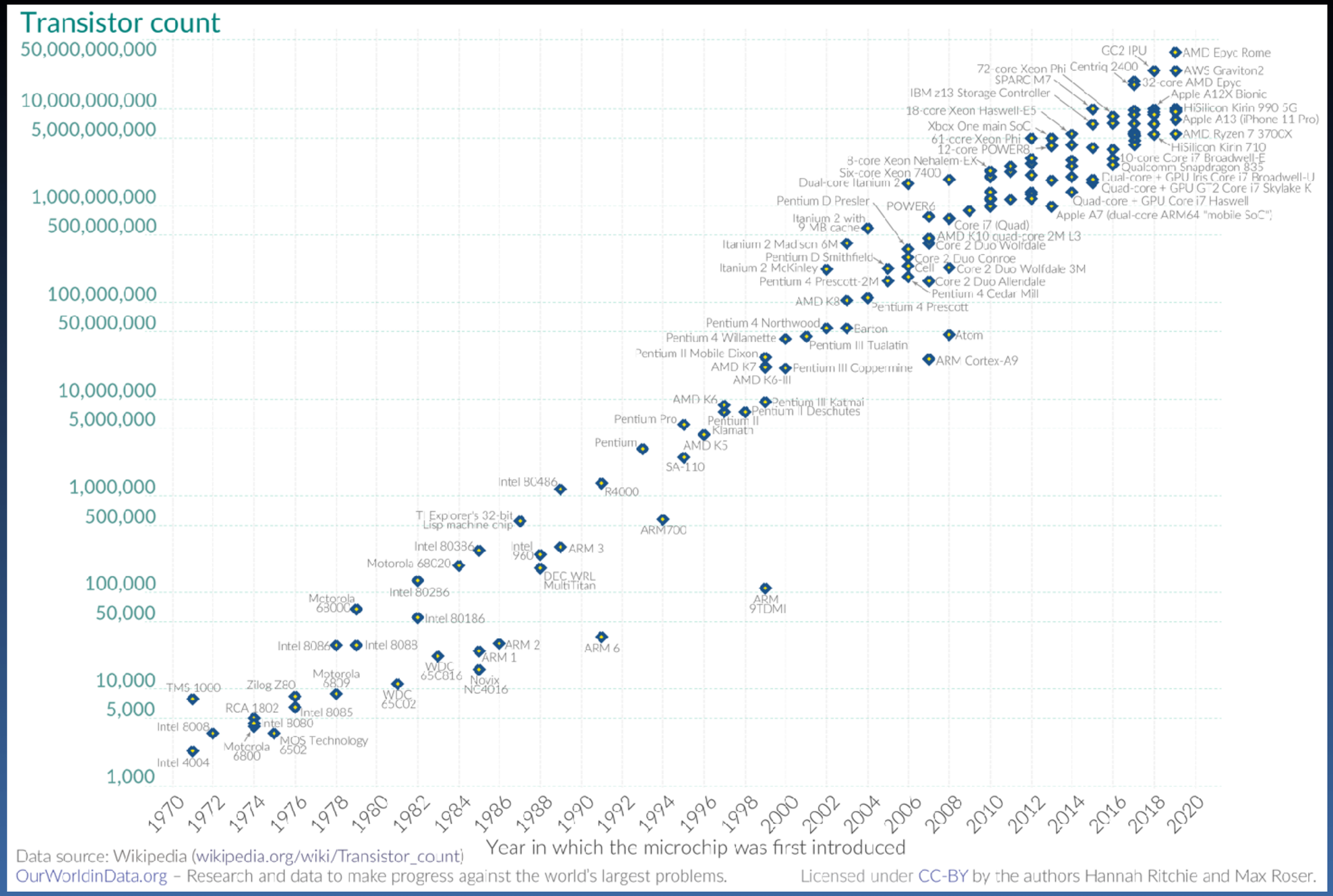
COL380

Introduction to
Parallel & Distributed Programming

Agenda

- Memory bottlenecks
  - ➡ and some solutions
- Instruction latency and overlap
- Core organization
- Inter-communication

Moore's Law



Subodh Kumar

Why Parallel

- Can’t clock faster ⇒ Do more per clock
  - ➡ Execute many simple instructions on many cores
  - ➡ Simpler operations are more general
  - ➡ Complex operations require hardware coordination across the chip
- Not just compute more things
  - ➡ Focus may be on parallelizing data access (Memory, IO)
  - ➡ Multiple processors can access memory in parallel, disrupt caches

Why Parallel

- Can’t clock faster ⇒ Do more per clock
  - ➡ Execute many simple instructions on many cores
  - ➡ Simpler operations are more general
  - ➡ Complex operations require hardware coordination across the chip

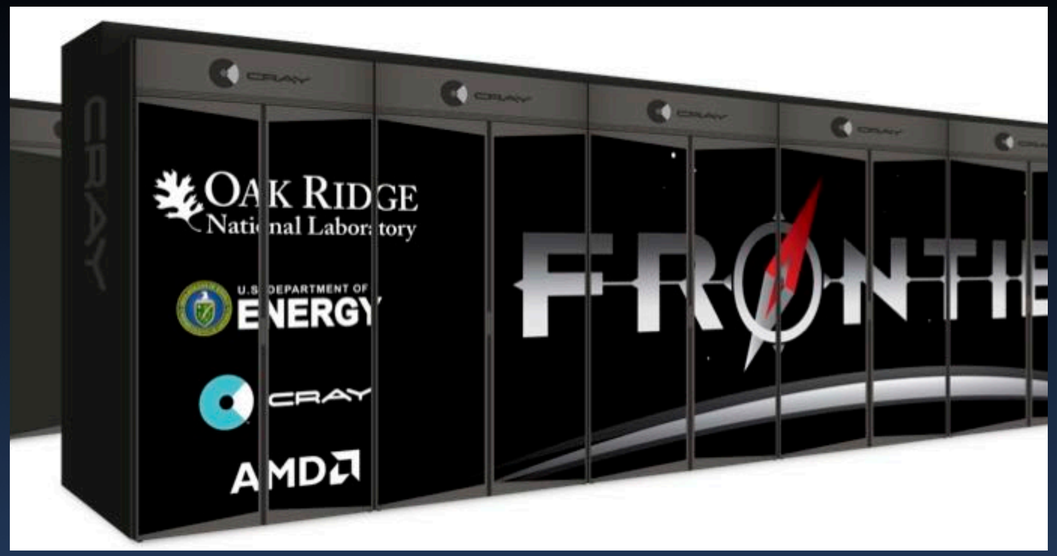
Software orchestration

- Not just compute more things
  - ➡ Focus may be on parallelizing data access (Memory, IO)
  - ➡ Multiple processors can access memory in parallel, disrupt caches

Application Areas

- Weather/Climate simulation: 3D-grid, Long duration simulation
- Energy modeling: Fusion energy
- Data science: Filter, Join, Cross, Sort
- Financial processing: Market prediction, Investing, Blockchain
- Computational biology: Drug design, Gene sequencing, Vaccines
- Distributed Service: DB and File systems, Traffic network, Archive

Supercomputer



Today: “Fastest supercomputer in the world”
[HPL Rmax >1EF: 1102000000000000000]
Nodes: 9,472
Cores: 8,730,112 (64c CPU, 4x GPU/node)
Memory: 4.6 PB of DRAM memory + Flash
Interconnect: Multiple 100 GB/s NIC
- Racks: 74 cabinets
- Space: 7300 SqFt
- Power consumption: 21.1 MW

FRONTIER

Supercomputer

Arm CPU
A64FX: 48 core CPU with 512-bit SIMD
Peak Flops: 3379G (DP) [70.4G/core]
Memory BW: 1 TB/s
Network: 6D Torus [68GB/s x2] *

Today: “Fastest supercomputer in the world”
[HPL Rmax >1EF: 1102000000000000000]
Nodes: 9,472
8,730,112 (64c CPU, 4x GPU/node)
ry: 4.6 PB of DRAM memory + Flash
nnect: Multiple 100 GB/s NIC
ks: 74 cabinets
ce: 7300 SqFt
Power consumption: 21.1 MW

FRONTIER

Supercomputer

Today: “Fastest supercomputer in the world”
[HPL Rmax >1EF: 1102000000000000000]
Nodes: 9,472
8,730,112 (64c CPU, 4x GPU/node)
ry: 4.6 PB of DRAM memory + Flash
onnect: Multiple 100 GB/s NIC
ks: 74 cabinets
ce: 7300 SqFt
ver consumption: 21.1 MW

FRONTIER

Arm CPU
A64FX: 48 core CPU with 512-bit SIMD
Peak Flops: 3379G (DP) [70.4G/core]
Memory BW: 1 TB/s
N

