Hardware and Software for AI
CPU and GPU

Tushar B. Kute,
http://tusharkute.com

AI Today

- Artificial intelligence (AI) is set to transform global productivity, working patterns, and lifestyles and create enormous wealth.
- Research firm Gartner expects the global AI economy to increase from about $1.2 trillion last year to about $3.9 Trillion by 2022, while McKinsey sees it delivering global economic activity of around $13 trillion by 2030.
- And of course, this transformation is fueled by the powerful Machine Learning (ML) tools and techniques such as Deep Reinforcement Learning (DRL), Generative Adversarial Networks (GAN), Gradient-boosted-tree models (GBM), Natural Language Processing (NLP), and more.

AI Today

- Most of the success in modern AI & ML systems is dependent on their ability to process massive amounts of raw data in a parallel fashion using task-optimized hardware.
- In fact, the modern resurgence of AI started with the 2012 ImageNet competition where deep-learning algorithms demonstrated an eye-popping increment in the image classification accuracy over their non-deep-learning counterparts (algorithms).
- However, along with clever programming and mathematical modeling, use of specialized hardware played a significant role in this early success.

AI Today

- As a shining example, advancements in computer vision (CV) continue to drive many modern AI & ML systems.
- CV is accelerating almost every domain in the industry enabling organizations to revolutionize the way machines and business systems work – manufacturing, autonomous driving, healthcare.
- Almost all CV systems have graduated from traditional rule-based programming paradigm to large-scale, data-driven ML paradigm.
- And, consequently, GPU-based hardware plays a critical role in ensuring high quality predictions and classification by helping crunching massive amounts of training data (often in the range of petabytes).

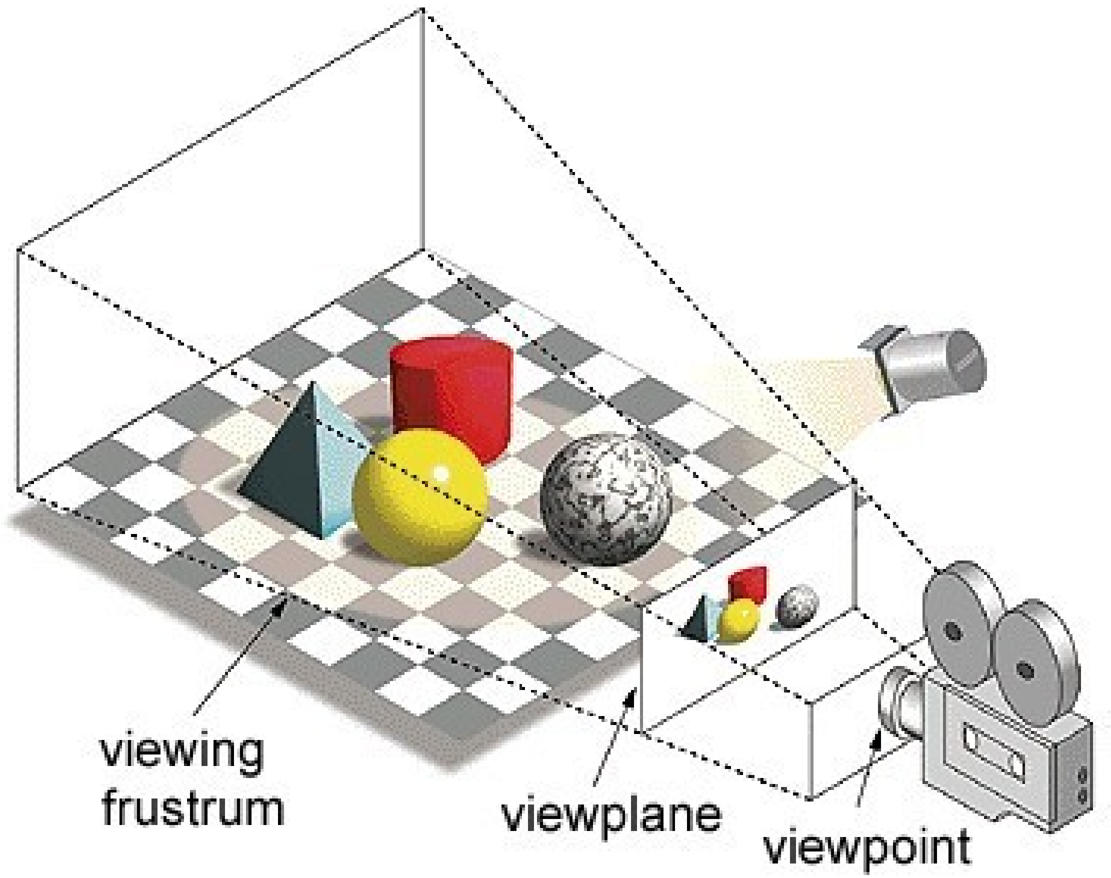
AI Today

- Some of the most talked about areas in AI & ML are:
  - Autonomous driving
  - Healthcare/ Medical imaging
  - Fighting disease, drug discovery
  - Environmental/Climate science

