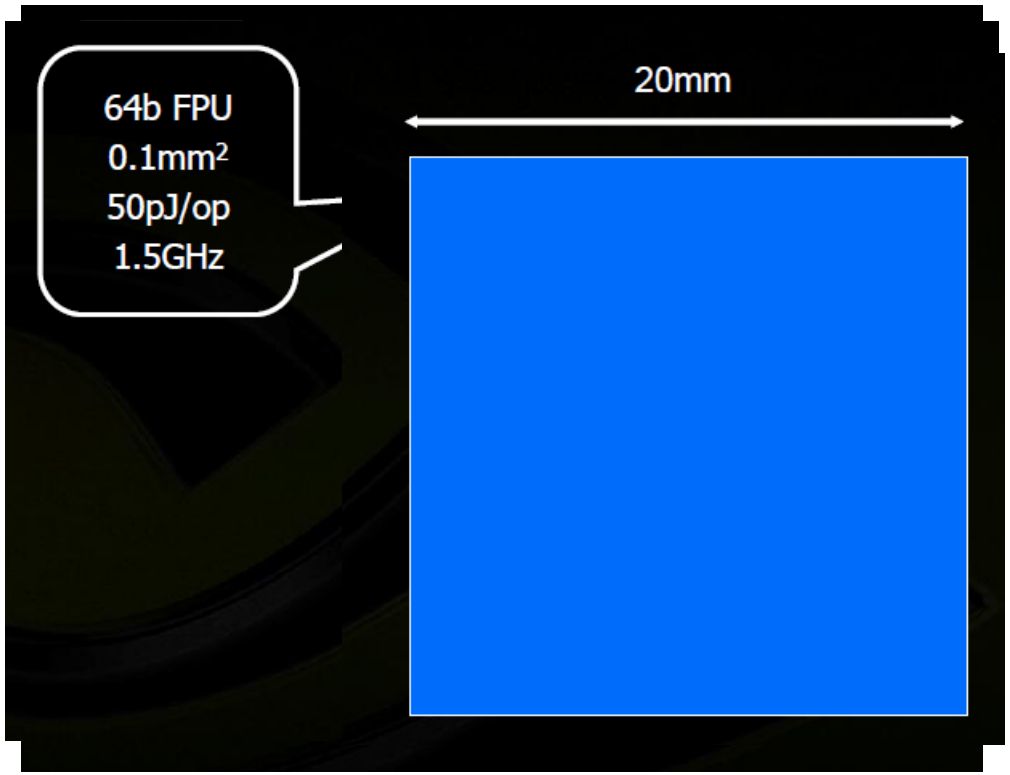
And the theory is…..

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

So what ?

APE group

- What are the differences?
- Why should we bother?

Today: at least a factor 2 in perf/price ratio

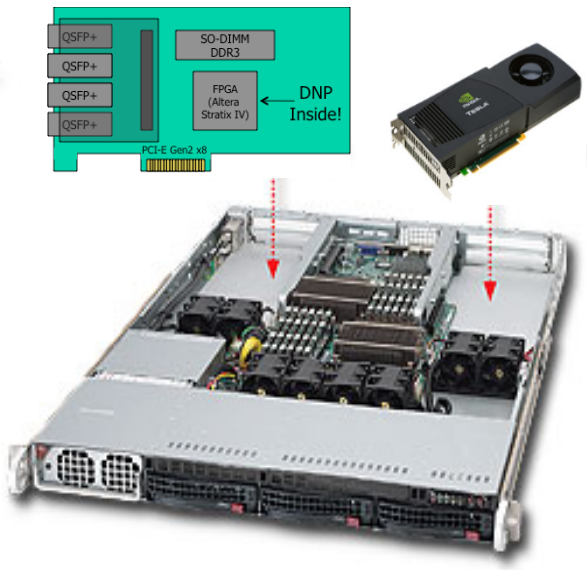
Tomorrow: CPU & GPU converging, see current ATI Fusion

Cost Effective processing solution…

APE group

GPU is an accelerator, it needs a GPU!!!