Intel CPU
2.3 GHz x40 cores (~3000 GFlop)
+ 2x AVX-512 FMA units
Maximum Memory Speed: 3200 MHz
Memory Channels: 8
Memory bandwidth: ~200 GB/s

Supercomputer

Today: “Fastest supercomputer in the world”
[HPL Rmax >1EF: 1102000000000000000]
Nodes: 9,472
8,730,112 (64c CPU, 4x GPU/node)
ry: 4.6 PB of DRAM memory + Flash
onnect: Multiple 100 GB/s NIC
ks: 74 cabinets
ce: 7300 SqFt

FRONTIER

Arm CPU
A64FX: 48 core CPU with 512-bit SIMD
Peak Flops: 3379G (DP) [70.4G/core]
Memory BW: 1 TB/s

Intel CPU
2.3 GHz x40 cores (~3000 GFlop)
+ 2x AVX-512 FMA units
Maximum Memory Speed: 3200 MHz
Memory Channels: 8
Memory bandwidth: ~200 GB/s

Nvidia GPU
1000+ Cores: 9.7 TF (DP)
19.5 TF with Tensor Core
GPU Memory Bandwidth: 1.6 TB/s
Network: NVLink 600 GB/s

Toy Supercomputer

Cray XMP-1

The New York Times

India and U.S. Agree On Supercomputer Sale

Give this article

By Steven R. Weisman, Special To the New York Times

Oct. 9, 1987

USD 20 MILLION

4x 64-bit Vector processor, ~117 MHz
400 MFLOPS (peak)
128 MB RAM

Toy Supercomputer

Arduino

Jetson Tx 472 GF

The New York Times

India and U.S. Agree On Supercomputer Sale

Give this article

By Steven R. Weisman, Special To the New York Times

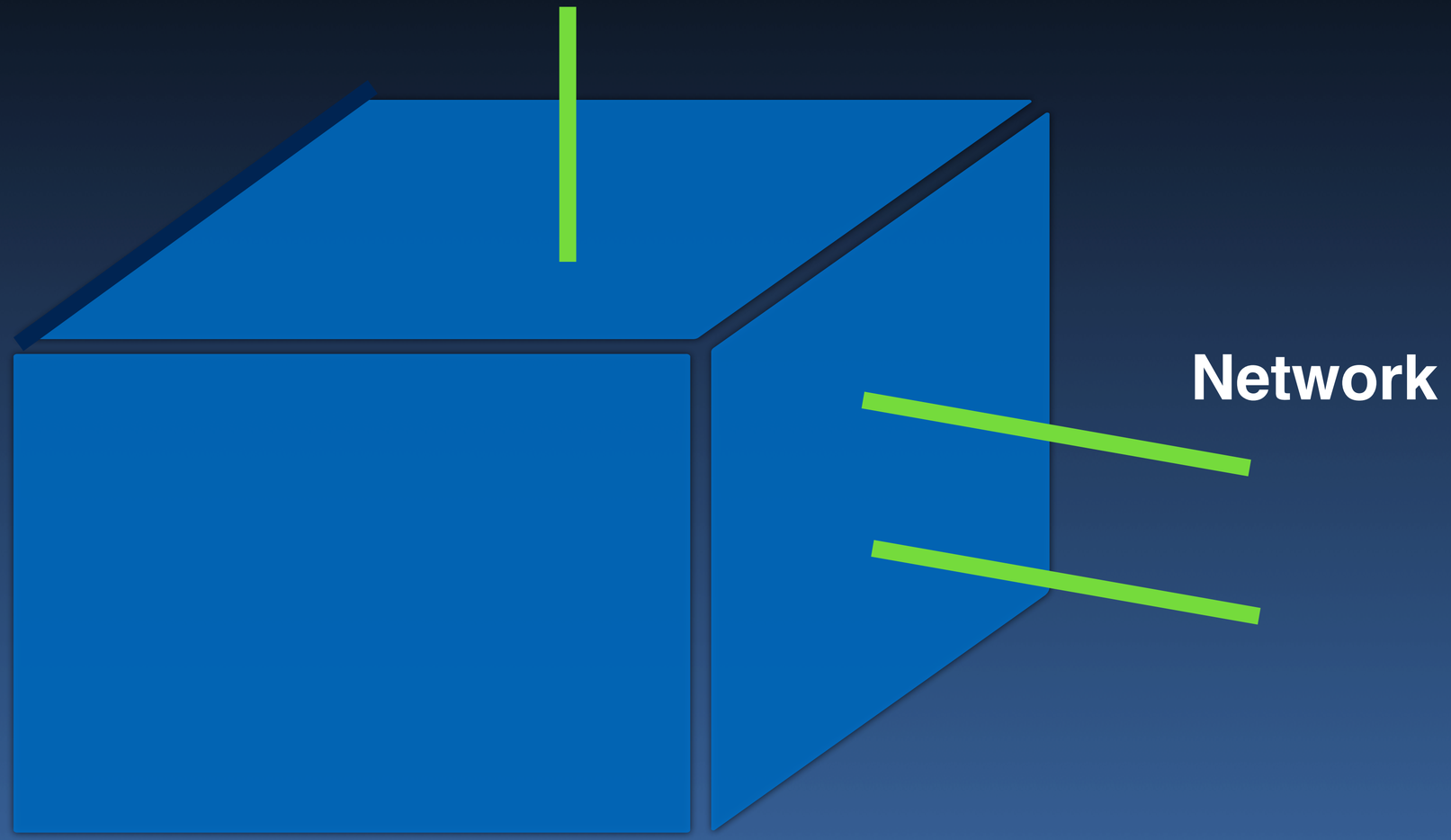
Oct. 9, 1987

USD 20 MILLION

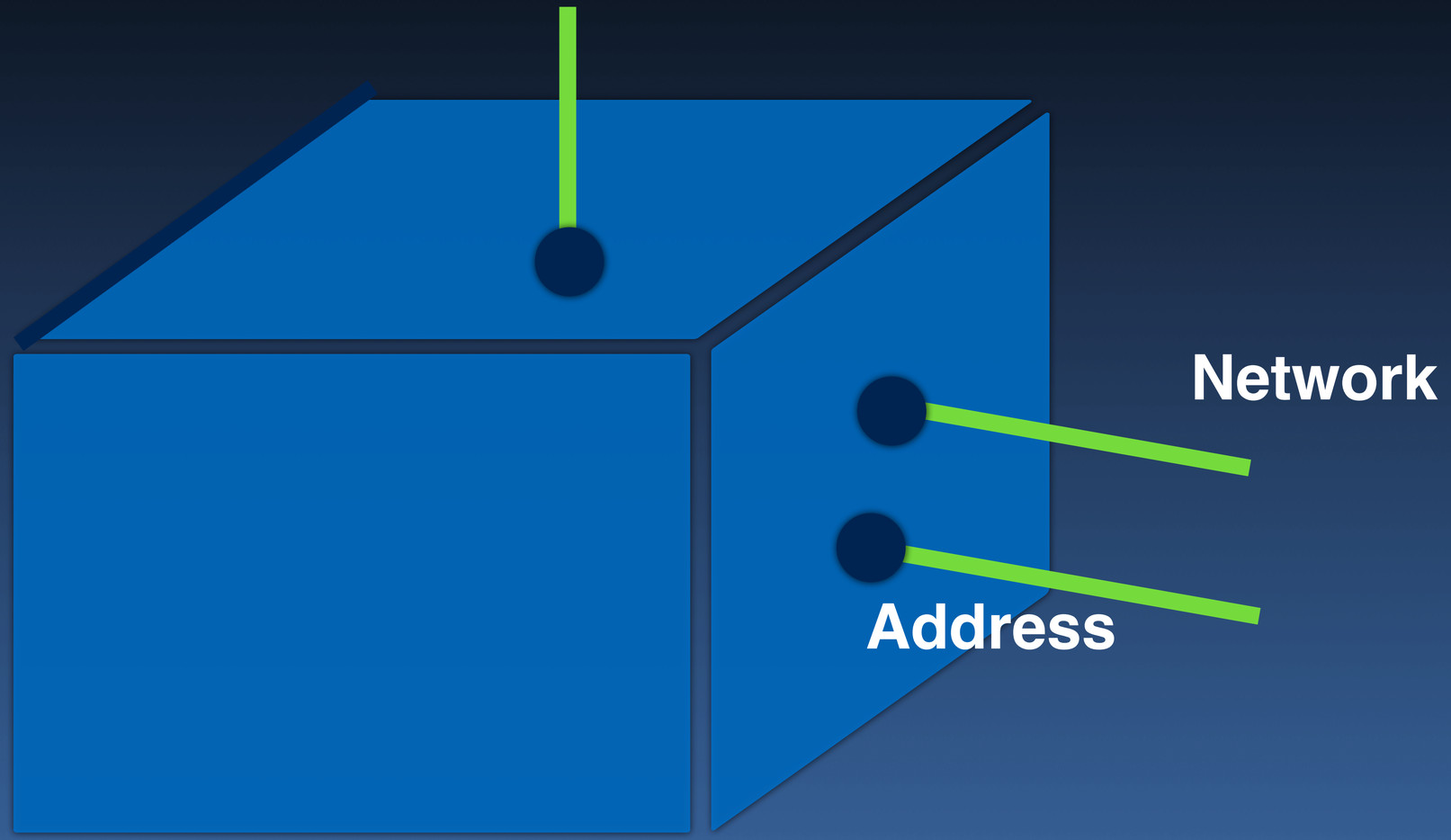
4x 64-bit Vector processor, ~117 MHz
400 MFLOPS (peak)
128 MB RAM