GPU

- A GPU or ‘Graphics Processing Unit’ is a mini version of an entire computer but only dedicated to a specific task.
- It is unlike a CPU that carries out multiple tasks at the same time. GPU comes with its own processor which is embedded onto its own motherboard coupled with v-ram or video ram, and also a proper thermal design for ventilation and cooling.

GPU



How they work?

- In the term ‘Graphics Processing Unit’, ‘Graphics’ refers to rendering an image at specified coordinates on a 2d or 3d space.
- A viewport or viewpoint is a viewer’s perspective of looking to an object depending upon the type of projection used.
- Rasterisation and Ray-tracing are some of the ways of rendering 3d scenes, both of these concepts are based on a type of a projection called as perspective projection. What is perspective projection?

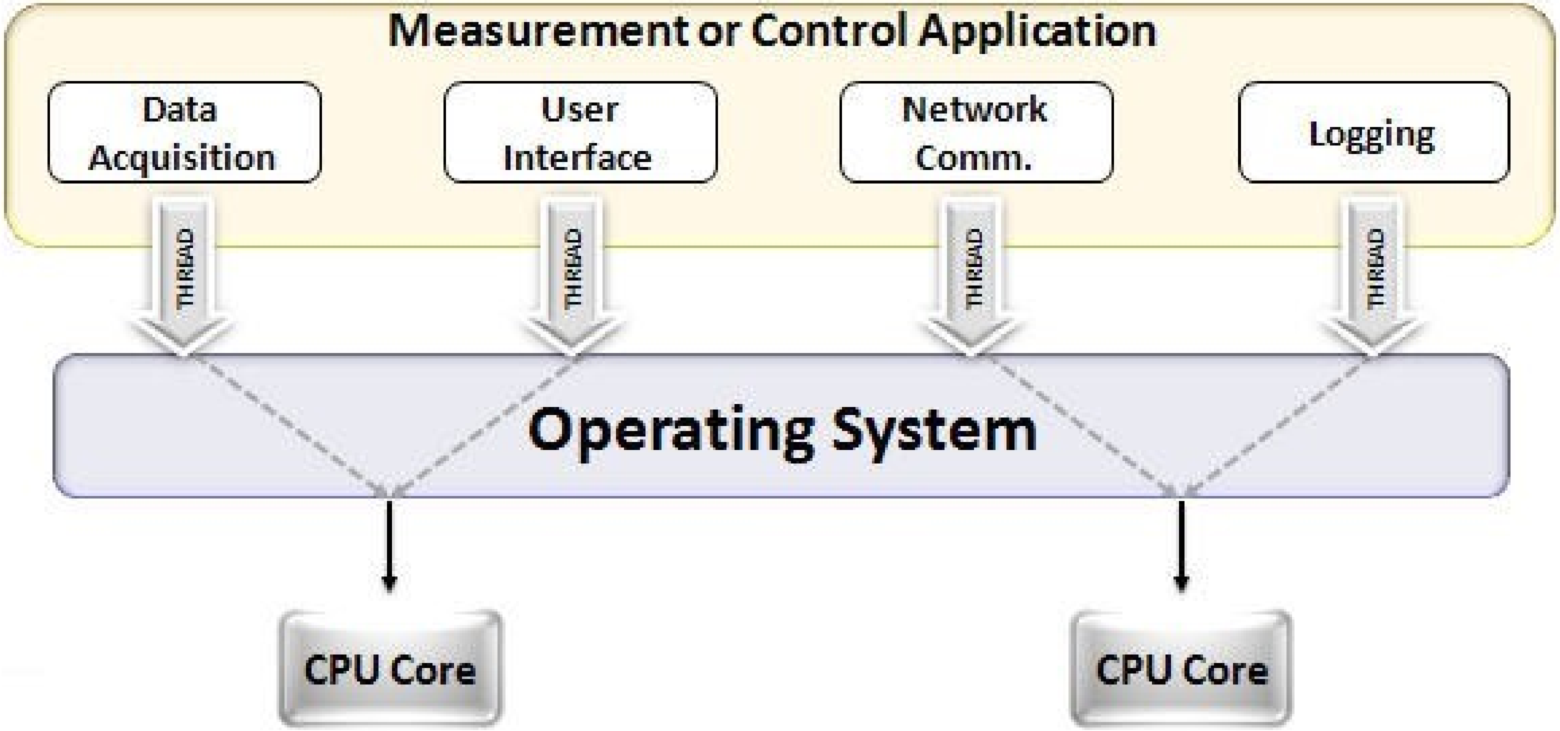
How they work?