- A cluster
- A dual-socket node
- Accelerated with 1-8 GPUs per node
- Network



Flexible:

- 1 GPU can be shared by multiple processes
- 1 process can use multiple GPUs

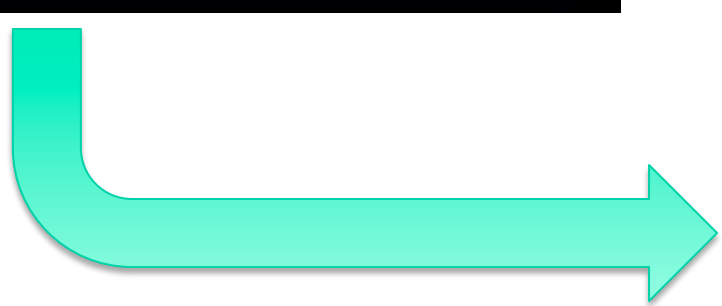
With latest top GPUs…

APE group

Slidecast: Nvidia Tesla 2090, World's Fastest HPC Processor
by RichReport

	M2090	M2070
CUDA Cores Count	512	448
Memory Bandwidth	178 GB/s	150 GB/s
Single Precision Perf	1330 GigaFlops	1030 GigaFlops
Double Precision Perf	665 GigaFlops	515 GigaFlops
Memory Size	6 GB	6 GB

Dell PowerEdge C410x



What is it for?

