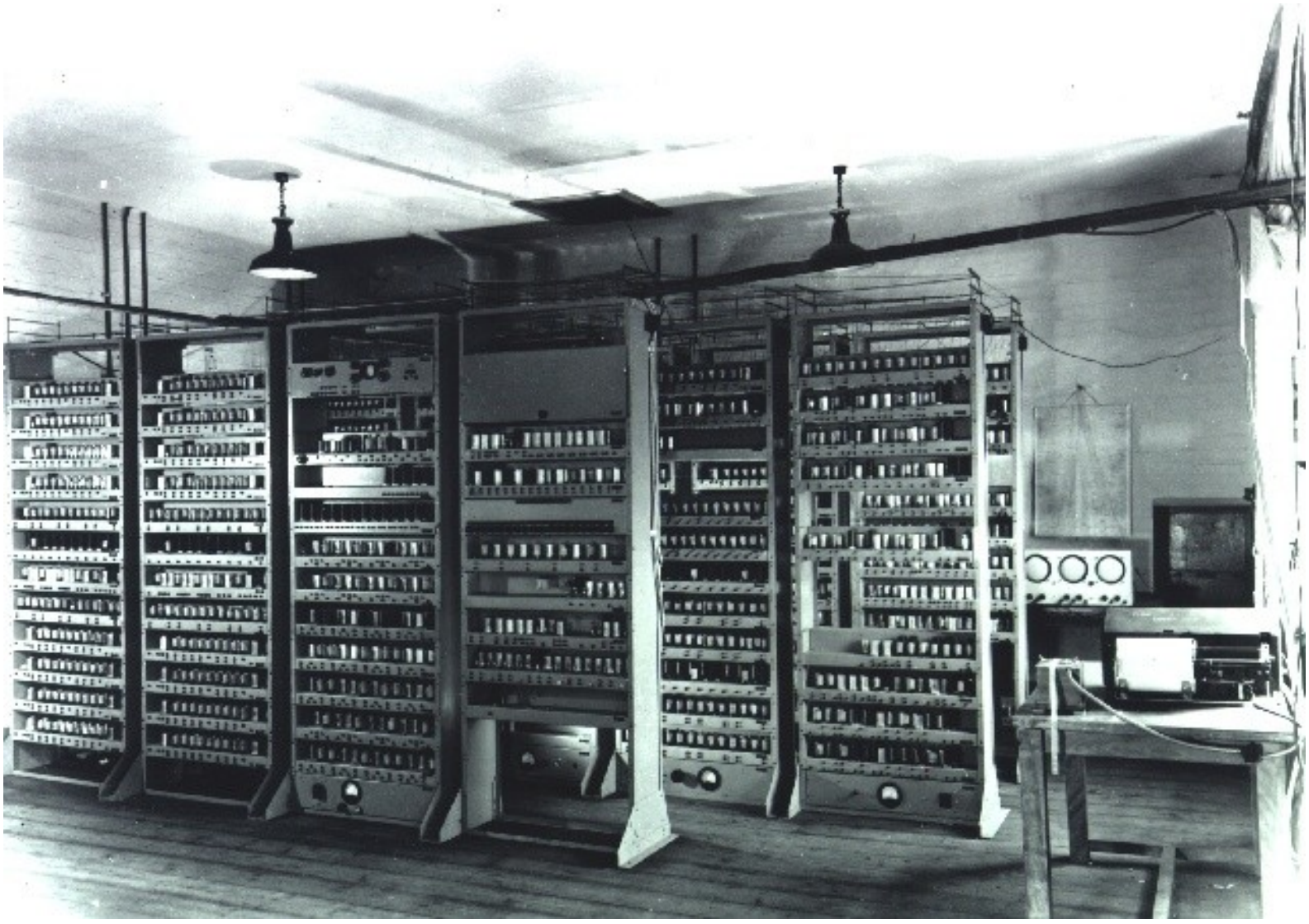
Stuart Rankin
High Performance Computing Service
University of Cambridge

sjr20@cam.ac.uk
www.hpc.cam.ac.uk

Cambridge HPCS

- Cambridge HPC: 1949 - 1996 - 2006 -
- Mission
  - Support and service delivery of a large HPC resource to the University of Cambridge research community
  - To be operated as a self-sustaining cost centre and funded as a MRF
- Multidisciplinary
- Cambridge and beyond



