HPC Architecture Overview

Tyson Whitehead

SHARCNET
The University of Western Ontario

May 28, 2010

Architectures/Programming models

- Standard computers
- Symmetric multiprocessor systems
- Beowulf style computing
- Accelerator based systems

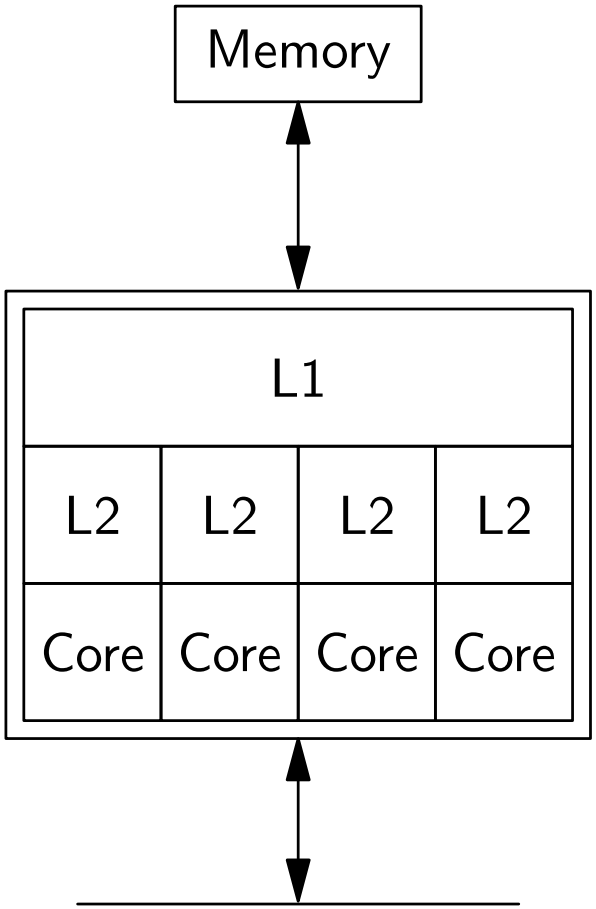
Basic Blocks

Concepts

- IO bound
- Latency/bandwidth
- Compute bound
- Floating point operations/s (FLOPs)

Locality is the key

Standard CPU Socket (i7)



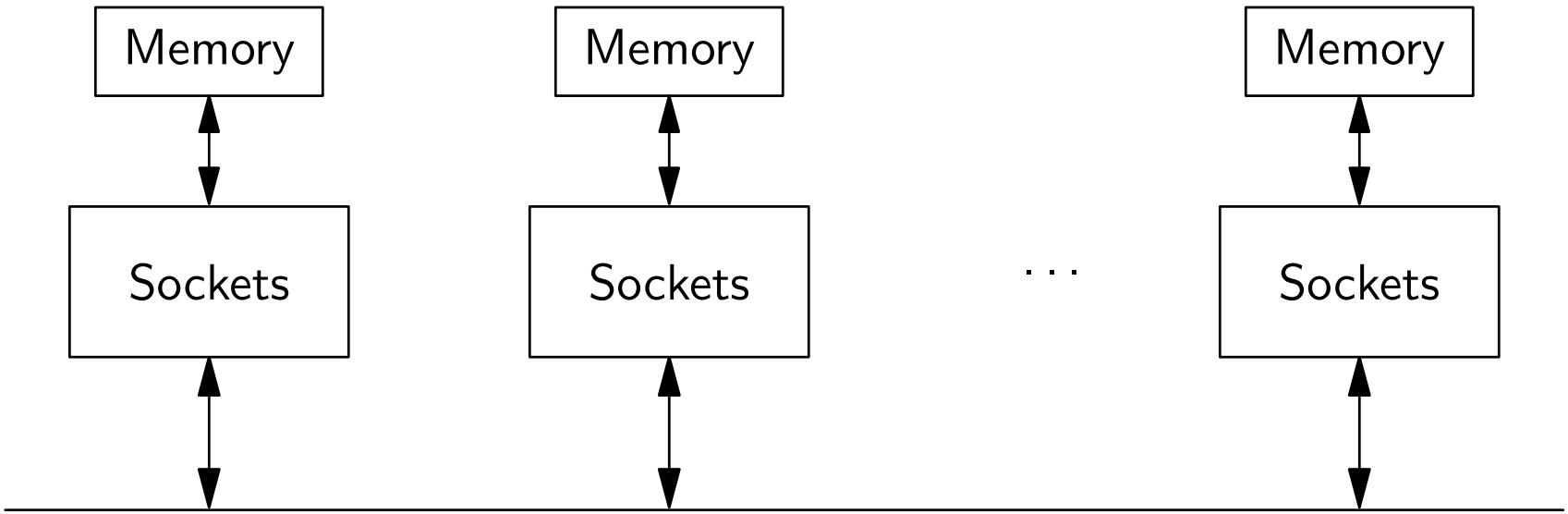
Cache sizes

- L3 8MB shared
- L2 256KB/core
- L1 32KB instruction/core
  32KB data/core

Cores are extremely general and work independently of each other.

