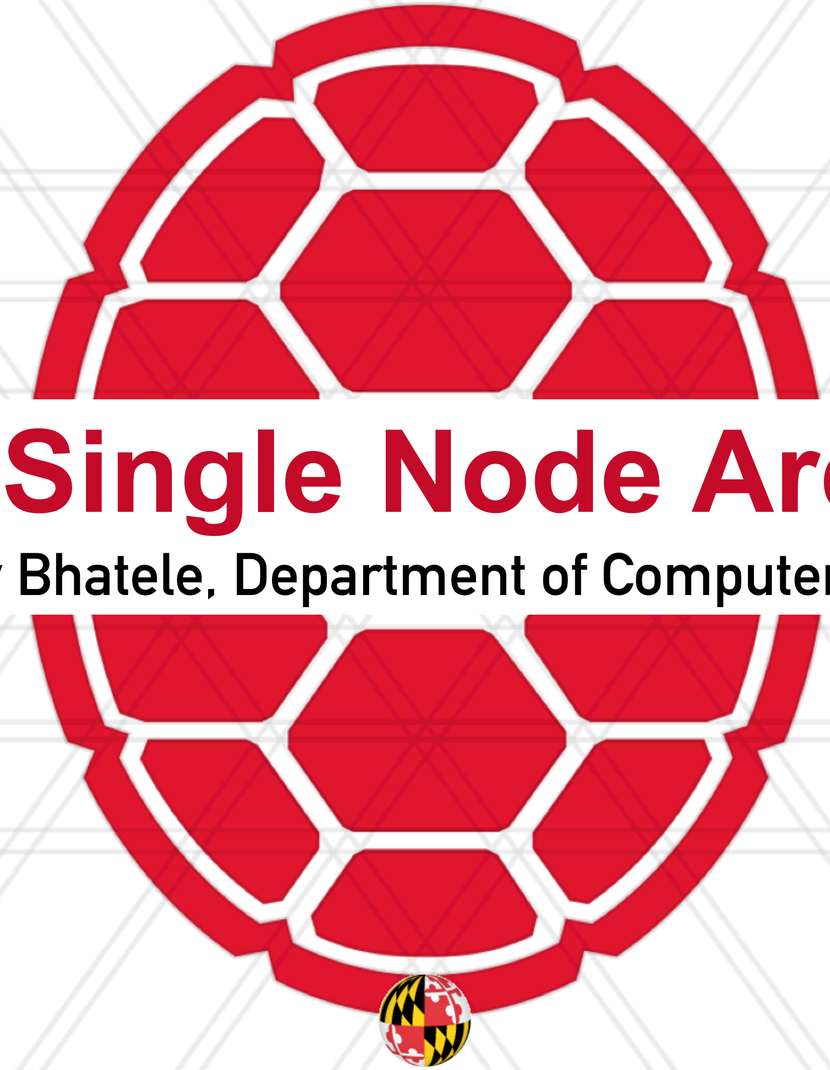
High Performance Computing Systems (CMSC714)



