Why Multiprocessors?

Moore's Law predicted a doubling of processor performance every couple of years

- true until about 2000

Limits on the performance of a single processor: what are they?

Why Multiprocessors

1. Utilizes coarser granularities than ILP
2. Lots of workload opportunity

- Scientific computing/supercomputing
  - Examples: weather simulation, aerodynamics, protein folding
  - Each processor computes for a part of the grid
- Server workloads
  - Example: airline reservation database
  - Many concurrent updates, searches, lookups, queries
  - Processors handle different requests
- Media workloads
  - Processors compress/decompress different parts of image/frames
- Desktop workloads …
- Gaming workloads …

3. Can now fit multiple processors on a chip; but each one is probably simpler

What would you do with a billion transistors on a chip? Or more?

What is a Parallel Architecture?

A parallel computer is a collection of processing elements that cooperate to solve large problems fast.

Some broad issues:

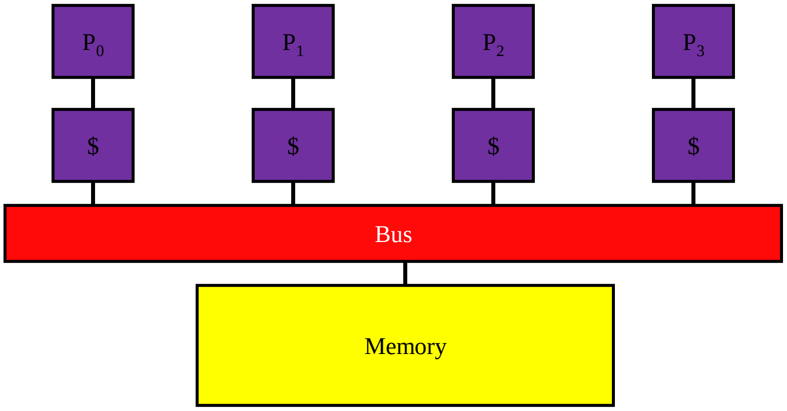
- Resource Allocation:
  - How many processing elements (PEs)?
  - How powerful are the PEs?
  - How much memory?
- Data access, Communication and Synchronization
  - How do the PEs cooperate and communicate?
  - How are data transmitted between PEs?
  - What are the abstractions and primitives for cooperation?
- Performance and Scalability
  - How does a design translate into performance?
  - How does it scale?

Multiprocessors

Low-end

- bus-based
  - simple, but a bottleneck
  - broadcast cache coherency protocol
- physically centralized memory
- uniform memory access (UMA machine)
- today's small CMPs:
  Intel Core i3, i5, i7 (2-6 cores), AMD Opteron "Bulldozer" (4-16 cores), Sun SPARC T4 (8 cores per processor, 4 processors per system), ARM Cortex A5 (2 cores), Nvidia Tegra 3 (4 cores)

Low-end MP

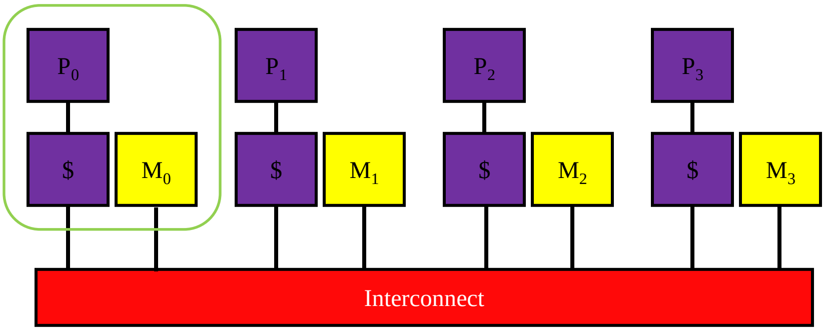


Multiprocessors

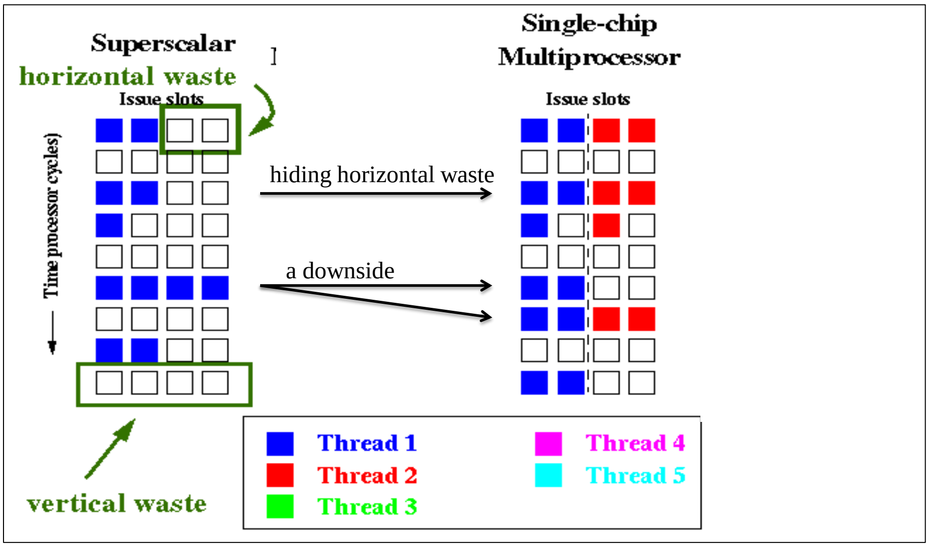
High-end

- multiple-path interconnect
  - higher bandwidth
  - longer memory latencies
  - more scalable
  - point-to-point cache coherency protocol
- physically distributed memory
- non-uniform memory access (NUMA machine)
- could have processor clusters
- today’s large MPs:
  SGI UV (256 10-core Xeon processors, 2D torus), Cray XE6 (1M Opteron 6200 cores), IBM BlueGene/Q (100K 16-core PowerPCs, 5D torus), Fujitsu K Computer (44K 16-core SPARCs)

High-end MP



Comparison of Issue Capabilities


Superscalar
horizontal waste
Issue slots
Time (processor cycles)
hiding horizontal waste
a downside
Single-chip Multiprocessor
Issue slots
vertical waste
Thread 1
Thread 2
Thread 3
Thread 4
Thread 5