Resources

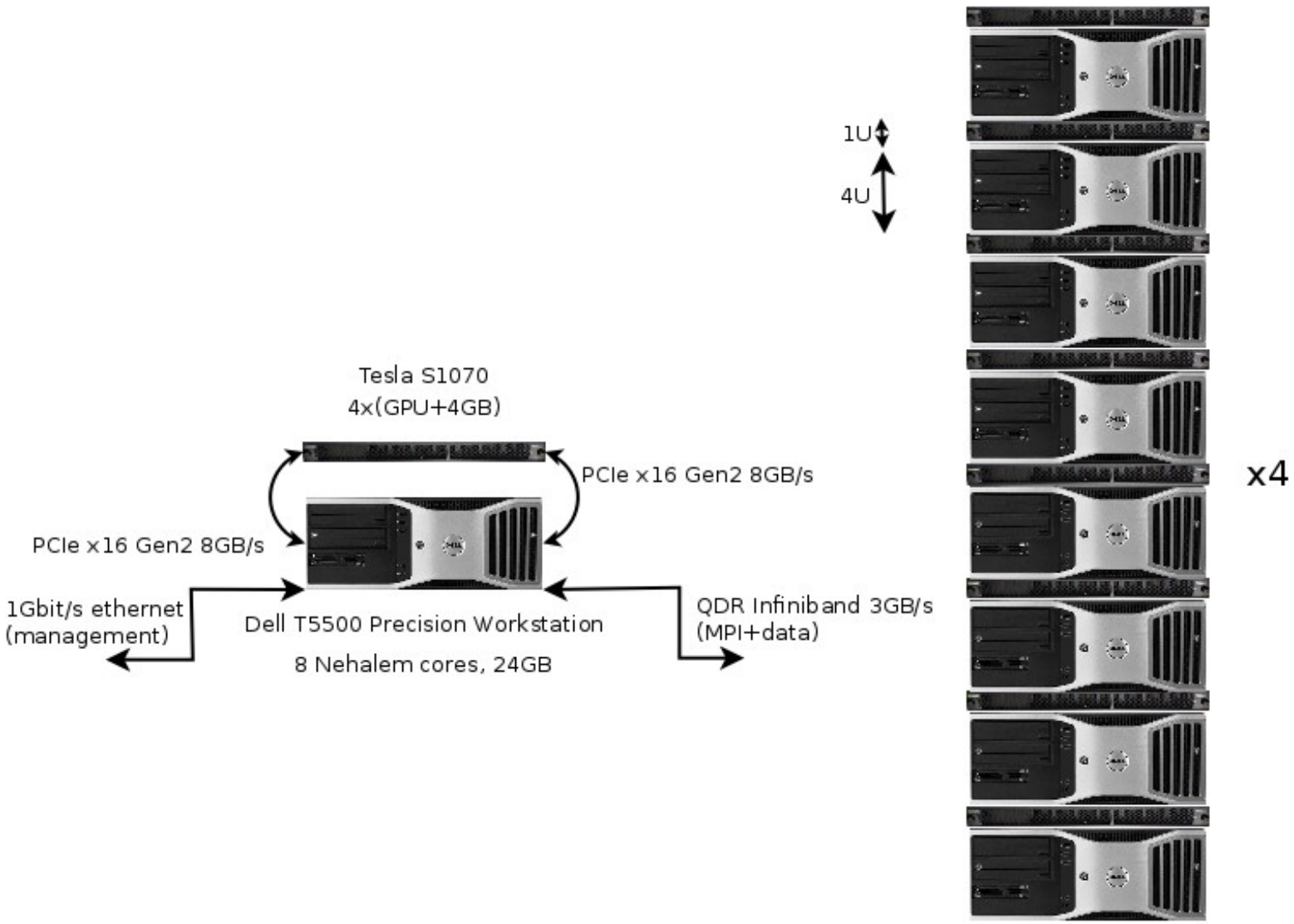
- Darwin
  - in production since February 2007
  - refreshed August 2010
  - 2048 3GHz Intel Woodcrest cores
    - Qlogic SDR infiniband
  - 1536 2.67GHz Intel Westmere cores
    - Mellanox ConnectX2 QDR infiniband
  - 440 TB shared storage (Lustre)
  - 32x Tesla S1070 GPU cluster with QDR
  - C61xx, C410x, Fermi M20xx test hardware

