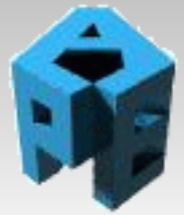
INFN Istituto Nazionale di Fisica Nucleare



APE group

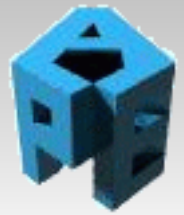
Remote Direct Memory Access between NVIDIA GPUs with the APEnet 3D Torus Interconnect

davide.rossetti@roma1.infn.it

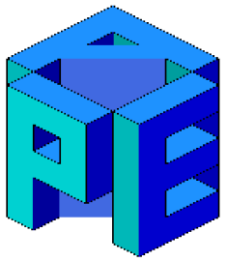
SC11

Seattle, Nov 14-17, 2011

Credits

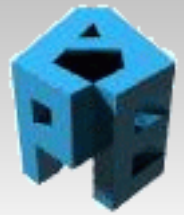


APE group



- APEnet design and development in INFN by the APE™ team ☺
- Partially supported by EU EURETILE project (euretile.roma1.infn.it)
- GPU support developed in collaboration with Massimiliano Fatica, Timothy Murray et al @ Nvidia

Index



APE group

- APEnet cluster interconnect
- GPU features
- Performance plots
- Programming model
- Status & Future plans

The APEnet+ History

APE group

- Custom HPC supercomputers: APE (86), APE100 (94), APEmille (99), apeNEXT (04)
- Cluster interconnects:
  - 2003-2004: APEnet V3
  - 2005: APEnet V3+, same HW with RDMA API
  - 2006-2009: DNP, or APEnet goes embedded
  - 2011: APEnet V4 aka APEnet+

APEnet interconnect

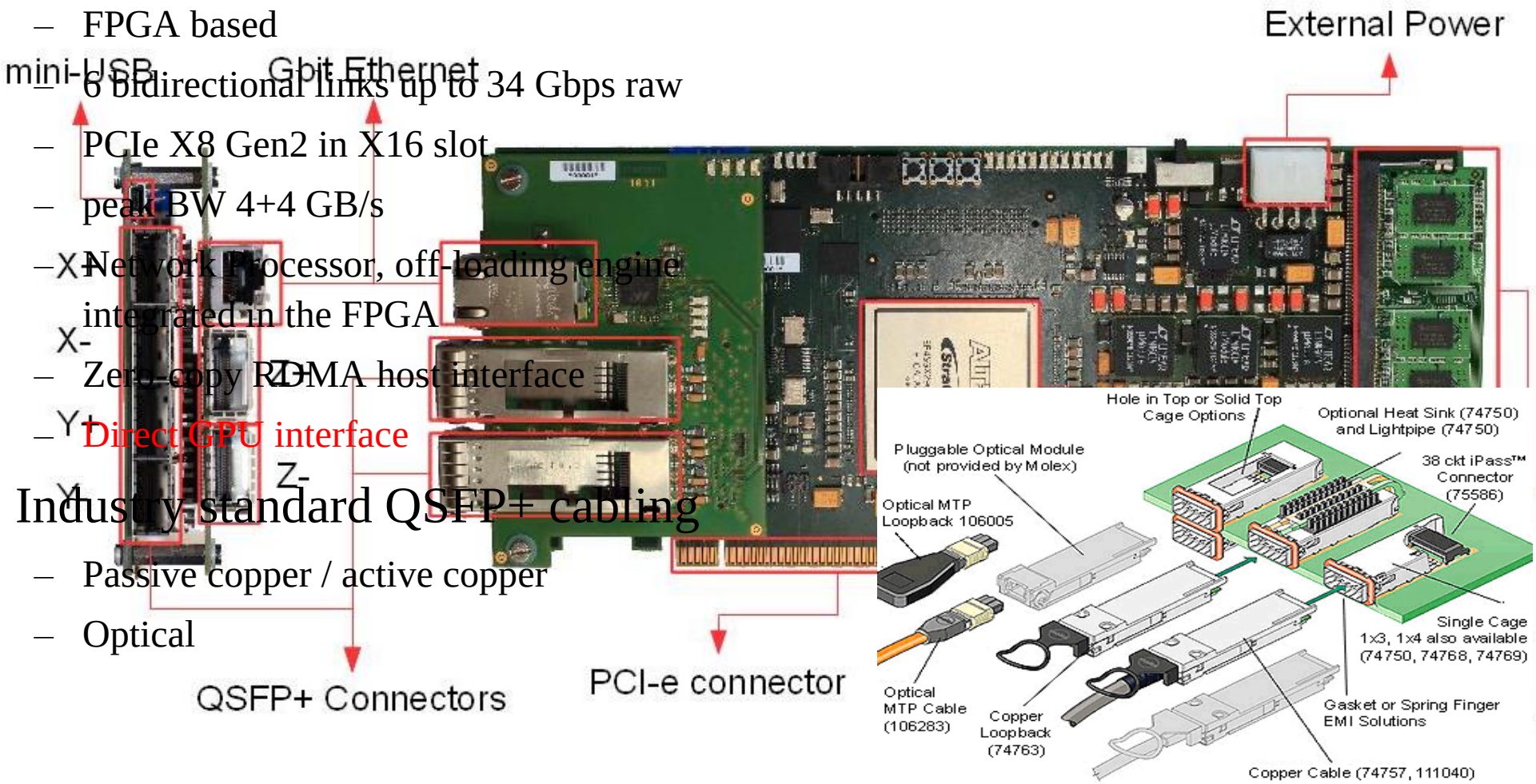
APE group

- APEnet 3D Torus network
  - ideal for large-scale scientific simulations (domain decomposition, stencil computation, …)
  - today scalable up to 32K nodes
  - No external switches! scaling cost: 1 card + 3 cables
- RDMA: Zero-copy RX & TX !
- Small latency & high bandwidth
- GPU clusters features:
  - RDMA support for GPUs! no buffer copies between GPU and host.
  - Very good GPU to GPU latency