Subodh Kumar

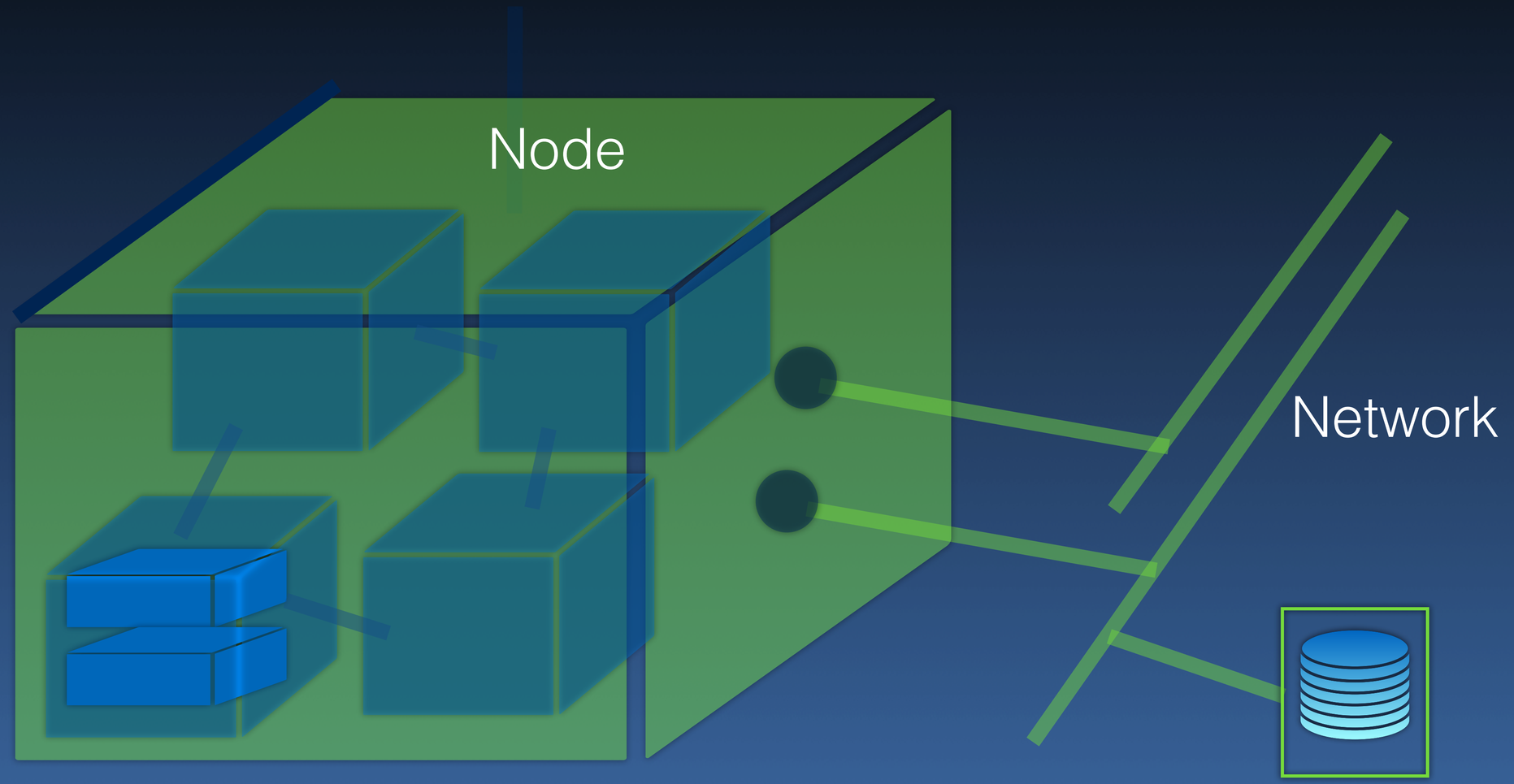
Supercomputer Node



Supercomputer Node

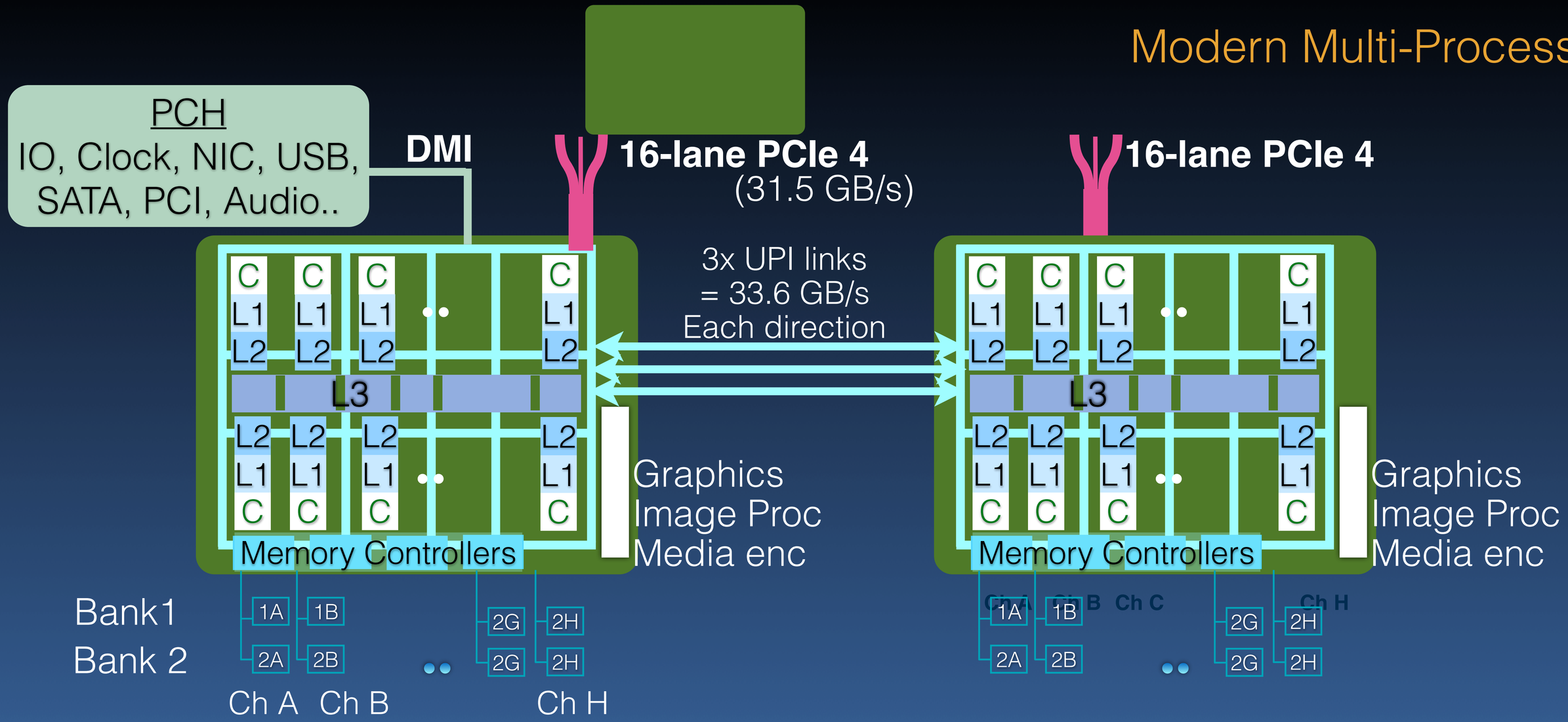


Supercomputer Node

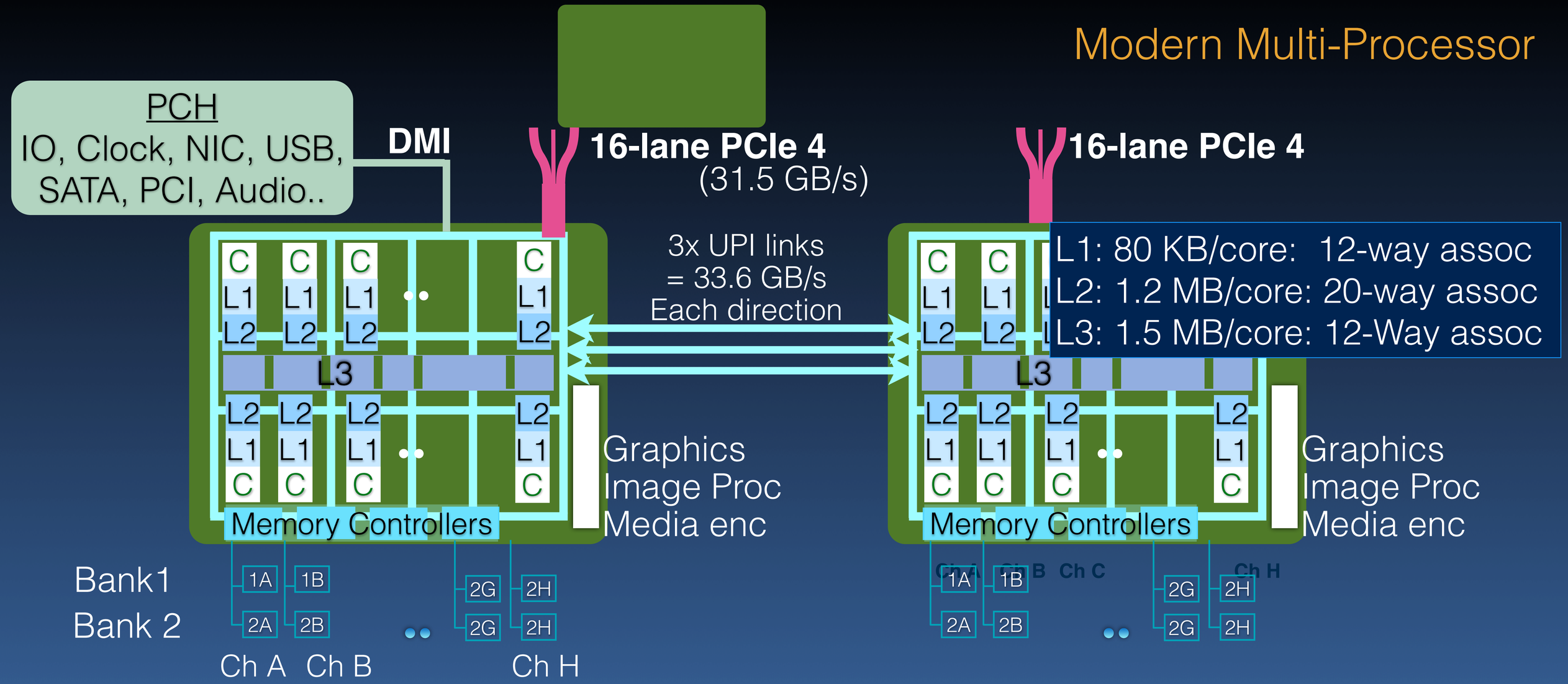


Subodh Kumar

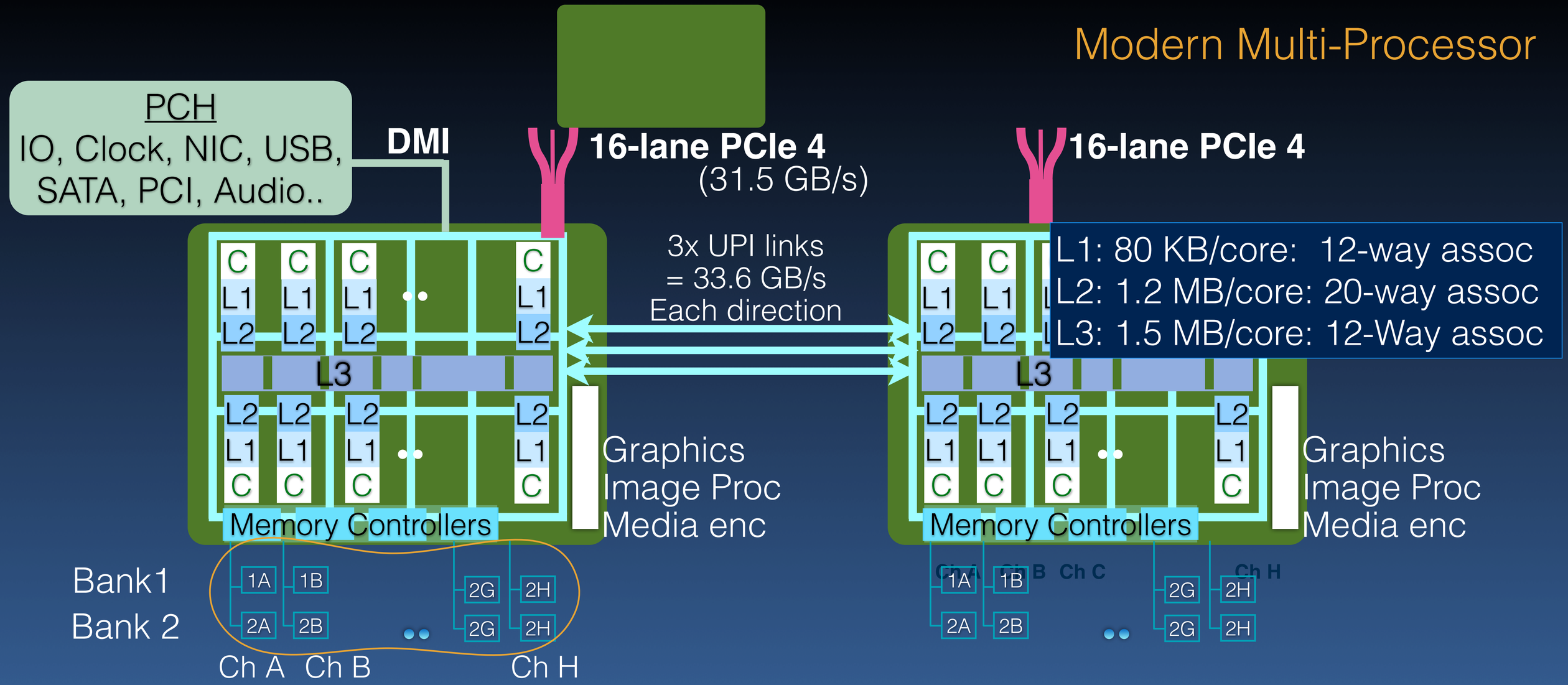
Modern Multi-Processor



Modern Multi-Processor



Modern Multi-Processor

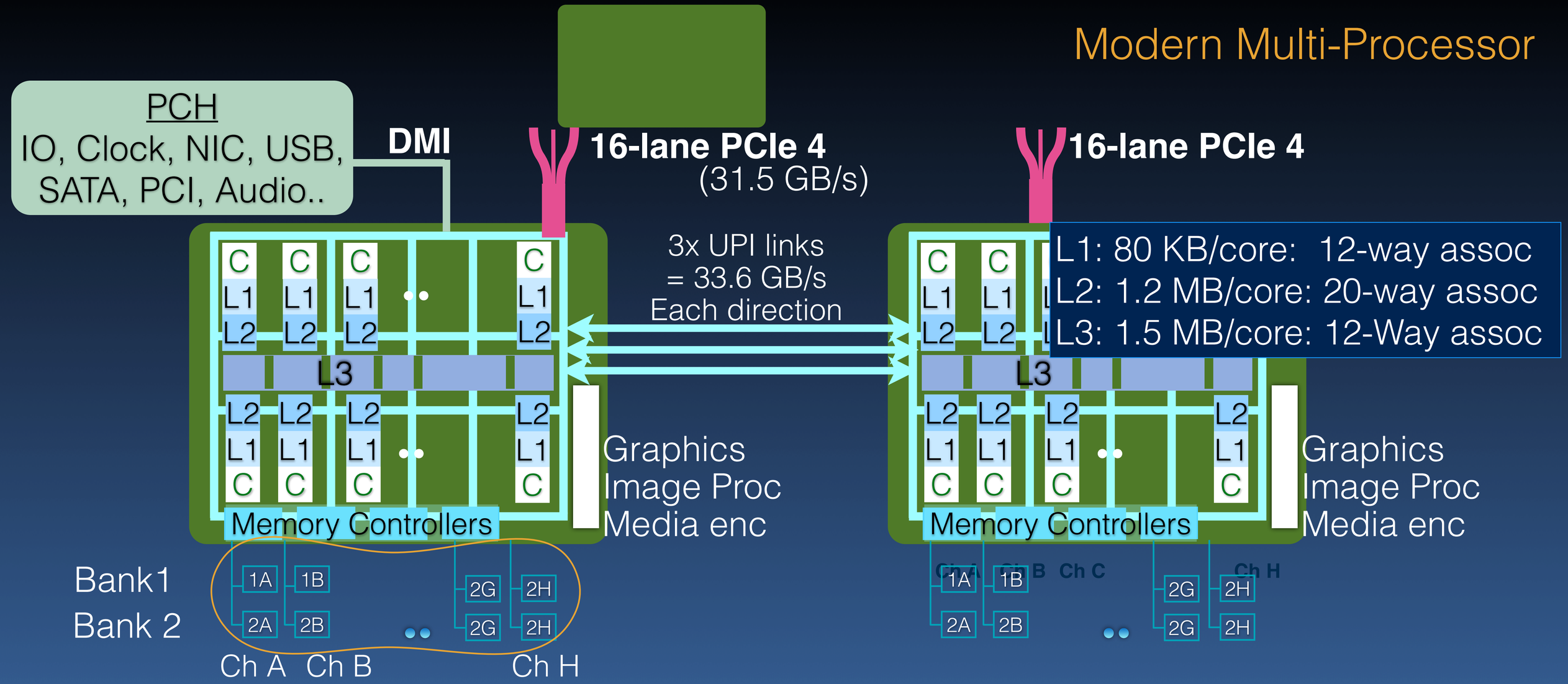


8 channels of DDR4-3200 DIMM

25.6 * 8 GB/s

Subodh Kumar

Modern Multi-Processor