Lecture 7: Single Node Architectures

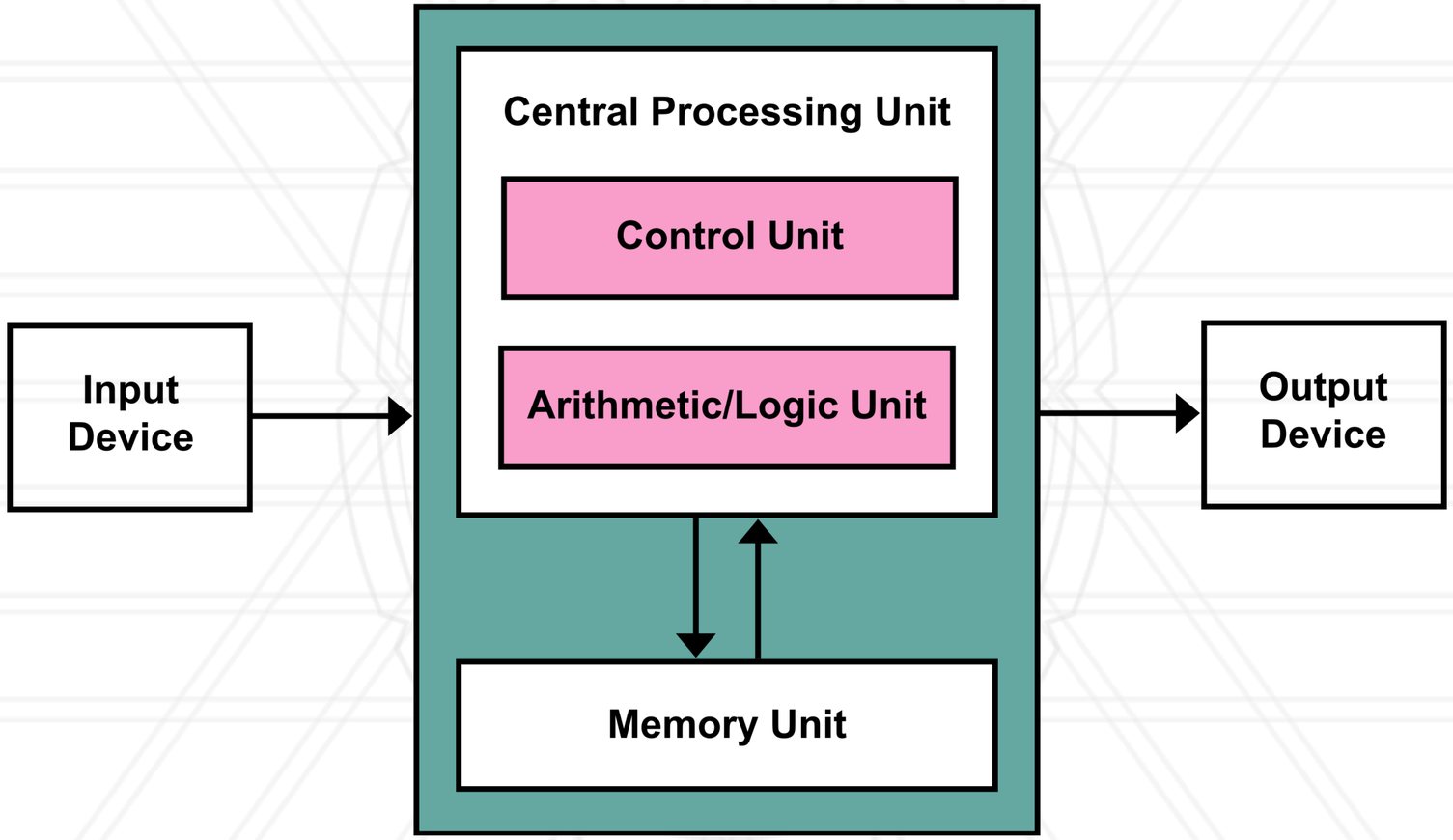
Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Summary of last lecture

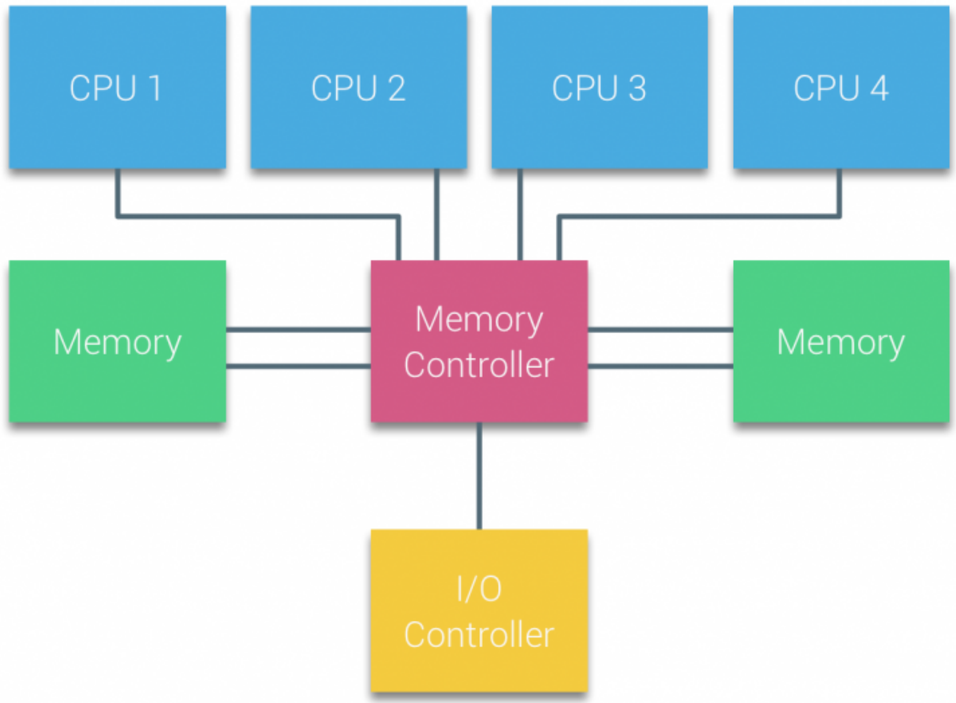
- Task-based programming models and Charm++
- Key principles:
  - Over-decomposition, virtualization
  - Message-driven execution
- Automatic load balancing, checkpointing, fault tolerance

