High Performance Computing Systems (CMSC714)

Lecture 9: Single Node Architectures

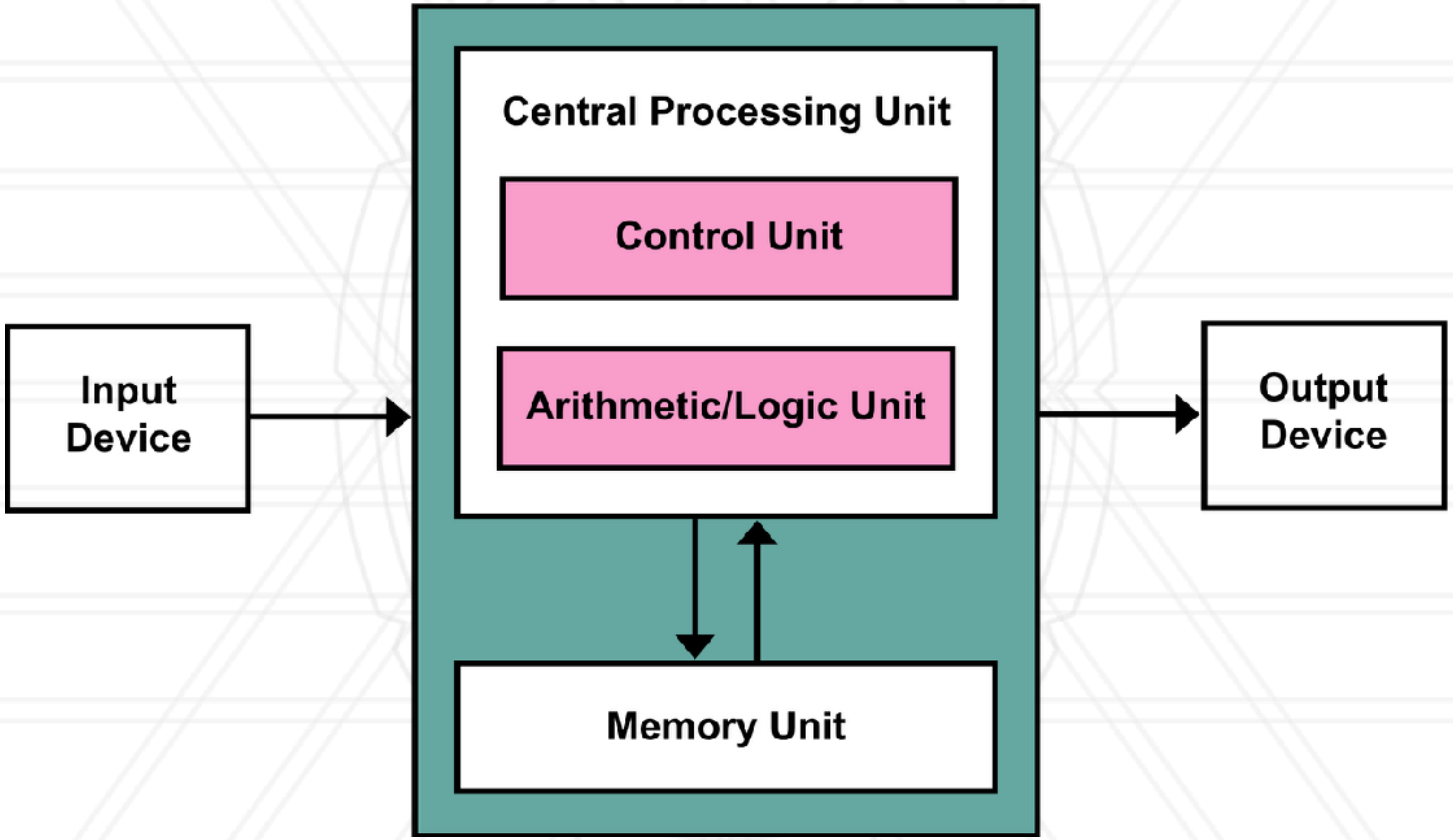
Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Summary of last lecture