Dell/Cambridge Solution Centre

- Consultation…
  - integration
  - operation
  - storage
  - GPGPU
  - visualisation
  - application optimisation and benchmarking
- Meetings, whitepapers, POCs

GPU HPC Cluster

Cambridge GPU/Nehalem Cluster



Tesla Production Cluster

- In use since March 2010
- 32 Dell T5500 workstations
  - 8 cores of Nehalem X5550, 24GB RAM
  - Mellanox IB (MPI and Lustre filesystem)
  - S1070 (4x GPUs; 2x PCIe Gen2 x16)
  - Max b/w to S1070 is 0.75*2*8GB/s=12GB/s
  - Max b/w to 1 GPU ~ 5700MB/s (with pinning)
- Successive NVIDIA drivers have improved stability
- 28 nodes still functional (December 2011)

Tesla S1070

- Four GPUs
  - Each with 4GB memory, 30 multiprocessors
  - 8 cores per multiprocessor
  - Memory not ECC
  - 933 GFlop/s (single precision)
  - 78 GFlop/s (double precision)
- CPU threads submit a *CUDA kernel* to a GPU
  - Grid of thread blocks, 1 block per multiprocessor (vector/SIMD parallelism)
  - Threads in a block share memory

Dell C6100/C410x

- Single C6100 server (4 per chassis)
  - 12 cores of Westmere X5650
  - Mellanox IB (MPI and Lustre filesystem)
  - 1x PCIe Gen2x16 interface per HIC
    - ~5900MB/s (with pinning)
- C410x can hold 16 GPUs (each PCIe x16) and 8 interfaces
- C6145 (2 per chassis) can hold more HICs

Fermi M20XY

- 3GB memory (M2050), 6GB (M2070)
- ECC support
  - removes 12.5% of memory when enabled
  - can be disabled
- 14 multiprocessors
- 32 cores per multiprocessor
- 1030 GFlop/s (single precision)
- 515 GFlop/s (double precision)

PBS Issues

- Health checks
  - non-invasive (don’t disturb running jobs)
- Prologues
  - aggressive; verify performance
- Epilogues
  - cuda_memtest and cuda_memscrubber
- Default and exclusive(_thread) modes
  - now also exclusive_process
  - Turbostream wants exclusive, NAMD wants default
- We don't schedule individual GPUs
  - our smallest unit of allocation is (Node+Tesla)

Memory Pinning

- *a.k.a GPUDirect v1*
- Performance of host to GPU communication improved by allocating pinned memory registered to the GPU driver
  - i.e. *cudaMallocHost* instead of *malloc*
  - reduces memory-memory copies on the host
- E.g. using mpi_pinned.c (download from NVIDIA)
  - 2700MB/s unpinned vs 5700MB/s pinned (T5500/S1070)
  - 4650MB/s unpinned vs 5950MB/s pinned (C6100/Fermi)

GPUDirect v1

- Performance benefits of eliminating memory to memory copies in host memory extend to MPI comms
  - *if* the IB NIC and GPU can share the same host memory
  - MPI send/recv using mpi_pinned.c
    - 900MB/s unpinned vs 3200MB/s pinned (T5500/S1070) [!]
    - 2800MB/s unpinned vs 3200MB/s pinned (C6100/Fermi)

