Workshop on Advanced Computing for Accelerators

Introduction to HPC and parallel architectures: hardware

Jonathan Follows (Hartree Centre)

Outline

- Motivation behind parallel architectures
- Architectural developments
- Technology trends
- The Hartree Centre's systems

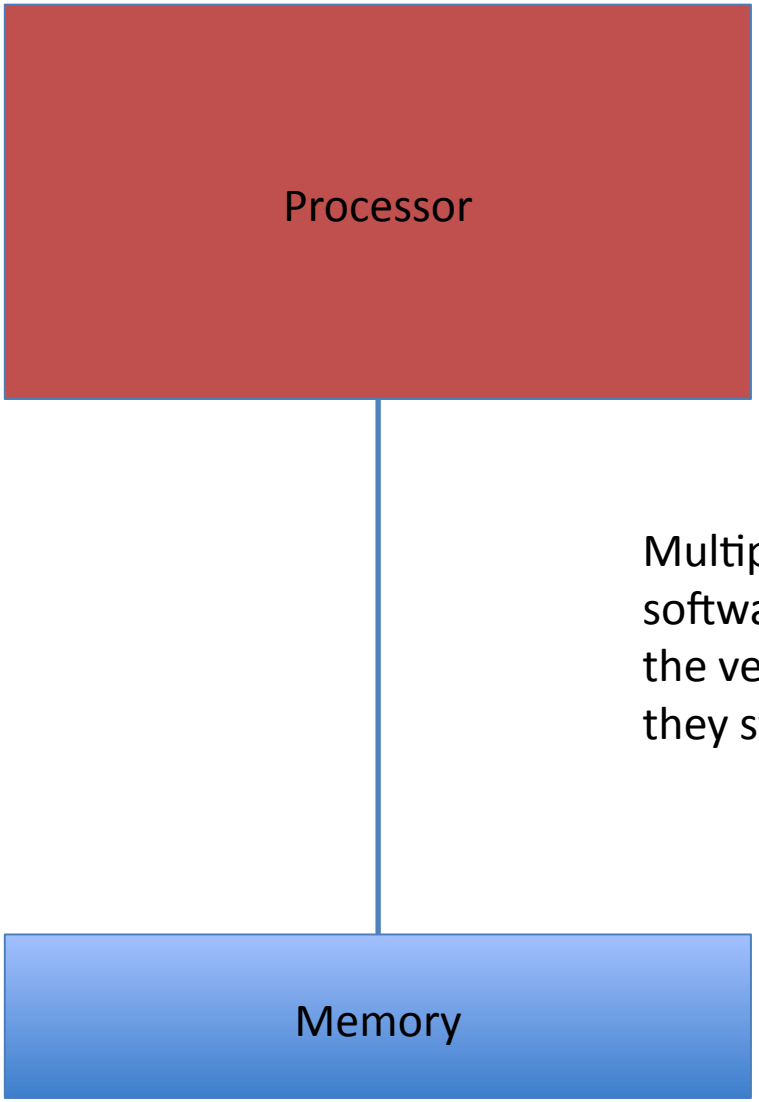
Motivation

- To harness computer processing power to simulate the world at some level
- To be able to perform simulations more quickly and cheaply than by actual experimentation
- To maximise performance by maximising the use of floating point arithmetic

Architectural developments

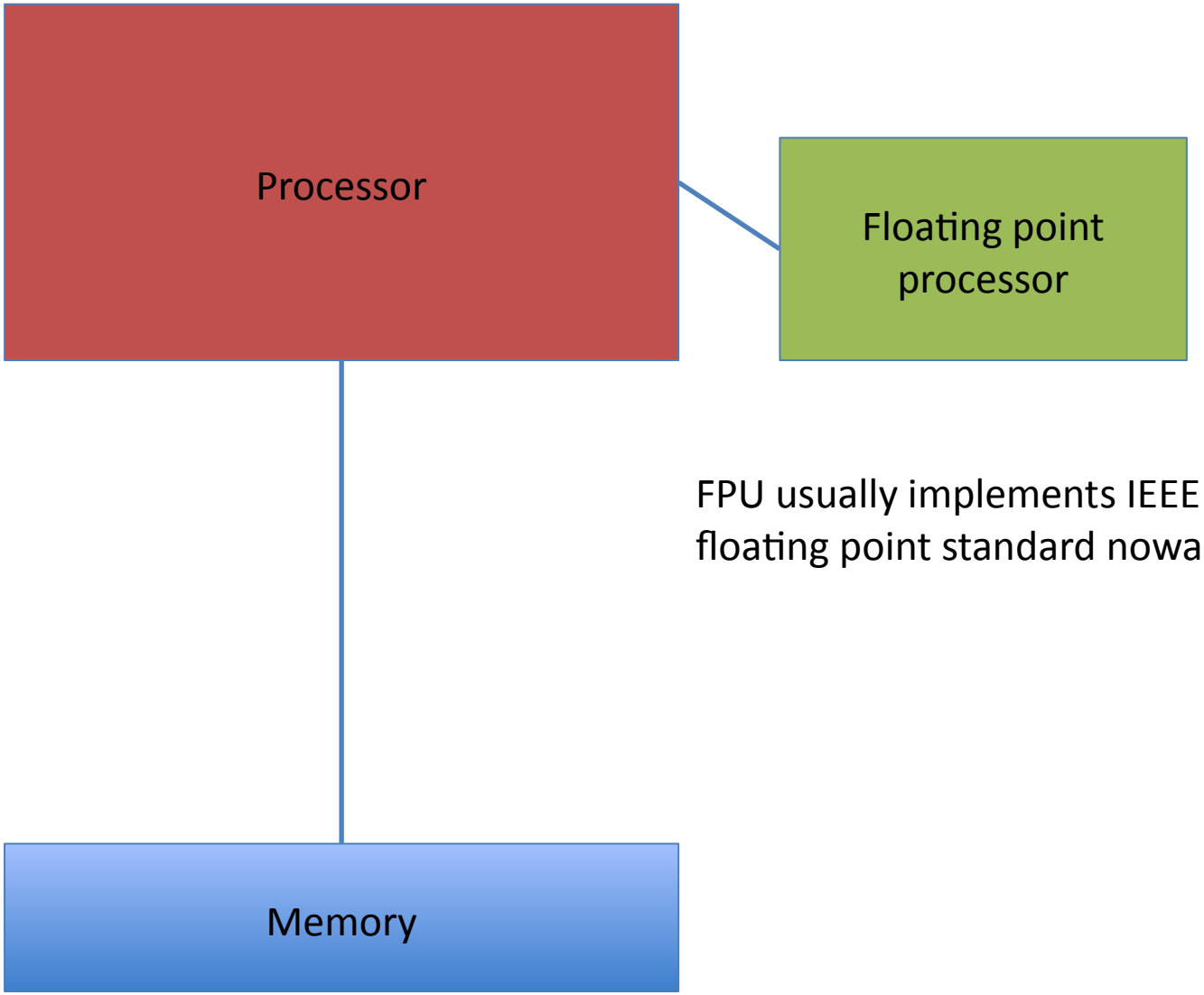
- Single processor systems
- Multi-processor shared memory systems
- Memory hierarchy and cache
- Floating point units
- Distributed memory systems

