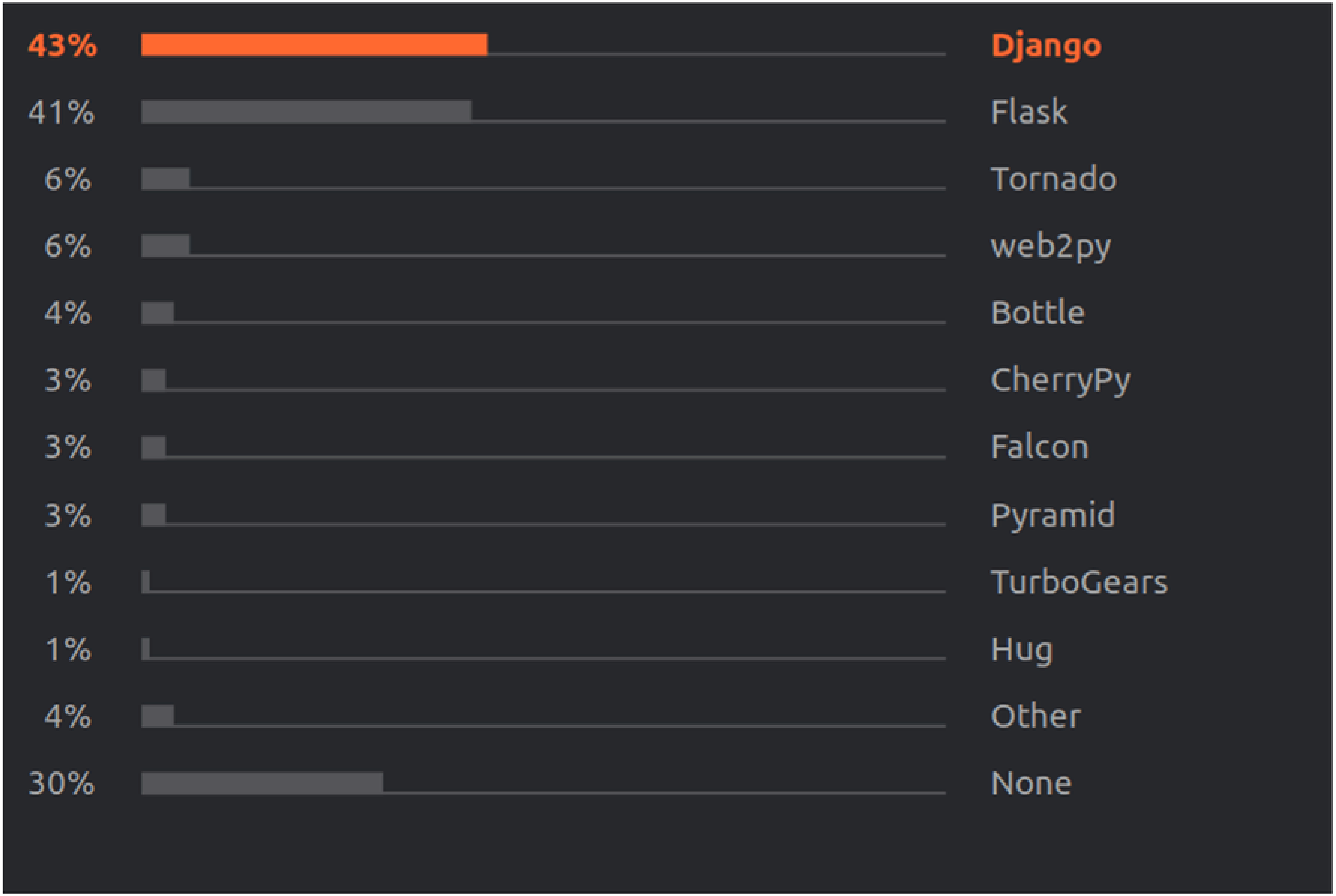
Deep Learning Model Deployment and Cloud Computing

DAY 10

Web frameworks

- A web framework (WF) or web application framework (WAF) is a software framework that is designed to support the development of web applications including web services, web resources, and web APIs.
- Web frameworks provide a standard way to build and deploy web applications on the World Wide Web. Web frameworks aim to automate the overhead associated with common activities performed in web development.
- For example, many web frameworks provide libraries for database access, templating frameworks, and session management, and they often promote code reuse.[1]
- Although they often target development of dynamic web sites, they are also applicable to static websites.


