GPU FUNDAMENTALS

TOPICS

CPUs and Serial Processing

Power of Parallel Processing with GPUs

GPU Computing in Enterprises

CPUS AND SERIAL PROCESSING

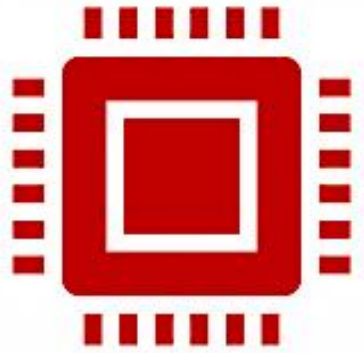
HOW A COMPUTER WORKS

Typical Computer

Input
- Keyboard
- Mouse
- Camera/Mic
- Touch pad

→

CPU



Memory

→

Output
- Solved Problem
- Games
- Image, Video or VR
- Control Signals

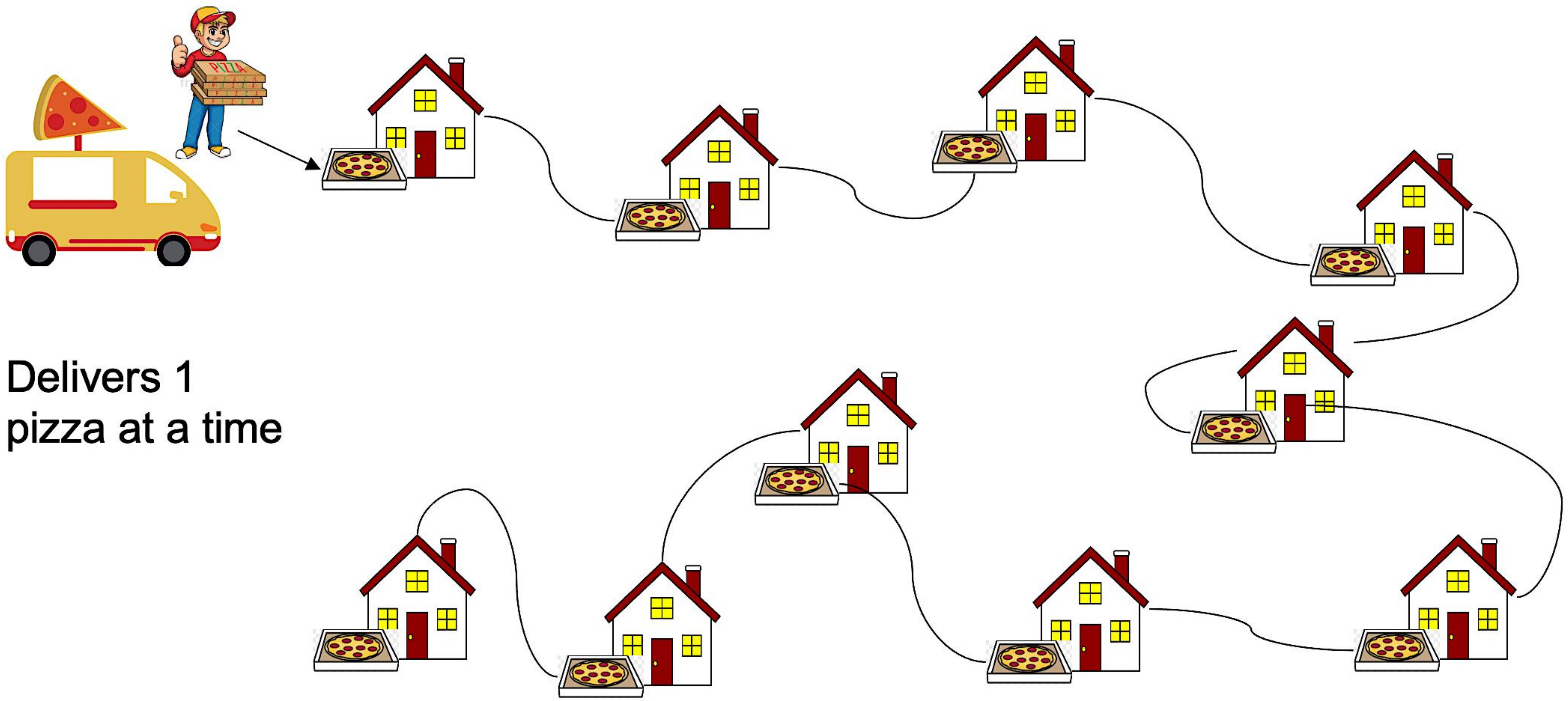
Ingest Data

Store & Process Data

Show Data

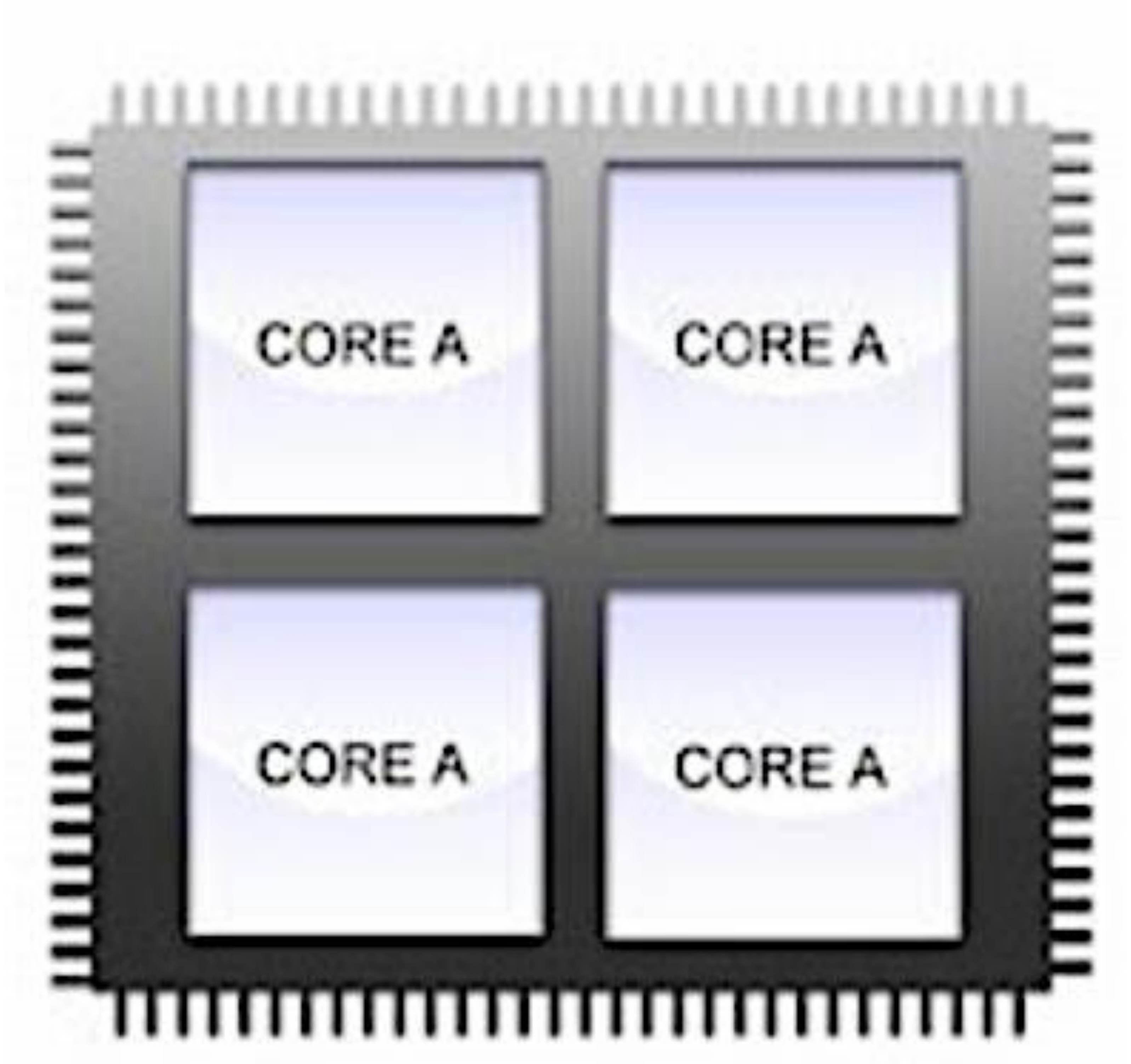
HOW A CPU WORKS

WHAT CPU-BASED SERIAL PROCESSING LOOKS LIKE



Delivers 1
pizza at a time

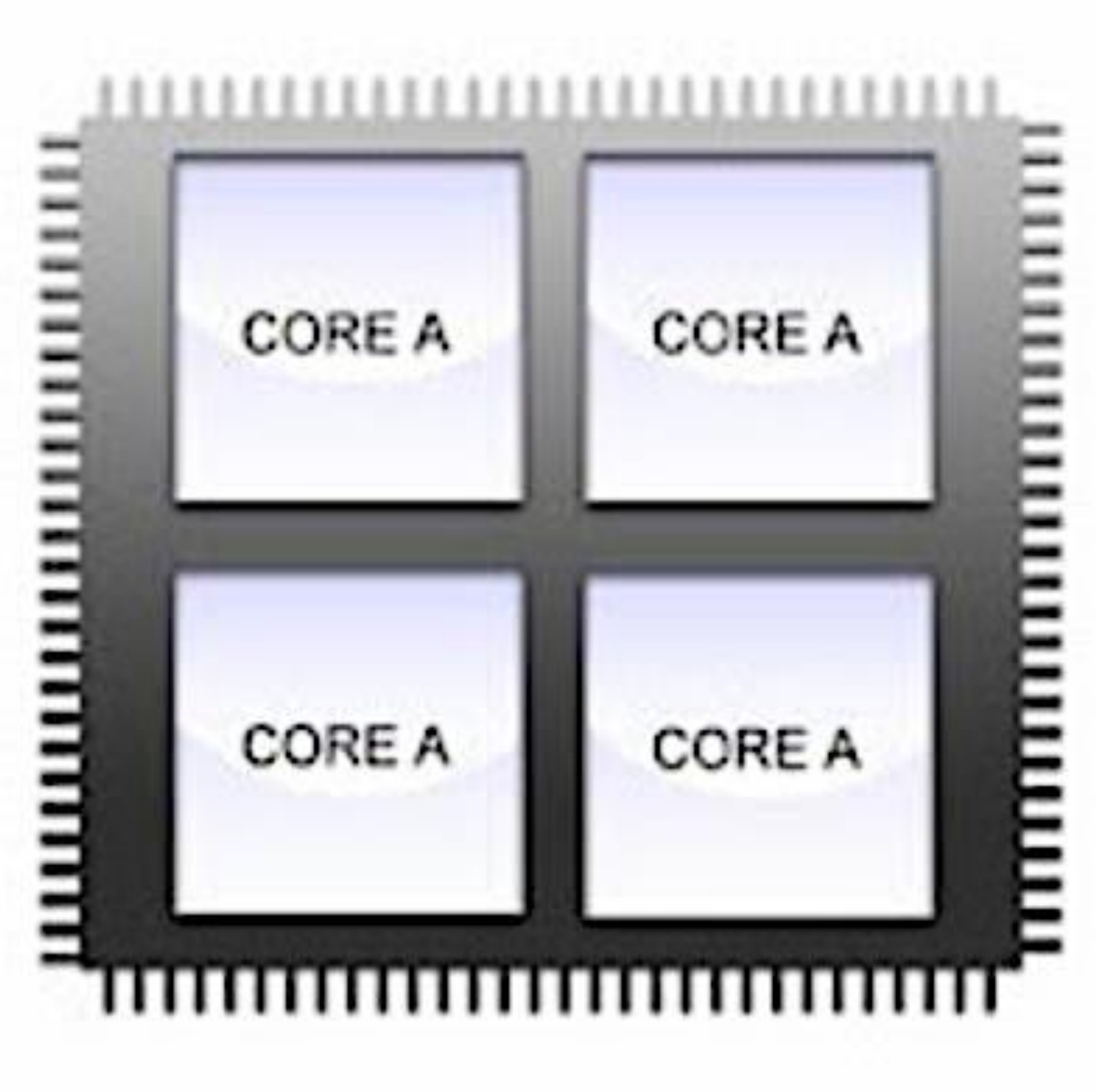
CPU-BASED MULTICORE PROCESSING

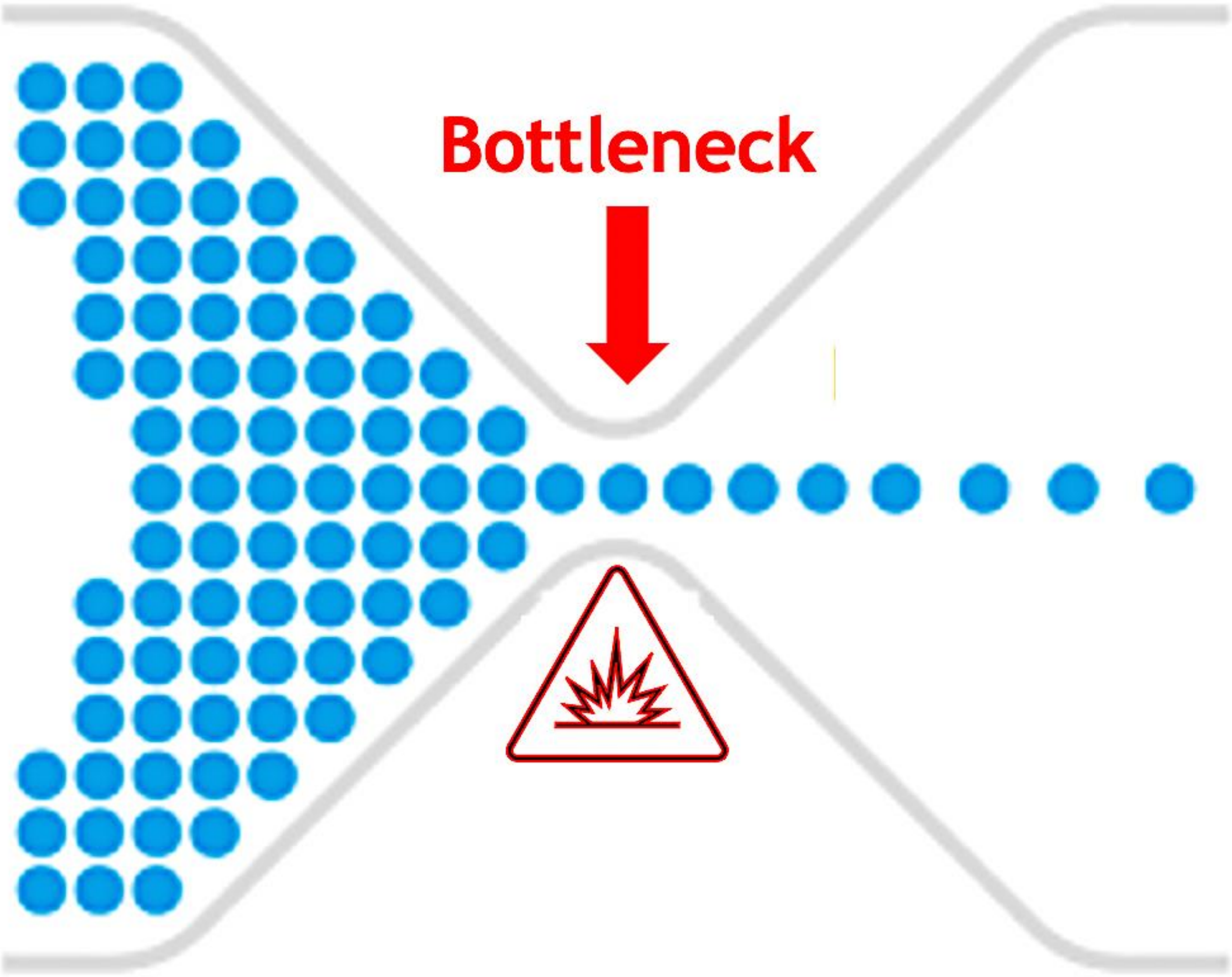


Multicore CPU

Many pizzas at a time

DRAWBACKS OF CPU-BASED MULTICORE PROCESSING





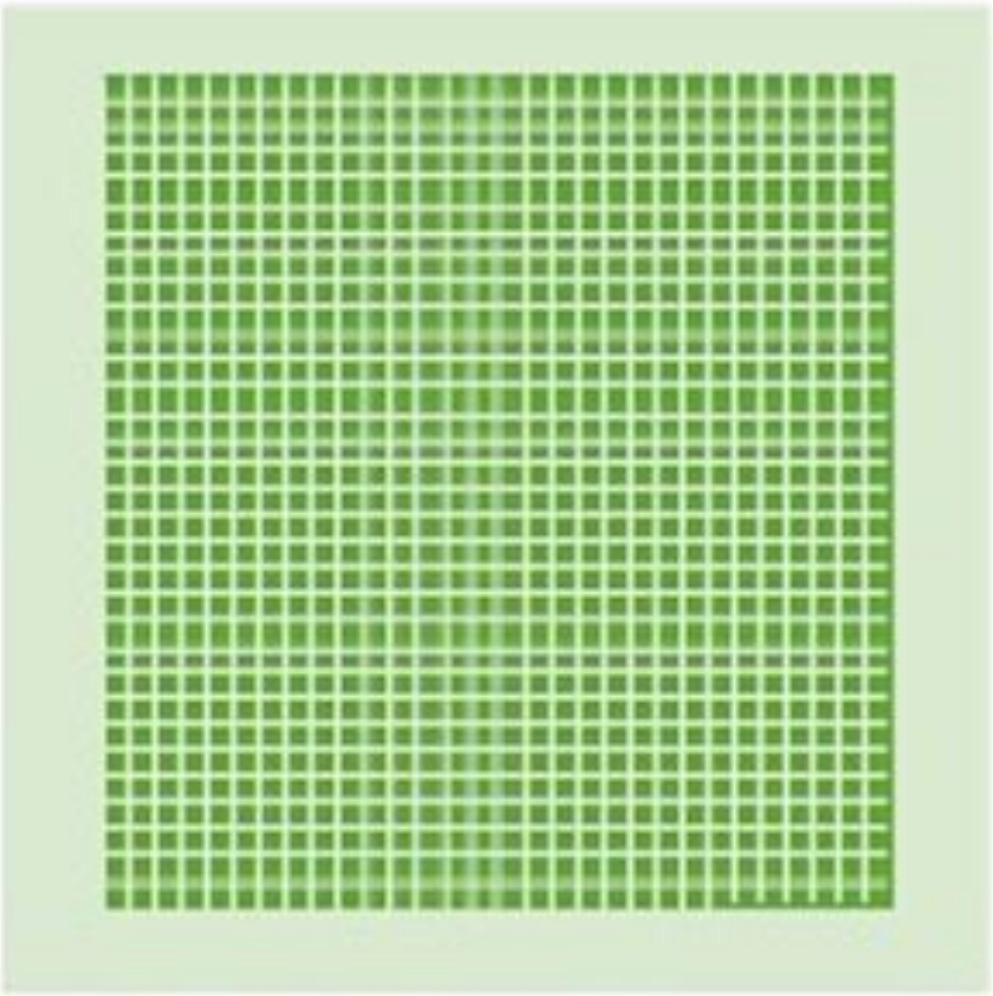
POWER OF PARALLEL PROCESSING WITH GPUS

INVENTION OF THE GPU

CPU Architecture

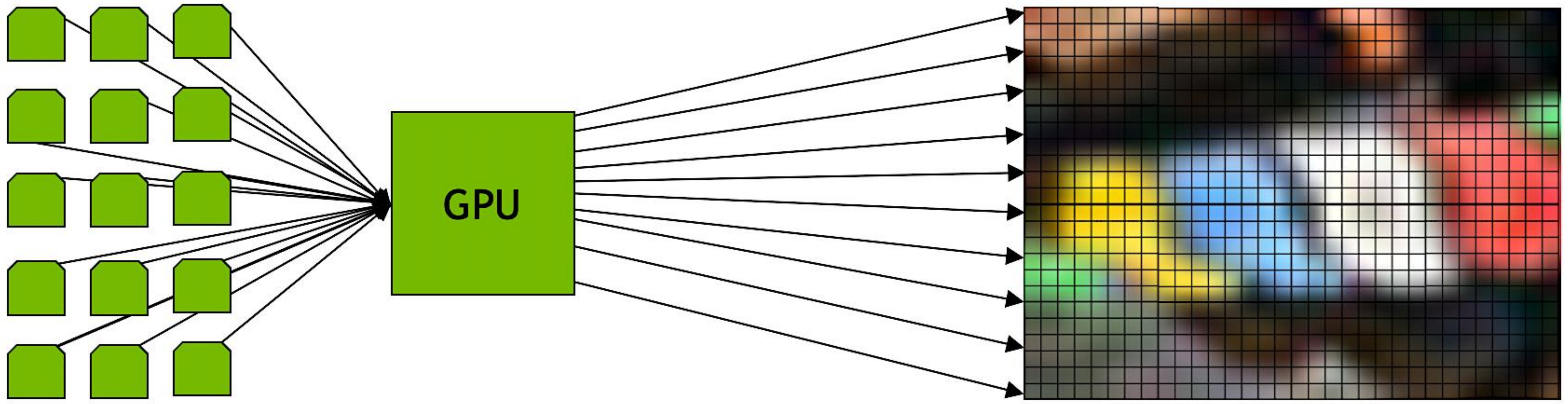
multiple
high performance
cores

GPU Architecture



1000s of
light weight cores

Parallel Computing on GPUs
Designed for many processes at the same time

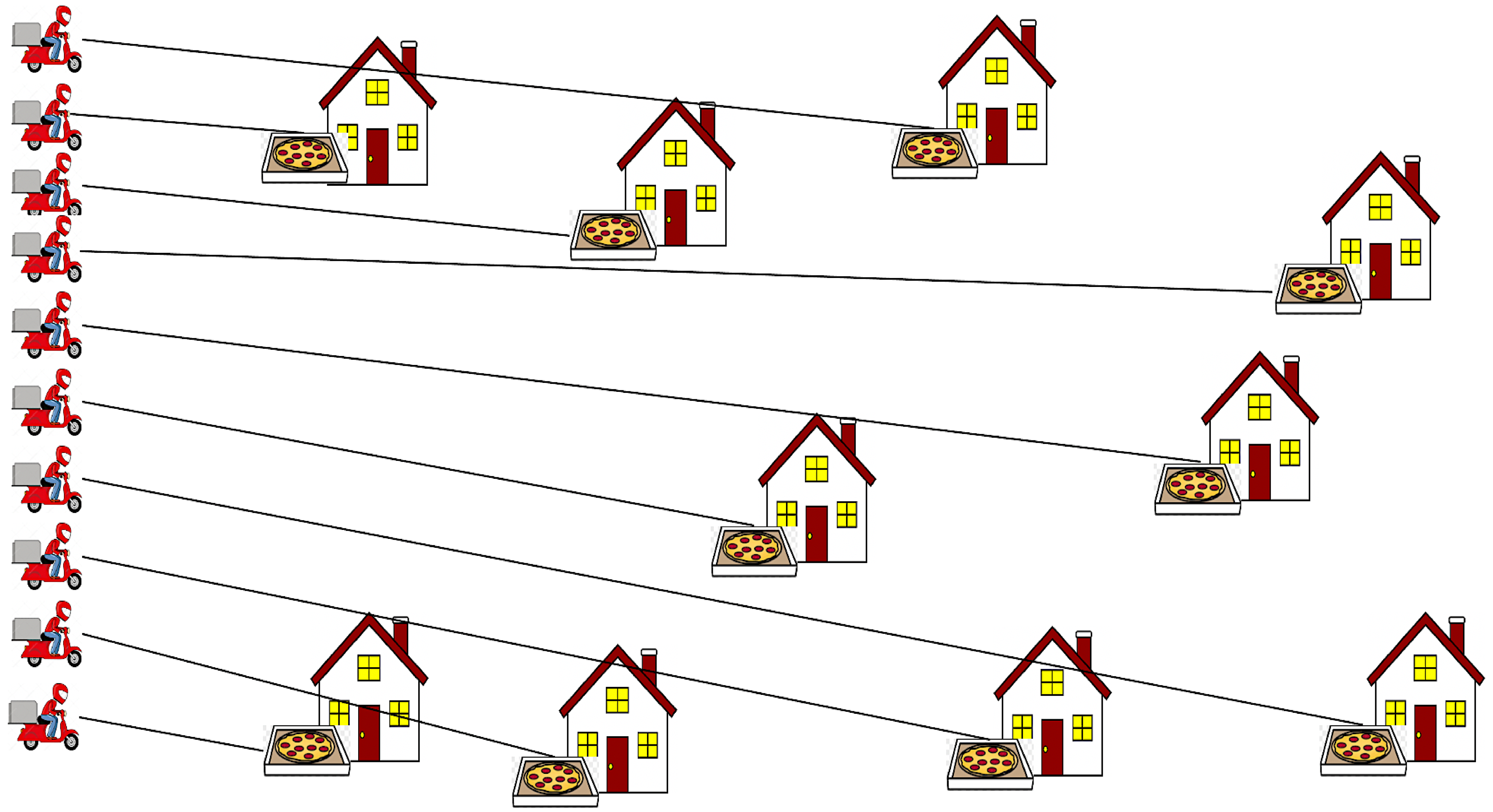


Parallel acceleration

SUPER FAST, SUPER EFFICIENT

SPEED AND EFFICIENCY OF GPU COMPUTING

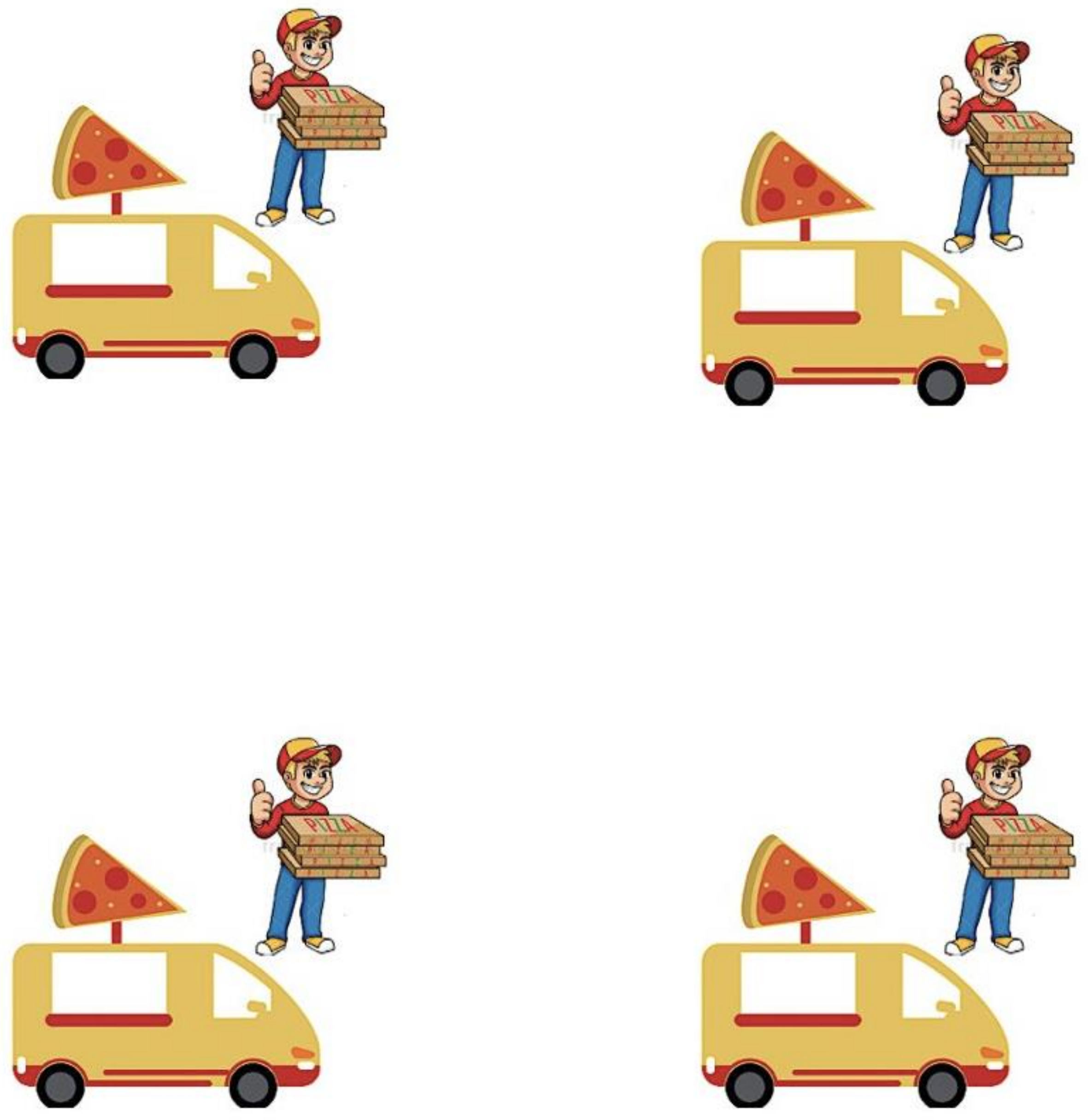
GPU-based Computing



Delivers many pizzas at the same time
Fast and efficient

Fast Scooters
+
Huge Trucks
=
Best Performance

CPU-based Computing



Delivers larger pizza orders
Reliable but slow

CPU AND GPU COLLABORATION

Central Processing Unit
4-8 Cores
Good for serial processing
Low Latency

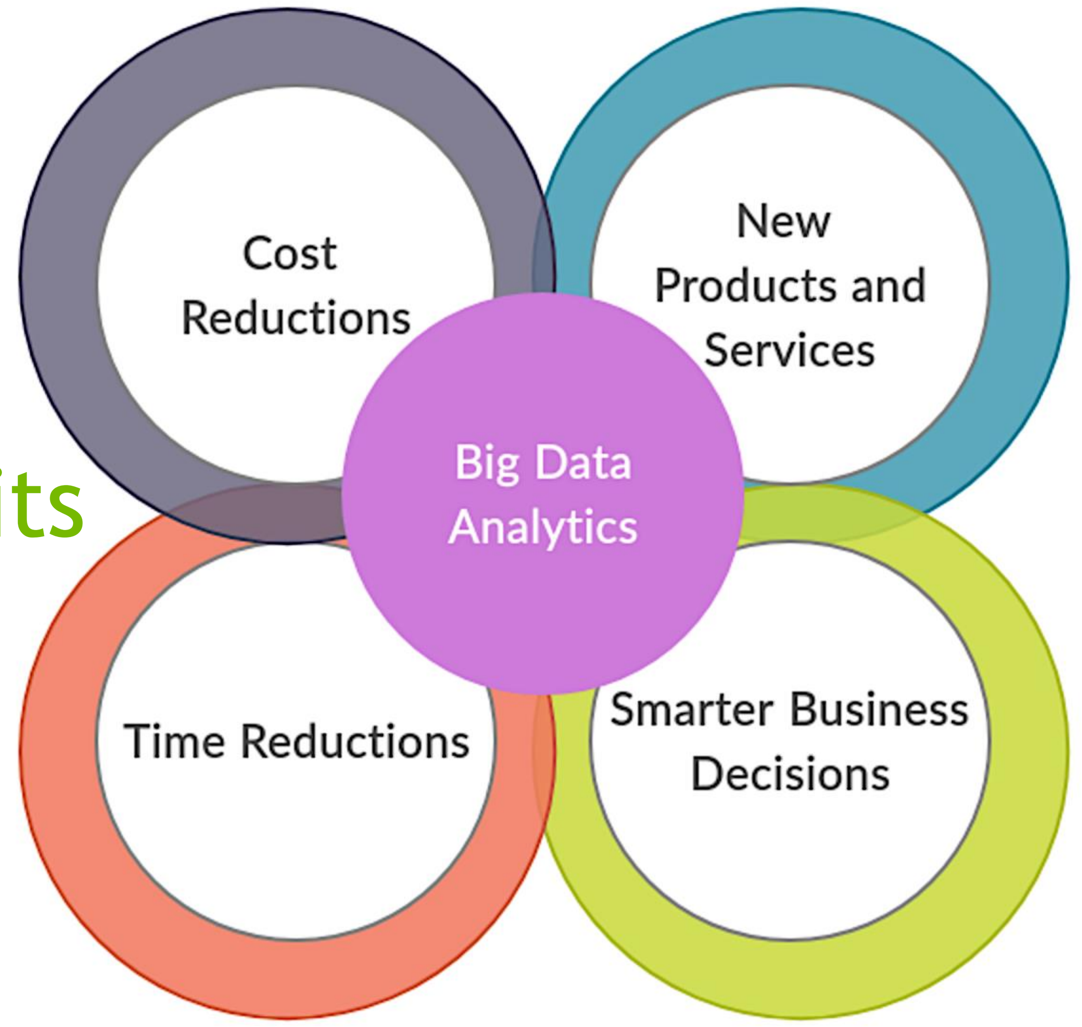
Graphics Processing Unit
100s or 1000s of Cores
Good for parallel processing
High Throughput

GPU COMPUTING IN ENTERPRISES

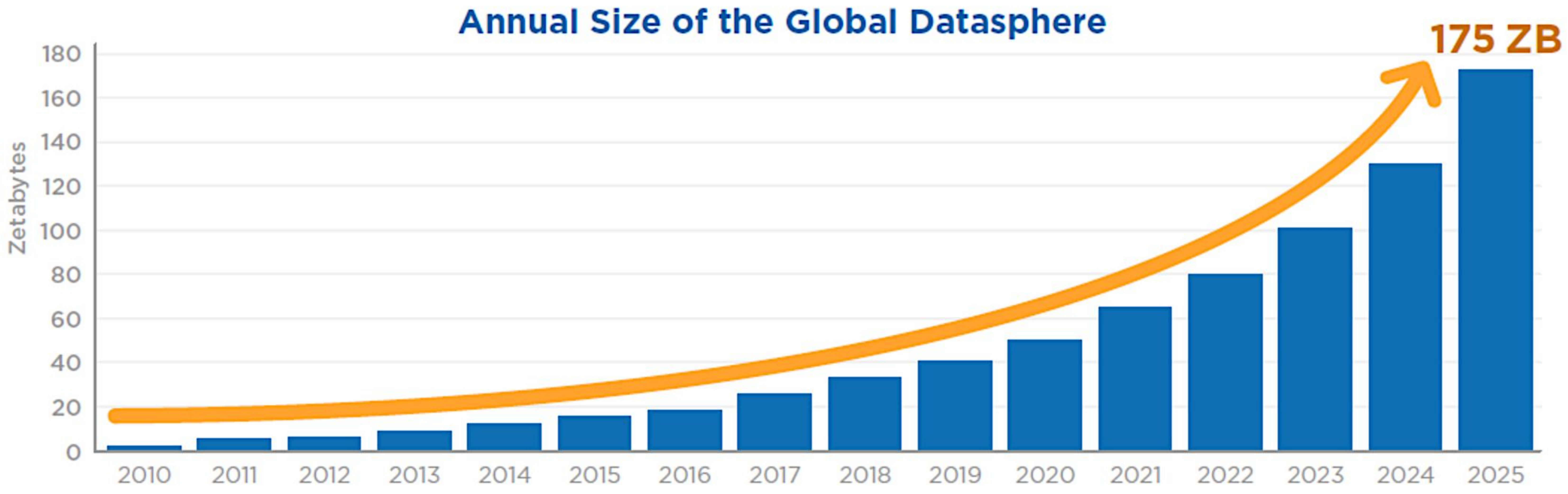
WHY ENTERPRISES NEED GPUS

Valuable Hidden Insights

Indispensable Business Benefits

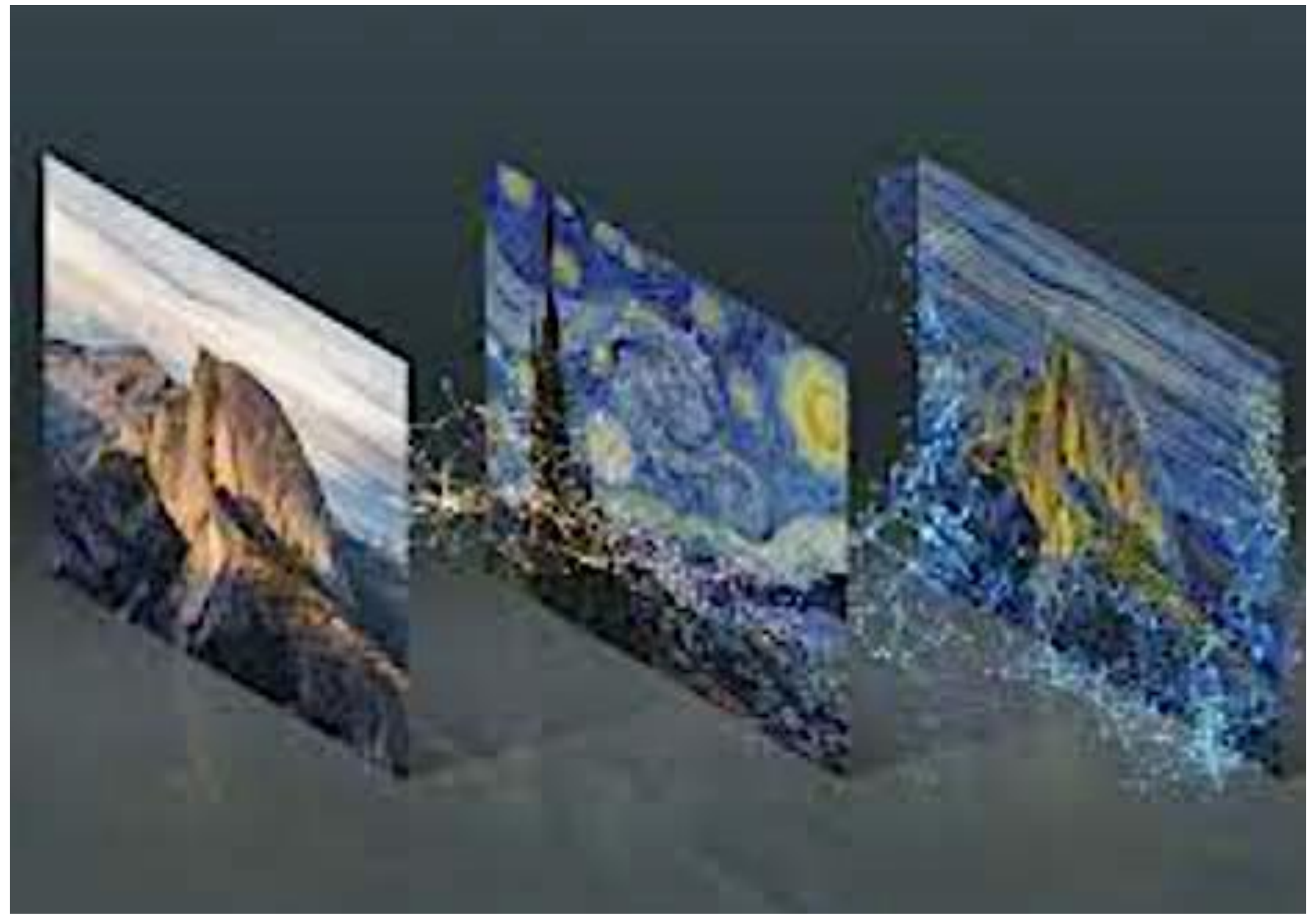


Source: IDC DataAge 2025 Whitepaper

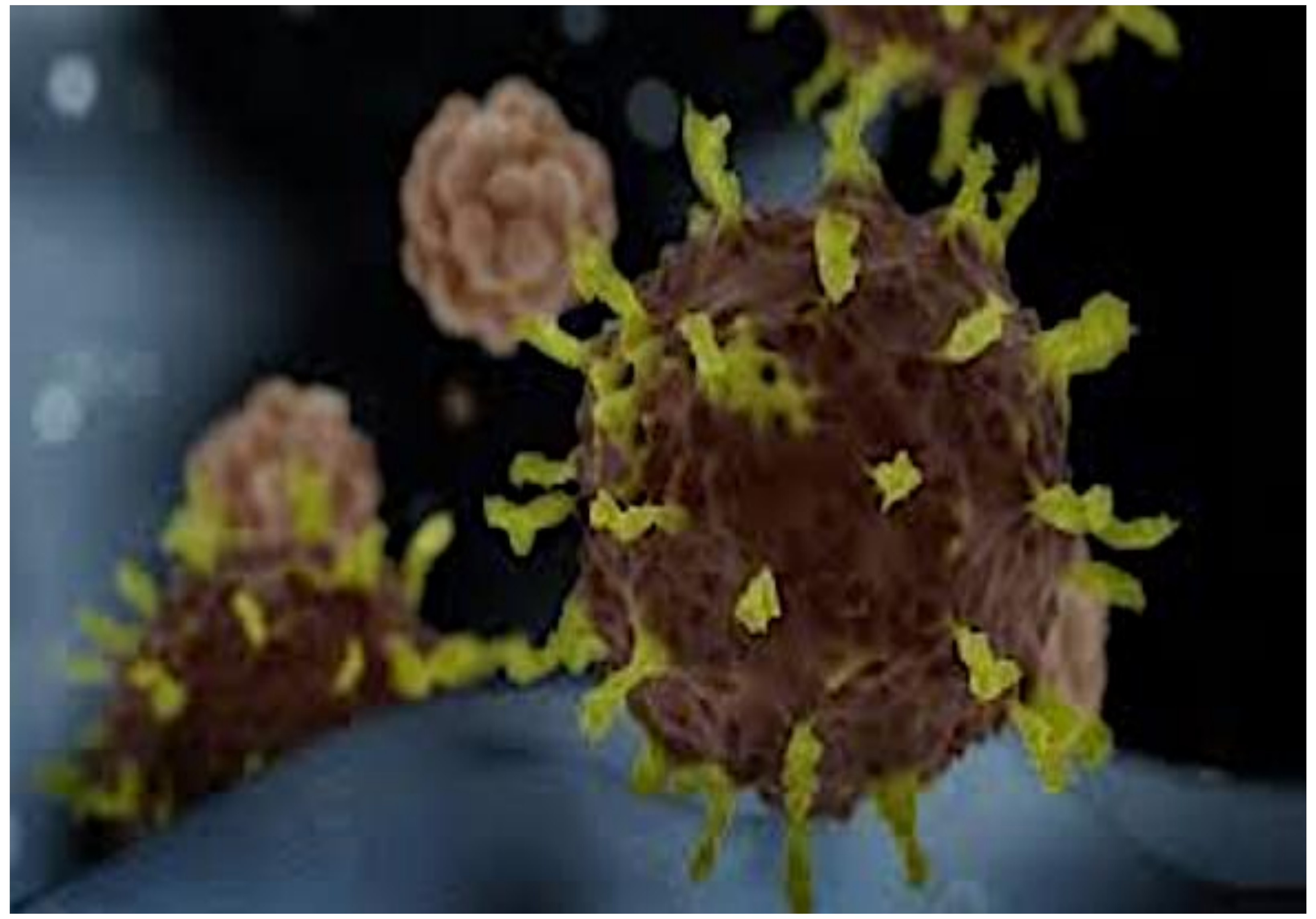


GPU-ACCELERATED ENTERPRISE WORKLOADS & USE CASES

Graphics
- Movie Animation
- Gaming



AI & Data Analytics
- Natural Language Processing
- Recommender Systems



High Performance Computing
- Scientific Computing
- Industrial HPC

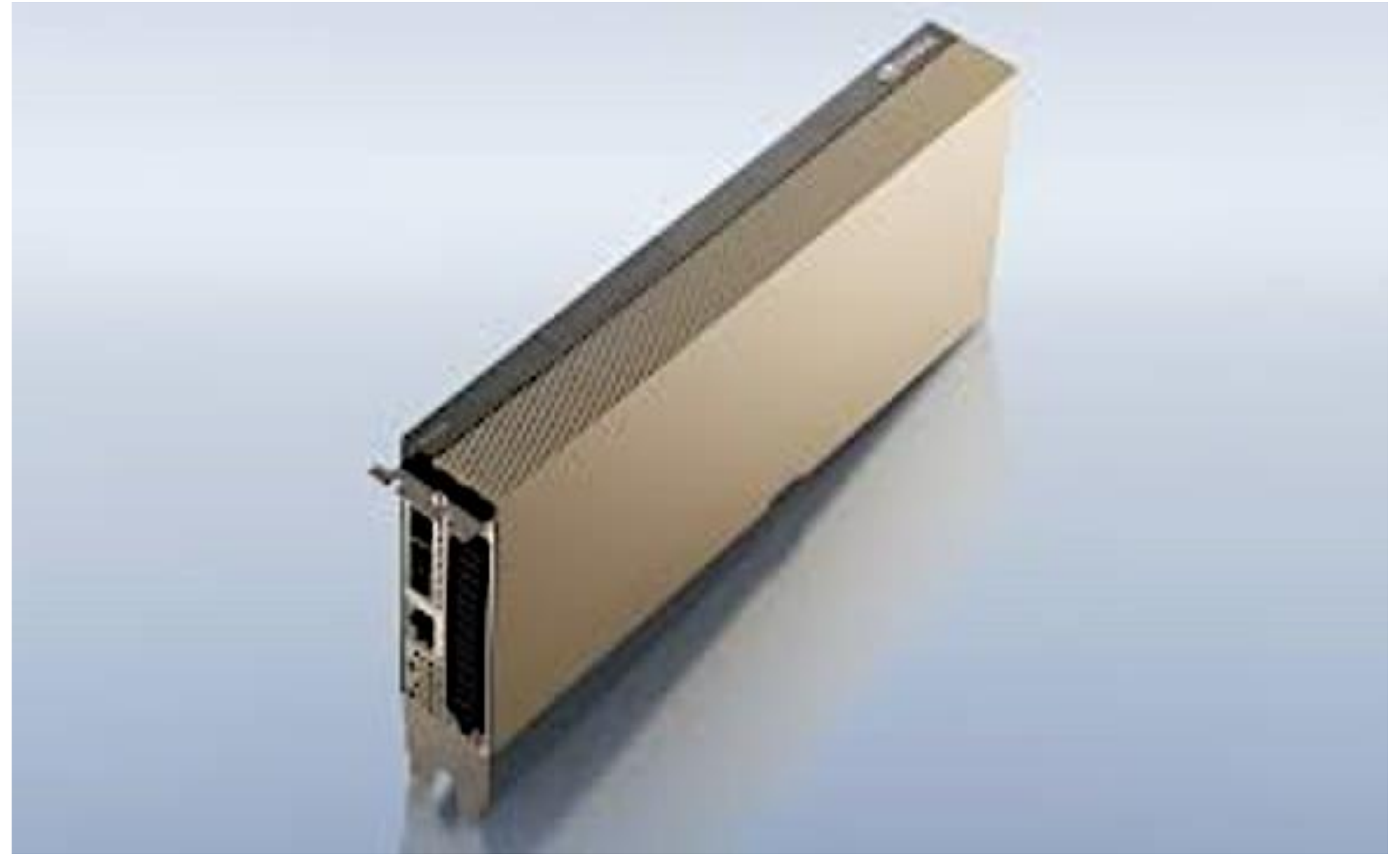
Visualization

Compute

A GPU FOR EVERY WORKLOAD

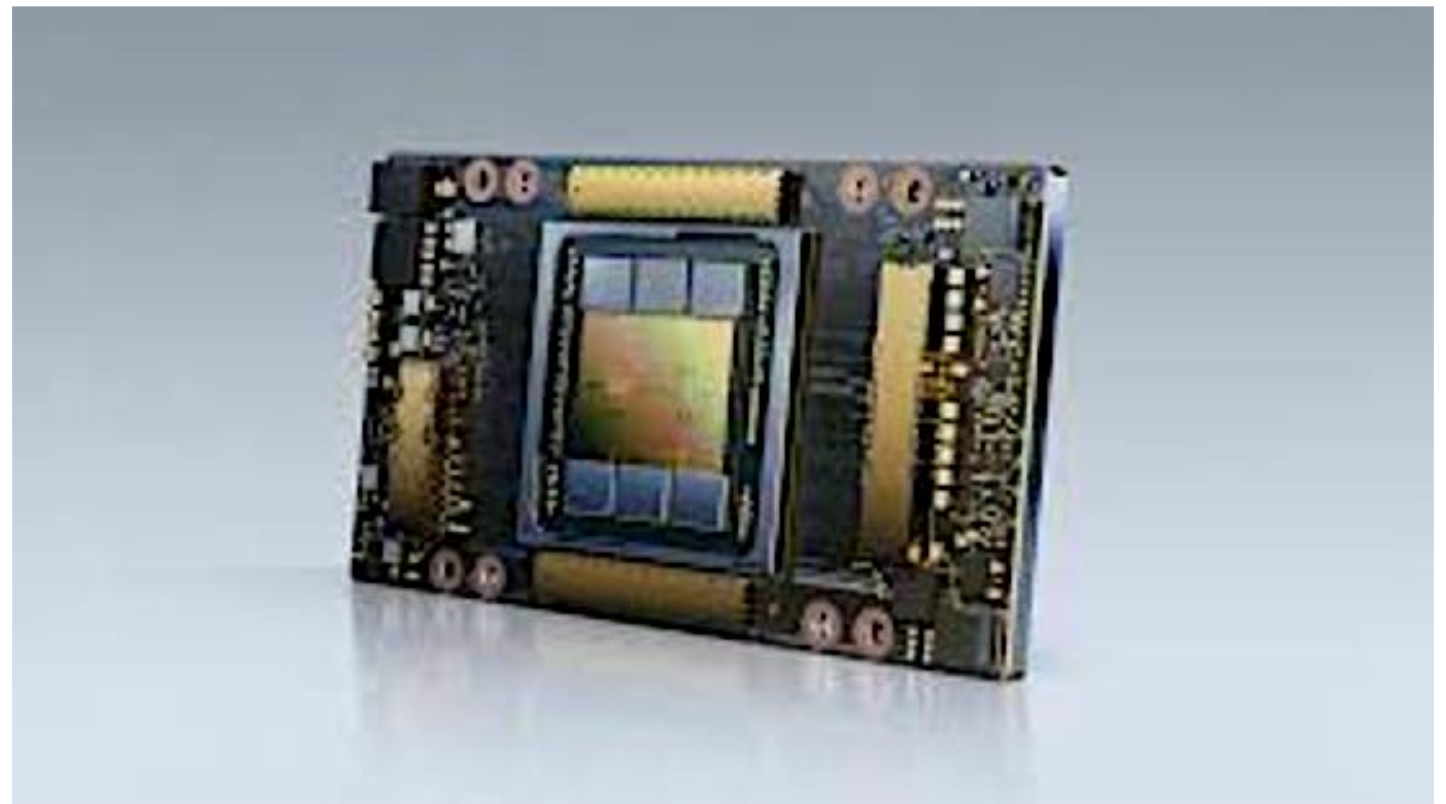
Visualization GPUs

NVIDIA A40
Highest Perf Graphics
Visual Computing

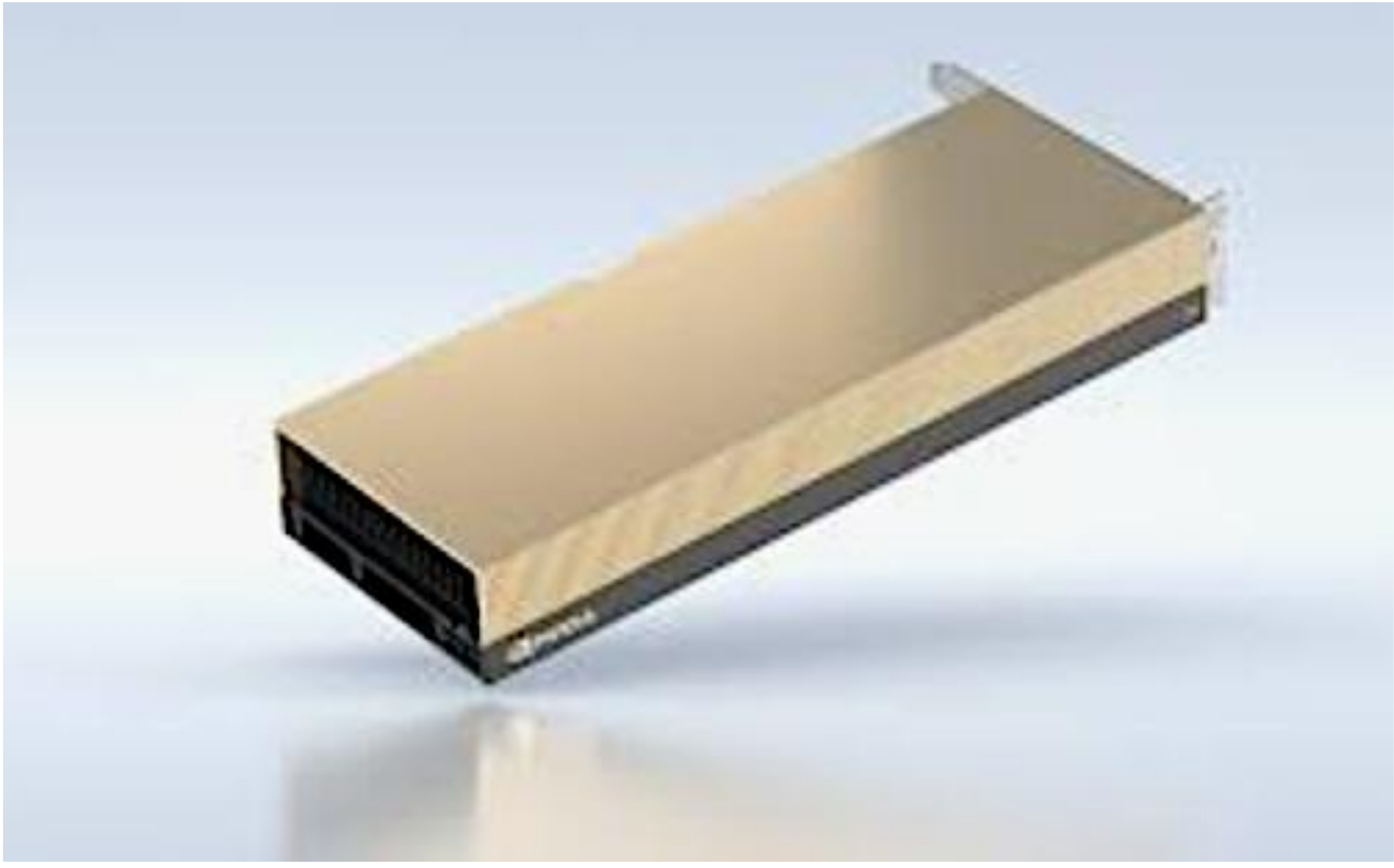


NVIDIA A16
High Density,
Best Experience VDI

Compute GPUs



NVIDIA A100
Highest Perf Compute
AI, HPC & Data Analytics

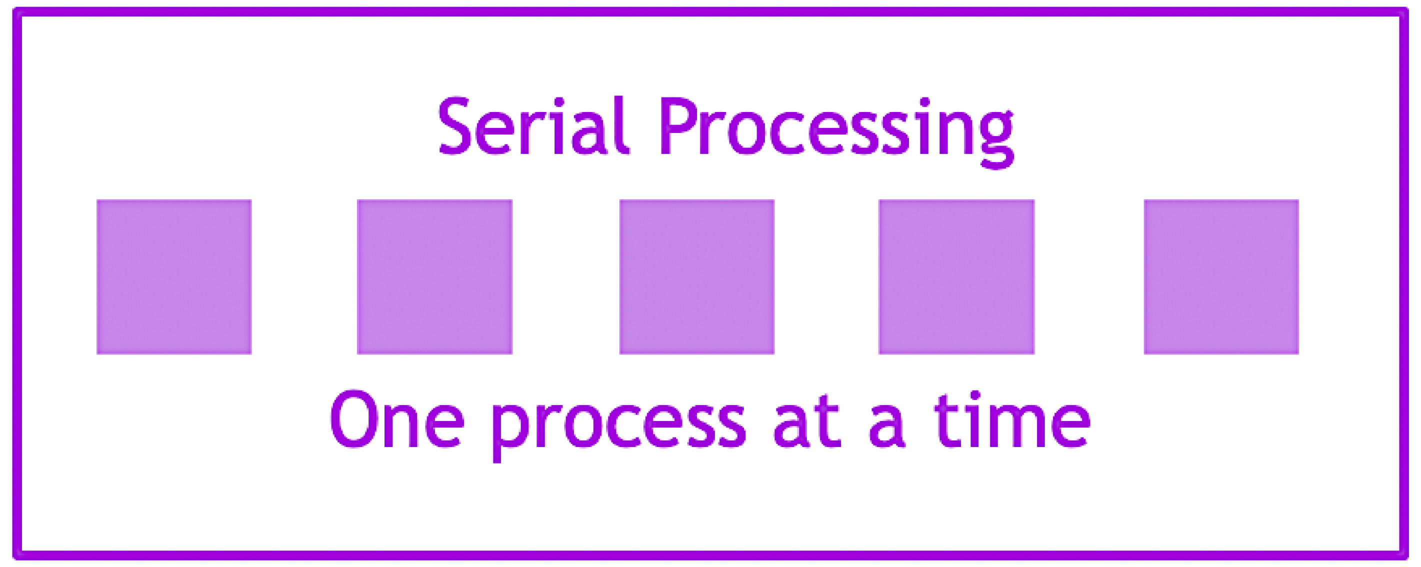


NVIDIA A30
AI Inference
Mainstream Compute

KEY TAKEAWAYS

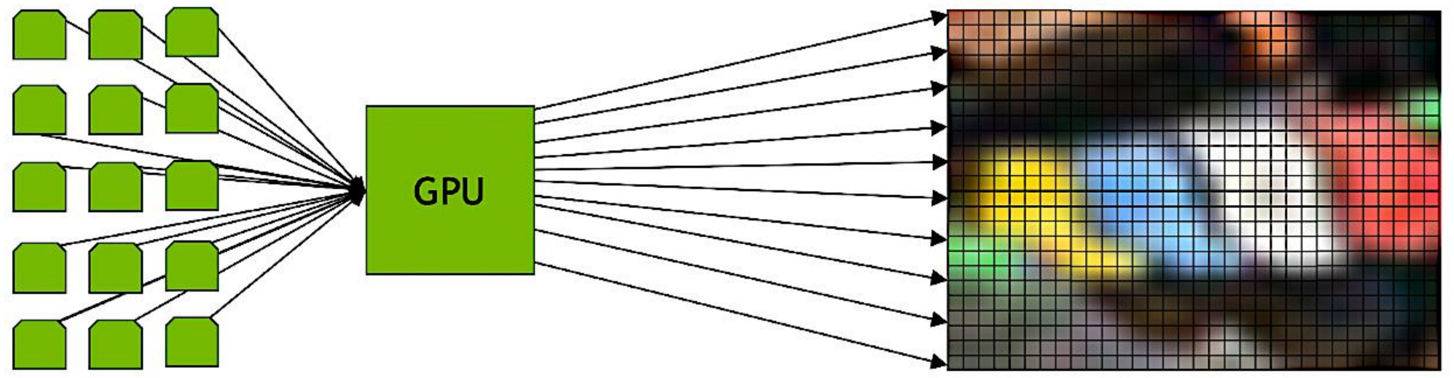
CPUs & Serial Processing

- A computer does its primary work in the Central Processing Unit
- A CPU is awesome at serial processing but is less effective at multicore processing
- CPU-based multicore processing causes bottlenecks and affects performance



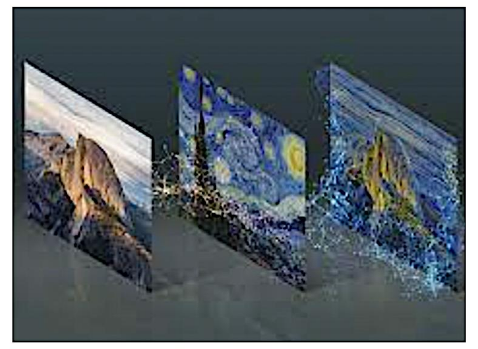
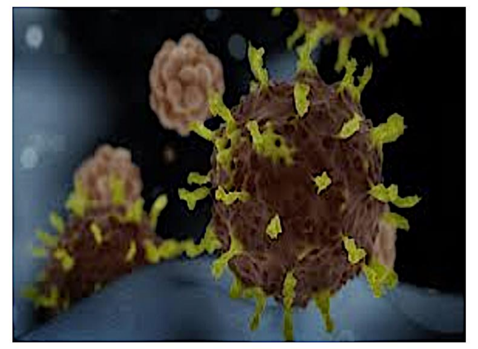
GPUs & Parallel Processing

- GPUs are great at performing a few simple mathematical calculations
- A GPU has up to 1000s of light weight cores that process data in parallel
- GPUs process large amounts of data at tremendous speed



GPU Computing In Enterprises

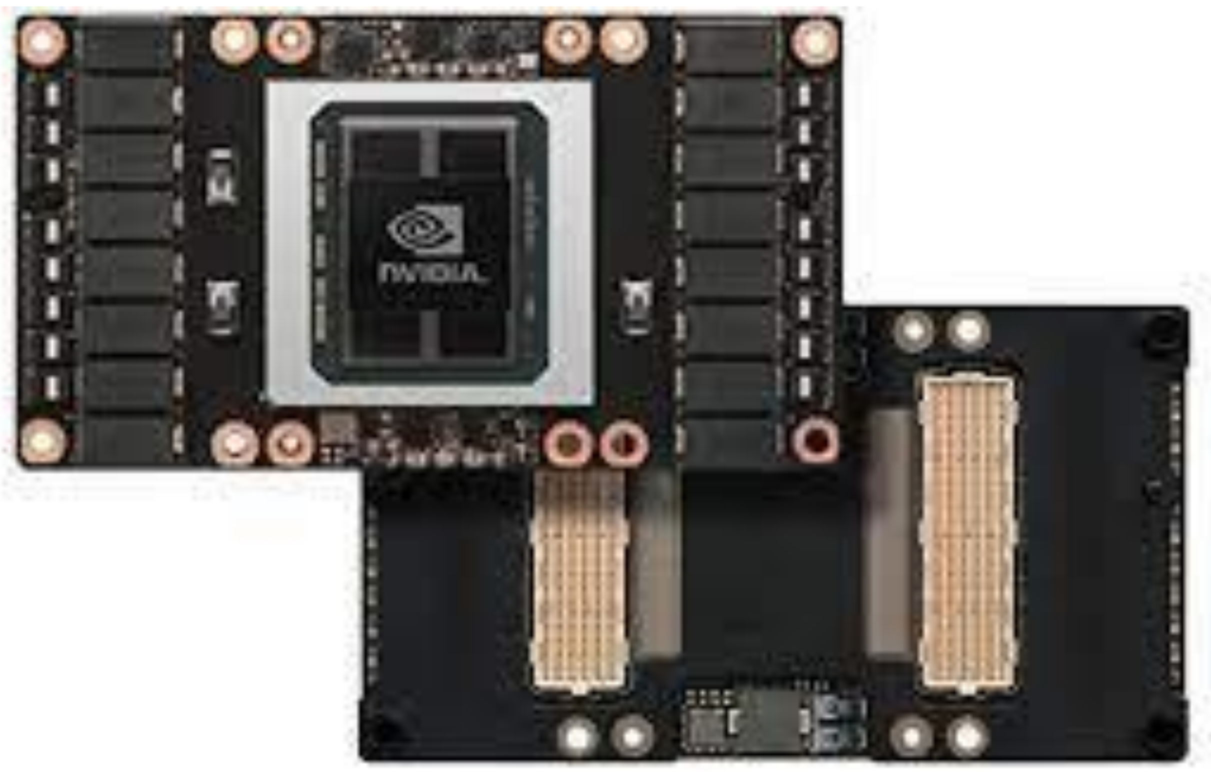
- Modern enterprise computing focuses on extracting insights from Big Data
- GPUs are indispensable for accelerating AI and Data Analytics, HPC and Graphics
- NVIDIA's GPUs specialize in accelerating all enterprise computing workloads for a variety of use cases

NVIDIA GPU ARCHITECTURE EVOLUTION

PASCAL ARCHITECTURE (2016)

A schematic of one P100 SM



P100 GPU

Features

- Hardware support for float16 calculations for high performance
- P100 GPU contains 56 SMs (streaming multiprocessor)
- Each SM consist of 2 processing blocks
- One P100 SM is composed of two identical processing blocks
- The green blocks represent CUDA cores
- Yellow blocks represent CUDA cores dedicated for double precision calculations
- SFUs are blocks for special function units which compute functions like sine, cosine etc
- Performance of P100:
  - Single Precision - **10.6 teraFLOPS**
  - Half Precision - **21.2 teraFLOPS**

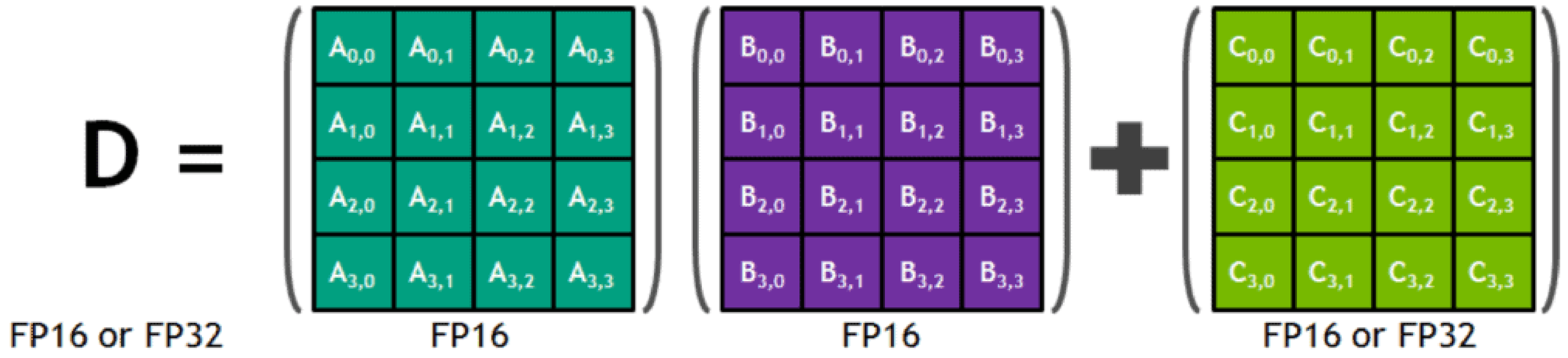
VOLTA ARCHITECTURE (2018)

A schematic of one V100 SM

Features

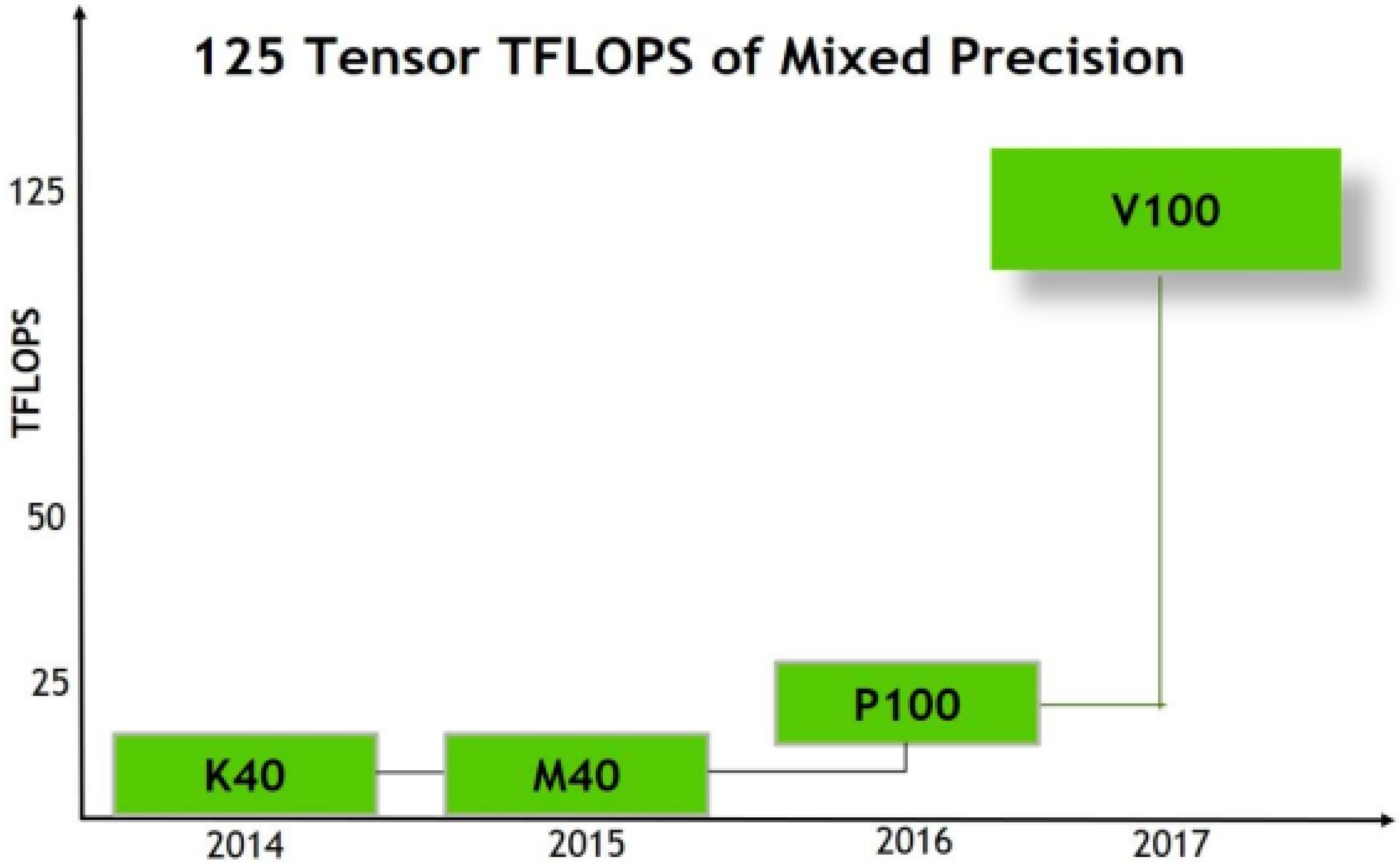
- V100 consists of 80 SMs
- Each SM consist of four processing blocks
- Introduced Tensor Cores
- A tensor core is a special type of CUDA core designed for Multiply Accumulate (MAC) operations
  - D= A x B + C where A~D are all 4×4 matrices
- MAC calculations are used in most **deep learning layers**
- Performance of V100:
  - Single Precision - **16.4 teraFLOPS**
  - Half Precision - **32.8 teraFLOPS**

V100 GPU



Tensor Core Operations

VOLTA ARCHITECTURE PERFORMANCE (2018)



AMPERE ARCHITECTURE (2020)

A schematic of one A100 SM

Features

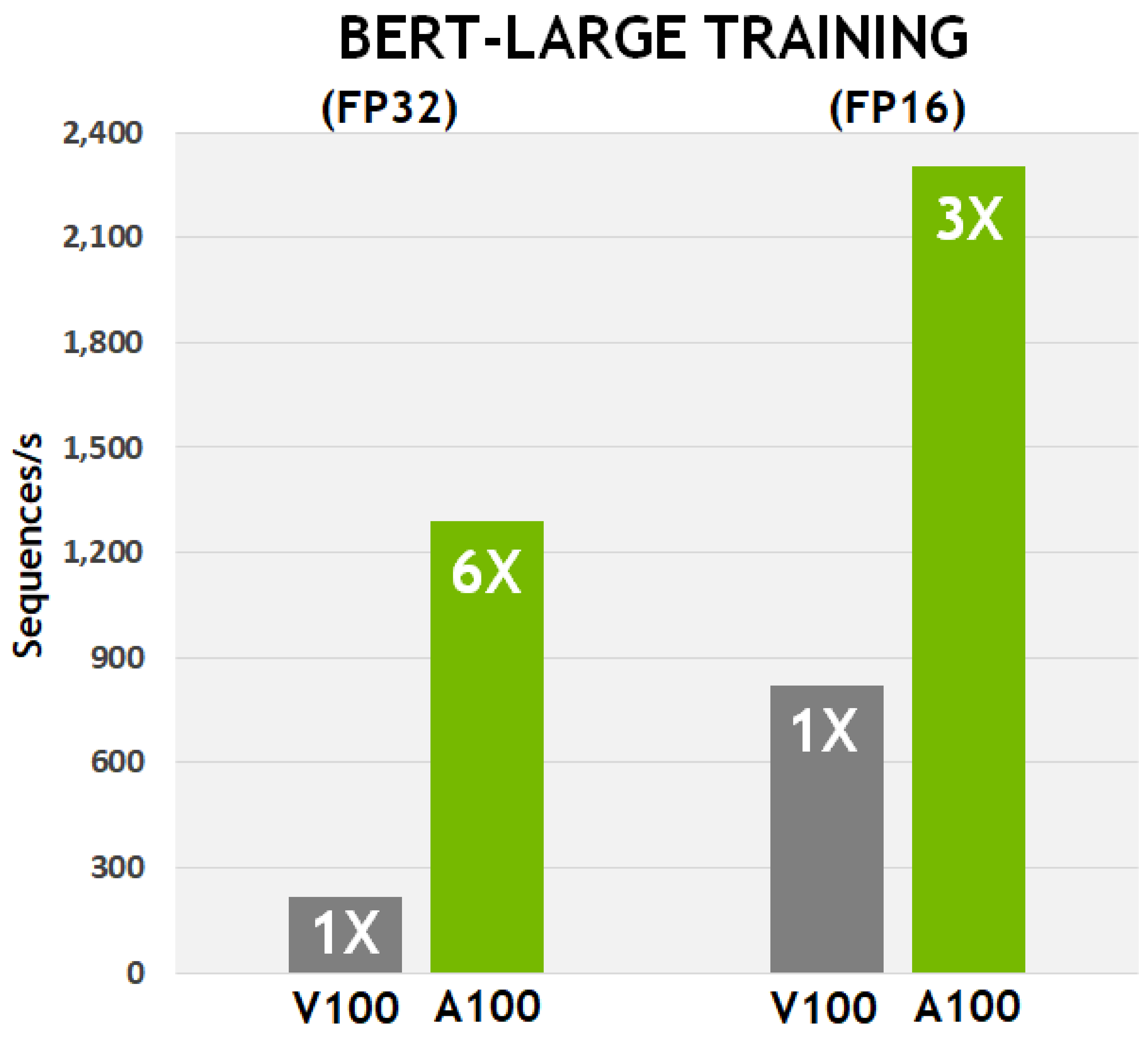
- A100 consists of 108 SMs
- Each SM consist of four processing blocks
- Introduced Multi Instance GPUs (MIGs)
- Introduced third generation Tensor Cores
  - Support all data types from binary, INT4, INT8, FP16, TF32 and even FP64.
  - Particularly useful for Deep Learning practitioners
- Introduced hardware support for **structured sparsity (SS)**
- Performance of A100:
  - Single Precision – **19.5 teraFLOPS**
  - Half Precision - **78 teraFLOPS**

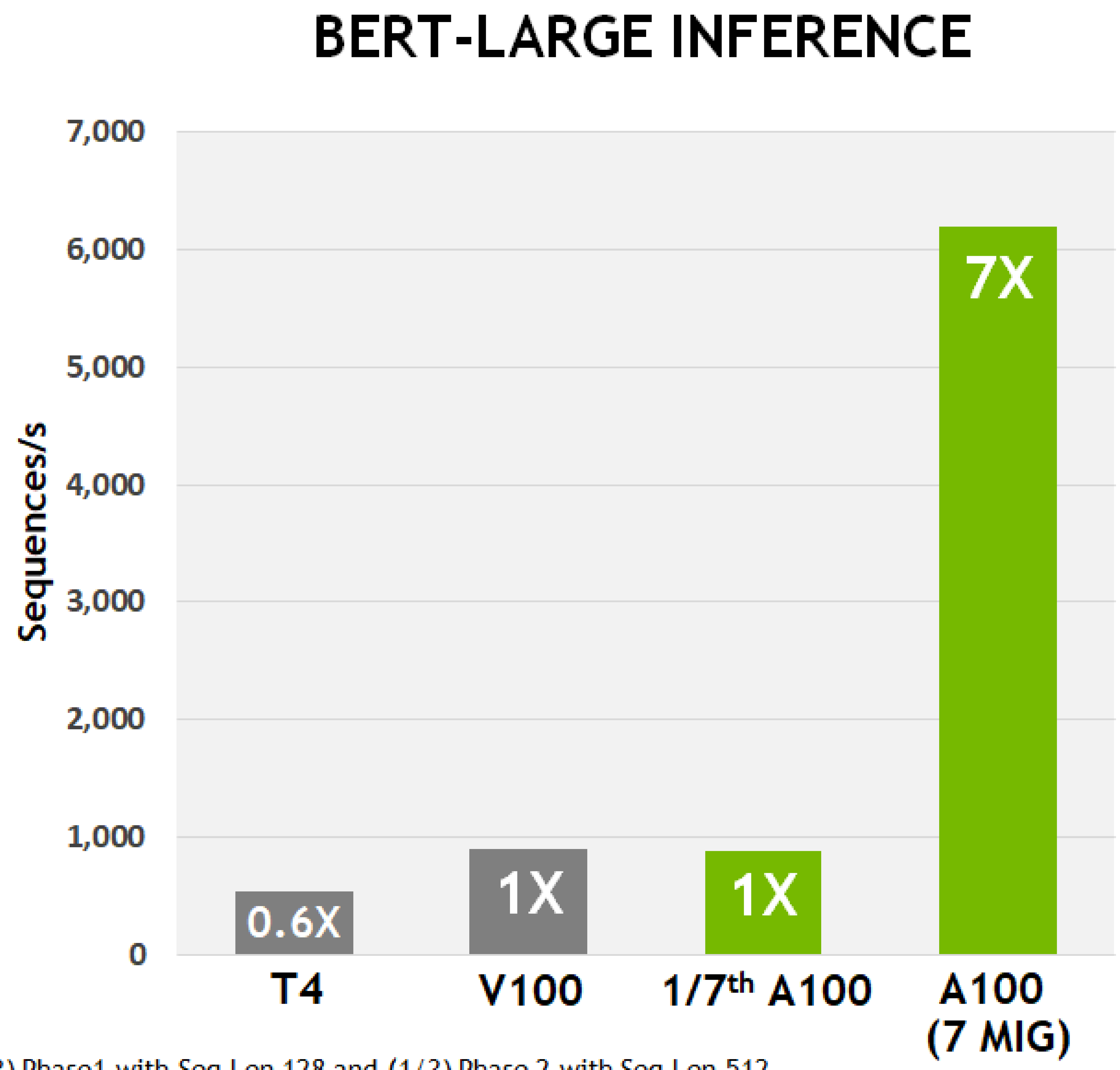
A100 GPU

DGX A100 with 8 A100 GPUs

AMPERE ARCHITECTURE PERFORMANCE (2020)

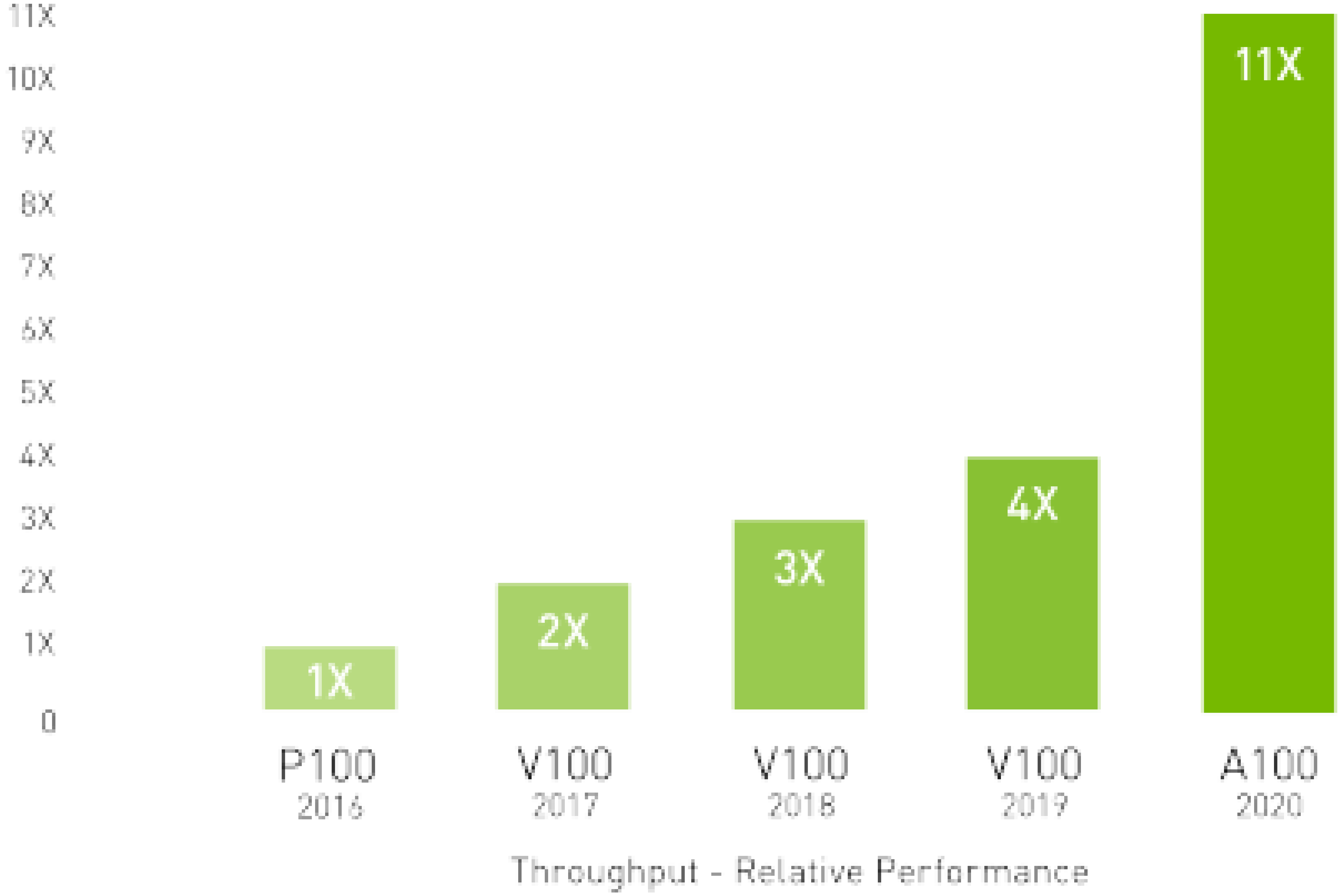
UNIFIED AI ACCELERATION





All results are measured
BERT Large Training (FP32 & FP16) measures Pre-Training phase, uses PyTorch including (2/3) Phase1 with Seq Len 128 and (1/3) Phase 2 with Seq Len 512,
V100 is DGX1 Server with 8xV100, A100 is DGX A100 Server with 8xA100, A100 uses TF32 Tensor Core for FP32 training
BERT Large Inference uses TRT 7.1 for T4/V100, with INT8/FP16 at batch size 256. Pre-production TRT for A100, uses batch size 94 and INT8 with sparsity

P100 VS V100 VS A100



Geometric mean of application speedups vs. P100: Benchmark application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT-Large Fine Tuner], Quantum Espresso [AUSURF112-jR]; Random Forest FP32 [make_blobs (160000 x 64 : 10)], TensorFlow [ResNet-50], VASP 6 [Si Huge] | GPU node with dual-socket CPUs with 4x NVIDIA P100, V100, or A100 GPUs.