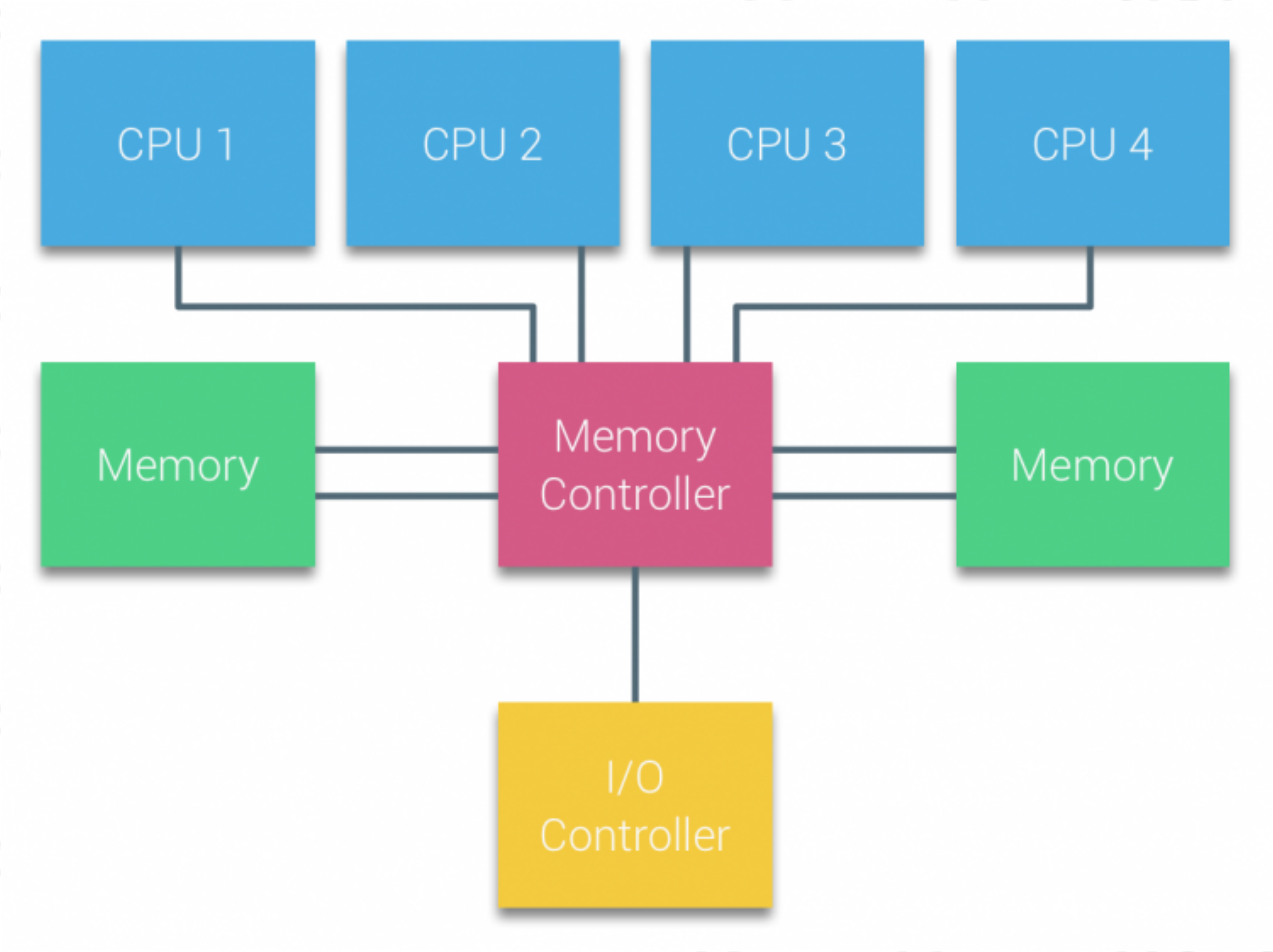
- MPI trace visualization
- Projections performance analysis tool
- Hatchet: programmable by the user