von Neumann architecture

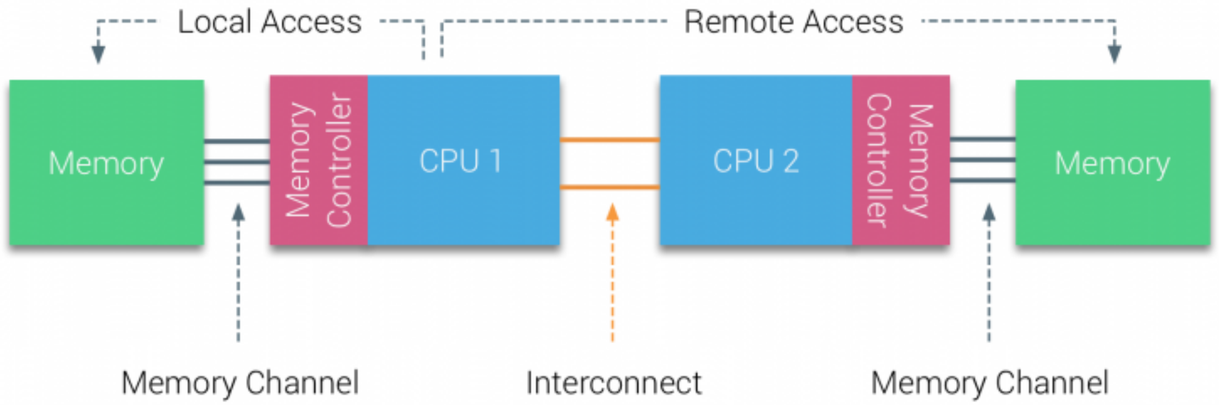


https://en.wikipedia.org/wiki/Von_Neumann_architecture

UMA vs. NUMA



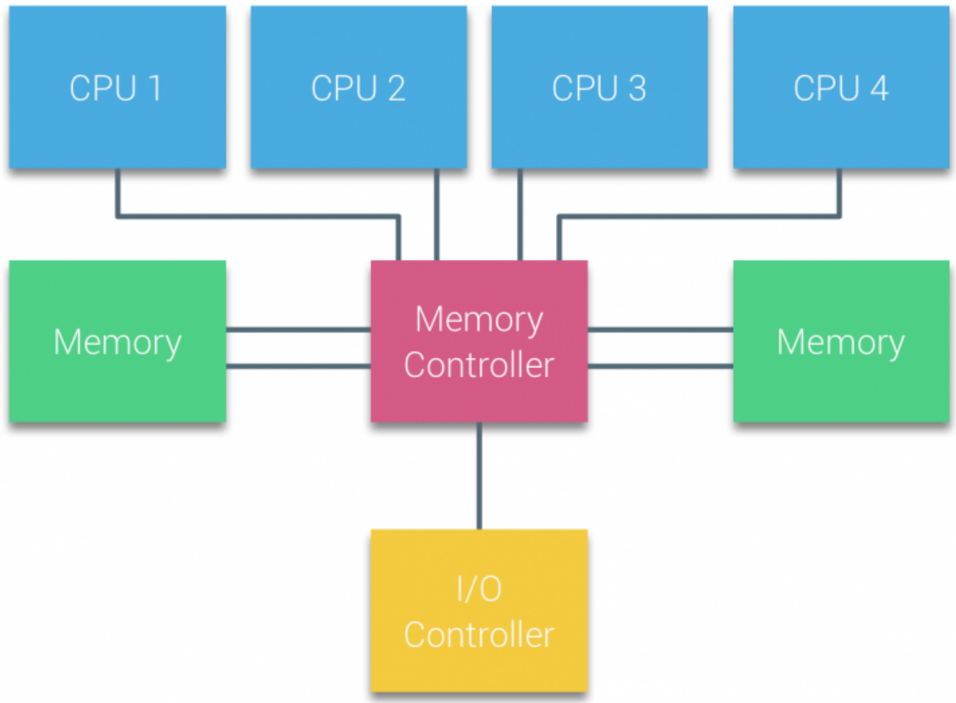
Uniform Memory Access



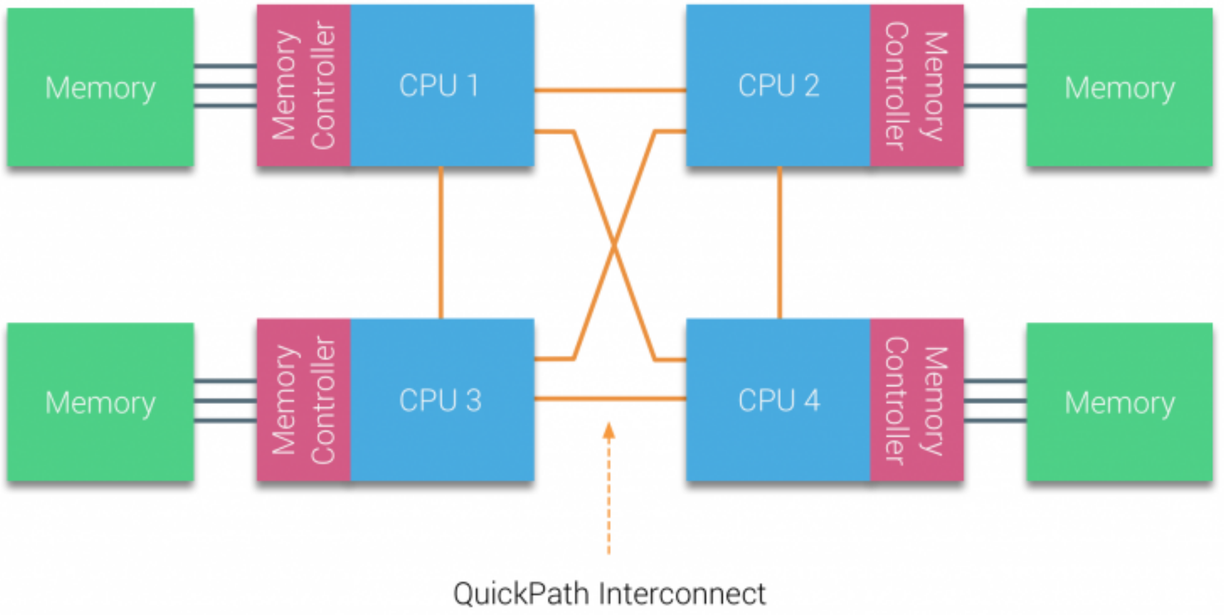