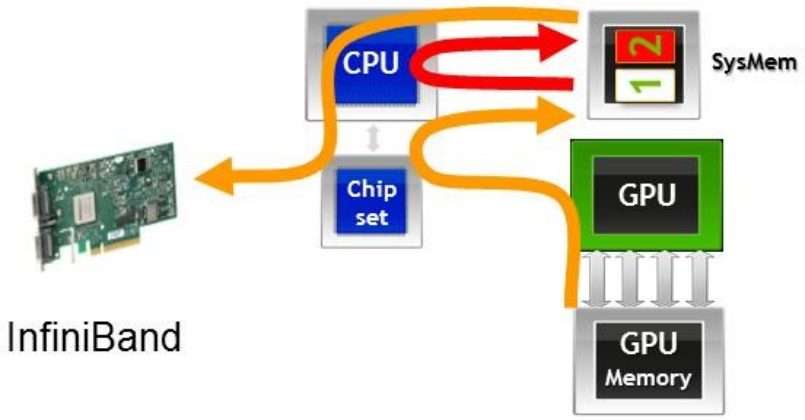
GPUDirect v1

(figure from Nvidia)

Without GPUDirect

Same data copied three times:

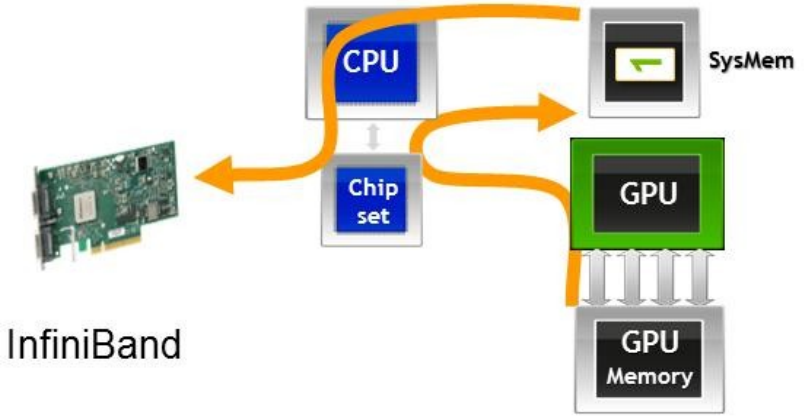
1. **GPU writes to pinned sysmem1**
2. **CPU copies from sysmem1 to sysmem2**
3. **InfiniBand driver copies from sysmem2**