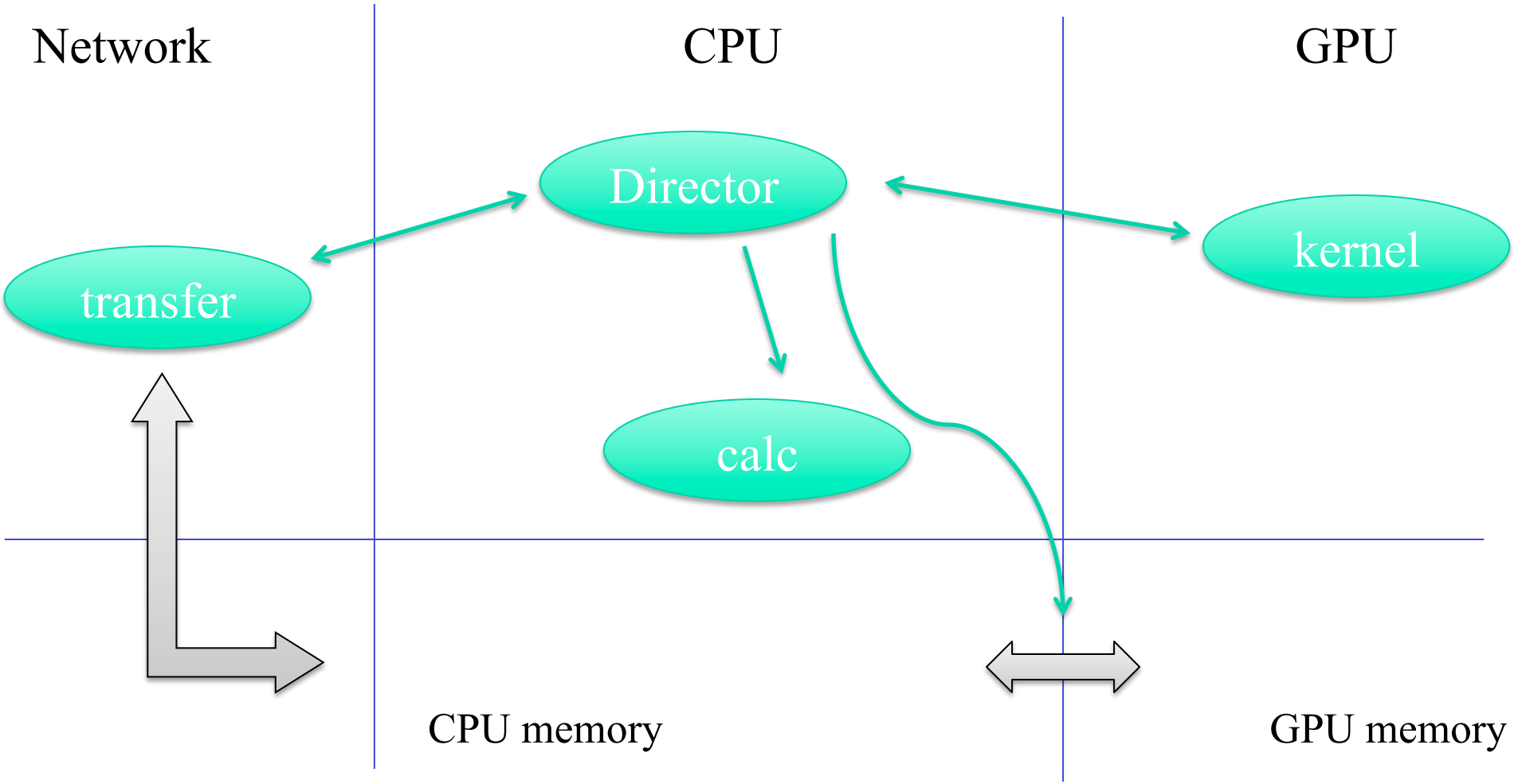
APE group

- Previous platform is good for throughput computing (capacity)
  - SuperB design ?
  - SuperB offline ?
  - Some Lattice QCD
- What else ?
  - SuperB online ? ROM ?
  - Grand challenge LQCD