8 channels of DDR4-3200 DIMM

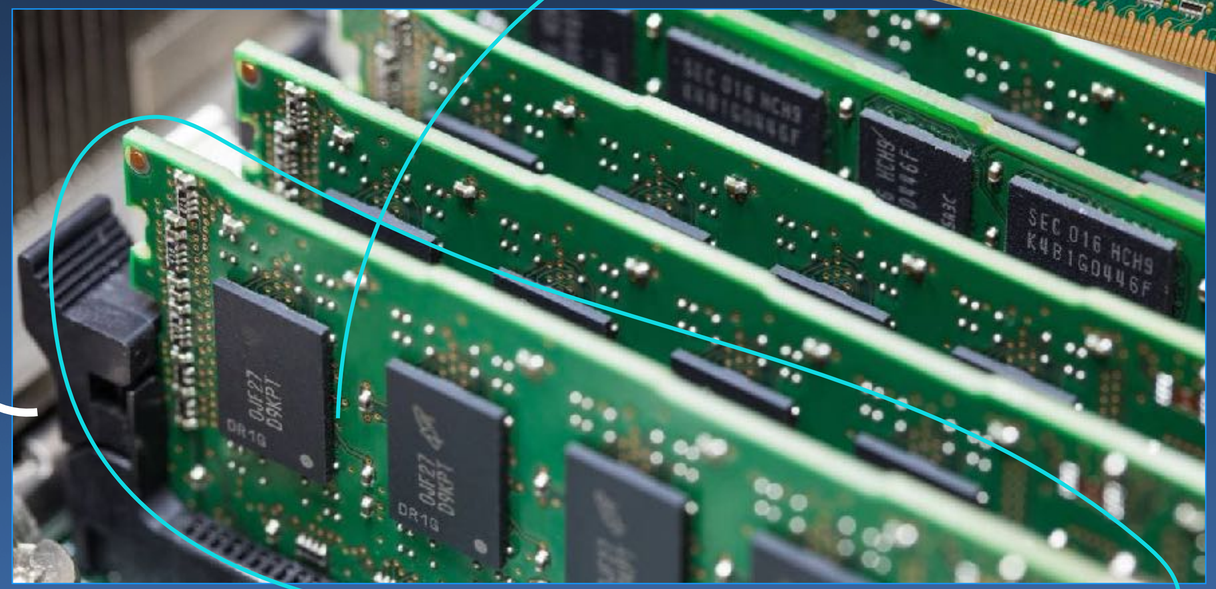
25.6 * 8 GB/s

100+ GF/core
32 cores = 3TF

Subodh Kumar

Memory

Memory
Controller



Subodh Kumar

Memory Bottleneck

- Fused Multiply Add
  - Double Precision A += B * C
  - 2 FLOPS, 3+1 Operands (32 bytes)
- Example
  - 2x AVX512 on Intel core = 32 FLOP/cycle
  - 96 GF/core @3GHz
    - Needed 1536 GB/s memory bandwidth/core
  - Compare DDR4: Throughput ~25GB/s, Latency ~10ns