Single processor system



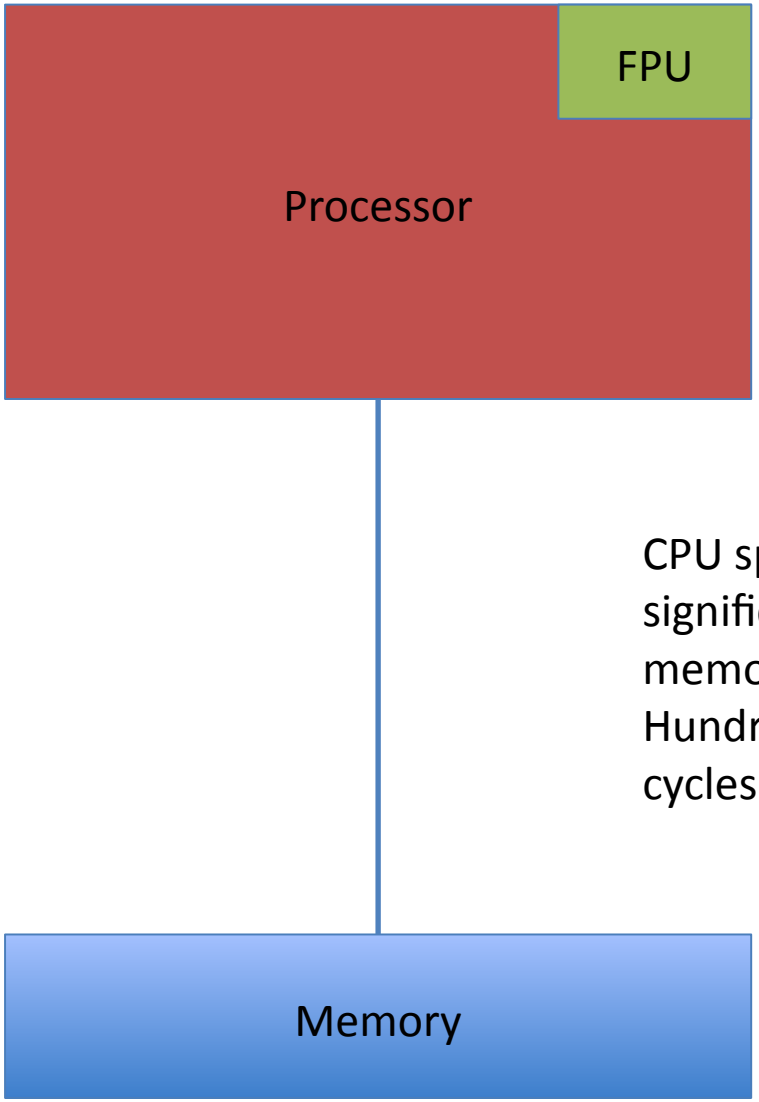
Multiply/divide/floating point in software even? If you're lucky, maybe the vendor might supply libraries, but they still won't be very fast!

Single processor system



FPU usually implements IEEE floating point standard nowadays

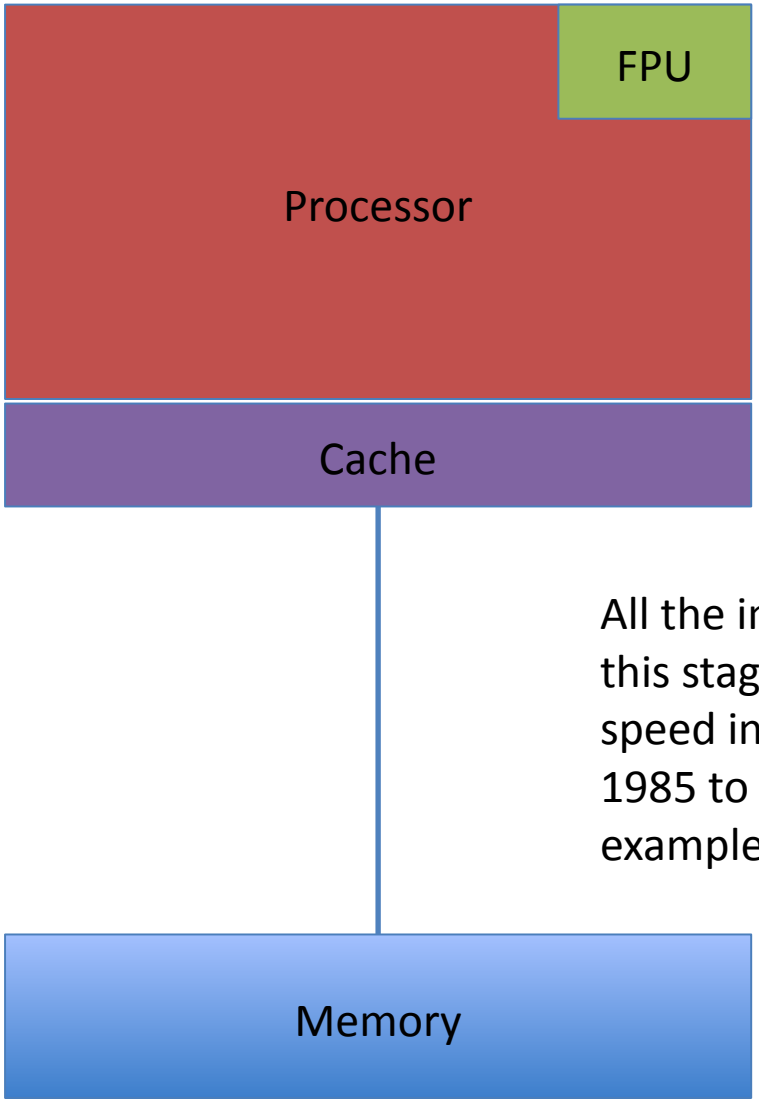
Single processor system



Essentially techniques improve and there's room on the base fabric to incorporate the FPU, eg Intel 80486 in 1989

CPU speeds increase at significantly greater rate than memory speed.
Hundreds of processor clock cycles to read memory today.

