LAPP
Laboratoire d'Annecy de Physique des Particules

Retour GTC 2023 : GPU Technical Conference

Pierre Aubert

ESCAPE
cta cherenkov telescope array
cnrs
UNIVERSITÉ SAVOIE MONT BLANC

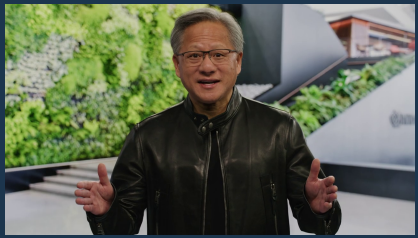
Introduction

GTC 2023 Presentation Catalog

This is not possible to summarize **525** talks !

Only focus on :

- **Hardwares** evolution
- **Compilers** evolution
- **Programming Languages** evolution
- **Data Compression** on **GPU**
- **Machine Learning** / **Deep Learning** evolution and use
- **Confidential** computing
- **GPU** consumption VS **Clock rate**
- **Photolithography**

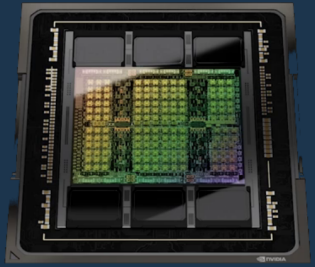


Keynotes video link

Hardware evolution

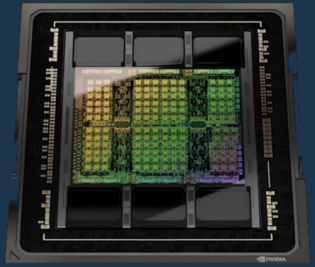
Hardware evolution

H100 (Hopper)



Hardware evolution

H100 (Hopper)

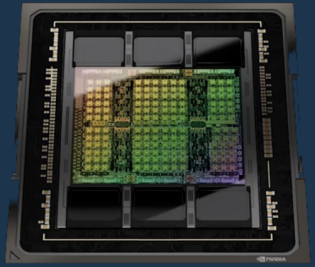


H100 SXM5 and PCie

Hardware evolution

4 nm (A100 7 nm)

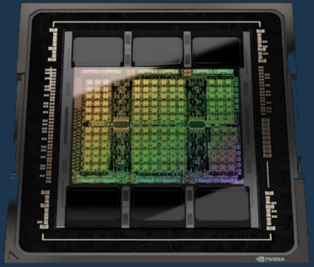
H100 (Hopper)



H100 SXM5 and PCie

Hardware evolution

H100 (Hopper)



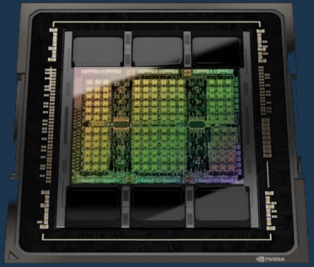
H100 SXM5 and PCie

4 nm (A100 7 nm)

Compute Capabilites 9.0

Hardware evolution

H100 (Hopper)



H100 SXM5 and PCie

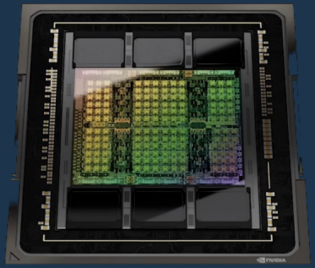
4 nm (A100 7 nm)

Compute Capabilites 9.0

132 SMs (tensor core Gen4) 8448 Cores

Hardware evolution

H100 (Hopper)



H100 SXM5 and PCie

4 nm (A100 7 nm)

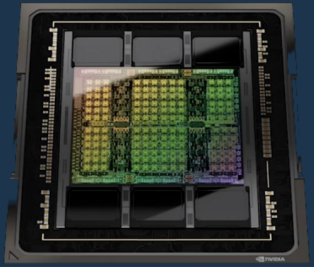
Compute Capabilites 9.0

132 SMs (tensor core Gen4) 8448 Cores

80 GB DRAM

Hardware evolution

H100 (Hopper)



H100 SXM5 and PCie

4 nm (A100 7 nm)

Compute Capabilites 9.0

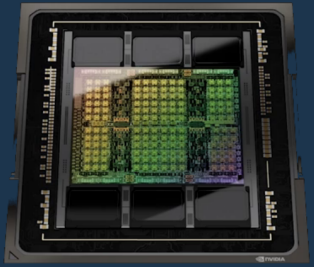
132 SMs (tensor core Gen4) 8448 Cores

80 GB DRAM

3 TB/s Bandwidth (HBM3)

Hardware evolution

H100 (Hopper)



H100 SXM5 and PCie

4 nm (A100 7 nm)

Compute Capabilites 9.0

132 SMs (tensor core Gen4) 8448 Cores

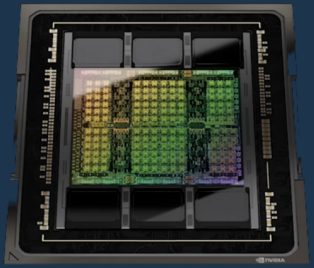
80 GB DRAM

3 TB/s Bandwidth (HBM3)

NVLink Gen 4 (900 GB/s)

Hardware evolution

H100 (Hopper)



H100 SXM5 and PCie

4 nm (A100 7 nm)

Compute Capabilites 9.0

132 SMs (tensor core Gen4) 8448 Cores

80 GB DRAM

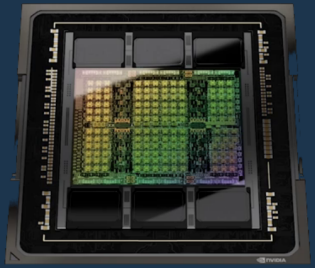
3 TB/s Bandwidth (HBM3)

NVLink Gen 4 (900 GB/s)

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

Hardware evolution

H100 (Hopper)



H100 SXM5 and PCie

4 nm (A100 7 nm)

Compute Capabilites 9.0

132 SMs (tensor core Gen4) 8448 Cores

80 GB DRAM

3 TB/s Bandwidth (HBM3)

NVLink Gen 4 (900 GB/s)

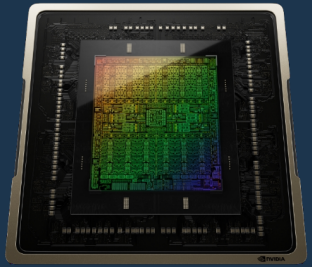
MIG Gen 2 with confidential computing
(all instances have Image/Video decoding)

A lot a computing precisions

Hardware evolution

Hardware evolution

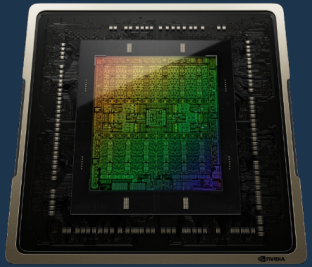
GeForce 4090
(Ada Lovelace)



Hardware evolution

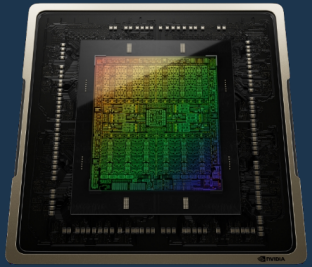
4 nm (A100 7 nm)

GeForce 4090
(Ada Lovelace)



Hardware evolution

GeForce 4090
(Ada Lovelace)

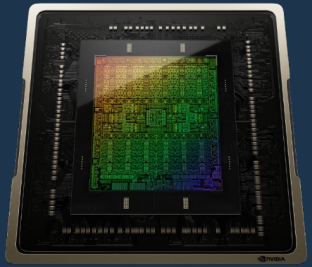


4 nm (A100 7 nm)

128 SMs (tensor core Gen4) 16384 Cuda Cores

Hardware evolution

GeForce 4090
(Ada Lovelace)



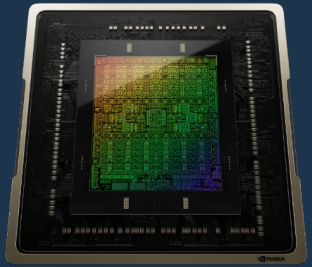
4 nm (A100 7 nm)

128 SMs (tensor core Gen4) 16384 Cuda Cores

24 GB GDDR6

Hardware evolution

GeForce 4090
(Ada Lovelace)



4 nm (A100 7 nm)

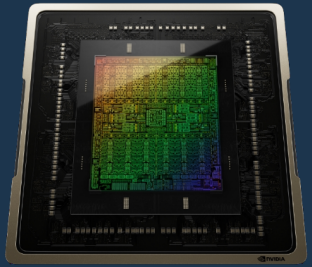
128 SMs (tensor core Gen4) 16384 Cuda Cores

24 GB GDDR6

1 TB/s Bandwidth (HBM3)

Hardware evolution

GeForce 4090
(Ada Lovelace)



4 nm (A100 7 nm)

128 SMs (tensor core Gen4) 16384 Cuda Cores

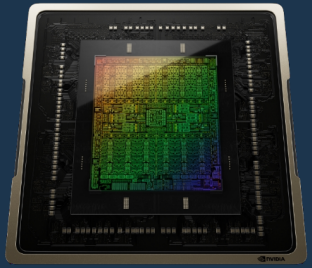
24 GB GDDR6

1 TB/s Bandwidth (HBM3)

Peak 82 TFlops (FP32 without tensor cores)

Hardware evolution

GeForce 4090
(Ada Lovelace)



4 nm (A100 7 nm)

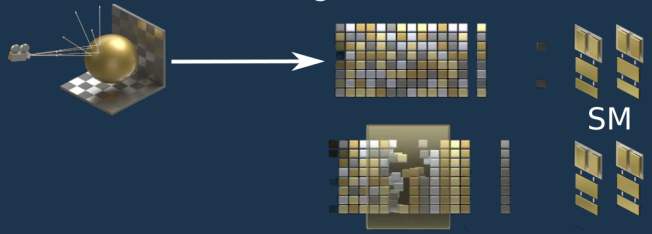
128 SMs (tensor core Gen4) 16384 Cuda Cores

24 GB GDDR6

1 TB/s Bandwidth (HBM3)

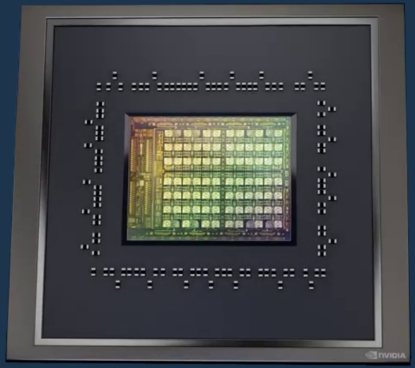
Peak 82 TFlops (FP32 without tensor cores)

Shader Reordering



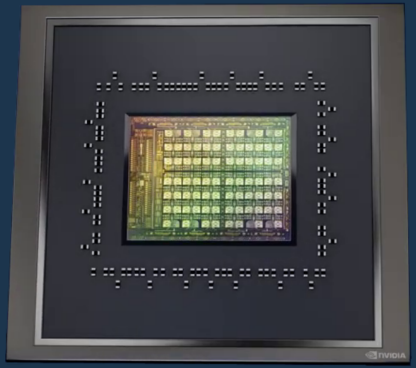
Hardware evolution

Grace



Hardware evolution

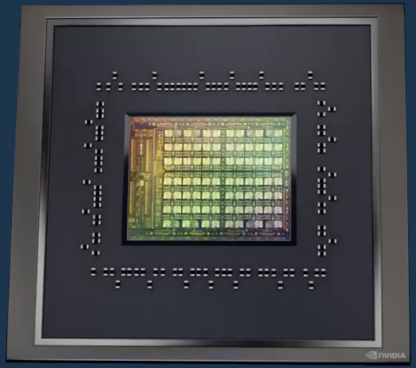
Grace



72 Neoverse V2 cores, Armv9
SVE2 with 4x128b

Hardware evolution

Grace

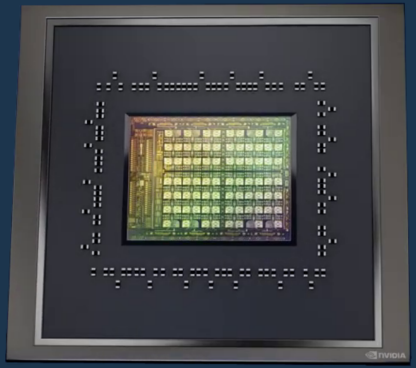


72 Neoverse V2 cores, Armv9 SVE2 with 4x128b

240 GB RAM (LPDDR5X)

Hardware evolution

Grace



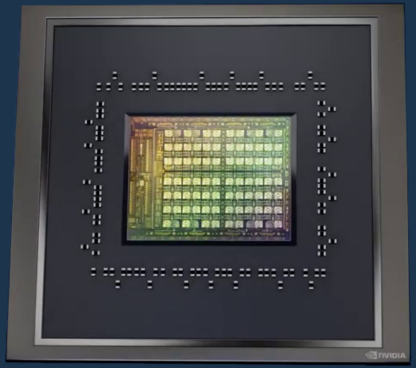
72 Neoverse V2 cores, Armv9 SVE2 with 4x128b

240 GB RAM (LPDDR5X)

3.2 TB/s Bi-section BW

Hardware evolution

Grace



72 Neoverse V2 cores, Armv9
SVE2 with 4x128b

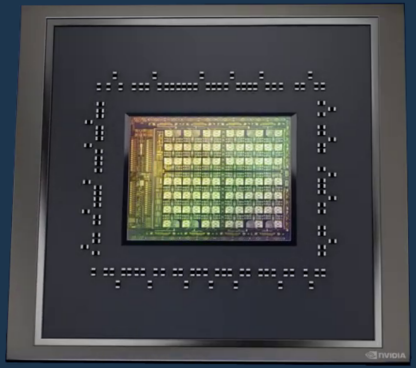
240 GB RAM (LPDDR5X)

3.2 TB/s Bi-section BW

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

Hardware evolution

Grace



72 Neoverse V2 cores, Armv9
SVE2 with 4x128b

240 GB RAM (LPDDR5X)

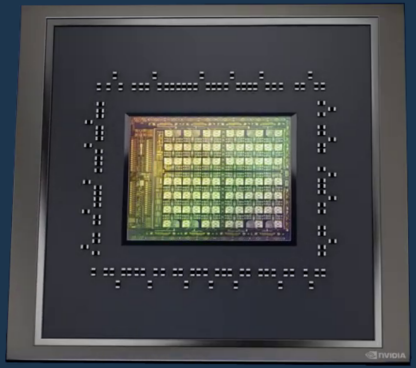
3.2 TB/s Bi-section BW

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

Hardware evolution

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)

Hardware evolution

Grace Superchip

Hardware evolution

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b SVE2

Hardware evolution

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b SVE2

144 Cores

Hardware evolution

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b SVE2

144 Cores

1 TB/s Bandwidth (HBM3)

Hardware evolution

Grace Superchip

Neoverse V2 Cores: Armv9 with 4x128b SVE2

144 Cores

1 TB/s Bandwidth (HBM3)

960 GB RAM (LPDDR5X)

Hardware evolution

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

Hardware evolution

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)

Hardware evolution

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)

Hardware evolution

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

Hardware evolution

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

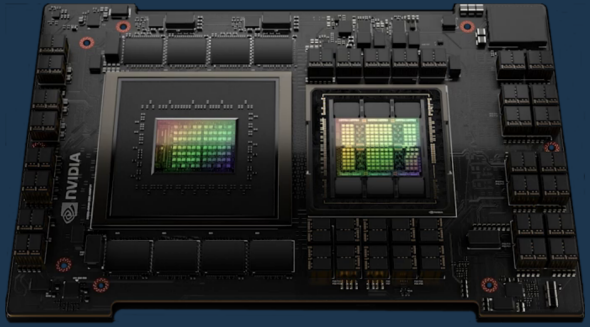
NVLink Gen 4 (900 GB/s)

7.1 TFlops (FP64 computing peak)

500 W

Hardware evolution

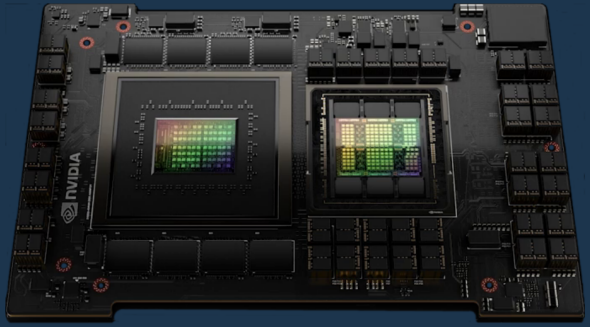
Grace Hopper Superchip



Hardware evolution

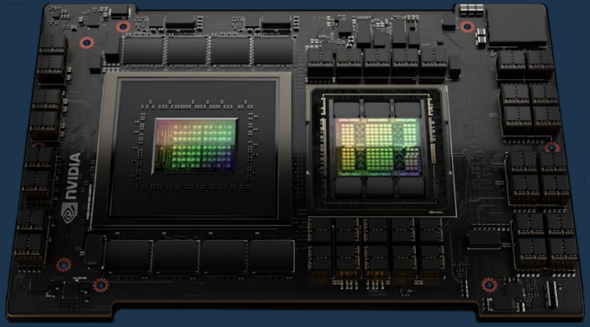
Grace Hopper Superchip

4 TB/s Bandwidth



Hardware evolution

Grace Hopper Superchip

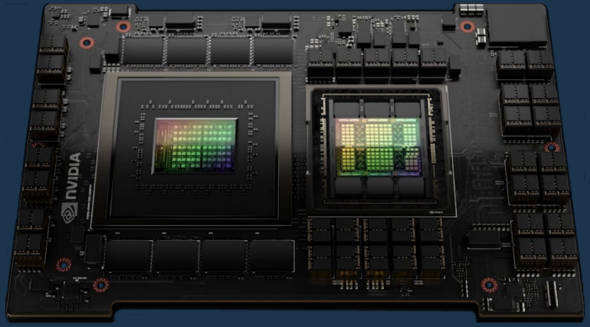


4 TB/s Bandwidth

600 GB RAM (LPDDR5X)

Hardware evolution

Grace Hopper Superchip



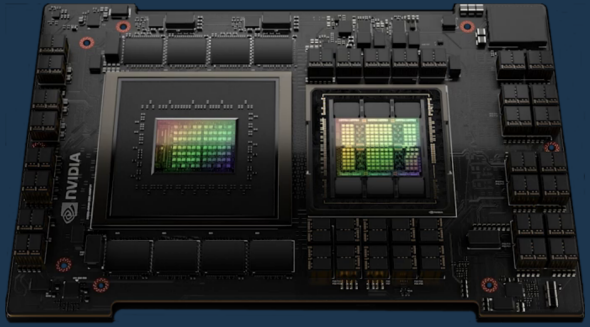
4 TB/s Bandwidth

600 GB RAM (LPDDR5X)

NVLink C2C (900 GB/s)

Hardware evolution

Grace Hopper Superchip



4 TB/s Bandwidth

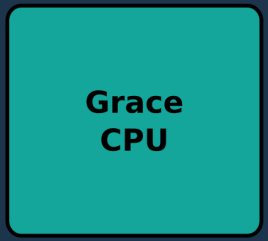
600 GB RAM (LPDDR5X)

NVLink C2C (900 GB/s)

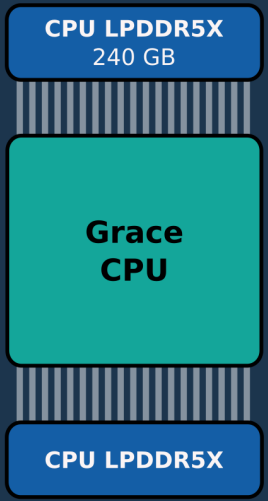
Unified Memory

Grace Hopper Superchip

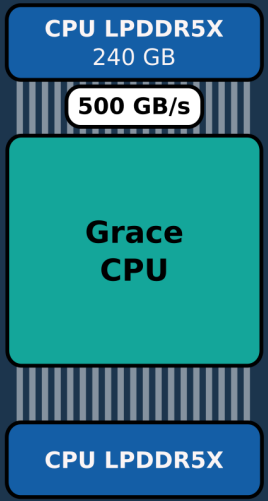
Grace Hopper Superchip



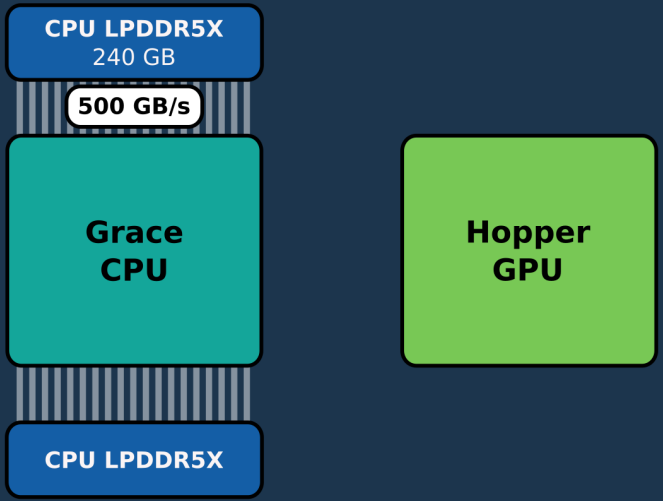
Grace Hopper Superchip



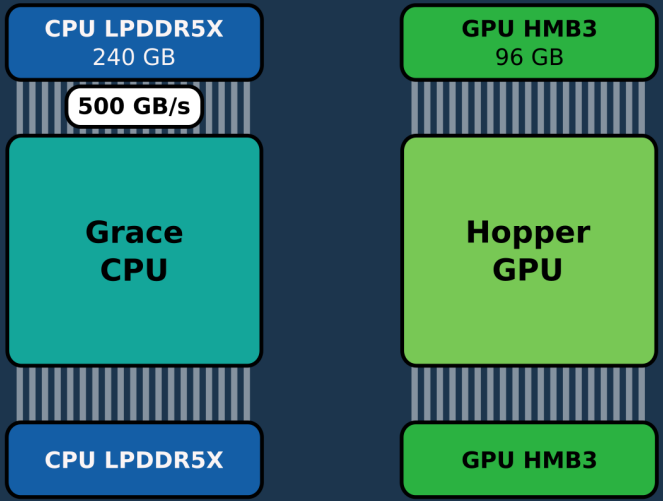
Grace Hopper Superchip



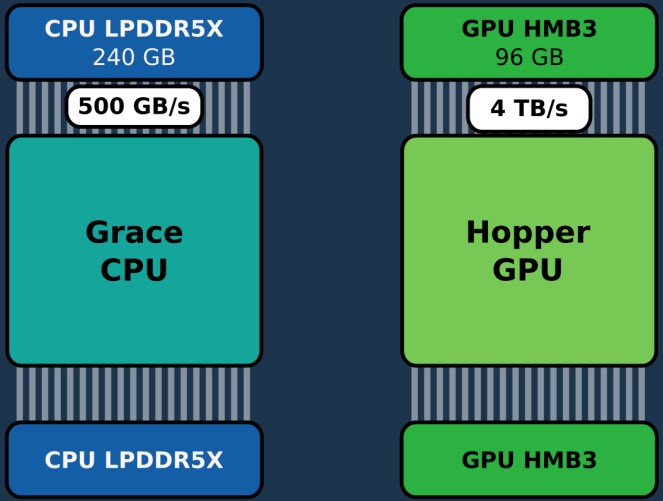
Grace Hopper Superchip



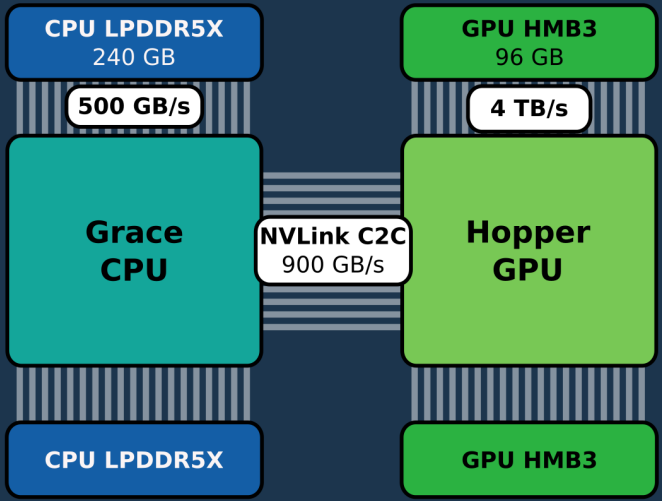
Grace Hopper Superchip



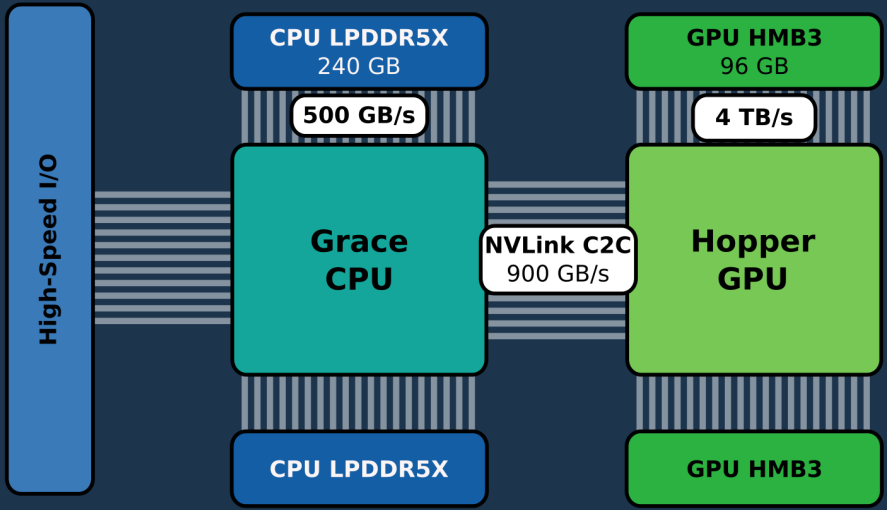
Grace Hopper Superchip



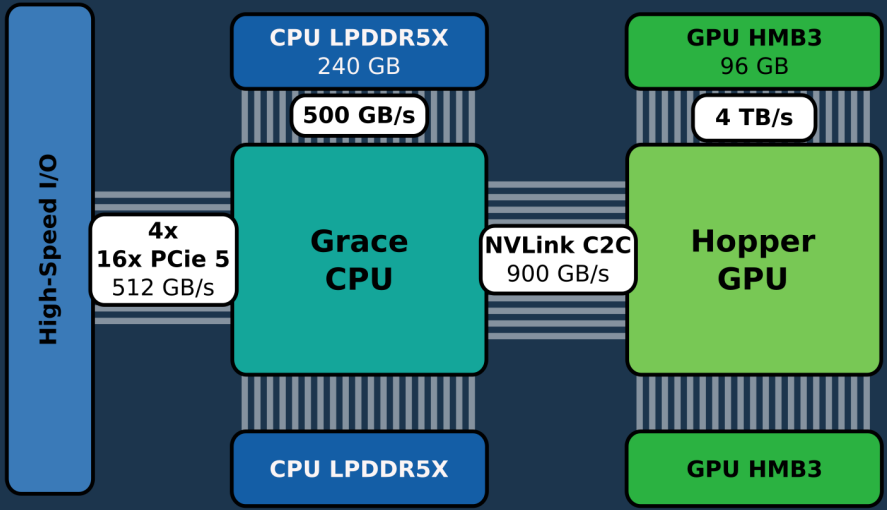
Grace Hopper Superchip



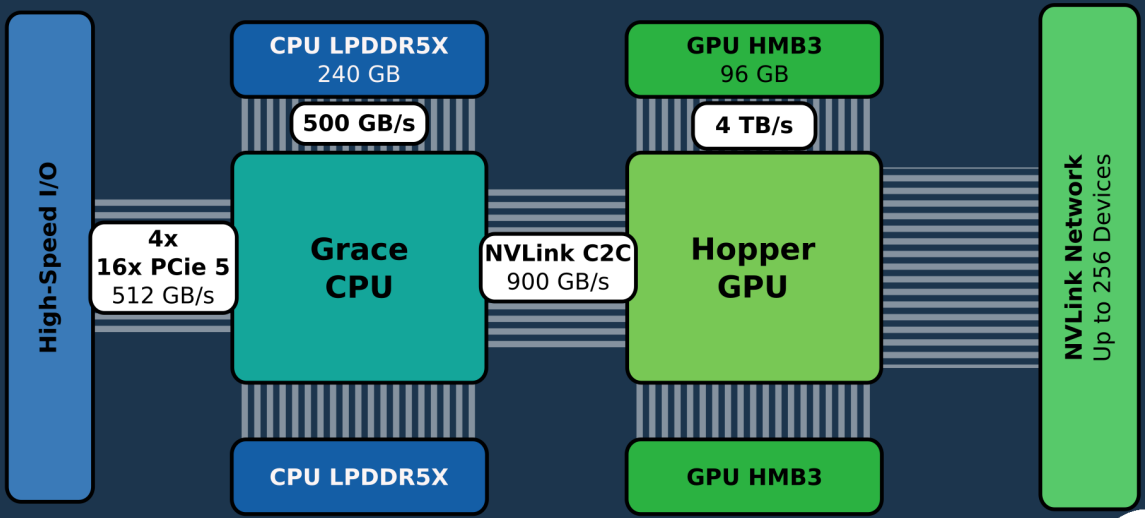
Grace Hopper Superchip



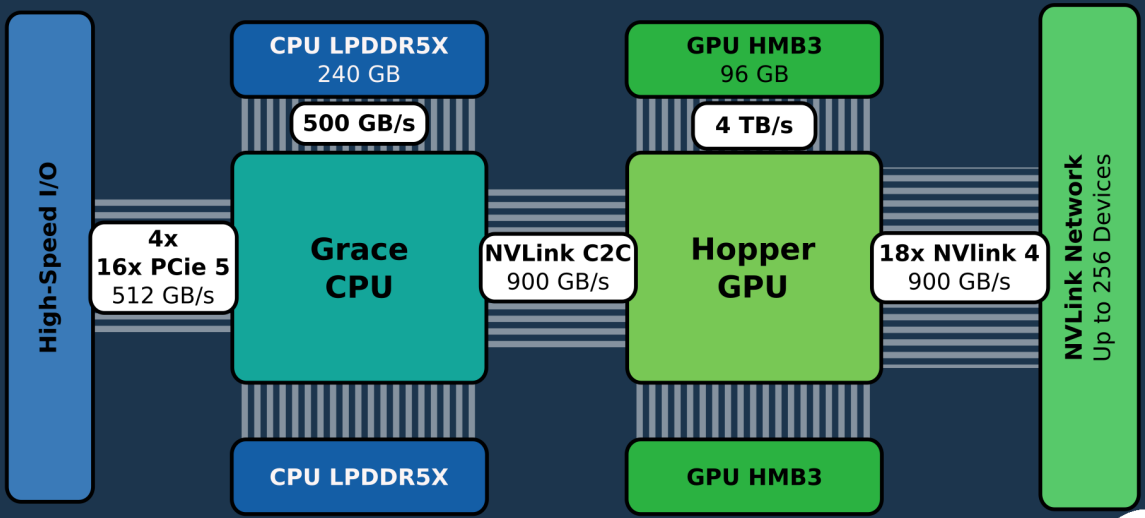
Grace Hopper Superchip



Grace Hopper Superchip



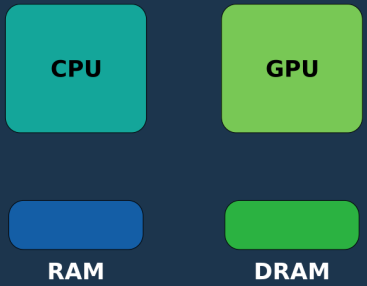
Grace Hopper Superchip



Memory Management

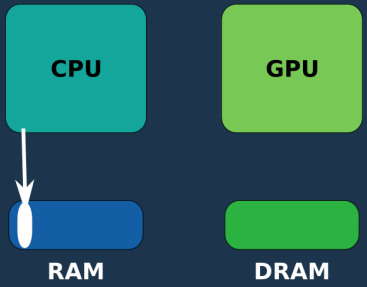
Memory Management

CUDA



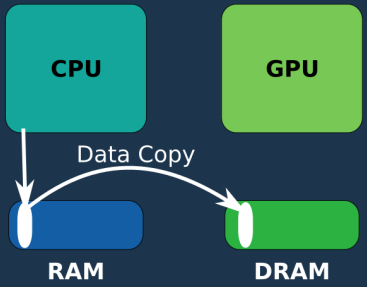
Memory Management

CUDA



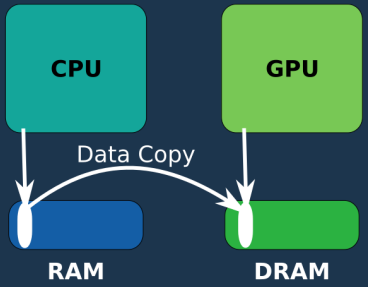
Memory Management

CUDA



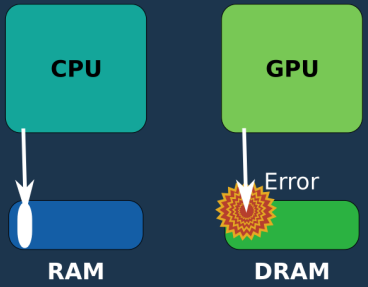
Memory Management

CUDA

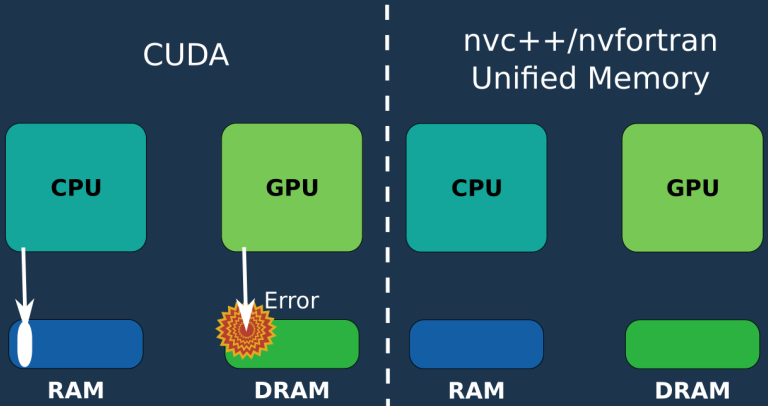


Memory Management

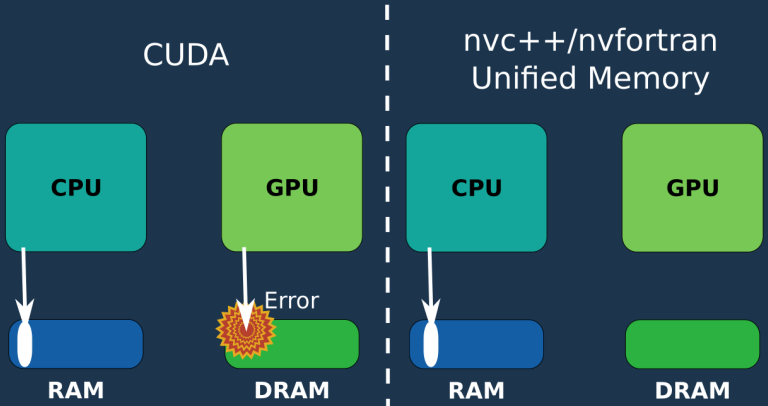
CUDA



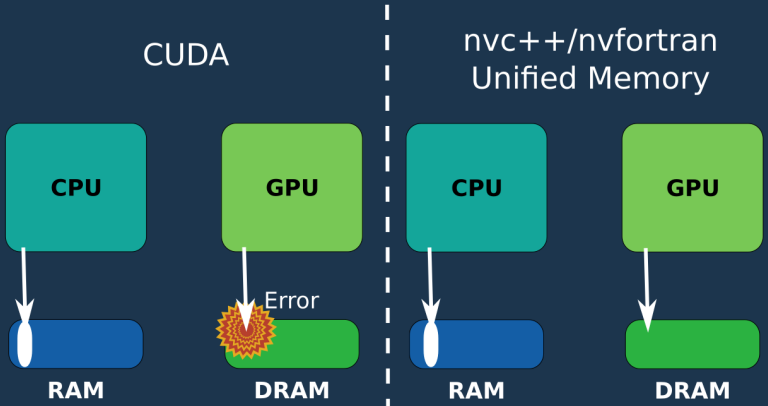
Memory Management



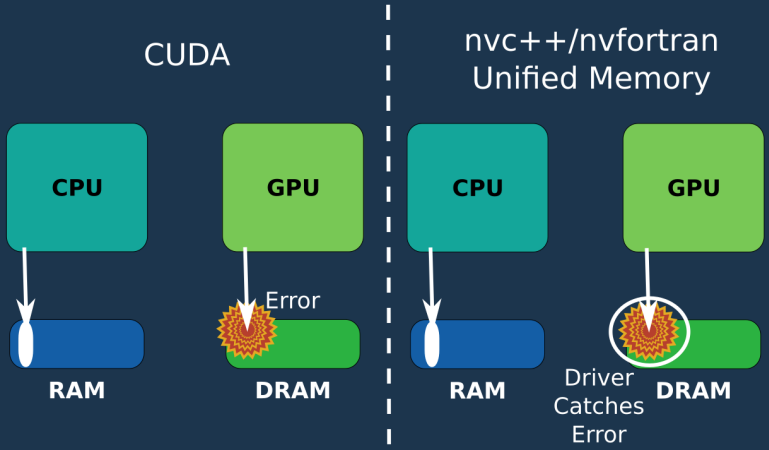
Memory Management



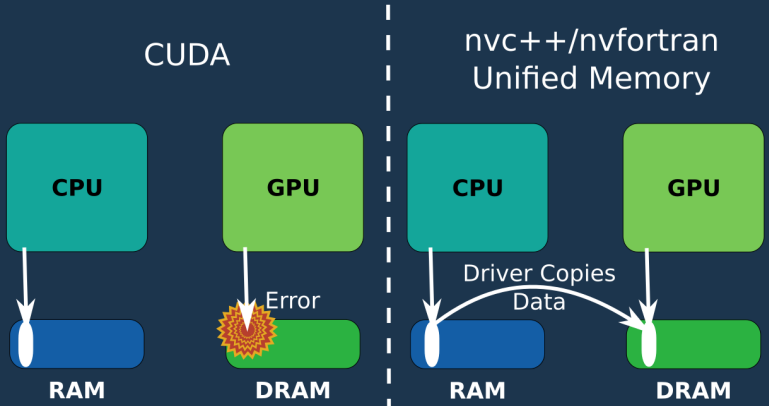
Memory Management



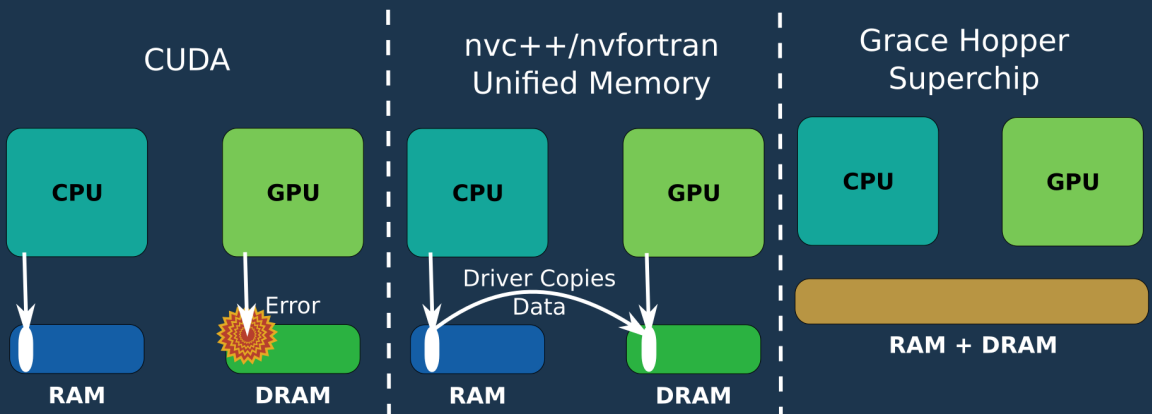
Memory Management



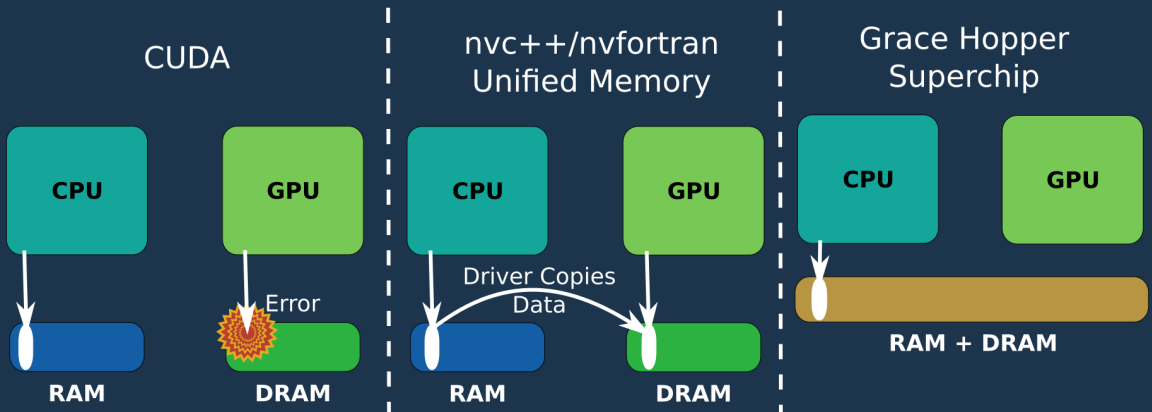
Memory Management



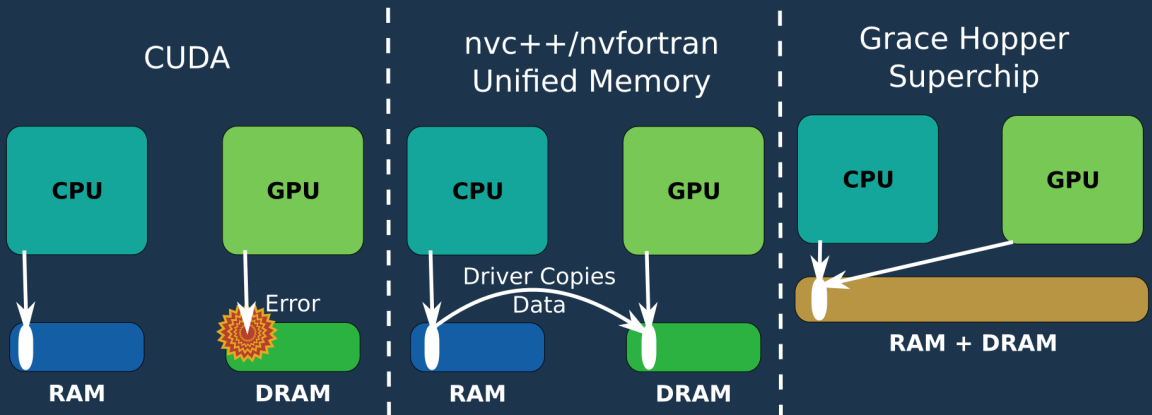
Memory Management



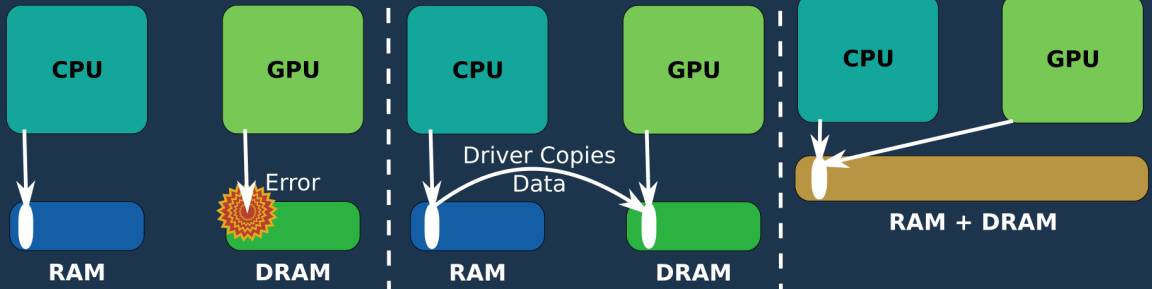
Memory Management



Memory Management

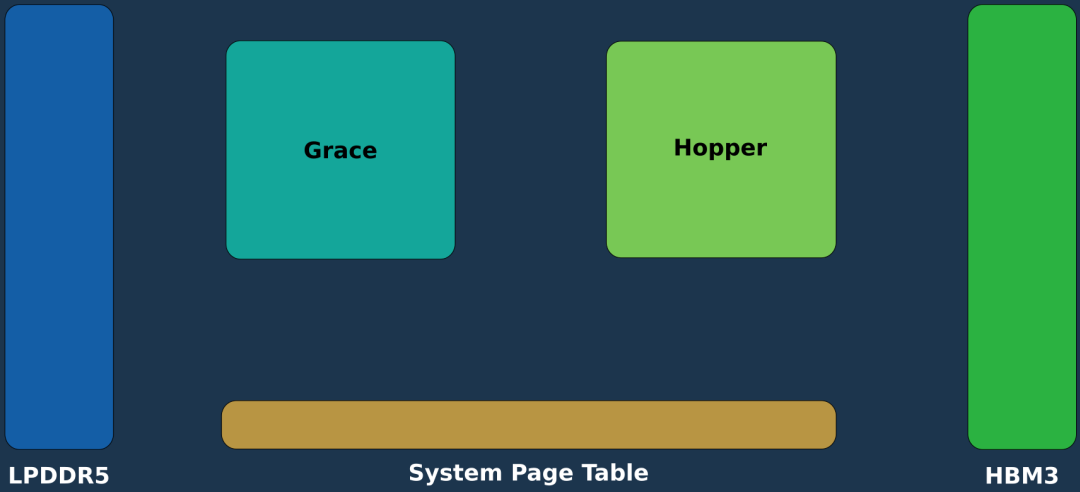


Memory Management

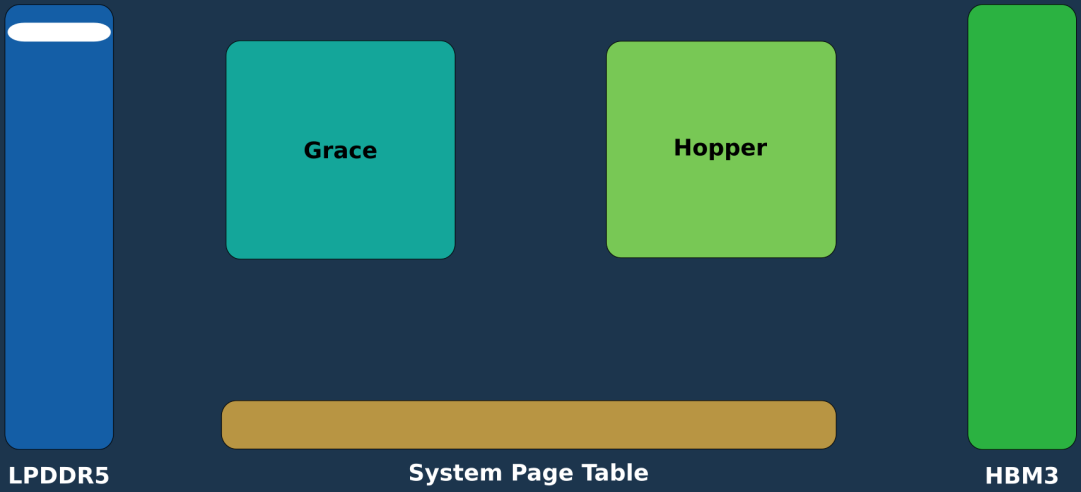


Yes but **RAM** and **DRAM** are **physically different**

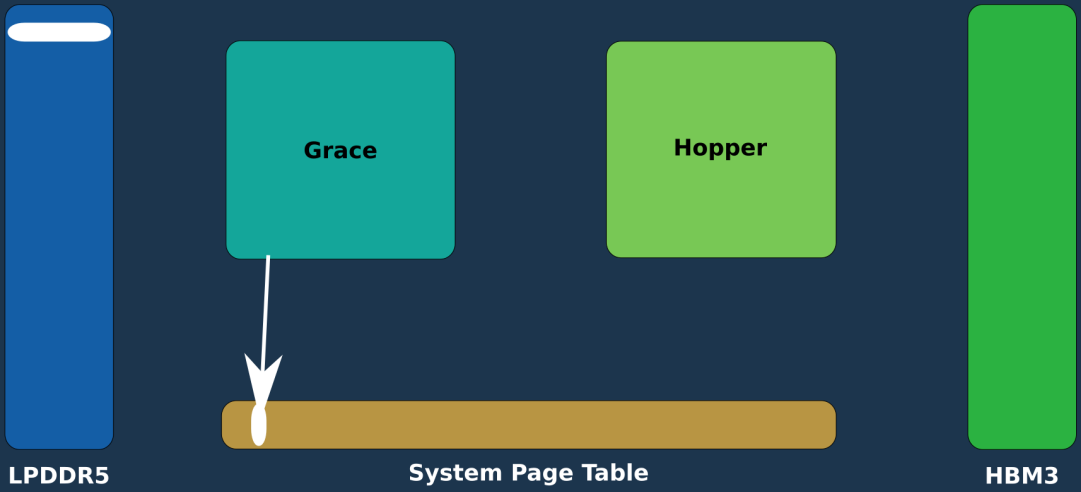
Memory Management Grace Hopper



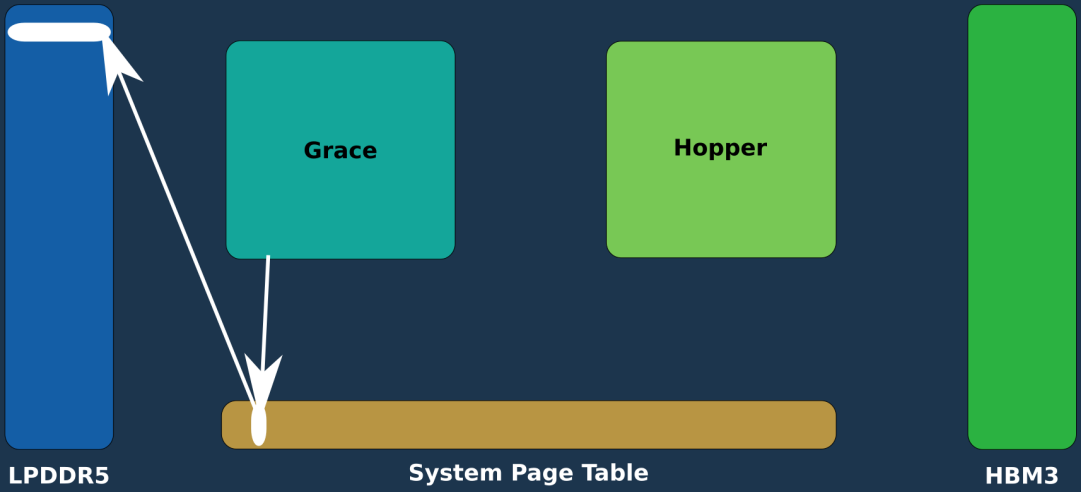
Memory Management Grace Hopper



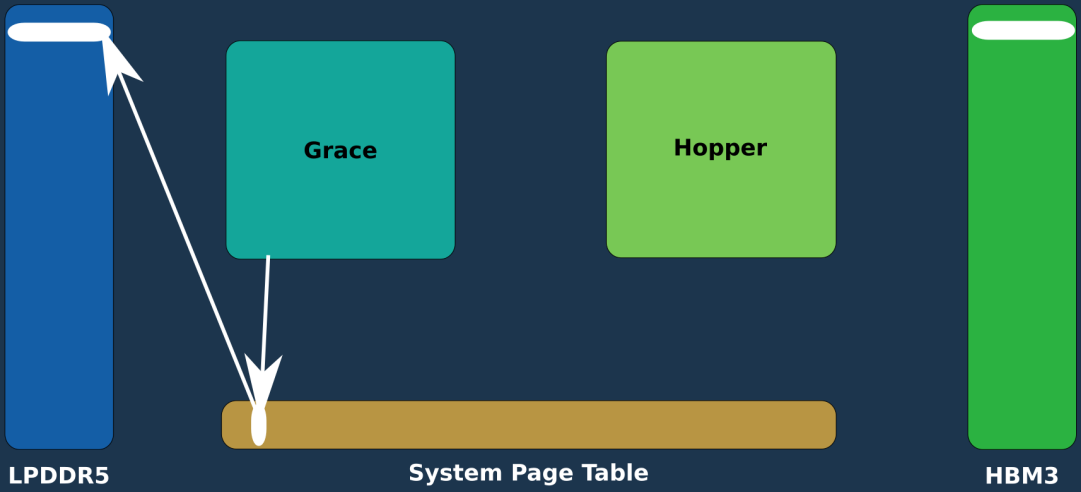
Memory Management Grace Hopper



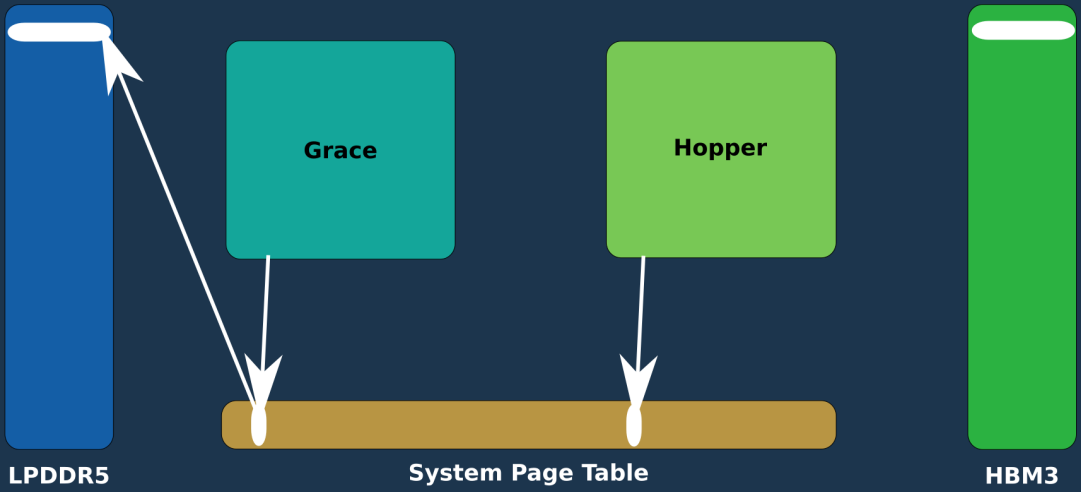
Memory Management Grace Hopper



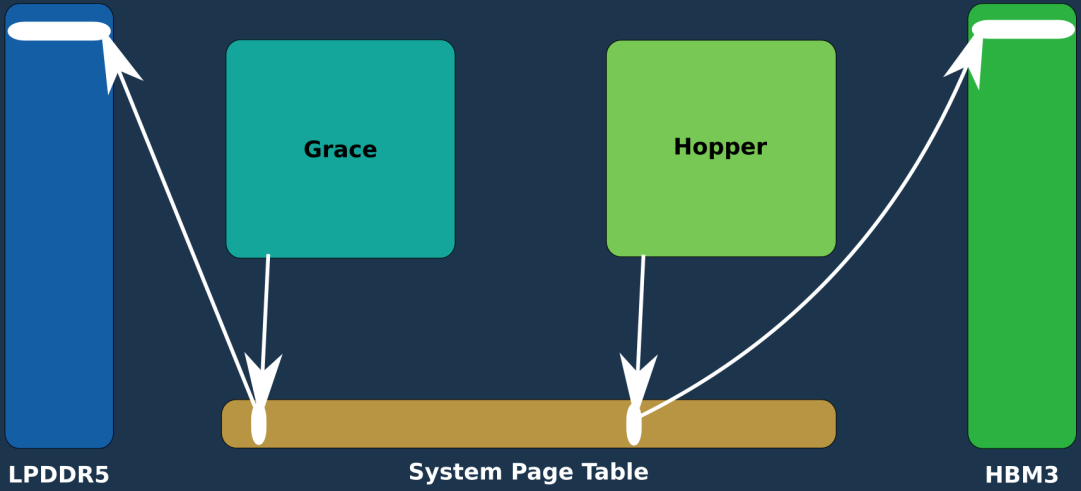
Memory Management Grace Hopper



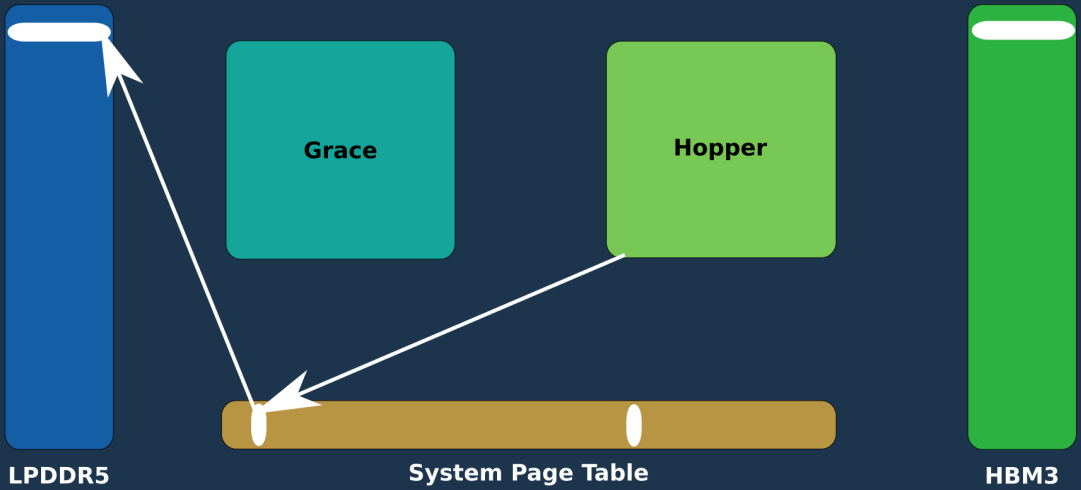
Memory Management Grace Hopper



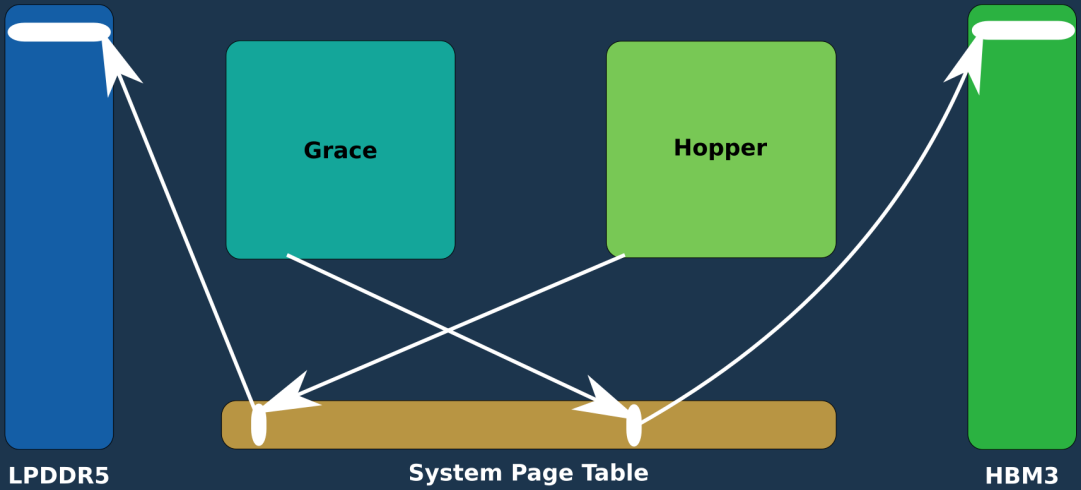
Memory Management Grace Hopper



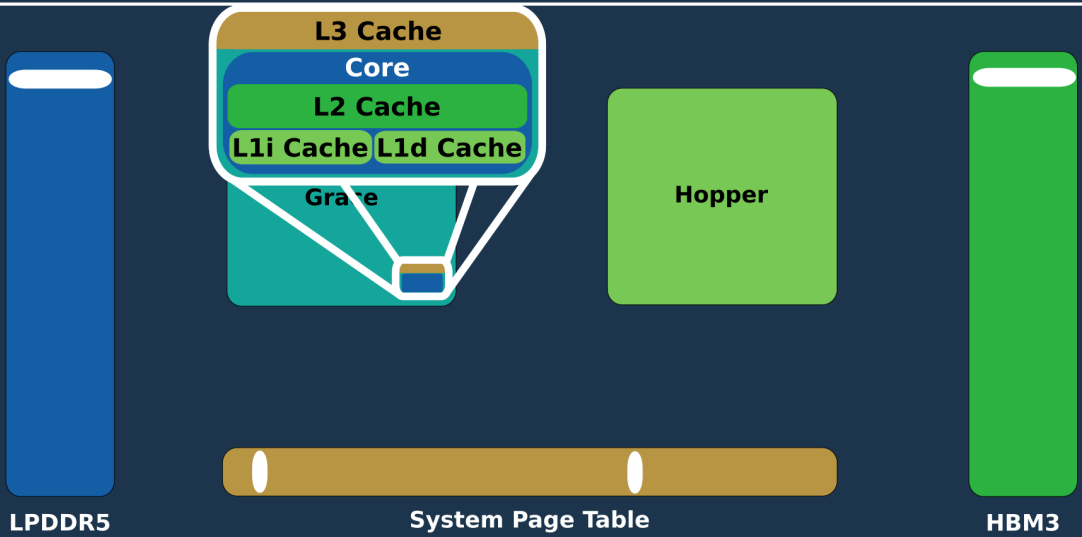
Memory Management Grace Hopper



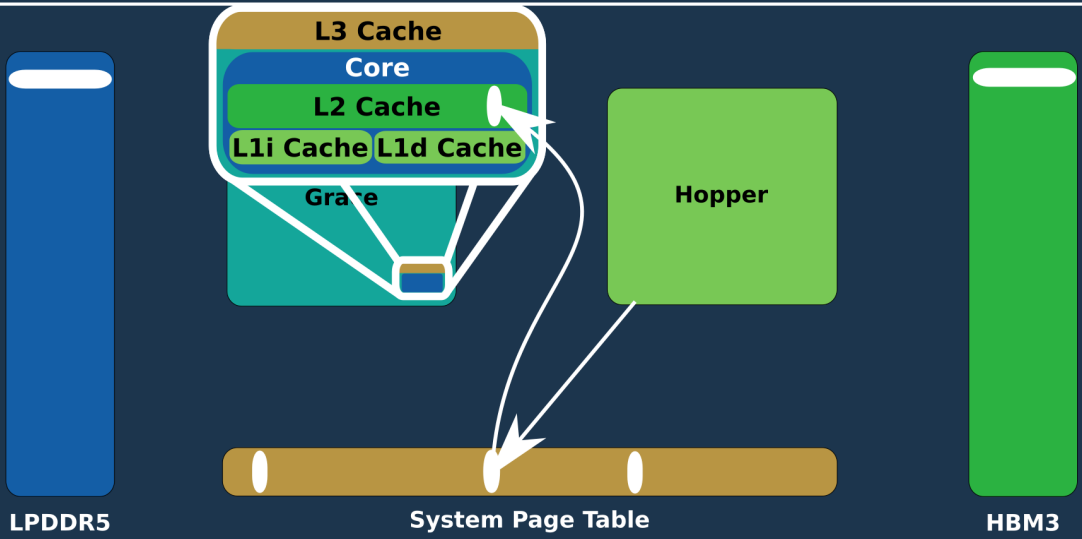
Memory Management Grace Hopper



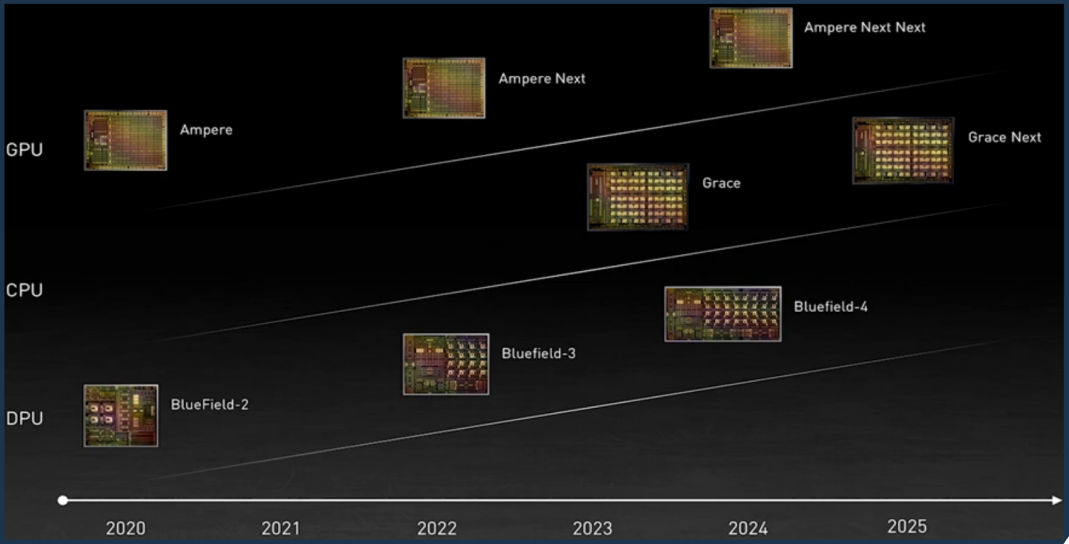
Memory Management Grace Hopper



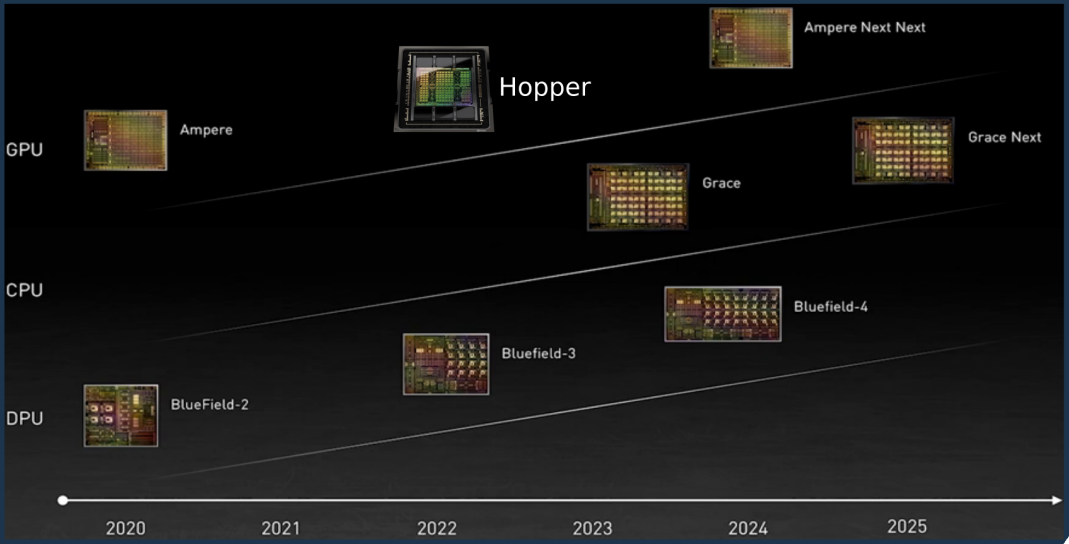
Memory Management Grace Hopper



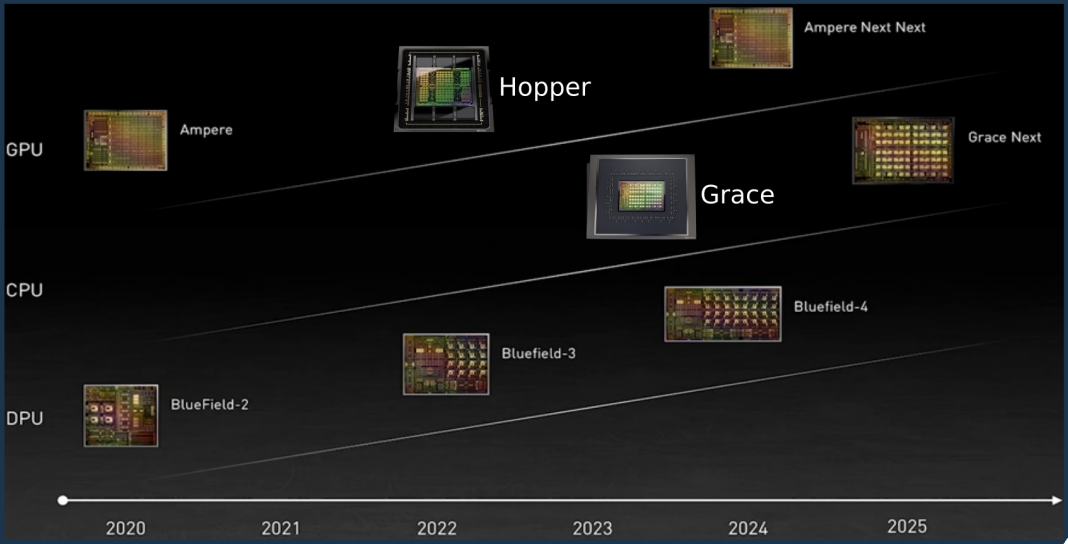
NVidia roadmap



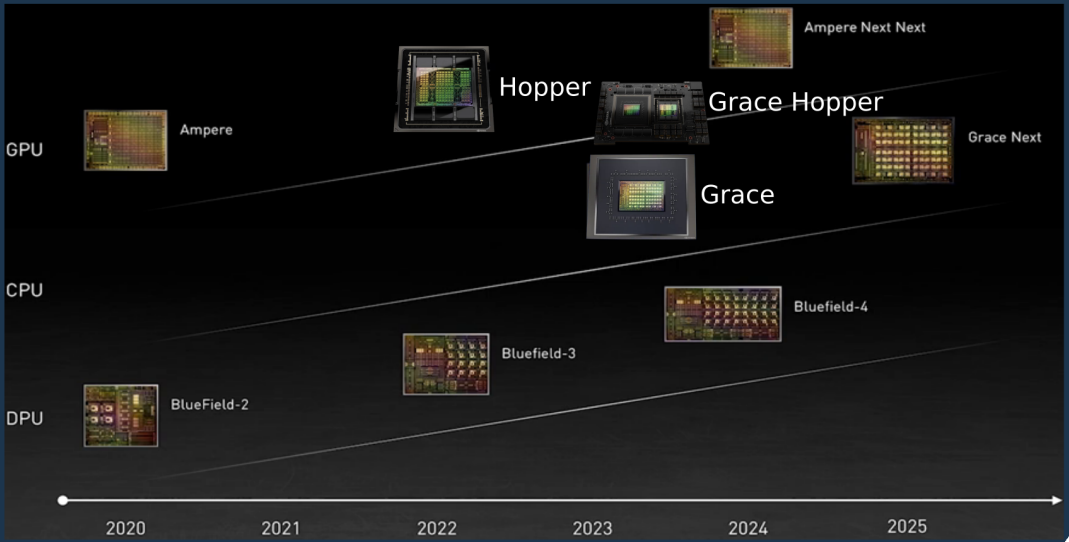
NVidia roadmap



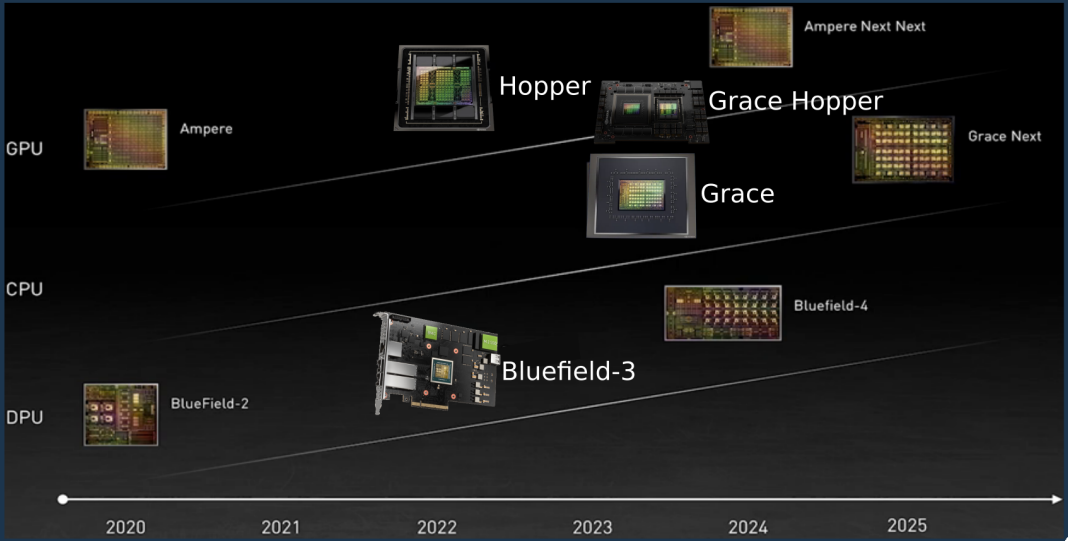
NVidia roadmap



NVidia roadmap



NVidia roadmap



C++ Evolution

C++ Evolution

- Parallel Algorithms*
- Forward Progress*
- New Memory Model*

C++ Evolution

- Parallel Algorithms*
- Forward Progress*
- New Memory Model*

- Modules
- Concepts*
- Coroutines
- Ranges*
- atomic wait*
- atomic_ref*
- barrier*

C++ Evolution

- Parallel Algorithms*
- Forward Progress*
- New Memory Model*

- Modules
- Concepts*
- Coroutines
- Ranges*
- atomic wait*
- atomic_ref*
- barrier*

- Ranges*
- mdspan*
- generator
- Extended FP

C++ Evolution

- Parallel Algorithms*
- Forward Progress*
- New Memory Model*

- Modules
- Concepts*
- Coroutines
- Ranges*
- atomic wait*
- atomic_ref*
- barrier*

- Ranges*
- mdspan*
- generator
- Extended FP

- Reflection
- Senders*
- Linear Algebra*
- SIMD
- RCU
- Hazard Pointers

C++ Evolution

- Parallel Algorithms*
- Forward Progress*
- New Memory Model*

- Modules
- Concepts*
- Coroutines
- Ranges*
- atomic wait*
- atomic_ref*
- barrier*

- Ranges*
- mdspan*
- generator
- Extended FP

- Reflection
- Senders*
- Linear Algebra*
- SIMD
- RCU
- Hazard Pointers

* = Available now in **NVC++** (**HPC SDK 22.7**)

Python : cuNumeric / Legate

Python : cuNumeric / Legate

```
import numpy as np
```

Python : cuNumeric / Legate

```
import numpy as np
```

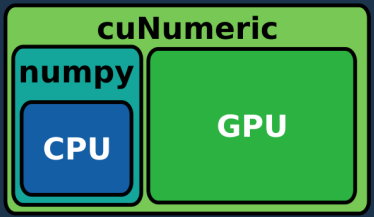
Python : cuNumeric / Legate

```
import numpy as np
```

↓

```
import cunumeric as np
```

60% of **Numpy** API on **GPU**



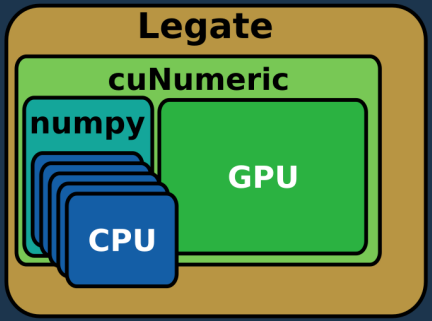
Python : cuNumeric / Legate

```
import numpy as np
```

↓

```
import cunumeric as np
```

60% of **Numpy** API on **GPU**



Legate :
- **Multi CPU**

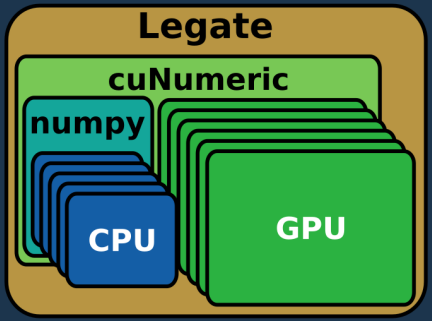
Python : cuNumeric / Legate

```
import numpy as np
```

↓

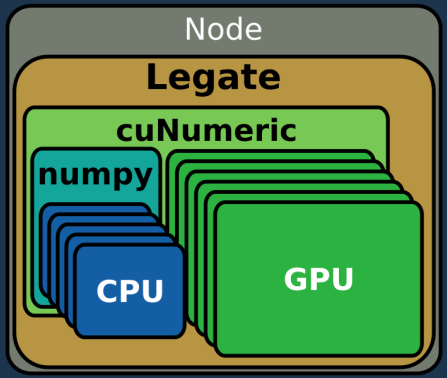
```
import cunumeric as np
```

60% of **Numpy** API on **GPU**



Legate :
- **Multi CPU**
- **Multi GPU**

Python : cuNumeric / Legate




```
import numpy as np
```

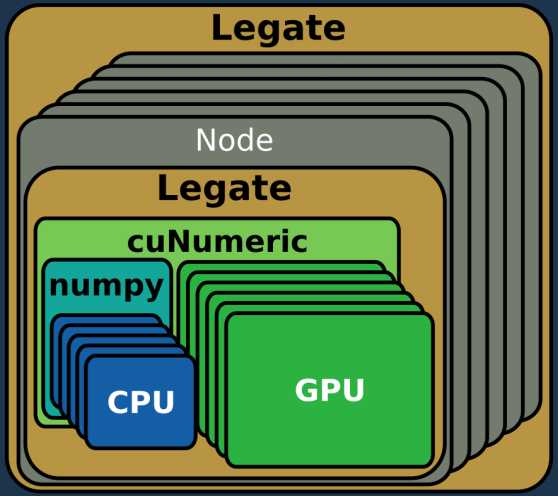
↓

```
import cunumeric as np
```

60% of **Numpy** API on **GPU**

Legate :
- **Multi CPU**
- **Multi GPU**

Python : cuNumeric / Legate




```
import numpy as np
```

↓

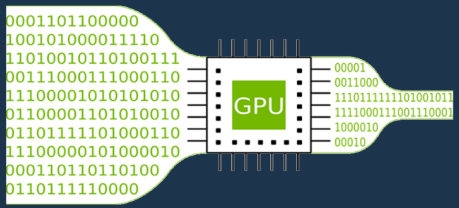
```
import cunumeric as np
```

60% of **Numpy** API on **GPU**

Legate :
- **Multi CPU**
- **Multi GPU**
- **Multi Node**

Data Compression : NVComp

C/C++ standard calls



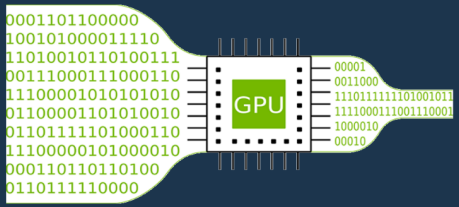
Compression / decompression algorithms :

- LZ4
- Snappy
- Deflate*/GDeflate*
- Cascaded*
- Bitcomp*
- ANS (Asymetric Numeral System)
- **ZStandard***
- **GZip***

* : might have errors with others decompressors

Data Compression : NVComp

C/C++ standard calls



Compression / decompression algorithms :

- LZ4
- Snappy
- Deflate*/GDeflate*
- Cascaded*
- Bitcomp*
- ANS (Asymetric Numeral System)
- **ZStandard***
- **GZip***

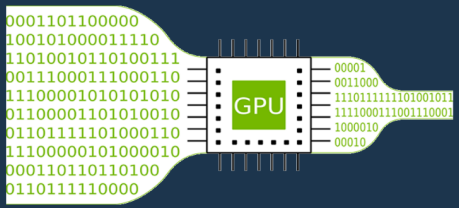
* : might have errors with others decompressors

Release 2.6 :

- Can be used with **GPU Direct Storage**
- High Level API simplified
- No kernel call available

Data Compression : NVComp

C/C++ standard calls



Compression / decompression algorithms :

- LZ4
- Snappy
- Deflate*/GDeflate*
- Cascaded*
- Bitcomp*
- ANS (Asymetric Numeral System)
- **ZStandard***
- **GZip***

* : might have errors with others decompressors

Release 2.6 :

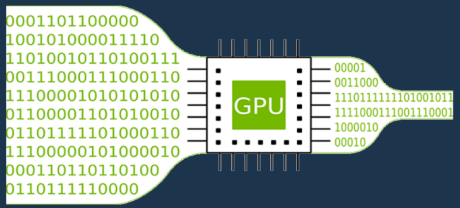
- Can be used with **GPU Direct Storage**
- High Level API simplified
- No kernel call available

Some advices :

- To be used on files of **few MB**
- With **chunks** > 8kB
- Decompressed data must be accessible by the GPU

Data Compression : NVComp

C/C++ standard calls



Compression / decompression algorithms :

- LZ4
- Snappy
- Deflate*/GDeflate*
- Cascaded*
- Bitcomp*
- ANS (Asymetric Numeral System)
- **ZStandard***
- **GZip***

* : might have errors with others decompressors

Release 2.6 :

- Can be used with **GPU Direct Storage**
- High Level API simplified
- No kernel call available

Things to be improved :

- Only low-level interface : **Deflate**, **ZStandard**, **GZip**
- **GZip** decompression with low-level interface only

Some advices :

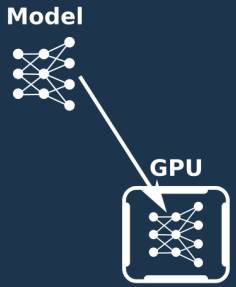
- To be used on files of **few MB**
- With **chunks** > 8kB
- Decompressed data must be accessible by the GPU

Deep Learning : Data Parallelism

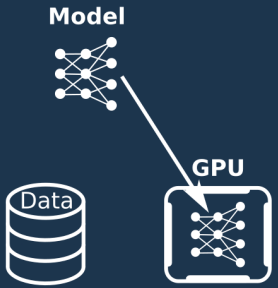
Deep Learning : Data Parallelism

Model

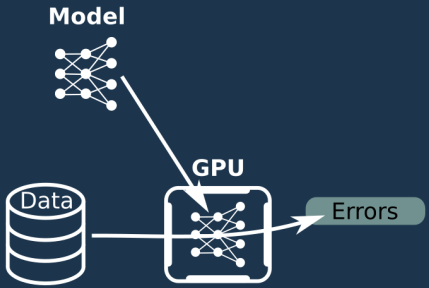
Deep Learning : Data Parallelism



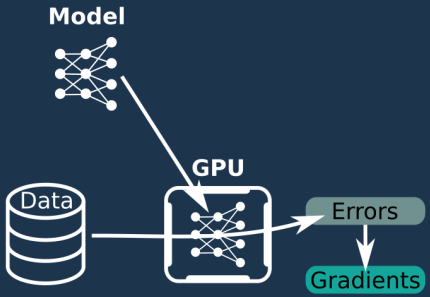
Deep Learning : Data Parallelism



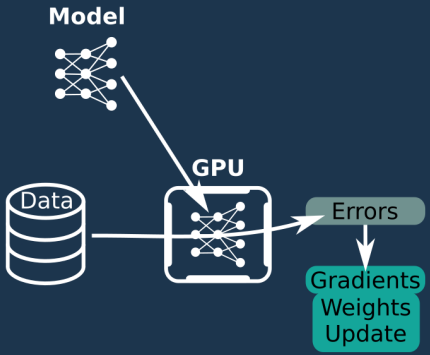
Deep Learning : Data Parallelism



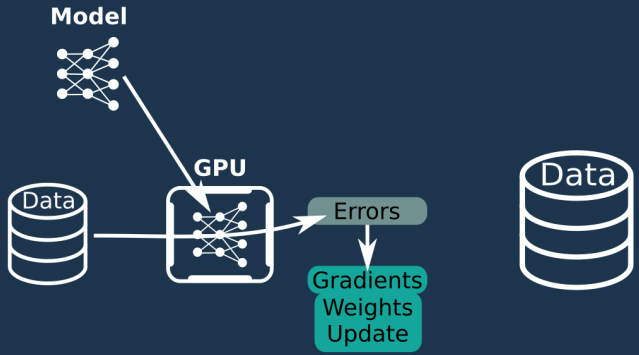
Deep Learning : Data Parallelism



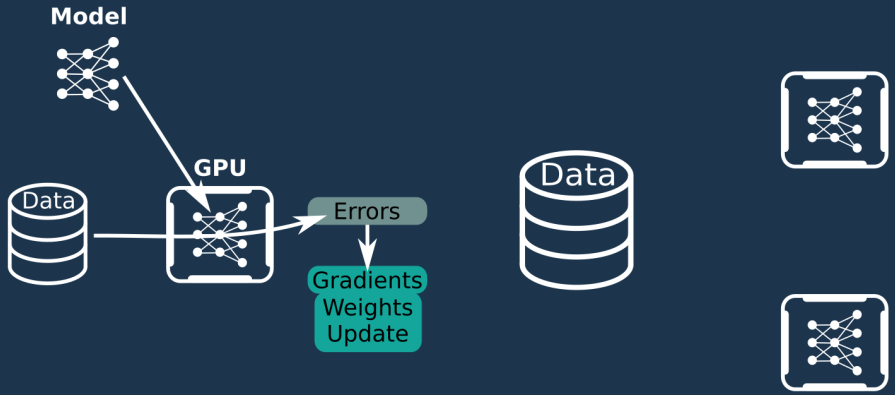
Deep Learning : Data Parallelism



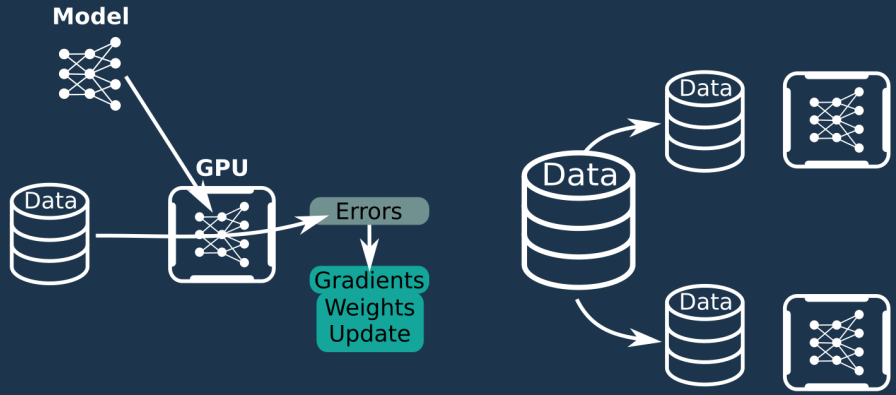
Deep Learning : Data Parallelism



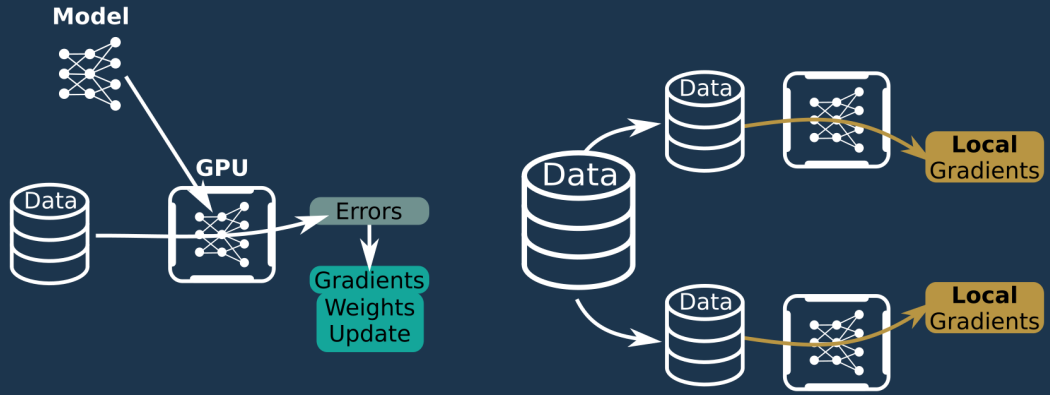
Deep Learning : Data Parallelism



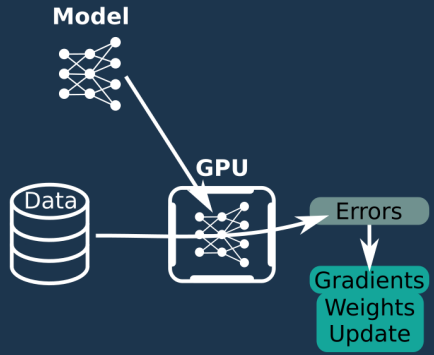
Deep Learning : Data Parallelism

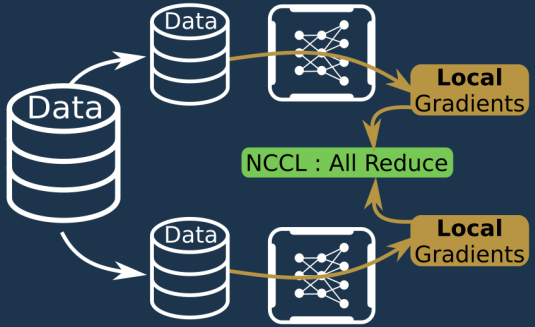


Deep Learning : Data Parallelism

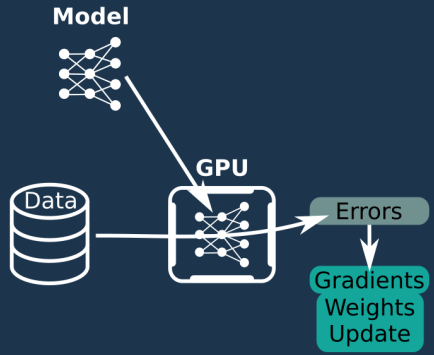


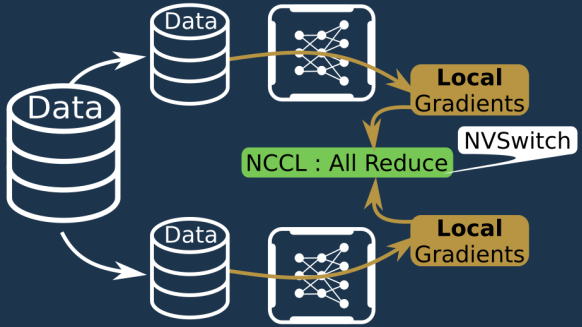
Deep Learning : Data Parallelism



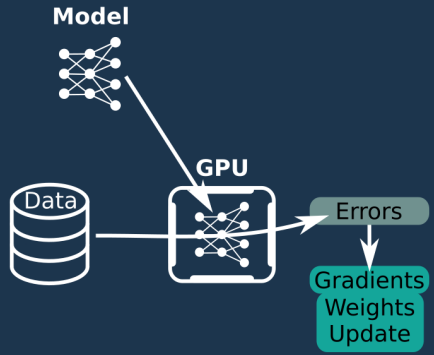


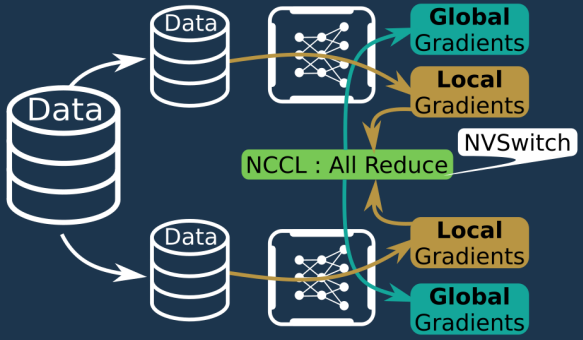
Deep Learning : Data Parallelism



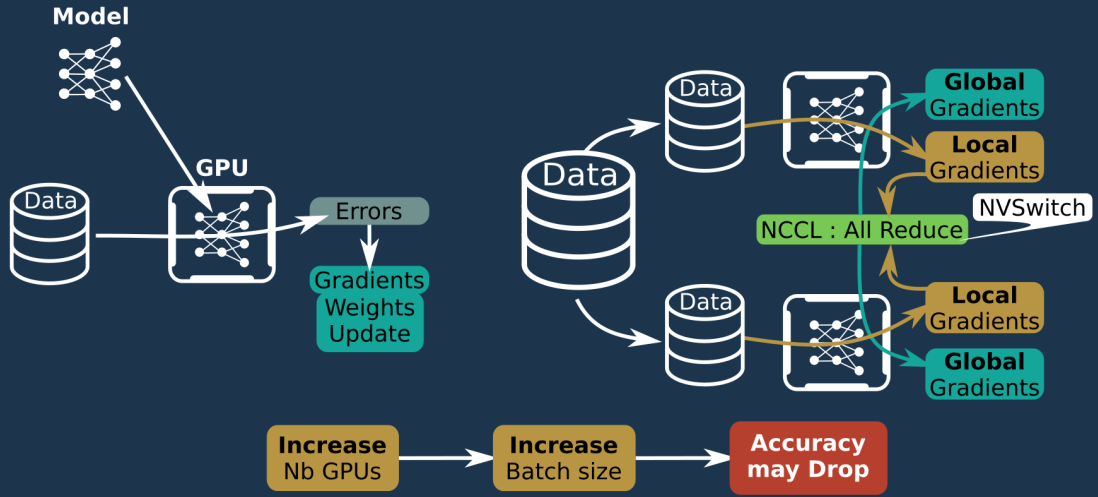


Deep Learning : Data Parallelism





Deep Learning : Data Parallelism

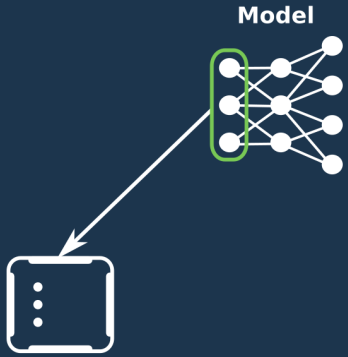


Deep Learning : Layer Parallelism

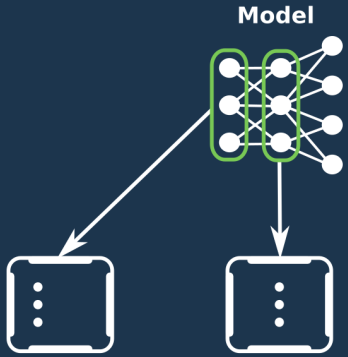
Deep Learning : Layer Parallelism

Model

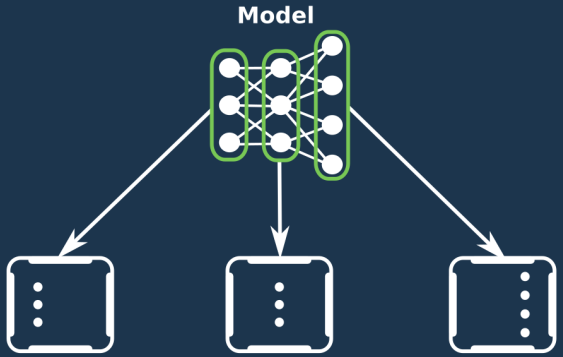
Deep Learning : Layer Parallelism



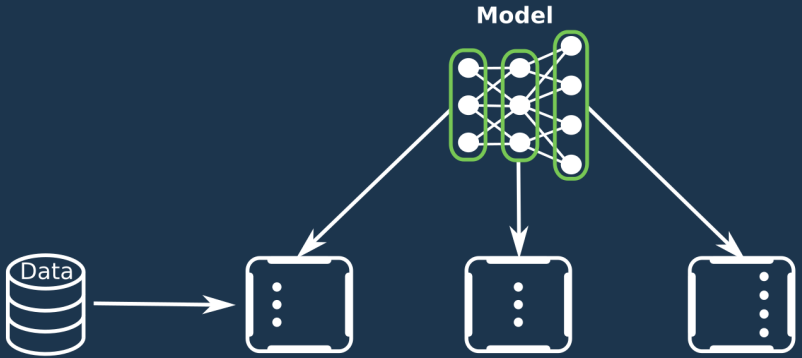
Deep Learning : Layer Parallelism



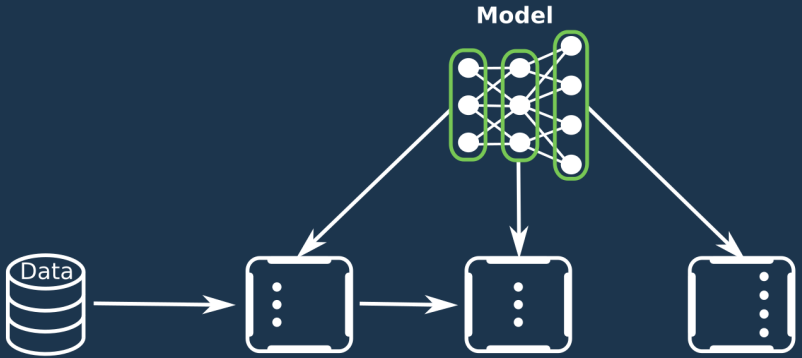
Deep Learning : Layer Parallelism



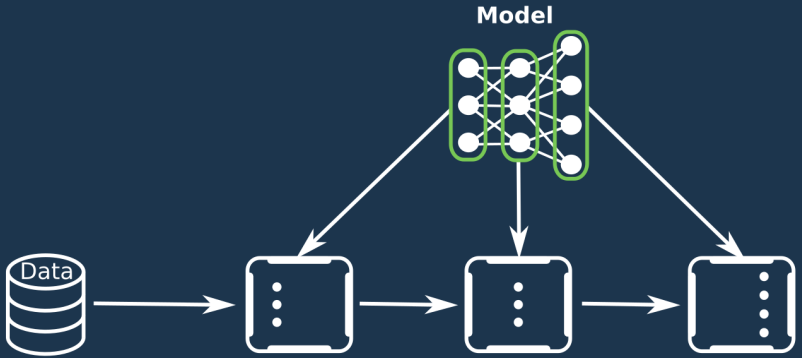
Deep Learning : Layer Parallelism



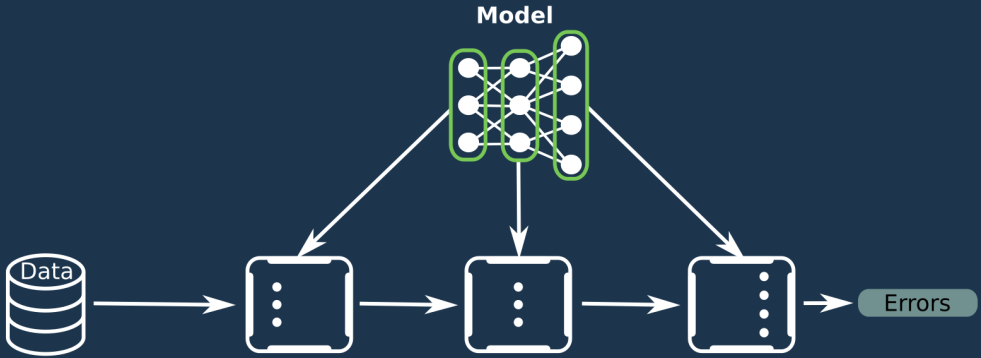
Deep Learning : Layer Parallelism



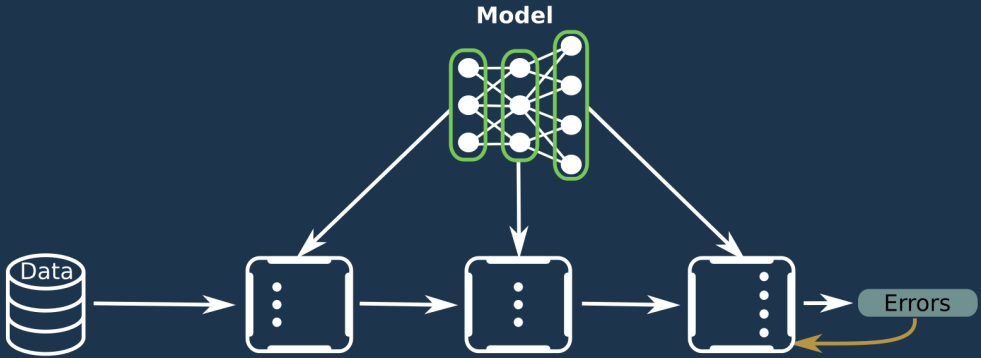
Deep Learning : Layer Parallelism



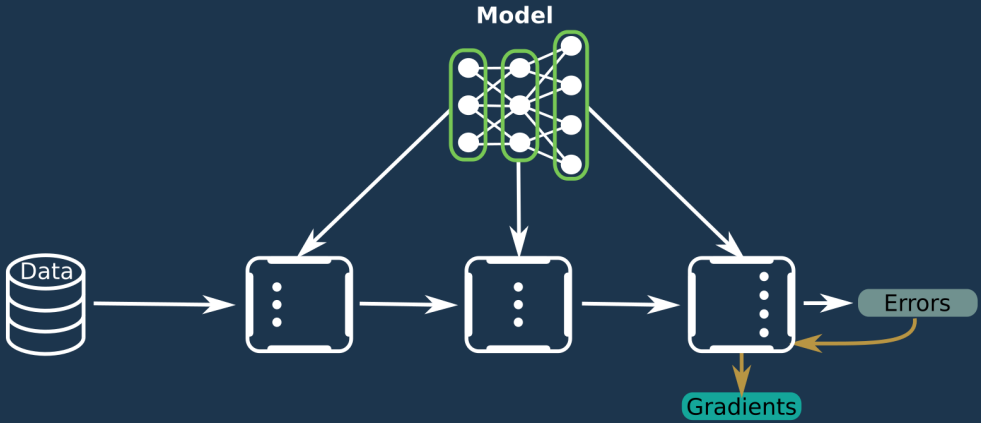
Deep Learning : Layer Parallelism



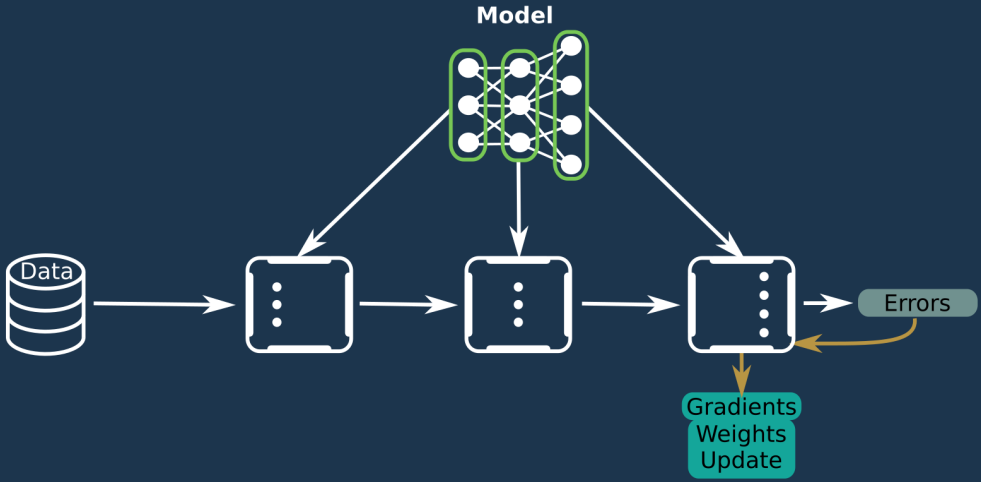
Deep Learning : Layer Parallelism



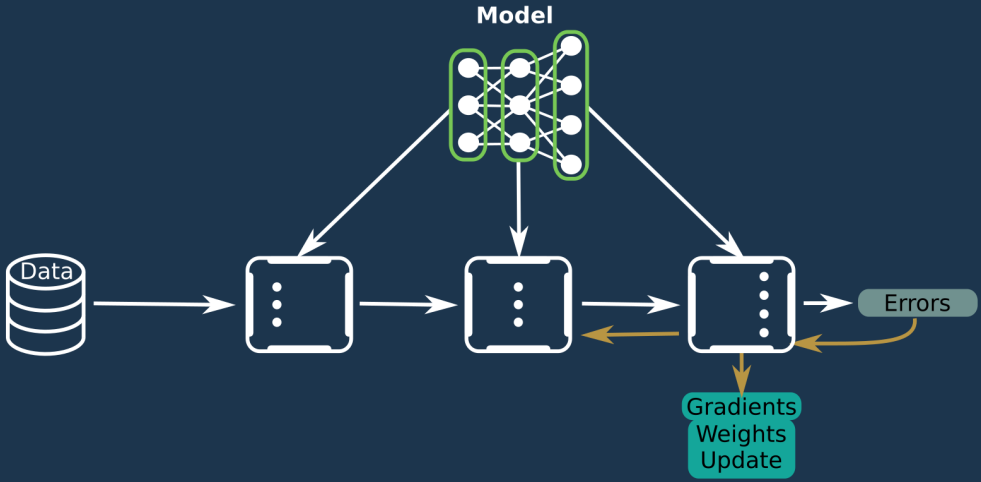
Deep Learning : Layer Parallelism



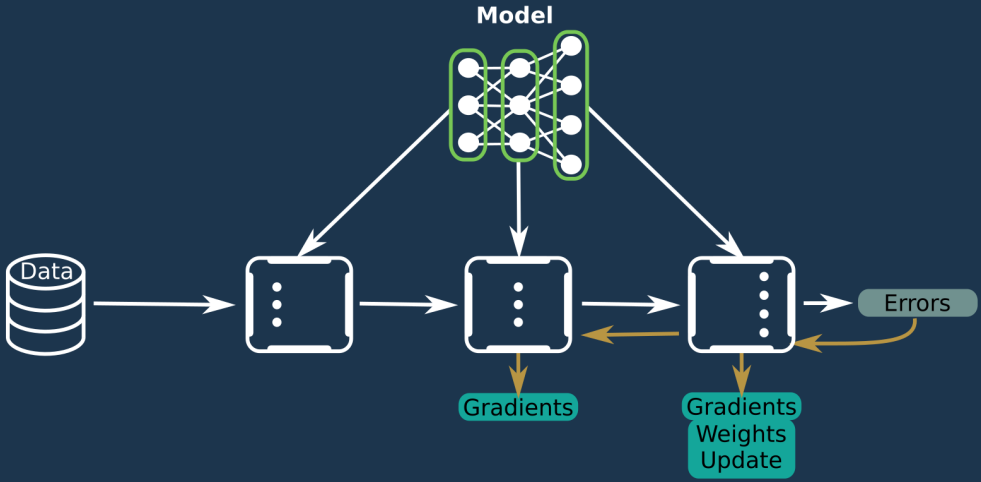
Deep Learning : Layer Parallelism



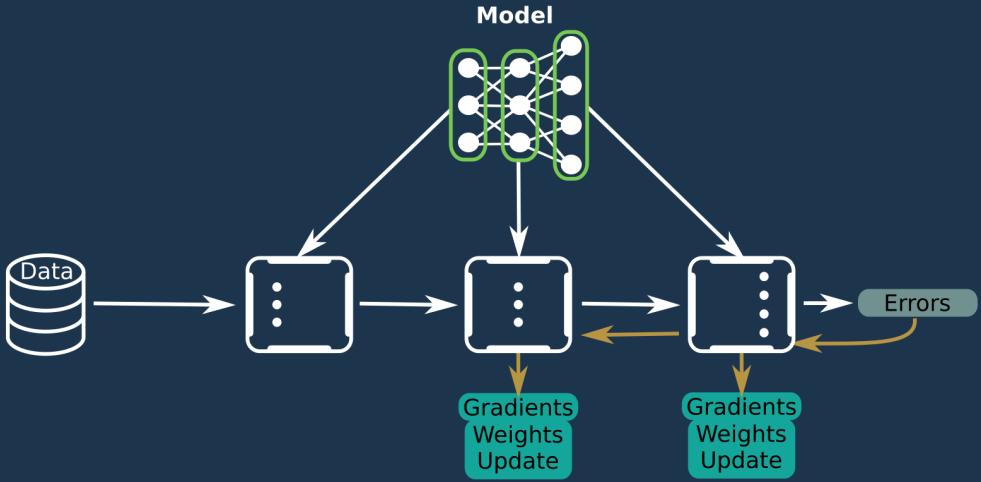
Deep Learning : Layer Parallelism



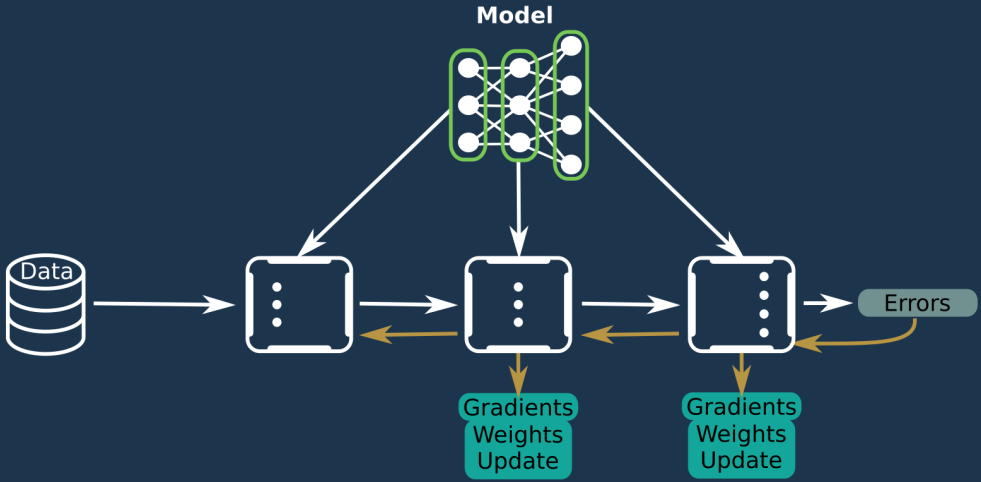
Deep Learning : Layer Parallelism



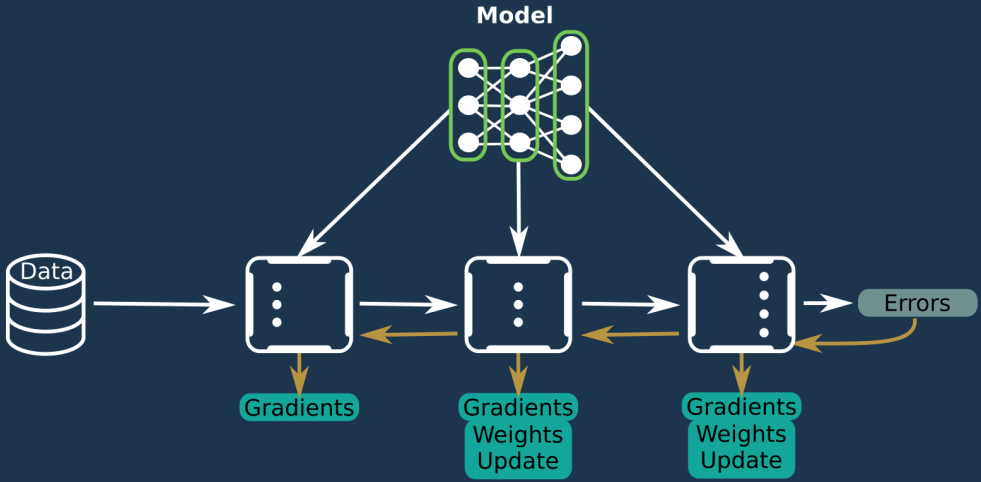
Deep Learning : Layer Parallelism



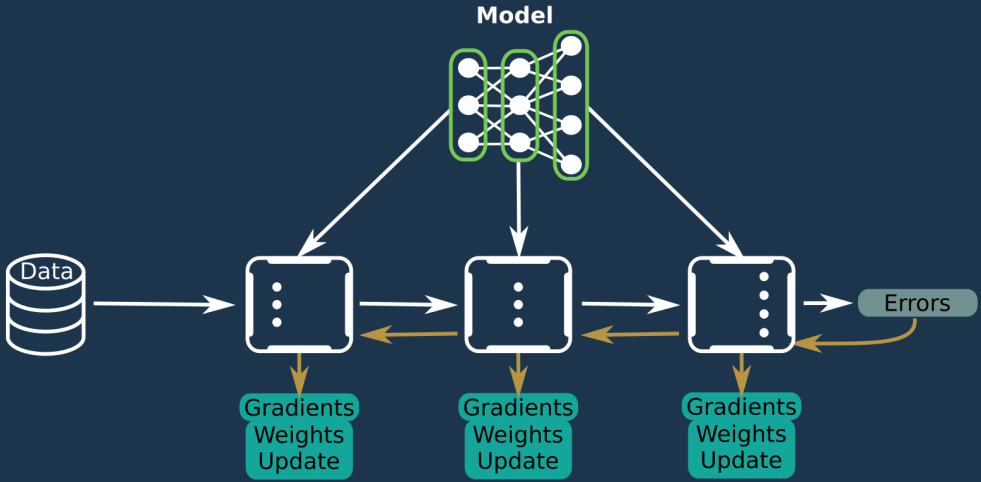
Deep Learning : Layer Parallelism



Deep Learning : Layer Parallelism

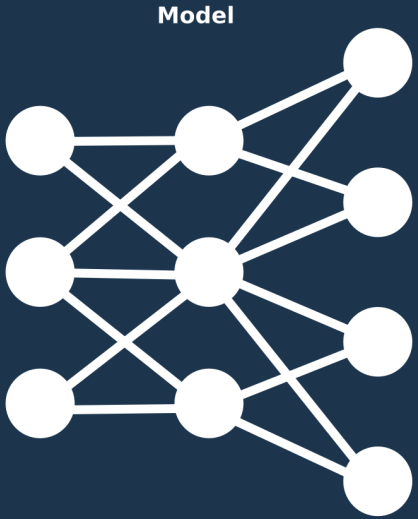


Deep Learning : Layer Parallelism

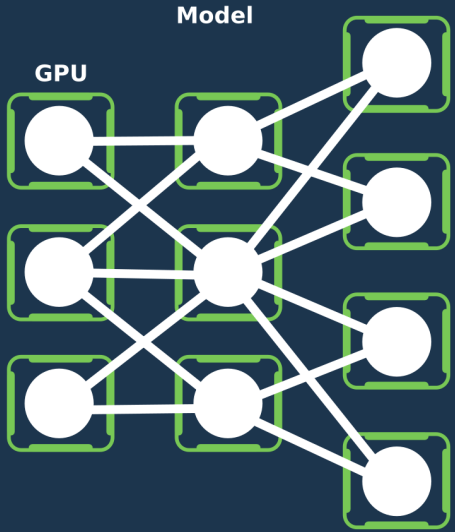


Deep Learning : Tensor Parallelism

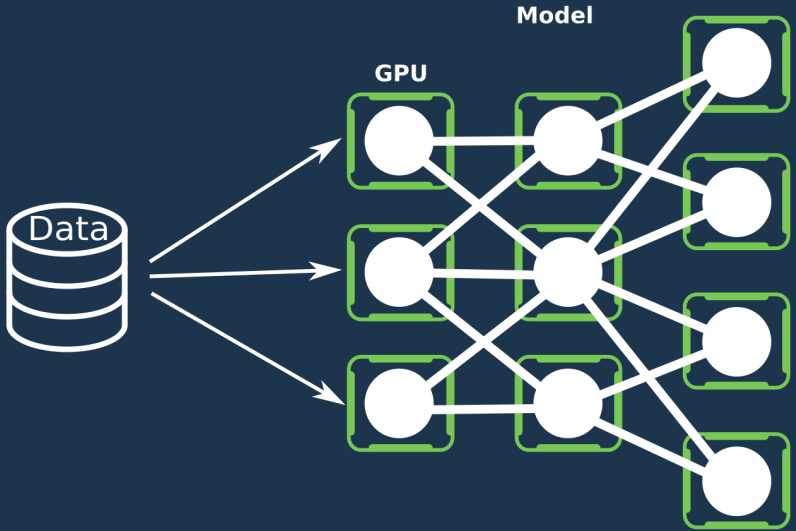
Deep Learning : Tensor Parallelism



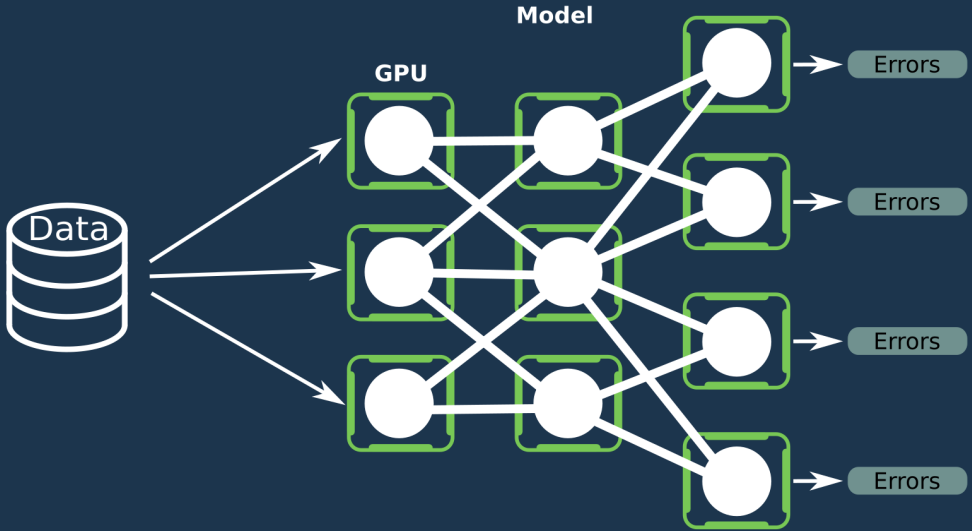
Deep Learning : Tensor Parallelism



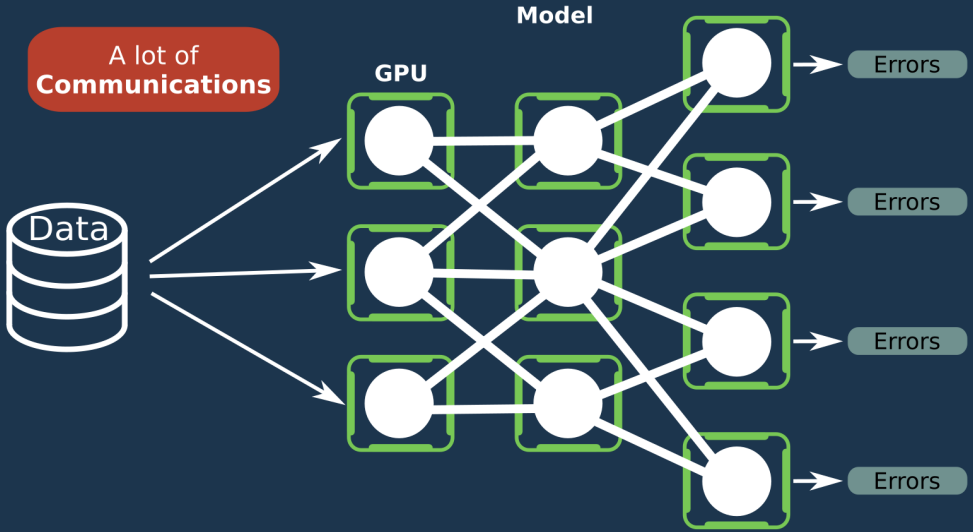
Deep Learning : Tensor Parallelism



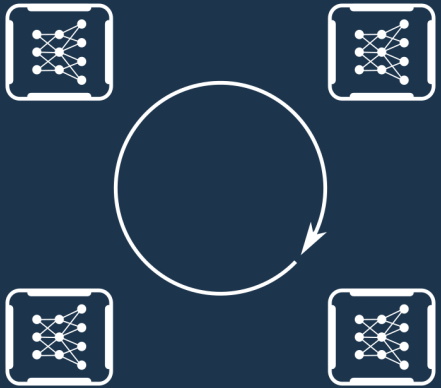
Deep Learning : Tensor Parallelism



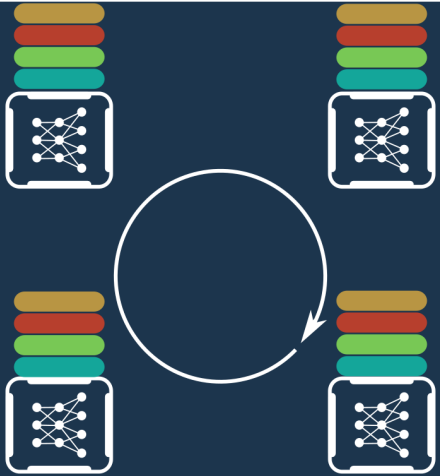
Deep Learning : Tensor Parallelism



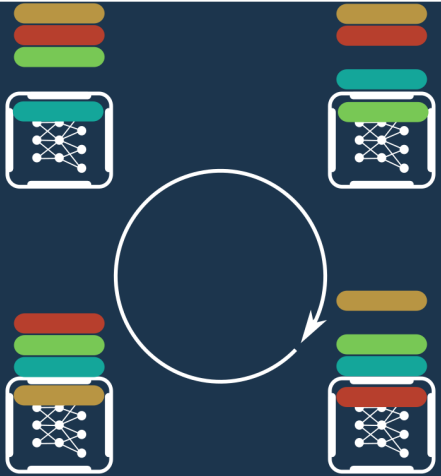
Deep Learning : Ring Algorithm



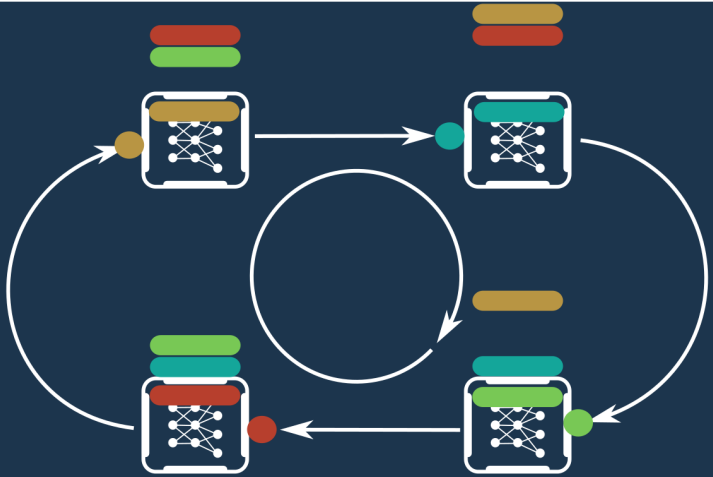
Deep Learning : Ring Algorithm



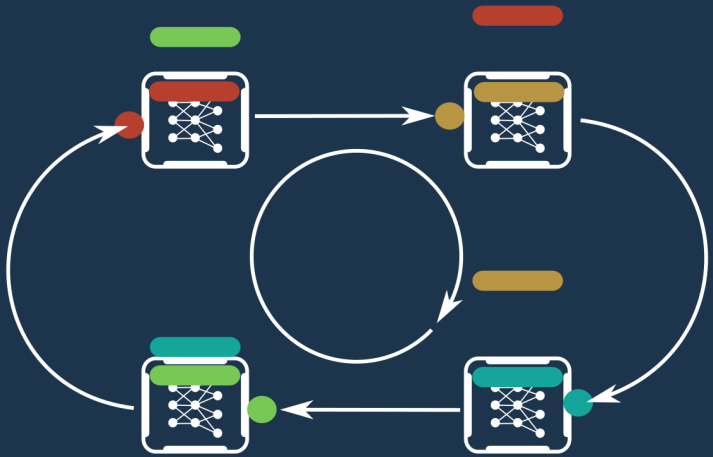
Deep Learning : Ring Algorithm



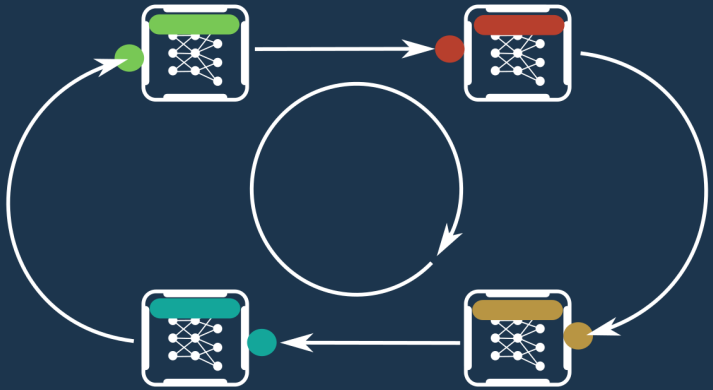
Deep Learning : Ring Algorithm



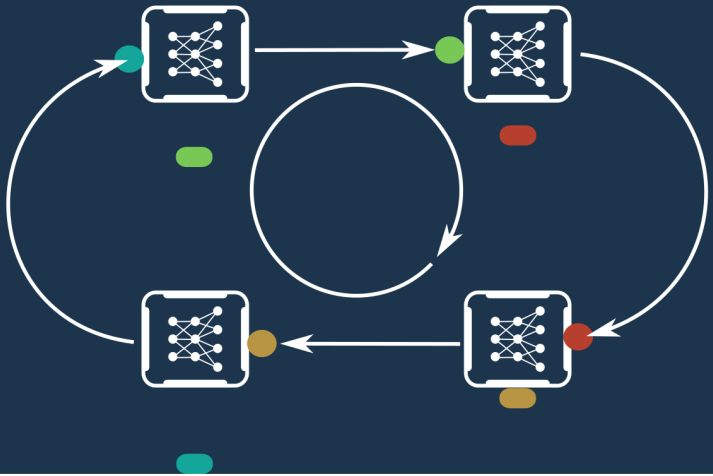
Deep Learning : Ring Algorithm



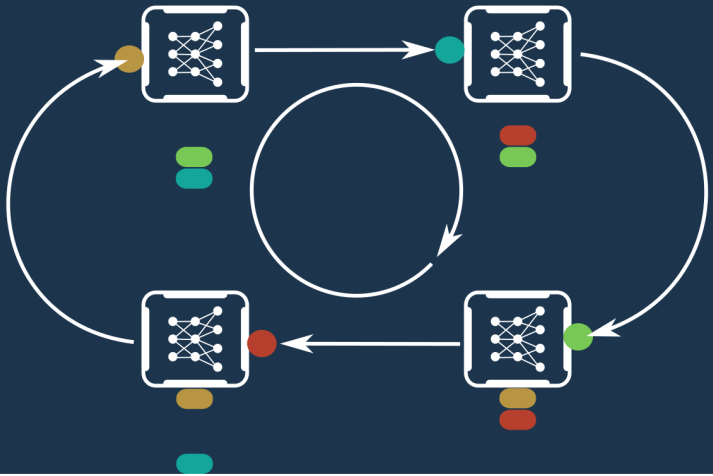
Deep Learning : Ring Algorithm



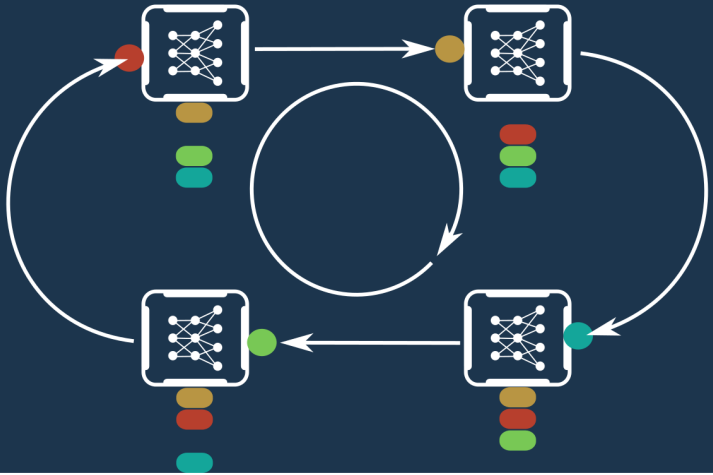
Deep Learning : Ring Algorithm



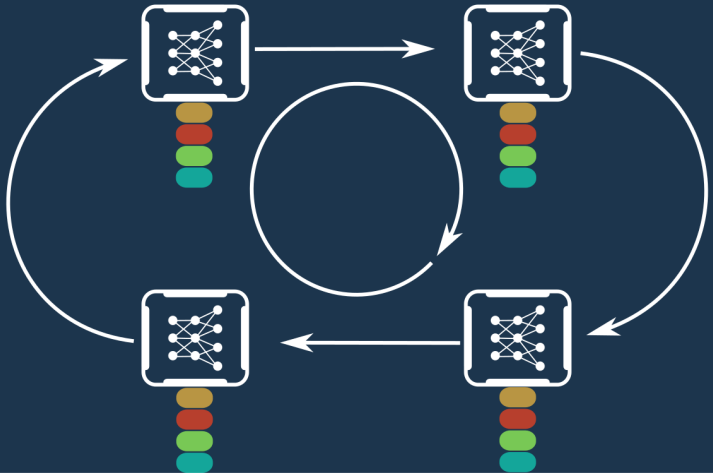
Deep Learning : Ring Algorithm



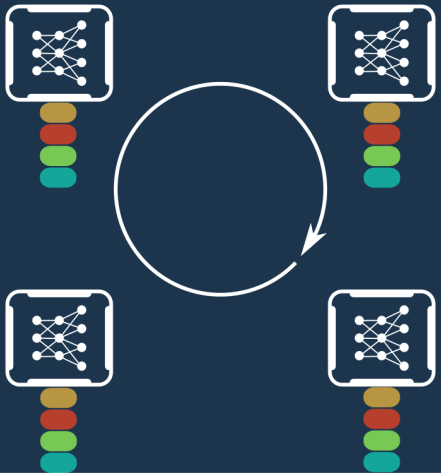
Deep Learning : Ring Algorithm



Deep Learning : Ring Algorithm

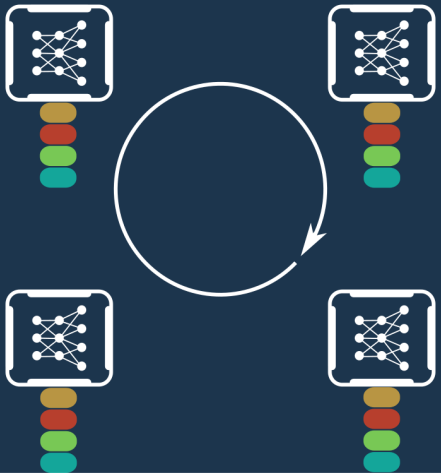


Deep Learning : Ring Algorithm



Simple Algorithm

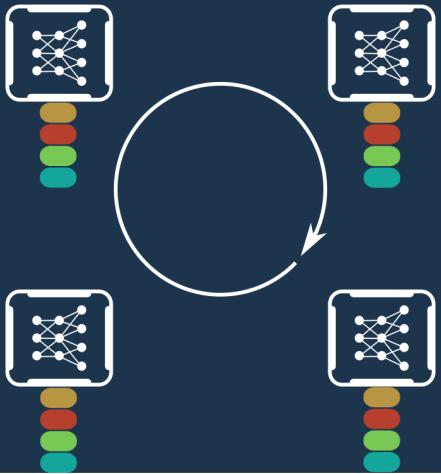
Deep Learning : Ring Algorithm



Simple Algorithm

2N - 1 steps

Deep Learning : Ring Algorithm



Simple Algorithm

2N - 1 steps

Scalability Issues

Deep Learning : Quantization

Large Language Models

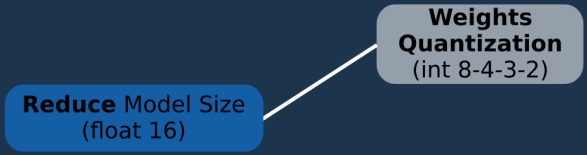
Deep Learning : Quantization

Large Language Models

Reduce Model Size
(float 16)

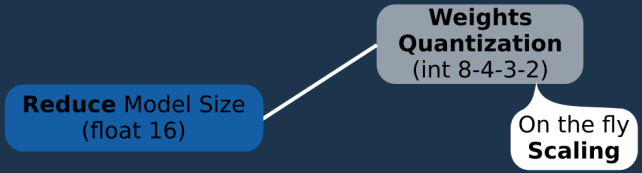
Deep Learning : Quantization

Large Language Models



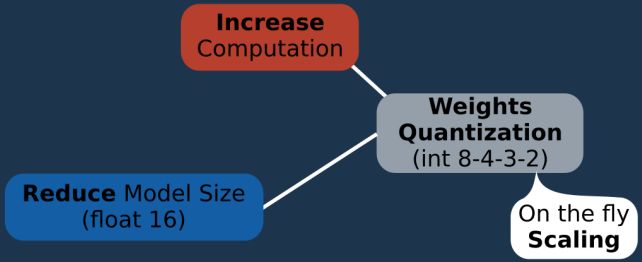
Deep Learning : Quantization

Large Language Models



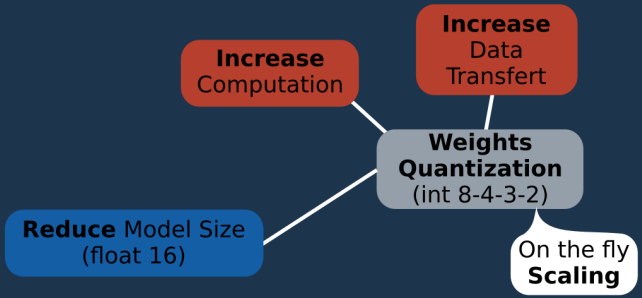
Deep Learning : Quantization

Large Language Models

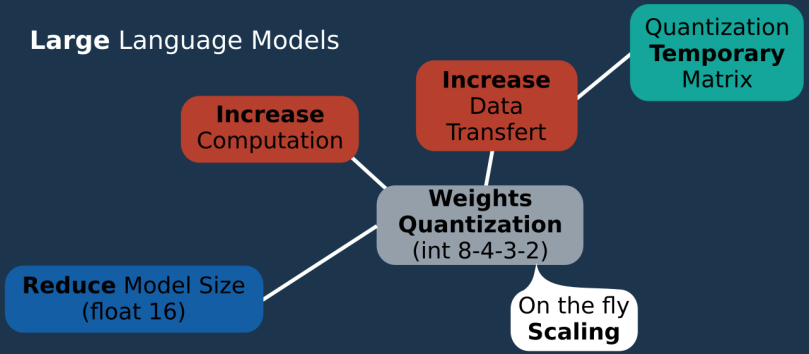


Deep Learning : Quantization

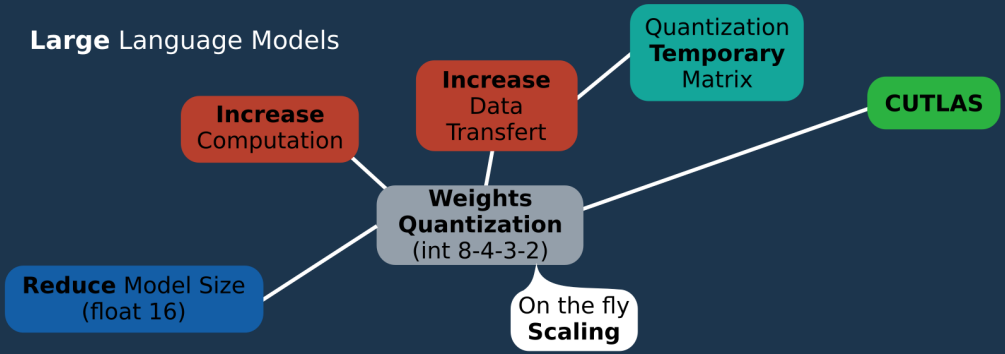
Large Language Models



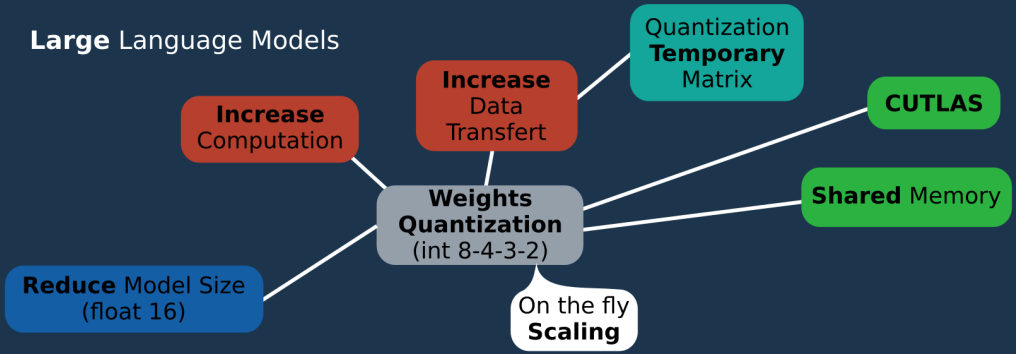
Deep Learning : Quantization



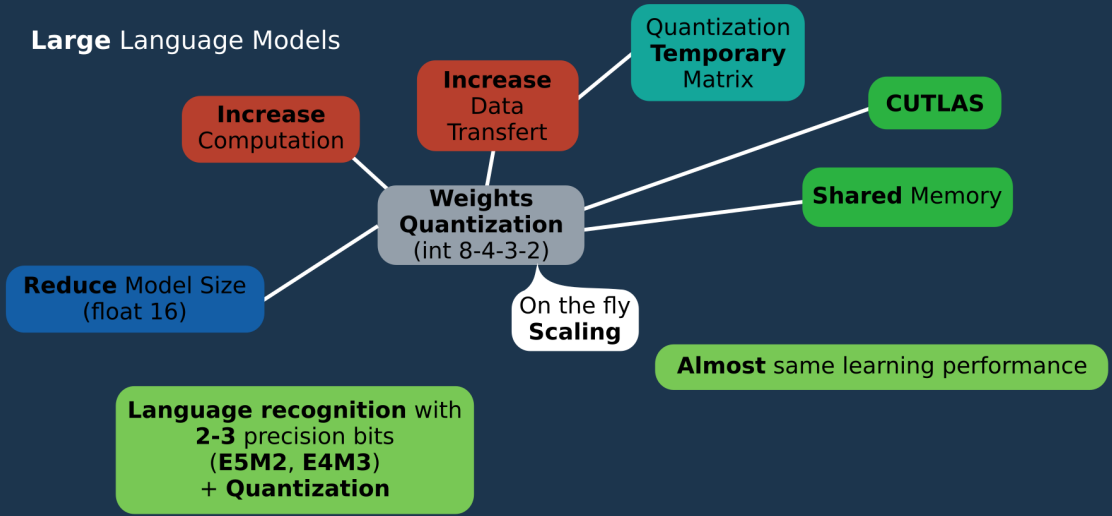
Deep Learning : Quantization



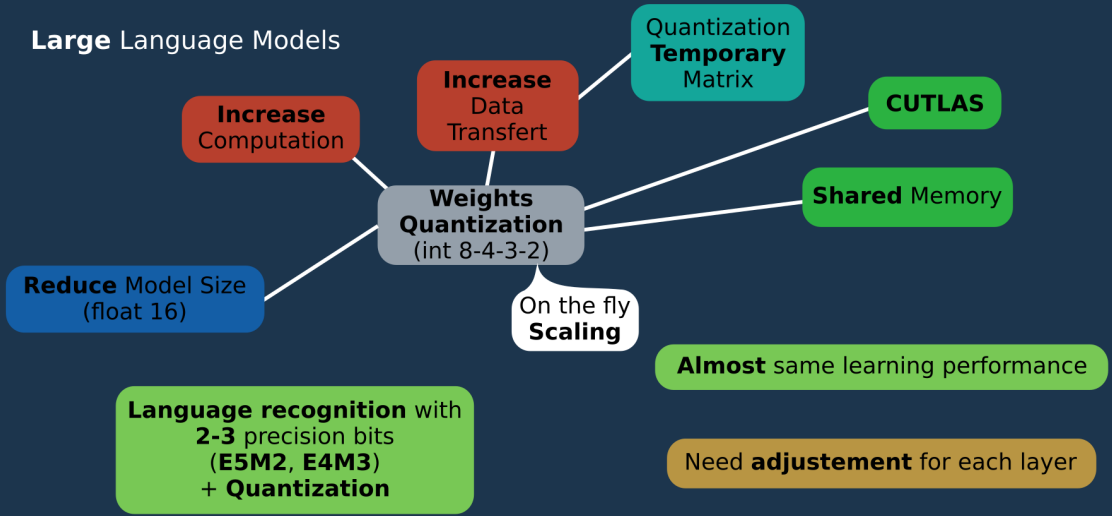
Deep Learning : Quantization



Deep Learning : Quantization



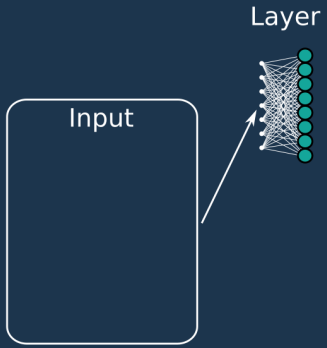
Deep Learning : Quantization



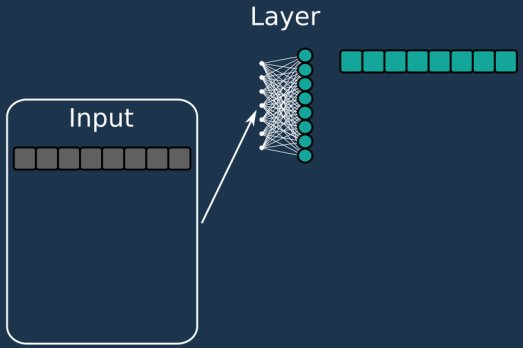
Transformers

Transformers

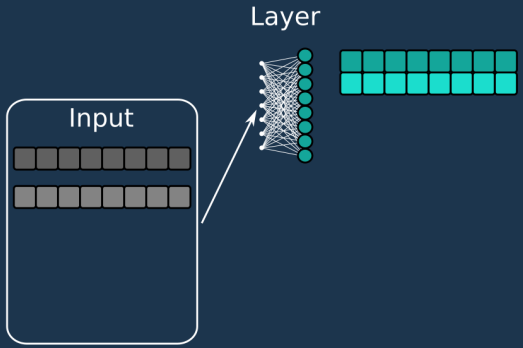
Transformers



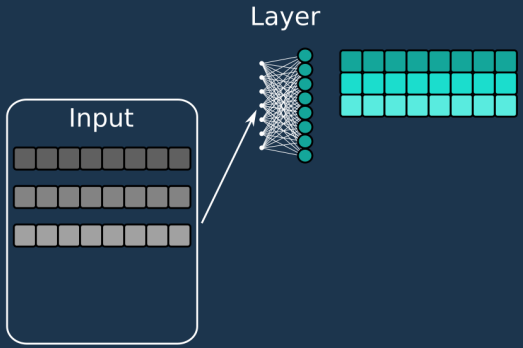
Transformers



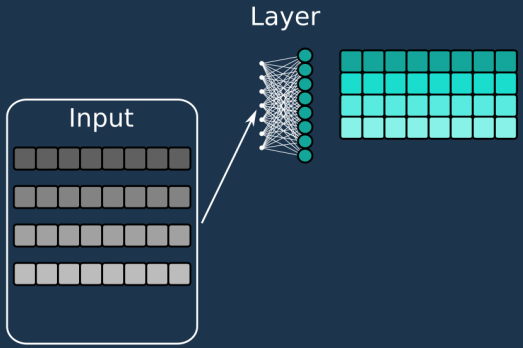
Transformers



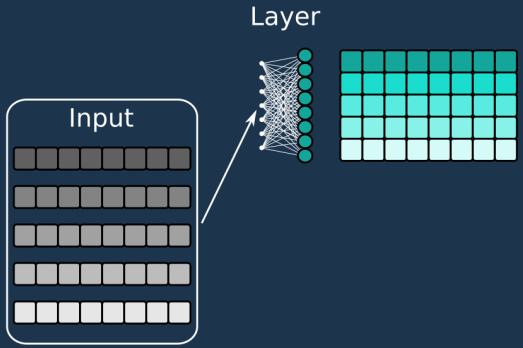
Transformers



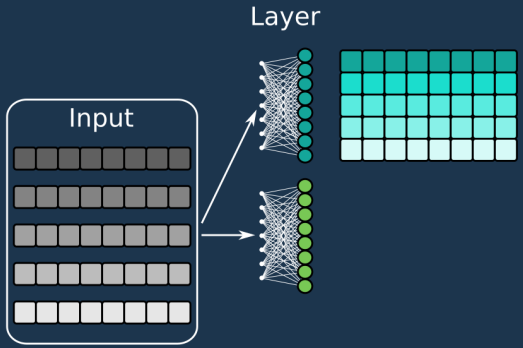
Transformers



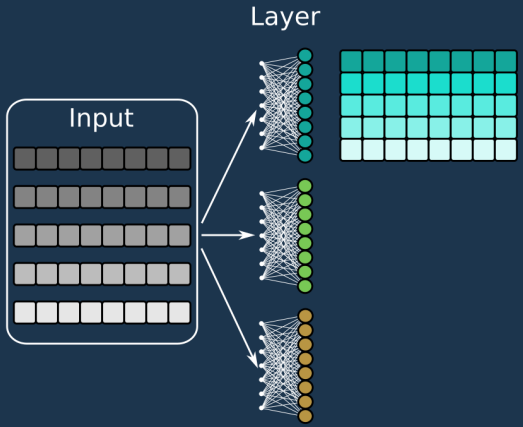
Transformers



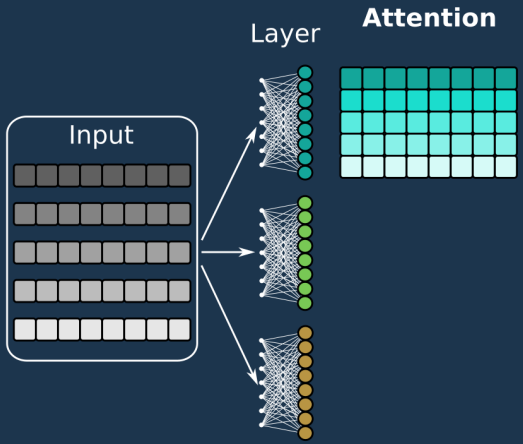
Transformers



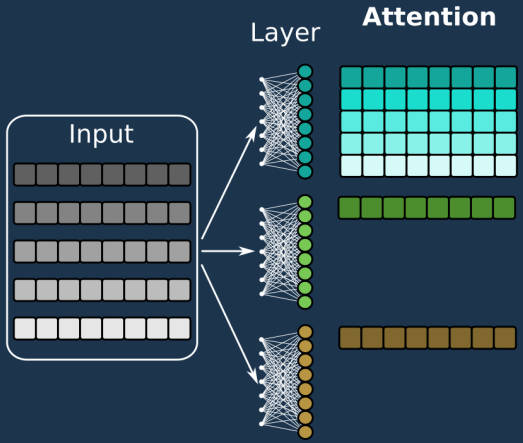
Transformers



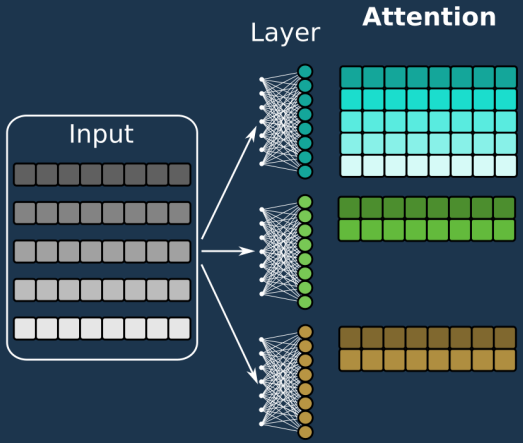
Transformers



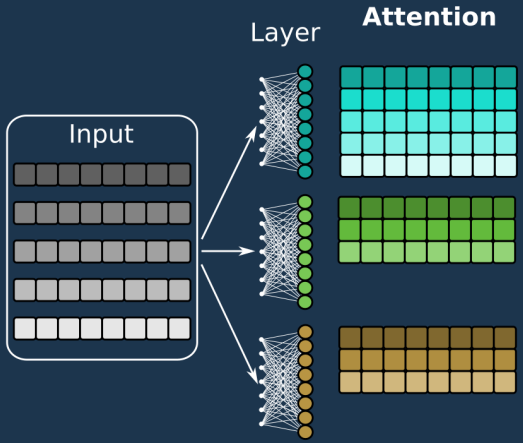
Transformers



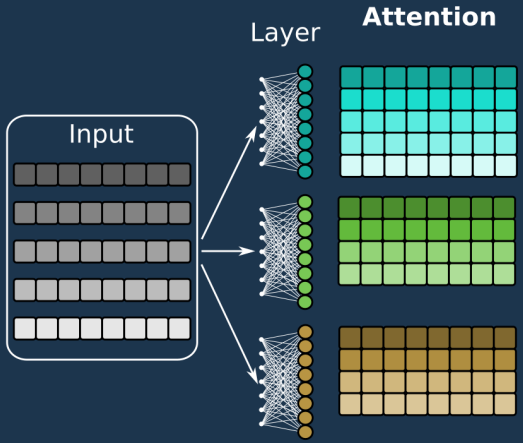
Transformers



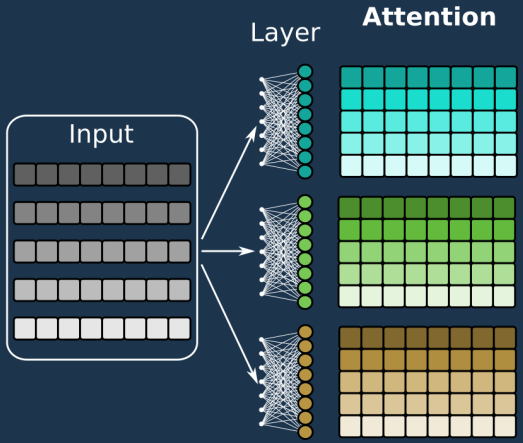
Transformers



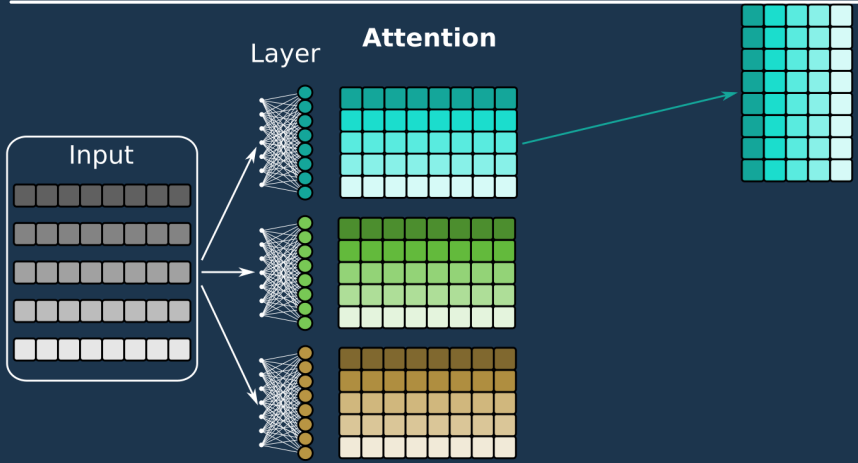
Transformers



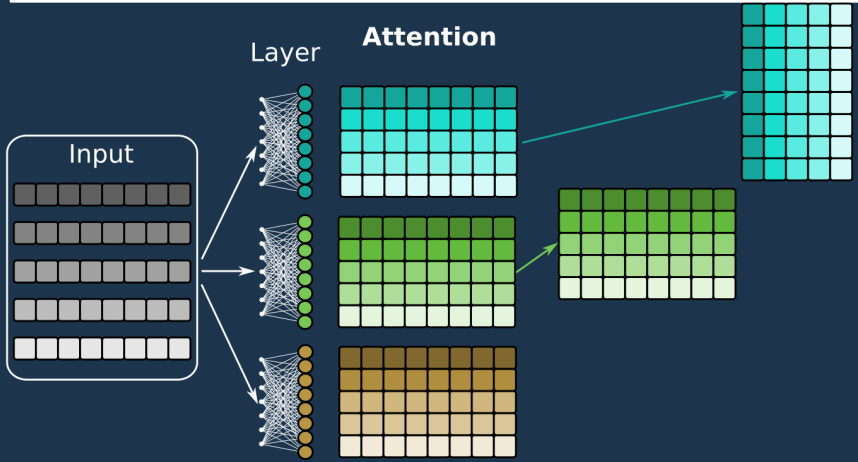
Transformers



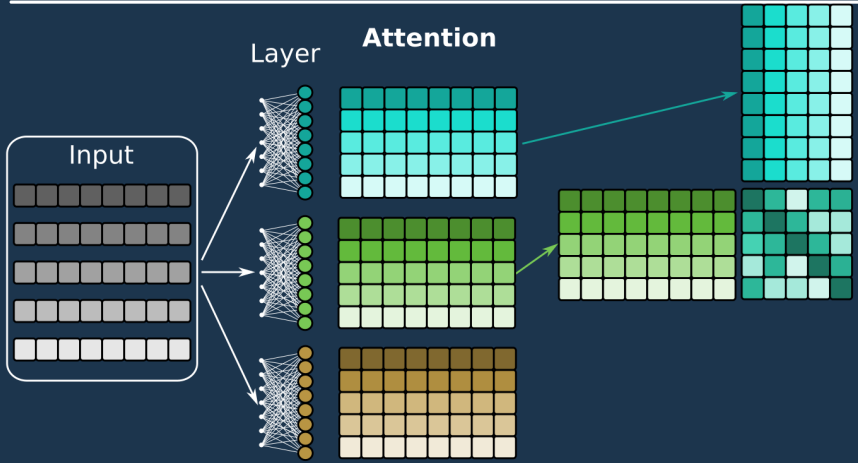
Transformers



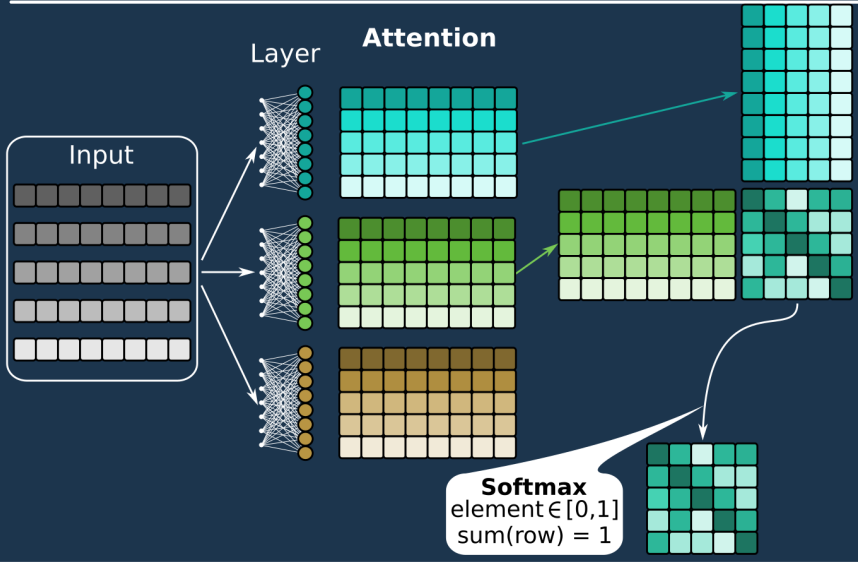
Transformers



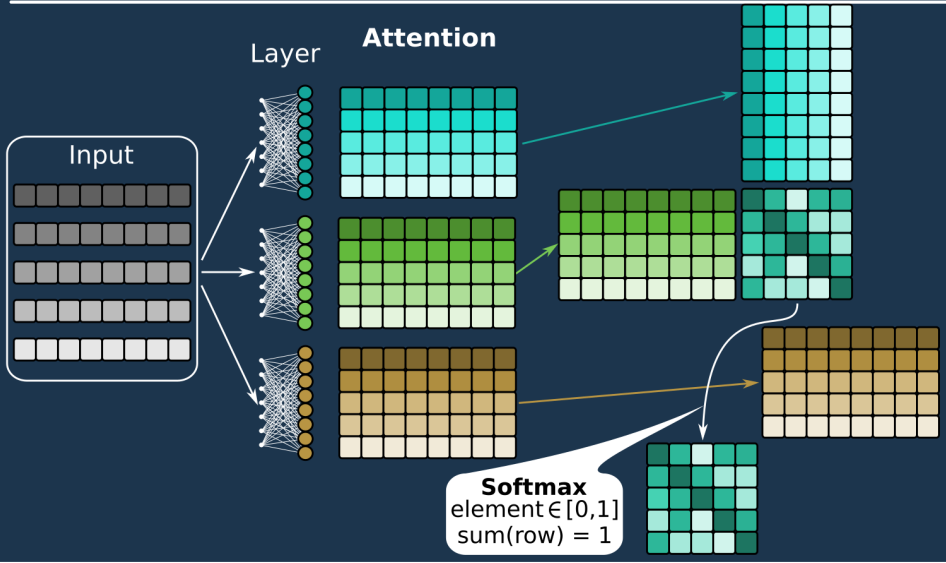
Transformers



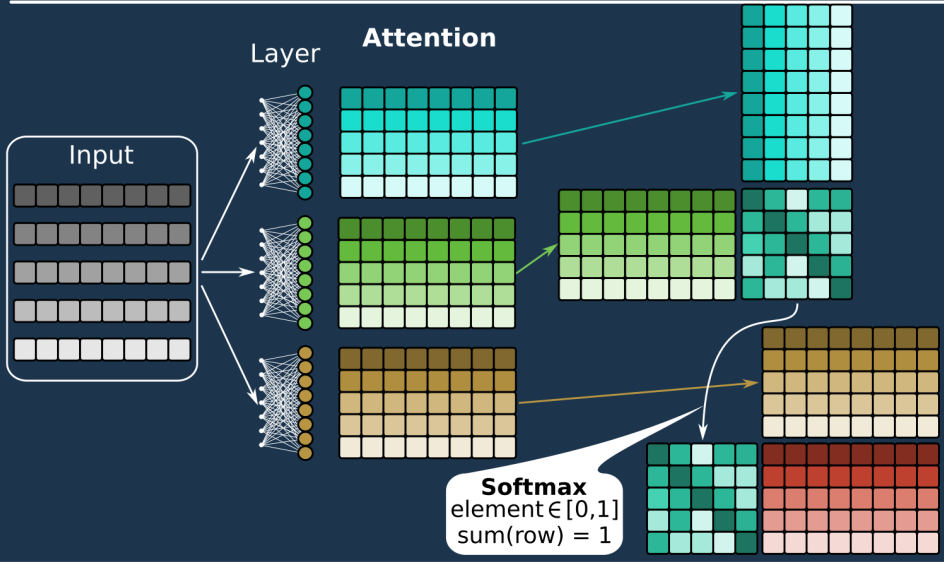
Transformers



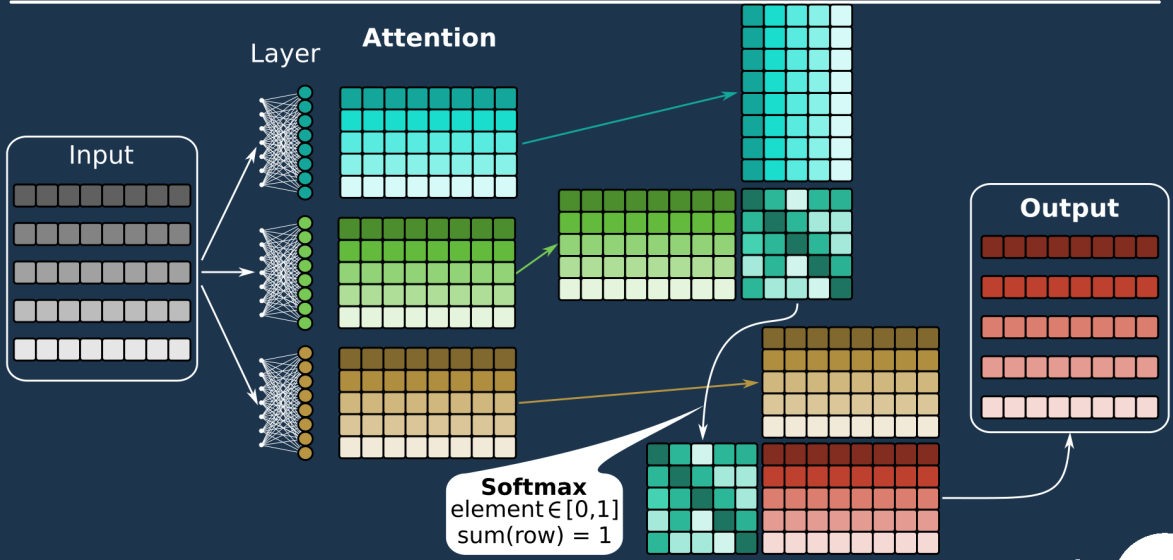
Transformers



Transformers



Transformers



Visual Transformers

Visual Transformers

Self-Attention :

Visual Transformers

Self-Attention :

Patch :

Swin Transformers

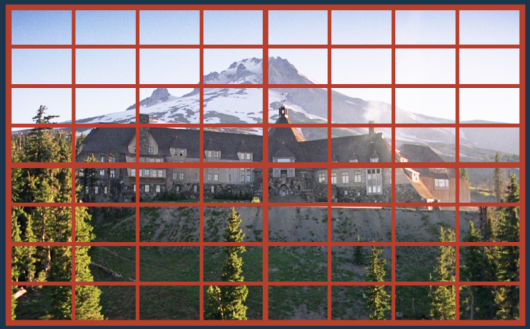
Swin Transformers

Self-Attention : **Patch :**

Swin Transformers

Self-Attention : **Patch :**

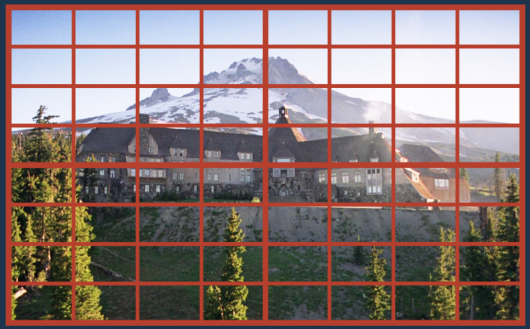
Visual Transformer



Swin Transformers

Self-Attention : Patch :

Visual Transformer

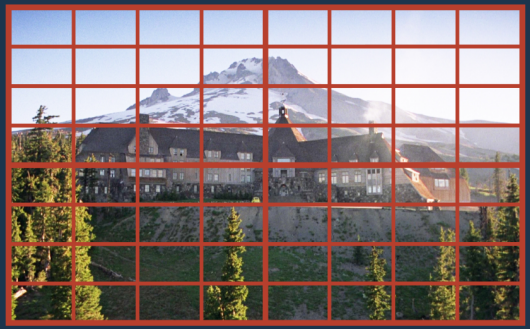


Global Attention

Swin Transformers

Self-Attention : Patch :

Visual Transformer



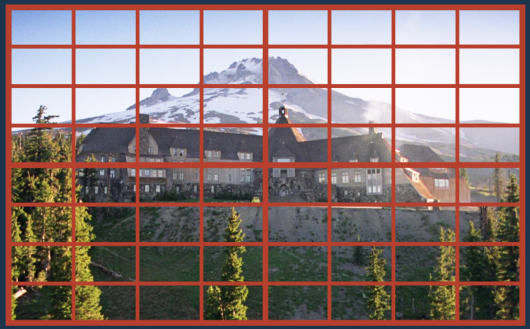
Global Attention

Swin Transformer

Swin Transformers

Self-Attention : Patch :

Visual Transformer



Global Attention

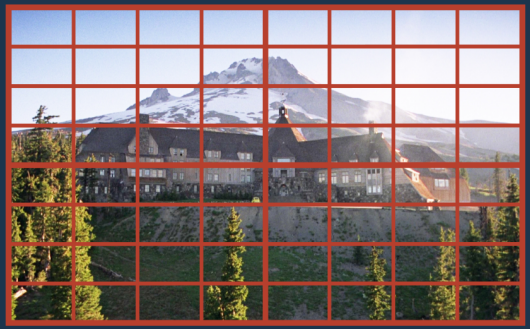
Swin Transformer

Local Attention

Swin Transformers

Self-Attention : Patch :

Visual Transformer



Global Attention

Swin Transformer

Local Attention
16x less computation

Swin Transformers

Self-Attention : Patch :

Visual Transformer

Swin Transformer

Swin Transformers

Self-Attention : Patch :

Visual Transformer

Swin Transformer

Sliding windows

Swin Transformers

Self-Attention : Patch :

Visual Transformer

Swin Transformer

Sliding windows

Swin Transformers

Self-Attention : Patch :

Visual Transformer

Sliding windows

Swin Transformer

Swin Transformers

Self-Attention : Patch :

Visual Transformer

Sliding windows

Swin Transformer

Shifted windows

Swin Transformers

Self-Attention : **Patch :**

Swin Transformers

Self-Attention : Patch :

Layer i

Swin Transformers

Swin Transformers

Self-Attention : Patch :

Layer i

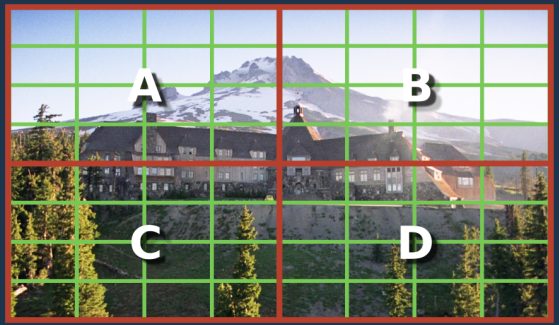
Layer i+1

Shifted non-overlapping windows

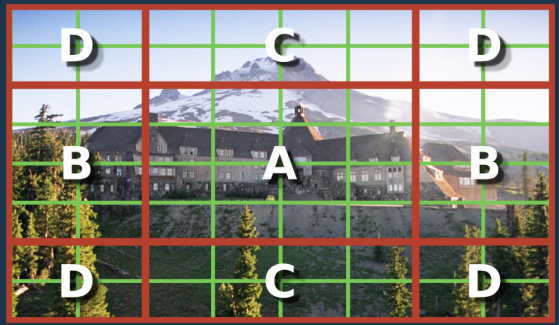
Swin Transformers

Self-Attention : Patch :

Layer i



Layer i+1



Shifted non-overlapping windows

Swin Transformers V2 : Performances

Swin V2 = Swin V1 + Computing Optimization (Nsight)

Swin Transformers V2 : Performances

Swin V2 = Swin V1 + Computing Optimization (Nsight)

Model **Size** and Training **Dataset**

Swin Transformers V2 : Performances

Swin V2 = Swin V1 + Computing Optimization (Nsight)

Model **Size** and Training **Dataset**

ViT-G :
- **1.8 G** parameters
- **3 G** images

CoAtNet-7 :
- **2.4 G** parameters
- **3 G** images

Swin Transformers V2 : Performances

Swin V2 = Swin V1 + Computing Optimization (Nsight)

Model **Size** and Training **Dataset**

ViT-G :
- **1.8 G** parameters
- **3 G** images

CoAtNet-7 :
- **2.4 G** parameters
- **3 G** images

Swin V2 :
- **3 G** parameters
- **70 M** images

Swin Transformers V2 : Performances

Swin V2 = Swin V1 + Computing Optimization (Nsight)

ViT-G :
- **1.8 G** parameters
- **3 G** images

CoAtNet-7 :
- **2.4 G** parameters
- **3 G** images

Model **Size** and Training **Dataset**

- **25** - **30 %** larger
- **40x** less labelled images
- **10x** lower training cost
- **Applicable** to richer tasks

Swin V2 :
- **3 G** parameters
- **70 M** images

Confidential Computing

Confidential Computing

No trust in computing centers

Confidential Computing

No trust in computing centers → **End-to-end encryption**

Confidential Computing

No trust in computing centers → **End-to-end encryption**

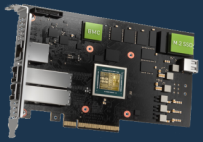
Hardware encryption

Confidential Computing

No trust in computing centers → **End-to-end encryption**

Hardware encryption

Available in **Bluefield DPU**



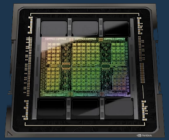
Confidential Computing

No trust in computing centers → **End-to-end encryption**

Hardware encryption

Available in **Bluefield DPU**

Available in **Hopper GPU**

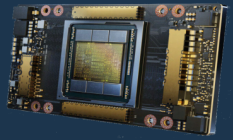


Confidential Computing

No trust in computing centers → **End-to-end encryption**

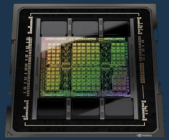
Hardware encryption

Not in **Ampere**



Available in **Bluefield DPU**

Available in **Hopper GPU**



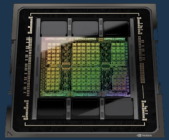
Confidential Computing

No trust in computing centers → **End-to-end encryption**

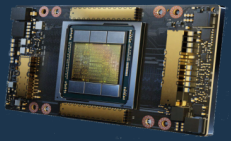
Hardware encryption

Available in **Bluefield DPU**

Available in **Hopper GPU**



Not in **Ampere**



Software encryption
even on **NVLink**

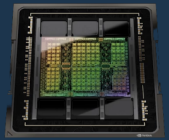
Confidential Computing

No trust in computing centers → **End-to-end encryption**

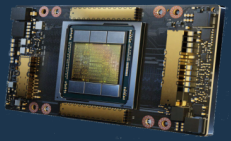
Hardware encryption

Available in **Bluefield DPU**

Available in **Hopper GPU**



Not in **Ampere**



Software encryption
even on **NVLink**

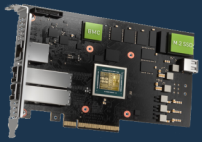
Homomorphic Encryption

Confidential Computing

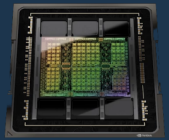
No trust in computing centers → **End-to-end encryption**

Hardware encryption

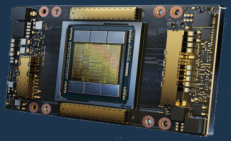
Available in **Bluefield DPU**



Available in **Hopper GPU**



Not in **Ampere**



Software encryption
even on **NVLink**

Homomorphic Encryption

Encrypted computation
on **encrypted Data**

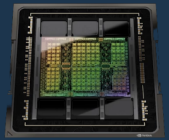
Confidential Computing

No trust in computing centers → **End-to-end encryption**

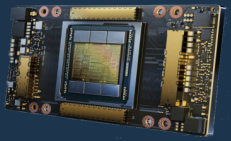
Hardware encryption

Available in **Bluefield DPU**

Available in **Hopper GPU**



Not in **Ampere**



Software encryption
even on **NVLink**

Homomorphic Encryption

Encrypted computation
on **encrypted Data**

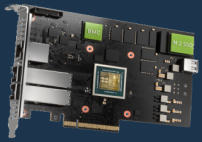
Computation -> logical gates

Confidential Computing

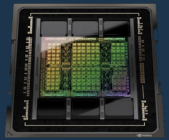
No trust in computing centers → **End-to-end encryption**

Hardware encryption

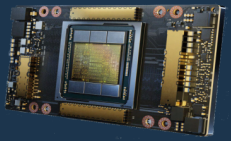
Available in **Bluefield DPU**



Available in **Hopper GPU**



Not in **Ampere**



Software encryption
even on **NVLink**

Homomorphic Encryption

Encrypted computation
on **encrypted Data**

Computation -> logical gates

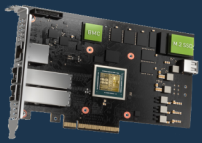
Very slow

Confidential Computing

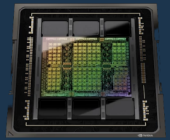
No trust in computing centers → **End-to-end encryption**

Hardware encryption

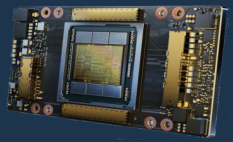
Available in **Bluefield DPU**



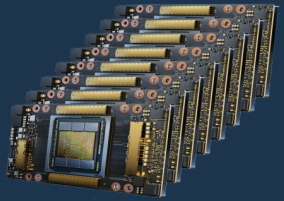
Available in **Hopper GPU**



Not in **Ampere**



Dot product of
2 vectors of **500** elements (int)
on **8 A100 GPU**



Software encryption
even on **NVLink**

Homomorphic Encryption

Encrypted computation
on **encrypted Data**

Computation -> logical gates

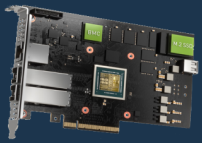
Very slow

Confidential Computing

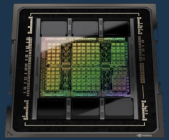
No trust in computing centers → **End-to-end encryption**

Hardware encryption

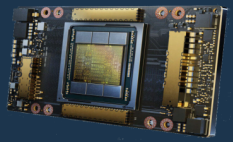
Available in **Bluefield DPU**



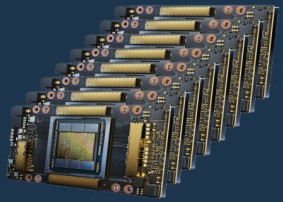
Available in **Hopper GPU**



Not in **Ampere**



Dot product of
2 vectors of **500** elements (int)
on **8 A100 GPU** takes **20 s**



Software encryption
even on **NVLink**

Homomorphic Encryption

Encrypted computation
on **encrypted Data**

Computation -> logical gates

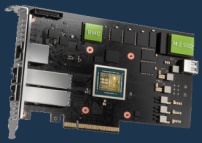
Very slow

Confidential Computing

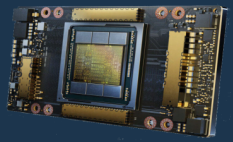
No trust in computing centers → **End-to-end encryption**

Hardware encryption

Available in **Bluefield DPU**



Not in **Ampere**



Software encryption
even on **NVLink**

Homomorphic Encryption

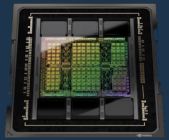
Encrypted computation
on **encrypted Data**

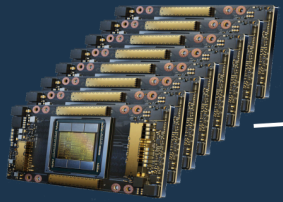
Computation -> logical gates

Very slow

Dot product of
2 vectors of **500** elements (int)
on **8 A100 GPU** takes **20 s**

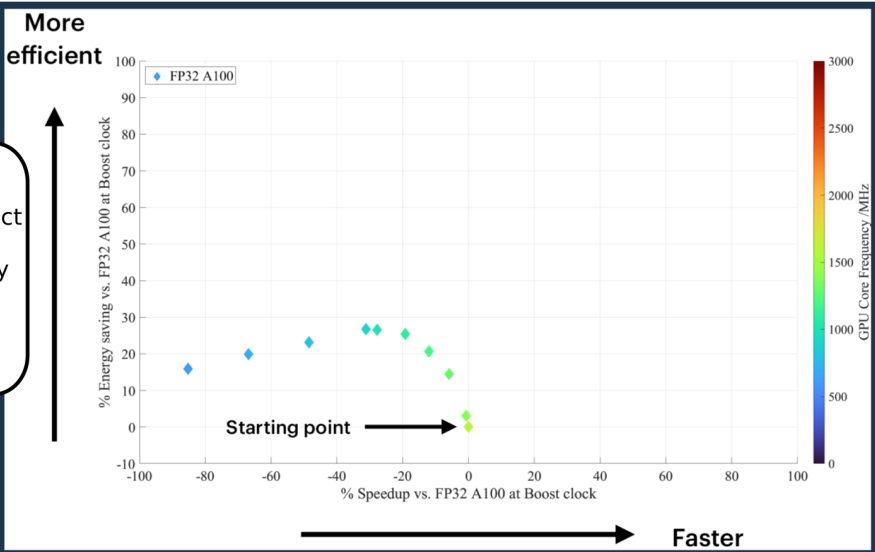
Available in **Hopper GPU**





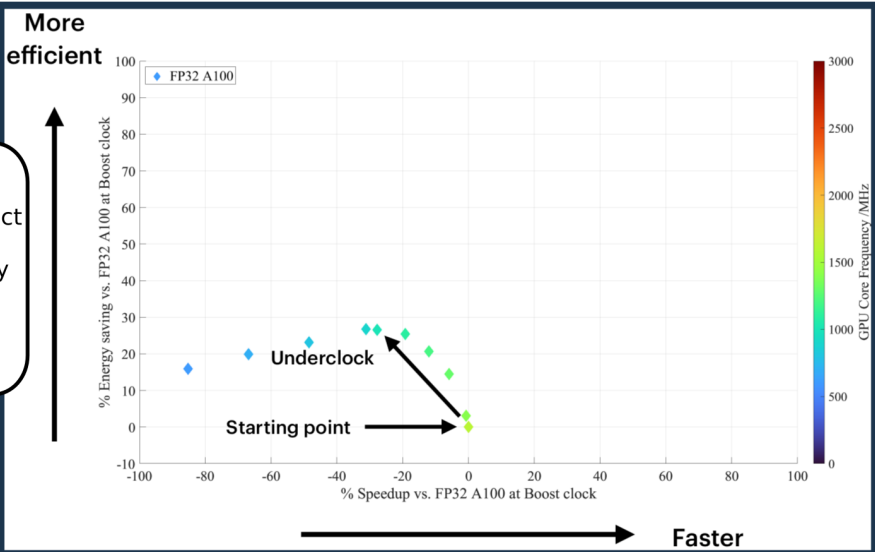
GPU Consumption VS Frequency

S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



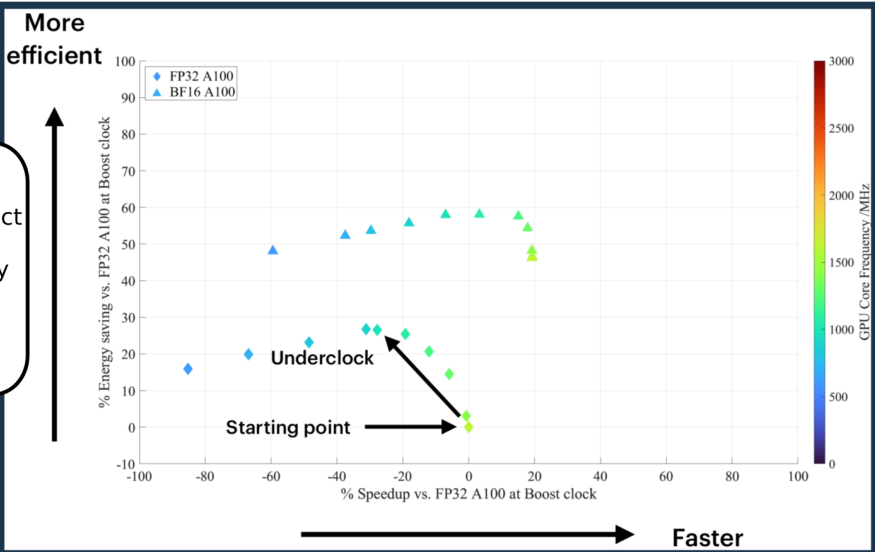
GPU Consumption VS Frequency

S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



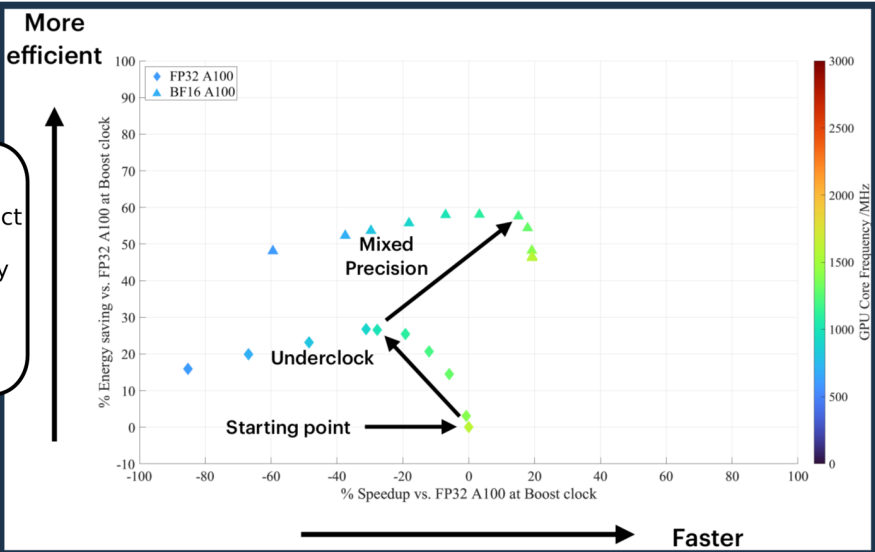
GPU Consumption VS Frequency

S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



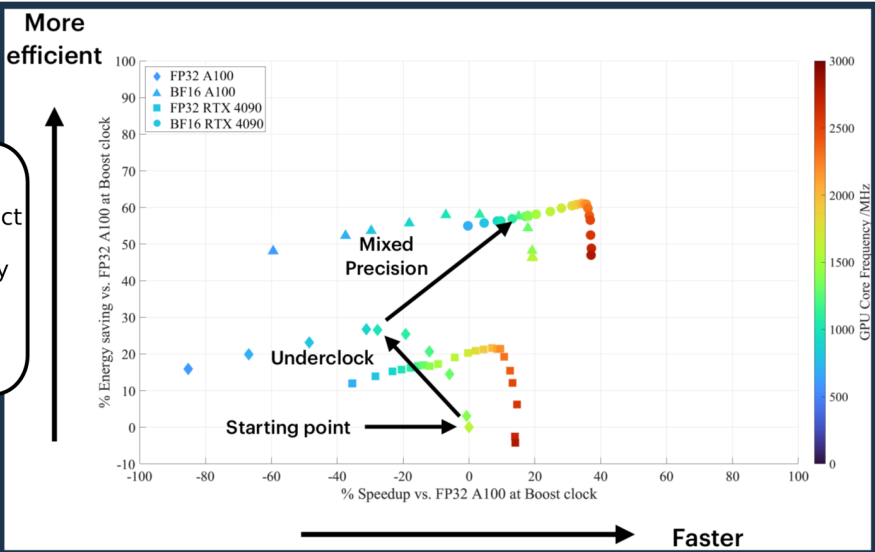
GPU Consumption VS Frequency

S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



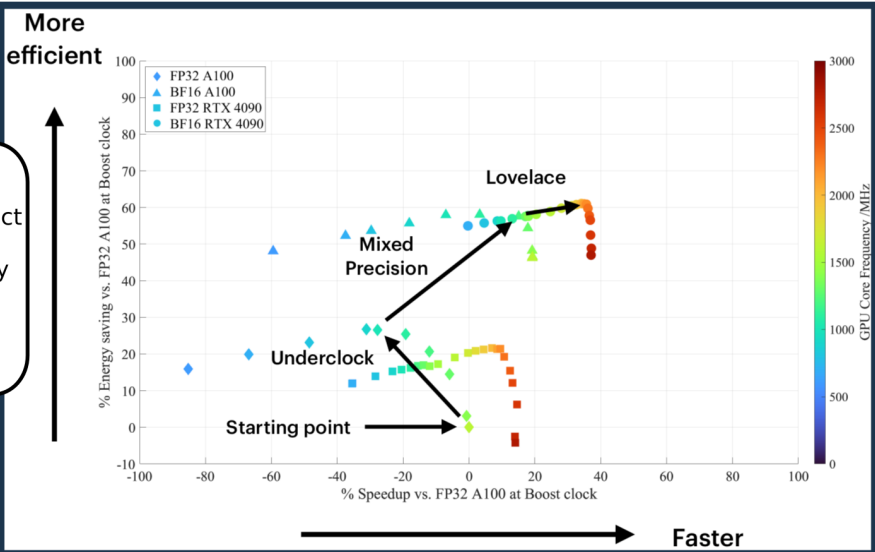
GPU Consumption VS Frequency

S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



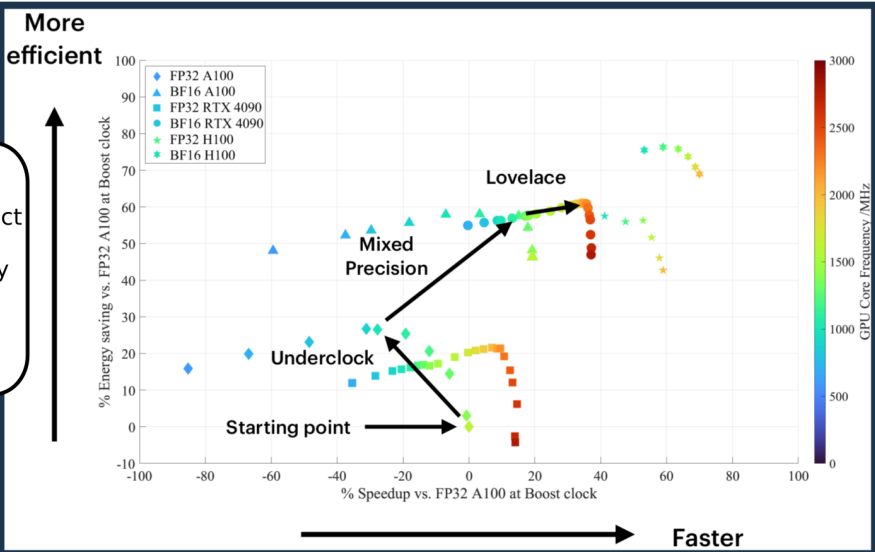
GPU Consumption VS Frequency

S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



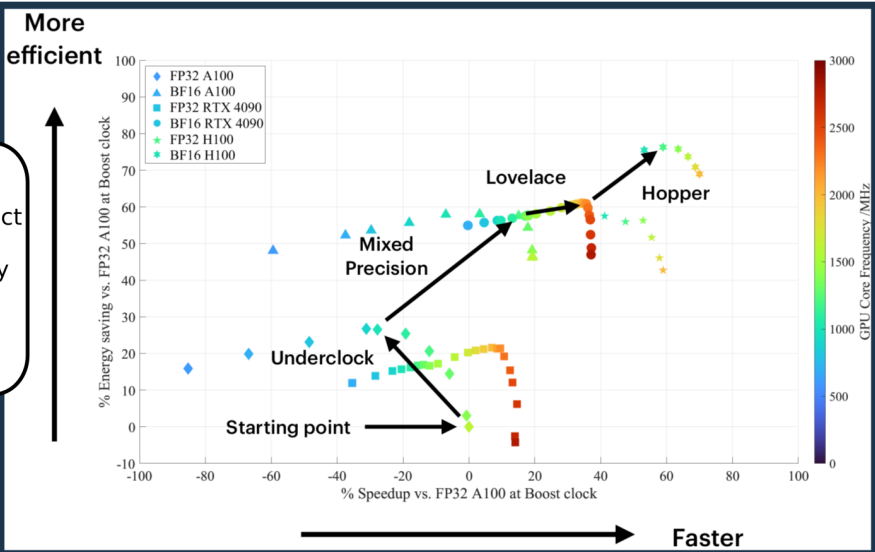
GPU Consumption VS Frequency

S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



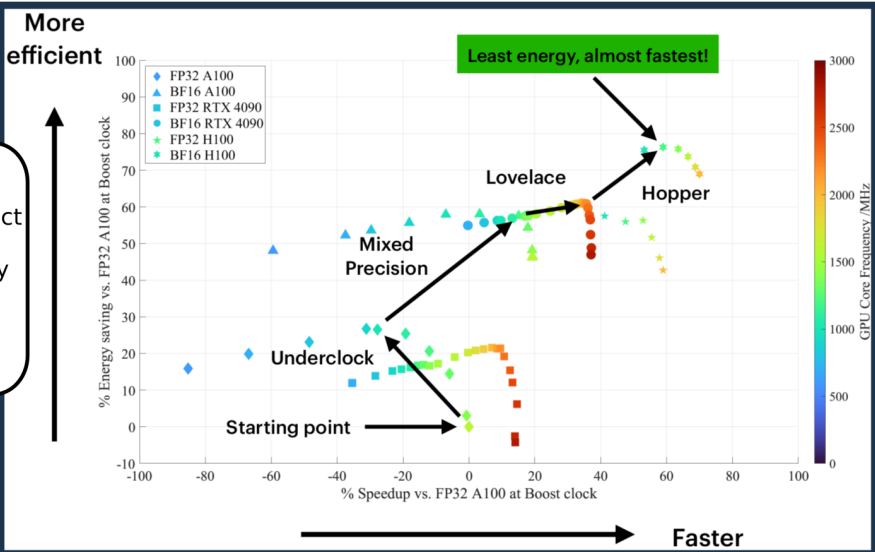
GPU Consumption VS Frequency

S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



GPU Consumption VS Frequency

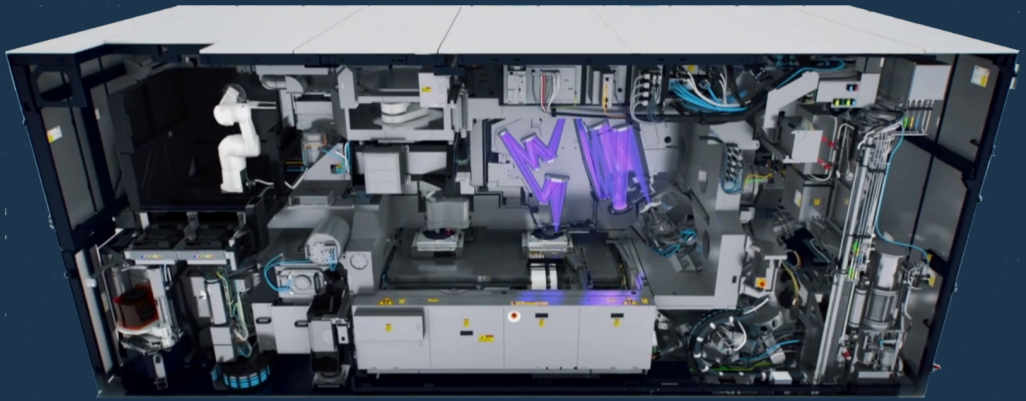
S51444
Reducing the Environmental Impact of HPC using Dynamic Frequency Scaling
Jack White
Dr. Karel Adámek
Prof. Wes Armour



Photolithography

Photolithography

ASML



Photolithography

Photolithography

Photolithography

Photolithography

Light Source

Photolithography

Photolithography

Light Source

Lens

Photolithography

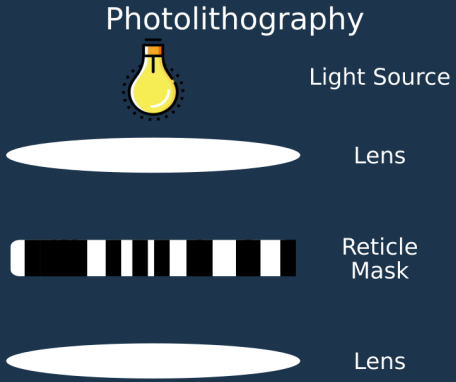
Photolithography

Light Source

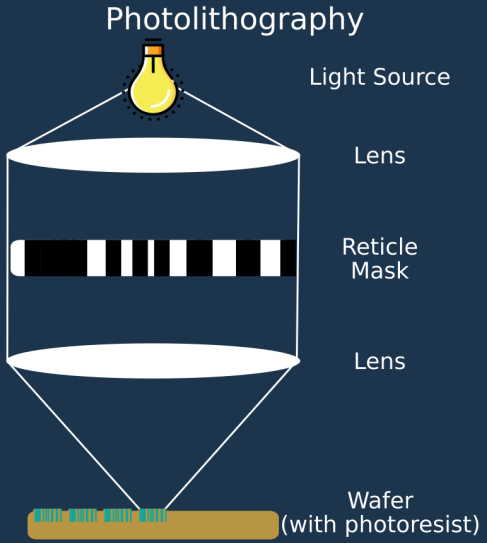
Lens

Reticle
Mask

Photolithography



Photolithography

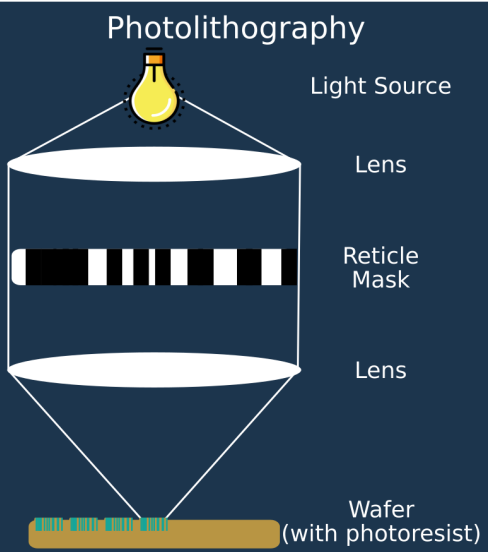


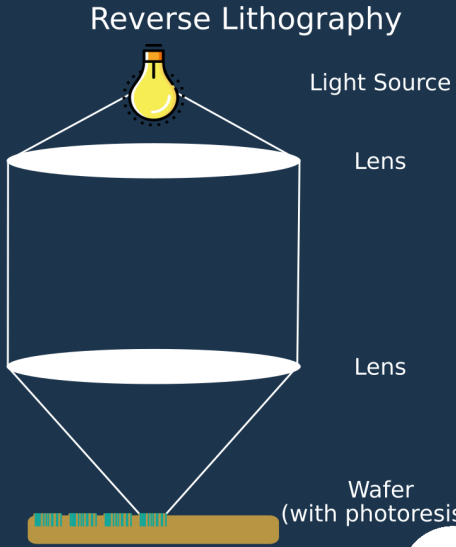
Photolithography

Photolithography

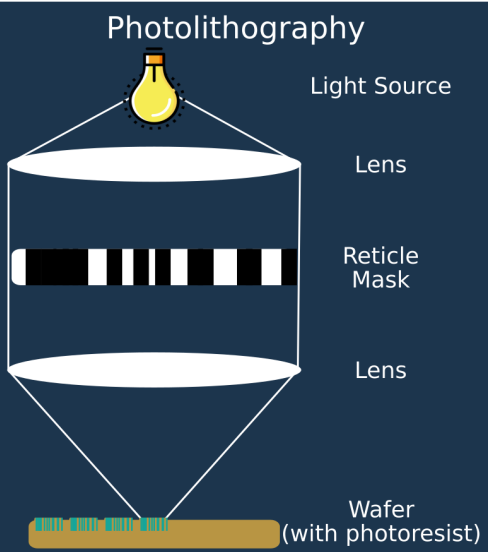
Reverse Lithography

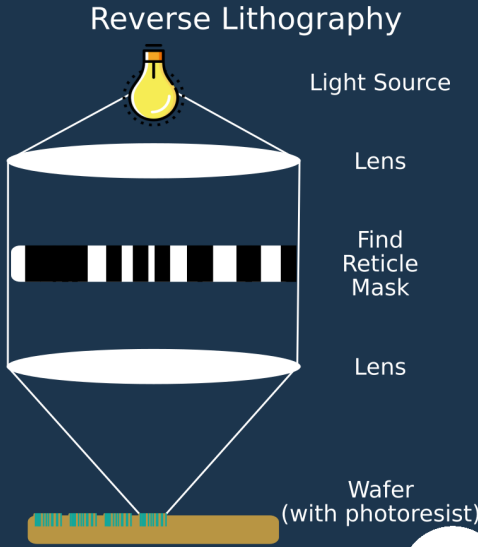
Photolithography



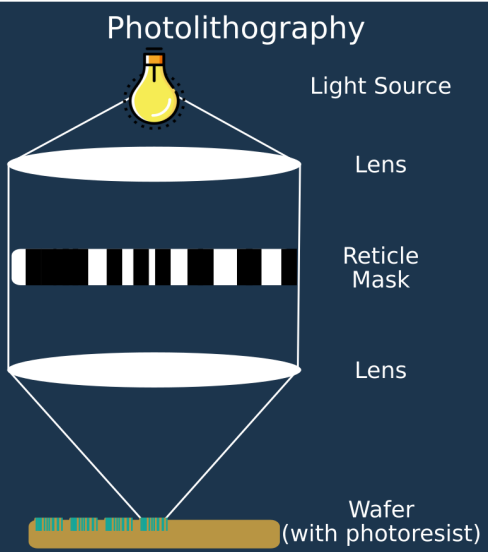


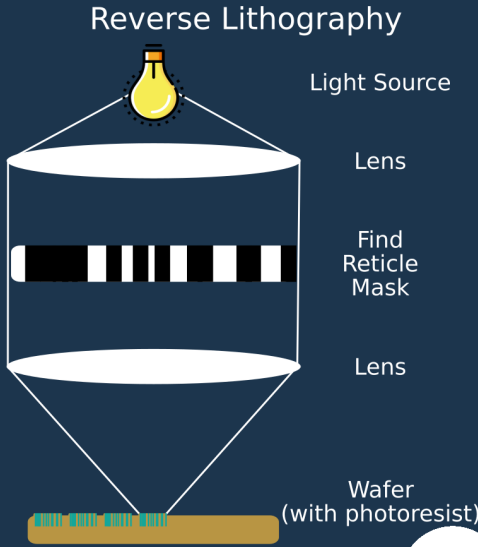
Photolithography



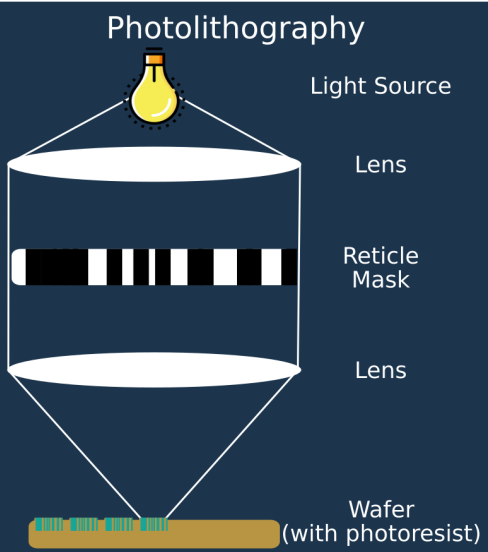


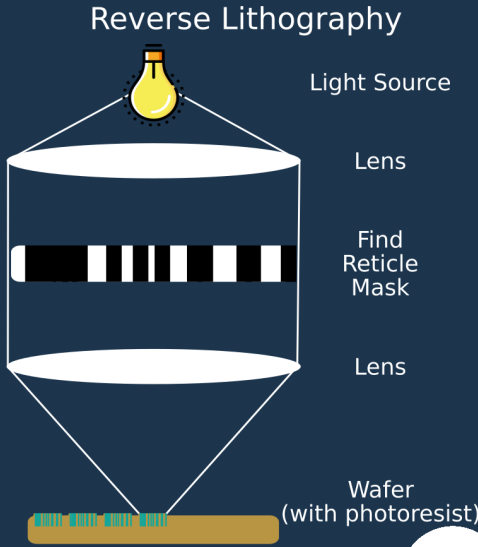
Photolithography



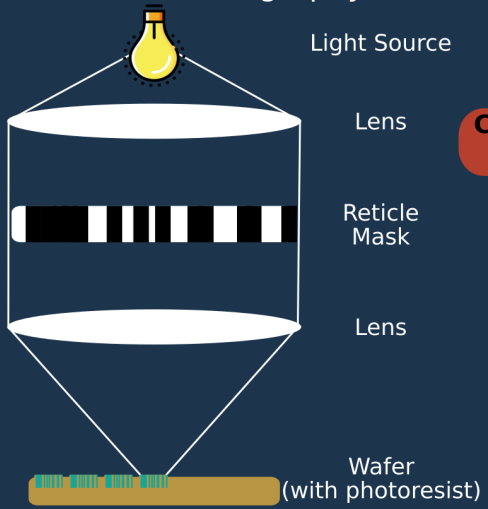


Photolithography





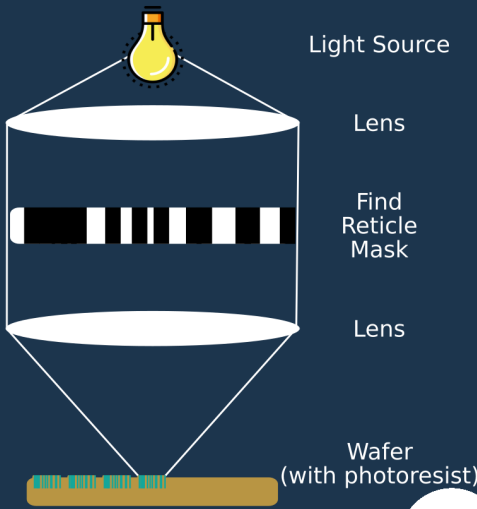
Photolithography



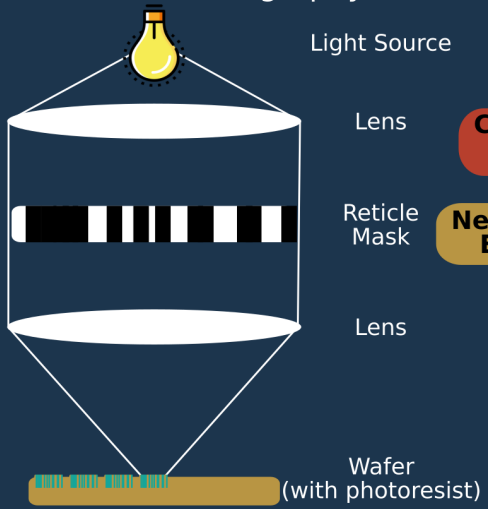
3 nm

UV

Cannot use Optic



Photolithography

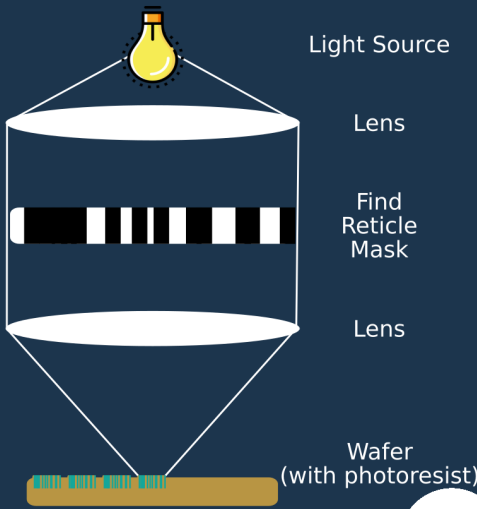


3 nm

UV

Cannot use Optic

Need Maxwell Equations



Photolithography

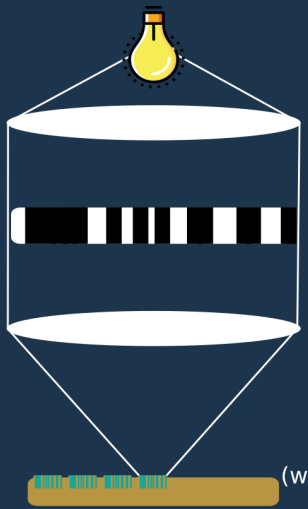
3 nm

UV

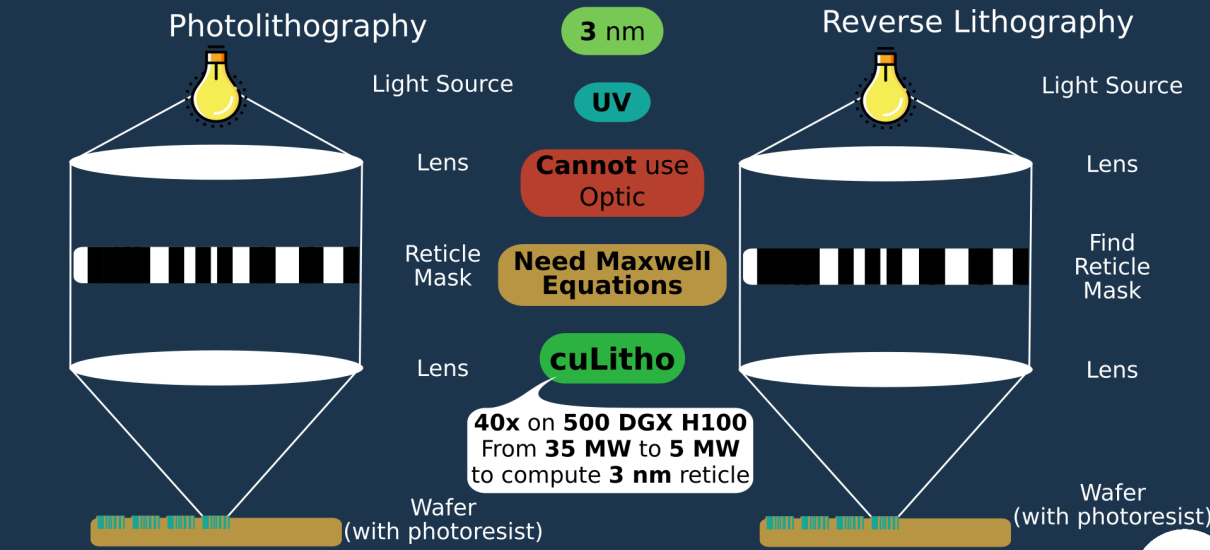
Cannot use Optic

Need Maxwell Equations

cuLitho



Photolithography



Link to a big summary

GTC 2023 notes from Reprises website