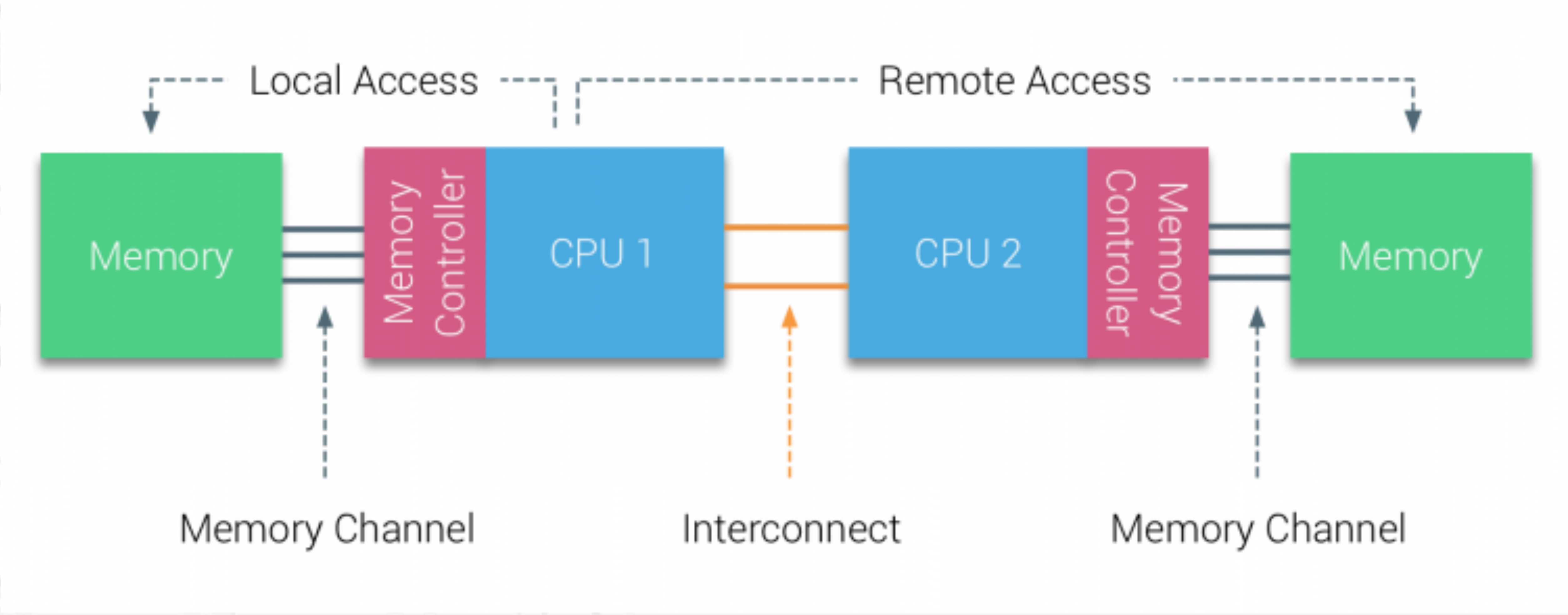
von Neumann architecture



https://en.wikipedia.org/wiki/Von_Neumann_architecture

UMA vs. NUMA



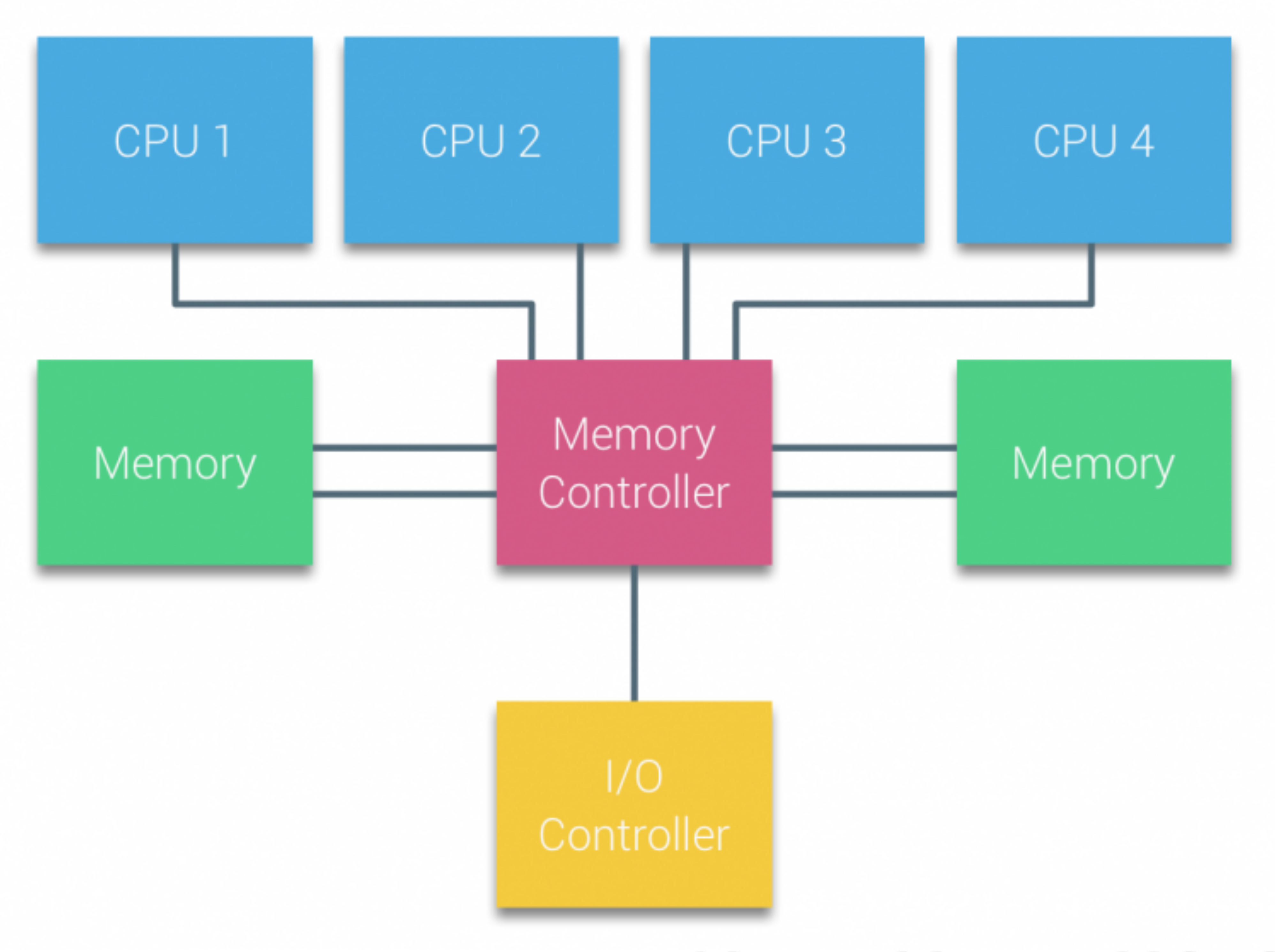


Uniform Memory Access