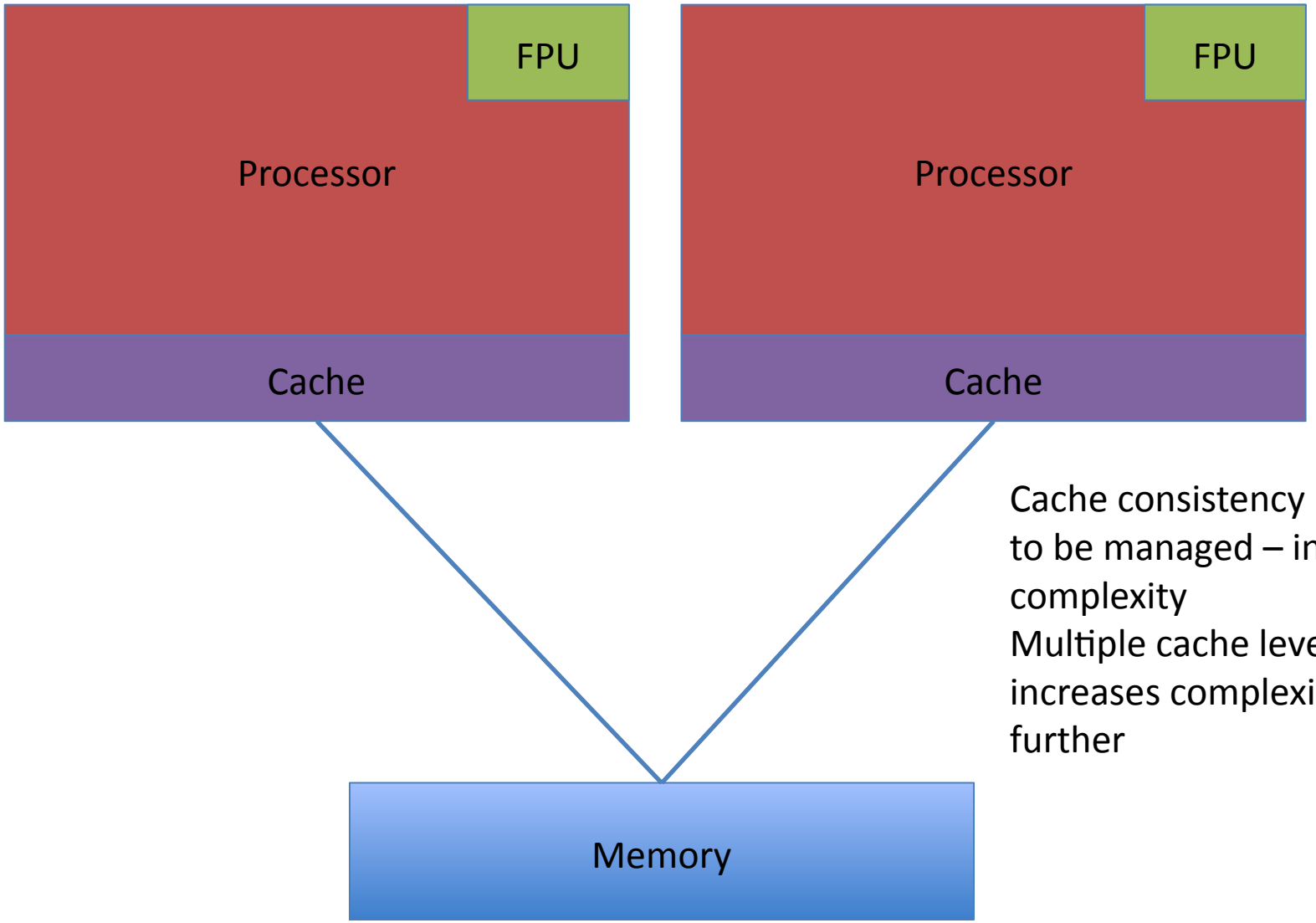
Single processor system



Cache is a work-around to the problem of disparity in speed between processors and memory

All the improvements in performance at this stage come from processor clock speed increases: Intel 80386 at 16MHz in 1985 to Pentium 4 at 2GHz in 2000, for example