With GPUDirect

Data only copied twice

Sharing pinned system memory makes sysmem-to-sysmem copy unnecessary

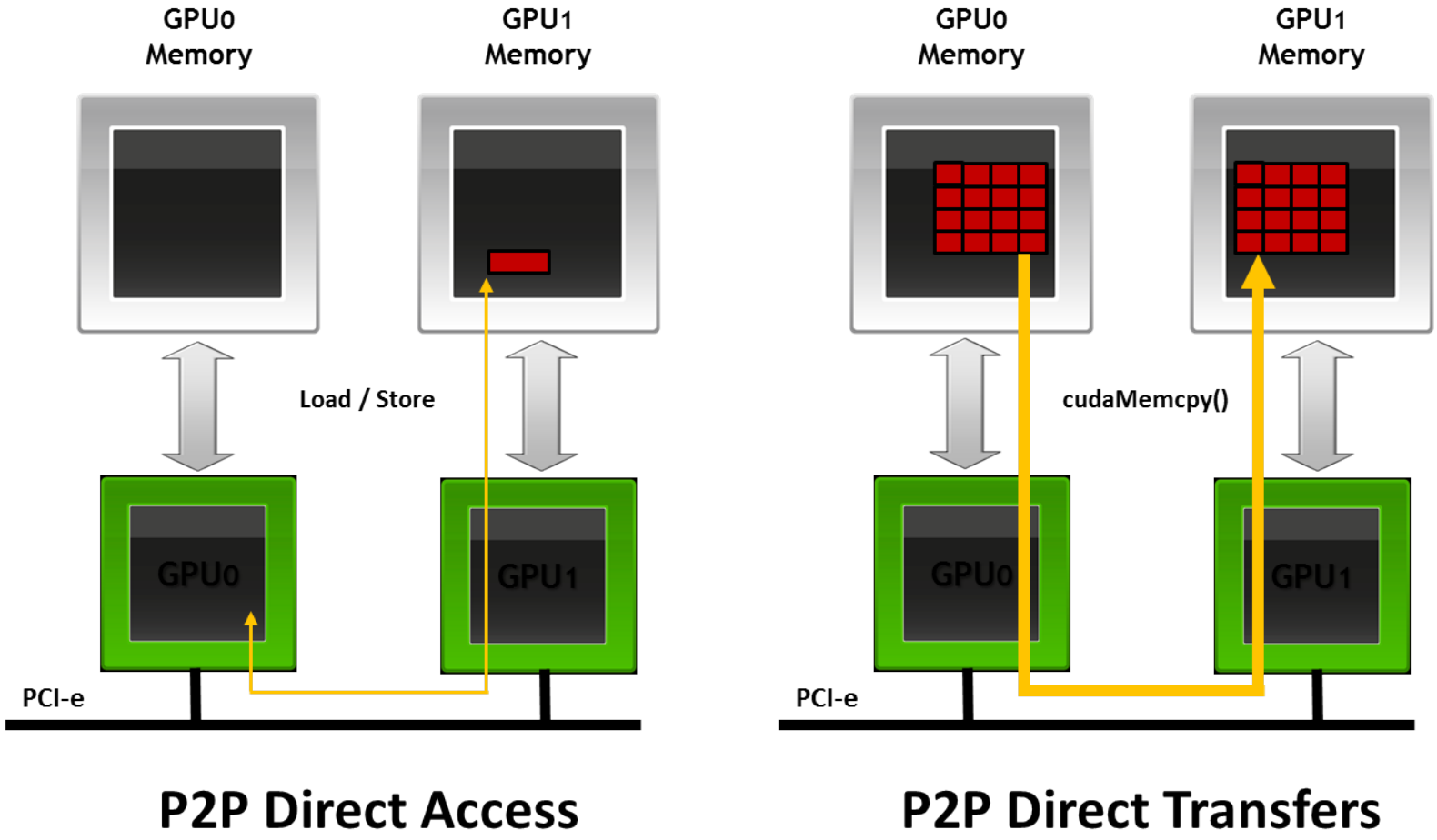


GPUDirect v1

- Before Nvidia driver 270.41.19
  - CUDA runtime pinned memory in a way incompatible with 3rd party drivers
  - worked around with modified kernel and IB drivers
- With Nvidia driver 270.41.19
  - environment variable CUDA_NIC_INTEROP=1 changes CUDA pinning to be compatible with unmodified kernels/drivers
- At Nvidia driver 285.05.09
  - CUDA_NIC_INTEROP has vanished but compatibility remains

GPUDirect v2 (CUDA 4.0)

(figure from Nvidia)



Affinity

- Must already worry about NUMA and process pinning
  - process should be physically near its memory to minimise slower remote memory accesses
- Also CPU-GPU access is not necessarily uniform
  - e.g. on T5500/S1070
    - ... numactl --cpunodebind=0 mpi_pinned
      - Host->device b/w 5051.525561 MB/s
    - ... numactl --cpunodebind=1 mpi_pinned
      - Host->device b/w 5724.753836 MB/s