APEnet at a glance

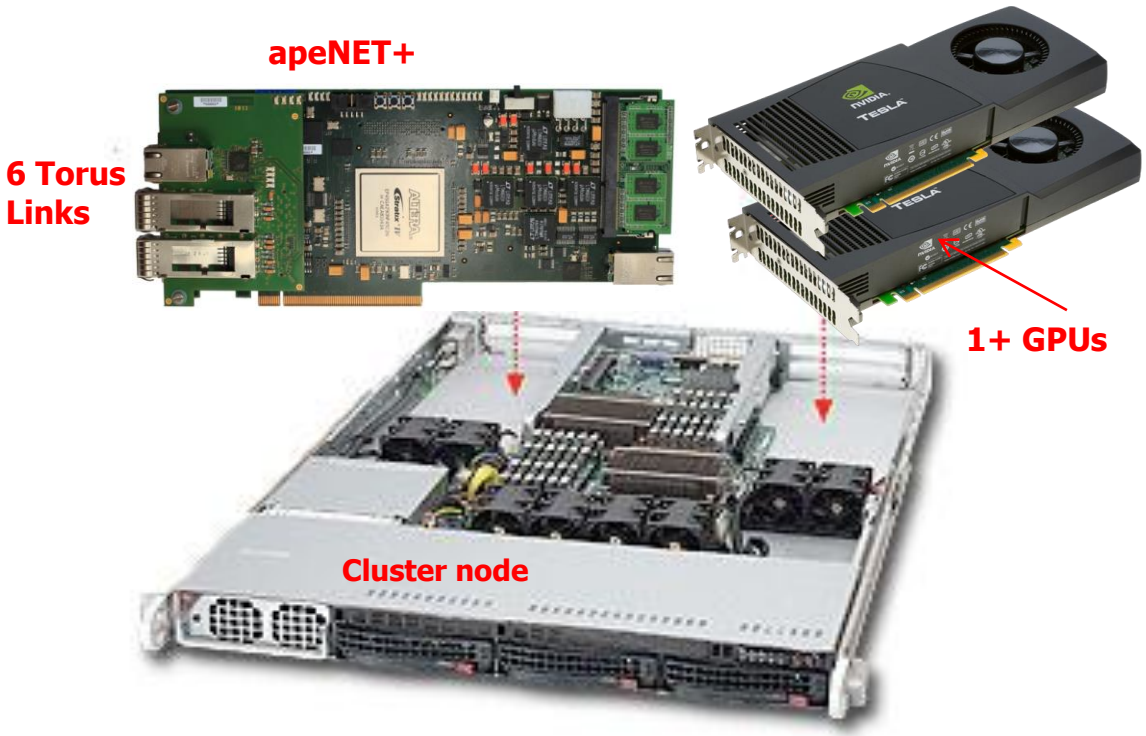
APE group

- APEnet+ card:
  - FPGA based
  - 6 bidirectional links up to 34 Gbps raw
  - PCIe X8 Gen2 in X16 slot
  - peak BW 4+4 GB/s
  - Network Processor, off-loading engine integrated in the FPGA
  - Zero-copy RDMA host interface
  - Direct GPU interface
- Industry standard QSFP+ cabling
  - Passive copper / active copper
  - Optical



