HPC Architectures

Types of HPC hardware platforms currently in use

Partners

Funding

Horizon 2020
European Union Funding
for Research & Innovation

Reusing this material

Outline

- Shared memory architectures
    - Symmetric Multi-Processing (SMP) architectures
    - Non-Uniform Memory Access (NUMA) architectures
- Distributed memory architectures
- Hybrid distributed / shared memory architectures
- Accelerators

Architectures

- Architecture is about how different hardware components are connected together to make up usable machines
- Many factors influence choice of architecture:
  - Performance, cost, scalability, use cases, …
- Focus here on the most important distinctions regarding how processors and memory are situated and connected in HPC
- Discuss the role this plays in how parallel computing can be done on different architectures and how it can be expected to perform

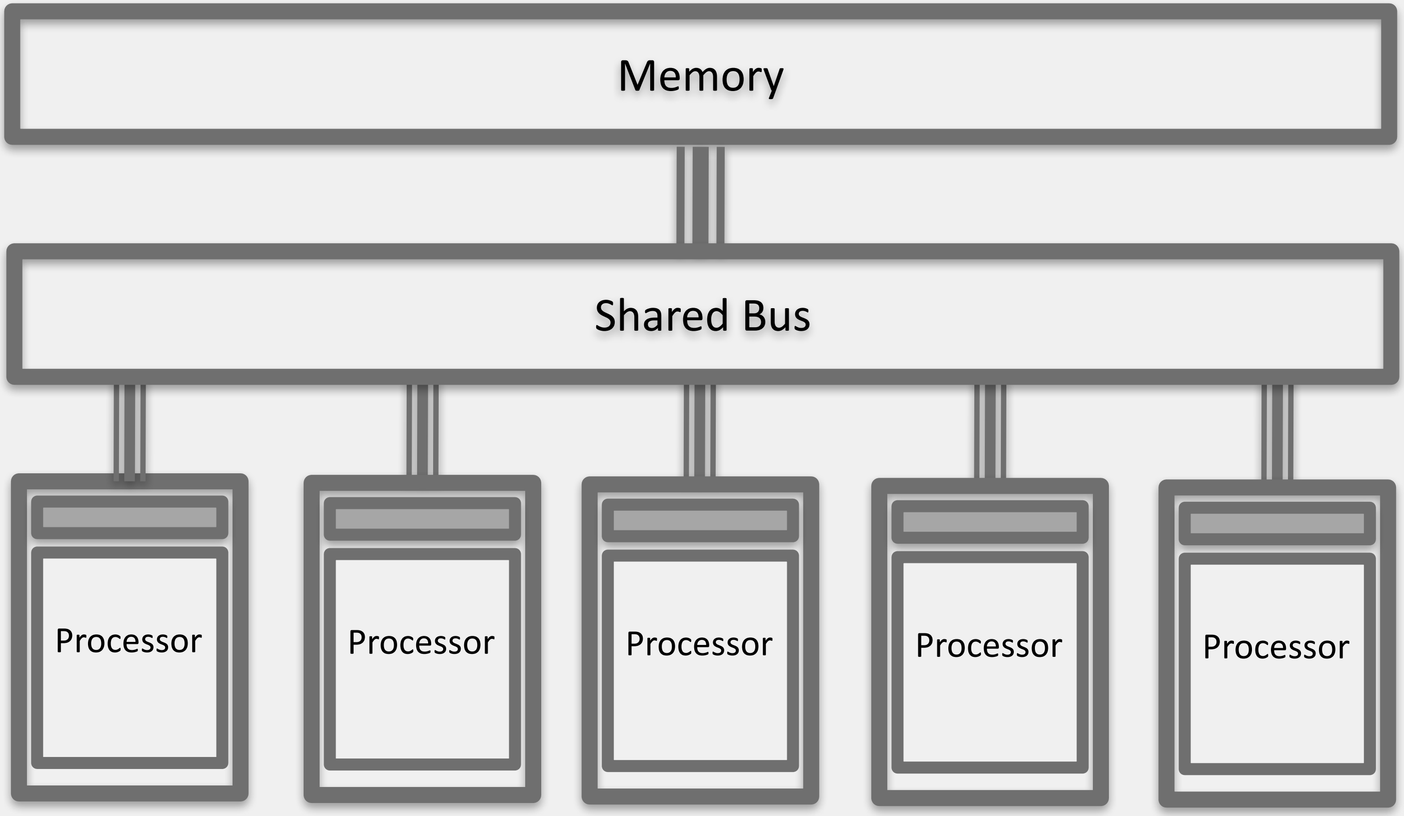
Shared memory architectures

Simplest to use, hardest to build

Shared-Memory Architectures

- Multi-processor shared-memory systems have been common since the early 90’s
  - originally built from many single-core processors
- A single OS controls the entire shared-memory system
- Modern multicore processors are really just shared-memory systems on a single chip
  - Nowadays can’t buy a single-core processor even if you wanted one!