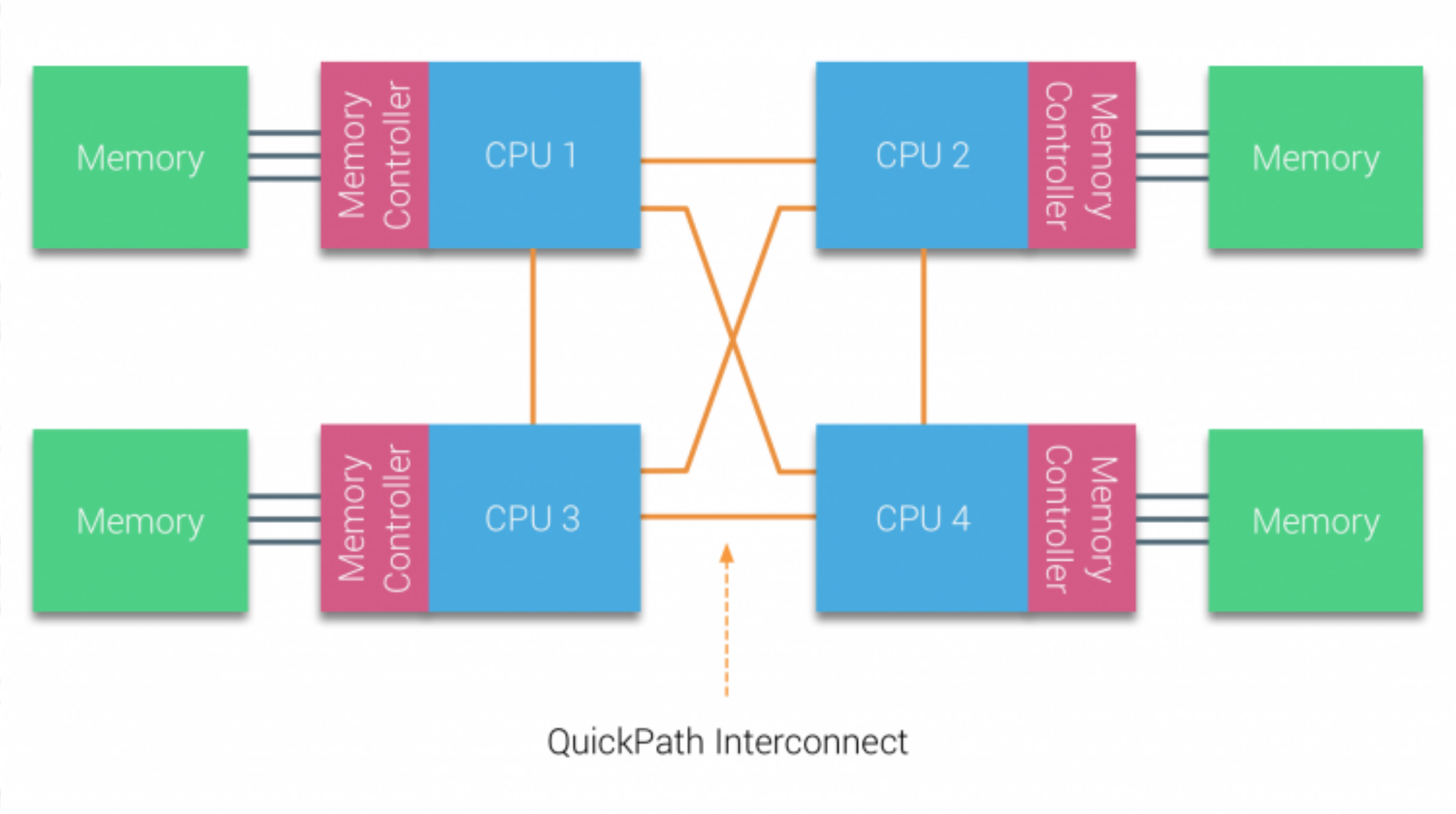
Non-uniform Memory Access

https://frankdenneman.nl/2016/07/07/numa-deep-dive-part-1-uma-numa/

UMA vs. NUMA



Uniform Memory Access