GPU cluster *a la* APEnet

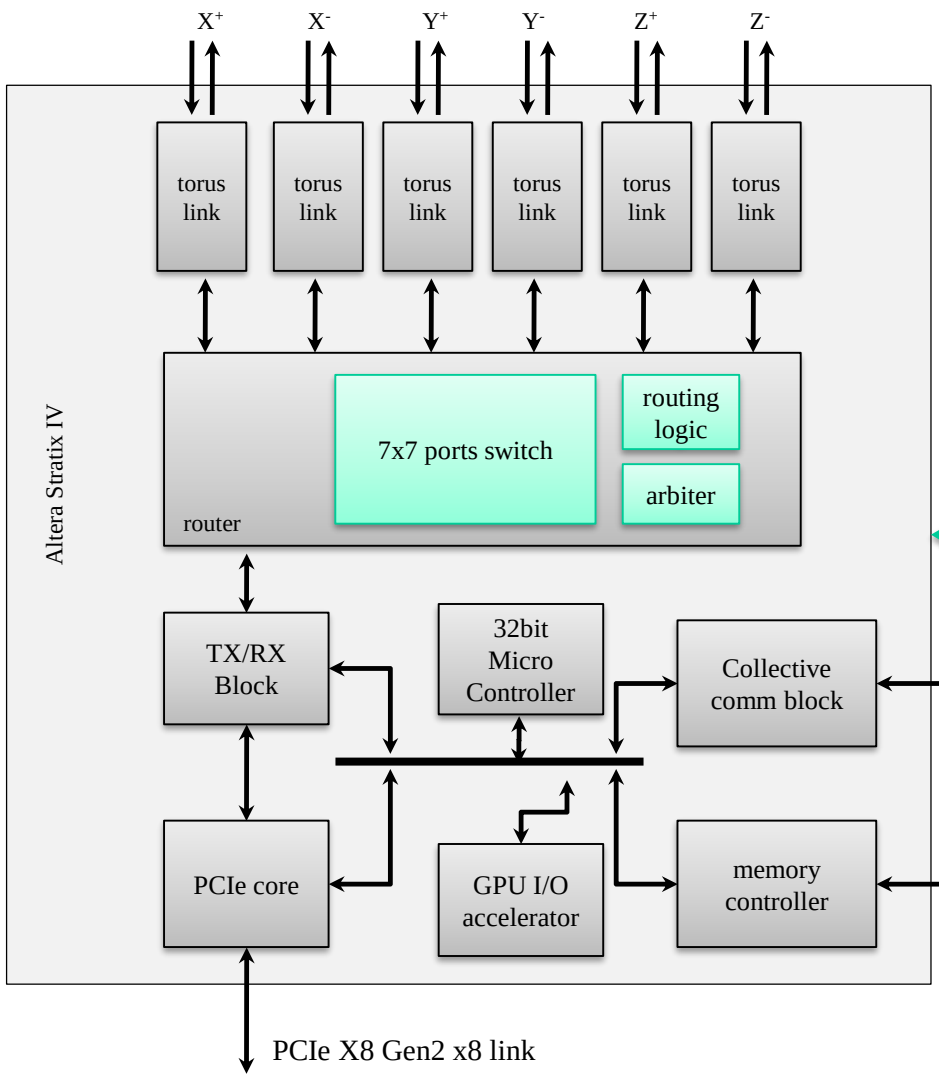
APE group

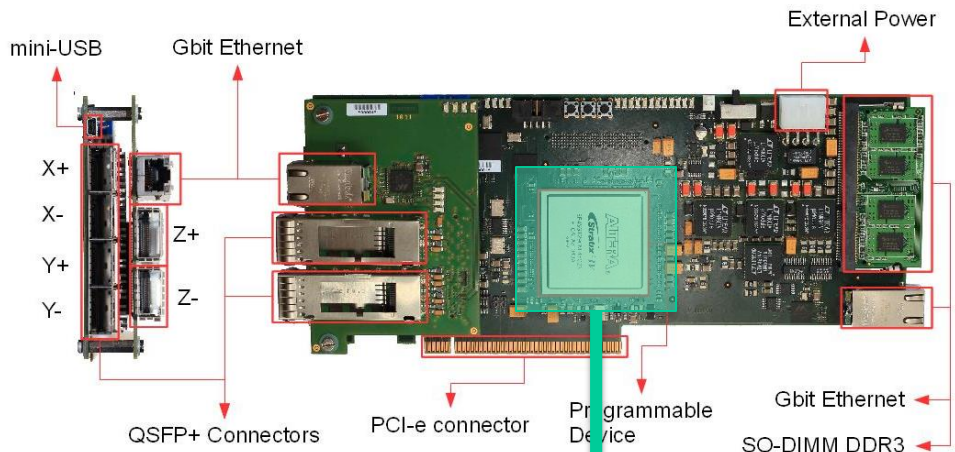


As simple as ☺

APEnet+ card architecture

APE group





1Gb Eth port

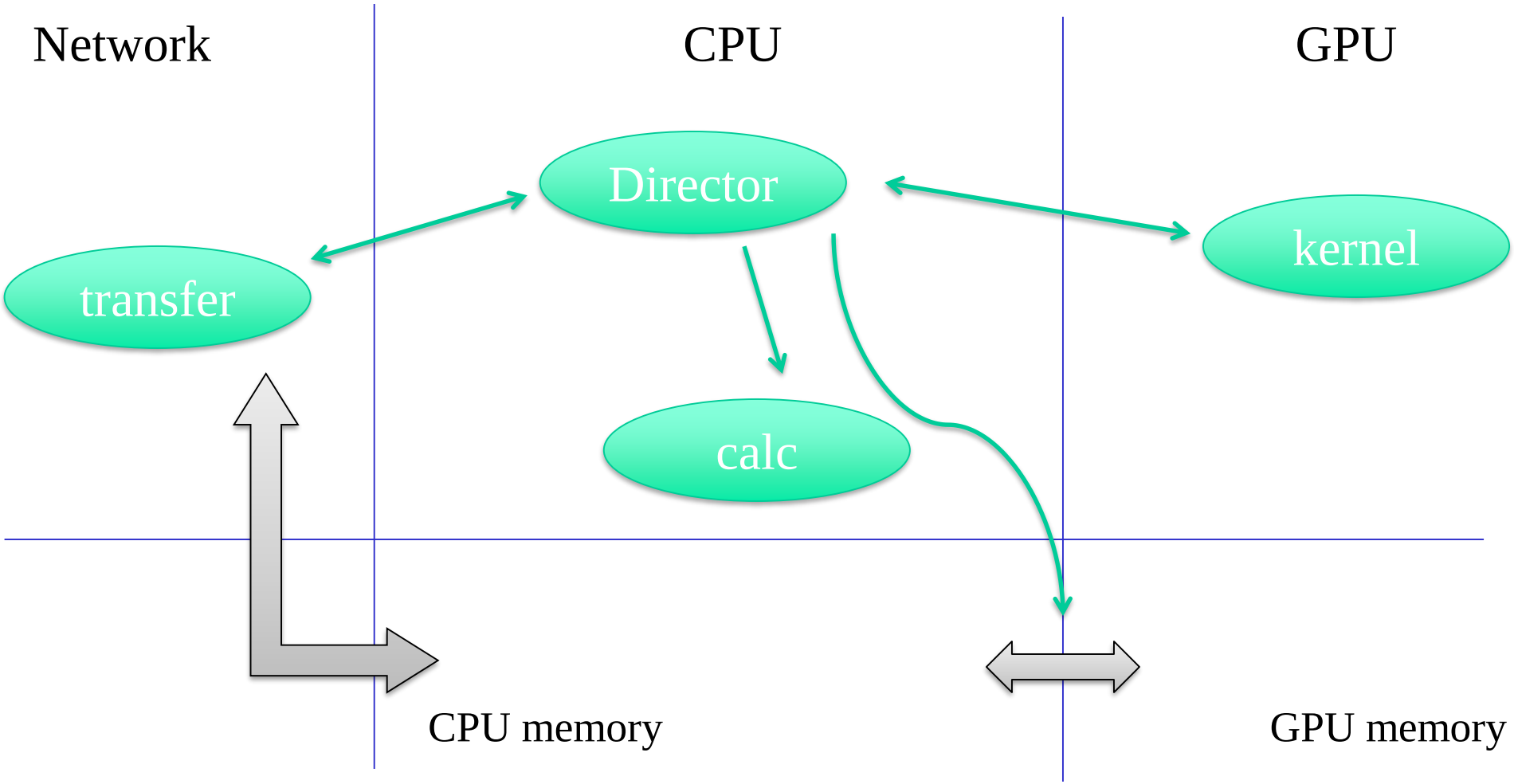
DDR3 Module

GPU support: P2P

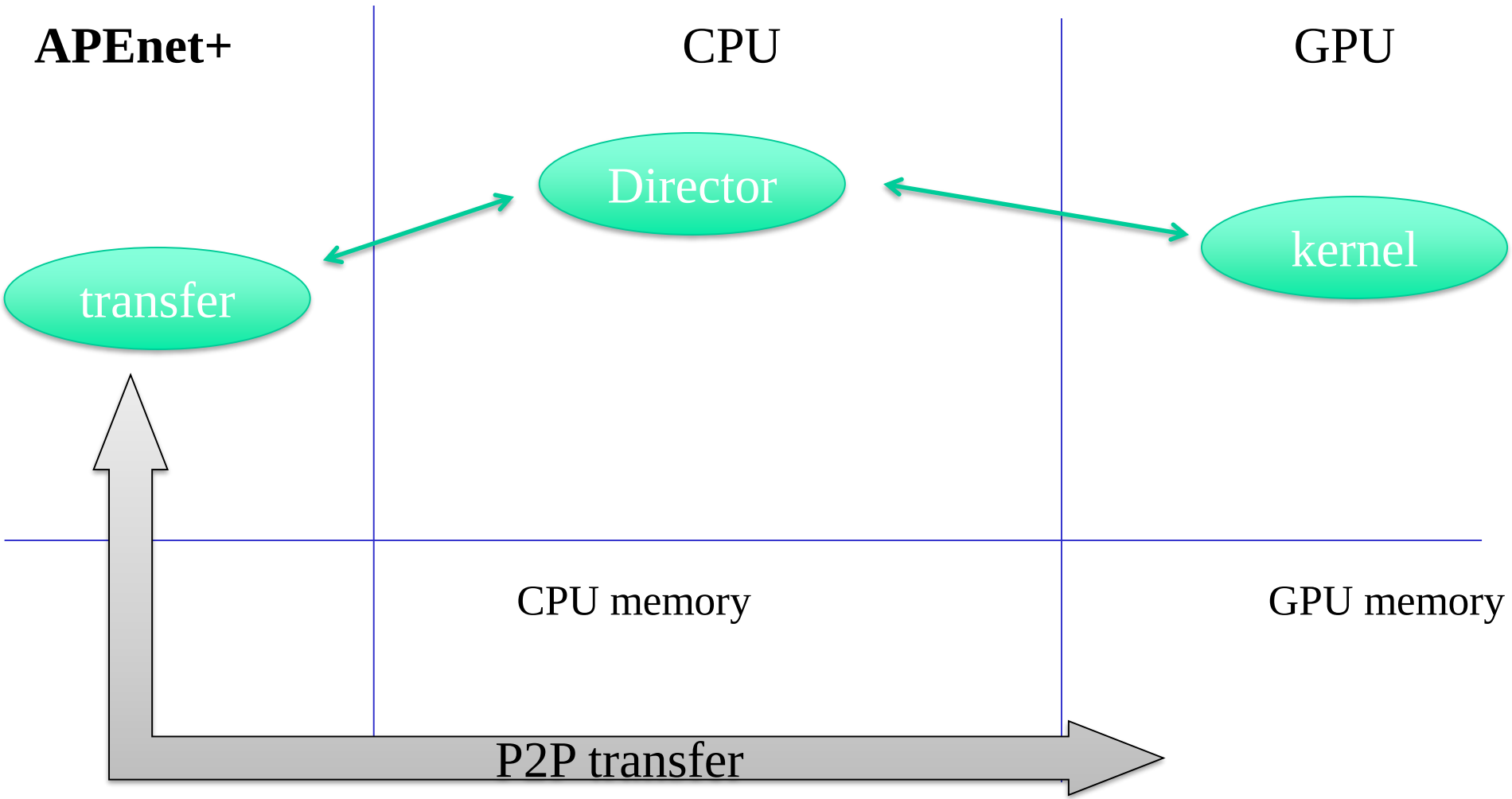
APE group

- CUDA 4.0:
  - Uniform Virtual Address space
  - GPUdirect 2.0 aka P2P among up to 8 GPUs
- CUDA 4.1: P2P protocol with *alien* devices
- P2P between Nvidia Fermi and APEnet+