Remote Visualization

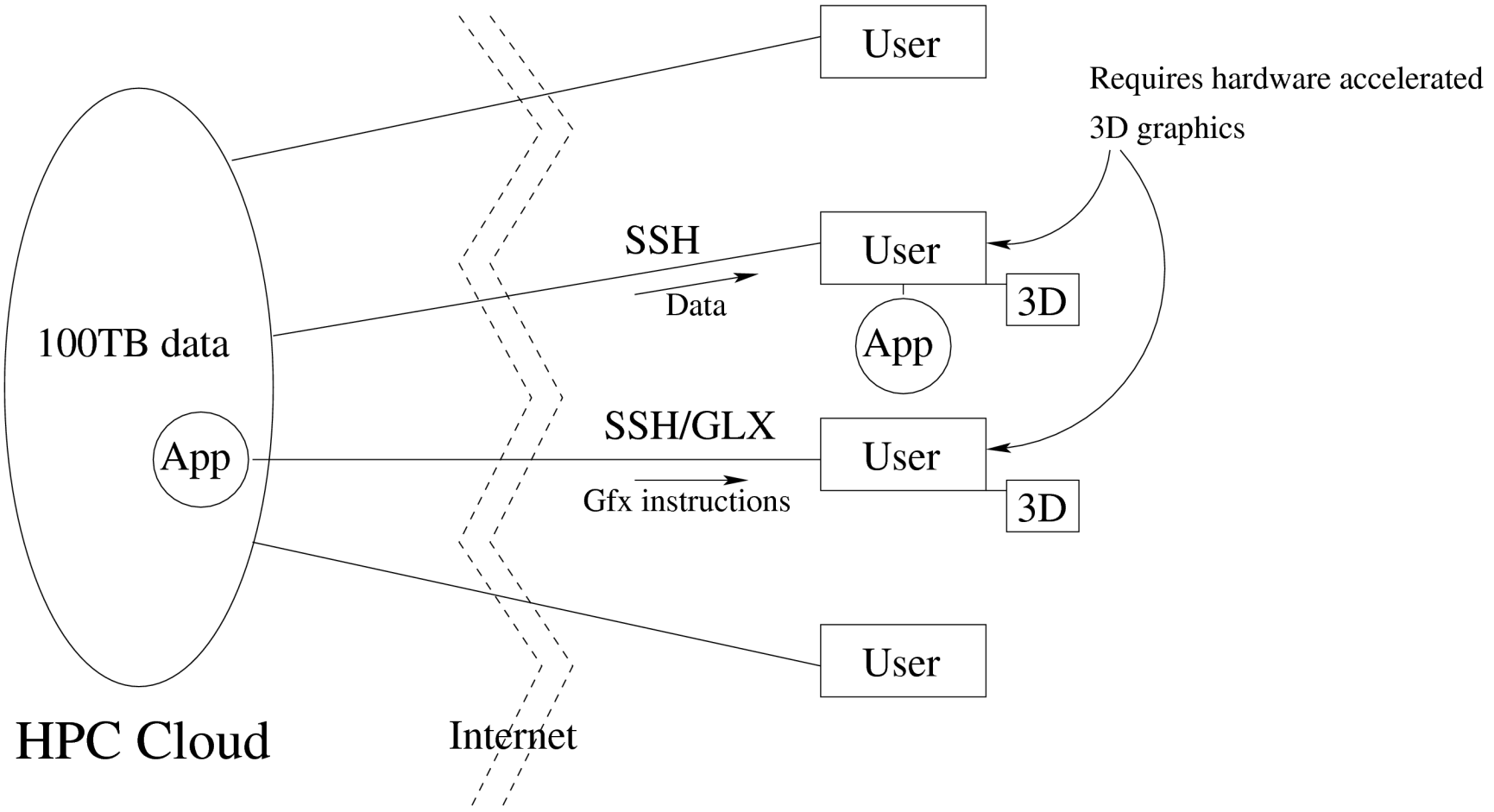
Visualization

- Many scientific disciplines require complex 3D graphics in real time from large datasets
  - Computational Fluid Dynamics (CFD)
  - Cosmology/Astrophysics
  - Molecular Modelling
  - Medical Imaging
- Require hardware-accelerated 3D rendering

Local Visualization

- *Locally* requires a well-provisioned workstation
  - CPU/memory
  - Storage
  - 3D graphics card
  - Good network