43%
Django
41%
Flask
6%
Tornado
6%
web2py
4%
Bottle
3%
CherryPy
3%
Falcon
3%
Pyramid
1%
TurboGears
1%
Hug
4%
Other
30%
None

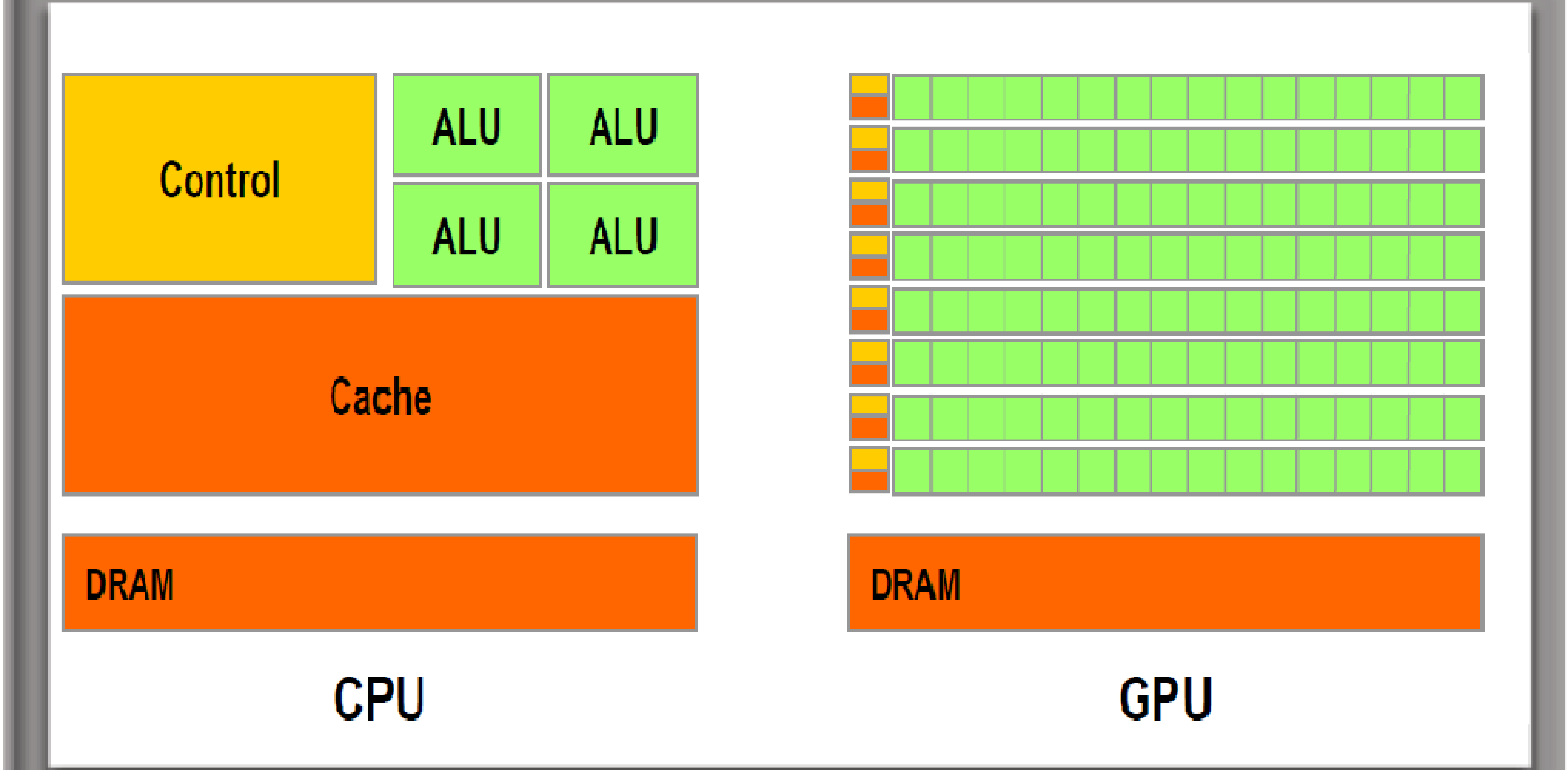
Flask

- Flask is a web application framework written in Python. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

Reference: https://www.tutorialspoint.com/flask/index.htm

Graphical Processing Units (GPUs)

- A CPU (central processing unit) works together with a GPU (graphics processing unit) to increase the throughput of data and the number of concurrent calculations within an application.
- GPUs were originally designed to create images for computer graphics and video game consoles, but since the early 2010's, GPUs can also be used to accelerate calculations involving massive amounts of data.
- A CPU can never be fully replaced by a GPU: a GPU complements CPU architecture by allowing repetitive calculations within an application to be run in parallel while the main program continues to run on the CPU.
- The CPU can be thought of as the taskmaster of the entire system, coordinating a wide range of general-purpose computing tasks, with the GPU performing a narrower range of more specialized tasks (usually mathematical). Using the power of parallelism, a GPU can complete more work in the same amount of time as compared to a CPU.