  - First non-Nvidia device to support it!!!
  - Jointly developed with NVidia
  - APElink+ card acts as a P2P peer
  - APElink I/O to/from GPU FB memory

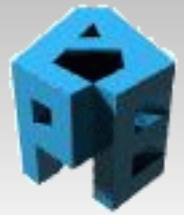
The traditional flow …



… and with APEnet P2P



P2P advantages



APE group

P2P means:

- Data exchange on the PCIe bus
- No bounce buffers on host

So:

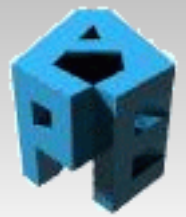
- Latency reduction for small msg
- Avoid host cache pollution for large msg
- Free GPU resources, e.g. for same host GPU-to-GPU memcpy
- Less load on host, more room for comp/comm overlap

Benchmarking platform

APE group

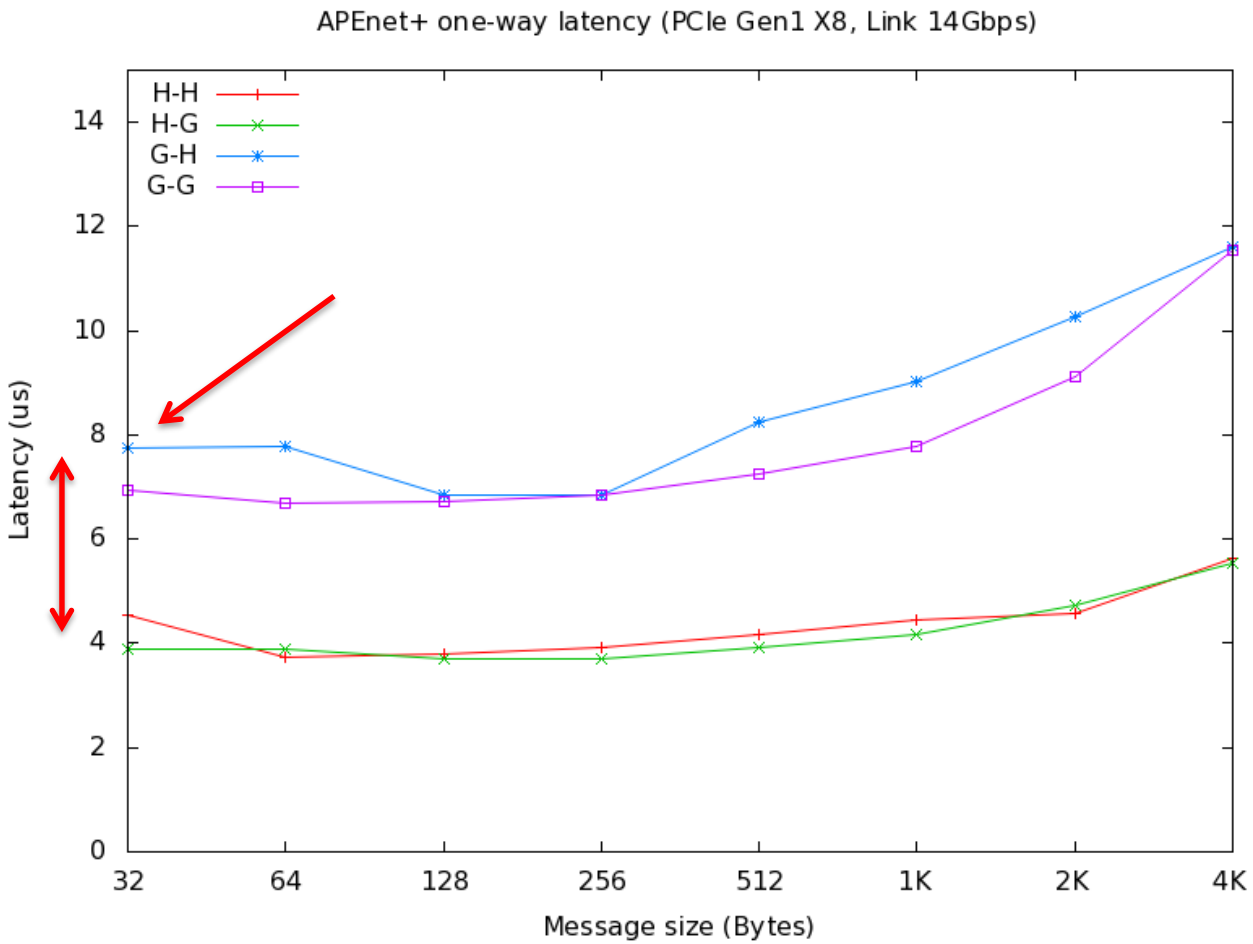
- Preliminary benchmarks:
  - Coded with APEnet RDMA API
  - One-way only but …
  - CUDA 4.1 pre-release
- Caveat: used APEnet test cards with reduced capabilities:
  - PCIe X8 Gen1
  - Link raw speed @14Gbps
- 2 slightly different servers
  - SuperMicro motherboards
  - CentOS 5.7 x86_64
  - Dual Xeon 56xx 24GB
  - Nvidia C2050 on X16 Gen2 slots