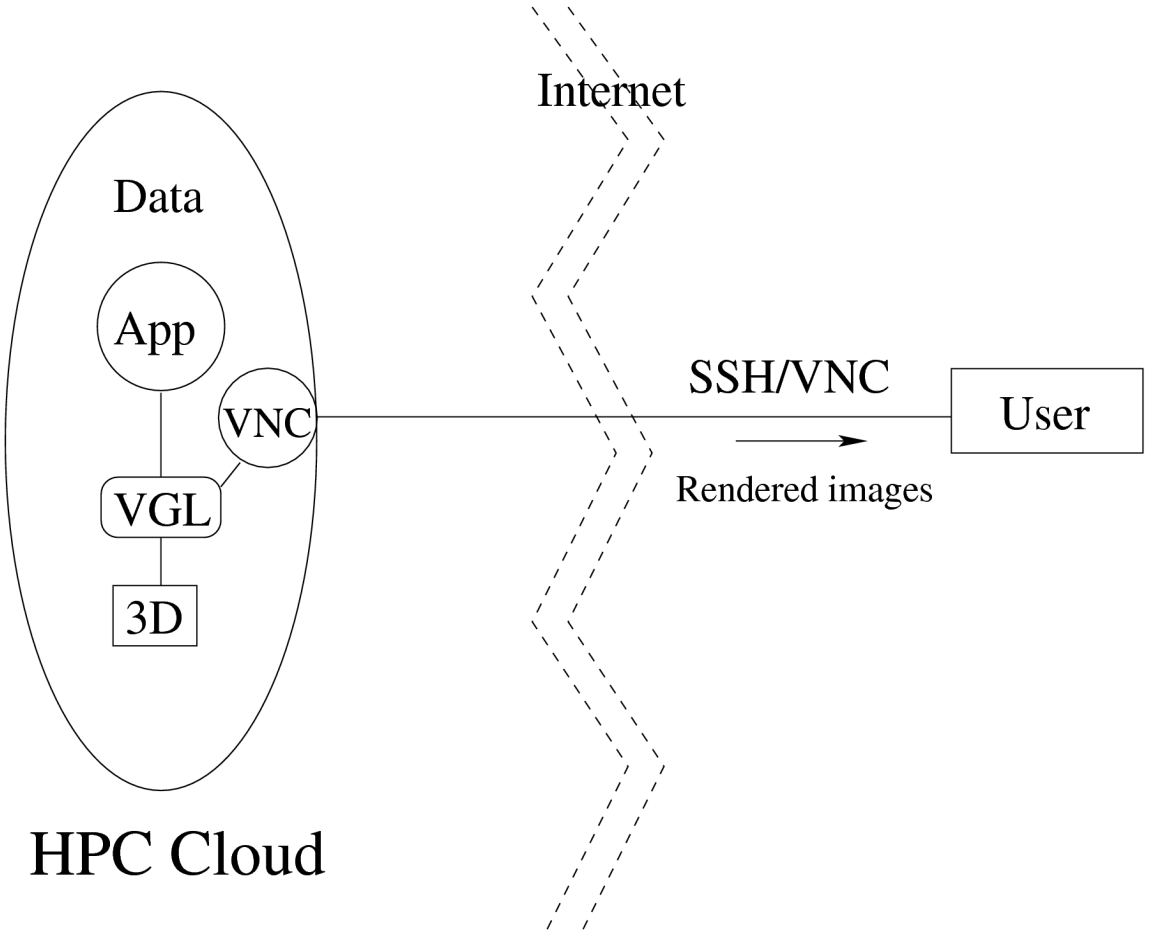
Remote Visualization



Remote Visualization - Traditional

- Or *remotely* run the analysis at the HPC end
  - Just send graphics instructions
  - Still need 3D graphics + good network
  - Technically problematic (may not work)
  - 3D not hardware accelerated!

Remote Visualization with VGL



VirtualGL (VGL)

- VirtualGL (www.virtualgl.org) redirects 3D rendering to a server-side card and returns the final image
- Anything the server can render, the client can see
- With VNC pushes all heavy client requirements including 3D capability back to the server
- Plus simplified setup, persistence, collaboration
- Accelerated JPEG codec is used to manage network bandwidth.