Memory Bottleneck

- **Fused Multiply Add**
  - Double Precision A += B * C
  - 2 FLOPS, 3+1 Operands (32 bytes)
- **Example**
  - 2x AVX512 on Intel core = 32 FLOP/cycle
  - 96 GF/core @3GHz
    - Needed 1536 GB/s memory bandwidth/core
  - Compare DDR4: Throughput ~25GB/s, Latency ~10ns

Latency can be hidden

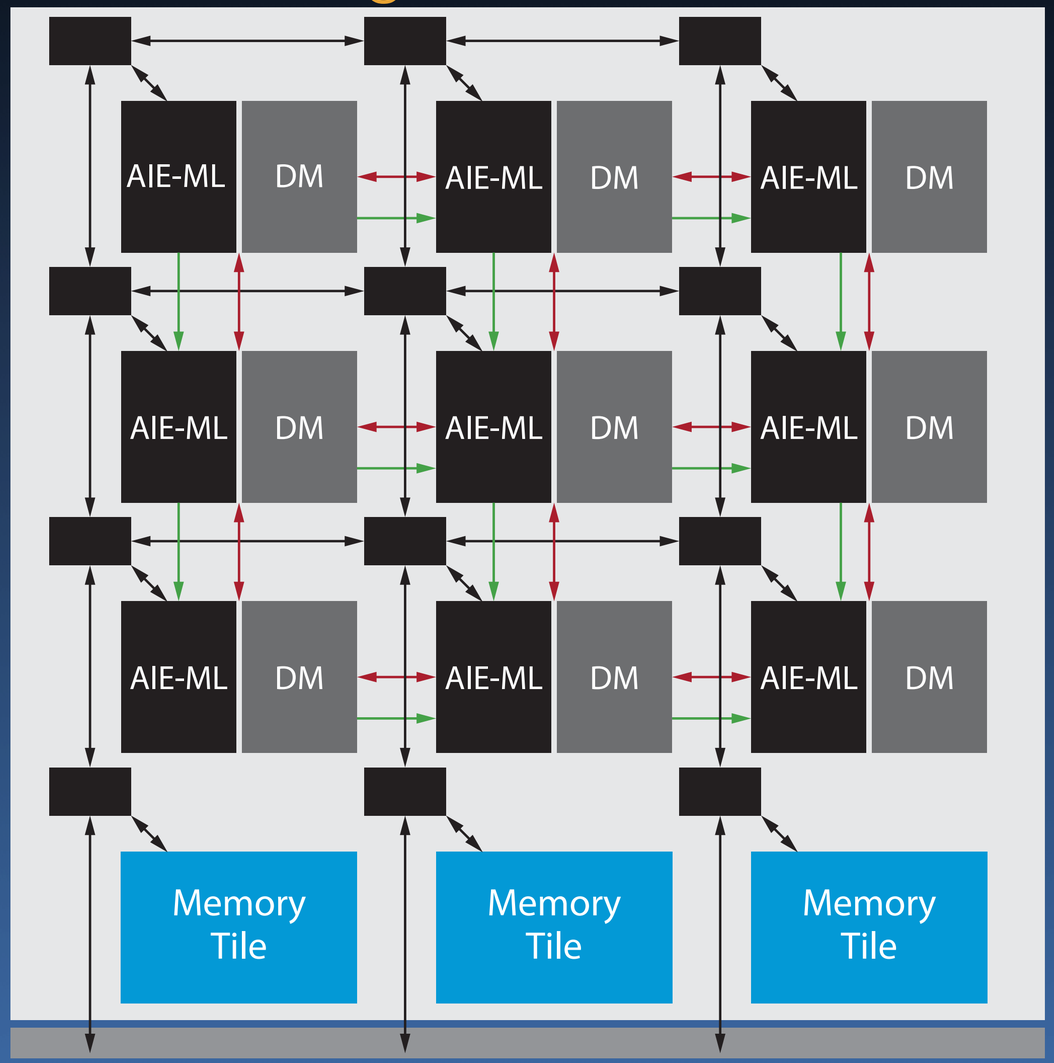
```
A[i] += B[i]*C[i]
A[i+1] += B[i+1]*C[i+1]
A[i+2] += B[i+2]*C[i+2]
A[i+3] += B[i+3]*C[i+3]
```