Latency benchmark

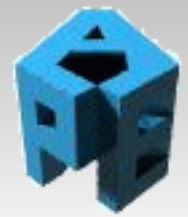


APE group

- OSU-like one-way latency test
  - re-coded using RDMA PUT
  - No small msg optimizations
  - No big difference with round-trip
- 6.7 us on GPU-GPU test!
- GPU TX demanding !! still …

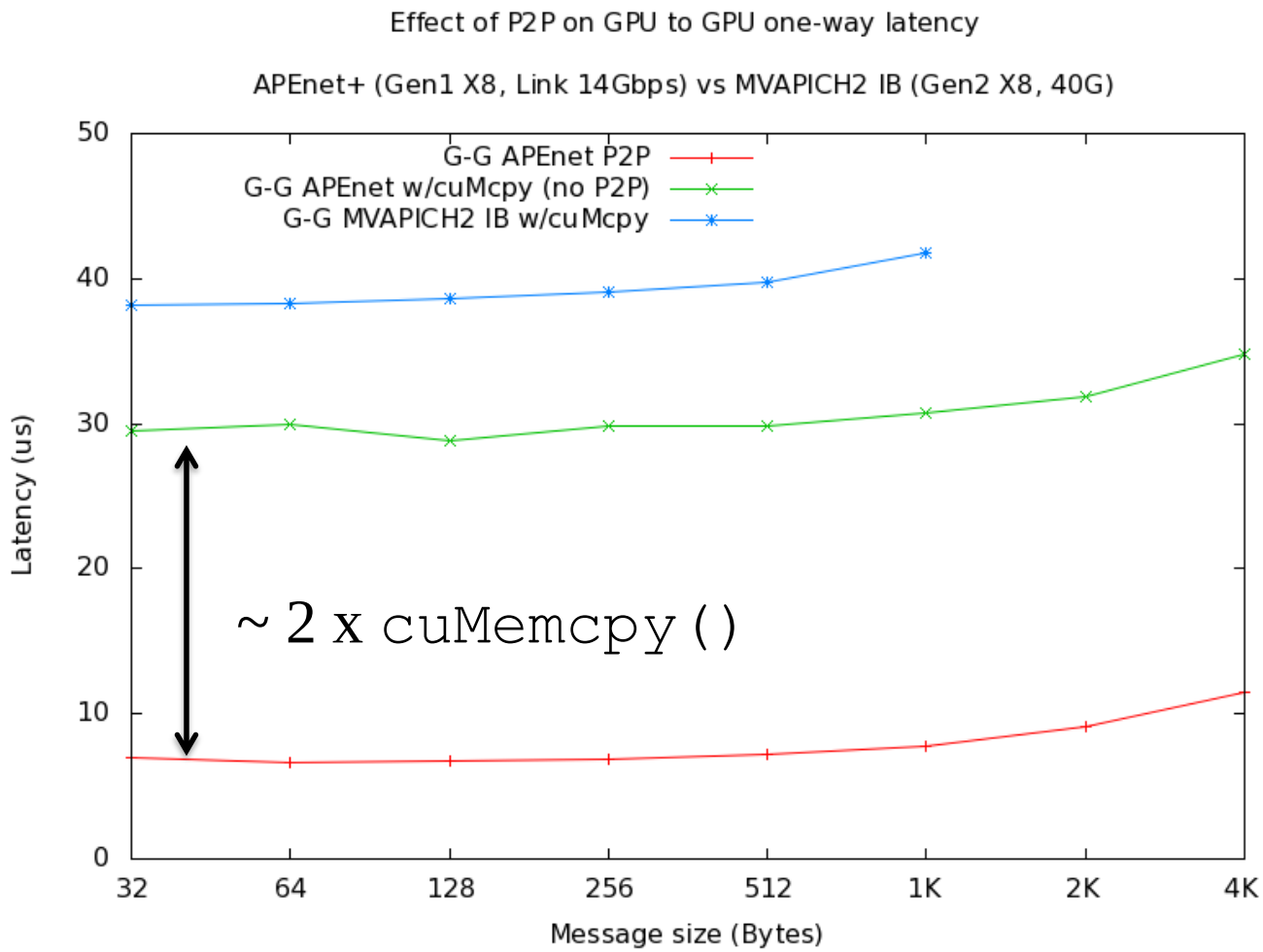


Latency benchmark: P2P effects



APE group

- No P2P = `cuMemcpyD2H/H2D()` on host bounce buffers
- Buffers pinned with `cuMemHostRegister`
- `cuMemcpy()` costs ~ 10us
- MVAPICH2 points from the Ohio State U. web site*