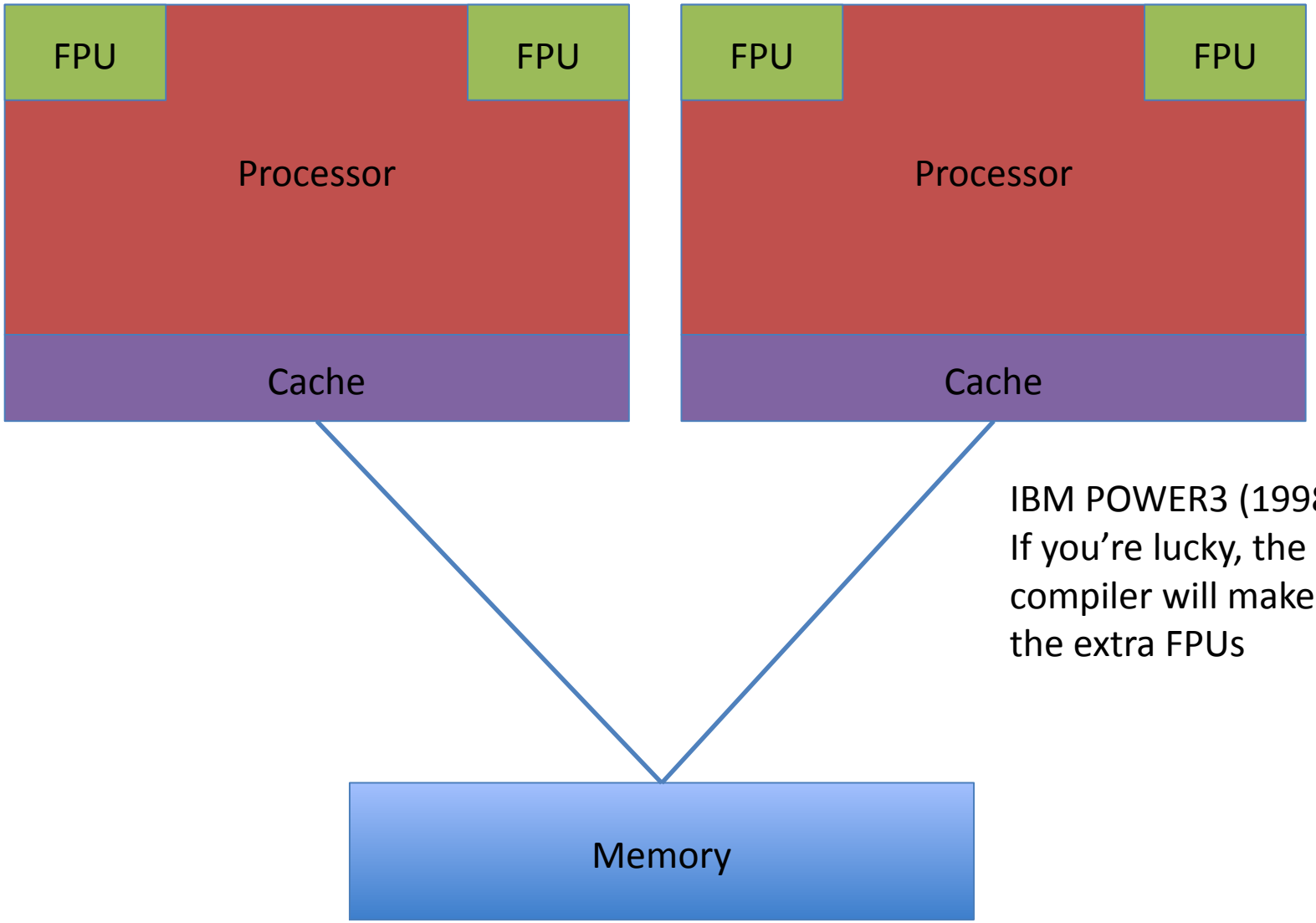
Multiple processor systems



Cache consistency needs to be managed – increases complexity

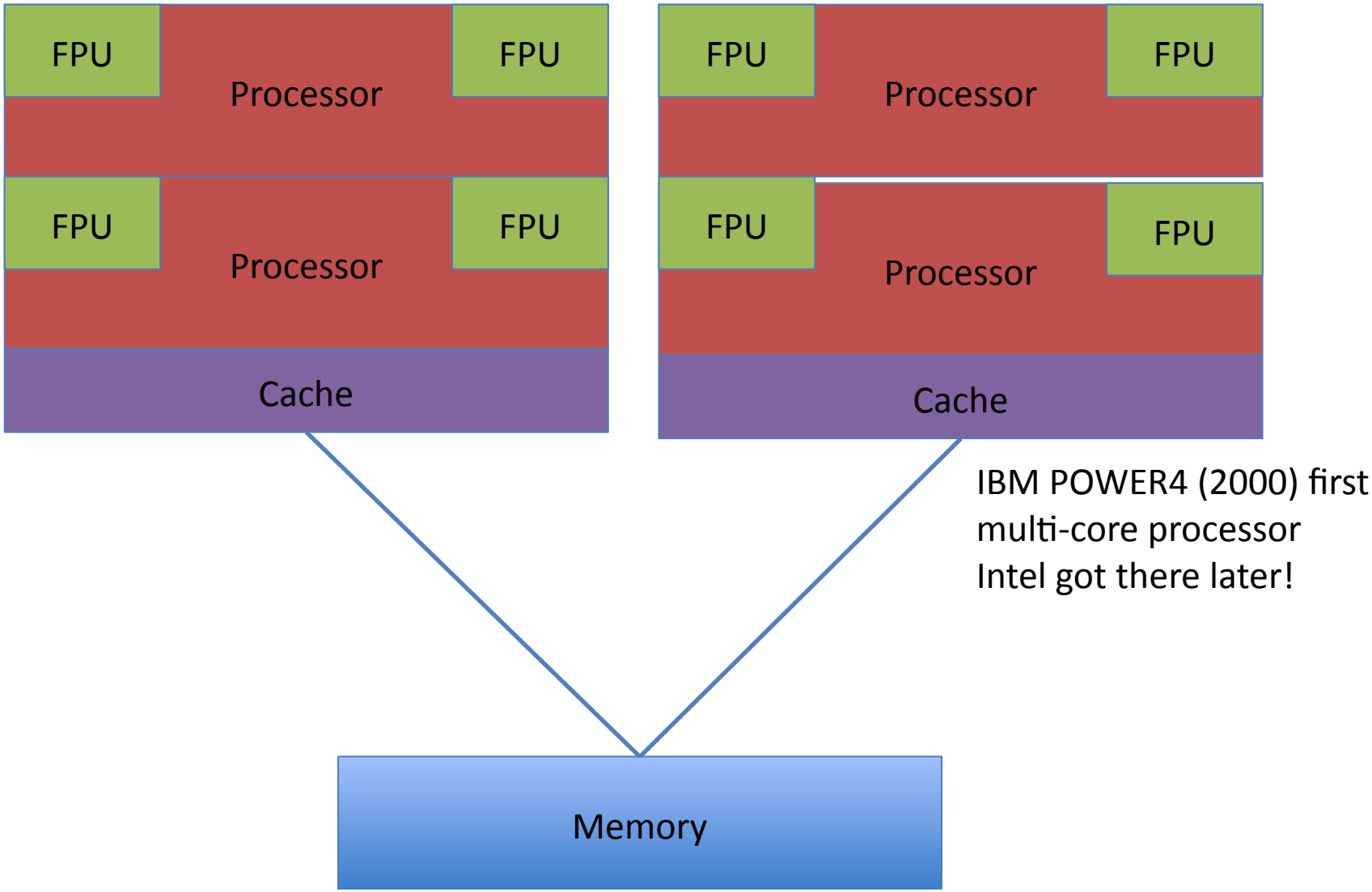
Multiple cache levels – increases complexity further

Add more floating point units



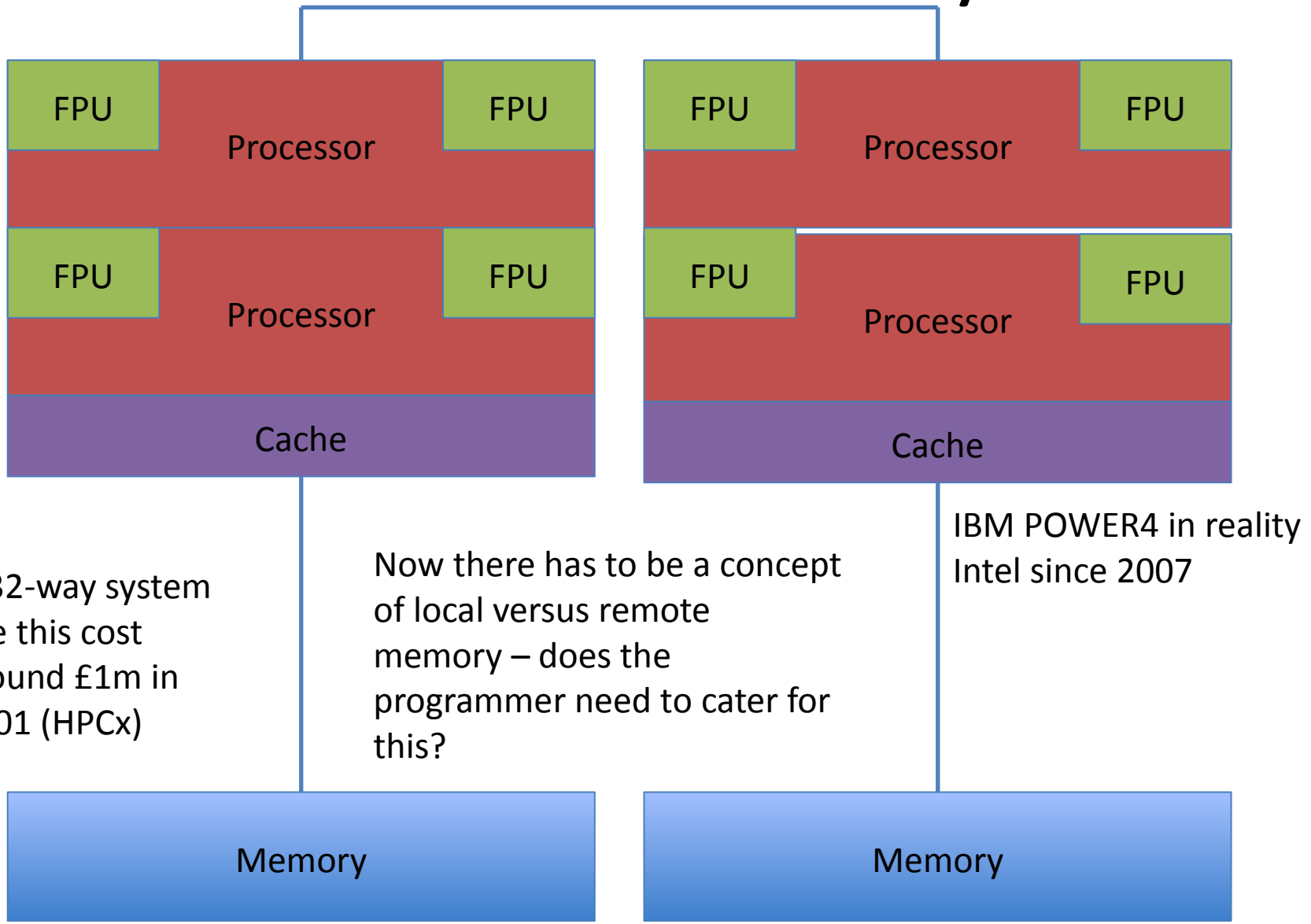
IBM POWER3 (1998)
If you're lucky, the compiler will make use of the extra FPUs