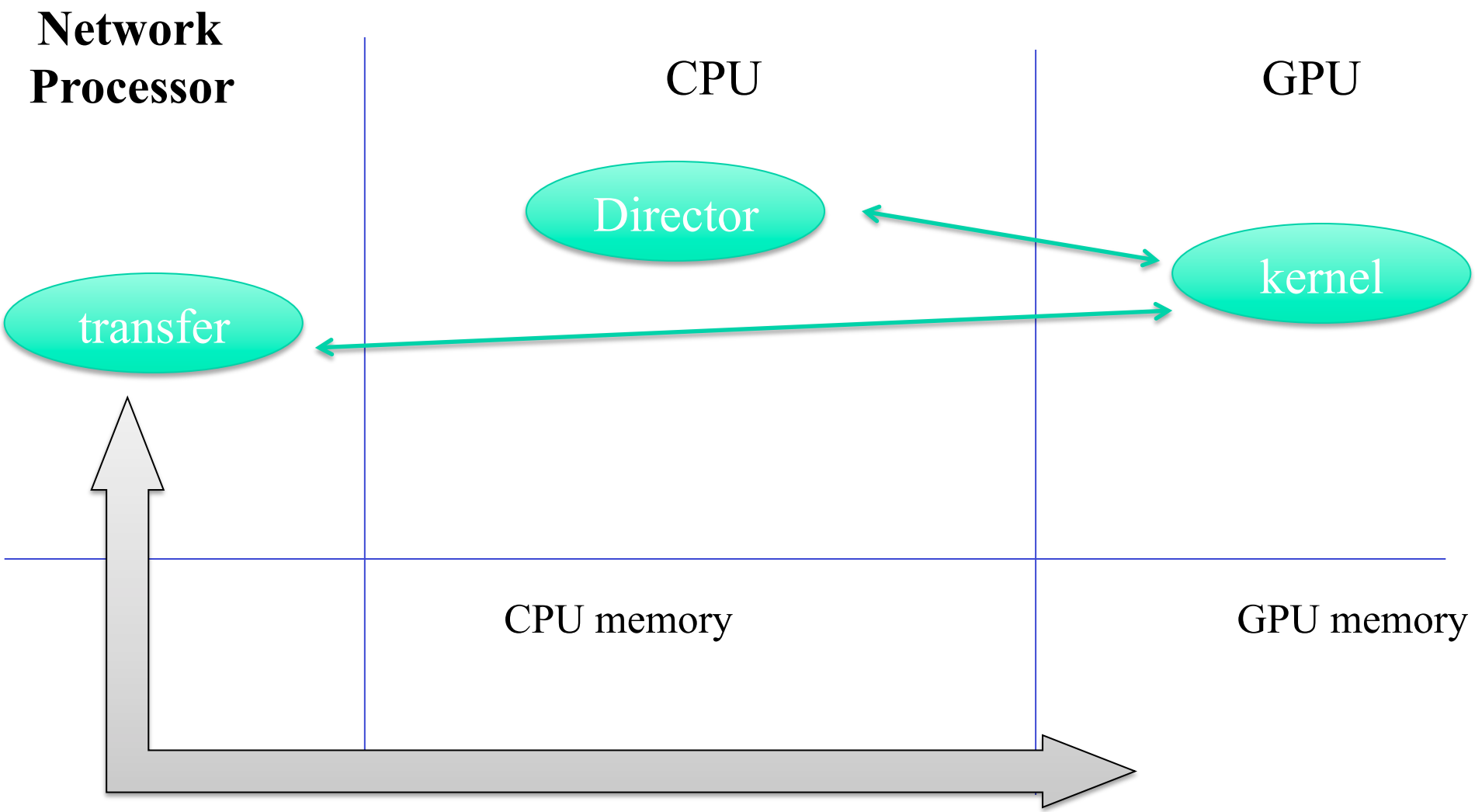
Capability

Strong scaling to multiple nodes on a single problem

The traditional flow

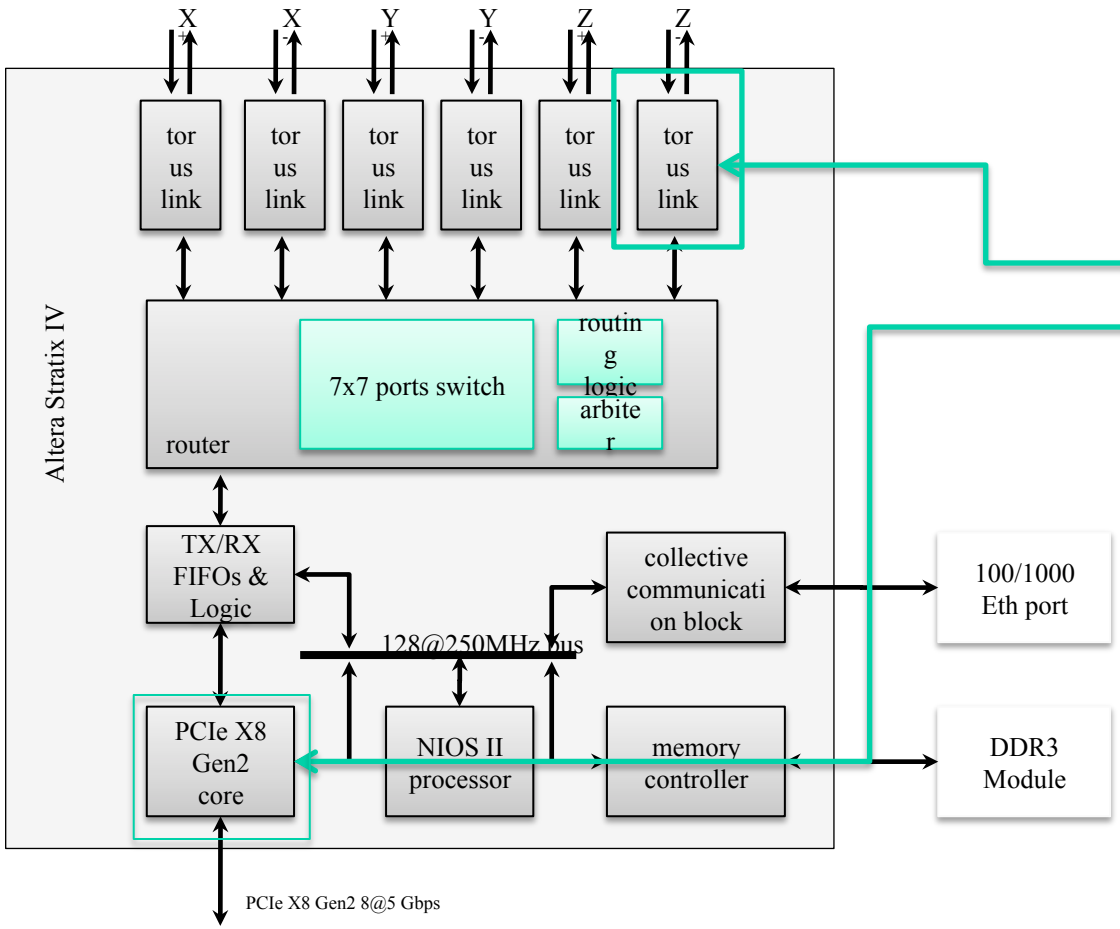


Improved network



APEnet+ interconnect

APE group



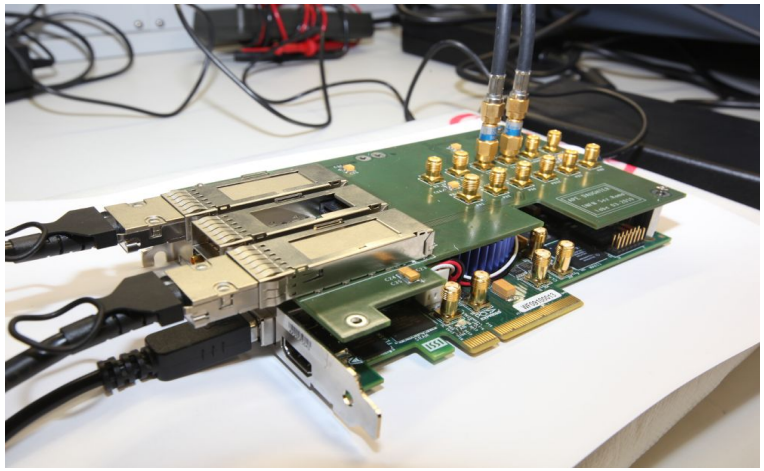
- 3D Torus, scaling up to thousands of nodes
  - packet auto-routing
  - 6 x 30+30Gbps links
- PCIe X8 gen2 (peak BW 4+4 GB/s)
- A *Network Processor*
  - Powerful zero-copy RDMA CPU interface
  - On-board processing
- Experimental direct GPU interface
- SW: MPI (high-level), RDMA API (low-level)