Non-uniform Memory Access

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

UMA vs. NUMA



Uniform Memory Access



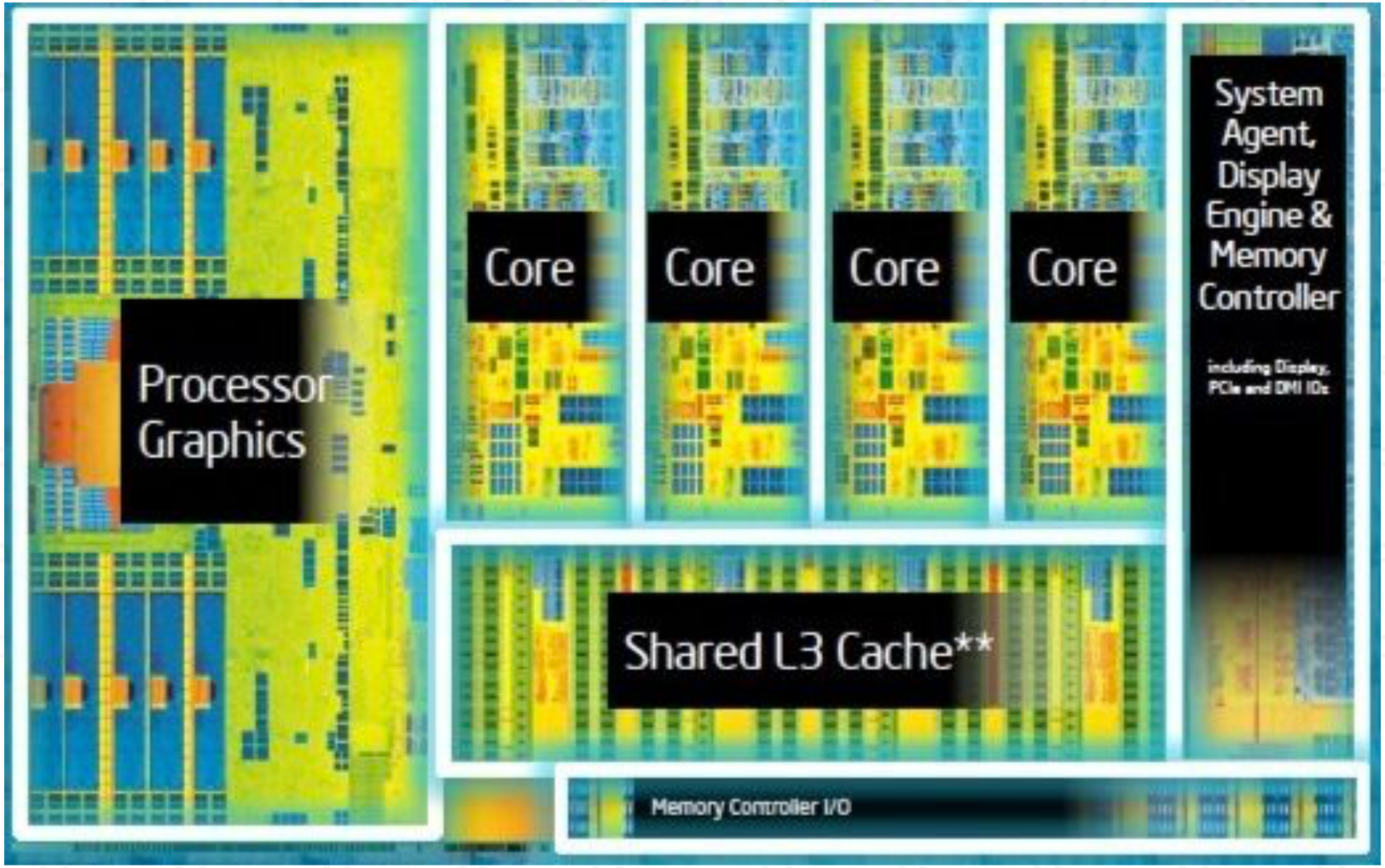
Non-uniform Memory Access

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