Add more processor cores



IBM POWER4 (2000) first multi-core processor
Intel got there later!

Move the memory

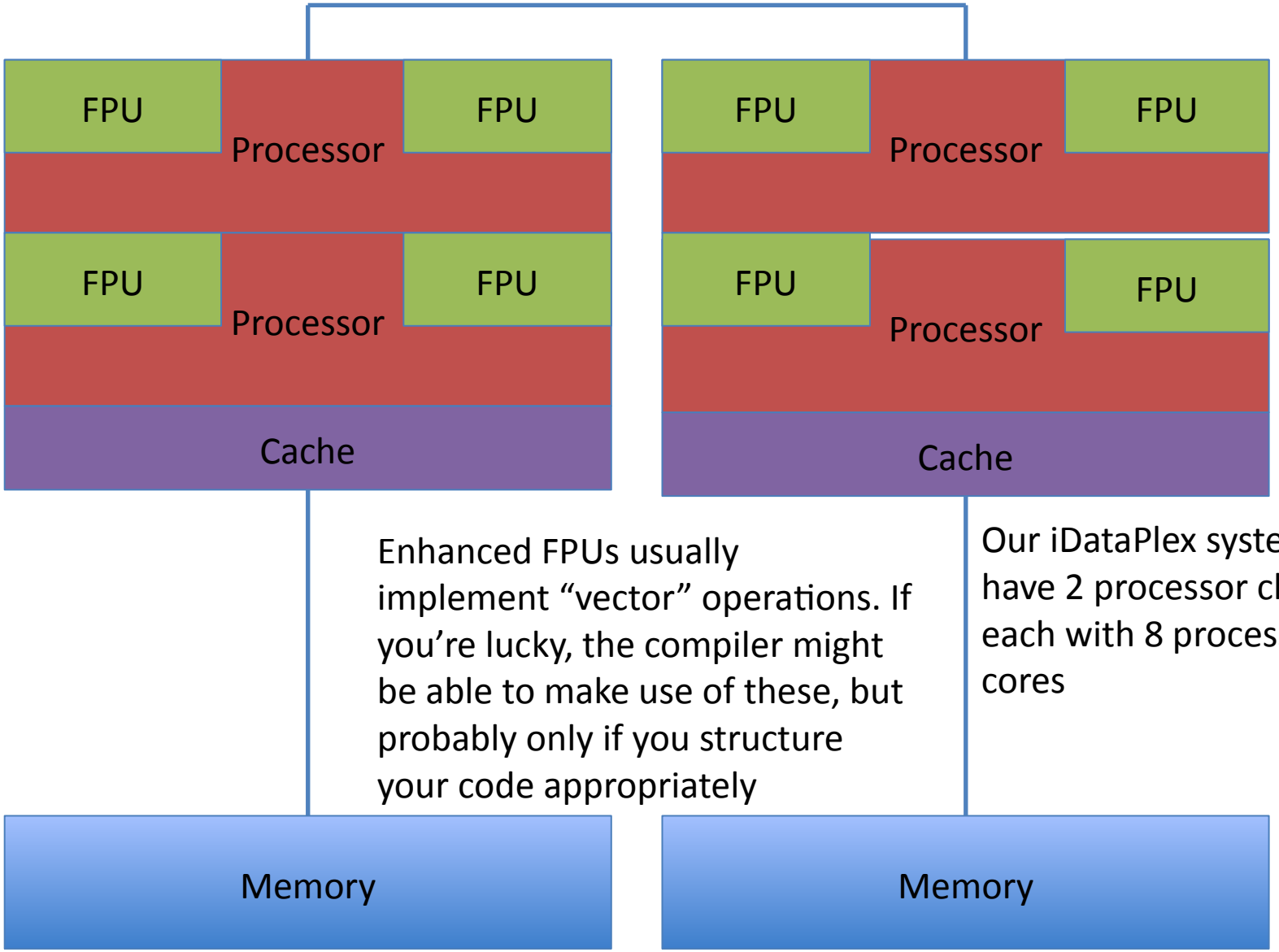


A 32-way system like this cost around £1m in 2001 (HPCx)

Now there has to be a concept of local versus remote memory – does the programmer need to cater for this?