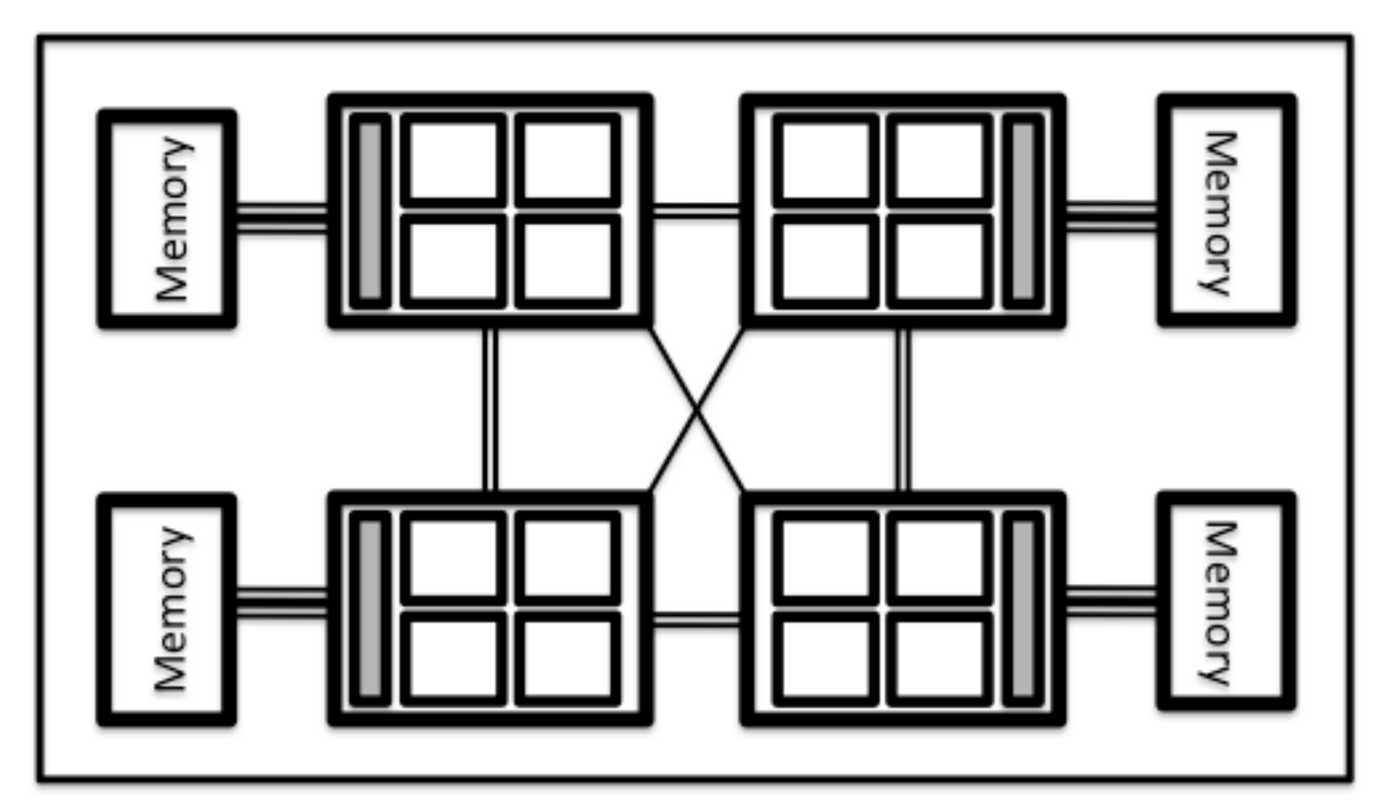
Symmetric Multi-Processing* Architectures



*SMP

All cores have access at the same speed to the same memory, e.g. a multicore laptop

Non-Uniform Memory Access* Architectures



*NUMA

Cores have access to memory used by other cores, but more slowly than access to their own local memory