- Pipeline hides latencies
- Requires predictability

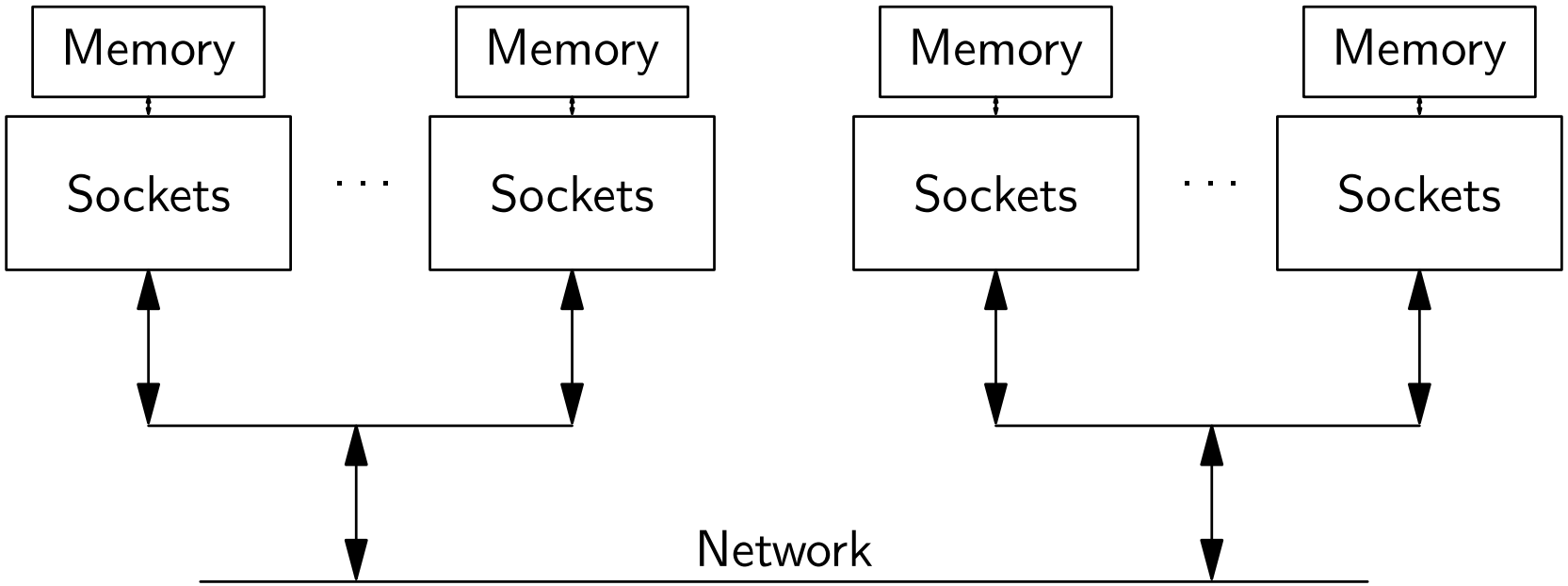
Symmetric Multiprocessor (SMP)



Highly specialized and expensive beyond stock machines.

- Implicitly move data
- Sensitive to location
- Synchronization is expensive

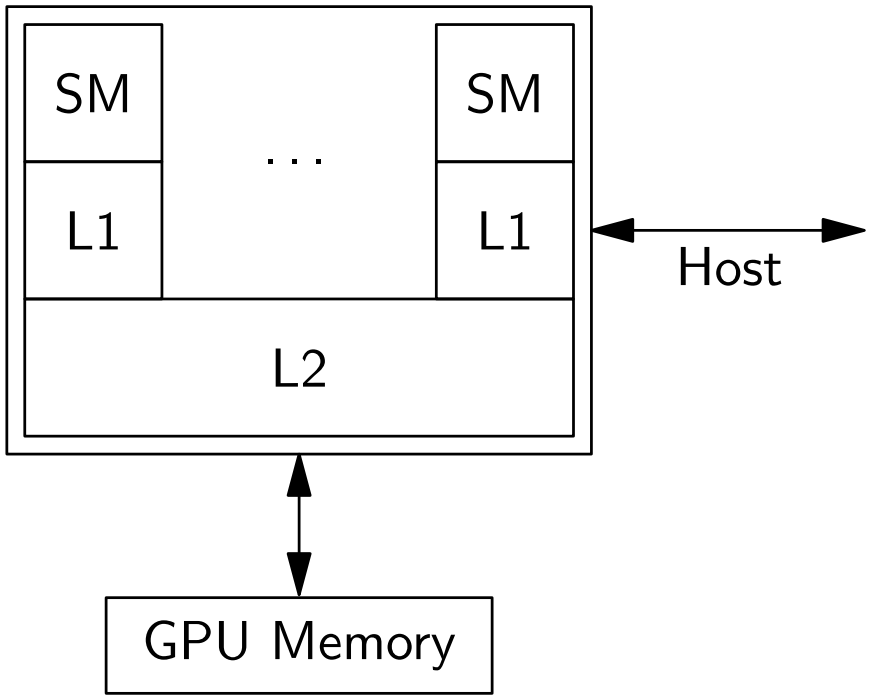
Beowulf style



Commodity hardware (except the network)

- Explicitly move data
- Keep processors busy
- Synchronization is expensive

Accelerator style (GPGPU – Fermi)



Cache sizes

- L2 768KB
- L1 16/48KB

L1 is a split local store/cache

The SMs are highly specialized processors based around very lightweight threads concurrently executing kernels.

- Single SM executes 32 threads simultaneously
- Threads execute in lock-step inside a SM