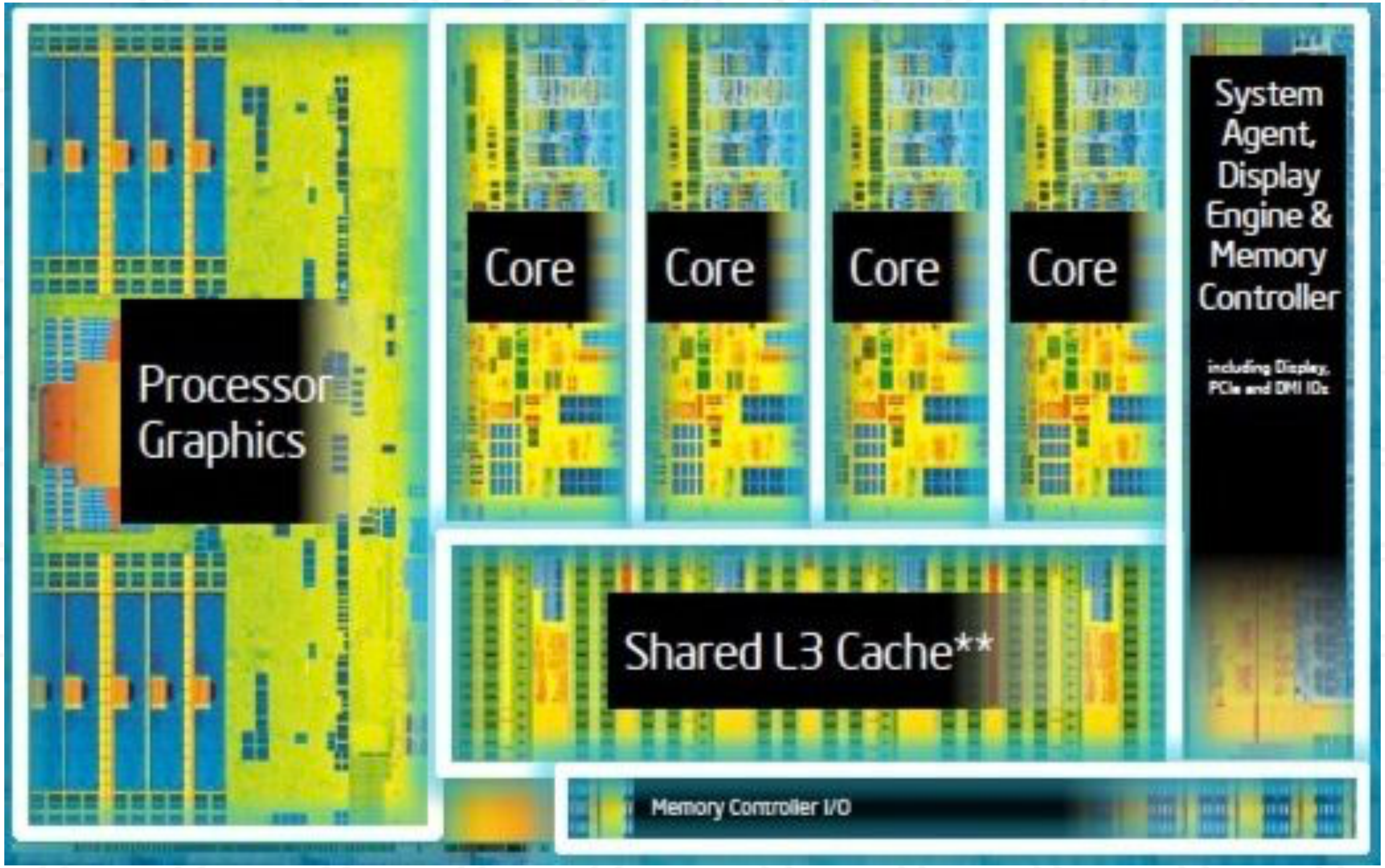
Non-uniform Memory Access

https://frankdenneman.nl/2016/07/07/numa-deep-dive-part-1-uma-numa/

Fast vs. slow cores

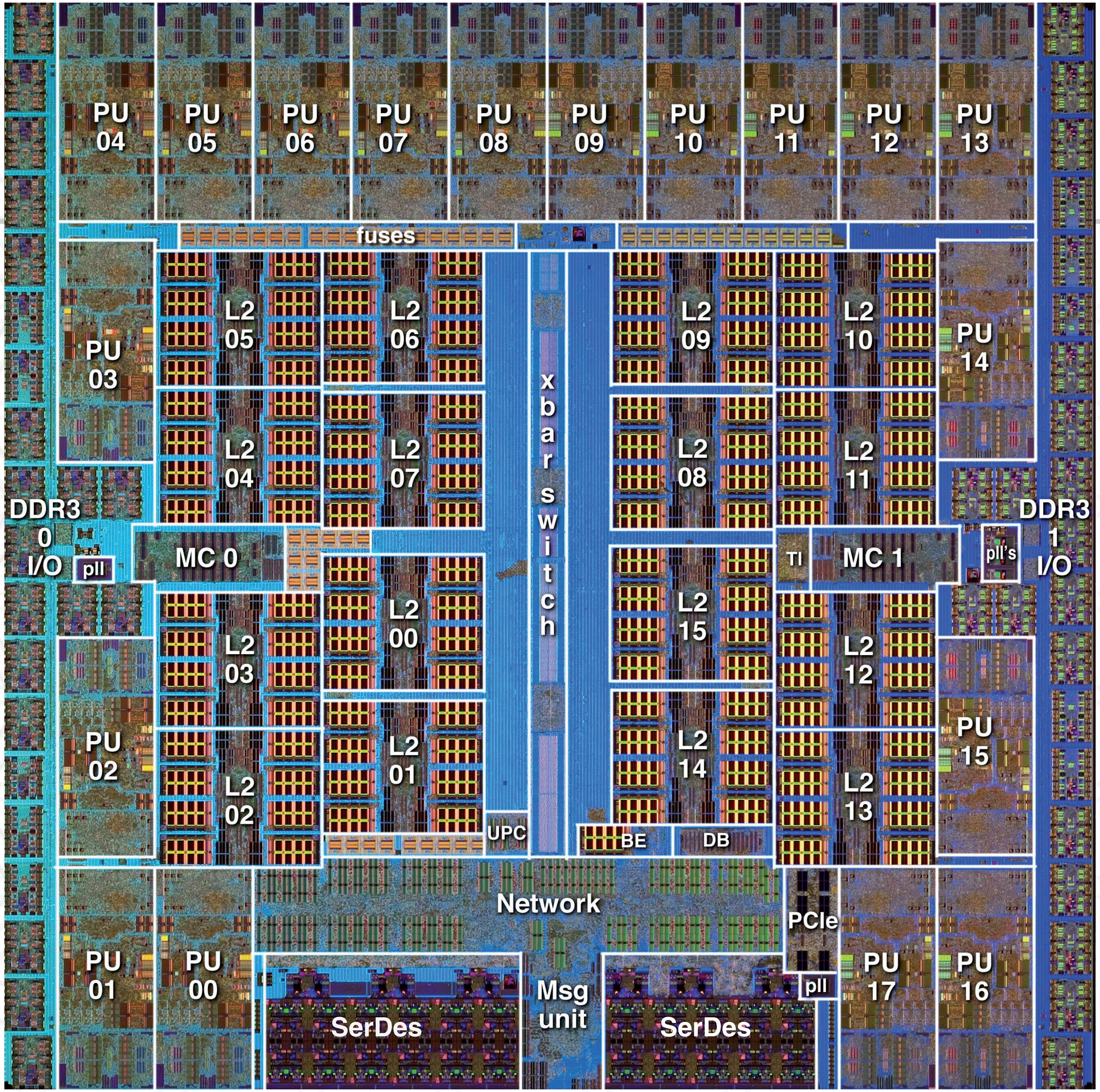
- Intel Core line (Nehalem, Sandy Bridge, Ivy Bridge, Haswell, Broadwell, …)
- AMD processors (Opteron, Athlon, Zen, …)
- IBM Power line
- Slower cores: Low frequency, low power
    - IBM PowerPC line (440, 450, A2, …)