FPGA blocks

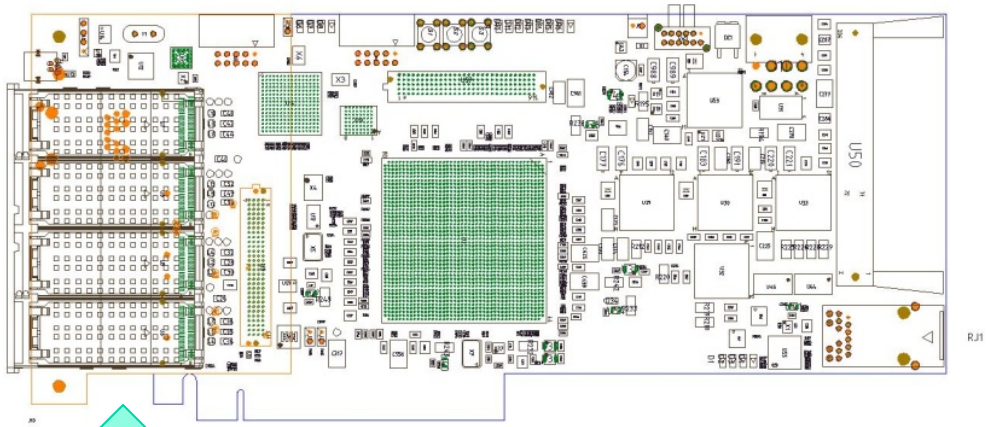
3 link test board

Some eye candies …

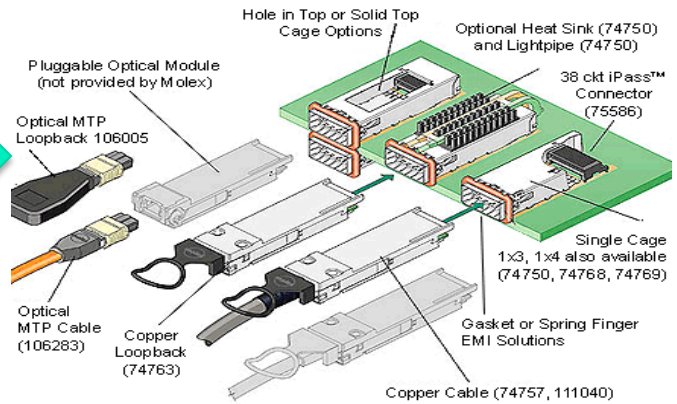
APE group



APEnet+ final board, 4+2 links



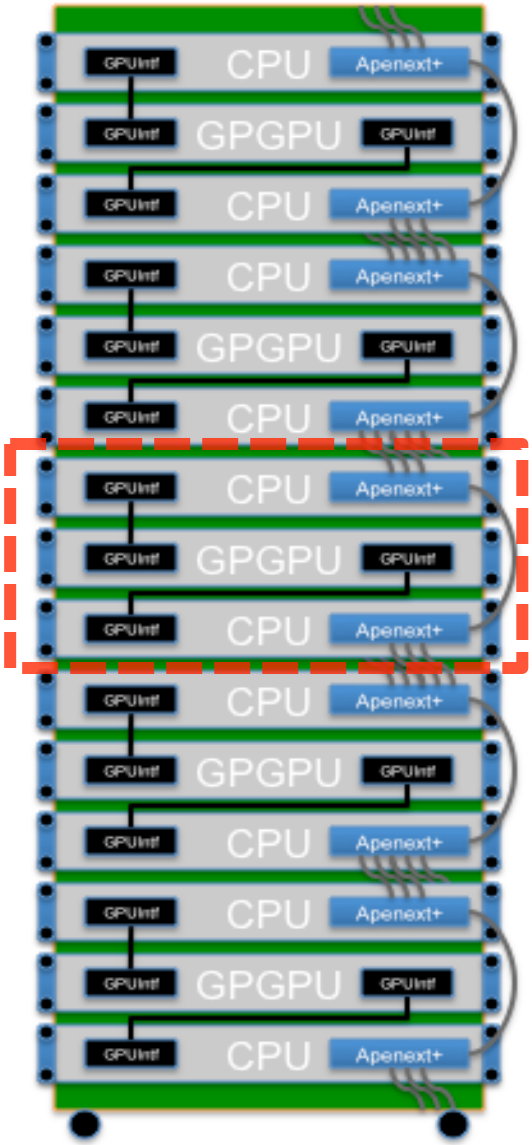
Cable options: copper or fibre



Our reference platform

APE group

- Today:
  - 7 GPU nodes with Infiniband for applications development:
    2 C1060 + 3 M2050 + S2050
  - 2 nodes HW devel:
    C2050 + 3 links card APEnet+
- Next steps, *green* and *cost effective system* within 2011
  - Elementary unit:
    - multi-core Xeon (packed in 2 1U rackable system)
    - S2070 FERMI GPU system (4 TFlops)
    - 2 APEnet+ board
  - 42U rack system:
    - 60 TFlops/rack peak
    - 25 kW/rack (i.e. 0.4 kW/TFlops)
    - 300 k€/rack (i.e. 5 K€/Tflops)



As of SuperB

APE group

A GPU-accelerated APEnet+ cluster:

- Highly compact MC simulation engine
- APEnet+ FPGA has lots of free space… with optical cables, usable for Readout Module test-bed ?
- Could it be the prototype of the SuperB computing platform ?

Game over…

APE group

Thank you for your patience!