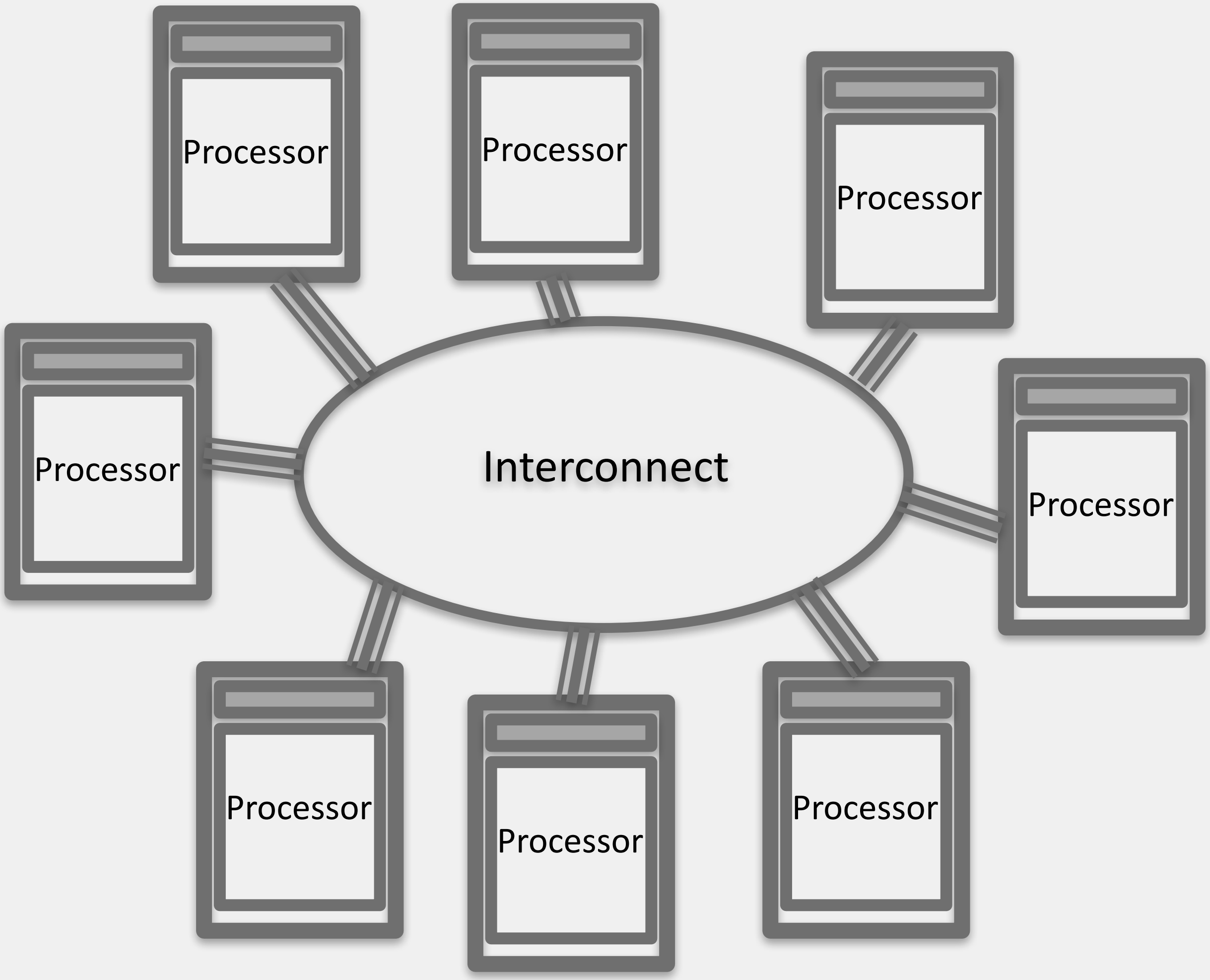
Shared-memory architectures

- Most computers are now shared memory machines due to multicore processors
- Some true SMP architectures…
  - *e.g.* BlueGene/Q nodes
- …but most are NUMA
  - Program NUMA as if they are SMP – details hidden from the user
  - all cores controlled by a single OS
- Difficult to build shared-memory systems with large core numbers (> 1024 cores)
  - Expensive and power hungry
  - Difficult to scale the OS to this level

Distributed memory architectures

Clusters and interconnects

Multiple Connected Computers



- Each self-contained part is called a *node*.
  - each node runs its own copy of the OS