IBM POWER4 in reality Intel since 2007

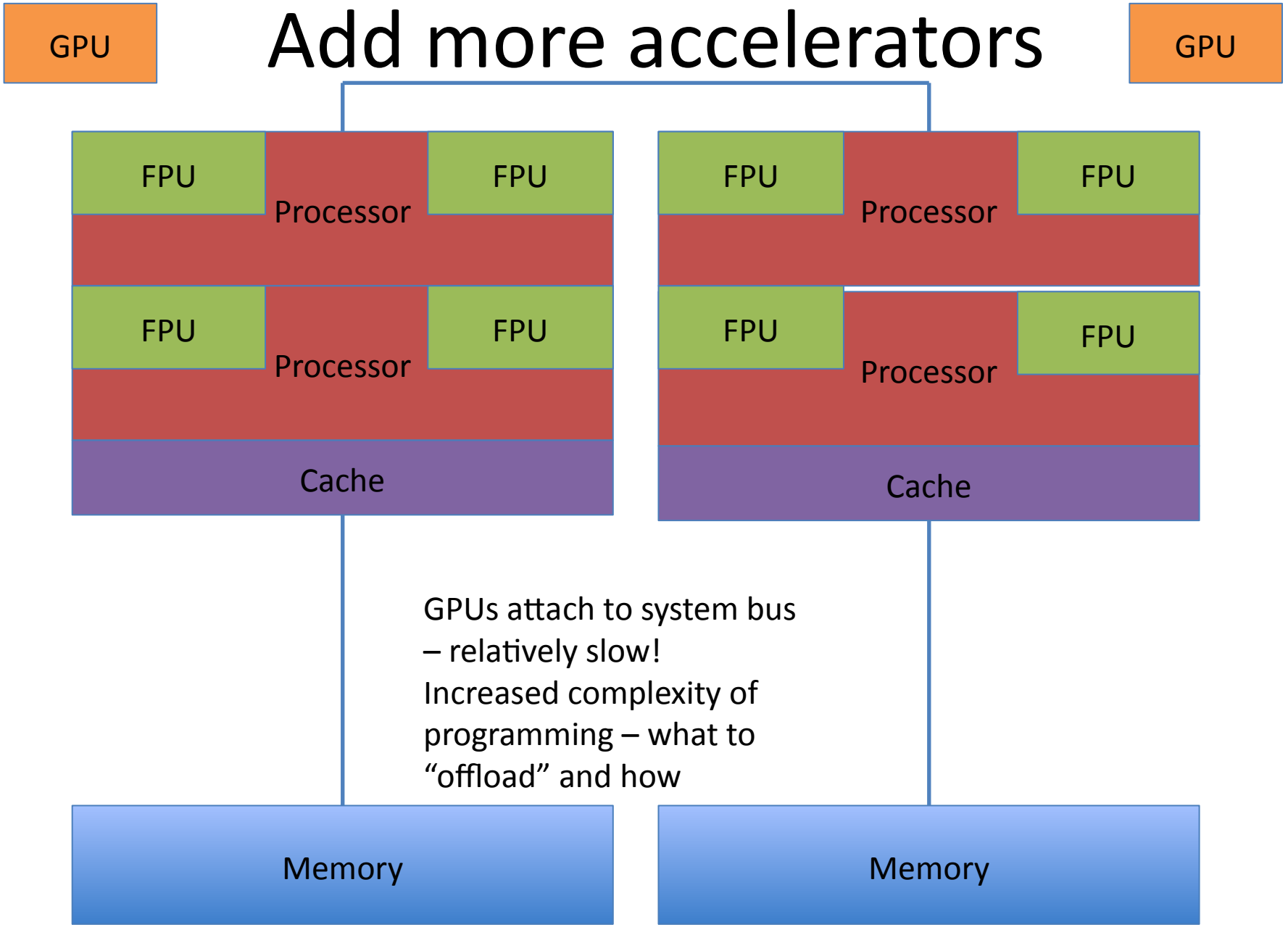
Enhance the FPUs



Enhanced FPUs usually implement “vector” operations. If you’re lucky, the compiler might be able to make use of these, but probably only if you structure your code appropriately

Our iDataPlex systems have 2 processor chips each with 8 processor cores

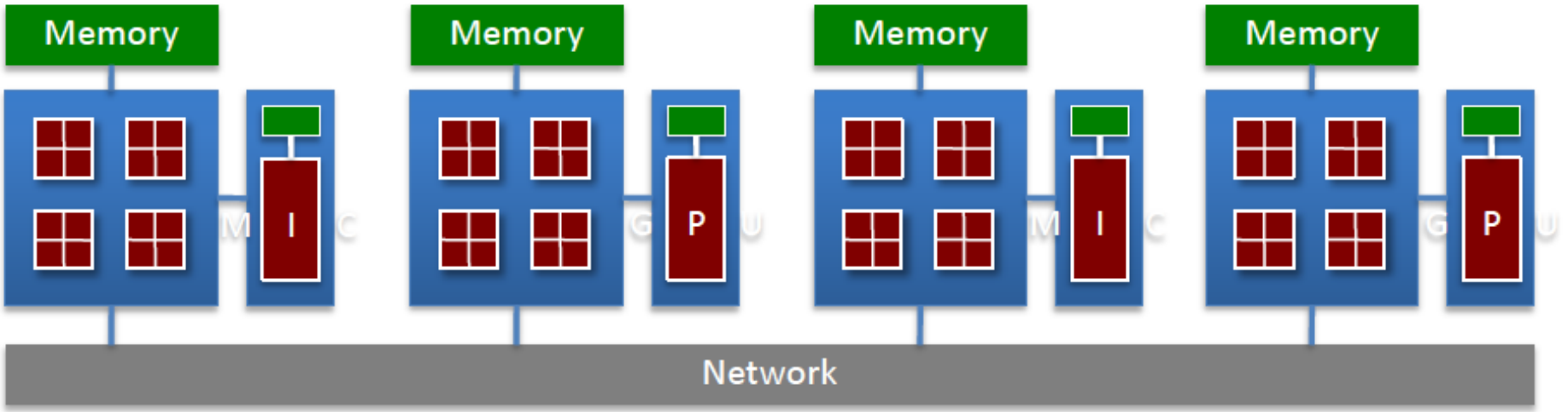
Add more accelerators



Conclusions

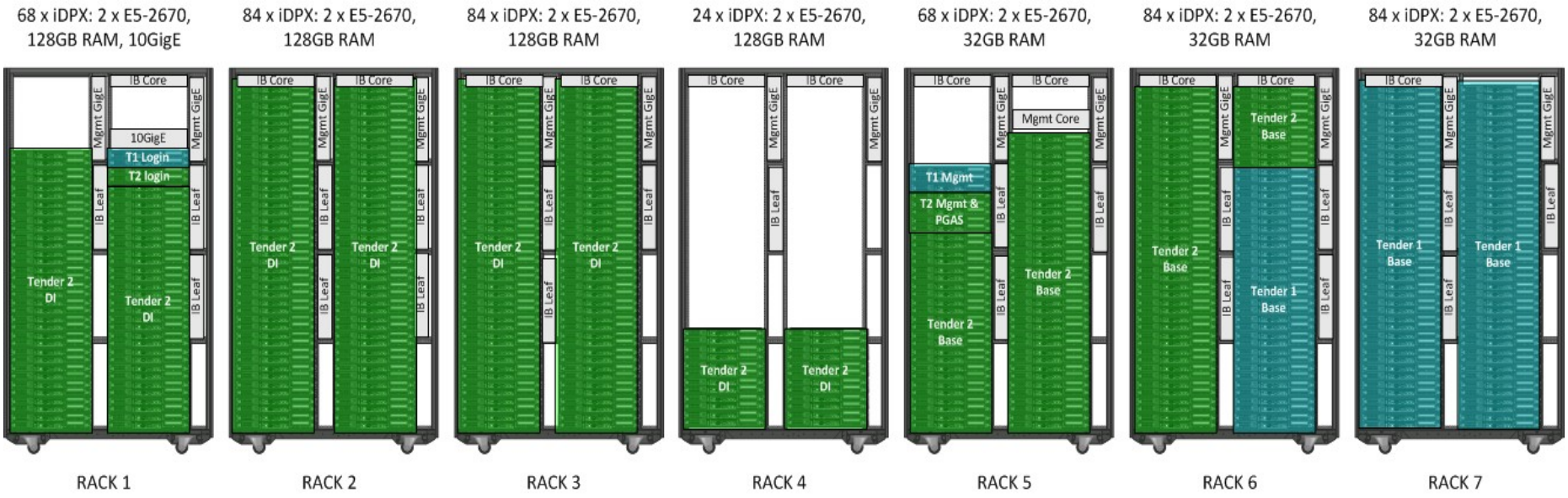
- Single systems with large numbers of CPUs are easier to program but get very expensive for large numbers of processors
- Every processor chip today is a multi-core processor, and processor numbers per chip will increase in future
- Clock speeds are no longer increasing and won't
- To increase aggregate performance, run multiple systems – parallel hardware and software
- Real applications may demonstrate 10% of theoretical peak of floating point performance available in the hardware