Fast vs. slow cores

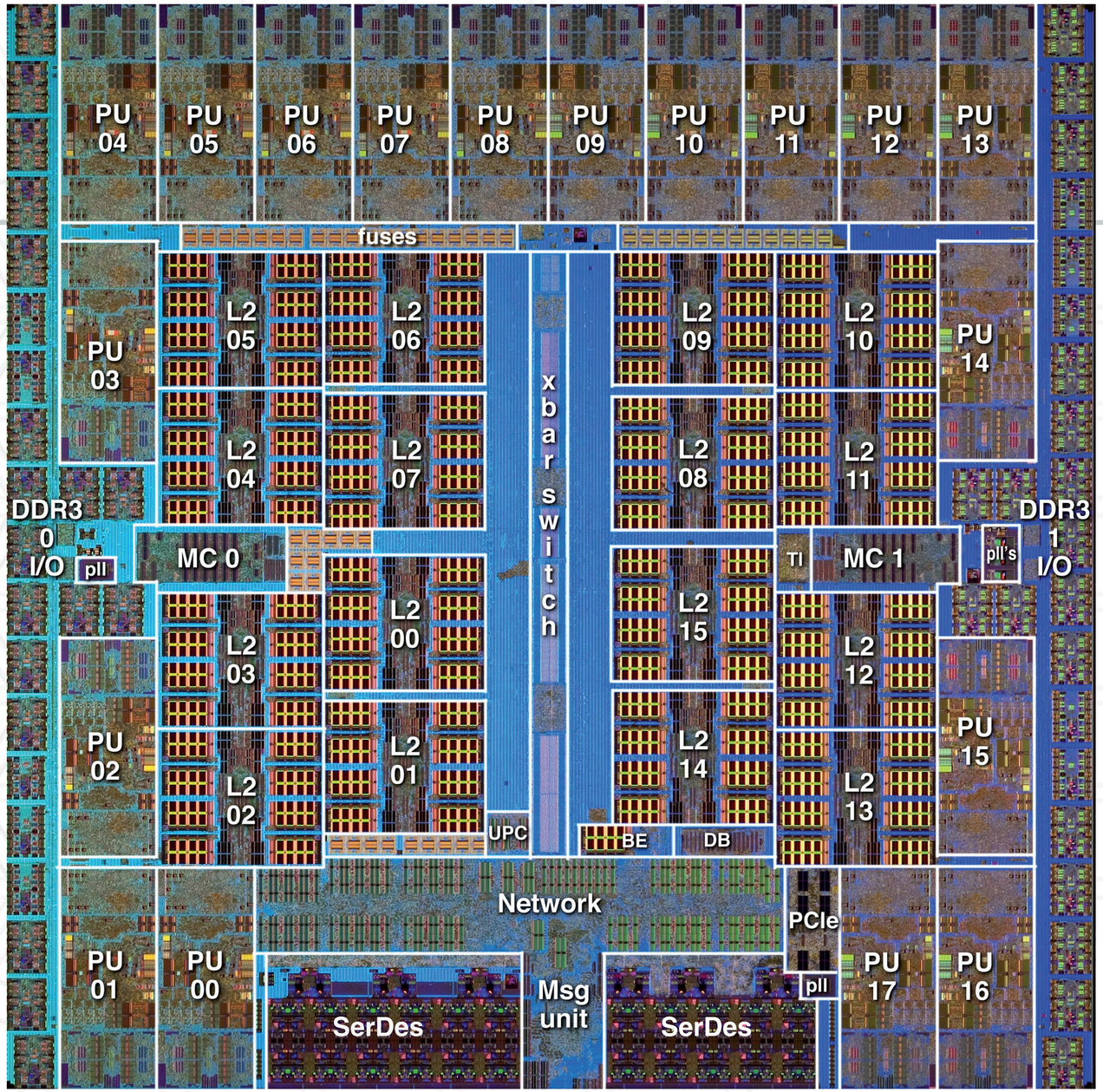
- Intel Core line (Nehalem, Sandy Bridge, Ivy Bridge, Haswell, Broadwell, …)
- AMD processors
- IBM Power line
- Slower cores: Low frequency, low power
  - IBM PowerPC line (440, 450, A2, …)