Intel Haswell Chip


Processor Graphics
Core
Core
Core
Core
System Agent, Display Engine & Memory Controller
including Display, PCIe and DMI IOs
Shared L3 Cache**
Memory Controller I/O

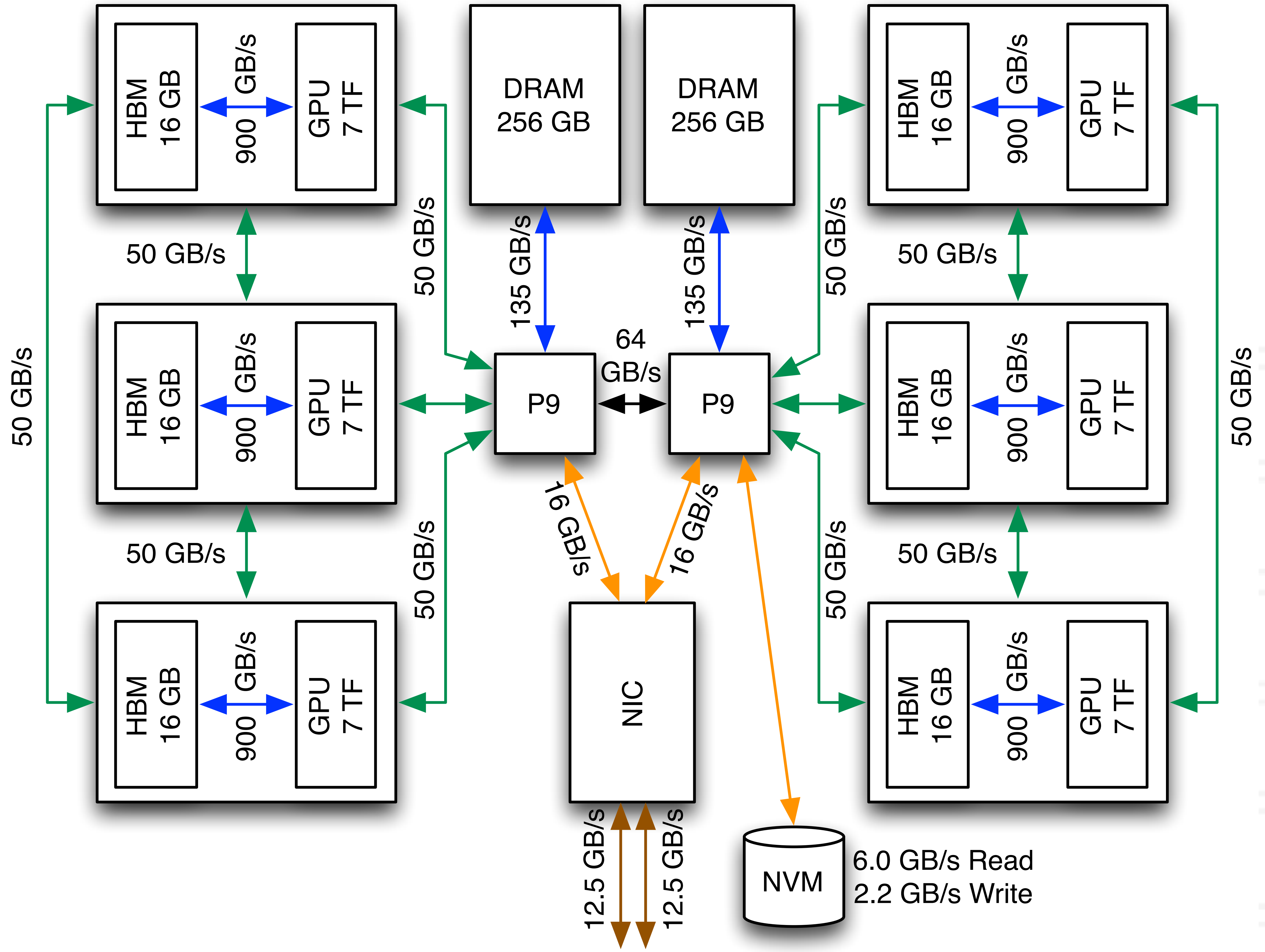
BQC Chip

- A2 processor core
  - Runs at 1.6 GHz
- Shared L2 cache
- Peak performance per core:
  - 12.8 Gflop/s
- Total performance per node: 204.8 Gflop/s



GPUs

- NVIDIA: Fermi, Kepler, Maxwell, Pascal, Volta, …
- AMD
- Intel
- Figure on the right shows a single node of Summit @ ORNL

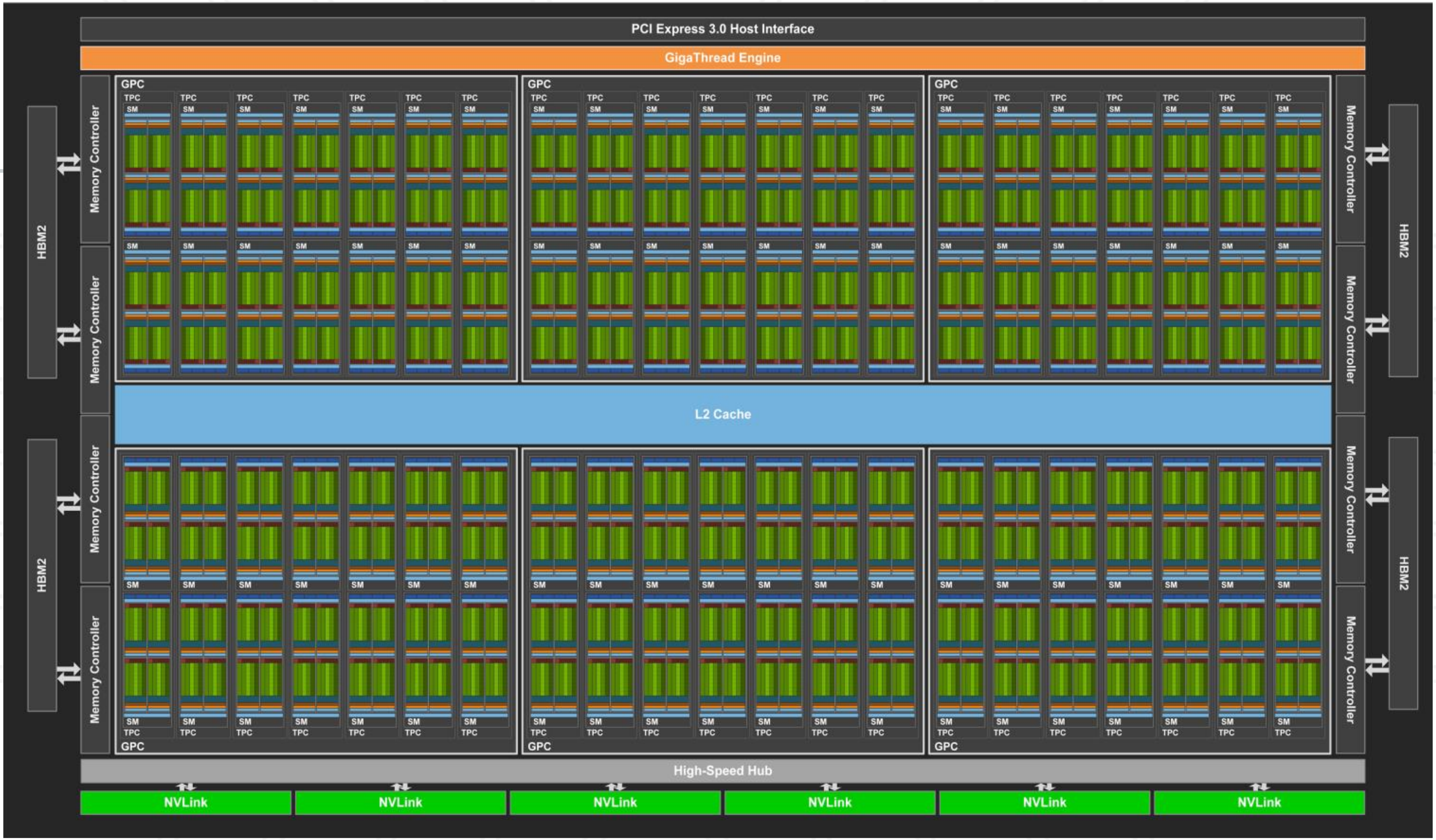


TF	42 TF (6x7 TF)
HBM	96 GB (6x16 GB)
DRAM	512 GB (2x16x16 GB)
NET	25 GB/s (2x12.5 GB/s)
MMsg/s	83

↔ HBM/DRAM Bus (aggregate B/W)
↔ NVLINK
↔ X-Bus (SMP)
↔ PCIe Gen4
↔ EDR IB

HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.

