ESCAPE 824064

Laboratoire d'Annecy de Physique des Particules

GPU Architecture

Pierre Aubert

UNIVERSITÉ SAVOIE MONT BLANC

Graphic Processing Unit

Graphic Processing Unit

1980s

Graphic Processing Unit

1980s

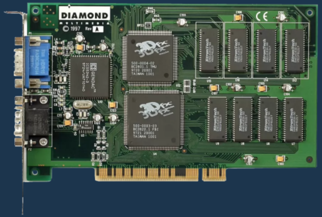
1993

Graphic Processing Unit

1980s

1993

1995

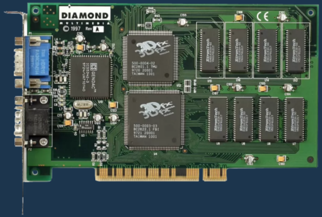


Graphic Processing Unit

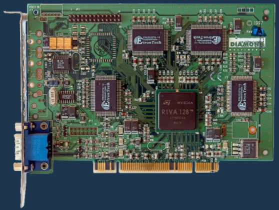
1980s

1993

1995



1997

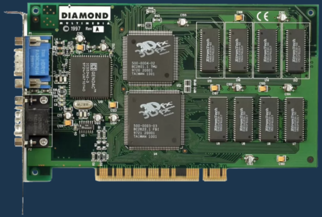


Graphic Processing Unit

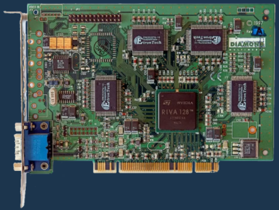
1980s

1993

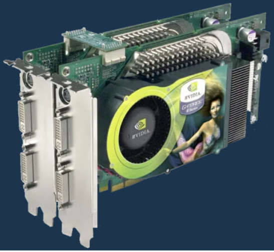
1995



1997



2005

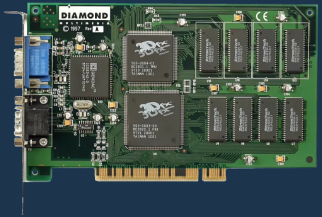


Graphic Processing Unit

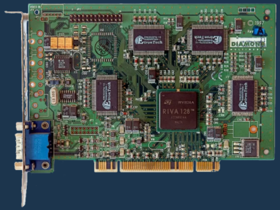
1980s

1993

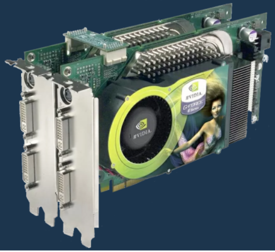
1995



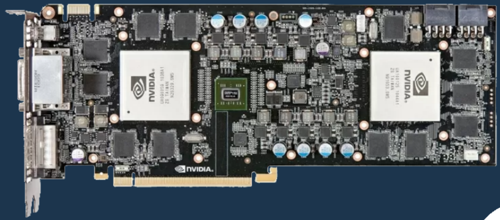
1997



2005



2011



Graphic Processing Unit

2015

Graphic Processing Unit

2015

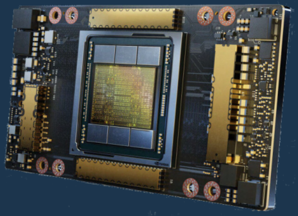
2018

Graphic Processing Unit

2015

2018

2020

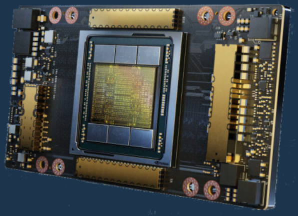


Graphic Processing Unit

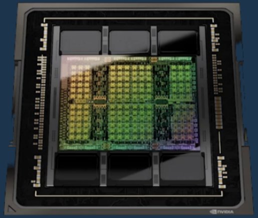
2015

2018

2020



2022-04

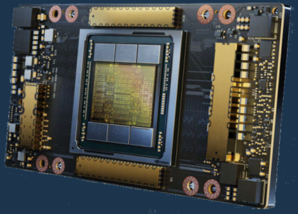


Graphic Processing Unit

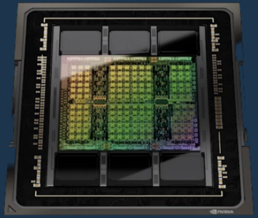
2015

2018

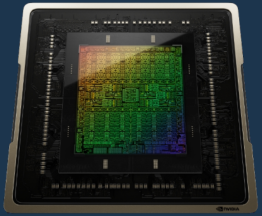
2020



2022-04



2022-09



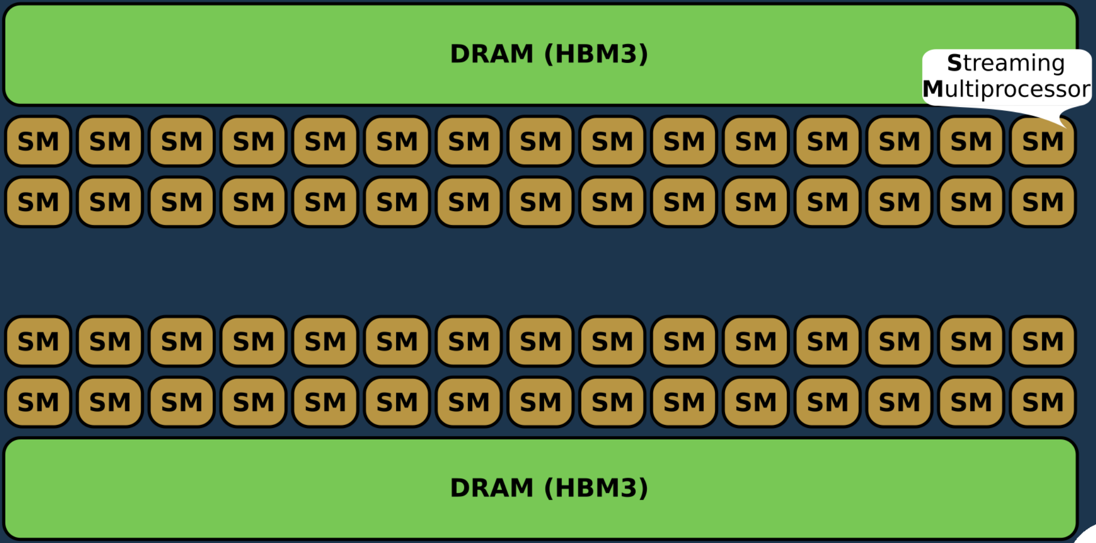
GPU Architecture

GPU Architecture

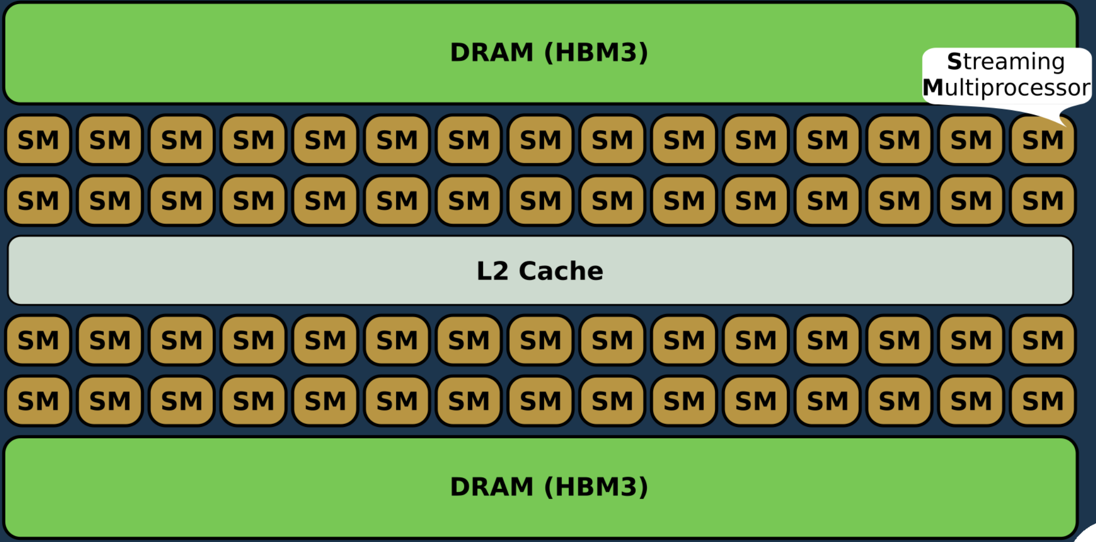
DRAM (HBM3)

DRAM (HBM3)

GPU Architecture



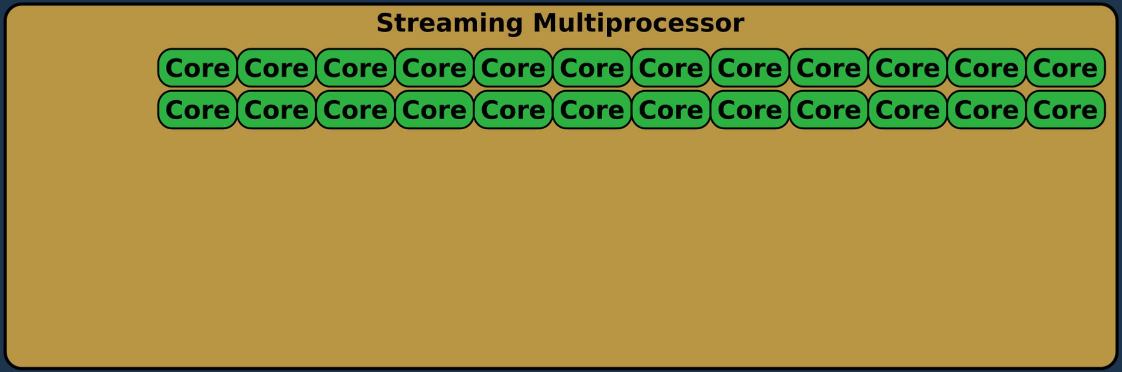
GPU Architecture



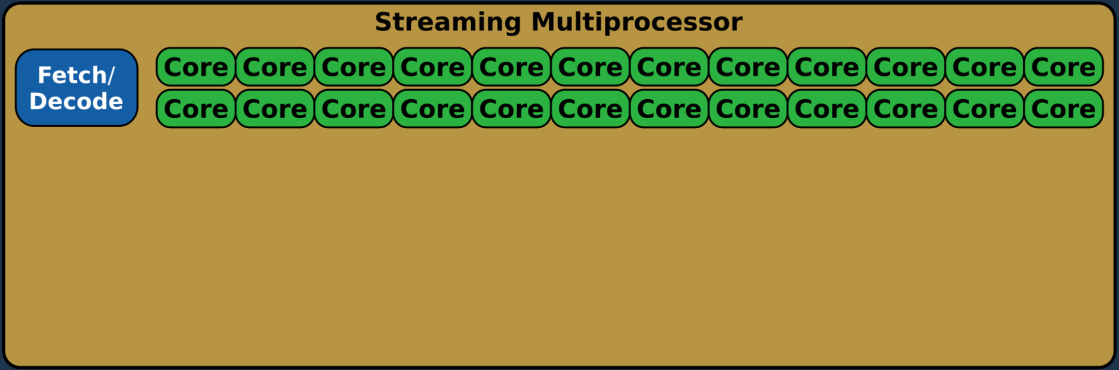
Streaming Multiprocessor

Streaming Multiprocessor

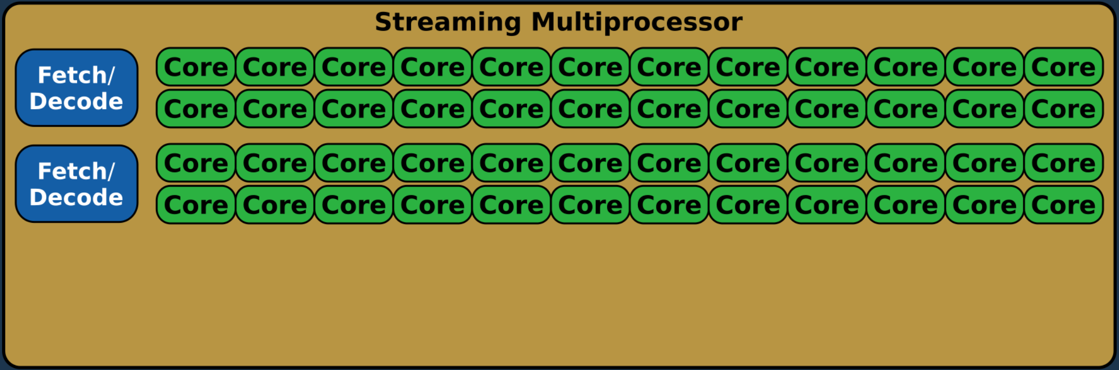
Streaming Multiprocessor



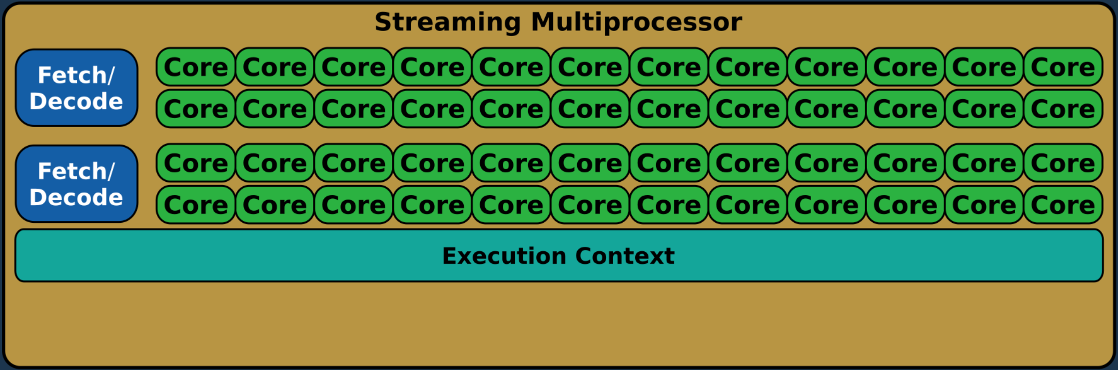
Streaming Multiprocessor



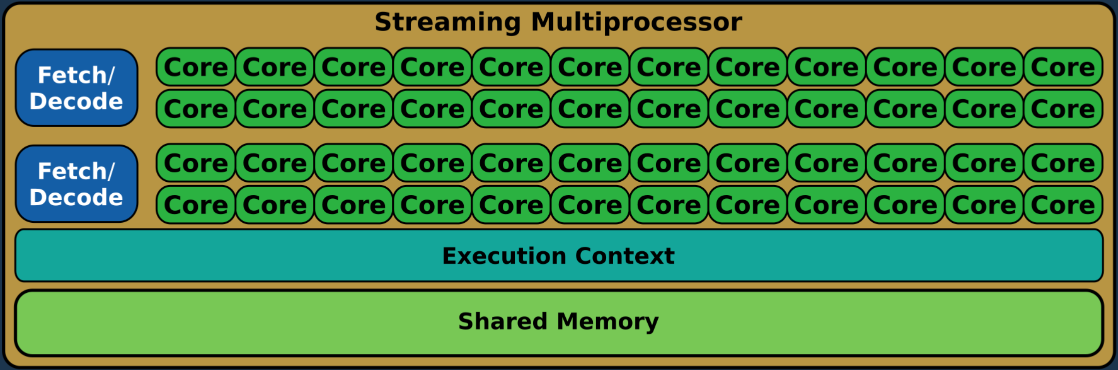
Streaming Multiprocessor



Streaming Multiprocessor

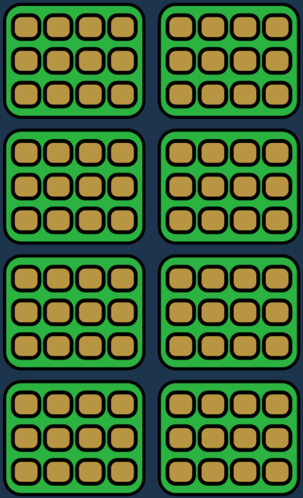


Streaming Multiprocessor

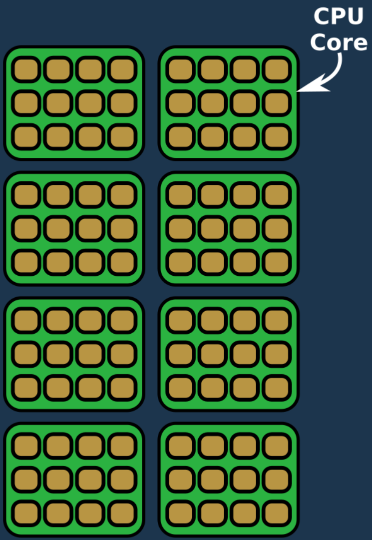


Threads ?

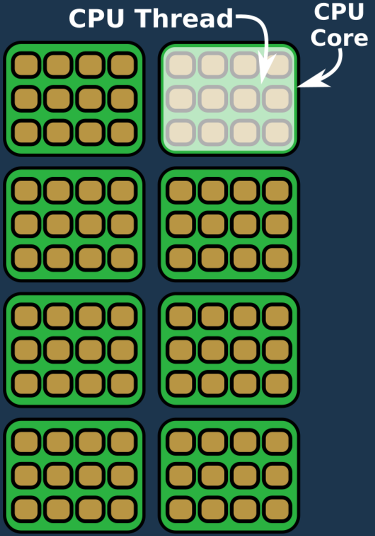
Threads ?



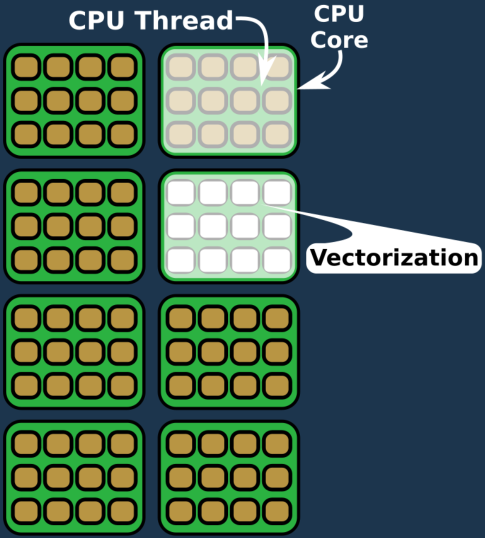
Threads ?



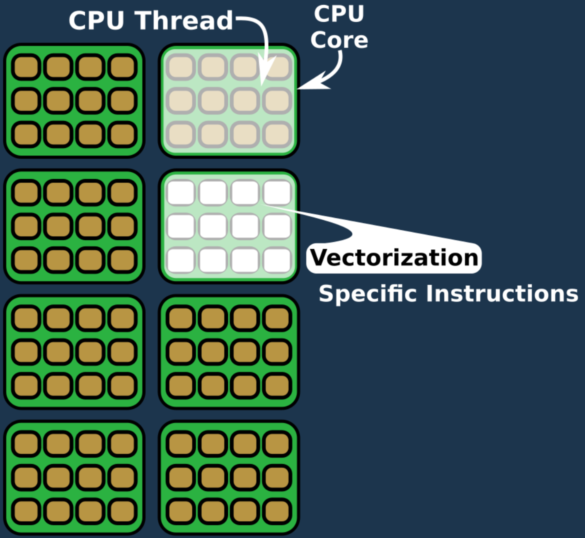
Threads ?



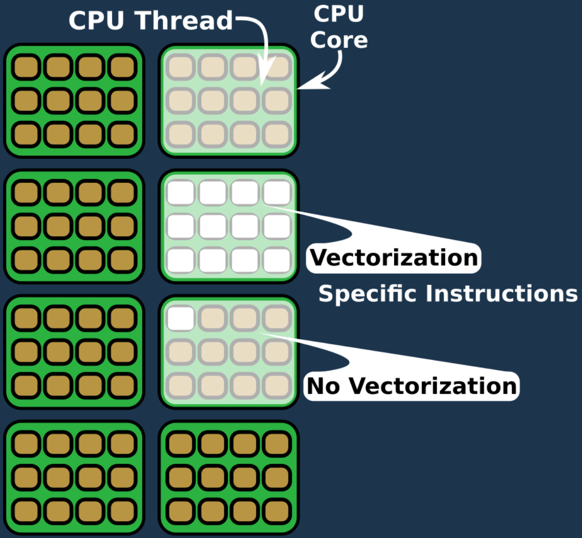
Threads ?



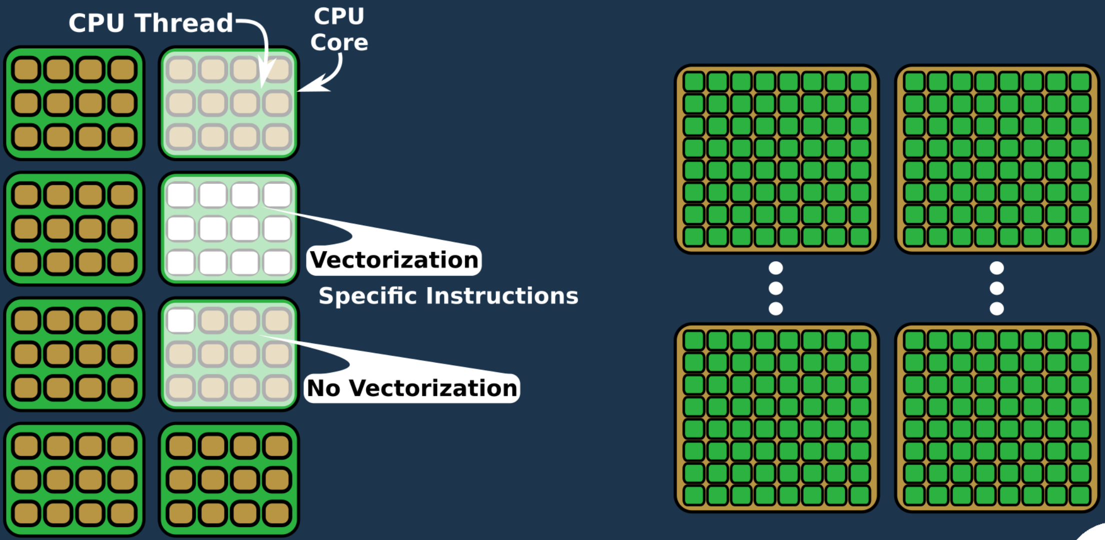
Threads ?



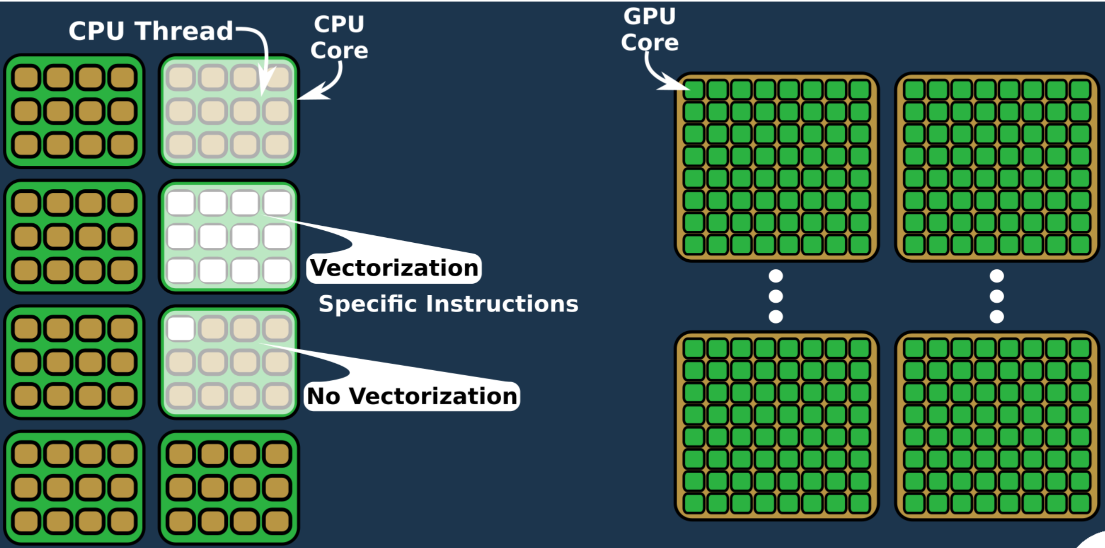
Threads ?


CPU Thread
CPU Core
Vectorization
Specific Instructions
No Vectorization

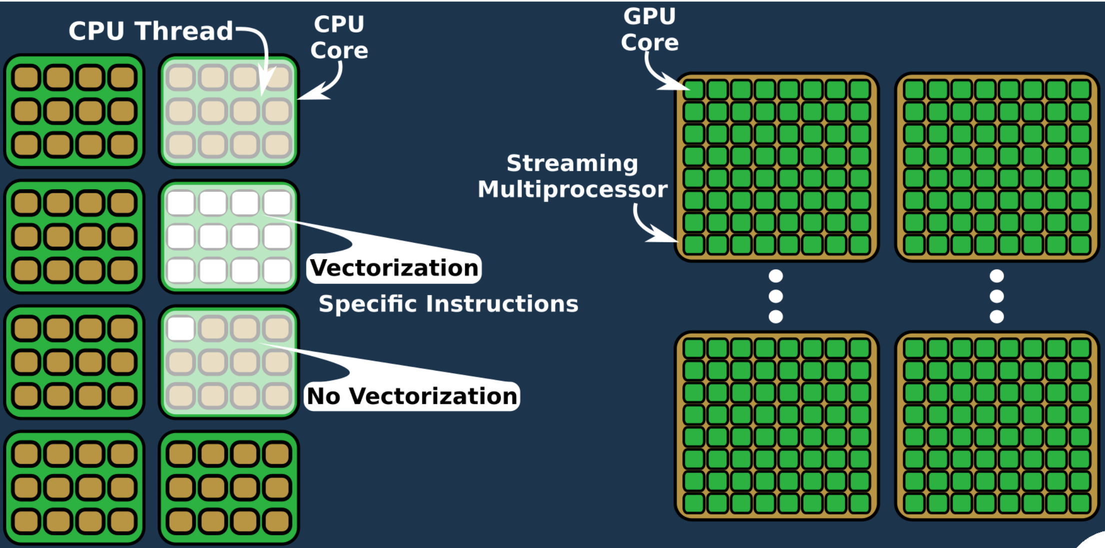
Threads ?


CPU Thread
CPU
Core
Vectorization
Specific Instructions
No Vectorization

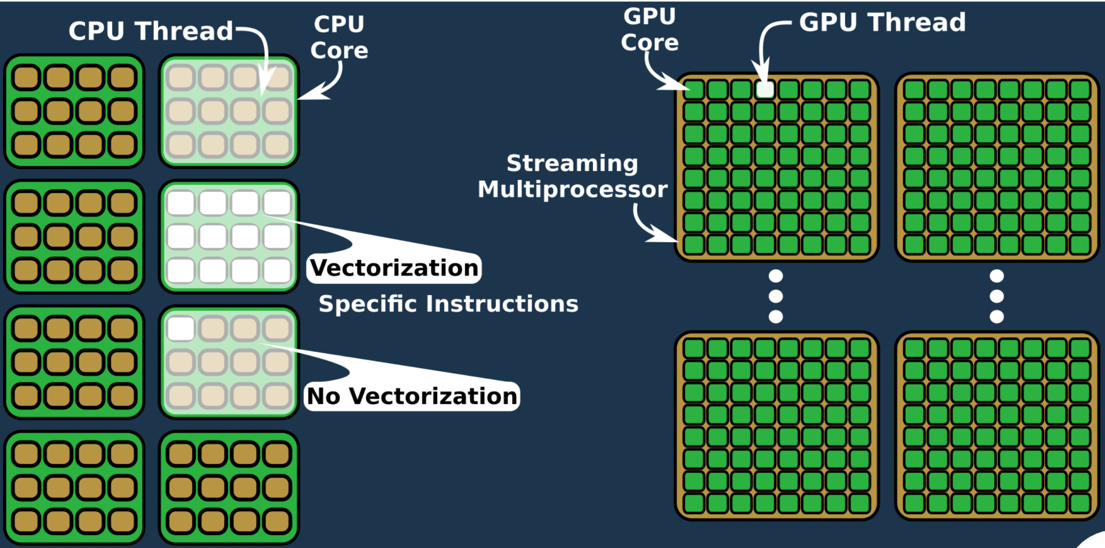
Threads ?


CPU Thread
CPU Core
Vectorization
Specific Instructions
No Vectorization
GPU Core

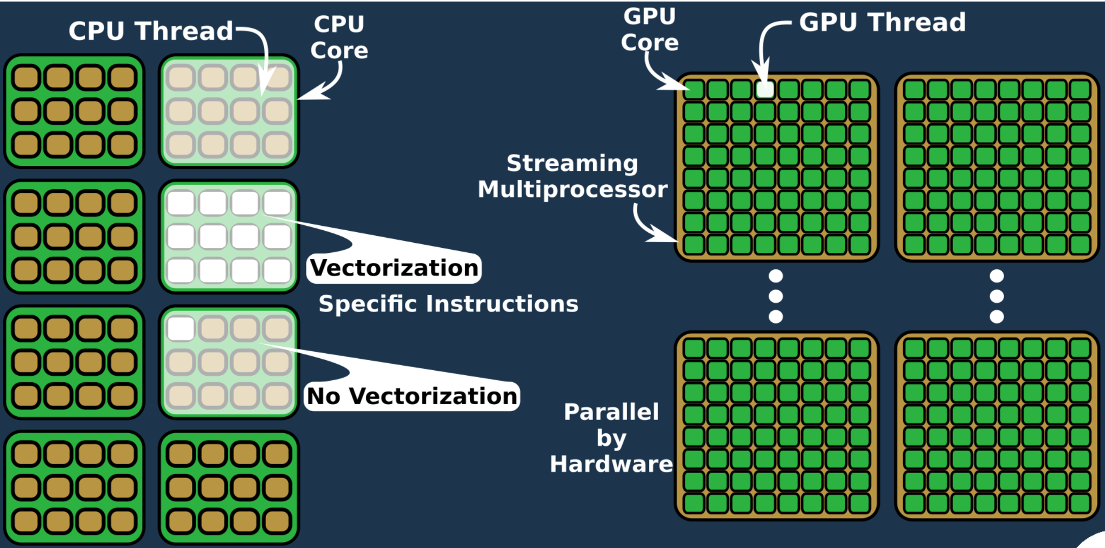
Threads ?


CPU Thread
CPU Core
Vectorization
Specific Instructions
No Vectorization
GPU Core
Streaming Multiprocessor

Threads ?


CPU Thread
CPU Core
Vectorization
Specific Instructions
No Vectorization
GPU Core
GPU Thread
Streaming Multiprocessor

Threads ?


CPU Thread
CPU Core
Vectorization
Specific Instructions
No Vectorization
GPU Core
GPU Thread
Streaming Multiprocessor
Parallel by Hardware

Threads Hierarchy

Threads Hierarchy

Thread

Threads Hierarchy

Thread Core

Threads Hierarchy

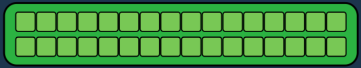
Thread	Core	Executes Computing Kernel

Threads Hierarchy

Warp

Thread Core Executes Computing Kernel

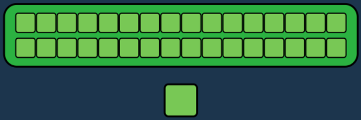
Threads Hierarchy



Warp Hardware Scheduler

Thread Core Executes Computing Kernel

Threads Hierarchy

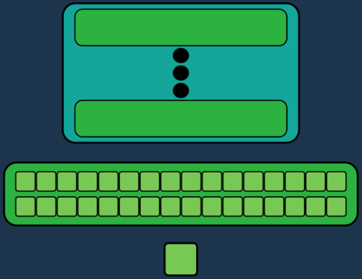


Warp	Hardware Scheduler	32 **Threads**
Thread	Core	Executes Computing Kernel

Threads Hierarchy

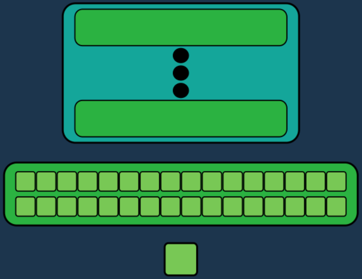
	**Warp**	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
	**Thread**	Core	Executes Computing Kernel	

Threads Hierarchy



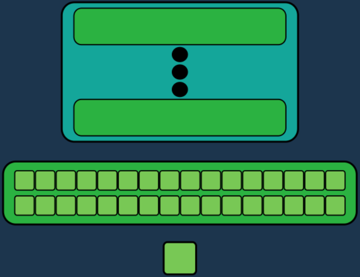
Block			
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Hierarchy



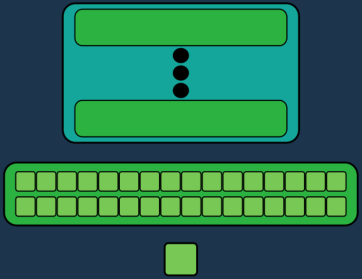
Block	Streaming Multiprocessor		
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Hierarchy



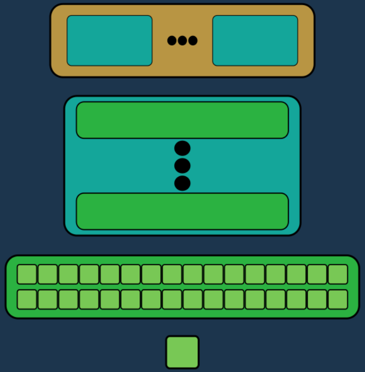
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Hierarchy



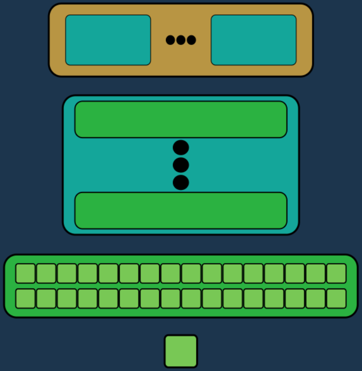
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	**Asynchronous** Same Instruction Multiple Data
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Hierarchy



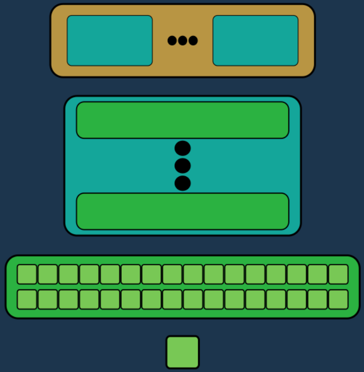
Cluster			
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	**Asynchronous** Same Instruction Multiple Data
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Hierarchy



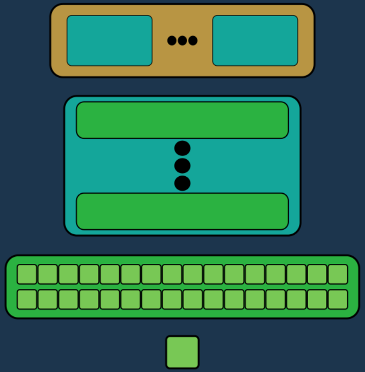
Cluster	Blocks Communication		
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	**Asynchronous** Same Instruction Multiple Data
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Hierarchy



Cluster	Blocks Communication	Since : - CUDA **12** - **Hopper**	
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	**Asynchronous** Same Instruction Multiple Data
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

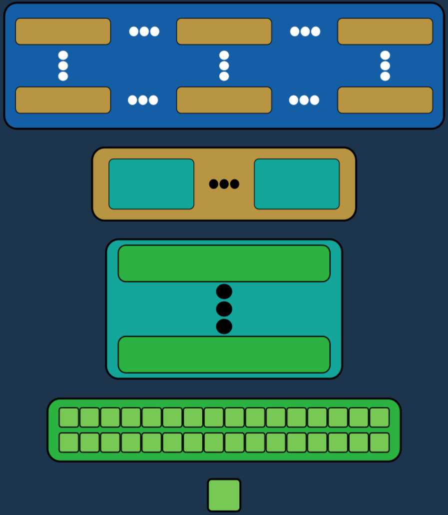
Threads Hierarchy



Cluster	Blocks Communication	Since : - CUDA **12** - **Hopper**	**Hopper** : 132 SMs **Ampere** : 108 SMs
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	**Asynchronous** Same Instruction Multiple Data
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

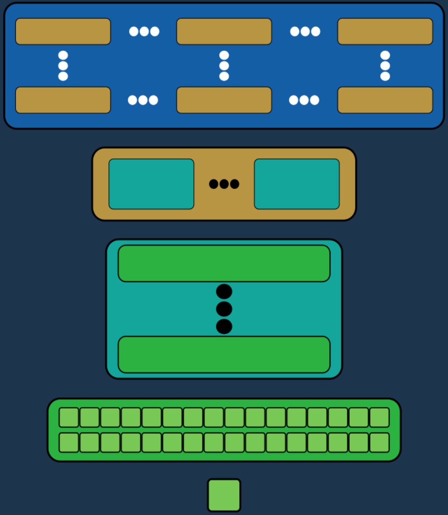
Laboratoire d'Annecy de Physique des Particules

Threads Hierarchy



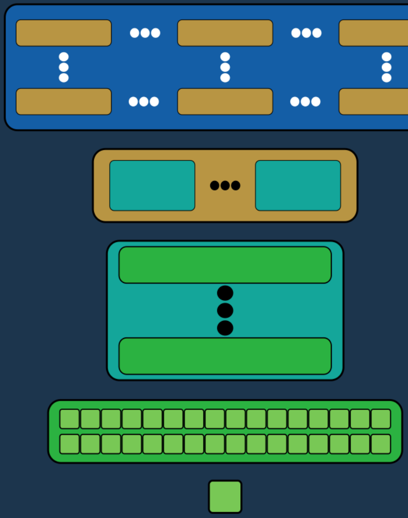
Grid			
Cluster	Blocks Communication	Since : - CUDA **12** - **Hopper**	**Hopper** : 132 SMs **Ampere** : 108 SMs
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	**Asynchronous** Same Instruction Multiple Data
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Hierarchy



Grid	Full **GPU**		
Cluster	Blocks Communication	Since : - CUDA **12** - **Hopper**	**Hopper** : 132 SMs **Ampere** : 108 SMs
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	**Asynchronous** Same Instruction Multiple Data
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Hierarchy



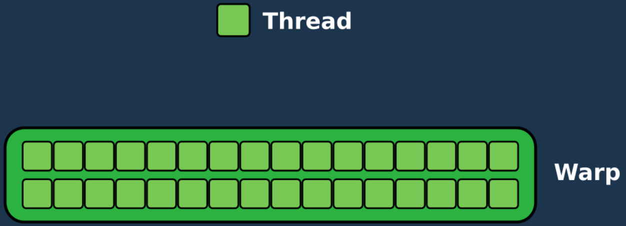
Grid	Full **GPU**	**Hopper** : 8448 Cores **Ampere** : 6912 Cores	
Cluster	Blocks Communication	Since : - CUDA **12** - **Hopper**	**Hopper** : 132 SMs **Ampere** : 108 SMs
Block	Streaming Multiprocessor	2 - 4 **Warps** 64 - 128 **Cores**	**Asynchronous** Same Instruction Multiple Data
Warp	Hardware Scheduler	32 **Threads**	**Synchronous** Same Instruction Multiple Data
Thread	Core	Executes Computing Kernel	

Threads Execution

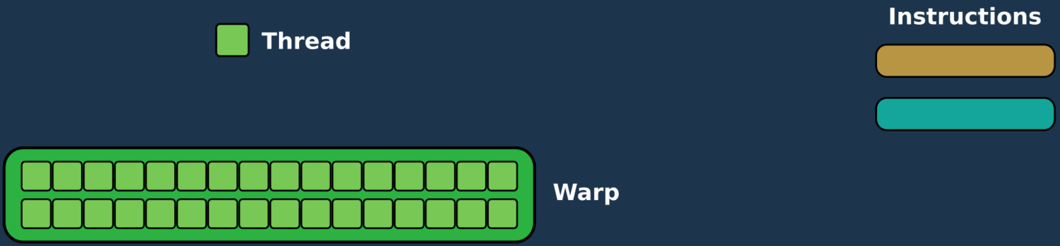
Threads Execution

Thread

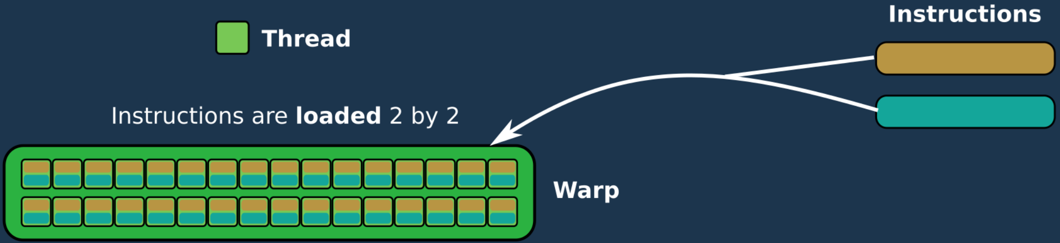
Threads Execution



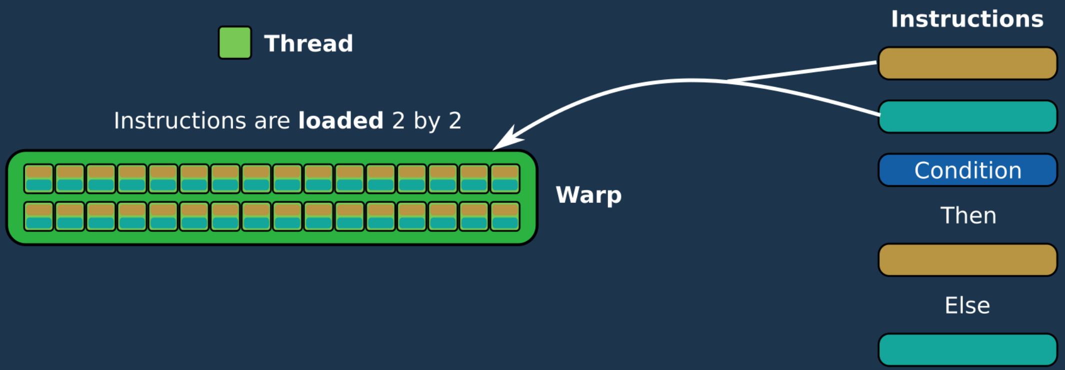
Threads Execution



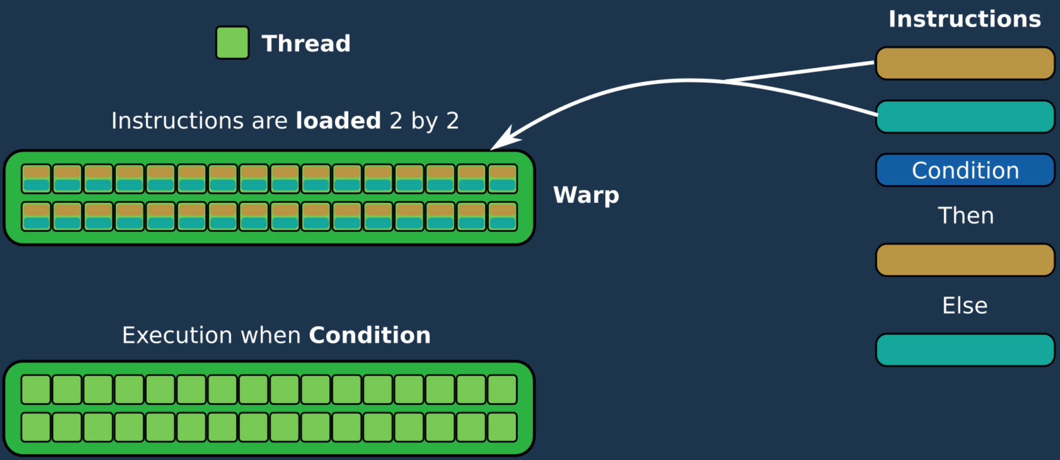
Threads Execution



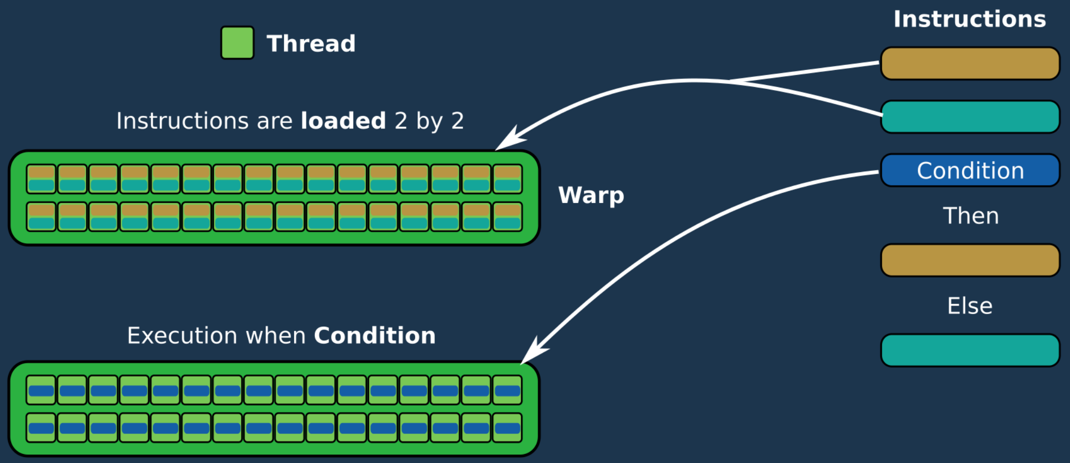
Threads Execution



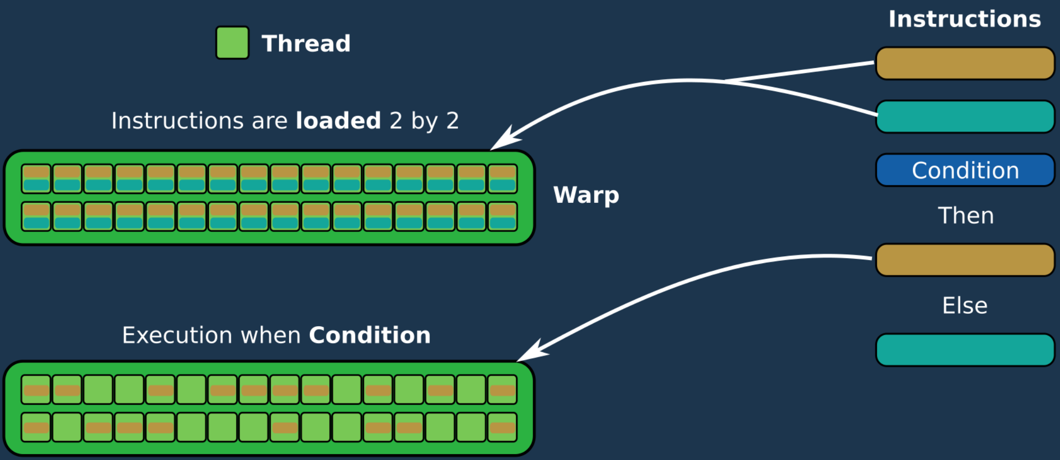
Threads Execution



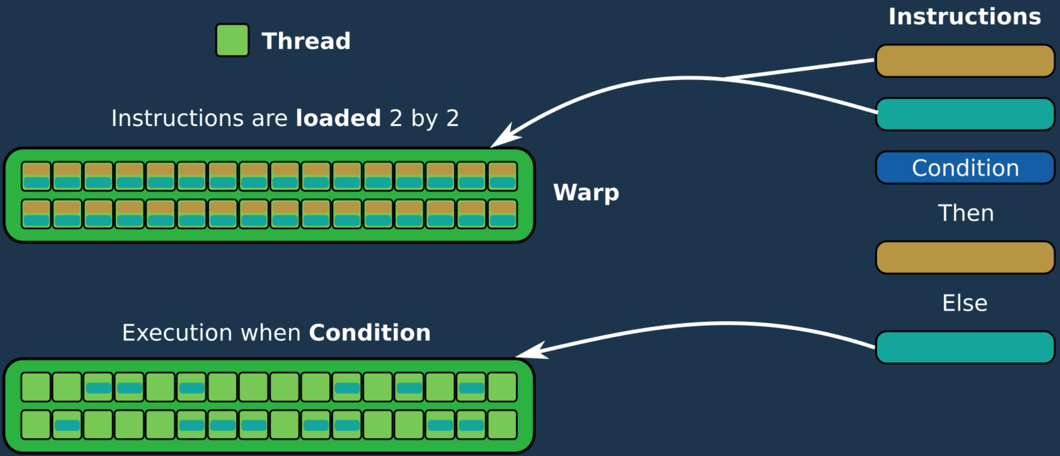
Threads Execution



Threads Execution



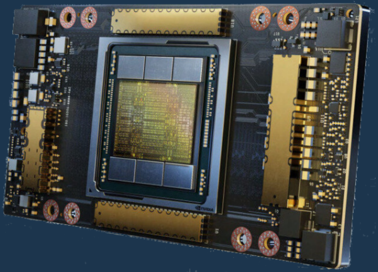
Threads Execution



Ampere GPU

Ampere GPU

A100 (Ampere)

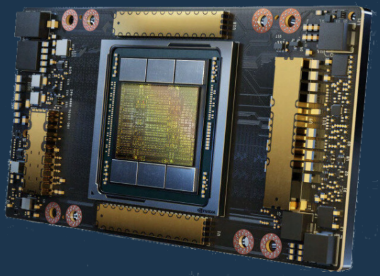


A100 SXM

Ampere GPU

7 nm

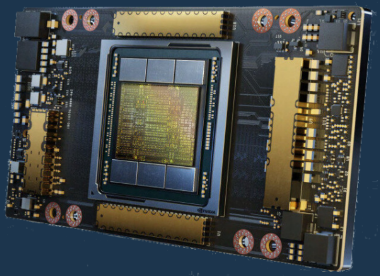
A100 (Ampere)



A100 SXM

Ampere GPU

A100 (Ampere)



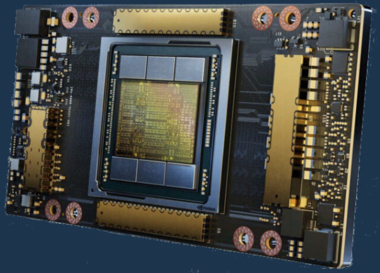
A100 SXM

7 nm

Compute Capabilites **8.0**

Ampere GPU

A100 (Ampere)



A100 SXM

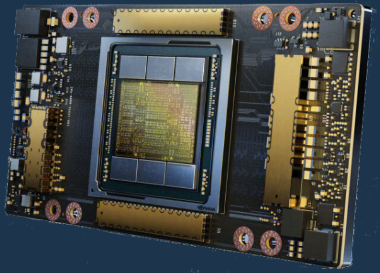
7 nm

Compute Capabilites **8.0**

108 SMs **6912** Cores

Ampere GPU

A100 (Ampere)



A100 SXM

7 nm

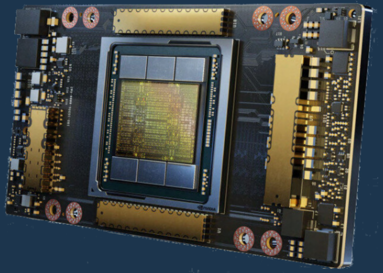
Compute Capabilites **8.0**

108 SMs **6912** Cores

Peak **19.5 TFlops** (without tensor cores)

Ampere GPU

A100 (Ampere)



A100 SXM

7 nm

Compute Capabilites **8.0**

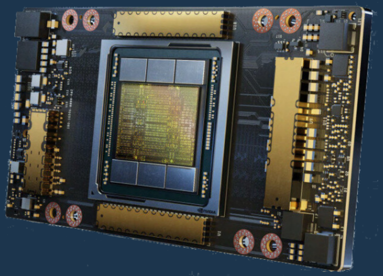
108 SMs **6912** Cores

Peak **19.5 TFlops** (without tensor cores)

40 - **80** GB DRAM

Ampere GPU

A100 (Ampere)



A100 SXM

7 nm

Compute Capabilites **8.0**

108 SMs **6912** Cores

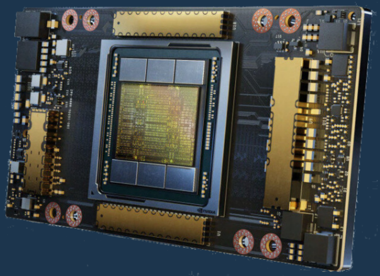
Peak **19.5 TFlops** (without tensor cores)

40 - **80** GB DRAM

1.3 TB/s Bandwidth (**HBM2**)

Ampere GPU

A100 (Ampere)



A100 SXM

7 nm

Compute Capabilites **8.0**

108 SMs **6912** Cores

Peak **19.5 TFlops** (without tensor cores)

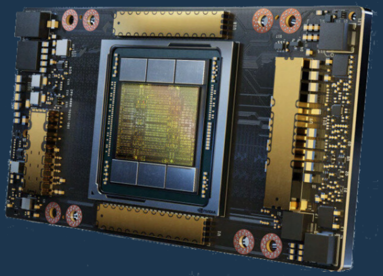
40 - **80** GB DRAM

1.3 TB/s Bandwidth (**HBM2**)

NVLink 600 GB/s

Ampere GPU

A100 (Ampere)



A100 SXM

7 nm

Compute Capabilites **8.0**

108 SMs **6912** Cores

Peak **19.5 TFlops** (without tensor cores)

40 - **80** GB DRAM

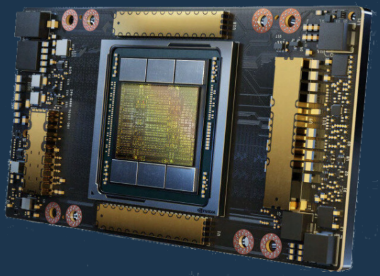
1.3 TB/s Bandwidth (**HBM2**)

NVLink 600 GB/s

MIG : **Multiple Instance GPU**

Ampere GPU

A100 (Ampere)



A100 SXM

7 nm

Compute Capabilites **8.0**

108 SMs **6912** Cores

Peak **19.5 TFlops** (without tensor cores)

40 - **80** GB DRAM

1.3 TB/s Bandwidth (**HBM2**)

NVLink 600 GB/s

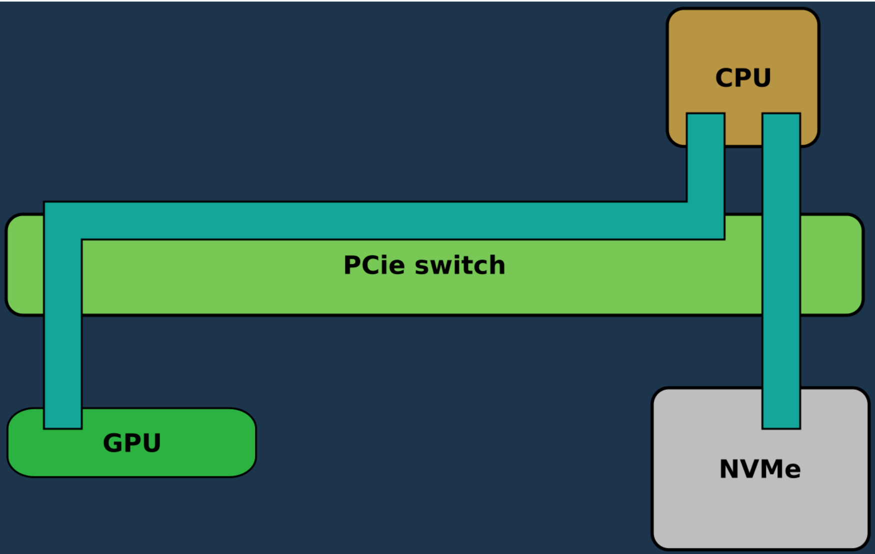
MIG : **Multiple Instance GPU**

Too much for **PCIe** ?

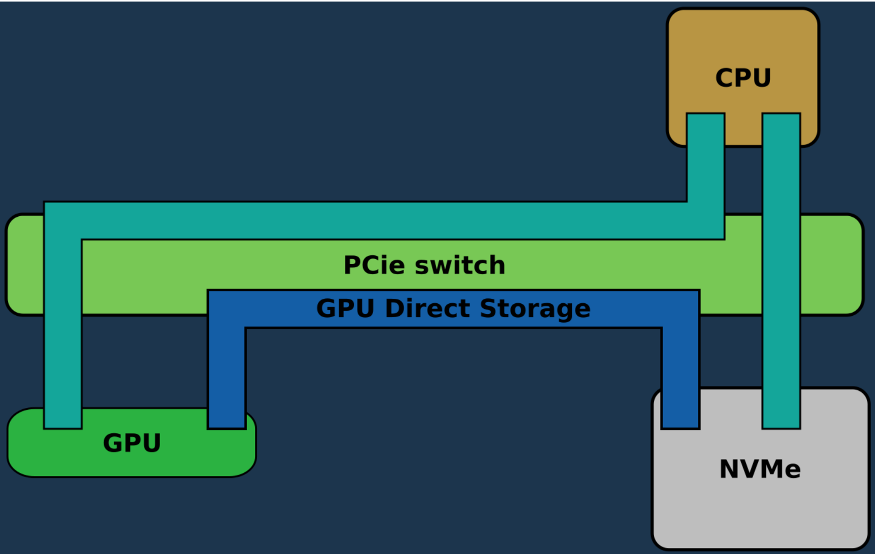
Hardware evolution (NVidia)

Hardware evolution (NVidia)

Hardware evolution (NVidia)



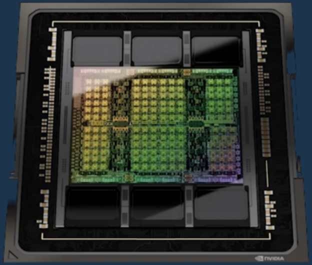
Hardware evolution (NVidia)



Hopper GPU

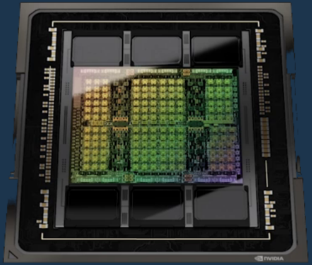
Hopper GPU

H100 (Hopper)



Hopper GPU

H100 (Hopper)

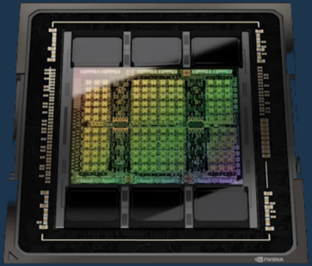


H100 SXM5 and **PCie**

Hopper GPU

4 nm (**A100 7** nm)

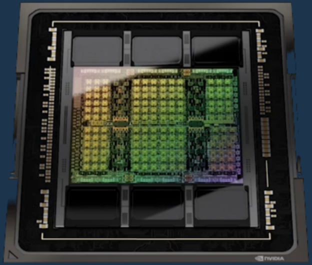
H100 (Hopper)



H100 SXM5 and **PCie**

Hopper GPU

H100 (Hopper)



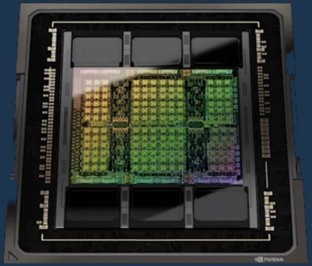
H100 SXM5 and **PCie**

4 nm (**A100 7** nm)

Compute Capabilites **9.0**

Hopper GPU

H100 (Hopper)



H100 SXM5 and **PCie**

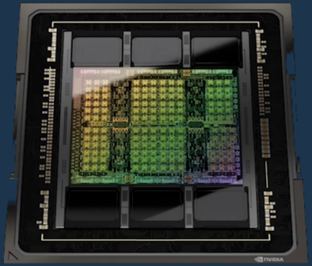
4 nm (**A100 7** nm)

Compute Capabilites **9.0**

132 SMs (tensor core Gen4) **8448** Cores

Hopper GPU

H100 (Hopper)



H100 SXM5 and **PCie**

4 nm (**A100 7** nm)

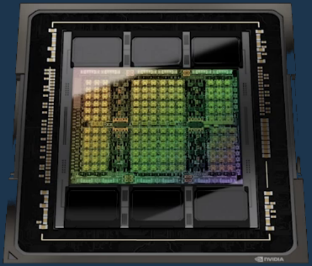
Compute Capabilites **9.0**

132 SMs (tensor core Gen4) **8448** Cores

Peak **30 TFlops** (without tensor cores)

Hopper GPU

H100 (Hopper)



H100 SXM5 and **PCie**

4 nm (**A100 7** nm)

Compute Capabilites **9.0**

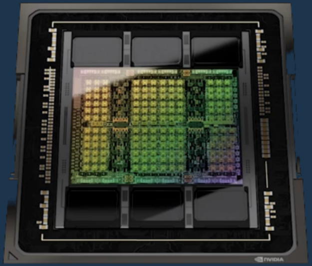
132 SMs (tensor core Gen4) **8448** Cores

Peak **30 TFlops** (without tensor cores)

80 GB DRAM

Hopper GPU

H100 (Hopper)



H100 SXM5 and **PCie**

4 nm (**A100 7** nm)

Compute Capabilites **9.0**

132 SMs (tensor core Gen4) **8448** Cores

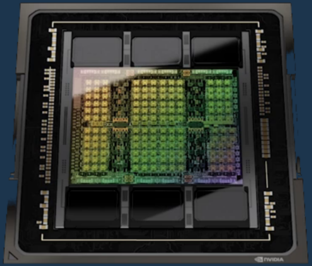
Peak **30 TFlops** (without tensor cores)

80 GB DRAM

3 TB/s Bandwidth (HBM3)

Hopper GPU

H100 (Hopper)



H100 SXM5 and **PCie**

4 nm (**A100 7** nm)

Compute Capabilites **9.0**

132 SMs (tensor core Gen4) **8448** Cores

Peak **30 TFlops** (without tensor cores)

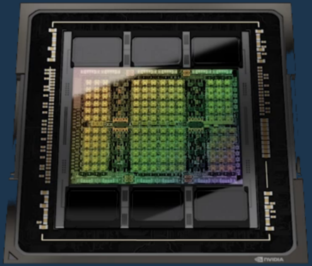
80 GB DRAM

3 TB/s Bandwidth (HBM3)

NVLink Gen 4 (**900** GB/s)

Hopper GPU

H100 (Hopper)



H100 SXM5 and **PCie**

4 nm (**A100 7** nm)

Compute Capabilites **9.0**

132 SMs (tensor core Gen4) **8448** Cores

Peak **30 TFlops** (without tensor cores)

80 GB DRAM

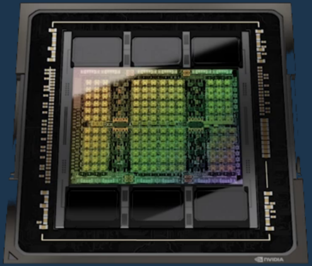
3 TB/s Bandwidth (HBM3)

NVLink Gen 4 (**900** GB/s)

MIG Gen 2 with **confidential computing** (**all** instances have **Image/Video decoding**)

Hopper GPU

H100 (Hopper)



H100 SXM5 and **PCie**

4 nm (**A100 7** nm)

Compute Capabilites **9.0**

132 SMs (tensor core Gen4) **8448** Cores

Peak **30 TFlops** (without tensor cores)

80 GB DRAM

3 TB/s Bandwidth (HBM3)

NVLink Gen 4 (**900** GB/s)

MIG Gen 2 with **confidential computing** (**all** instances have **Image/Video decoding**)

A lot of **computing precisions**

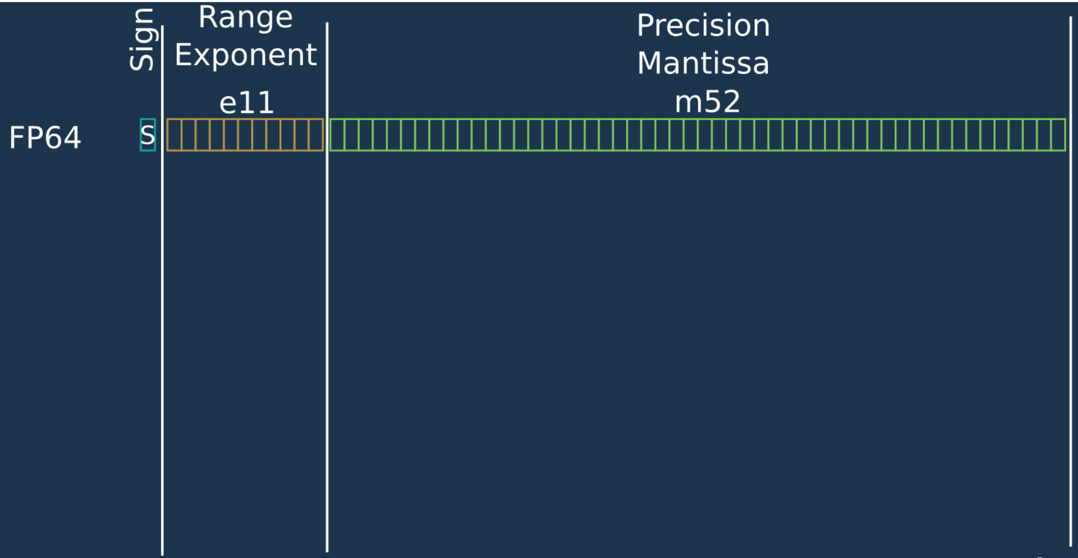
Computing Precision

Sign

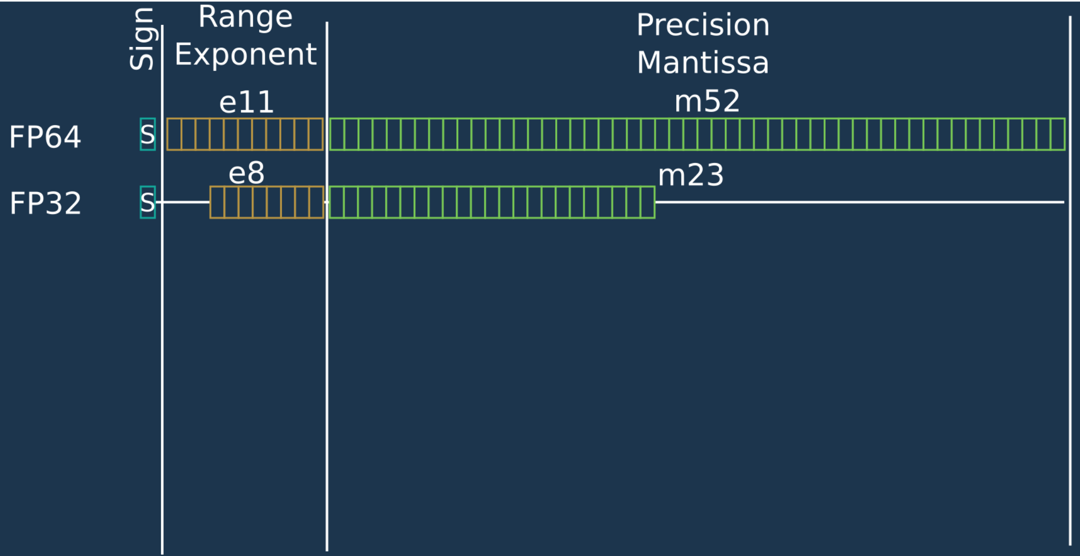
Range
Exponent

Precision
Mantissa

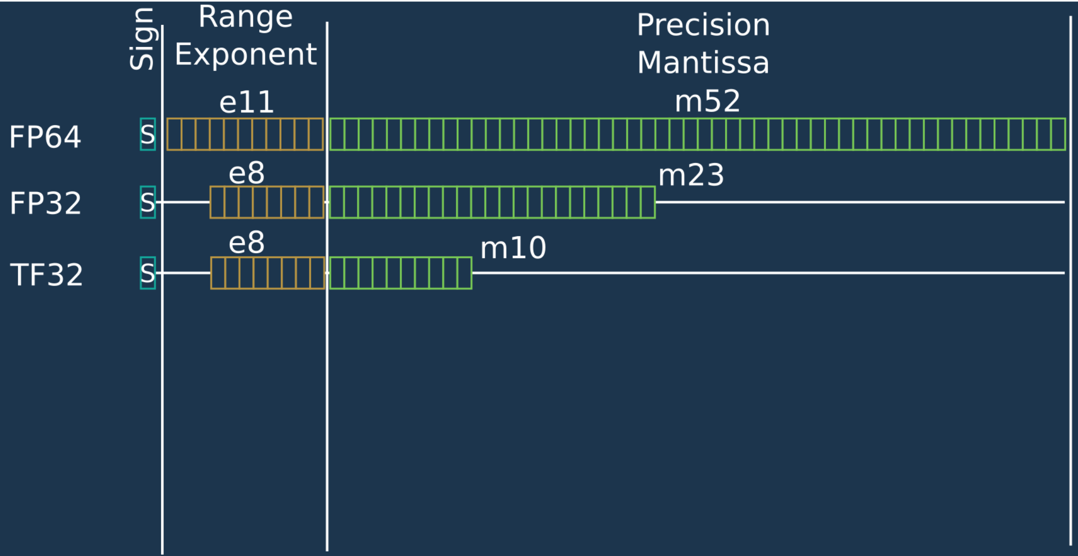
Computing Precision



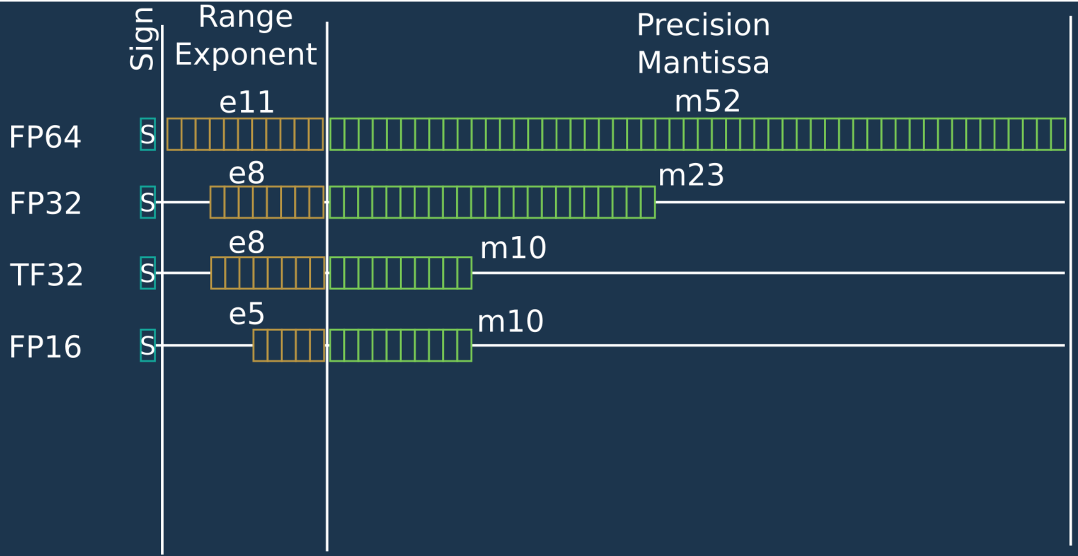
Computing Precision



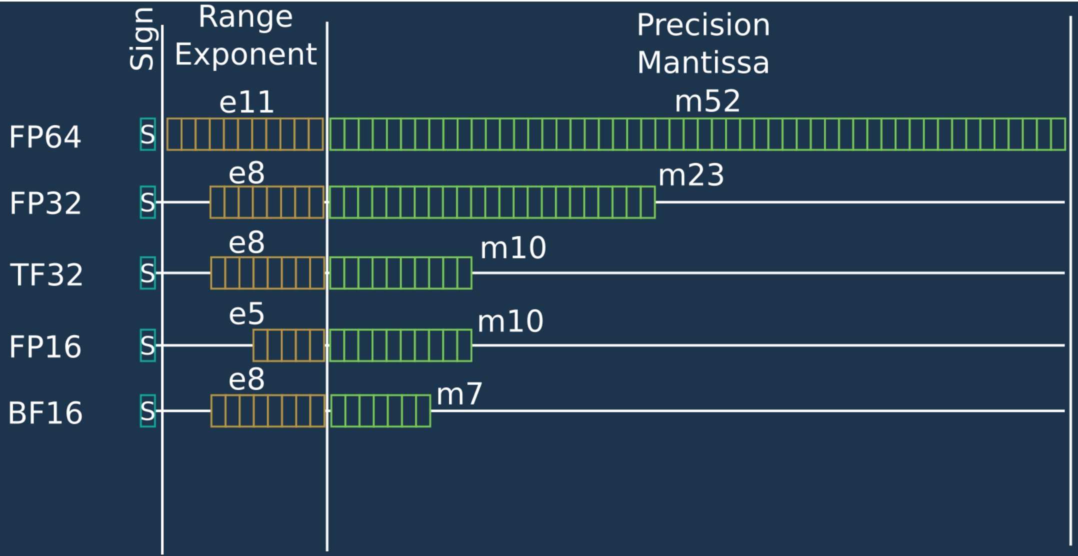
Computing Precision



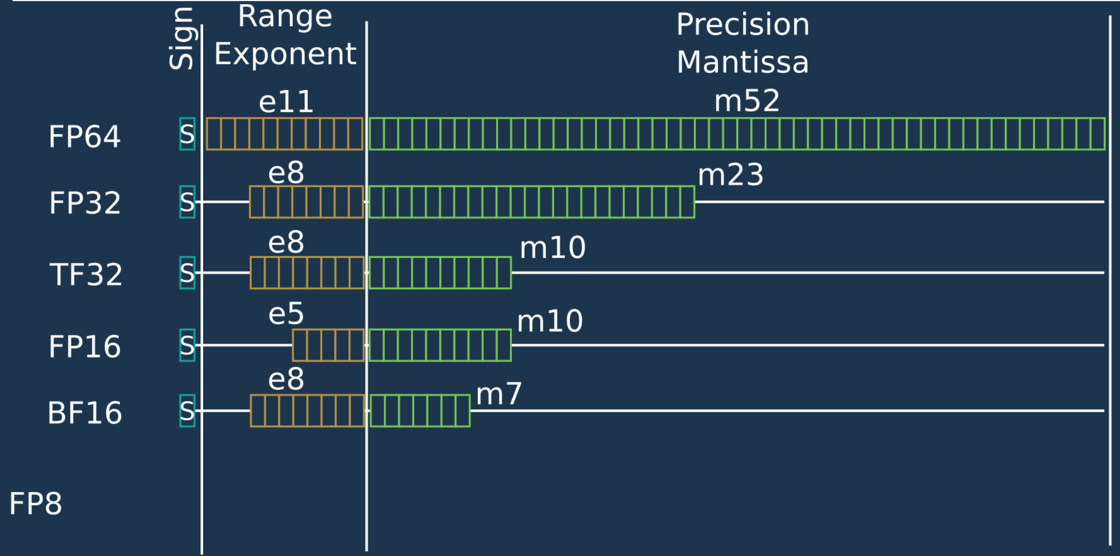
Computing Precision



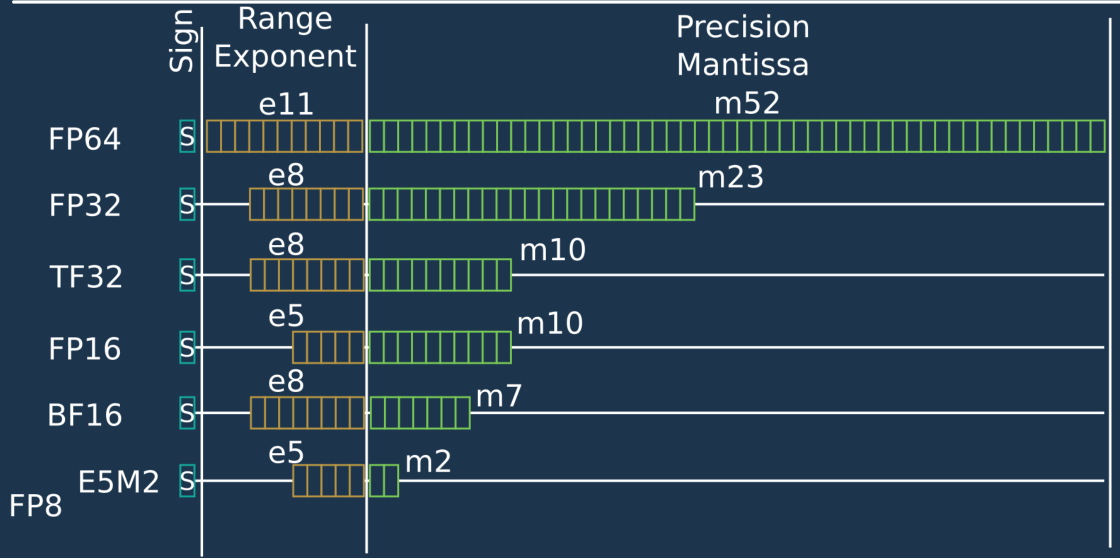
Computing Precision



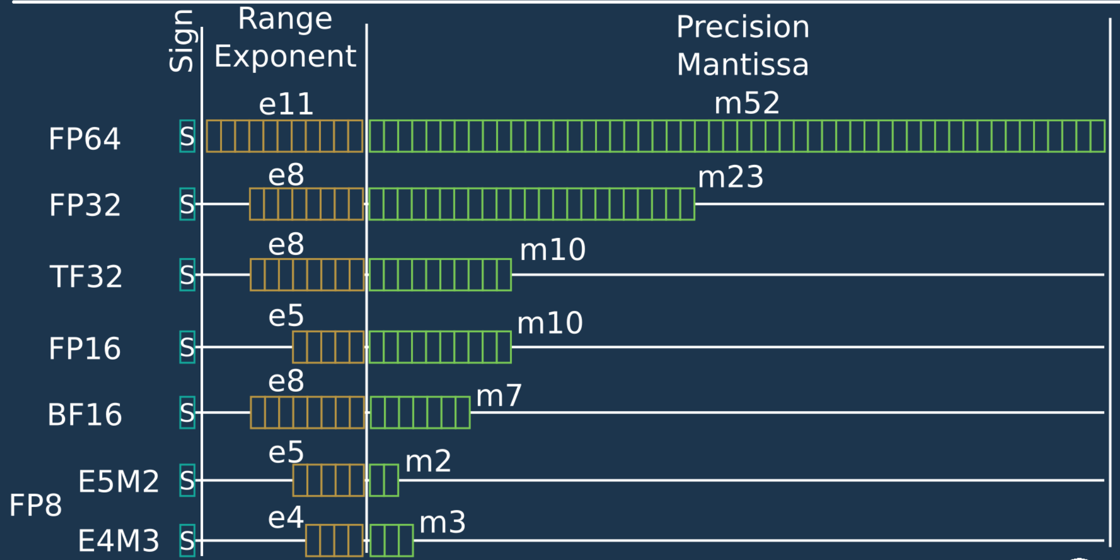
Computing Precision



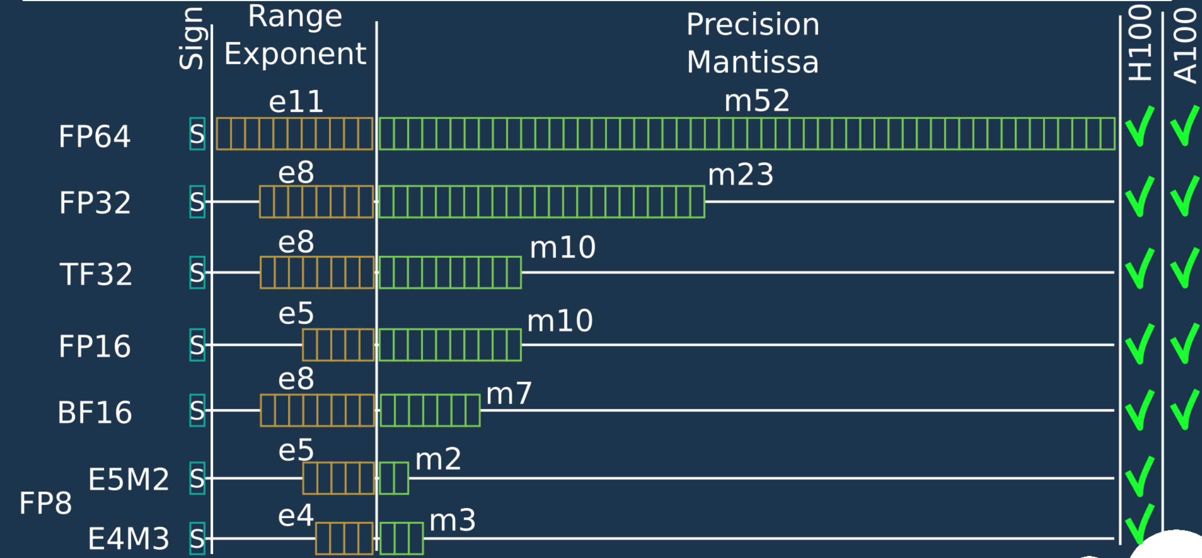
Computing Precision



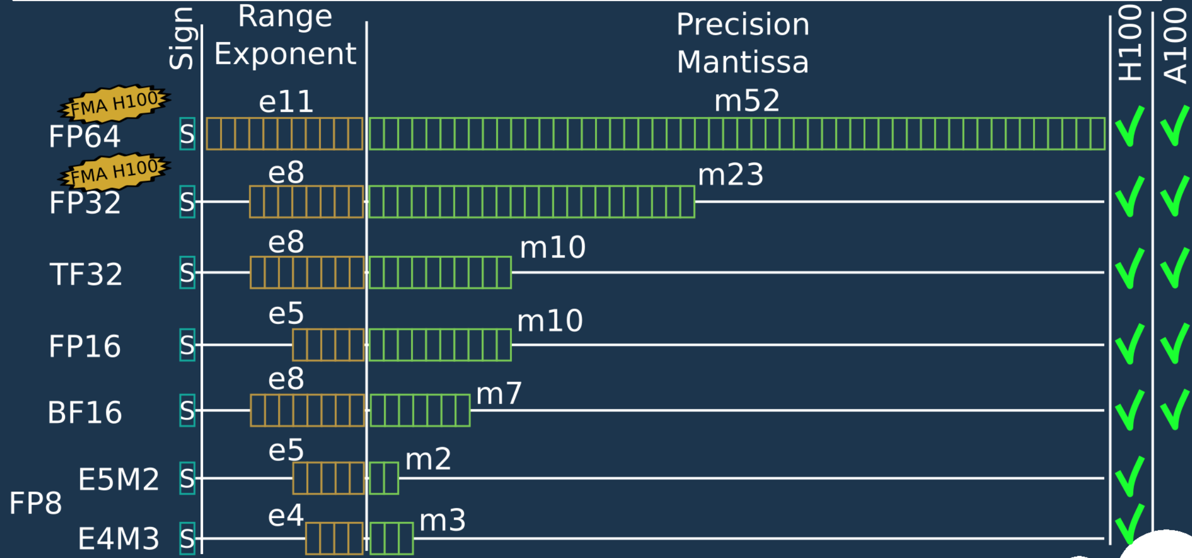
Computing Precision



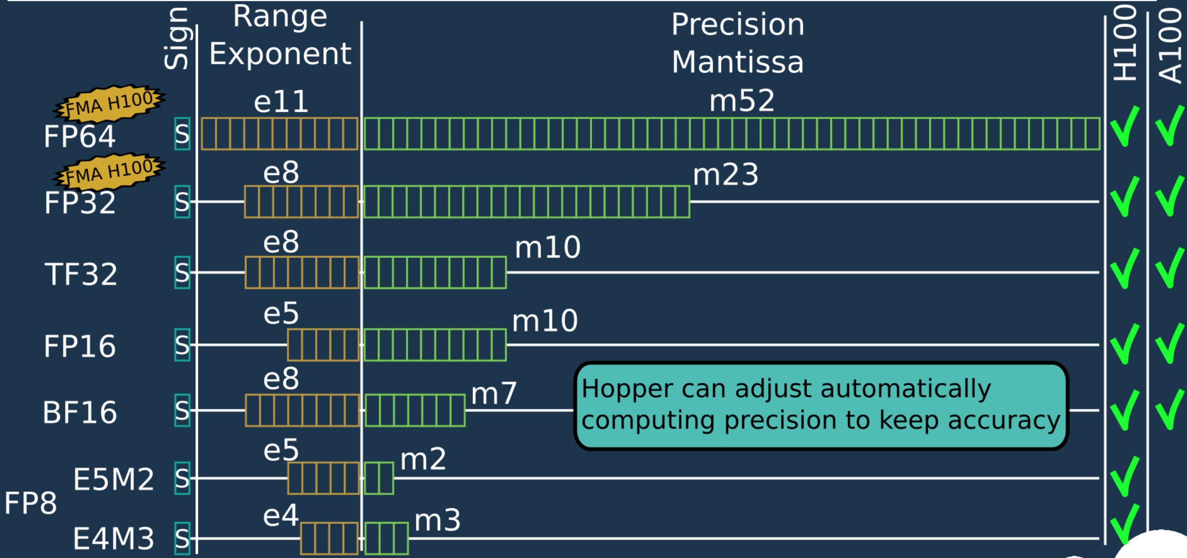
Computing Precision



Computing Precision



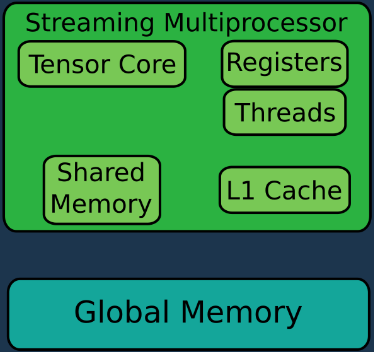
Computing Precision



Thread acceleration

Thread acceleration

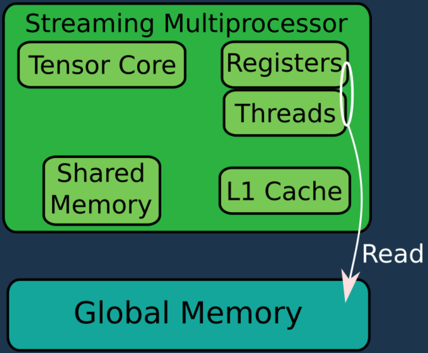
Tensor Memory
Acceleration (TMA)



Thread acceleration

Tensor Memory Acceleration (TMA)

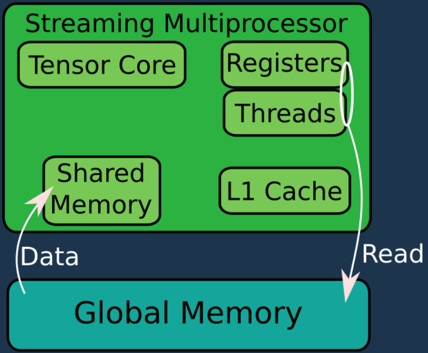
A100 (no TMA)



Thread acceleration

Tensor Memory
Acceleration (TMA)

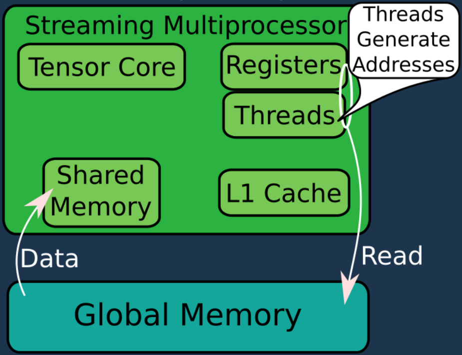
A100 (no TMA)



Thread acceleration

Tensor Memory Acceleration (TMA)

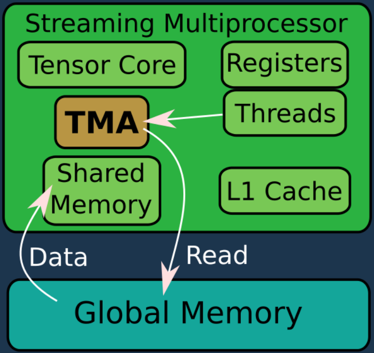
A100 (no TMA)



Thread acceleration

Tensor Memory Acceleration (TMA)

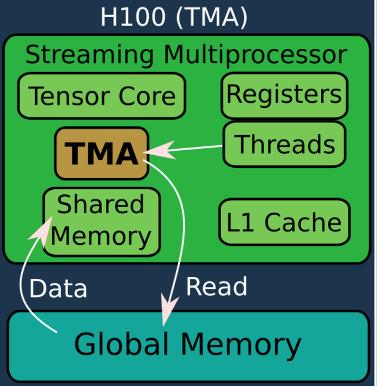
H100 (TMA)



Thread acceleration

Tensor Memory Acceleration (TMA)

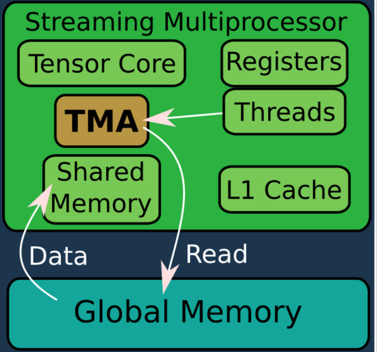
Distributed shared memory



Thread acceleration

Tensor Memory Acceleration (TMA)

H100 (TMA)

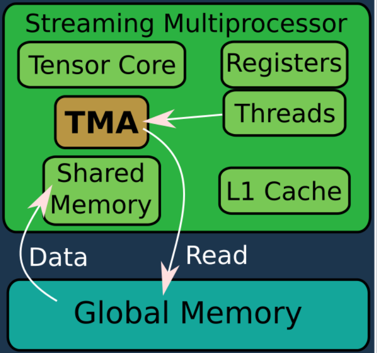


Distributed shared memory

Thread acceleration

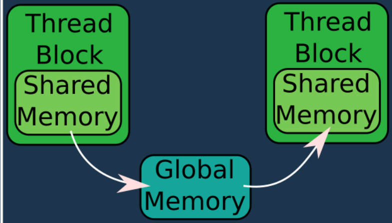
Tensor Memory Acceleration (TMA)

H100 (TMA)



Distributed shared memory

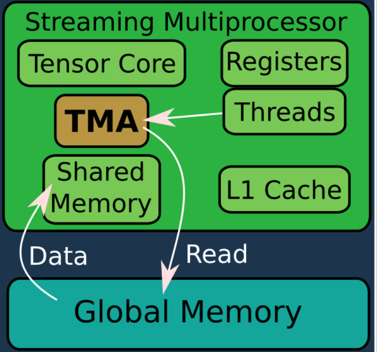
A100



Thread acceleration

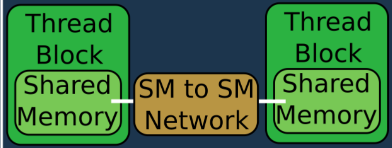
Tensor Memory
Acceleration (TMA)

H100 (TMA)



Distributed
shared memory

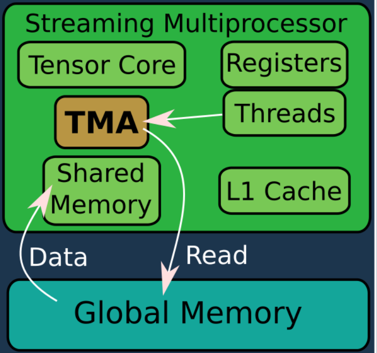
H100



Thread acceleration

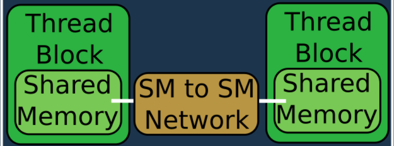
Tensor Memory Acceleration (TMA)

H100 (TMA)



Distributed shared memory

H100

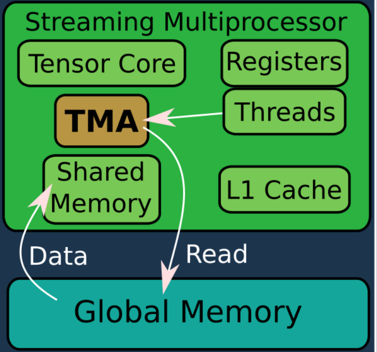


Asynchronous Transaction Barrier

Thread acceleration

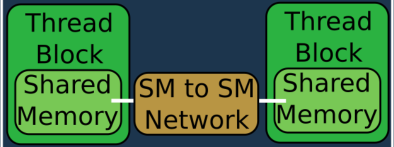
Tensor Memory Acceleration (TMA)

H100 (TMA)



Distributed shared memory

H100



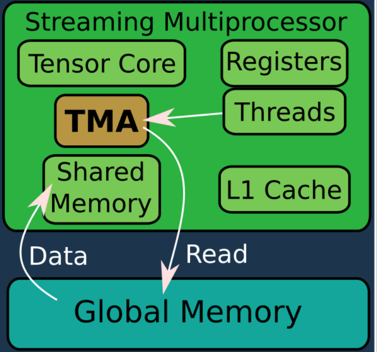
Asynchronous Transaction Barrier

A100

Thread acceleration

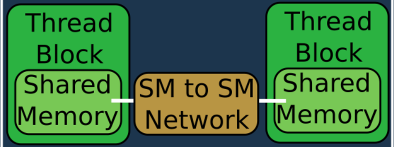
Tensor Memory Acceleration (TMA)

H100 (TMA)



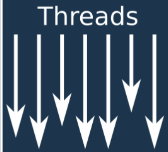
Distributed shared memory

H100



Asynchronous Transaction Barrier

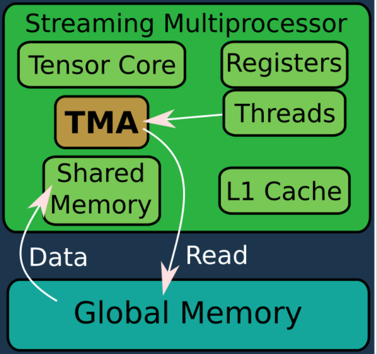
A100



Thread acceleration

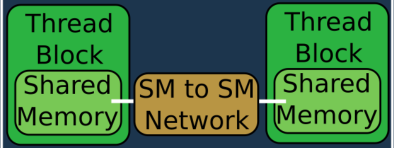
Tensor Memory Acceleration (TMA)

H100 (TMA)



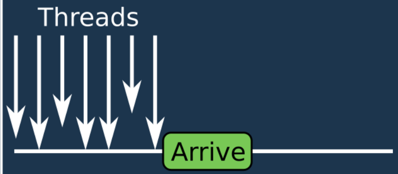
Distributed shared memory

H100



Asynchronous Transaction Barrier

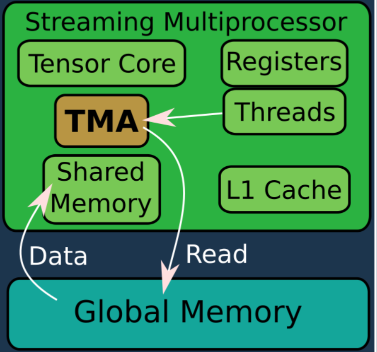
A100



Thread acceleration

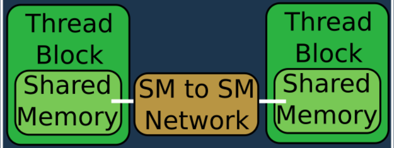
Tensor Memory Acceleration (TMA)

H100 (TMA)



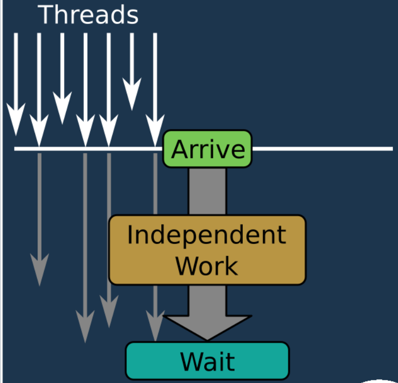
Distributed shared memory

H100



Asynchronous Transaction Barrier

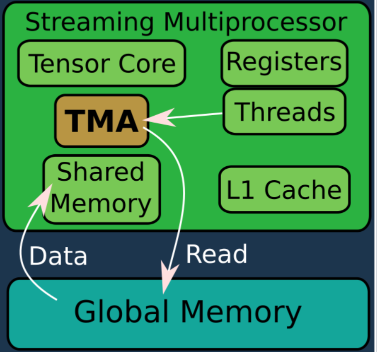
A100



Thread acceleration

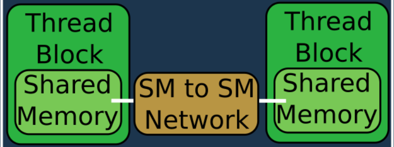
Tensor Memory Acceleration (TMA)

H100 (TMA)



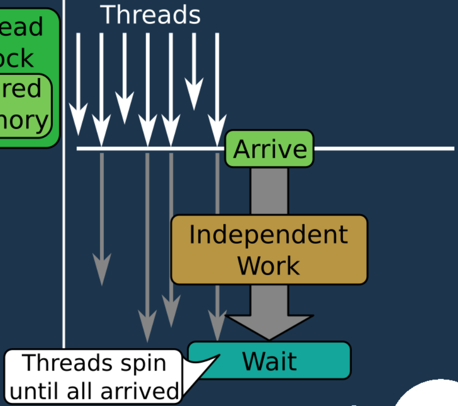
Distributed shared memory

H100



Asynchronous Transaction Barrier

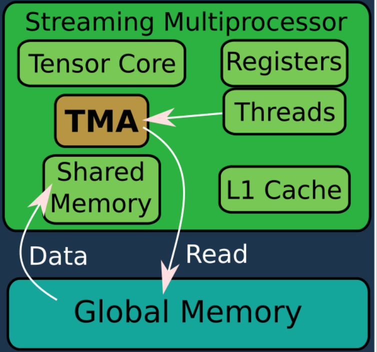
A100



Thread acceleration

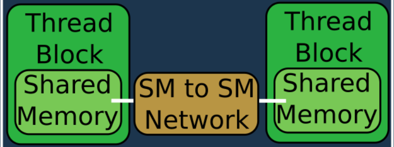
Tensor Memory Acceleration (TMA)

H100 (TMA)



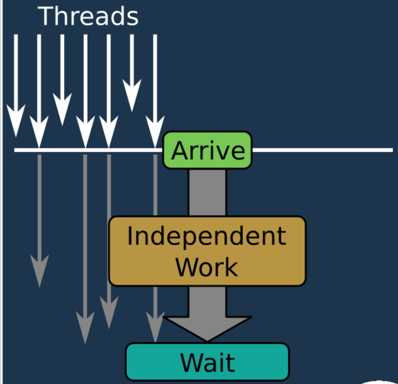
Distributed shared memory

H100



Asynchronous Transaction Barrier

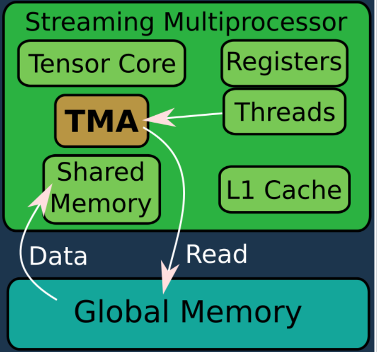
H100



Thread acceleration

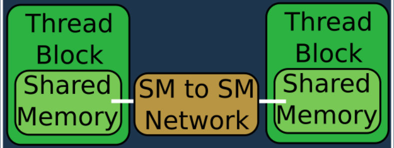
Tensor Memory Acceleration (TMA)

H100 (TMA)



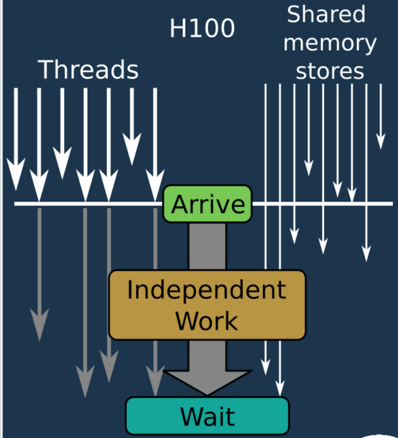
Distributed shared memory

H100



Asynchronous Transaction Barrier

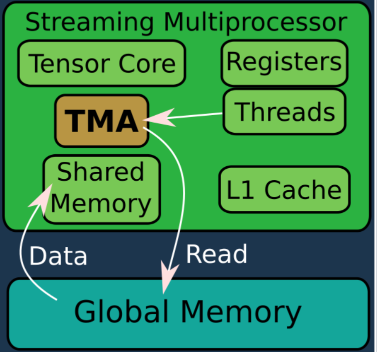
H100



Thread acceleration

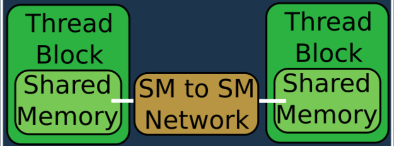
Tensor Memory Acceleration (TMA)

H100 (TMA)

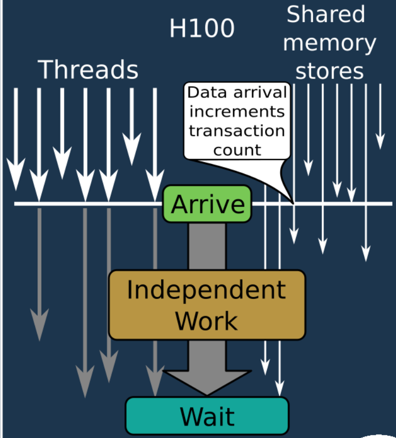


Distributed shared memory

H100

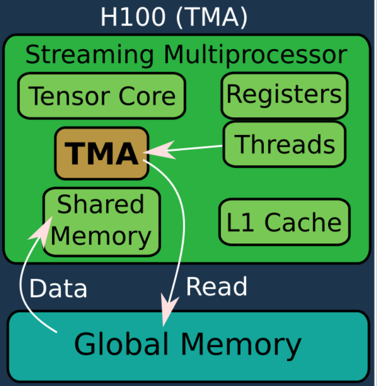


Asynchronous Transaction Barrier

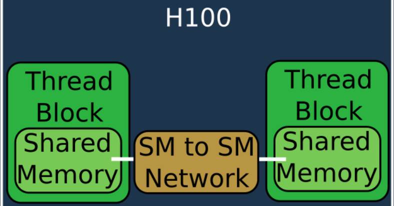


Thread acceleration

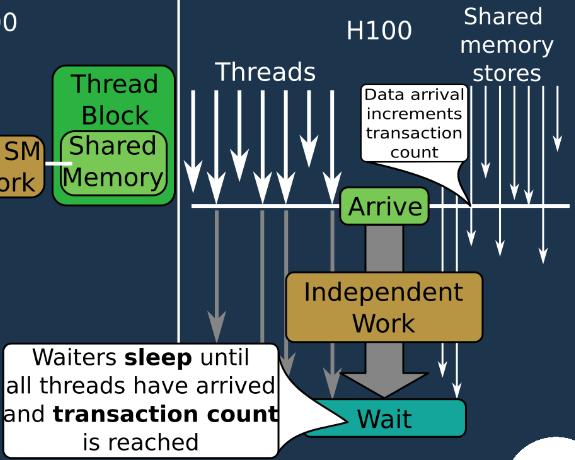
Tensor Memory Acceleration (TMA)



Distributed shared memory

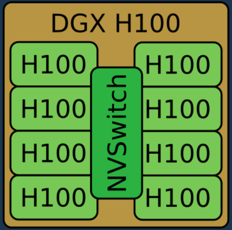


Asynchronous Transaction Barrier

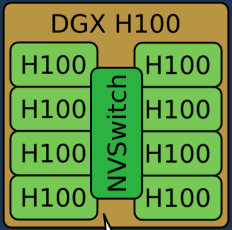


DGX H100 / DGX POD H100

DGX H100 / DGX POD H100



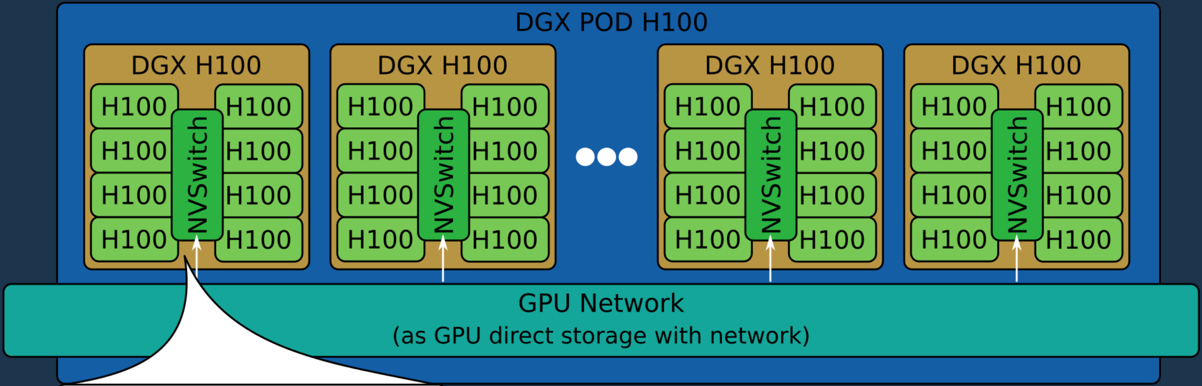
DGX H100 / DGX POD H100



DGX H100 :

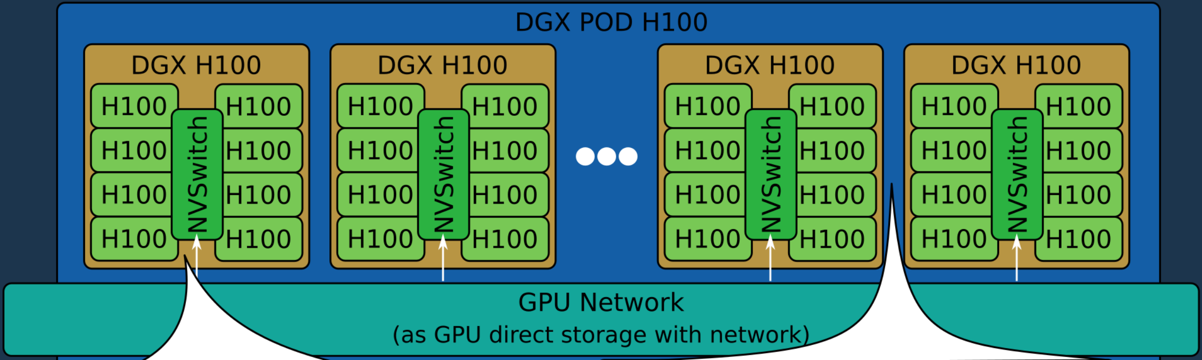
- 8 H100
- 32 PFlop (AI performance)
- 640 GB HBM3 Memory
- 24 TB/s memory Bandwidth

DGX H100 / DGX POD H100



DGX H100 :
- 8 H100
- 32 PFlop (AI performance)
- 640 GB HBM3 Memory
- 24 TB/s memory Bandwidth

DGX H100 / DGX POD H100



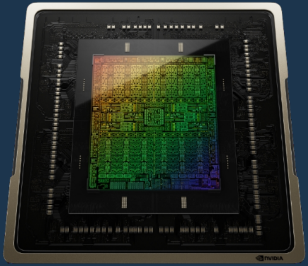
DGX H100 :
- 8 H100
- 32 PFlop (AI performance)
- 640 GB HBM3 Memory
- 24 TB/s memory Bandwidth

DGX POD H100 :
- 32 DGX H100
- 1 EFlop (AI performance)
- 20 TB HBM3 Memory
- 70 TB/s memory Bandwidth
- 192 TFlops of Sharp in-Network compute

Ada Lovelace GPU

Ada Lovelace GPU

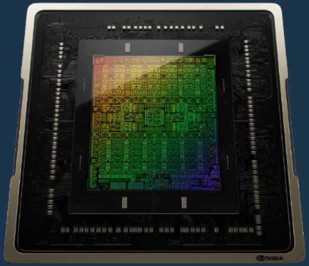
GeForce 4090
(Ada Lovelace)



Ada Lovelace GPU

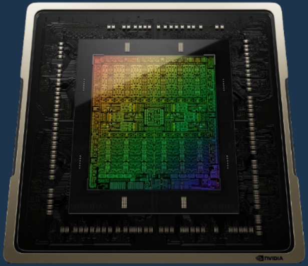
4 nm (A100 7 nm)

GeForce 4090
(Ada Lovelace)



Ada Lovelace GPU

GeForce 4090
(Ada Lovelace)

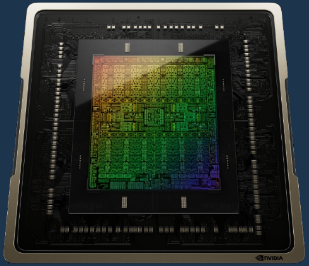


4 nm (A100 7 nm)

128 SMs (tensor core Gen4) **16384** Cuda Cores

Ada Lovelace GPU

GeForce 4090
(Ada Lovelace)



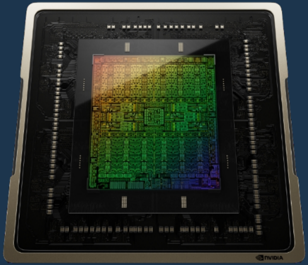
4 nm (A100 7 nm)

128 SMs (tensor core Gen4) **16384** Cuda Cores

24 GB GDDR6X

Ada Lovelace GPU

GeForce 4090
(Ada Lovelace)



4 nm (A100 7 nm)

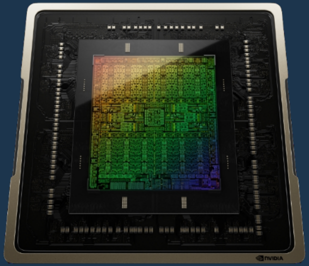
128 SMs (tensor core Gen4) **16384** Cuda Cores

24 GB GDDR6X

1 TB/s Bandwidth

Ada Lovelace GPU

GeForce 4090
(Ada Lovelace)



4 nm (A100 7 nm)

128 SMs (tensor core Gen4) **16384** Cuda Cores

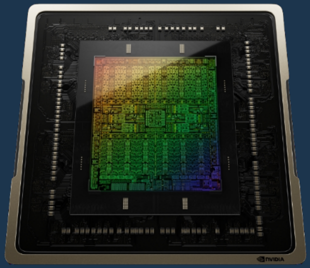
24 GB GDDR6X

1 TB/s Bandwidth

Peak **82** TFlops (FP32 without tensor cores)

Ada Lovelace GPU

GeForce 4090
(Ada Lovelace)



4 nm (A100 7 nm)

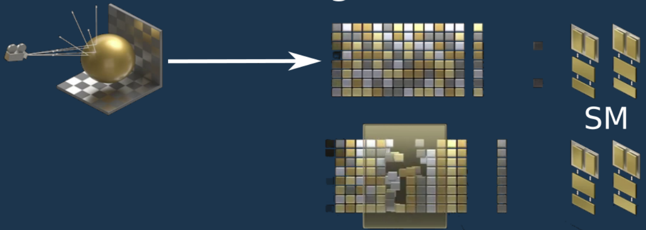
128 SMs (tensor core Gen4) **16384** Cuda Cores

24 GB GDDR6X

1 TB/s Bandwidth

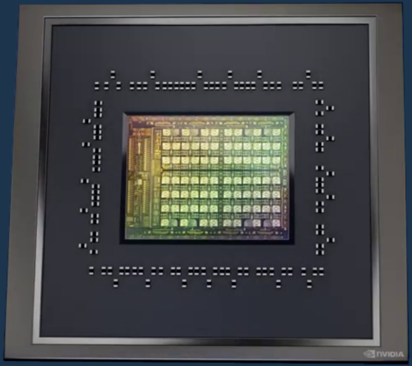
Peak **82** TFlops (FP32 without tensor cores)

Shader **Reordering**



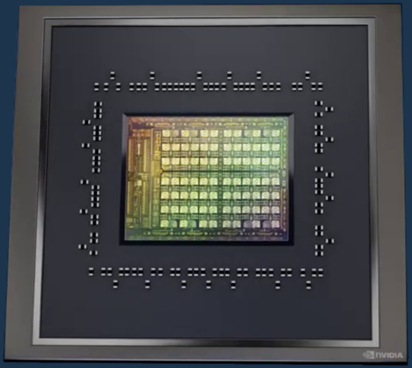
Grace CPU

Grace



Grace CPU

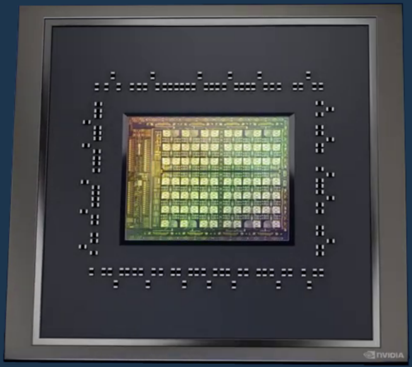
Grace



72 Neoverse V2 cores, **Armv9**
SVE2 with 4x128b

Grace CPU

Grace



72 Neoverse V2 cores, **Armv9 SVE2** with 4x128b

240 GB RAM (**LPDDR5X**)

Grace CPU

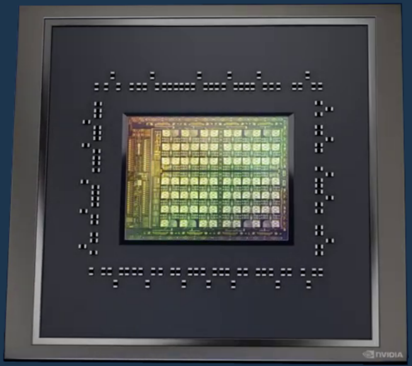
72 Neoverse V2 cores, **Armv9 SVE2** with 4x128b

240 GB RAM (**LPDDR5X**)

3.2 TB/s Bi-section BW

Grace CPU

Grace



72 Neoverse V2 cores, **Armv9 SVE2** with 4x128b

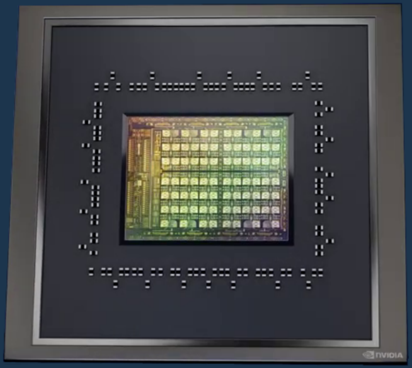
240 GB RAM (**LPDDR5X**)

3.2 TB/s Bi-section BW

Cache :
- **L1**: **64 KB** (I/D) per core
- **L2**: **1 MB** per core
- **L3**: **117 MB**

Grace CPU

Grace



72 Neoverse V2 cores, **Armv9**
SVE2 with 4x128b

240 GB RAM (**LPDDR5X**)

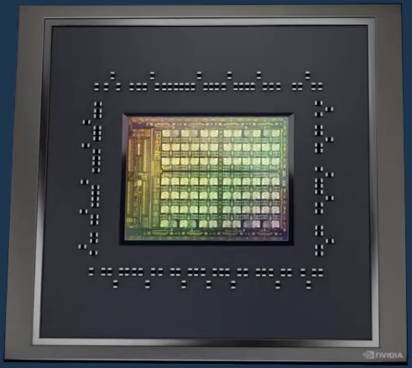
3.2 TB/s Bi-section BW

Cache :
- **L1**: **64 KB** (I/D) per core
- **L2**: **1 MB** per core
- **L3**: **117 MB**

Local caching of remote CPU/GPU memory
Supports up to 4 chip coherency over Coherent NVLINK

Grace CPU

Grace



72 Neoverse V2 cores, **Armv9**
SVE2 with 4x128b

240 GB RAM (**LPDDR5X**)

3.2 TB/s Bi-section BW

Cache :
- **L1**: **64 KB** (I/D) per core
- **L2**: **1 MB** per core
- **L3**: **117 MB**

Local caching of remote CPU/GPU memory
Supports up to 4 chip coherency over Coherent NVLINK

NVLink C2C (**900** GB/s)

Grace Superchip CPU

Grace Superchip

Grace Superchip CPU

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

Grace Superchip CPU

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

144 Cores

Grace Superchip CPU

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

144 Cores

1 TB/s Bandwidth (**HBM3**)

Grace Superchip CPU

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

144 Cores

1 TB/s Bandwidth (**HBM3**)

960 GB RAM (**LPDDR5X**)

Grace Superchip CPU

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

144 Cores

1 TB/s Bandwidth (**HBM3**)

960 GB RAM (**LPDDR5X**)

Cache :

- **L1**: **64 KB** (I/D) per core
- **L2**: **1 MB** per core
- **L3**: **234 MB** per superchip

Grace Superchip CPU

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

144 Cores

1 TB/s Bandwidth (**HBM3**)

960 GB RAM (**LPDDR5X**)

Cache :
- **L1**: **64 KB** (I/D) per core
- **L2**: **1 MB** per core
- **L3**: **234 MB** per superchip

NVLink Gen 4 (**900** GB/s)

Grace Superchip CPU

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

144 Cores

1 TB/s Bandwidth (**HBM3**)

960 GB RAM (**LPDDR5X**)

Cache :
- **L1**: **64 KB** (I/D) per core
- **L2**: **1 MB** per core
- **L3**: **234 MB** per superchip

NVLink Gen 4 (**900** GB/s)

7.1 TFlops (**FP64** computing peak)

Grace Superchip CPU

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

144 Cores

1 TB/s Bandwidth (**HBM3**)

960 GB RAM (**LPDDR5X**)

Cache :
- **L1**: **64 KB** (I/D) per core
- **L2**: **1 MB** per core
- **L3**: **234 MB** per superchip

NVLink Gen 4 (**900** GB/s)

7.1 TFlops (**FP64** computing peak)

500 W

Grace Superchip CPU

Grace Superchip

Passive air colling

Neoverse V2 Cores: Armv9 with 4x128b **SVE2**

144 Cores

1 TB/s Bandwidth (**HBM3**)

960 GB RAM (**LPDDR5X**)

Cache :

- **L1**: **64 KB** (I/D) per core
- **L2**: **1 MB** per core
- **L3**: **234 MB** per superchip

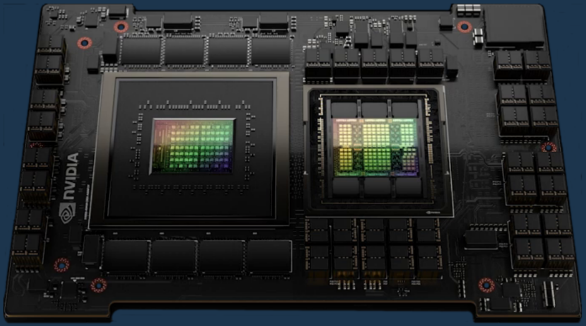
NVLink Gen 4 (**900** GB/s)

7.1 TFlops (**FP64** computing peak)

500 W

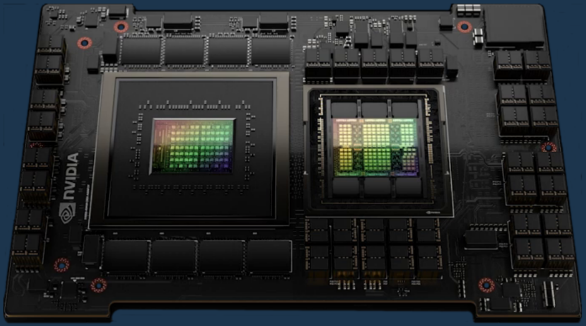
Grace Hopper CPU/GPU

Grace Hopper Superchip



Grace Hopper CPU/GPU

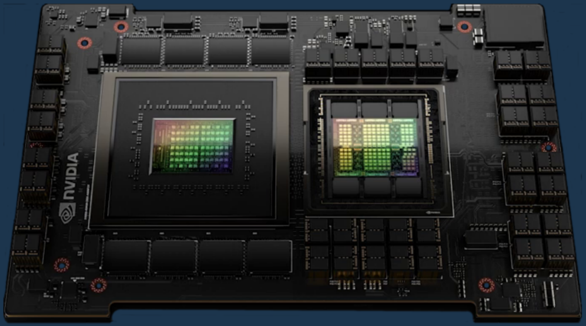
Grace Hopper Superchip



4 TB/s Bandwidth

Grace Hopper CPU/GPU

Grace Hopper Superchip

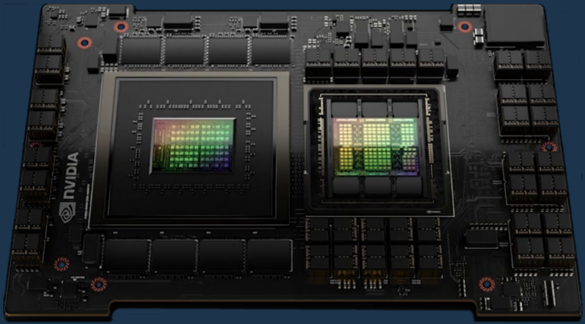


4 TB/s Bandwidth

600 GB RAM (**LPDDR5X**)

Grace Hopper CPU/GPU

Grace Hopper Superchip



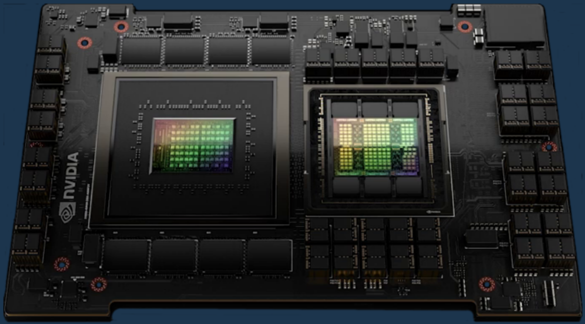
4 TB/s Bandwidth

600 GB RAM (**LPDDR5X**)

NVLink C2C (**900** GB/s)

Grace Hopper CPU/GPU

Grace Hopper Superchip



4 TB/s Bandwidth

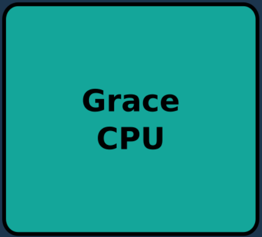
600 GB RAM (**LPDDR5X**)

NVLink C2C (**900** GB/s)

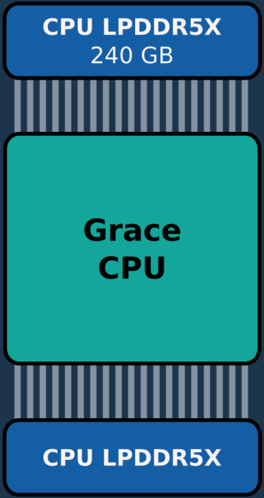
Unified Memory

Grace Hopper Superchip

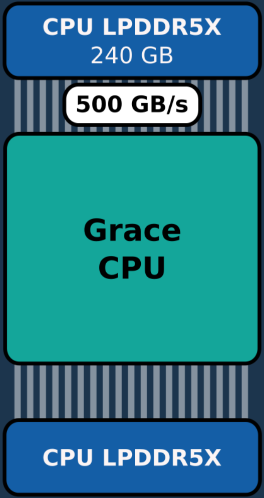
Grace Hopper Superchip



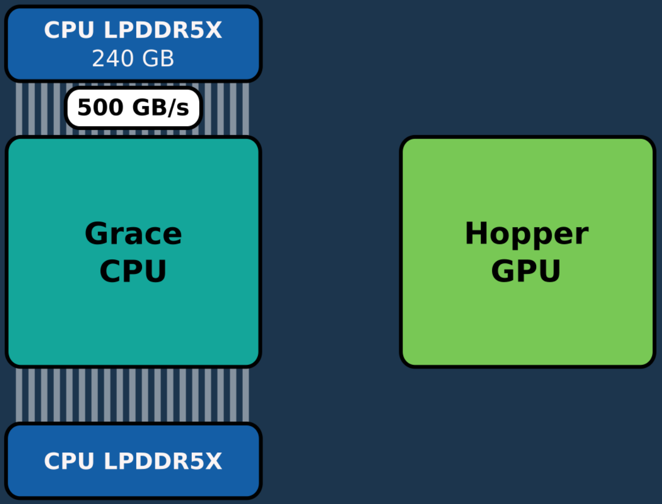
Grace Hopper Superchip



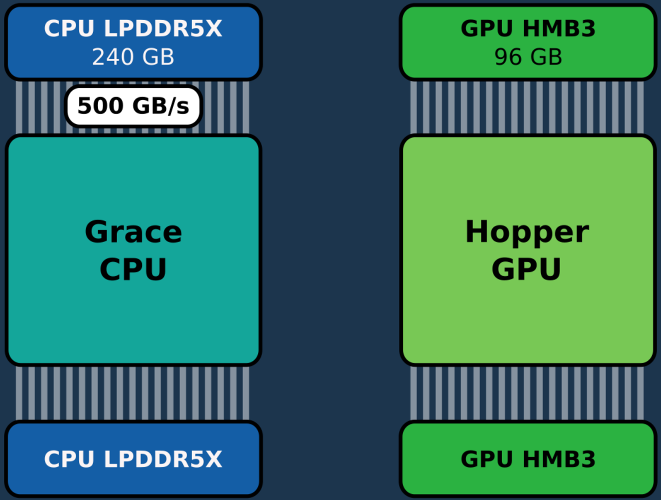
Grace Hopper Superchip



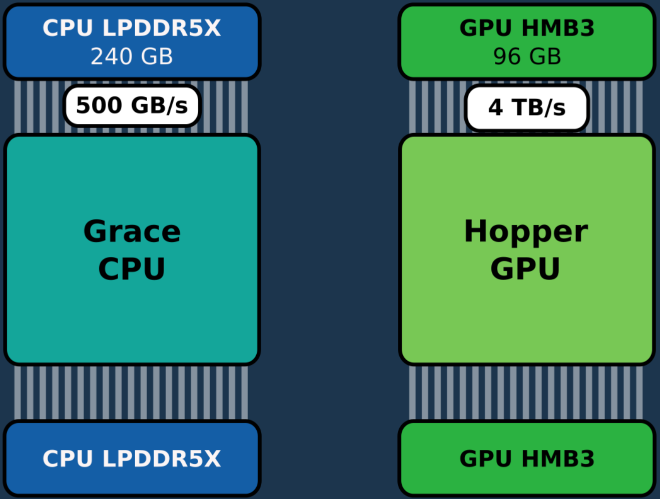
Grace Hopper Superchip



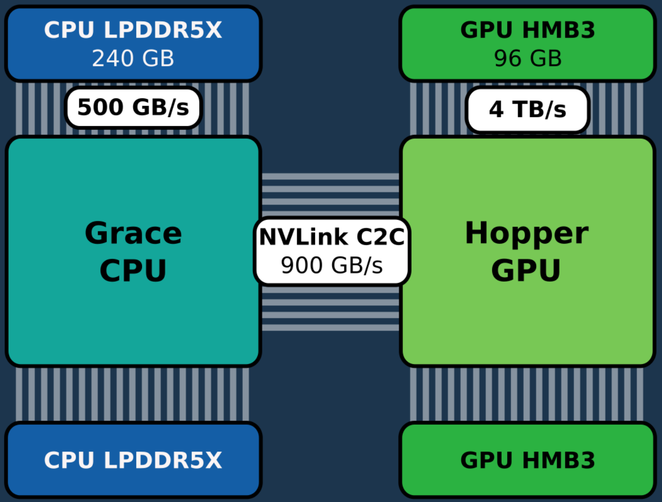
Grace Hopper Superchip



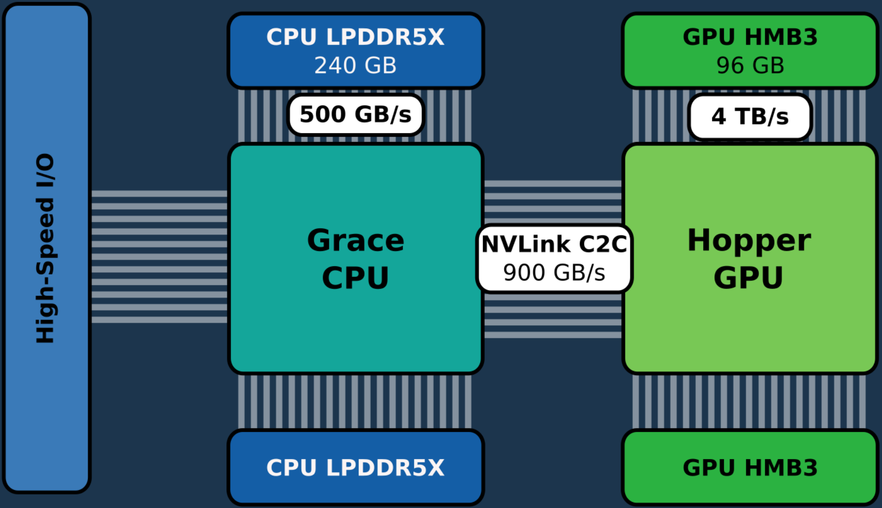
Grace Hopper Superchip



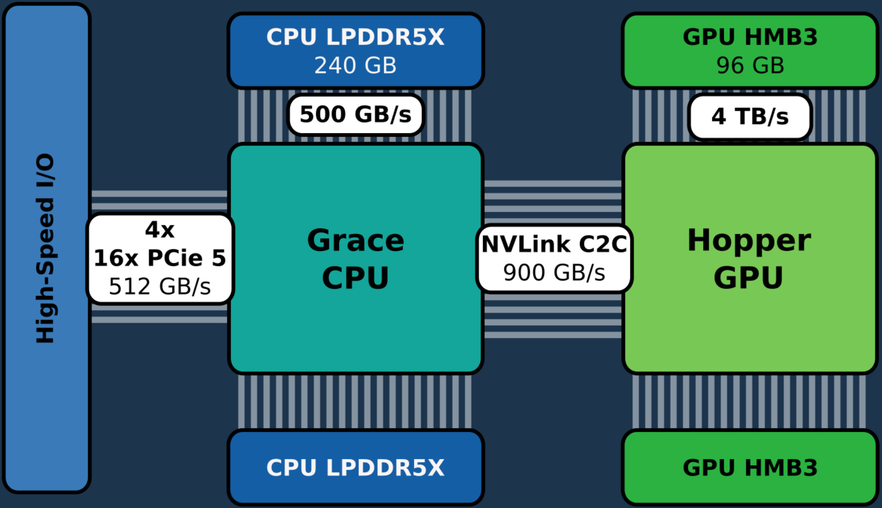
Grace Hopper Superchip



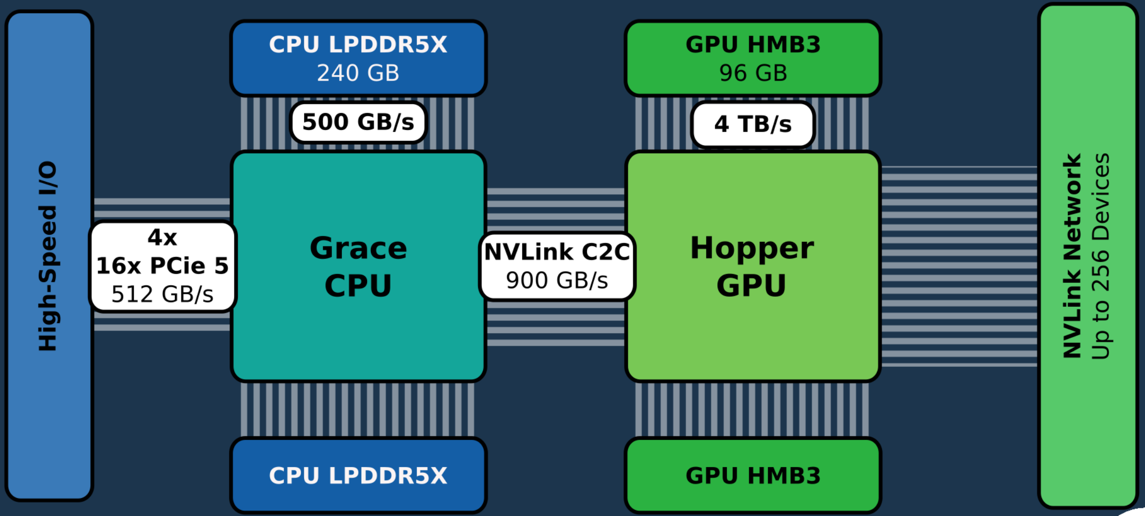
Grace Hopper Superchip



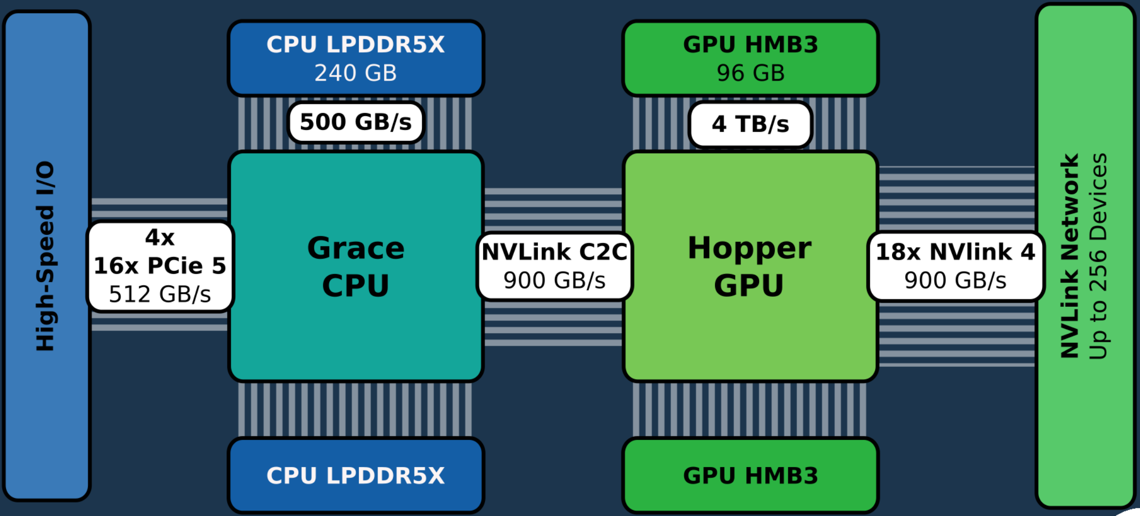
Grace Hopper Superchip



Grace Hopper Superchip

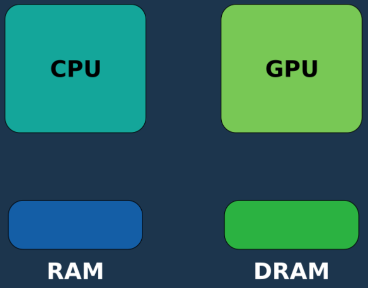


Grace Hopper Superchip



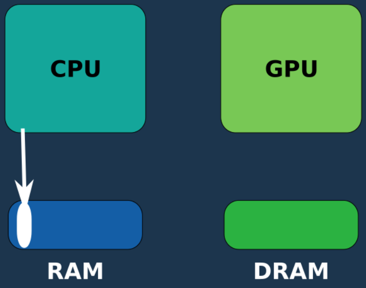
Memory Management

CUDA



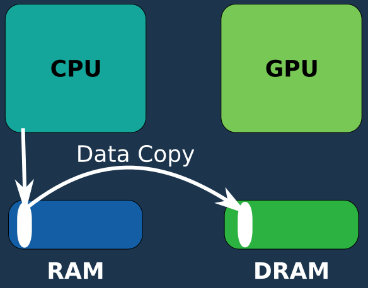
Memory Management

CUDA



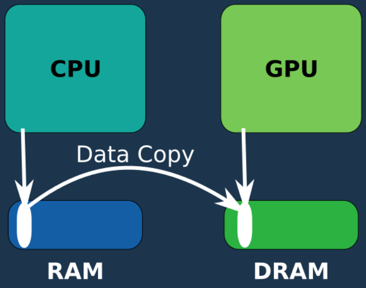
Memory Management

CUDA



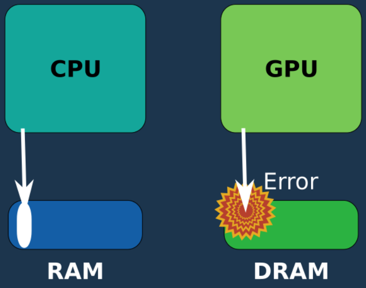
Memory Management

CUDA

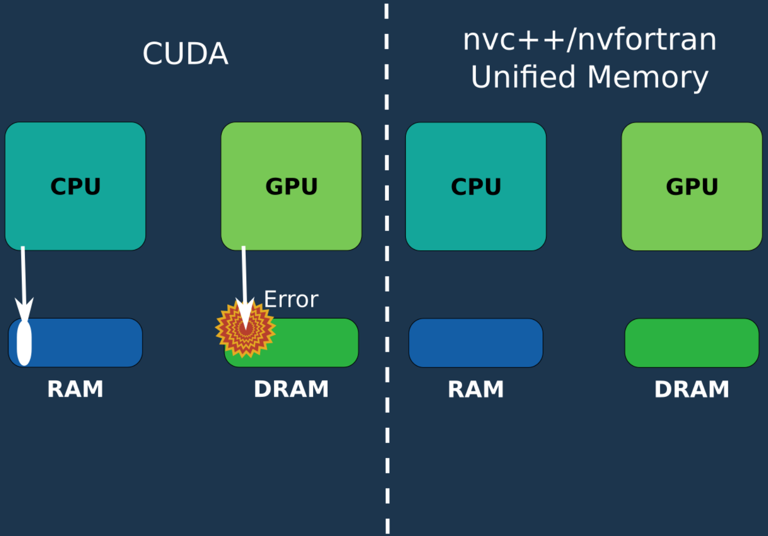


Memory Management

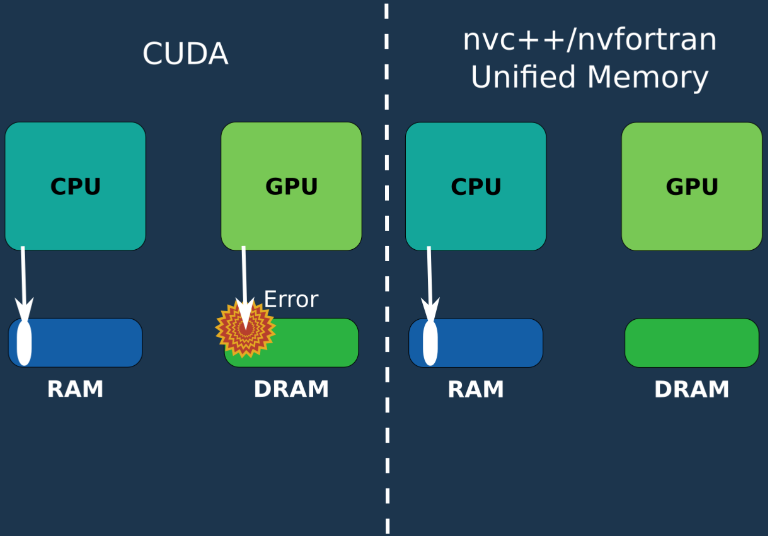
CUDA



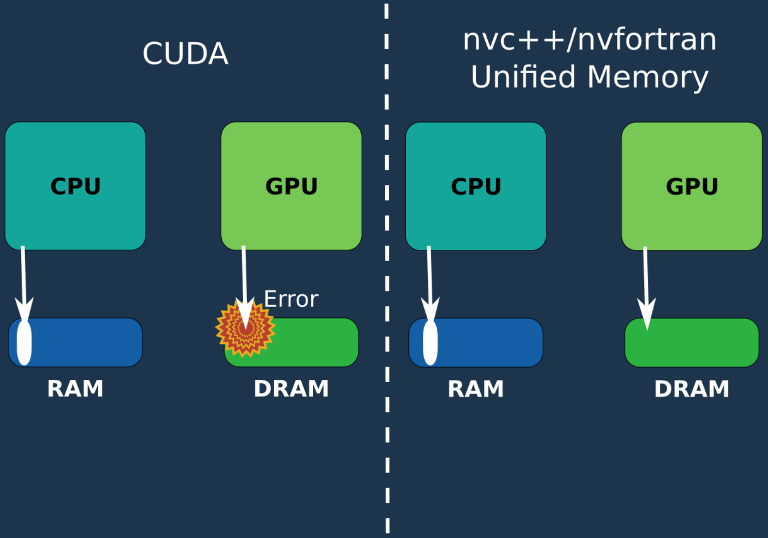
Memory Management



Memory Management



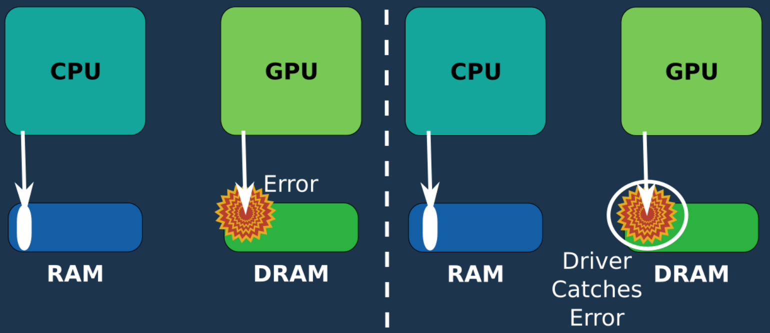
Memory Management



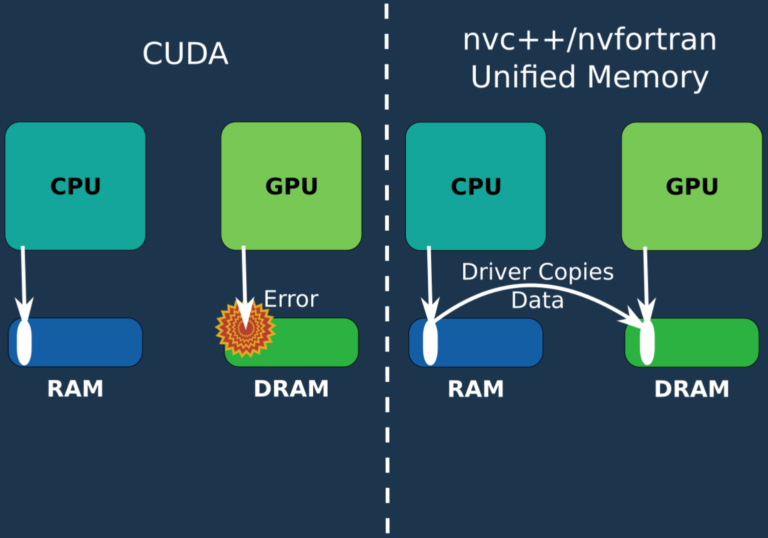
Memory Management

CUDA

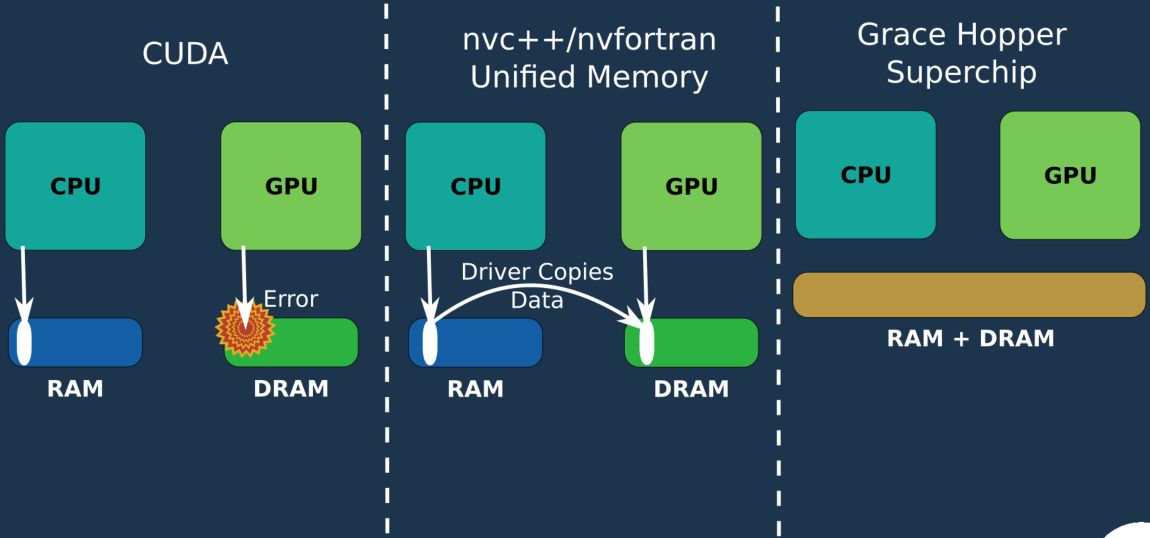
nvc++/nvfortran
Unified Memory



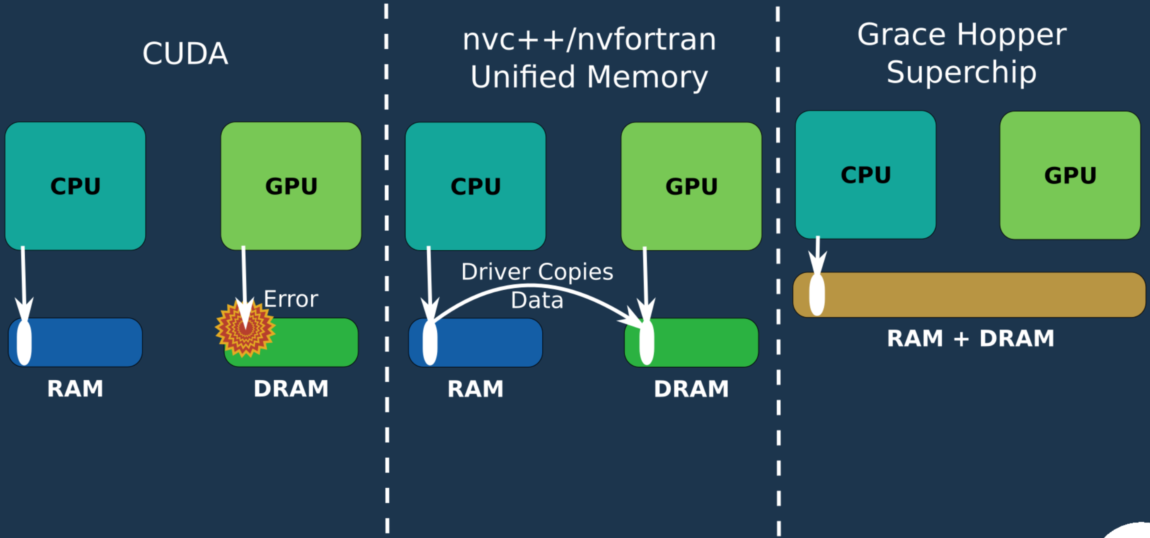
Memory Management



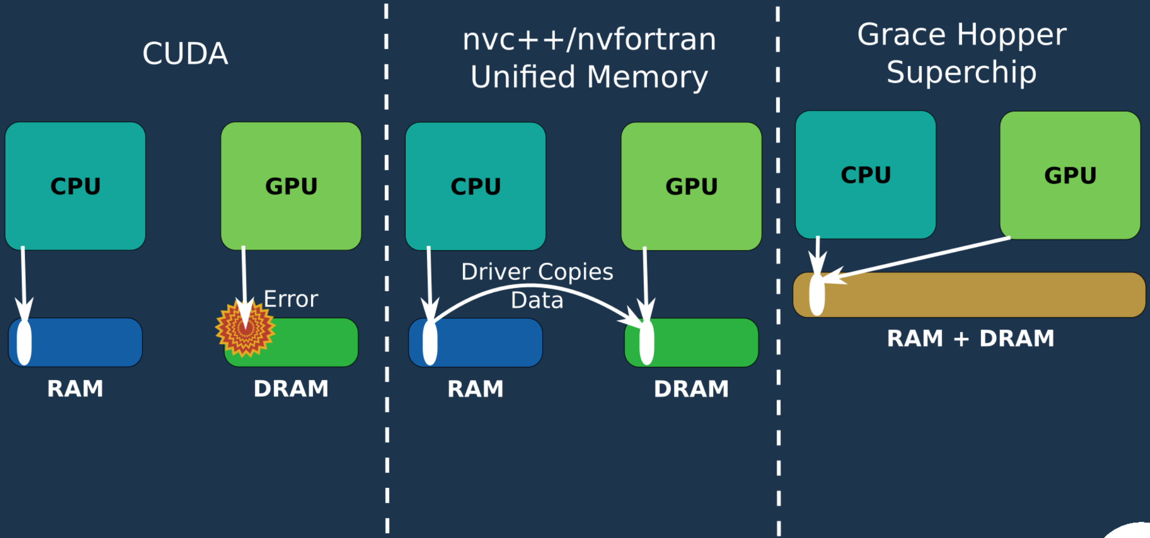
Memory Management



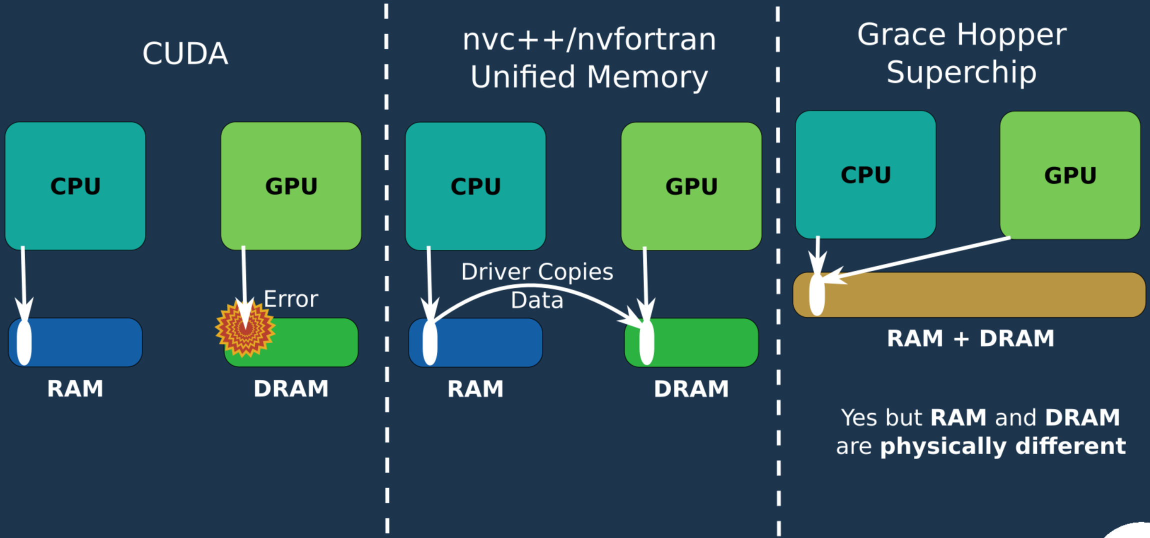
Memory Management



Memory Management

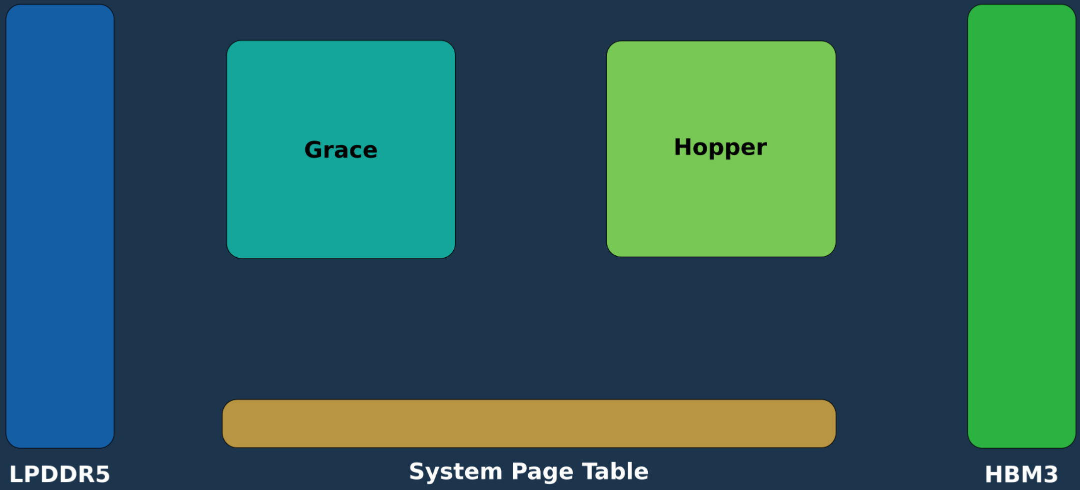


Memory Management

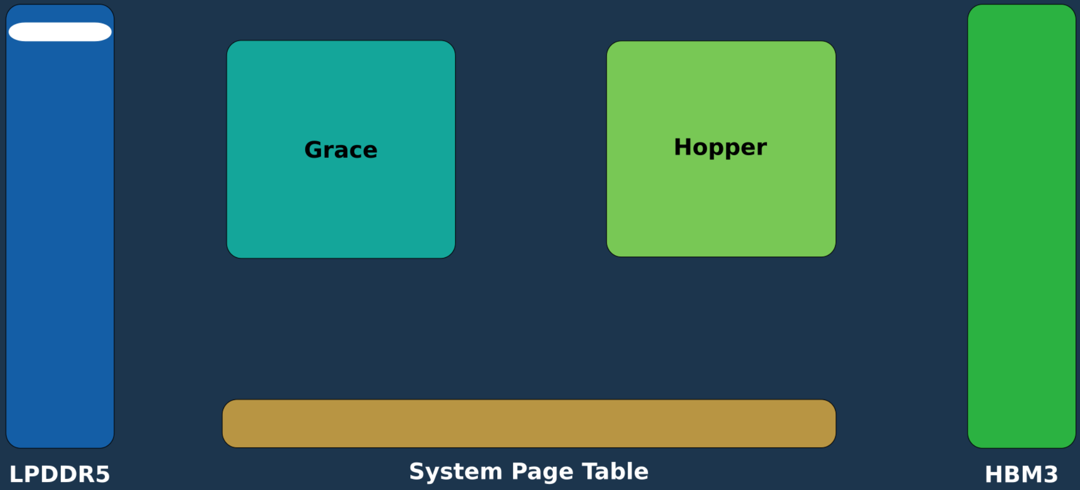


Yes but **RAM** and **DRAM** are **physically different**

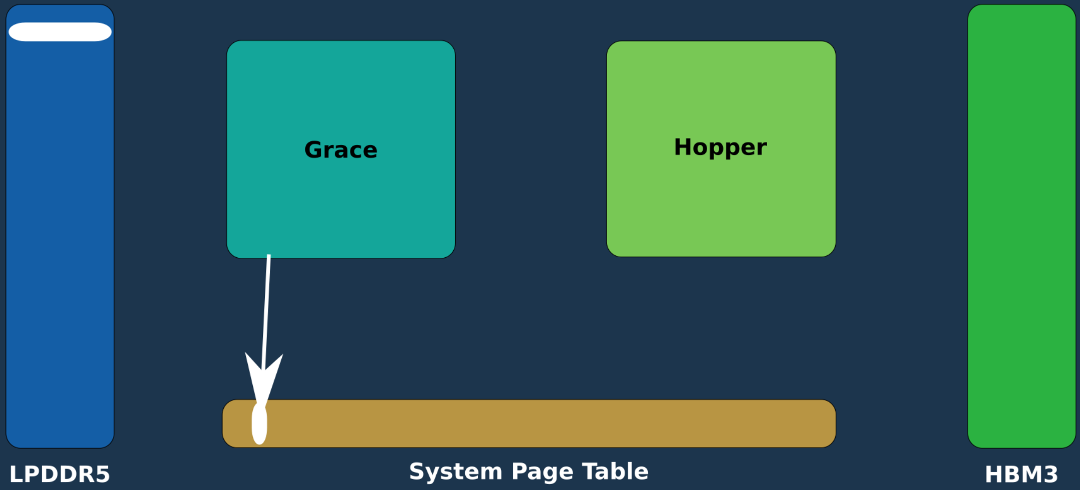
Memory Management Grace Hopper



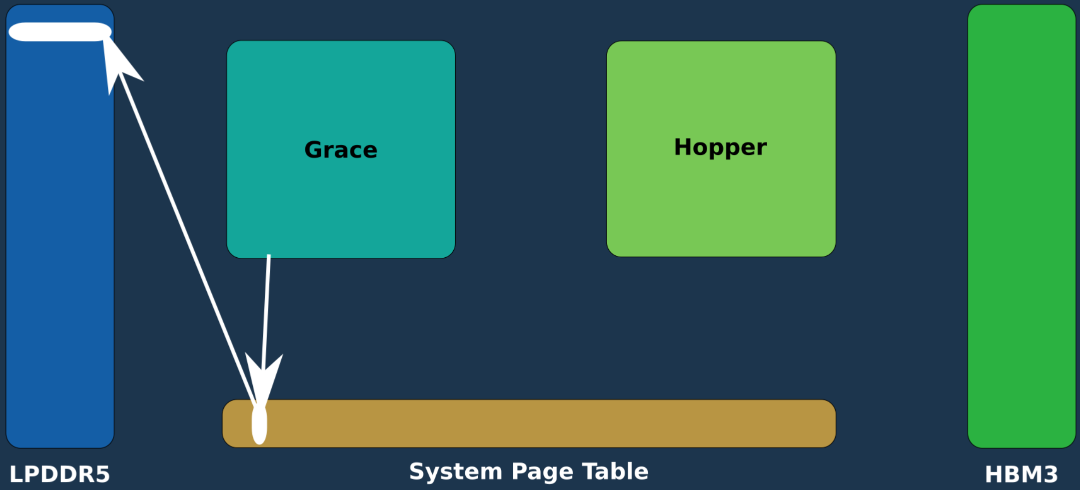
Memory Management Grace Hopper



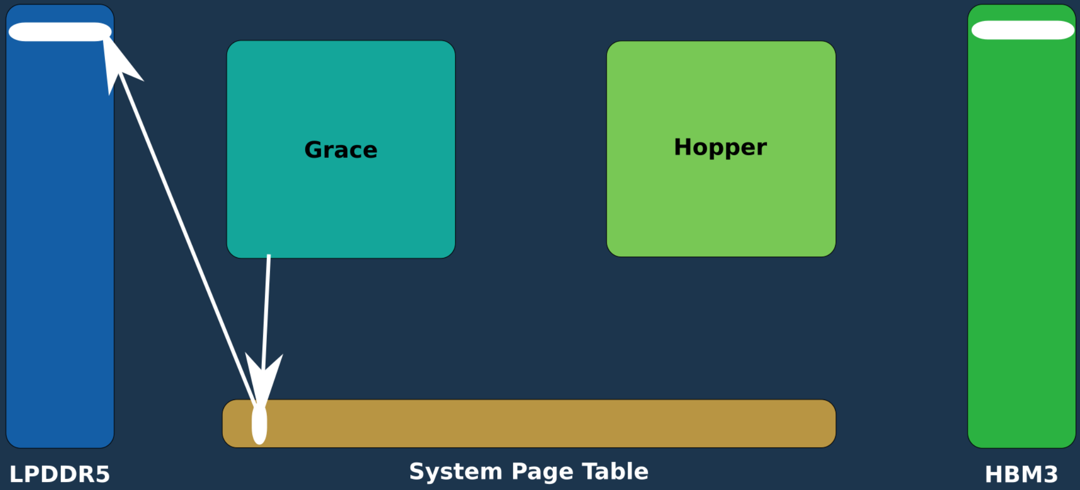
Memory Management Grace Hopper



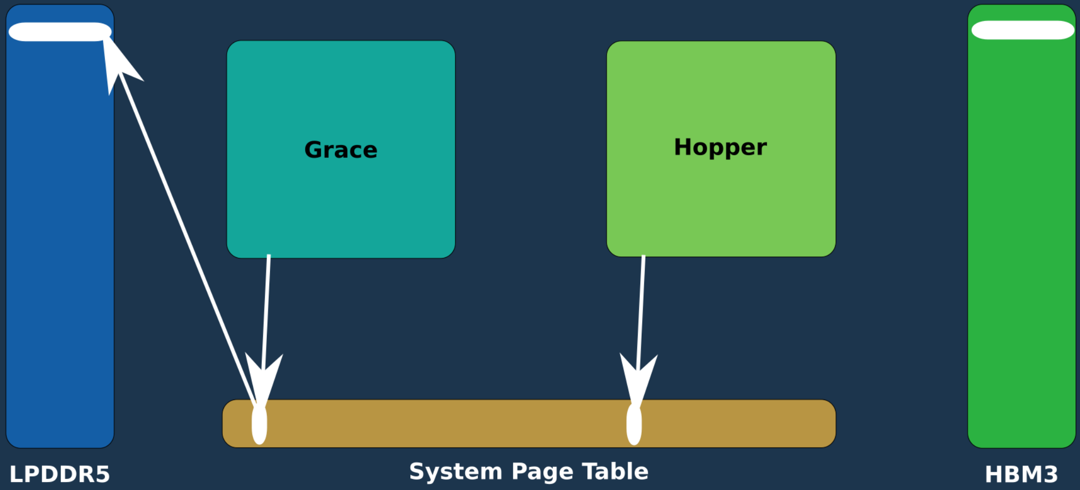
Memory Management Grace Hopper



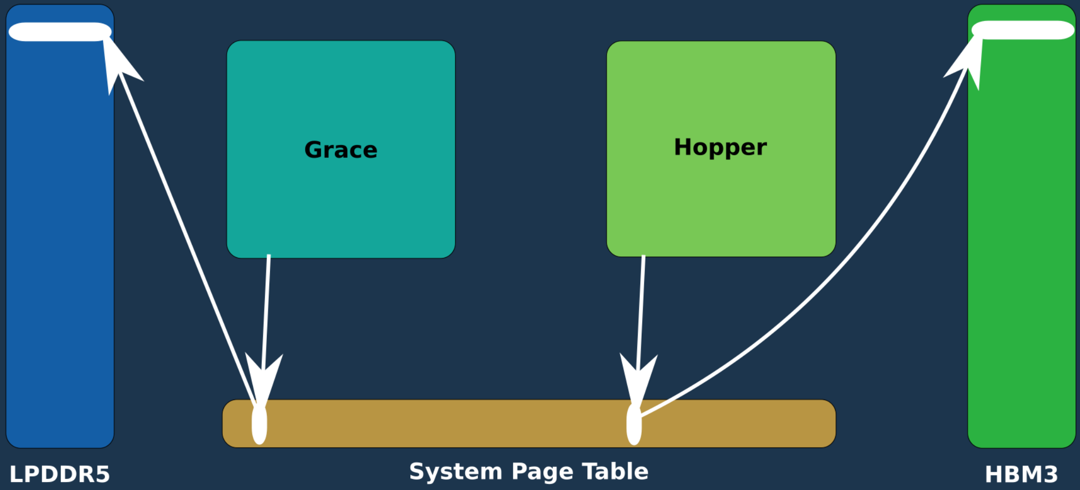
Memory Management Grace Hopper



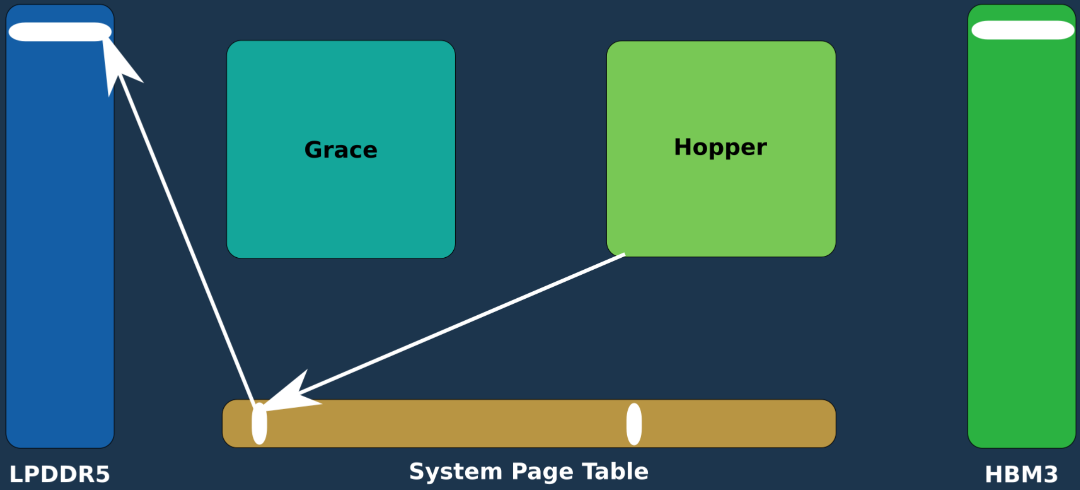
Memory Management Grace Hopper



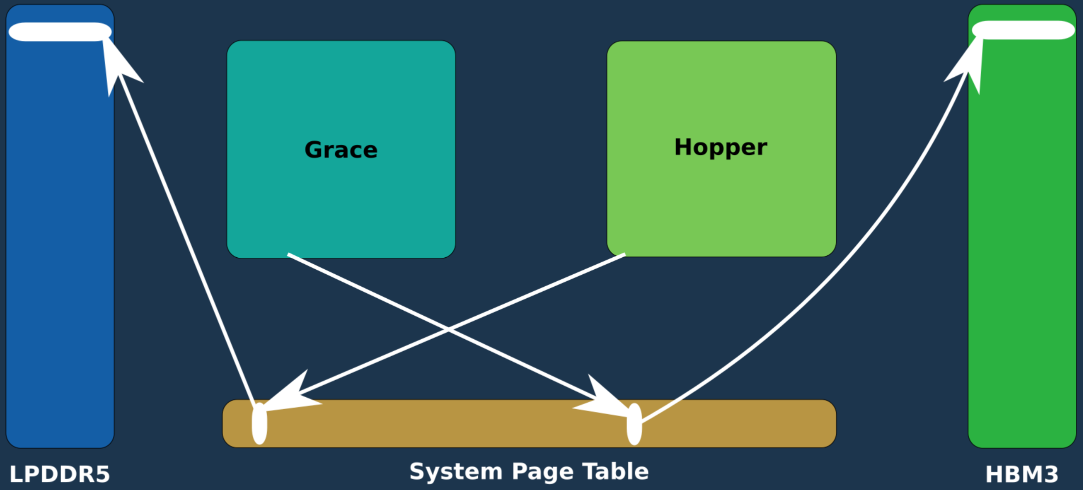
Memory Management Grace Hopper



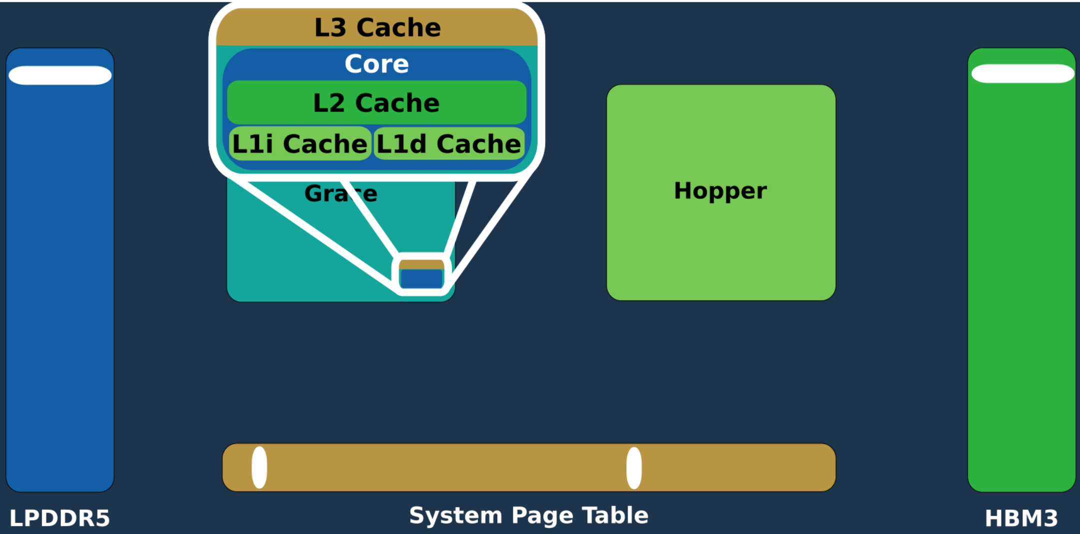
Memory Management Grace Hopper



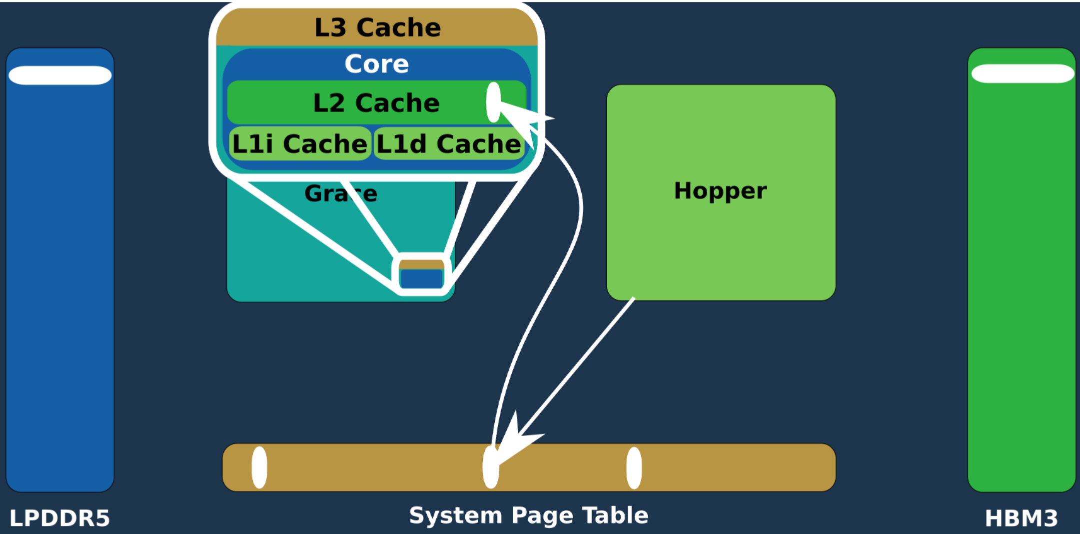
Memory Management Grace Hopper



Memory Management Grace Hopper

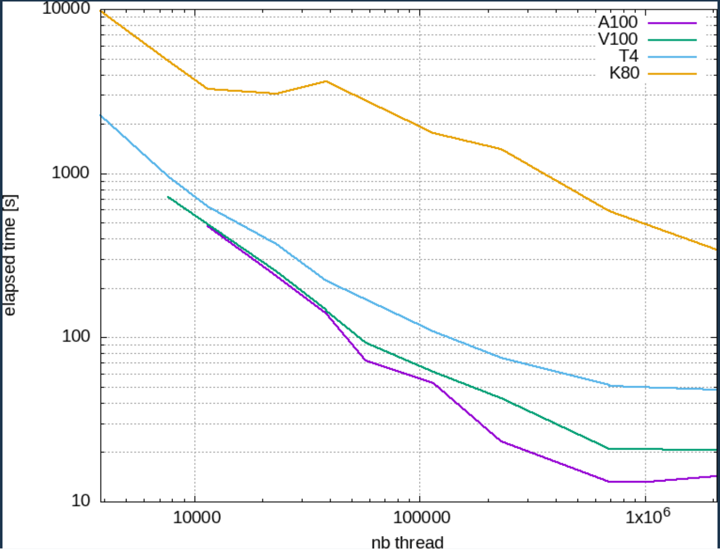


Memory Management Grace Hopper

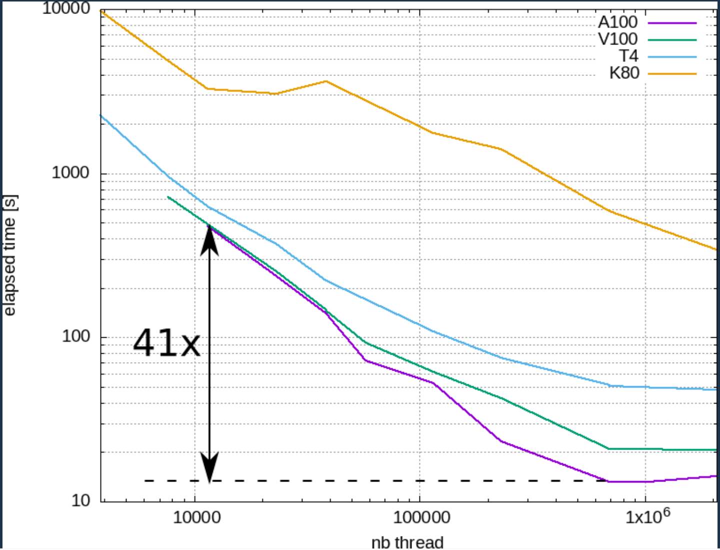


How many threads ?

How many threads ?



How many threads ?



Express Parallelism

Express Parallelism

Element

Express Parallelism

Element

Independent Elements => Independent Computing => Parallelism

Express Parallelism

Element

Independent Elements => Independent Computing => Parallelism

Particle

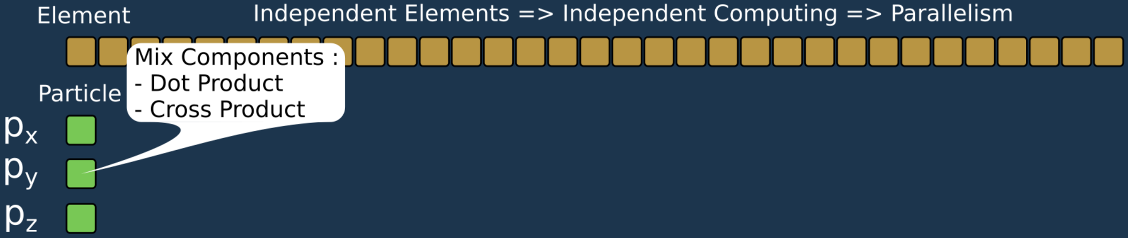
Express Parallelism

Element

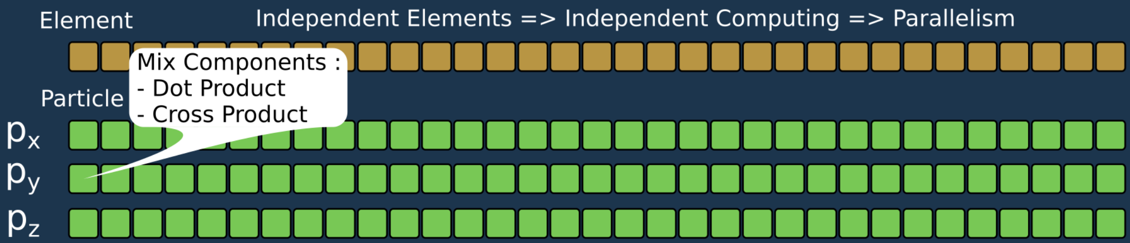
Independent Elements => Independent Computing => Parallelism

Express Parallelism

Independent Elements => Independent Computing => Parallelism

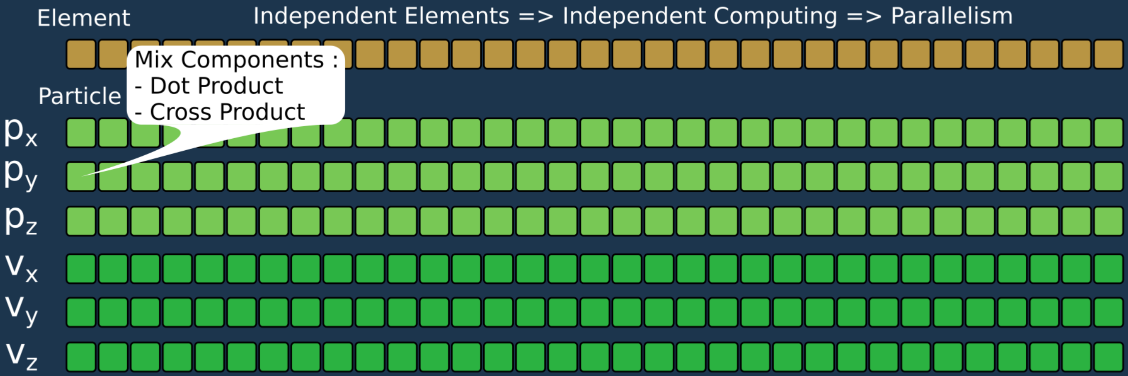


Express Parallelism



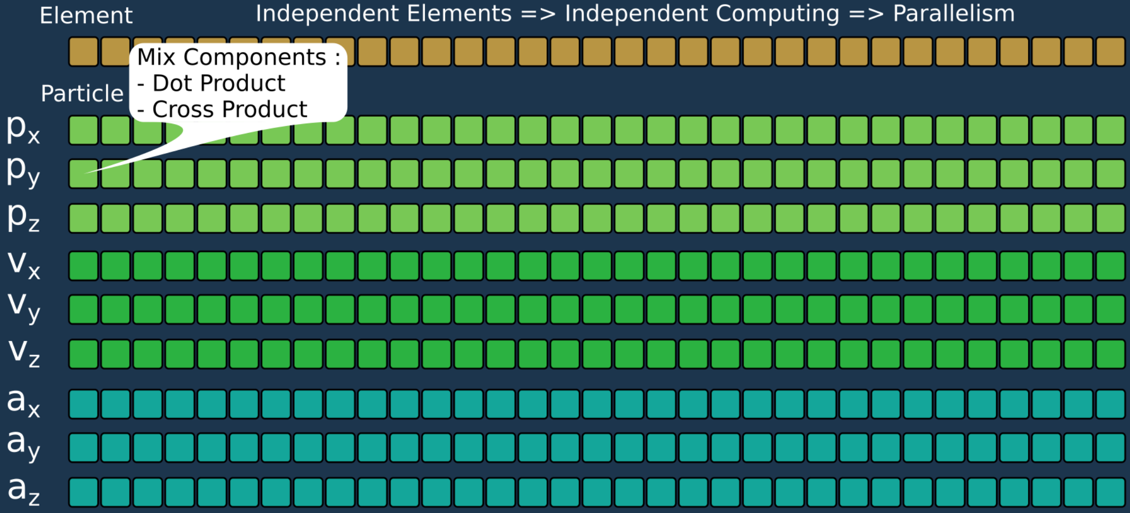
Express Parallelism

Independent Elements => Independent Computing => Parallelism

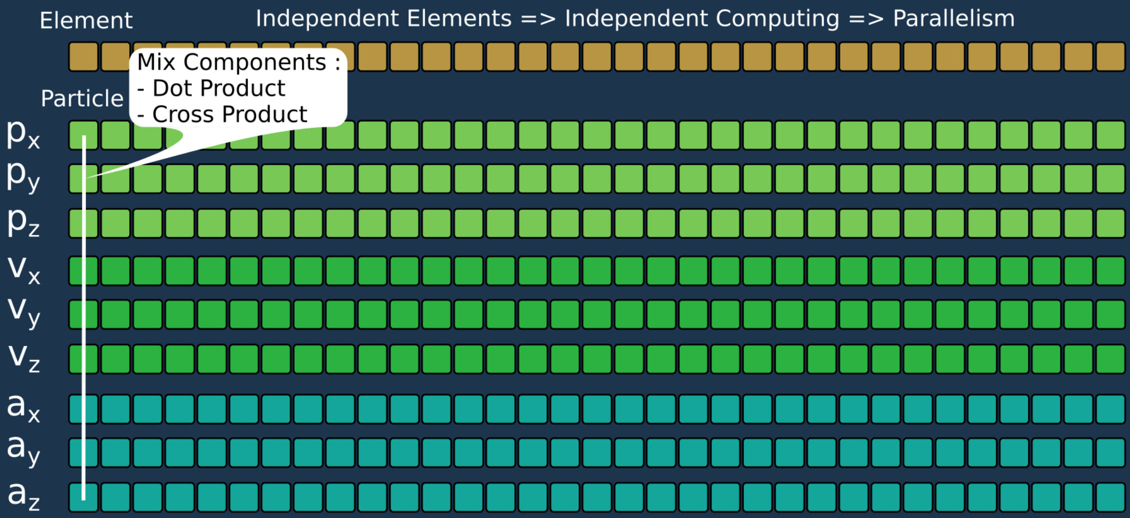


Express Parallelism

Independent Elements => Independent Computing => Parallelism

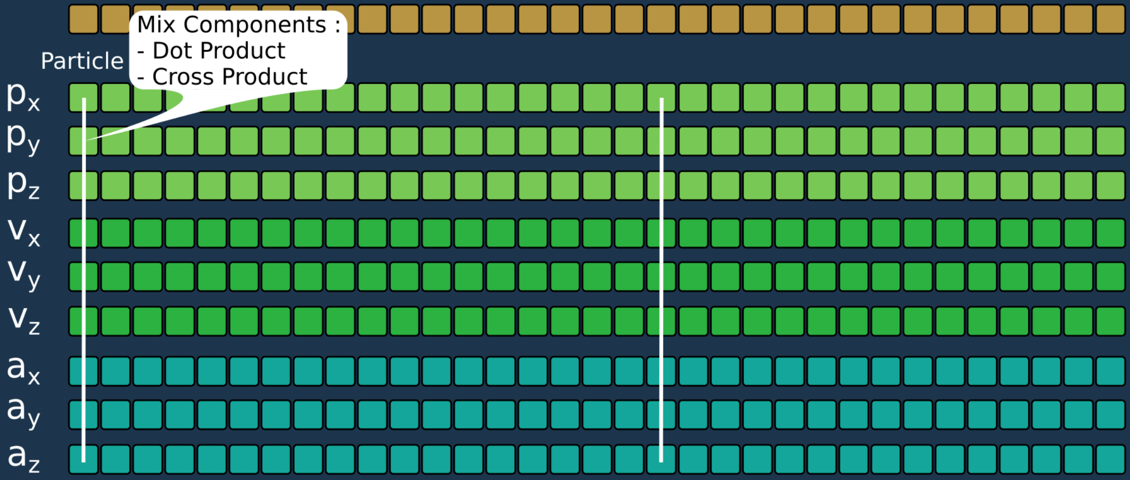


Express Parallelism

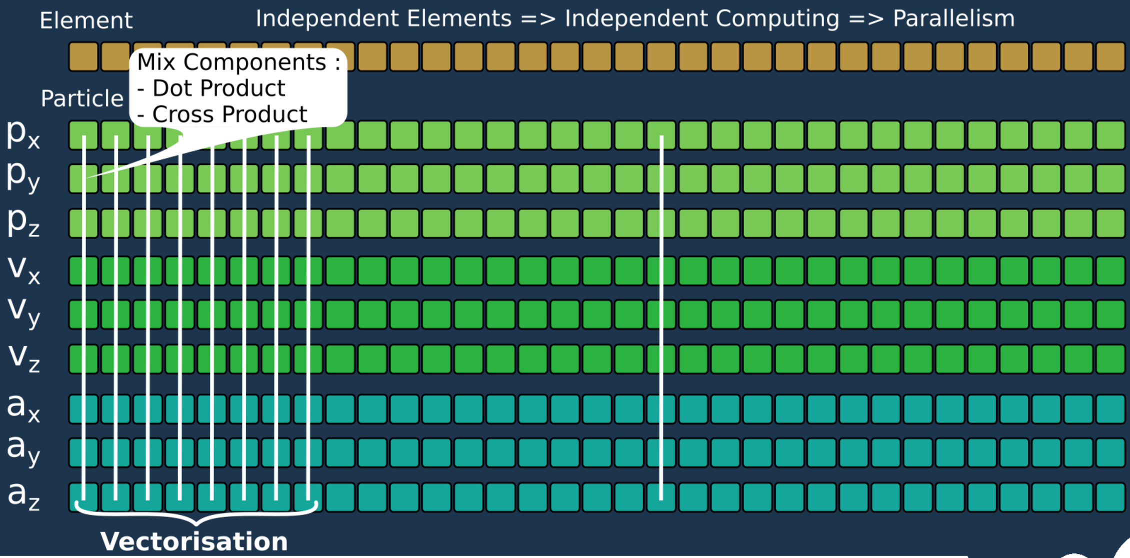


Express Parallelism

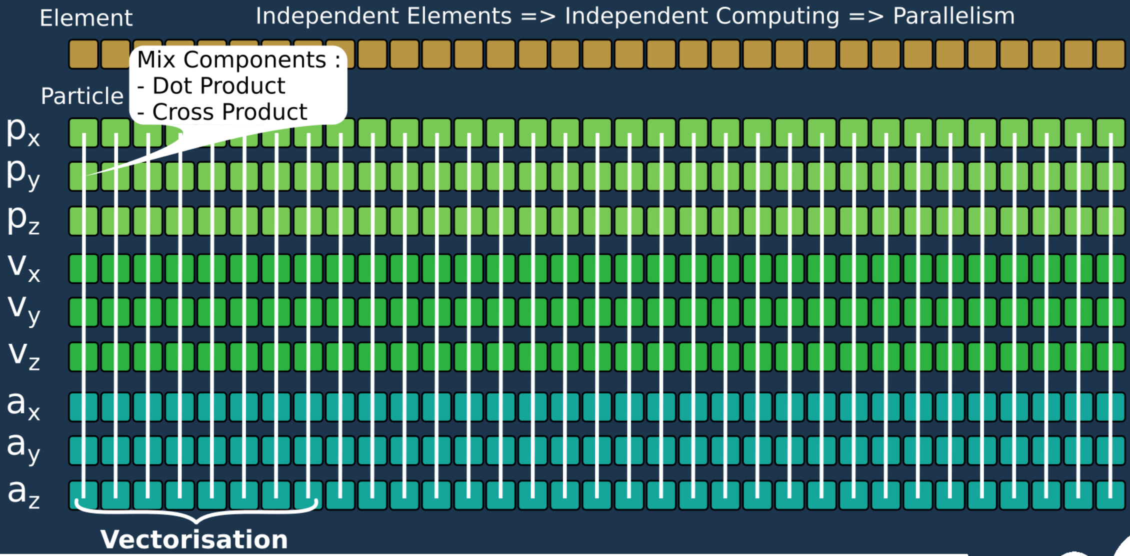
Element Independent Elements => Independent Computing => Parallelism



Express Parallelism



Express Parallelism



Express Parallelism

Express Parallelism

Event

Express Parallelism

Event

Independent Events => Independent Computing => Parallelism

Express Parallelism

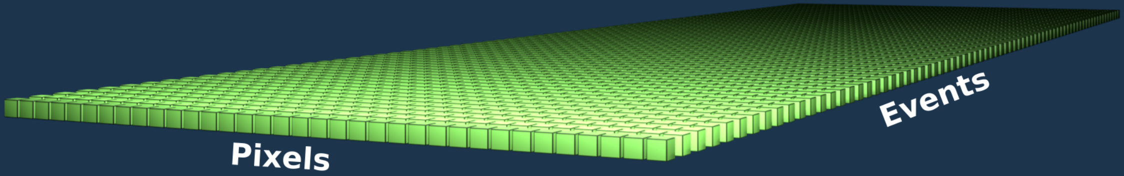
Event

Independent Events => Independent Computing => Parallelism

Express Parallelism

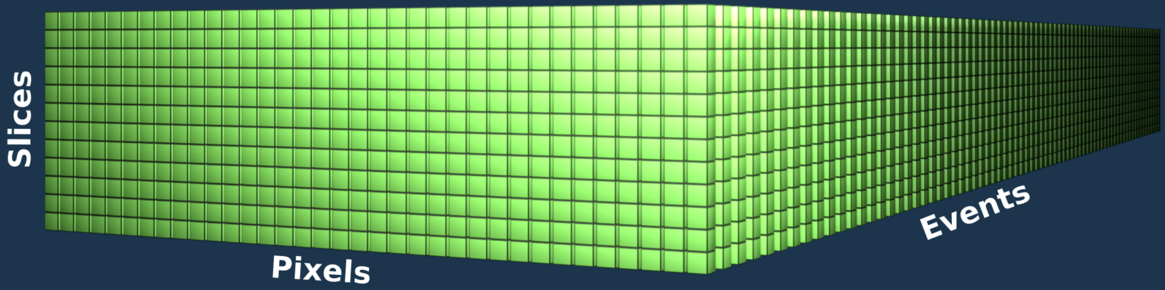
Event

Independent Events => Independent Computing => Parallelism



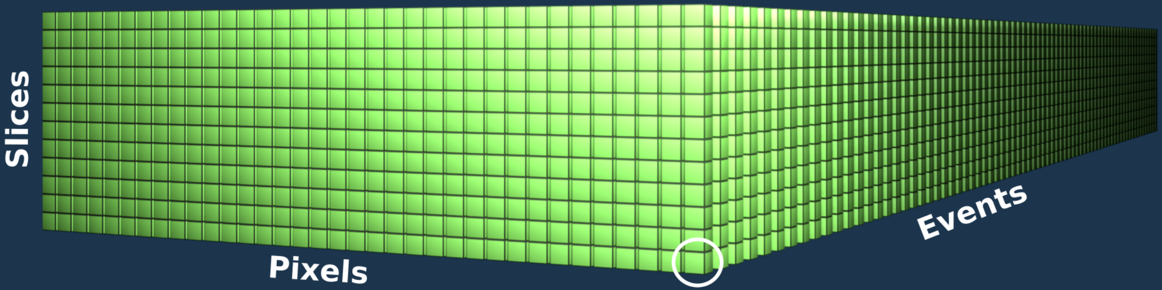
Express Parallelism

Event Independent Events => Independent Computing => Parallelism



Express Parallelism

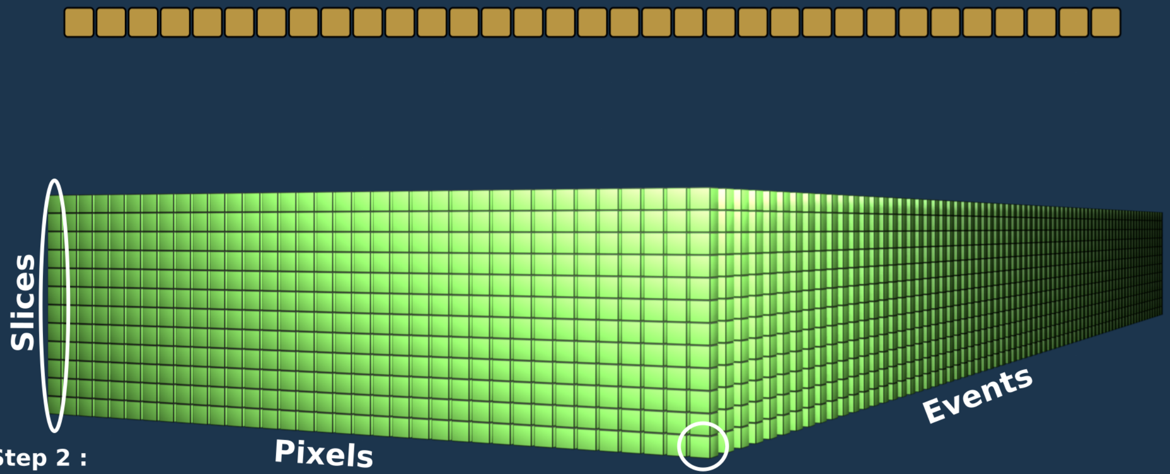
Event Independent Events => Independent Computing => Parallelism



Step 1 : Independent Computing in Calibration

Express Parallelism

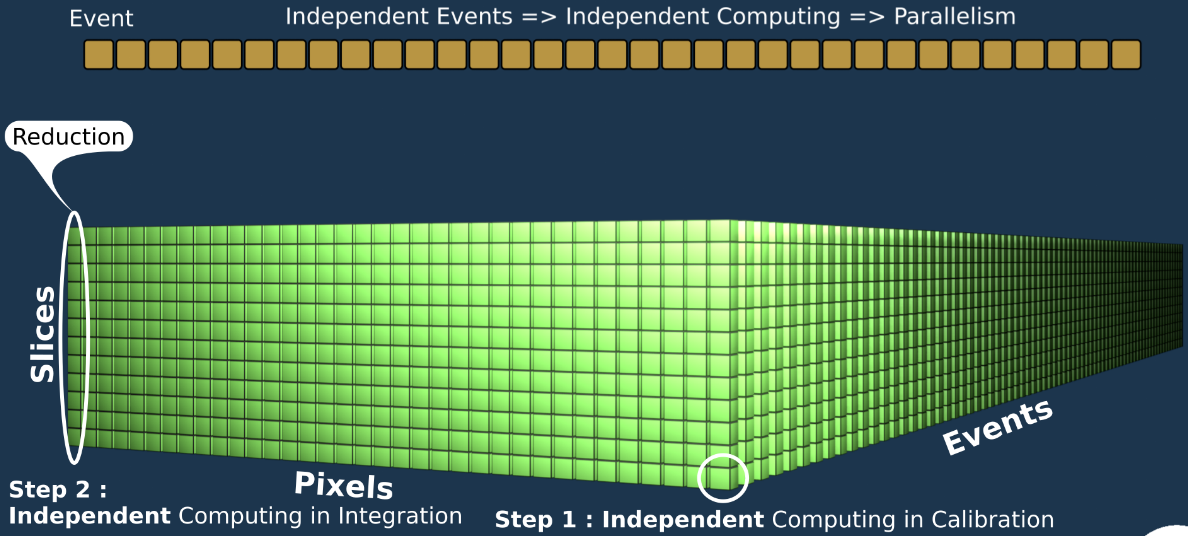
Event — Independent Events => Independent Computing => Parallelism



Step 2 :
Independent Computing in Integration

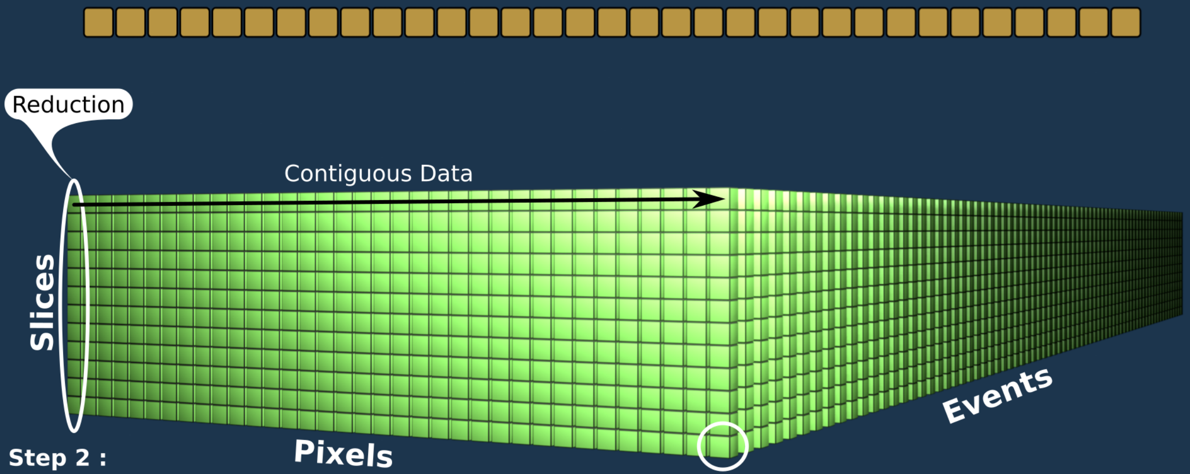
Step 1 : Independent Computing in Calibration

Express Parallelism



Express Parallelism

Event Independent Events => Independent Computing => Parallelism

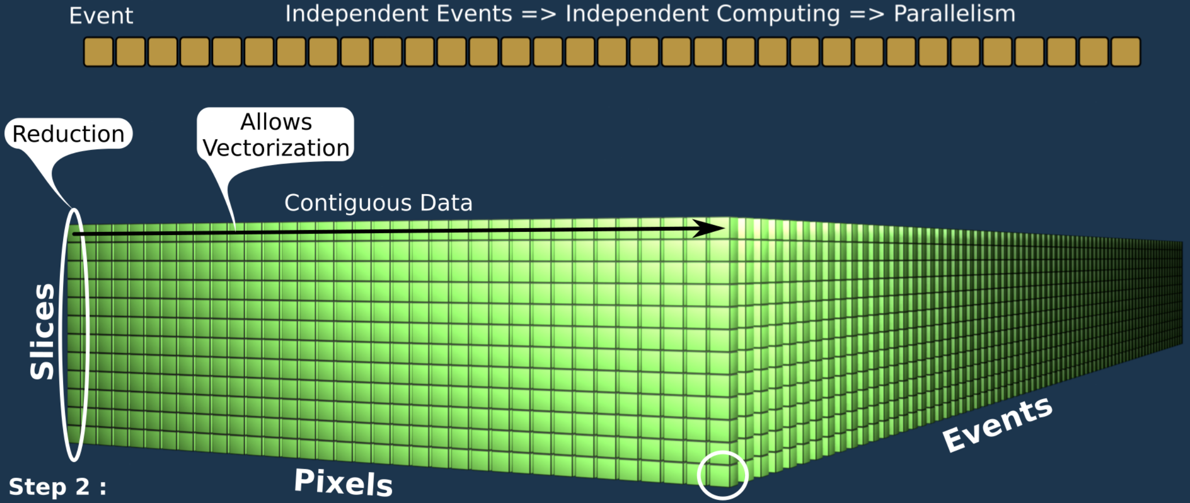


Step 2 :
Independent Computing in Integration

Step 1 : Independent Computing in Calibration

Express Parallelism

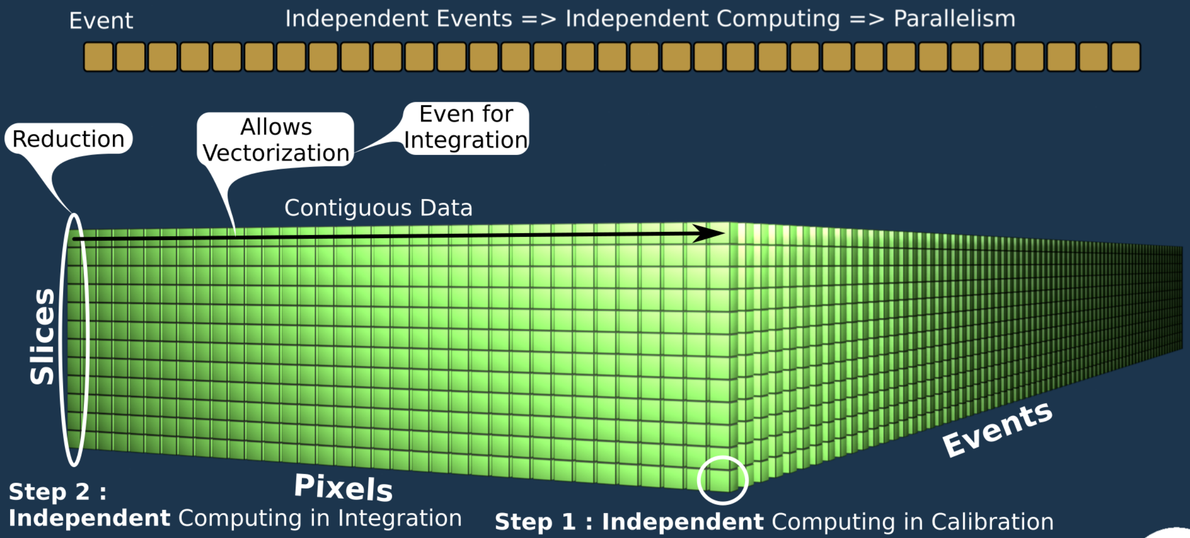
Independent Events => Independent Computing => Parallelism



Step 2 : Independent Computing in Integration

Step 1 : Independent Computing in Calibration

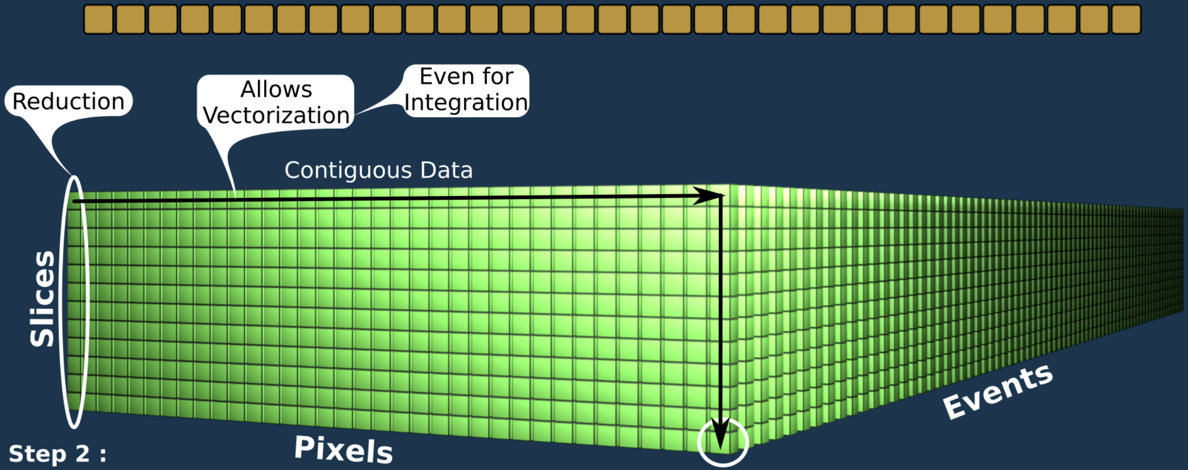
Express Parallelism



Express Parallelism

Event

Independent Events => Independent Computing => Parallelism



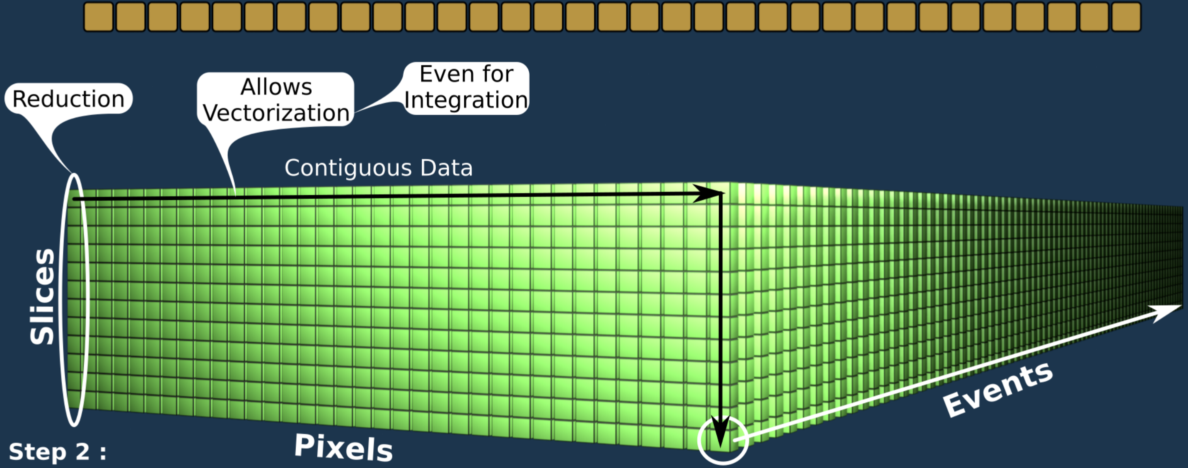
Step 2 :
Independent Computing in Integration

Step 1 : Independent Computing in Calibration

Express Parallelism

Event

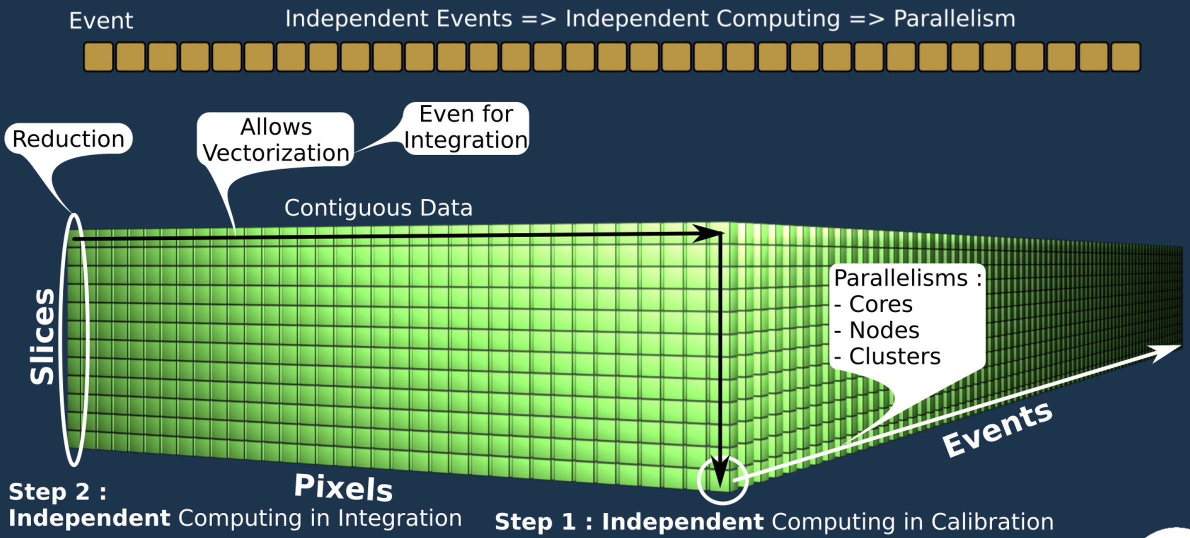
Independent Events => Independent Computing => Parallelism



Step 2 : Independent Computing in Integration

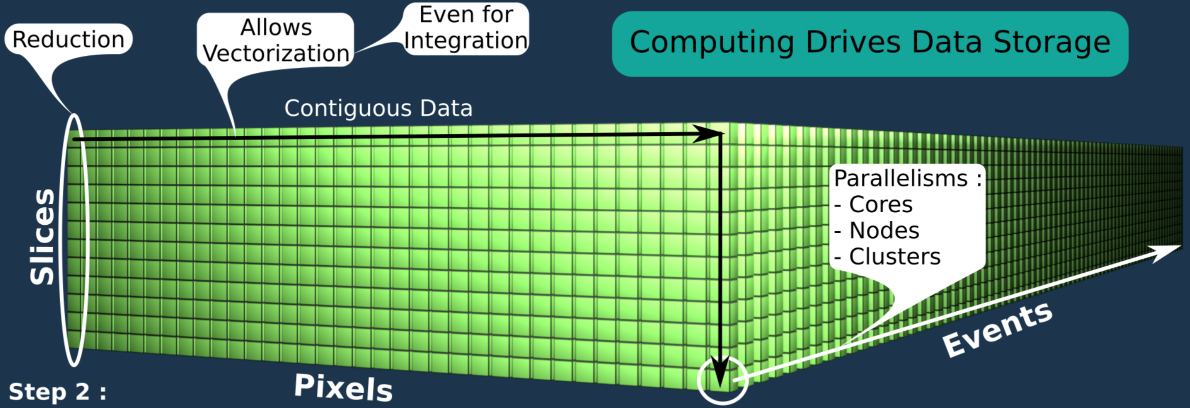
Step 1 : Independent Computing in Calibration

Express Parallelism



Express Parallelism

Independent Events => Independent Computing => Parallelism



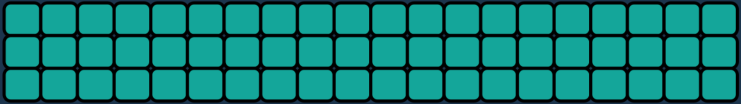
Step 2 :
Independent Computing in Integration

Step 1 : Independent Computing in Calibration

Computing Backend

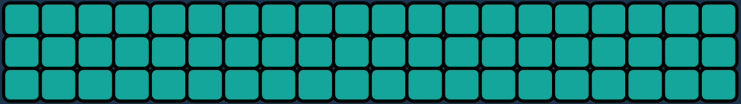
Computing Backend

Tensor



Computing Backend

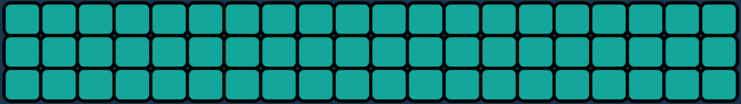
Tensor



Linear Algebra Backend

Computing Backend

Tensor



Linear Algebra Backend

CPU

Computing Backend

Tensor

Linear Algebra Backend

CPU

NUMA Node 1

NUMA Node 2

Computing Backend

Tensor

Linear Algebra Backend

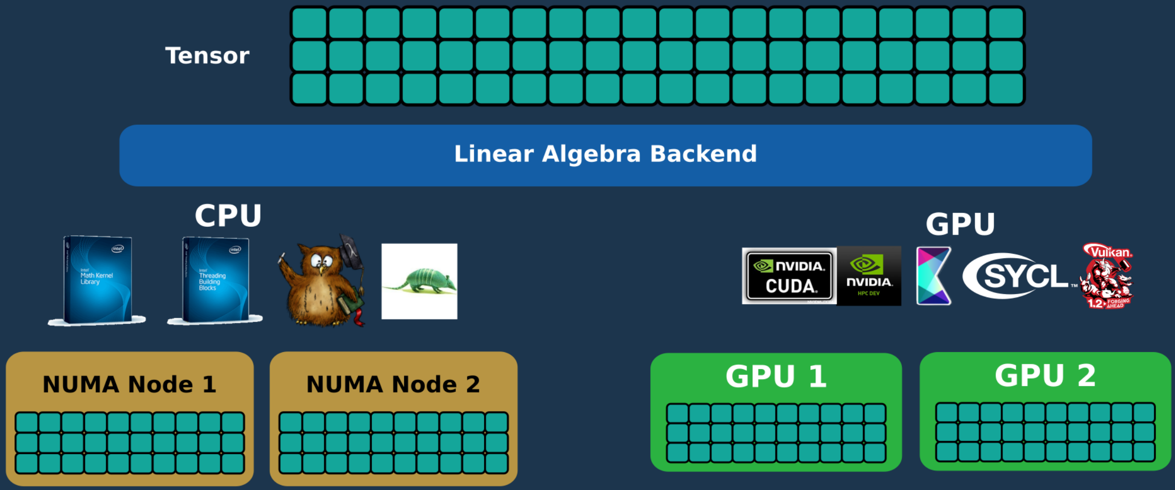
CPU

GPU

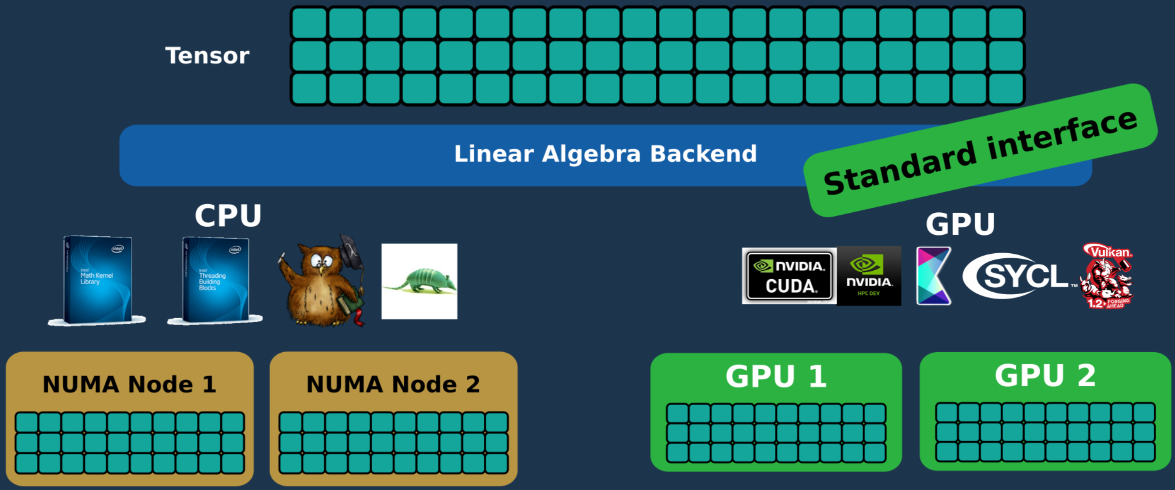
NUMA Node 1

NUMA Node 2

Computing Backend



Computing Backend



Conclusion

Conclusion

- Express **as much parallelism** as possible
(**700k** Threads for efficient **A100** computing)

Conclusion

- Express **as much parallelism** as possible
(**700k** Threads for efficient **A100** computing)

- Use **Vector**, **Matrix** and **Tensor** to store data

Conclusion

- Express **as much parallelism** as possible
(**700k** Threads for efficient **A100** computing)

- Use **Vector**, **Matrix** and **Tensor** to store data

- Use **Linear Algebra** computation :
    - **Dot product**
    - **Matrix Multiplication** (GEMM)
    - **Saxpy**
    - **Hadamard Product**

Conclusion

- Express **as much parallelism** as possible (**700k** Threads for efficient **A100** computing)

- Use **Vector**, **Matrix** and **Tensor** to store data

- Use **Linear Algebra** computation :
  - **Dot product**
  - **Matrix Multiplication** (GEMM)
  - **Saxpy**
  - **Hadamard Product**

Standard interface

Conclusion

- Express **as much parallelism** as possible
(**700k** Threads for efficient **A100** computing)

- Use **Vector**, **Matrix** and **Tensor** to store data

- Use **Linear Algebra** computation :
    - **Dot product**
    - **Matrix Multiplication** (GEMM)
    - **Saxpy**
    - **Hadamard Product**

Standard interface

Efficient on **multi cores** (**CPU**)

Conclusion

- Express **as much parallelism** as possible
(**700k** Threads for efficient **A100** computing)

- Use **Vector**, **Matrix** and **Tensor** to store data

- Use **Linear Algebra** computation :
  - **Dot product**
  - **Matrix Multiplication** (GEMM)
  - **Saxpy**
  - **Hadamard Product**

Standard interface

Efficient on **multi cores** (**CPU**)

Efficient on **many cores** (**GPU**)

NVIDIA CUDA

NVIDIA HPC DEV

SYCL

Vulkan 1.2 FORGING AHEAD