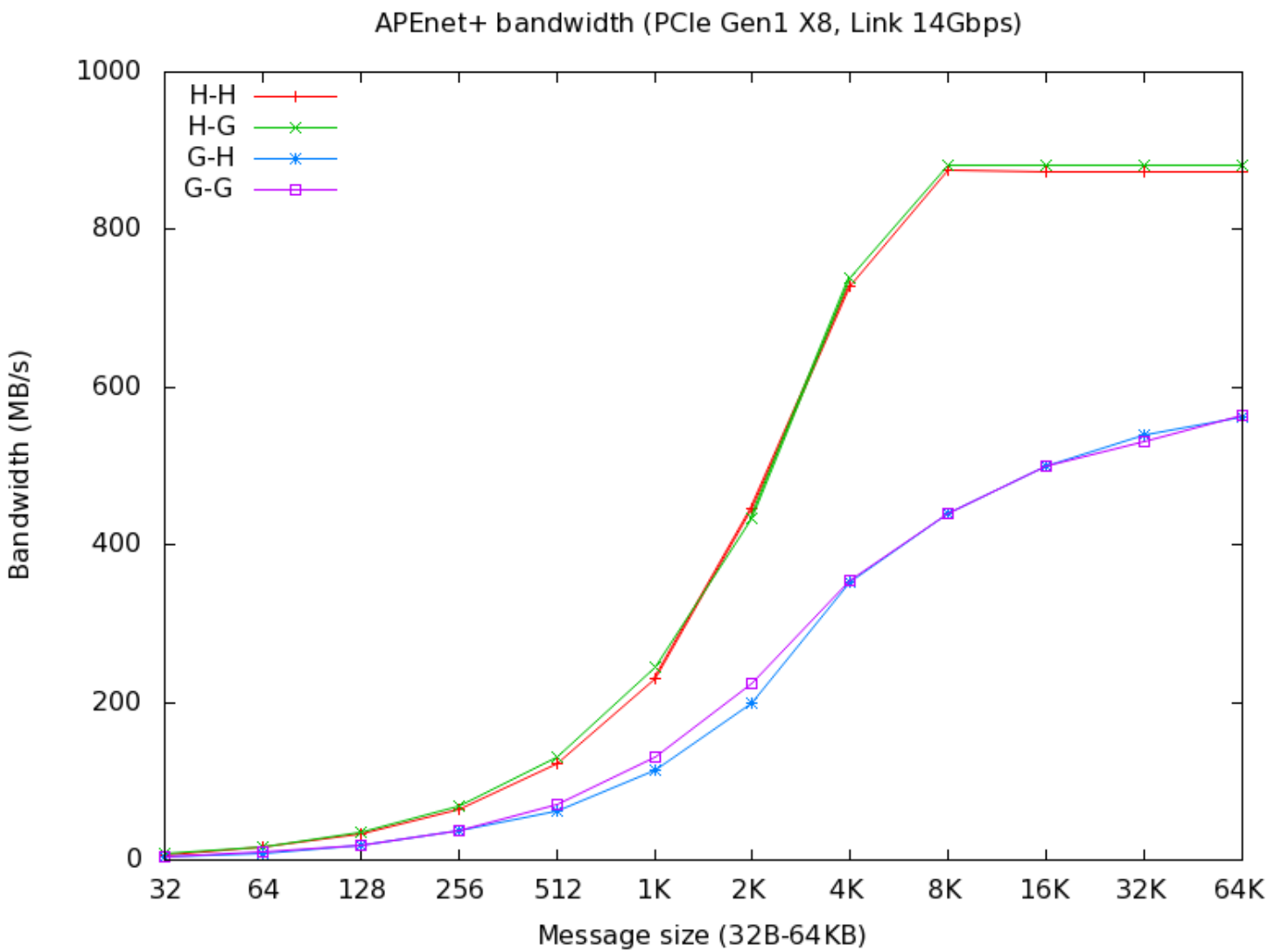


* http://mvapich.cse.ohio-state.edu/performance/mvapich2/inter_gpu.shtml

Bandwidth benchmark

APE group

- Very preliminary
- GPU RX is great! Better than Host RX
- Host TX is better but suffers PCIe Gen1 & link speed cap of APEnet test card
- Link speed over 2GB/s on final APEnet HW