Intel Haswell Chip



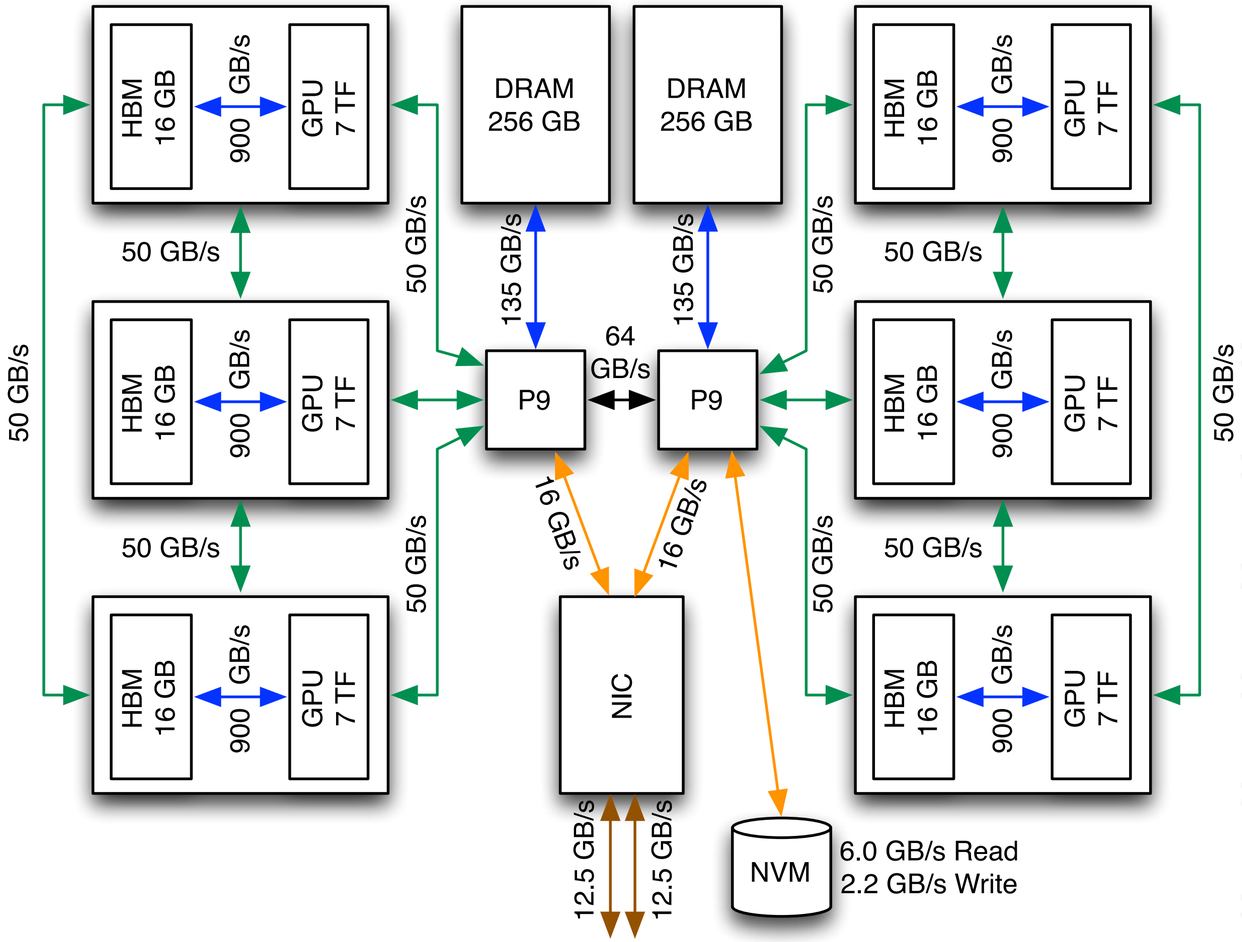
BQC Chip

- 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

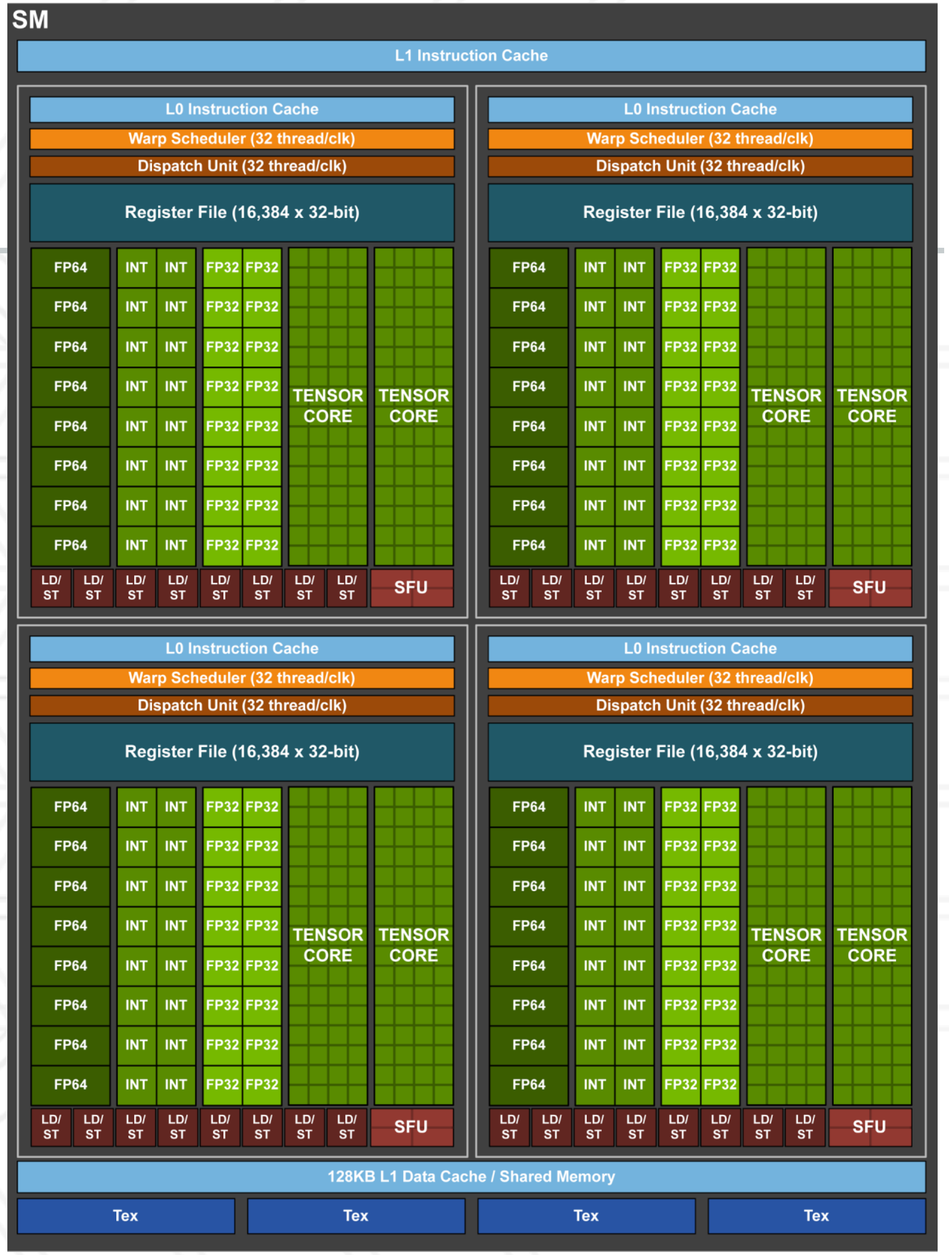
HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.

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

Questions

The IBM Blue Gene/Q Compute Chip

- Why are the L2 caches sitting on the center of the chip? Why not vise-versa? Is this a standard design?
- Why is this paper, Blue Gene/Q, or A2 processor, so important?
- Are there new significant prefetching methods other than list and stream prefetching in recent architectures?
- Is "multiply add pipeline" a commonly used operation, or is the architecture just trying to increase its FLOP count? What are other commonly used operations that get pipelined in other architectures?

Questions

Debunking the 100X GPU vs. CPU myth

- The paper is from 2010 and this is rather old. It seems that GPUs have evolved a lot in the last decade. How would it compare today?
  - The GPU in the first paper is 1.5 years older than the CPU, what would be the results if they were both from the same time? How does Moore's law apply to the GPUs, do they get 2x faster every 2 years?
- GPUs have several types of caches (shared buffer, constant cache, texture cache). How should these caches be differentiated (chosen) for a purpose?
- Where did the "myth" come from? Is the CPU more difficult to optimize?
- Have the features, recommended by the author, become true in current CPUs/GPUs?
- Why radix sort is chosen as a benchmark metric, while it's not used as the default algorithm in most programming languages? (java has mergesort, python timsort, C++ implements quicksort) Is it used more in HPC?
- The paper says they discarded the delays related to memory bandwidth because GPU have 5x faster b/w than CPU. What would be the approximate real life speeds with the memory included? How important is that to optimize bandwidth?

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