Introduction to Parallel Computing using MPI, OpenMP, & CUDA

(CME213)

GPU {Graphics processing Unit}

→ Specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.

TPU {Tensor processing Unit}

→ AI accelerator, application-specific IC developed by google specifically for neural network.

MPI {Message passing Interface}

OpenMP {Open Multi-Processing}

CUDA {Compute Unified device Architecture}

→ It is a parallel computing platform & API model created by Nvidia.

→ It allows software developers to use a CUDA-enabled GPU for general purpose processing.

Gordon Moore (1965)

→ The number of transistors on a chip shall double every 18-24 months.

Shared memory Parallel Computing

→ Computer with many cores.

Distributed Parallel Computing

→ Connecting many computers over a fast network.

⇒ An crease in transistor density is limited by:
- → Leakage current increases
- → Power consumption increases
- → Heat generated increased.

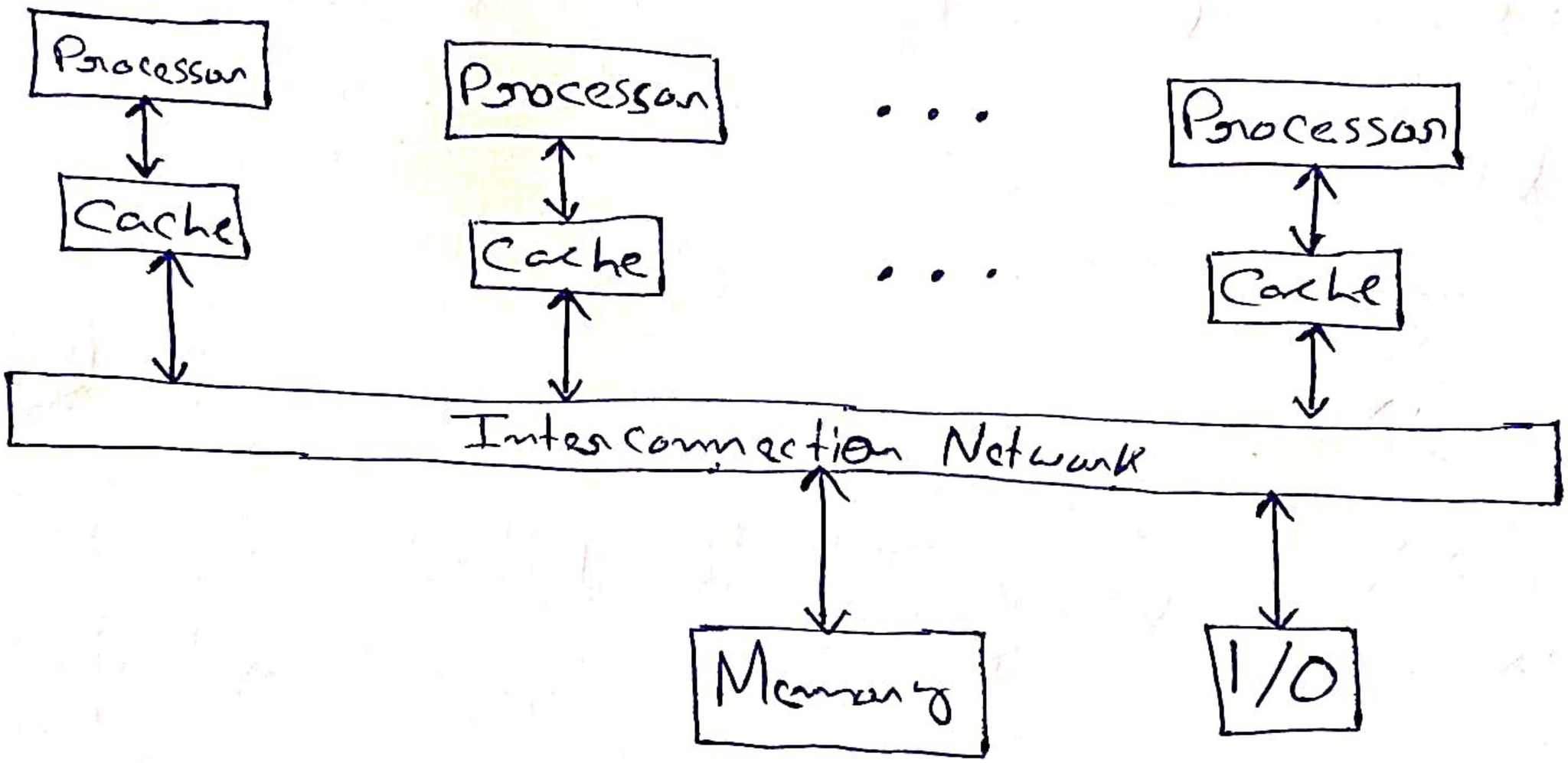
⇒ Memory access time has not been reduced at a rate comparable to the processing speed.

Multicore

- One/few but fast
- Many, but slower cores GPUs

★ Shared Memory Processor

→ A number of processors or cores
→ A shared physical memory (global memory)
→ A interconnection network to connect the processors with the memory.



→ GPU is great for:
- → Dense linear algebra
- → Finite-difference
- → Neural Network.