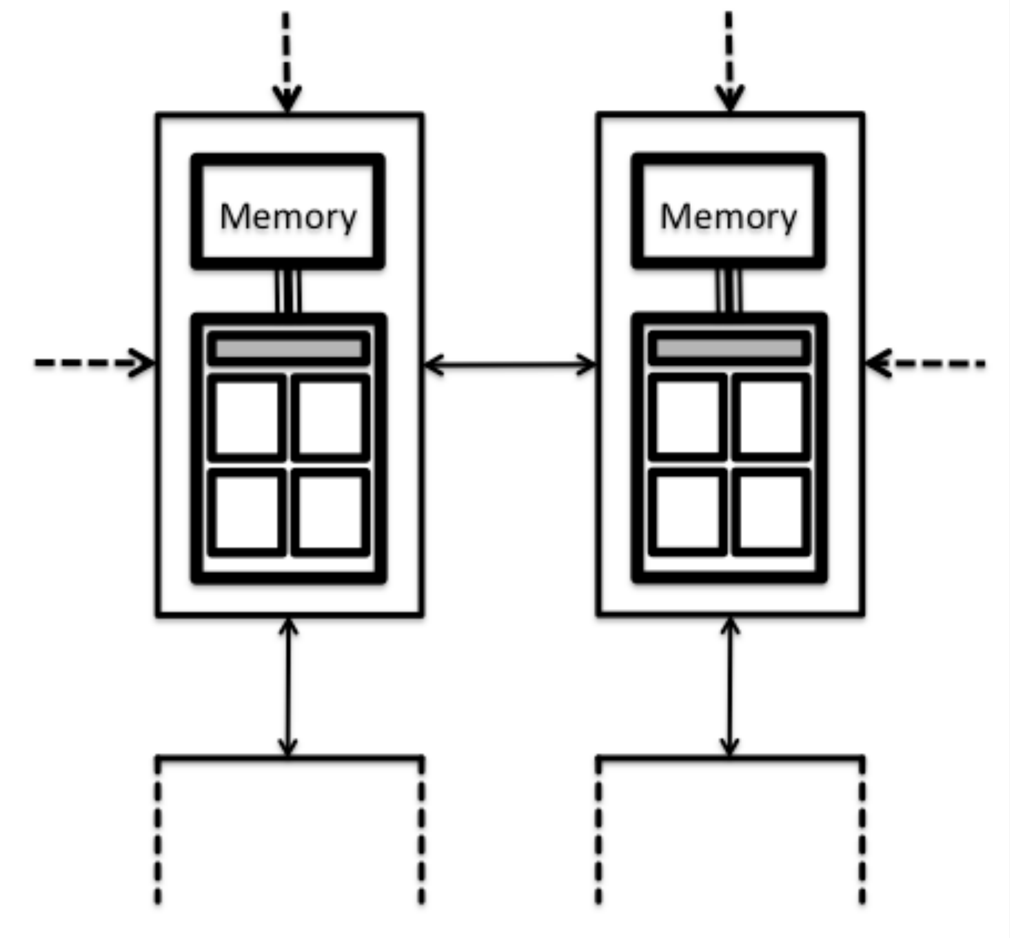
Distributed-memory architectures

- Almost all HPC machines are distributed memory
- The performance of parallel programs often depends on the *interconnect* performance
  - Although once it is of a certain (high) quality, applications usually reveal themselves to be CPU, memory or IO bound
  - Low quality interconnects (e.g. 10Mb/s – 1Gb/s Ethernet) do not usually provide the performance required
  - Specialist interconnects are required to produce the largest supercomputers. *e.g.* Cray Aries, IBM BlueGene/Q
  - Infiniband is dominant on smaller systems.
- High bandwidth relatively easy to achieve
  - low latency is usually more important and harder to achieve