Control
ALU
ALU
ALU
ALU
Cache
DRAM
DRAM
CPU
GPU

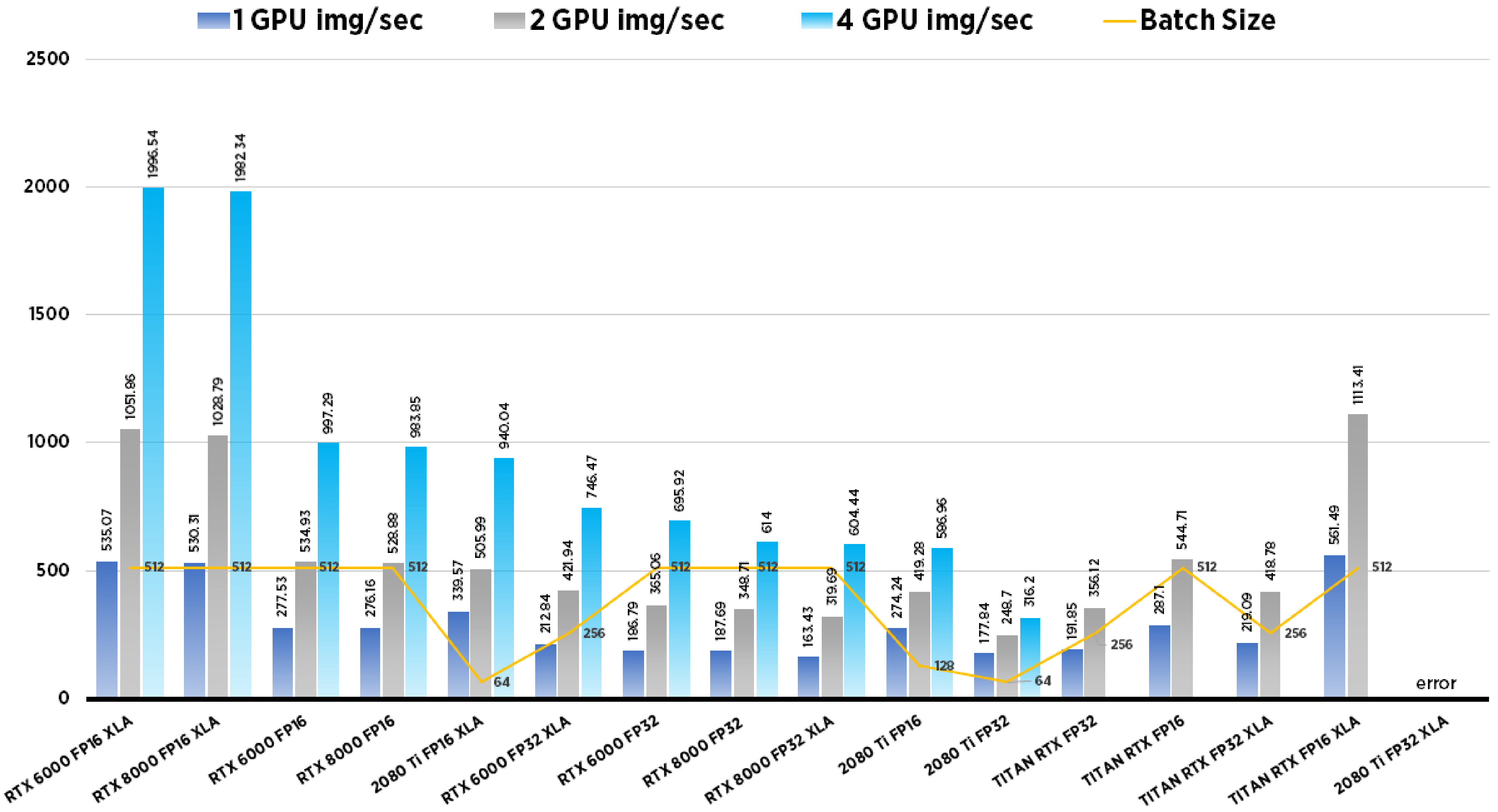
GPUs

- Traditionally, the training phase of the deep learning pipeline takes the longest to achieve. This is not only a time-consuming process, but an expensive one.
- The most valuable part of a deep learning pipeline is the human element – data scientists often wait for hours or days for training to complete, which hurts their productivity and the time to bring new models to market.
- To significantly reduce training time, you can use deep learning GPUs, which enable you to perform AI computing operations in parallel. When assessing GPUs, you need to consider the ability to interconnect multiple GPUs, the supporting software available, licensing, data parallelism, GPU memory use and performance.

Using Consumer GPUs for Deep Learning

- **RTX 8000**: 48 GB VRAM, ~$5,500.
- **RTX 6000**: 24 GB VRAM, ~$4,000.
- **Titan RTX**: 24 GB VRAM, ~$2,500.
- **RTX 2080 Ti**: 11 GB VRAM, ~$1,150. *
- **GTX 1080 Ti**: 11 GB VRAM, ~$800 *
- **RTX 2080**: 8 GB VRAM, ~$720. *
- **RTX 2070**: 8 GB VRAM, ~$500
- **RTX 2060**: 6 GB VRAM, ~$359.

TensorFlow GPU Performance Comparison: Inception VGG16 XLA, no XLA



Cloud GPUs

AWS Pricing: https://aws.amazon.com/ec2/instance-types/p3/
Azure Pricing: https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/
Google Pricing: https://cloud.google.com/compute/all-pricing#gpus
Linode Pricing: https://www.linode.com/pricing/