Today's systems such as ours

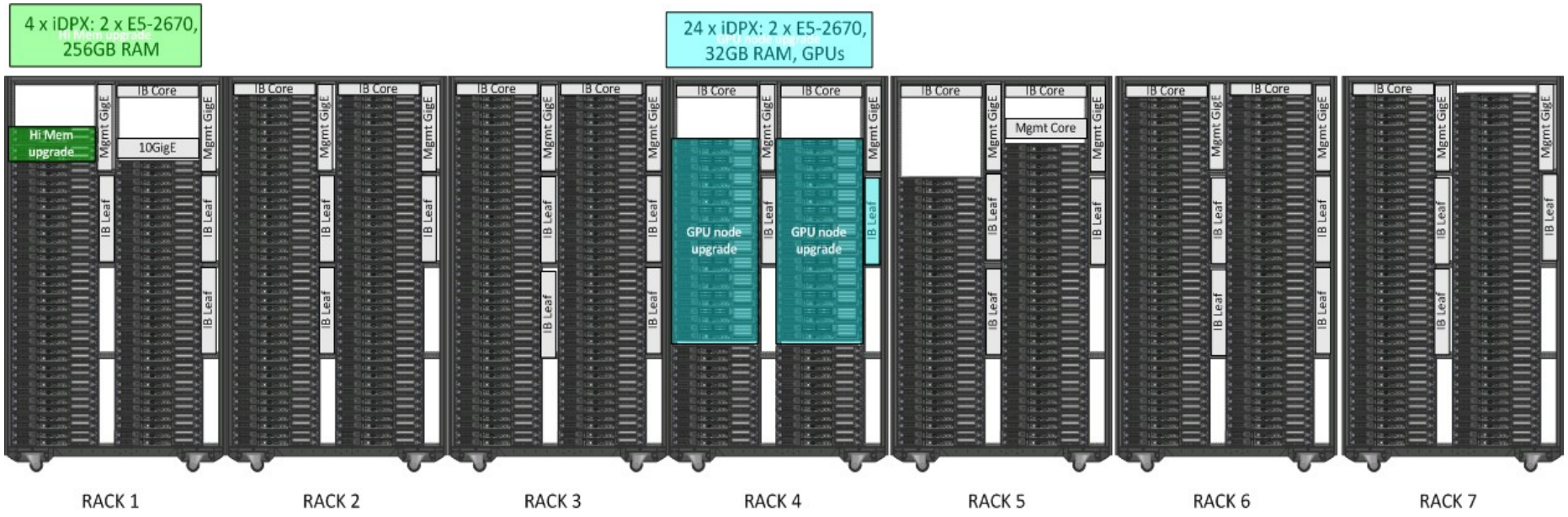


Communicate between systems by message passing – MPI
Communicate inside systems by shared memory – OpenMP
Communicate with accelerators with new models – CUDA, OpenCL
Network needs to be low latency and high bandwidth
The real bottlenecks are the speed/bandwidth of memory and the latency/bandwidth of the message passing, and in many cases there will be more floating point hardware available than can be used – CPUs are "free"

Our iDataPlex system



Additional large memory and GPU nodes installed later in 2012



Science & Technology Facilities Council

Hartree Centre
IBM iDataPlex Blue Wonder

TOP500
#158 in the Nov 2012 list

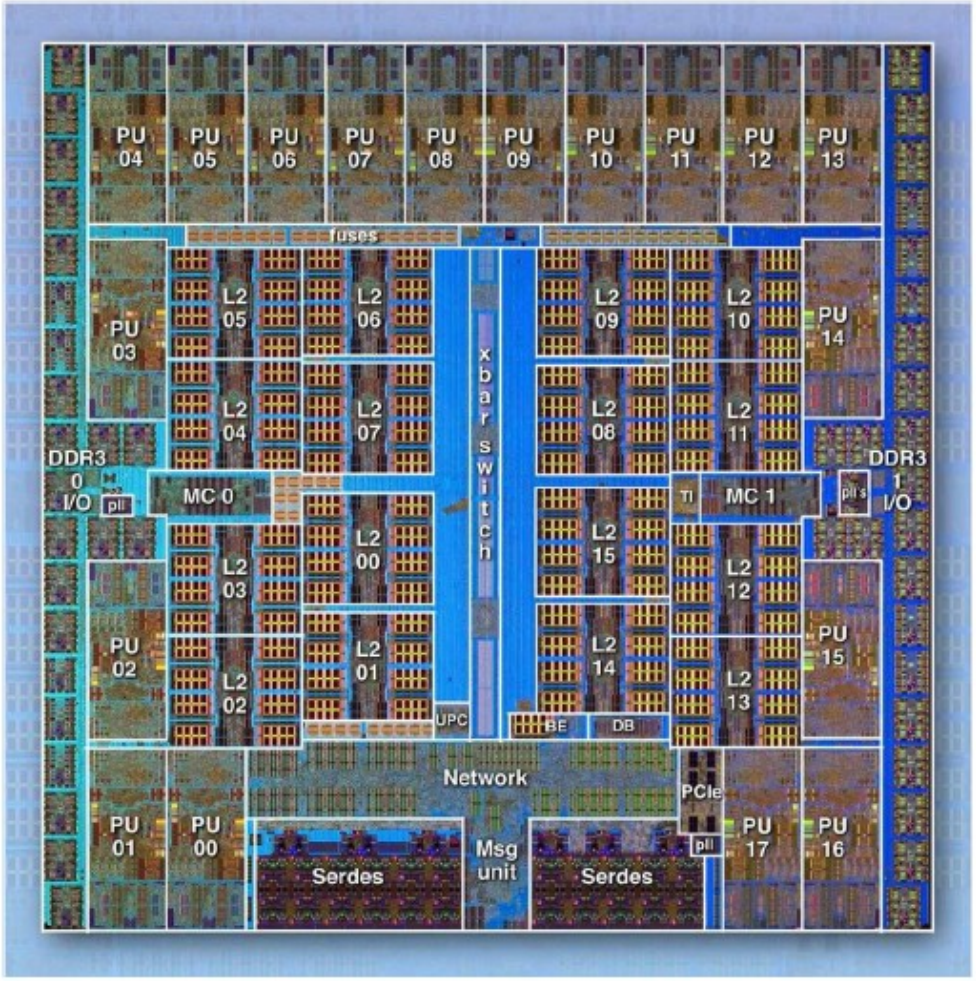
8192 cores, 196 Tflop/s peak
node has 16 cores, 2 sockets
Intel Sandy Bridge (AVX etc.)