Distributed/shared memory hybrids

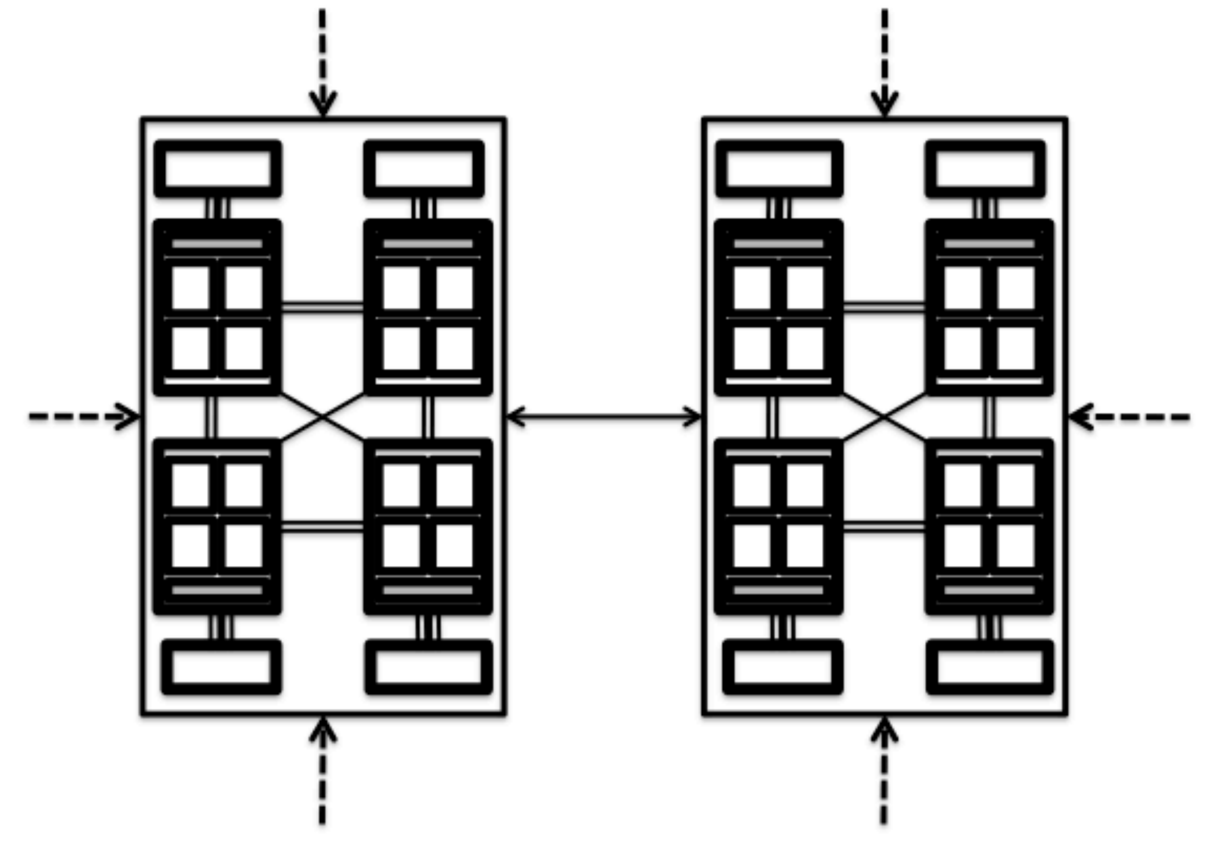
Almost everything now falls into this class

Multicore nodes



- In a real system:
  - each node will be a shared-memory system
    - e.g. a multicore processor
  - the network will have some specific topology
    - e.g. a regular grid

Hybrid architectures



- Now normal to have NUMA nodes
  - e.g. multi-socket systems with multicore processors
- Each node still runs a single copy of the OS

Hybrid architectures

- Almost all HPC machines fall in this class
- Most applications use a message-passing (MPI) model for programming
  - Usually use a single process per core
- Increased use of hybrid message-passing + shared memory (MPI+OpenMP) programming
  - Usually use 1 or more processes per NUMA region and then the appropriate number of shared-memory threads to occupy all the cores
- Placement of processes and threads can become complicated on these machines

Example: ARCHER

- ARCHER has two 12-way multicore processors per node
  - 2 x 2.7 GHz Intel E5-2697 v2 (Ivy Bridge) processors
  - each node is a 24-core, shared-memory, NUMA machine
  - each node controlled by a single copy of Linux
  - 4920 nodes connected by the high-speed ARIES Cray network

Accelerators

How are they incorporated?

Including accelerators

- Accelerators are usually incorporated into HPC machines using the hybrid architecture model
    - A number of accelerators per node
    - Nodes connected using interconnects
- Communication from accelerator to accelerator depends on the hardware:
    - NVIDIA GPU support direct communication
    - AMD GPU have to communicate via CPU memory
    - Intel Xeon Phi communication via CPU memory
    - Communicating via CPU memory involves lots of extra copy operations and is usually very slow

Summary

- Vast majority of HPC machines are shared-memory nodes linked by an interconnect.
  - Hybrid HPC architectures – combination of shared and distributed memory
  - Most are programmed using a pure MPI model (more later on MPI) - does not really reflect the hardware layout
- Accelerators are incorporated at the node level
  - Very few applications can use multiple accelerators in a distributed memory model