Caches can help with throughput but working set can be large

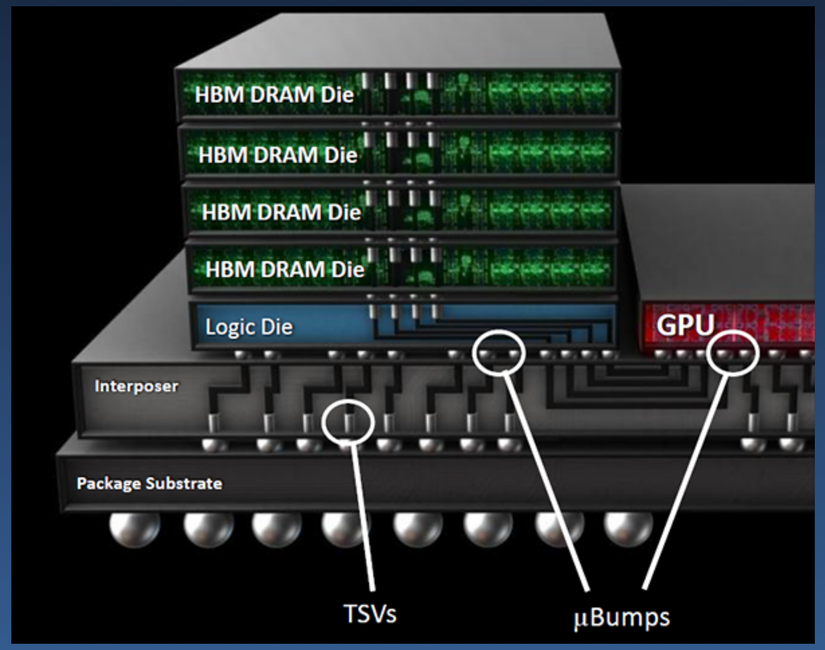
✳ Must be used wisely

Memory Gap Mitigation

AMD AI Engine

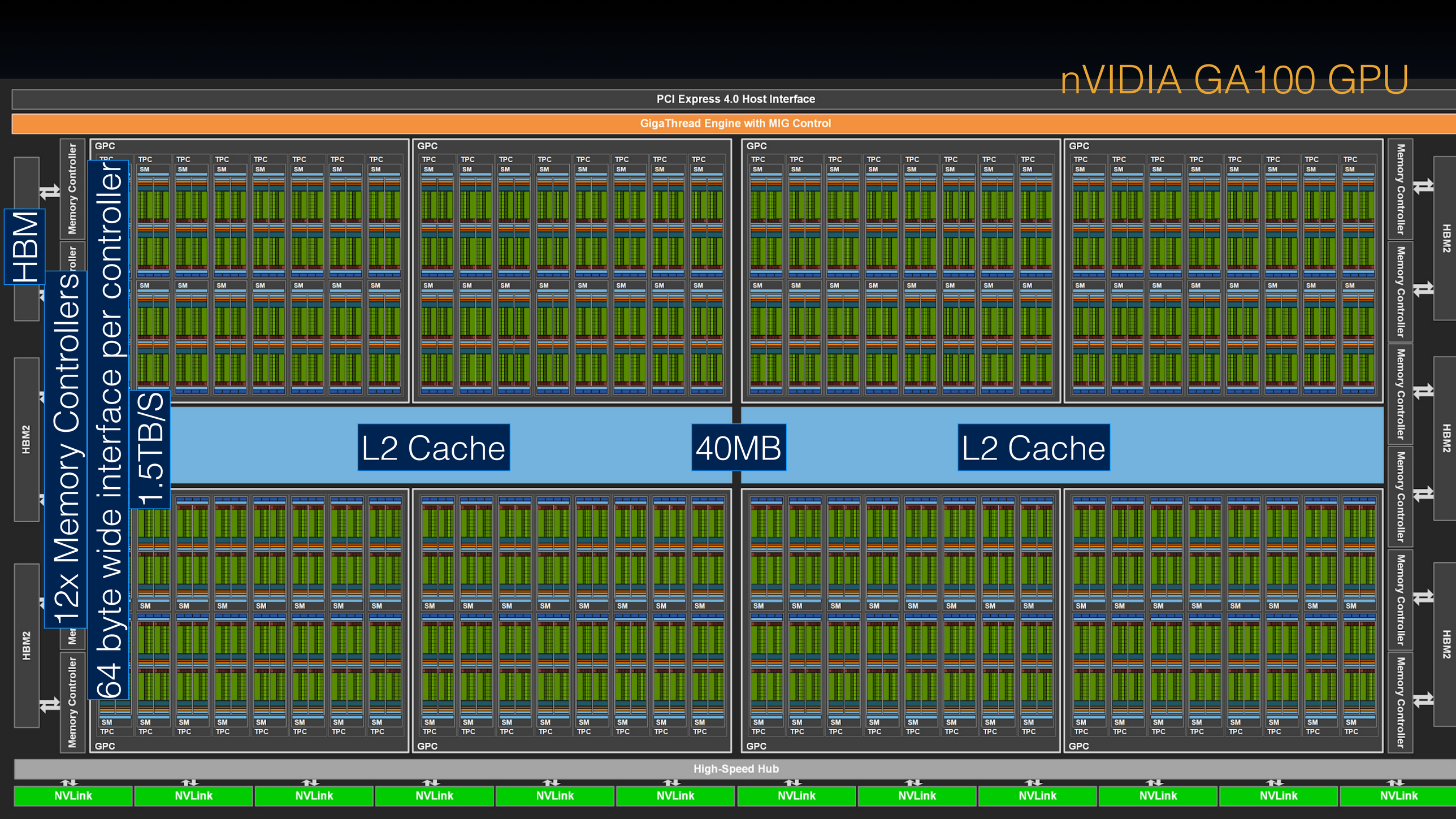


High Bandwidth Memory

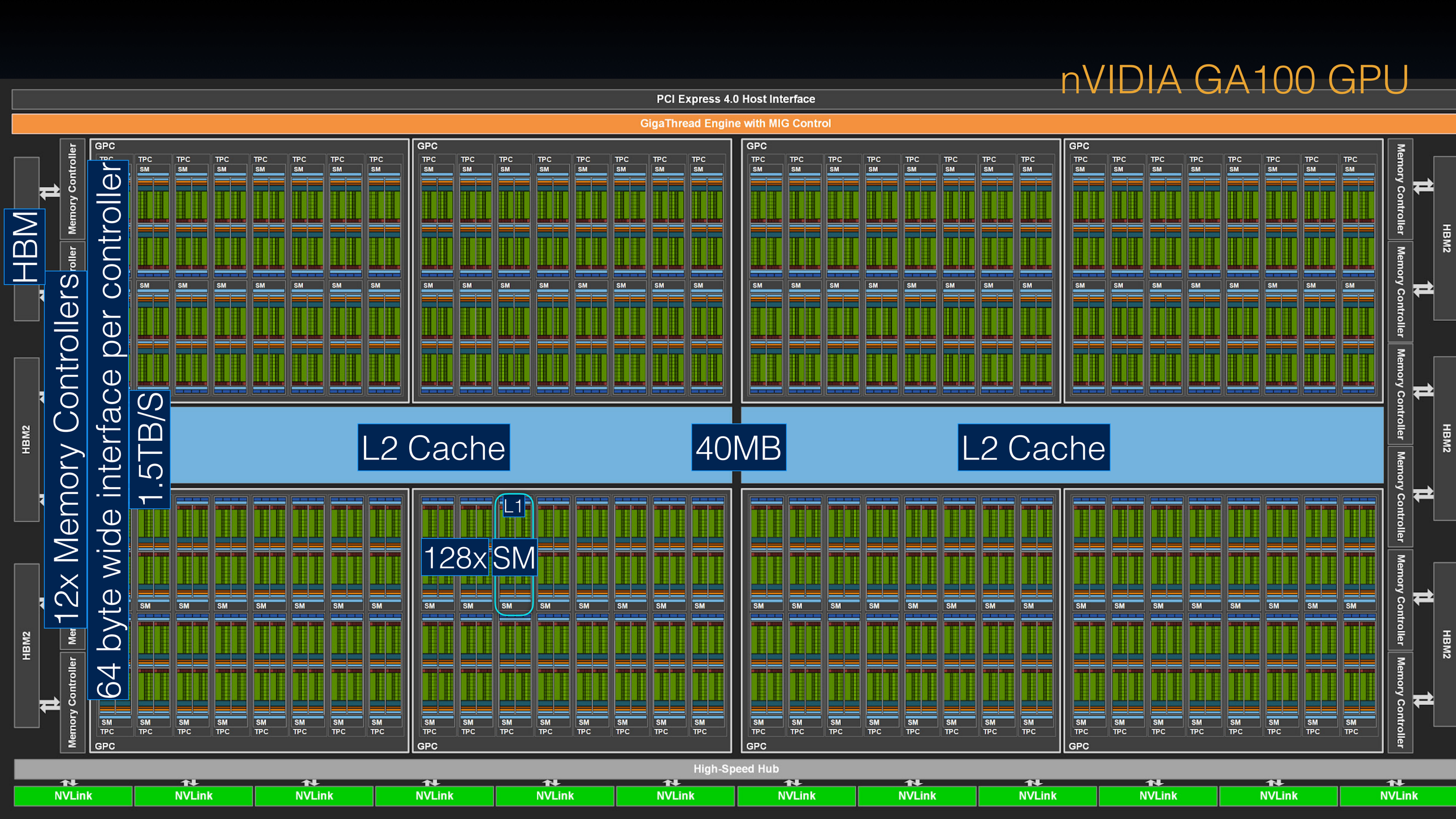


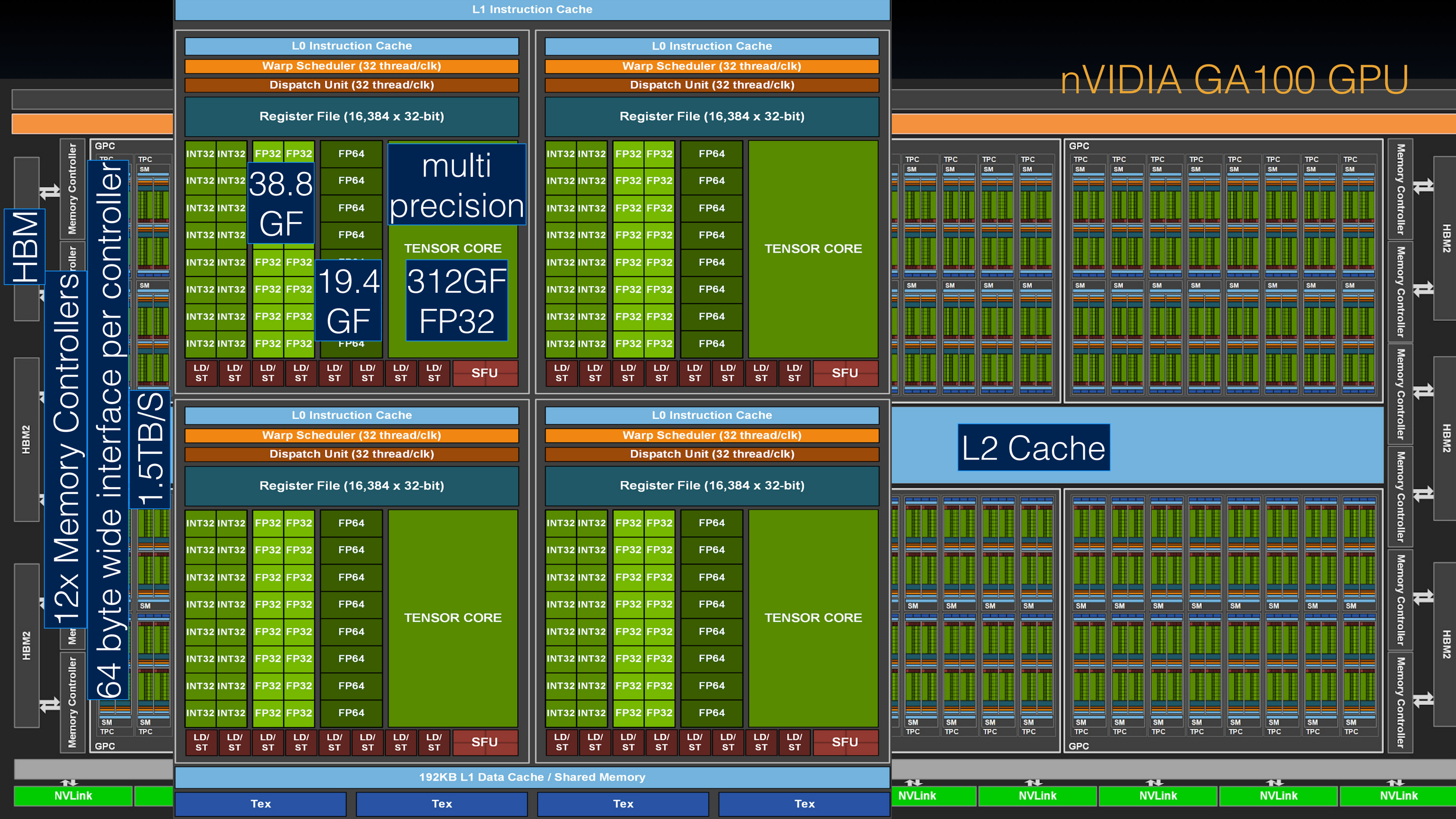
~600 GB/site

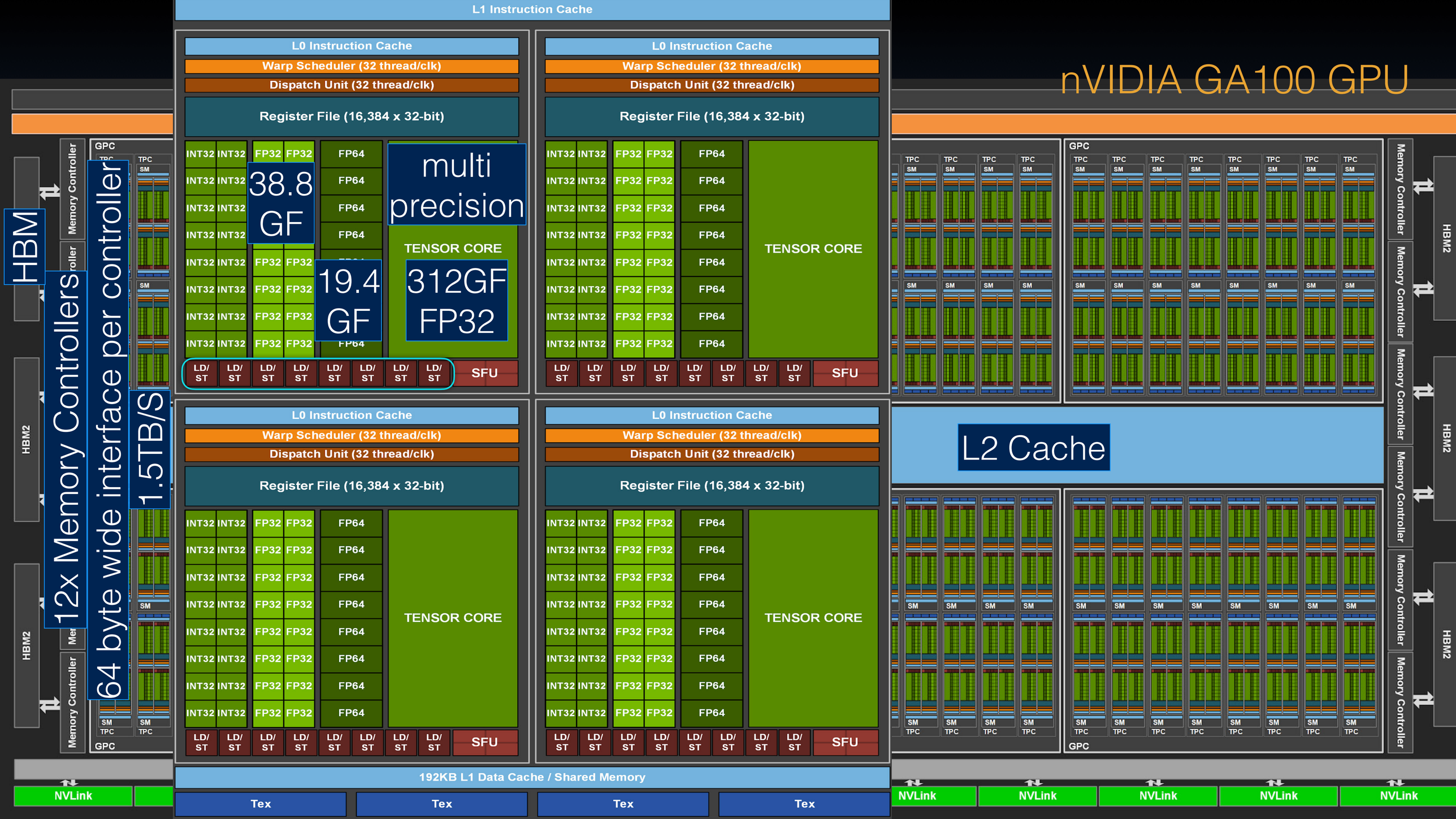
nVIDIA GA100 GPU



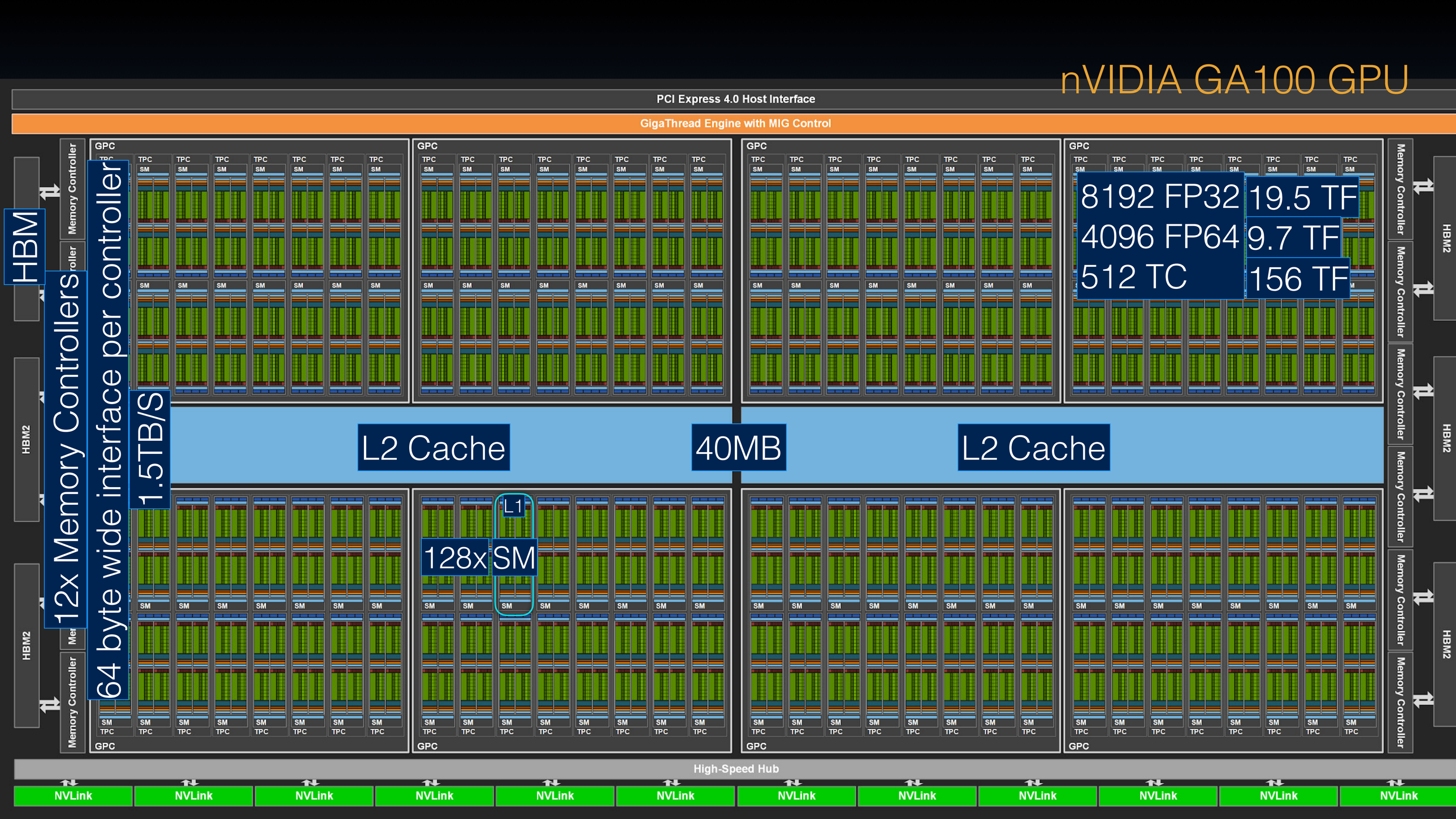