252 nodes with 32 GB
4 nodes with 256 GB
12 nodes with X3090 GPUs

256 nodes with 128 GB
ScaleMP virtualization software up to 4TB virtual shared memory

BlueGene/Q Compute chip

System-on-a-Chip design : integrates processors, memory and networking logic into a single chip

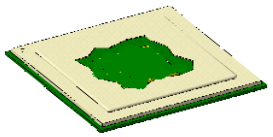


- 360 mm² Cu-45 technology (SOI)
- 16 user + 1 service PPC processors
  - plus 1 redundant processor
  - all processors are symmetric
  - 11 metal layer
  - each 4-way multi-threaded
  - 64 bits
  - 1.6 GHz
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
  - each processor has Quad FPU (4-wide double precision, SIMD)
  - peak performance 204.8 GFLOPS @ 55 W
- Central shared L2 cache: 32 MB
  - eDRAM
  - multiversioned cache – supports transactional memory, speculative execution.
  - supports scalable atomic operations
- Dual memory controller
  - 16 GB external DDR3 memory
  - 42.6 GB/s DDR3 bandwidth (1.333 GHz DDR3) (2 channels each with chip kill protection)
- Chip-to-chip networking
  - 5D Torus topology + external link
    → 5 x 2 + 1 high speed serial links
  - each 2 GB/s send + 2 GB/s receive
  - DMA, remote put/get, collective operations
- External (file) IO -- when used as IO chip.
  - PCIe Gen2 x8 interface (4 GB/s Tx + 4 GB/s Rx)
  - re-uses 2 serial links
  - interface to Ethernet or Infiniband cards

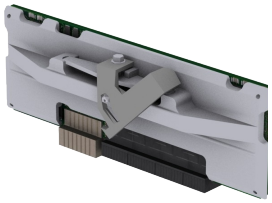
Blue Gene/Q

1. Chip
16 cores

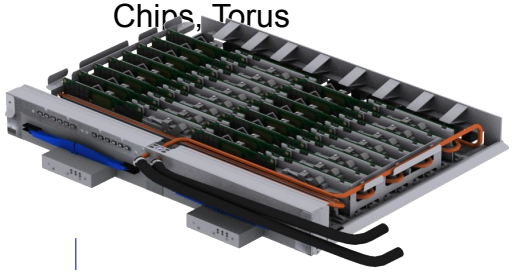
2. Module
Single Chip



3. Compute Card
One single chip module,
16 GB DDR3 Memory



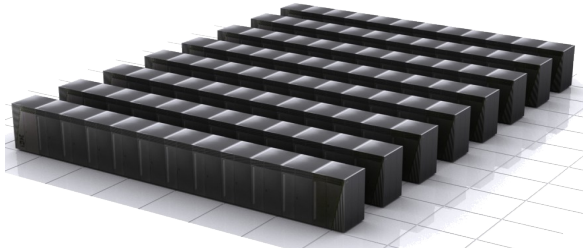
4. Node Card
32 Compute Cards,
Optical Modules, Link
Chips, Torus



5. Midplane
16 Node Cards

6. Rack
2 Midplanes
1, 2 or 4 I/O Drawers

7. System
96 racks @ 20PF/s



Per Rack	
Peak Performance	209 TF
Sustained *(Linpack)*	~170+ TF
Power	~100 kW
Power Efficiency	~1.7 GF/W

Science & Technology Facilities Council

Hartree Centre
IBM BG/Q Blue Joule

TOP500
#16 in the Nov 2012 list
#5 in Europe
#1 system in UK

6 racks

- 98,304 cores
- 6144 nodes
- 16 cores & 16 GB per node
- 1.26 Pflop/s peak

1 rack of BGAS (Blue Gene Advanced Storage)

- 16,384 cores
- Up to 1PB Flash memory

Science & Technology Facilities Council

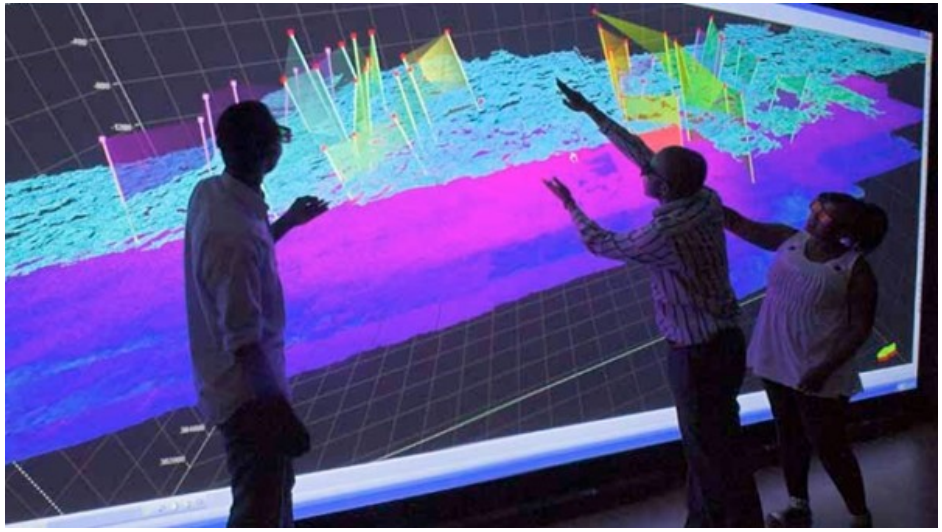
Hartree Centre Datastore

Storage:

5.76 PB usable disk storage
15 PB tape store

Science & Technology Facilities Council

Hartree Centre Visualization



Four major facilities:

Hartree Vis-1: a large visualization "wall" supporting stereo
Hartree Vis-2: a large surround and immersive visualization system
Hartree ISIC: a large visualization "wall" supporting stereo at ISIC
Hartree Atlas: a large visualization "wall" supporting stereo in the Atlas Building at RAL, part of the Harwell Imaging Partnership (HIP)

Virtalis is the hardware supplier

Science & Technology Facilities Council

Hartree Centre architecture