- In short, it is the way in which how an image is formed on a view plane or canvas where the parallel lines converge to a converging point called as ‘center of projection’ also as the object moves away from the viewpoint it appears to be smaller, exactly how our eyes portray in real-world and this helps in understanding depth in an image as well, that is the reason why it produces realistic images.

Why GPU for Deep Learning?

- One of the most admired characteristics of a GPU is the ability to compute processes in parallel. This is the point where the concept of parallel computing kicks in.
- A CPU in general completes its task in a sequential manner. A CPU can be divided into cores and each core takes up one task at a time. Suppose if a CPU has 2 cores.
- Then two different task's processes can run on these two cores thereby achieving multitasking.

Why GPU for Deep Learning?



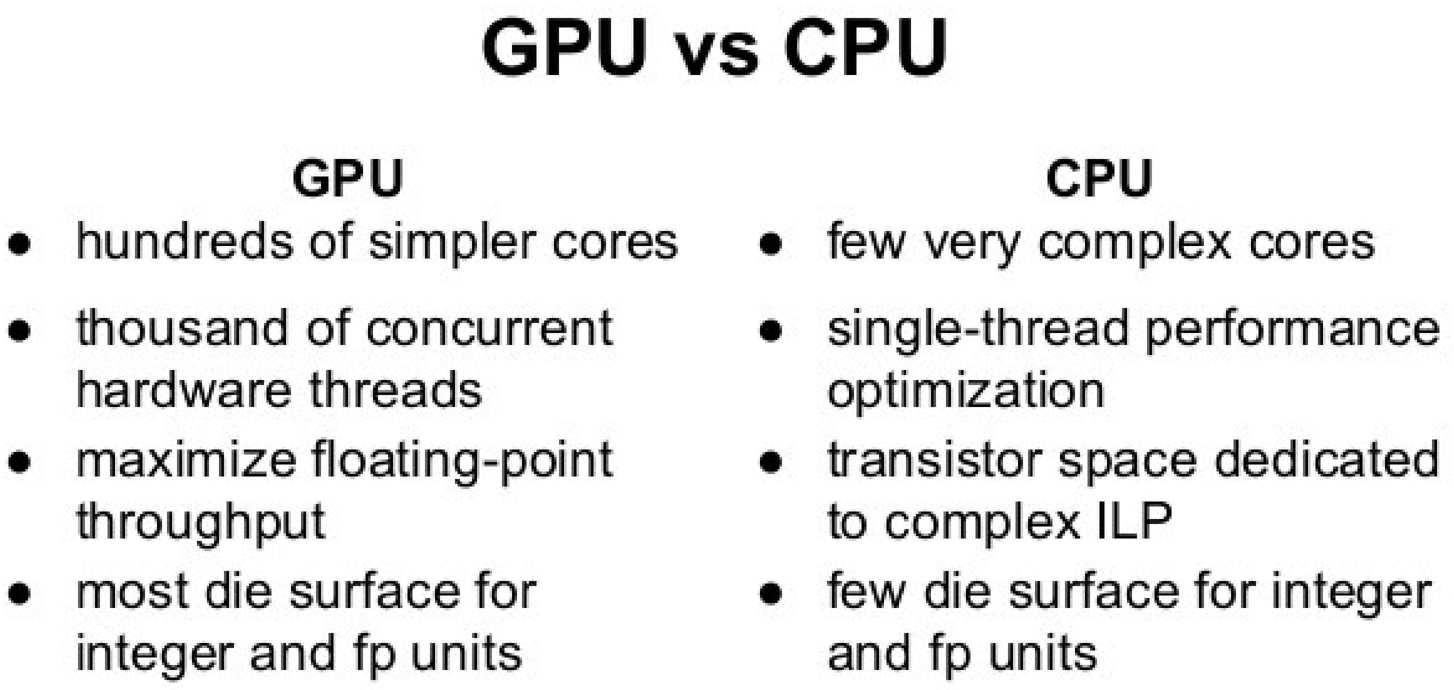
Why GPU?

- General-purpose CPUs struggle when operating on a large amount of data e.g., performing linear algebra operations on matrices with tens or hundreds thousand floating-point numbers.
- Under the hood, deep neural networks are mostly composed of operations like matrix multiplications and vector additions.

Why GPU?

- GPUs were developed (primarily catering to the video gaming industry) to handle a massive degree of parallel computations using thousands of tiny computing cores.
- They also feature large memory bandwidth to deal with the rapid dataflow (processing unit to cache to the slower main memory and back), needed for these computations when the neural network is training through hundreds of epochs.
- This makes them the ideal commodity hardware to deal with the computation load of computer vision tasks.

CPU vs. GPU



GPU vs CPU

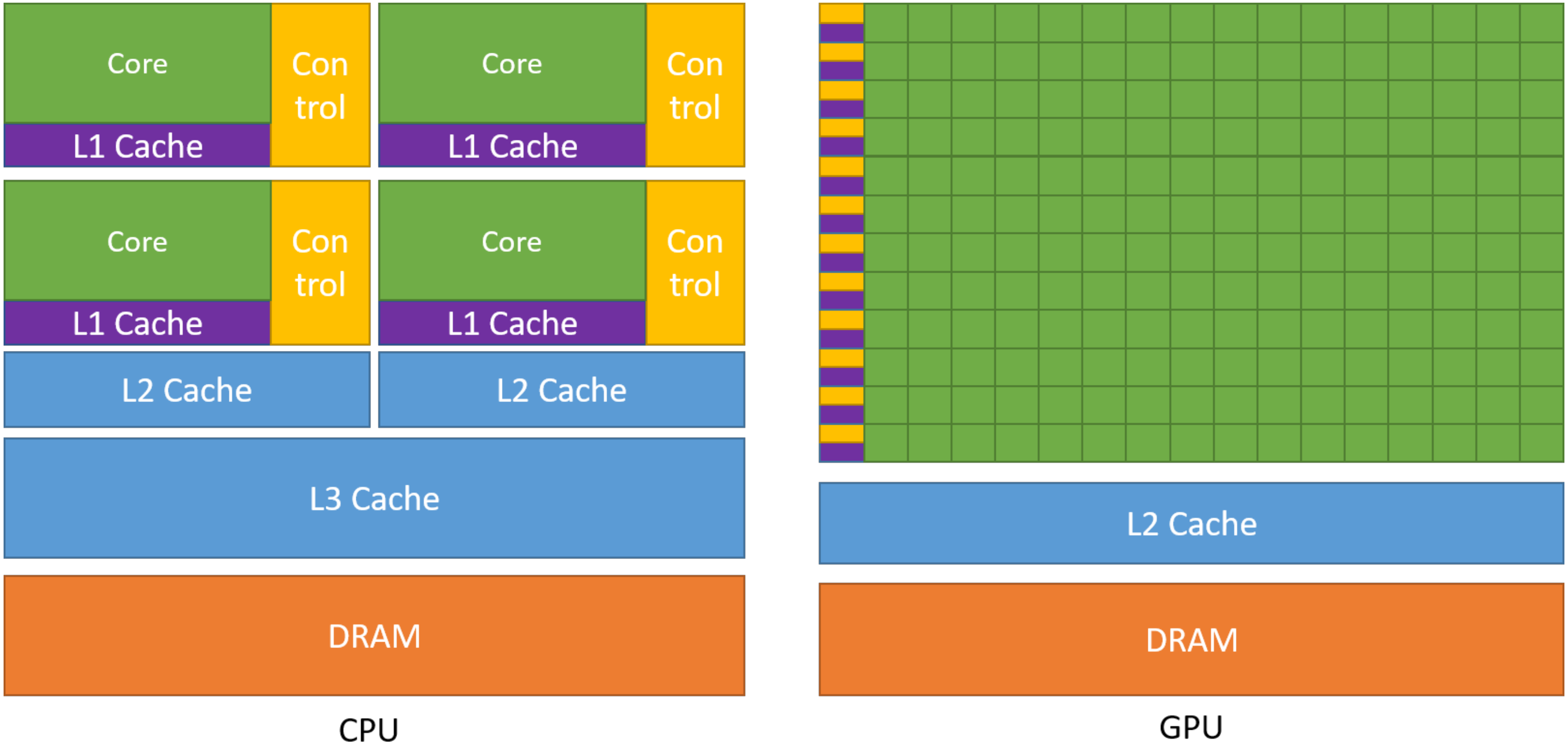
GPU

- hundreds of simpler cores
- thousand of concurrent hardware threads
- maximize floating-point throughput
- most die surface for integer and fp units

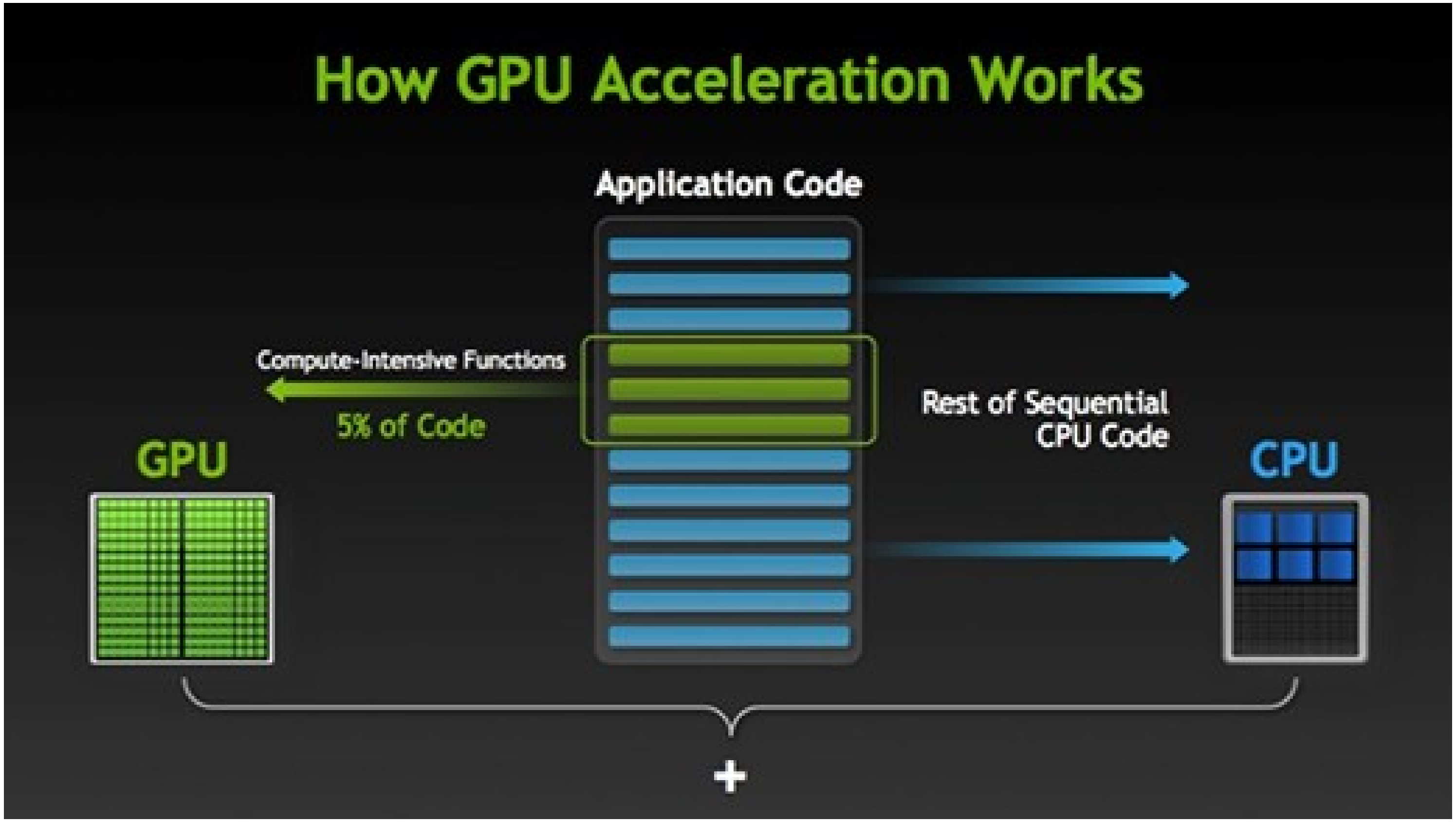
CPU

- few very complex cores
- single-thread performance optimization
- transistor space dedicated to complex ILP
- few die surface for integer and fp units

CPU vs. GPU


Core
Con trol
L1 Cache
Core
Con trol
L1 Cache
Core
Con trol
L1 Cache
Core
Con trol
L1 Cache
L2 Cache
L2 Cache
L3 Cache
DRAM
CPU
L2 Cache
DRAM
GPU

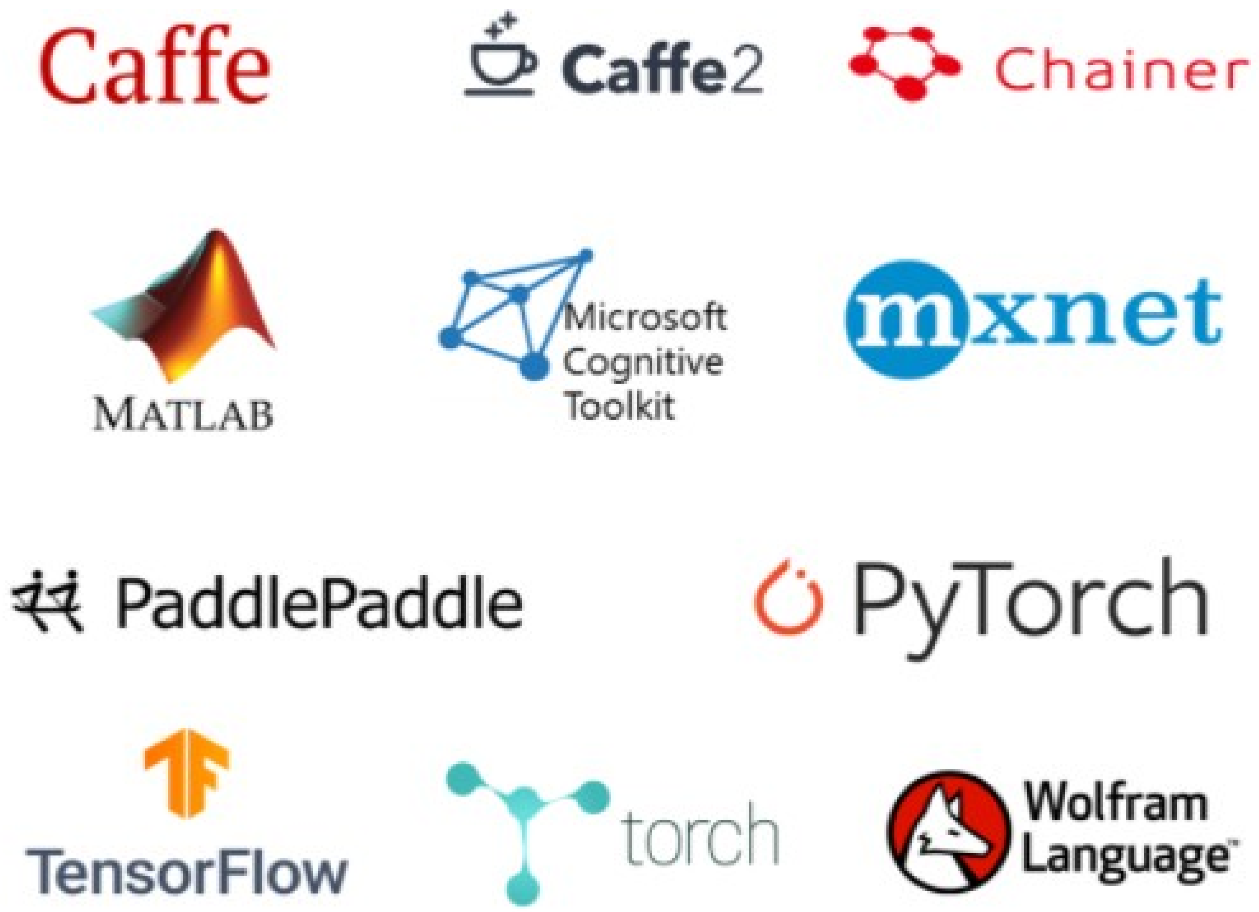
CPU vs. GPU


How GPU Acceleration Works
Application Code
Compute-Intensive Functions
5% of Code
GPU
Rest of Sequential CPU Code
CPU
+

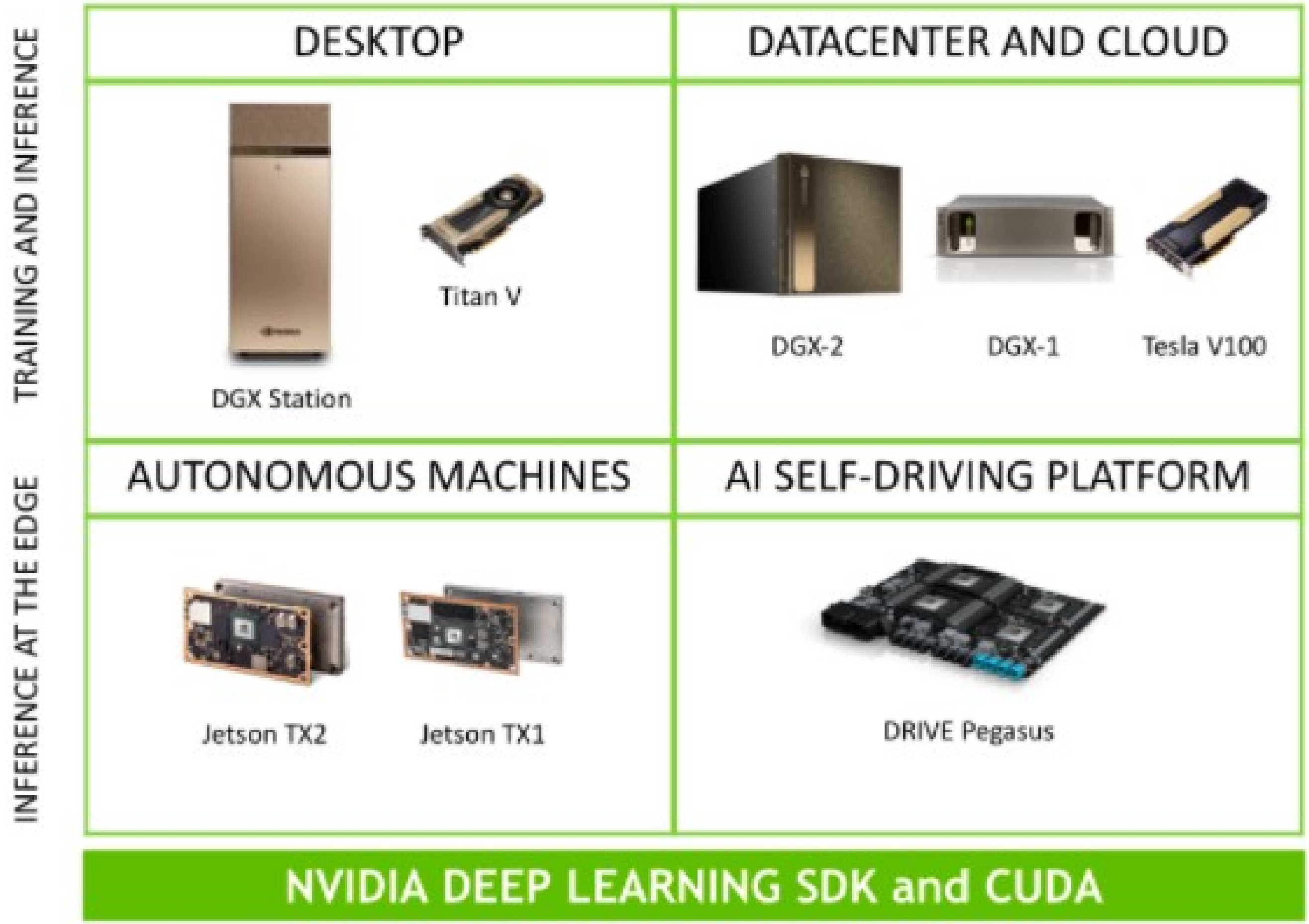
GPU

- The general architecture of GPUs was suitable for the particular type of computing tasks that are at the heart of deep learning algorithms.
- However, once this synergy was fully exploited by academic researchers and demonstrated beyond any doubt, corporations which produce GPU, such as Nvidia, invested a tremendous amount of R&D and human capital to develop more high-performing and highly optimized GPUs for a variety of applications.

Frameworks using GPUs



Frameworks using GPUs



TRAINING AND INFERENCE

INFERENCE AT THE EDGE

DESKTOP

DGX Station

Titan V

DATACENTER AND CLOUD

DGX-2

DGX-1

Tesla V100

AUTONOMOUS MACHINES

Jetson TX2

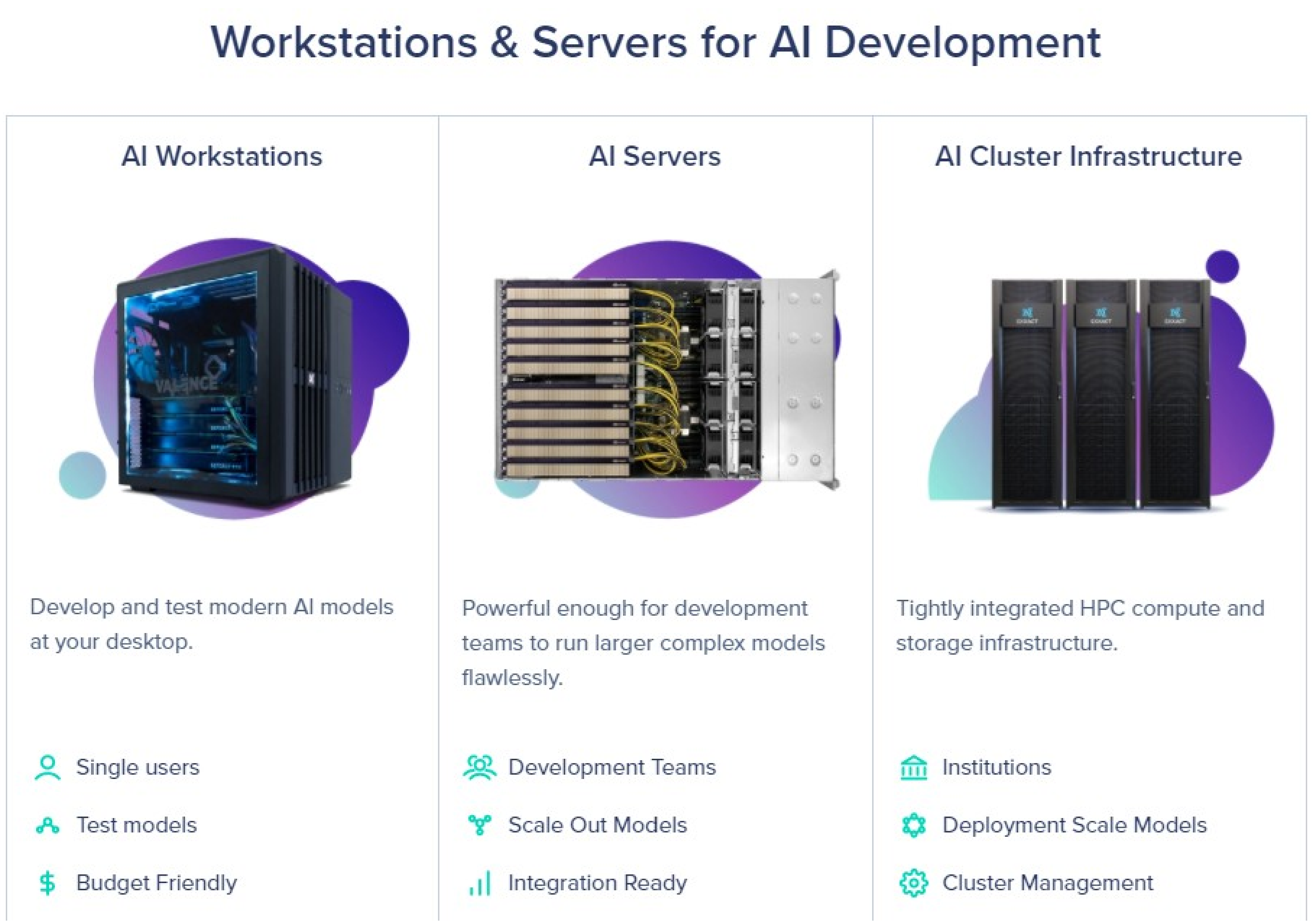
Jetson TX1

AI SELF-DRIVING PLATFORM

DRIVE Pegasus

NVIDIA DEEP LEARNING SDK and CUDA

More Hardware



Workstations & Servers for AI Development

AI Workstations

Develop and test modern AI models at your desktop.

- Single users
- Test models
- Budget Friendly

AI Servers

Powerful enough for development teams to run larger complex models flawlessly.

- Development Teams
- Scale Out Models
- Integration Ready

AI Cluster Infrastructure

Tightly integrated HPC compute and storage infrastructure.

- Institutions
- Deployment Scale Models
- Cluster Management

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
http://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com