nVIDIA GA100 GPU



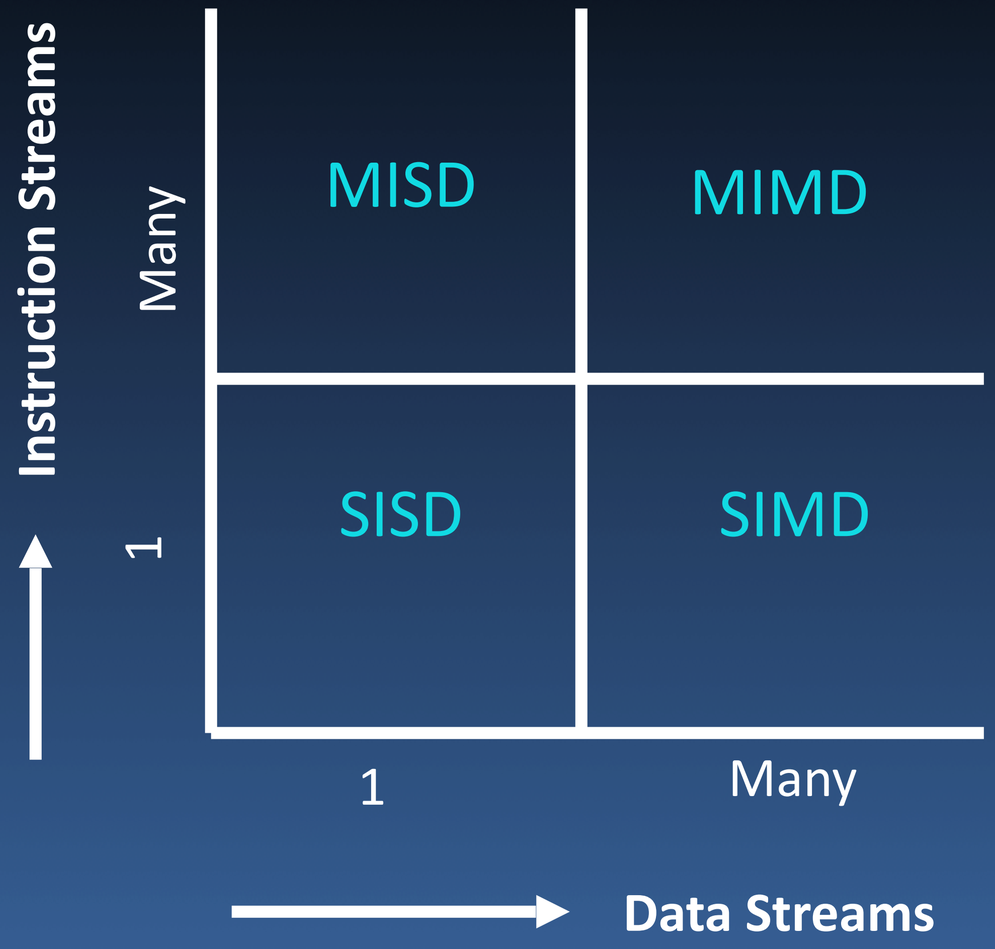

nVIDIA GA100 GPU
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
FP64
38.8 GF
19.4 GF
multi precision
TENSOR CORE
312GF FP32
LD/ST
SFU
192KB L1 Data Cache / Shared Memory
Tex
L2 Cache
GPC
TPC
SM
HBM
HBM2
Memory Controller
12x Memory Controllers
64 byte wide interface per controller
1.5TB/S
NVLink


nVIDIA GA100 GPU
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
FP64
38.8 GF
19.4 GF
multi precision
TENSOR CORE
312GF FP32
LD/ ST
SFU
192KB L1 Data Cache / Shared Memory
Tex
L2 Cache
GPC
TPC
SM
HBM
HBM2
Memory Controller
12x Memory Controllers
64 byte wide interface per controller
1.5TB/S
NVLink

nVIDIA GA100 GPU



Flynn's Classification



- A number of instruction threads
- A number of data of data threads

Review

- Parallel computers
- Memory bottlenecks
- Parallel computer Organization