Low-level profiling

APE group

OSU-like oneway latency benchmark

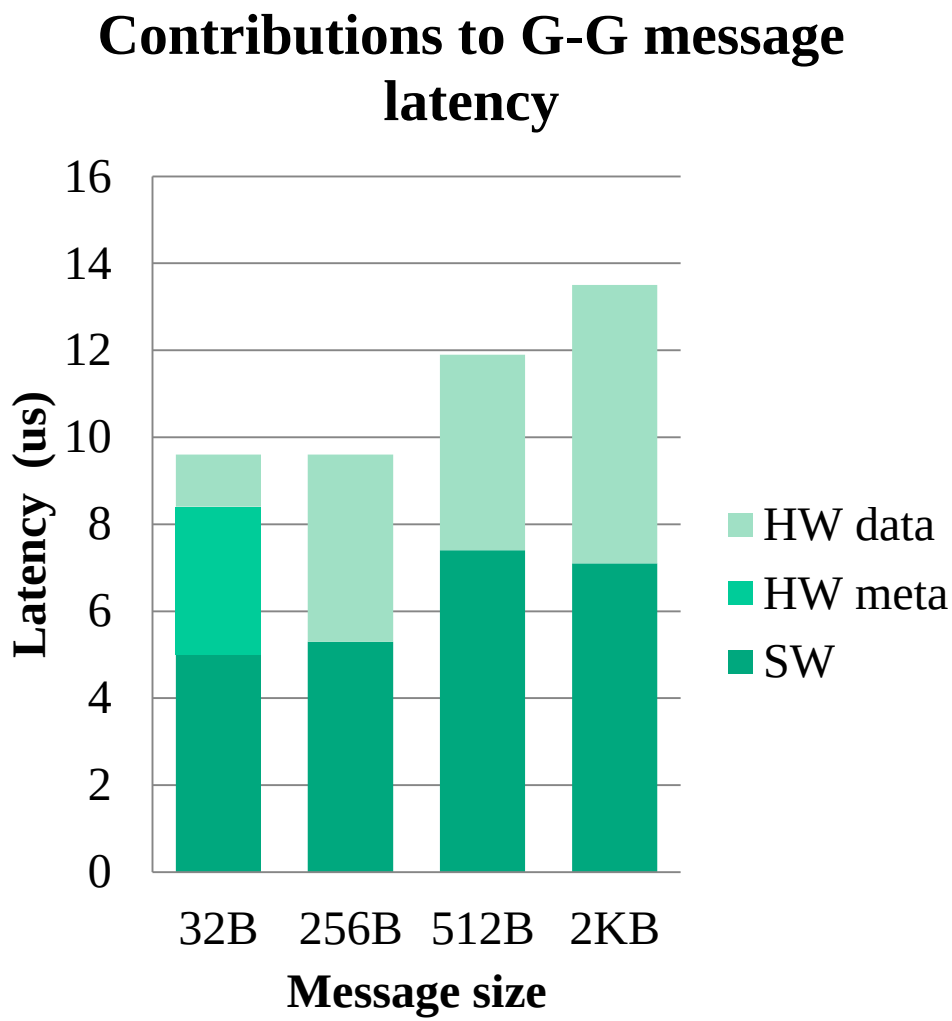
GPU to GPU

Showing TX side

Sizes: 32B to 2KB

Break-out contributions to message time:

- Software: application, user-space library, kernel driver
- HW data: data related time
- HW metadata: remaining time



SW: RDMA API

APE group

- RDMA Buffer management:
  - expose memory buffers to remote access
  - am_register_buf(), am_unregister_buf()
  - 2 types: SBUF use-once, PBUF are targets of RDMA_PUT
  - Typically at app init time
- Comm primitives:
  - Non blocking, async progress
  - am_send() to SBUF
  - am_put() to remote PBUF via buffer virtual address
  - am_get() from remote PBUF (future work)
- Event delivery:
  - am_wait_event()
  - Generated on: comm primitives completion, RDMA buffers access

SW: RDMA API

APE group

Typical stencil app

- Init:
  - Allocate buffers for *ghost cells*
  - Register buffers
  - Exchange buffers host address
- Computation loop:
  - Calc boundary
  - cuMemcpy of boundary to buffer
  - am_put() buffer to neighbors
  - Calc bulk
  - Wait for put done and local ghost cells written
  - cuMemcpy of rx buffer to GPU ghost cells

Same app with P2P

- Init:
  - cudaMalloc() buffers on GPU
  - Register GPU buffers
  - Exchange GPU buffers address
- Computation loop:
  - Launch calc_bound kernel on stream0
  - Launch calc_bulk kernel on stream1
  - cudaStreamSync(stream0)
  - am_put(rem_gpu_addr)
  - Wait for put done and buffer written
  - cudaStreamSync(stream1)

Thanks to P2P!

SW: MPI

APE group

OpenMPI 1.5

- APEnet BTL-level module
- 2 protocols, based on threshold size:
  - Eager: small message size, uses plain send, no sync
  - Rendezvous: pre-register dest buffer, use RDMA_PUT, synch needed
- Working on integration of P2P support
  - Use of CUDA 4.x UVA

Status & future plans

APE group

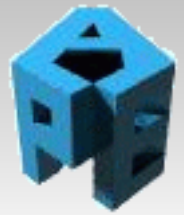
Status:

- Bring-up phase for the APEnet card
- Mature FPGA firmware (beta)
- OpenMPI coming soon (alpha)

Future:

- 8-16 node GPU+APEnet cluster available 1Q'12
- HW roadmap: Gen2 x16, Gen3
- GPU initiated communications
- fault tolerance
- Application acceleration via reconf. comp. & new comm primitives

Game over



APE group

- See you at IGI booth #752 at 12:30PM for
  - Close view of the real APEnet+ card
  - Going through sample code
  - Q&A
- APEnet web site: http://apegate.roma1.infn.it/
- Contact us at apenet@apegate.roma1.infn.it