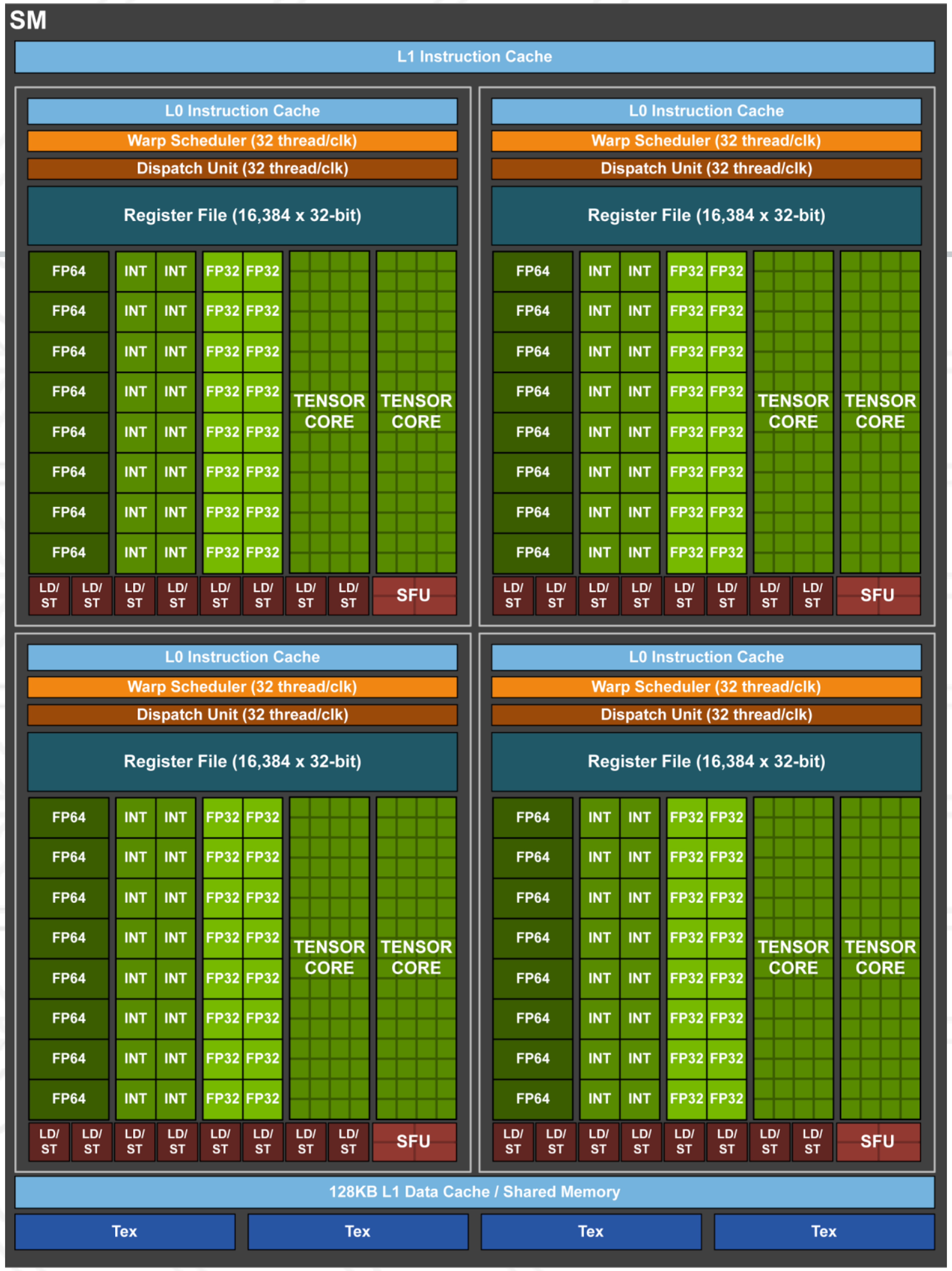
Volta GV100


PCI Express 3.0 Host Interface
GigaThread Engine
GPC
TPC
SM
Memory Controller
HBM2
L2 Cache
High-Speed Hub
NVLink

Volta GV100 SM

- Each Volta Streaming Multiprocessor (SM) has:
  - 64 FP32 cores
  - 64 INT32 cores
  - 32 FP64 cores
  - 8 Tensor cores

https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf


SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex

Questions?

UNIVERSITY OF MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu