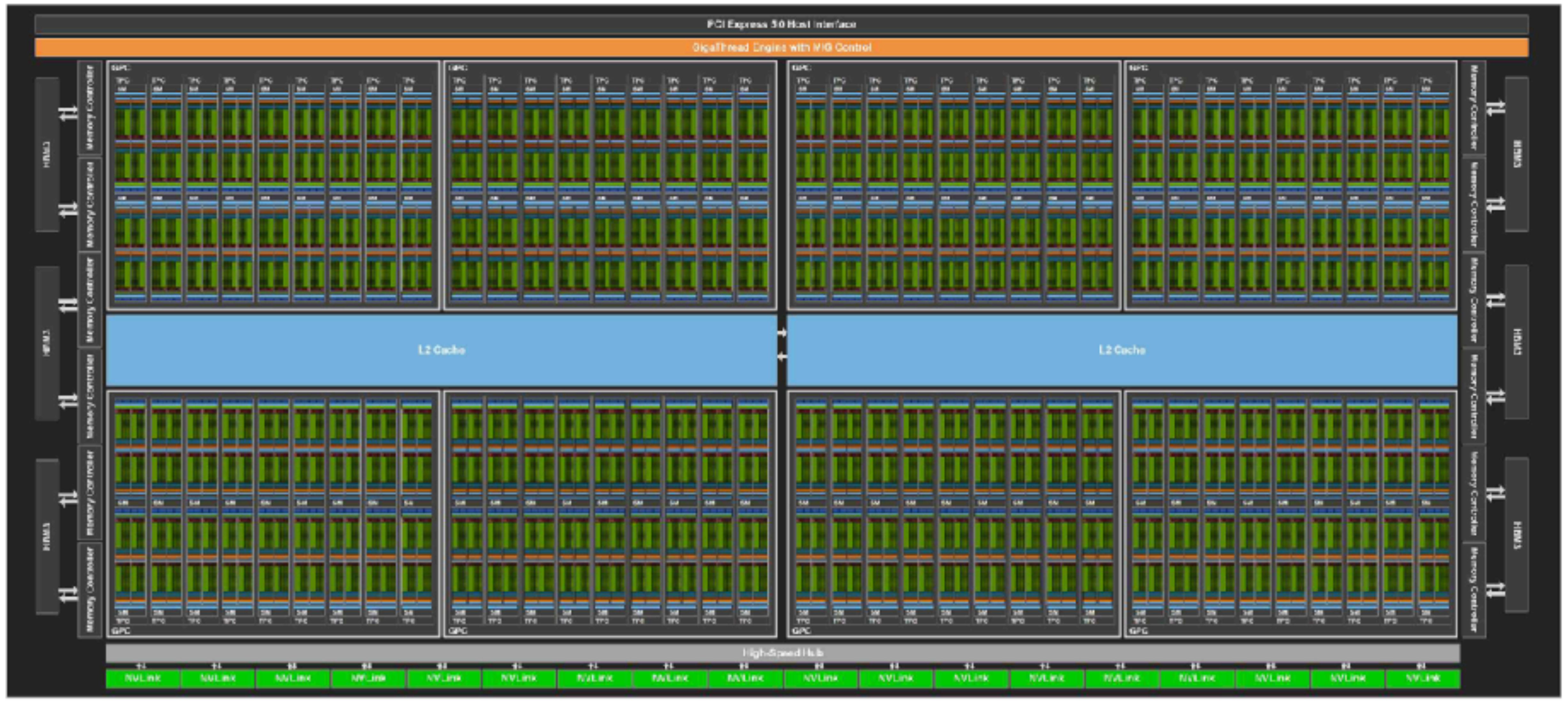
Modern GPU architectures

Philipp Krähenbühl, UT Austin

GPUs

- Massively parallel processors
- H100 SXM5
  - 132 Streaming Multiprocessors (SM) per GPU
  - 128 FP32 cores per SM
  - 80GB HBM3 ram
  - 228 KB shared memory / SM



GH100 Full GPU with 144 SMs [1]

[1] NVIDIA. NVIDIA H100 Tensor Core GPU Architecture. 2022.

GPUs - SM

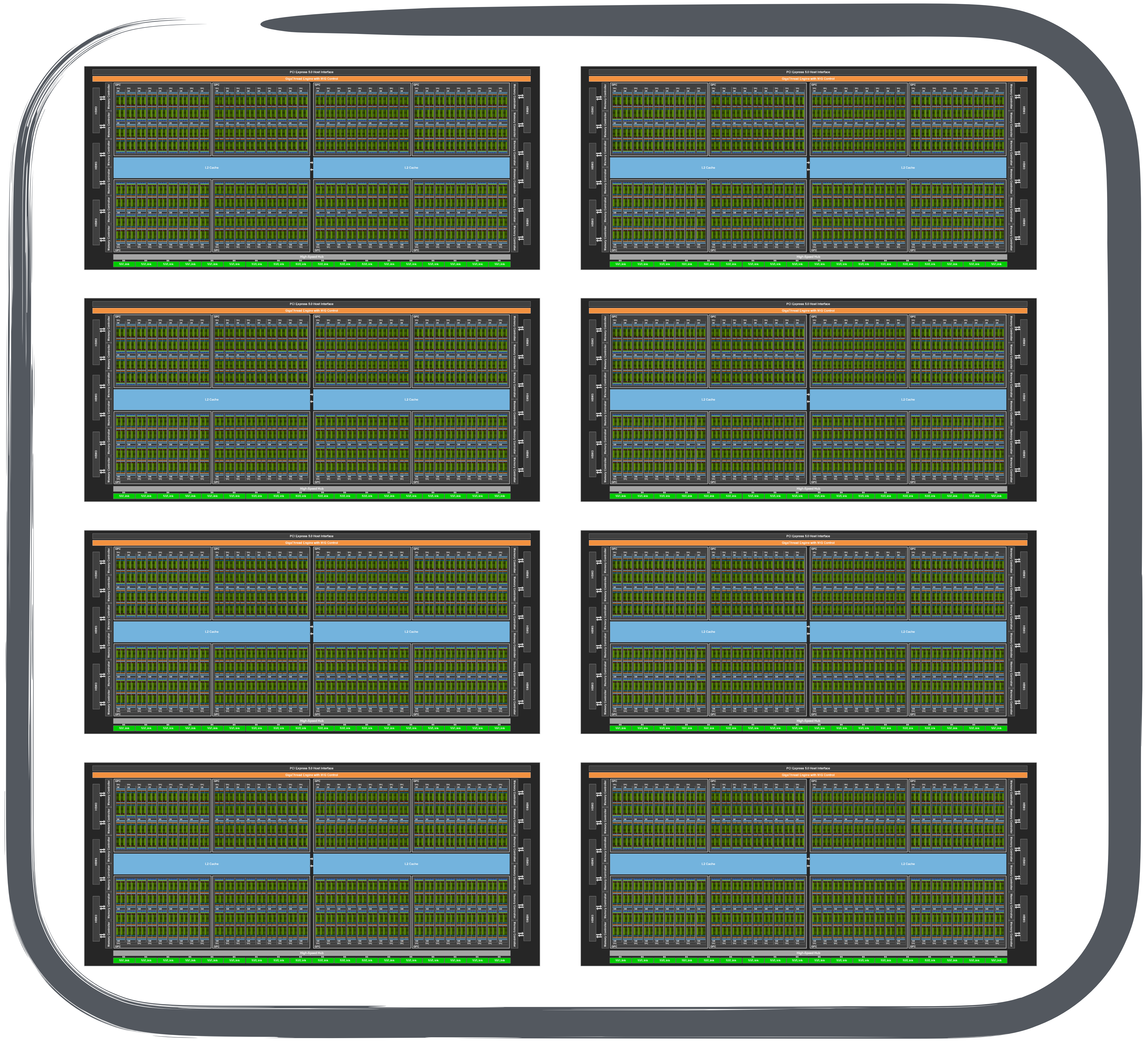
- Streaming Multiprocessors (SM)
  - Individual "CPUs" on chip
  - 4 warps (similar to CPU cores)
- Each warp
  - Tensor Core (matrix multiplier)
  - 32 threads (shared scheduler, dispatcher)

GH100 Streaming Multiprocessor (SM) [1]

[1] NVIDIA. NVIDIA H100 Tensor Core GPU Architecture. 2022.

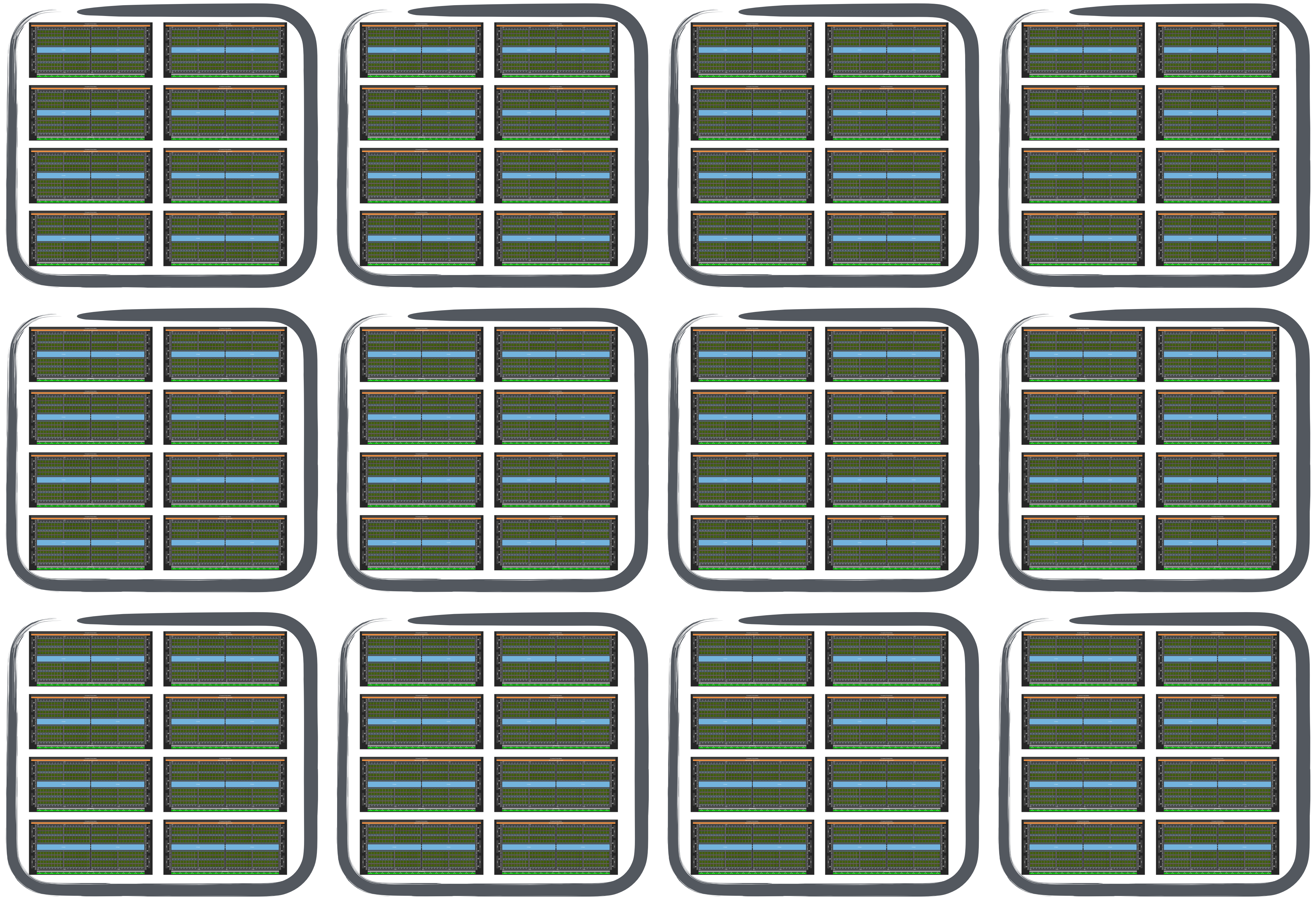
GPUs in a node

- Compute node
  - 8-16 GPUs per server / node
- Fast / specialized communication between GPUs (NVlink)



Node

GPUs in a datacenter

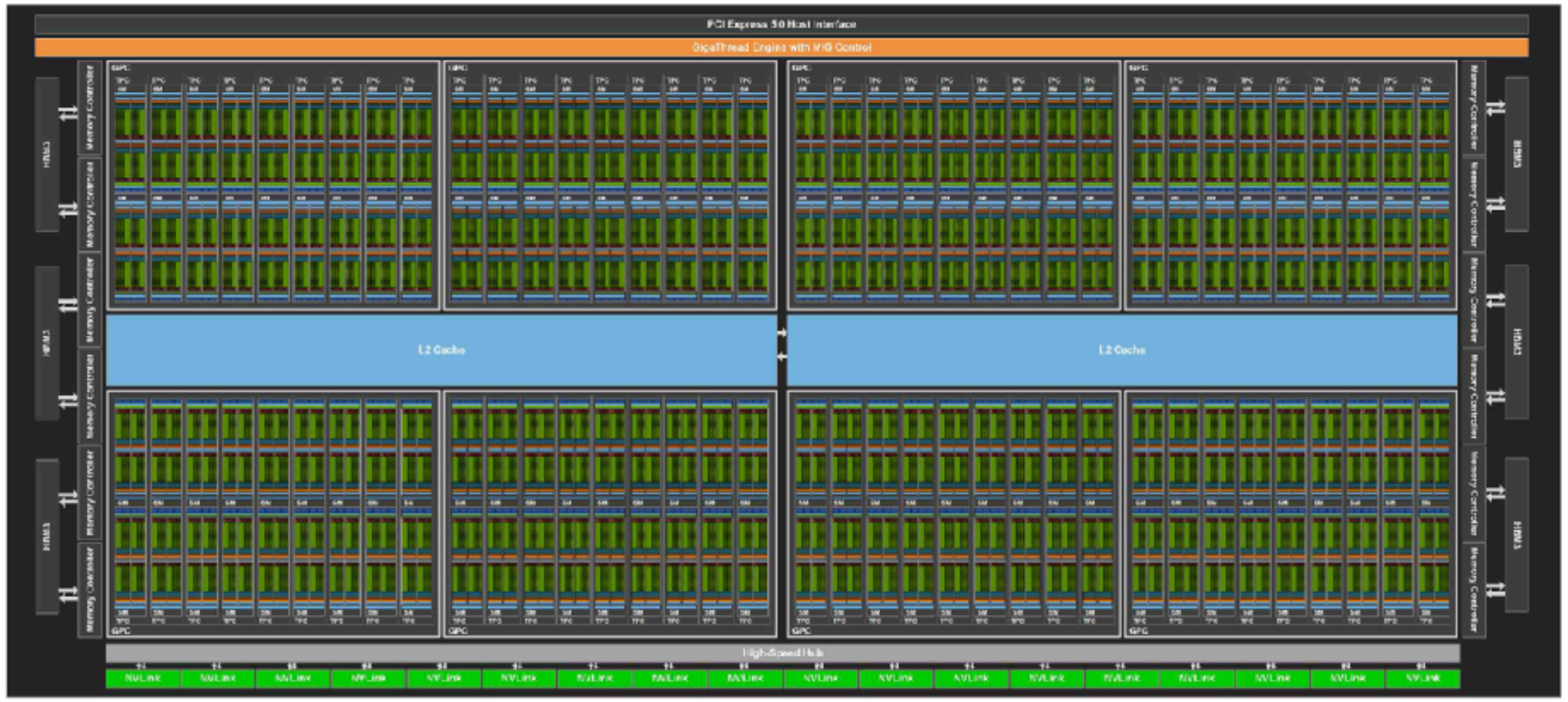


- Nodes networks in a datacenter
- Up to 40k nodes with 16 GPUs each
    - 0.42 GigaWatt
    - 40% of nuclear power plant, excluding cooling, other hardware
- We have peaked

[1] Meta. Building Meta's GenAI Infrastructure. 2024.
[2] https://en.wikipedia.org/wiki/Nuclear_power.

GPUs - Mental model

- Massively parallel processors
  - Intuitions from CPUs and theoretical CS are often wrong
  - Nearly endless compute
    - On a restricted set of operations
  - Limited memory and memory bandwidth



GH100 Full GPU with 144 SMs [1]

[1] NVIDIA. NVIDIA H100 Tensor Core GPU Architecture. 2022.

GPUs - Mental model

A simple example

x_1	x_2	x_3	x_4	x_5	x_6

- You are given a series of numbers $\mathbf{x}$ and a **fixed** window size W
- Find the maximum number value for all possible windows
  - $e_i = \max(x_i, x_{i+1}, \ldots, x_{i+W-1})$
- What deep learning operation is this?

GPUs - Mental model

A simple example

x_1	x_2	x_3	x_4	x_5	x_6

- You are given a series of numbers $\mathbf{x}$ and a **fixed** window size W

- Find the maximum number value for all possible windows

  - $e_i = \max(x_i, x_{i+1}, \ldots, x_{i+W-1})$

- What deep learning operation is this?

```python
def maxpool_1d_brute(x: torch.Tensor, window_size: int):
    """A windowed maximum pooling operation for 1D tensors."""
    output = x.new_zeros(x.size(0) - window_size + 1)
    for i in range(output.size(0)):
        for j in range(window_size):
            output[i] = max(output[i], x[i + j])
    return output
```

Compute: $O(|\mathbf{x}|W)$

GPUs - Mental model

A simple example

x_1	x_2	x_3	x_4	x_5	x_6

- You are given a series of numbers $\mathbf{x}$ and a **fixed** window size W
- Find the maximum number value for all possible windows
  - $e_i = \max(x_i, x_{i+1}, \ldots, x_{i+W-1})$
- What deep learning operation is this?

```python
def maxpool_1d_heap(x: torch.Tensor, window_size: int):
    """A windowed maximum pooling operation for 1D
tensors."""
    output = x.new_zeros(x.size(0) - window_size + 1)

    h = []
    for i in range(x.size(0)):
        heapq.heappush(h, (-x[i].item(), i))
        if i >= window_size - 1:
            while h[0][1] <= i - window_size:
                heapq.heappop(h)
            output[i - window_size + 1] = -h[0][0]
    return output
```

Compute: $O(|\mathbf{x}| \log W)$

GPUs - Mental model

A simple example in CUDA

x_1	x_2	x_3	x_4	x_5	x_6

- You are given a series of numbers $\mathbf{x}$ and a **fixed** window size W
- Find the maximum number value for all possible windows
  - $e_i = \max(x_i, x_{i+1}, \ldots, x_{i+W-1})$
- What deep learning operation is this?

Compute: $O(|\mathbf{x}|W)$

Memory access: $O(|\mathbf{x}|W)$

GPUs - Mental model

A simple example in CUDA

x_1	x_2	x_3	x_4	x_5	x_6

- You are given a series of numbers $\mathbf{x}$ and a **fixed** window size W
- Find the maximum number value for all possible windows
  - $e_i = \max(x_i, x_{i+1}, \ldots, x_{i+W-1})$
- What deep learning operation is this?

Compute: $O(|\mathbf{x}|W/G)$

Memory access: $O(|\mathbf{x}|W/G)$

GPUs - Mental model

A simple example in CUDA

x_1	x_2	x_3	x_4	x_5	x_6

- You are given a series of numbers $\mathbf{x}$ and a **fixed** window size W
- Find the maximum number value for all possible windows
  - $e_i = \max(x_i, x_{i+1}, \ldots, x_{i+W-1})$
- What deep learning operation is this?

Compute: $O(|\mathbf{x}|W)$

Memory access: $O\left(|\mathbf{x}|\frac{W}{S}\right)$

S: shared memory size

GPUs - Mental model

What have we learned?

x_1	x_2	x_3	x_4	x_5	x_6

- Memory access matters
  - Reads from global memory are expensive
  - Computation is cheap

GPUs - Mental model

The secret solution

x_1	x_2	x_3	x_4	x_5	x_6

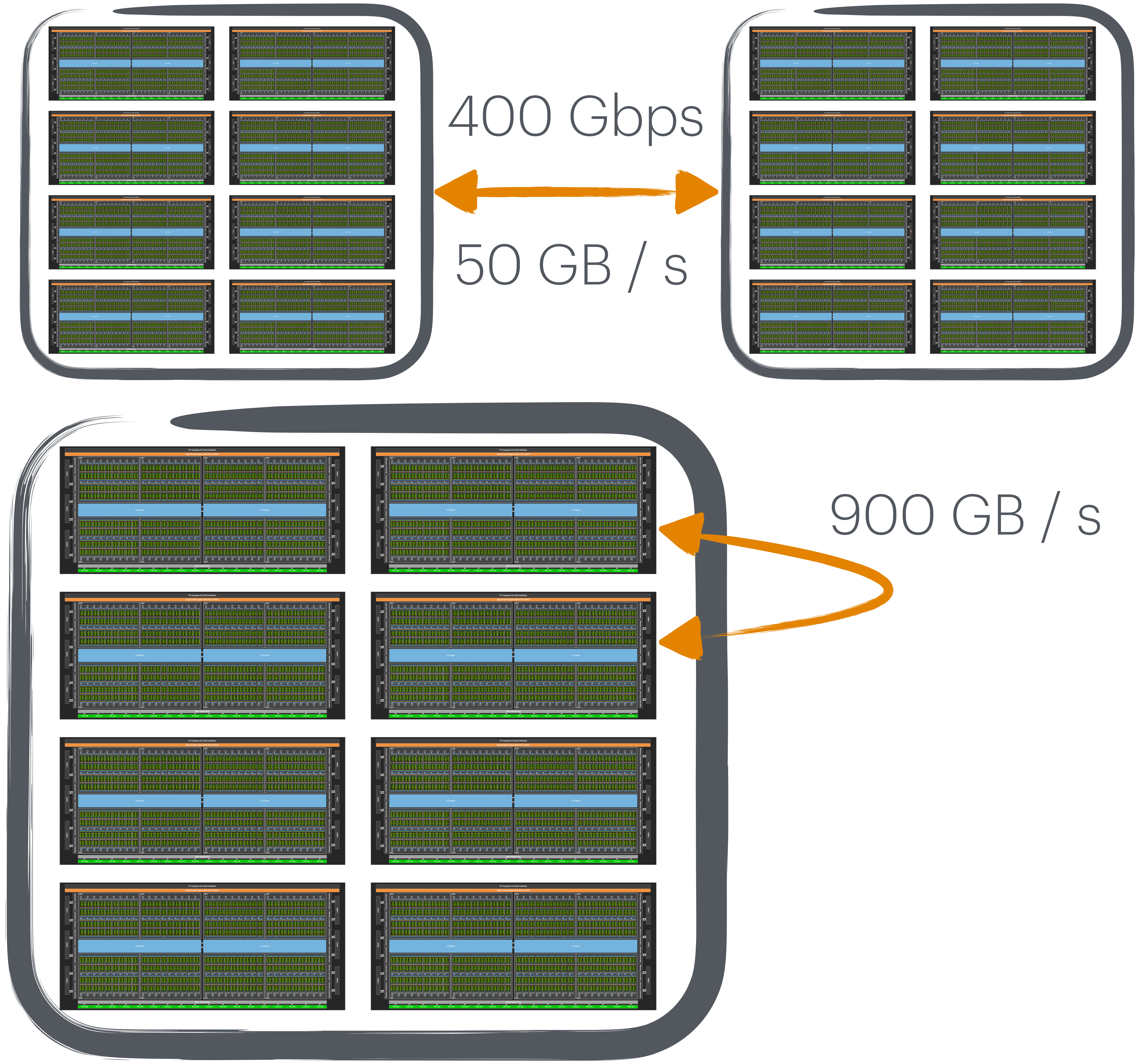
- $e_i = \max(x_i, x_{i+1}, \ldots, x_{i+W-1})$
- $e_i = \max\left(\max(x_i, \ldots, x_K), \max(x_{K+1}, \ldots, x_{i+W-1})\right)$
- $e_{i+1} = \max\left(\max(x_{i+1}, \ldots, x_K), \max(x_{K+1}, \ldots, x_{i+W})\right)$

Compute: $O(|\mathbf{x}|)$

Memory access: $O(|\mathbf{x}|)$

GPUs - Memory Bandwidth

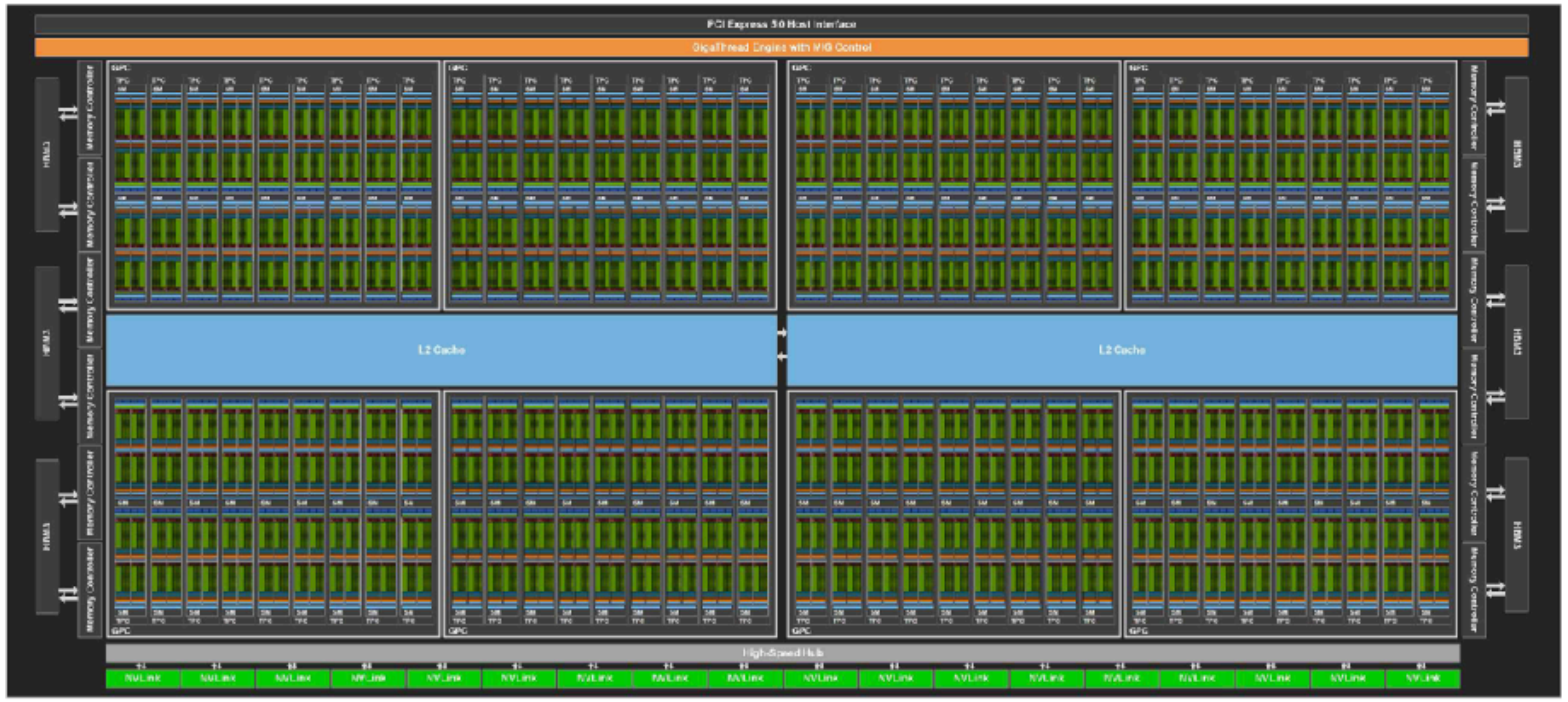
- Node to node communication
  - RDMA/IB: 50GB / s
- GPU to GPU communication (within node)
  - NVLink: 900 GB / s
- GPU memory bandwidth
  - HBM3->shared mem: 3.35 TB / s
- Peak flops: 130-1000 teraFLOPS @ BF16



[1] NVIDIA. NVIDIA H100 Tensor Core GPU Architecture. 2022.

Modern GPU architectures

- Near infinite compute
- Memory bandwidth and size limits
  - Order of magnitude slower
    GPU -> Node -> Datacenter
- Approaching limits of power consumption, and physical limits in manufacture



GH100 Full GPU with 144 SMs [1]

[1] NVIDIA. NVIDIA H100 Tensor Core GPU Architecture. 2022.

References

- [1] NVIDIA. NVIDIA H100 Tensor Core GPU Architecture. 2022. (link)
- [2] Meta. Building Meta's GenAI Infrastructure. 2024 (link)