Accelerate Machine Learning Workloads Using Amazon EC2 P3 Instances

Chetan Kapoor, Senior Product Manager – AWS EC2

July 31st, 2018

ML @ AWS
OUR MISSION

Put machine learning in the hands of **every developer** and data scientist

Machine Learning at Amazon: A long heritage

Personalized Recommendations

Fulfillment automation & Inventory Management

Drones

Voice driven Interactions

Inventing entirely new Customer experiences

The AWS Machine Learning Stack

APPLICATION SERVICES


REKOGNITION


REKOGNITION VIDEO


POLLY


TRANSCRIBE


TRANSLATE


COMPREHEND


LEX

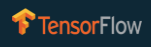
PLATFORMS


Amazon SageMaker


Amazon Mechanical Turk


Amazon Deep Learning AMIs

FRAMEWORKS & INFRASTRUCTURE



P3

NVIDIA Tesla V100 GPUs

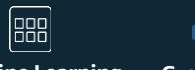
(14x faster than P2)

5,120 Tensor cores

1 Petaflop of compute

128GB of memory

NVLink 2.0


Machine Learning AMIs


Greengrass ML

aws

Amazon EC2 P3 Instances (October 2017)

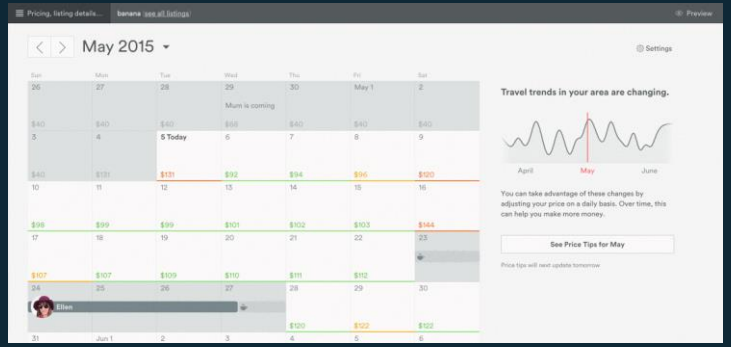
One of the fastest, most powerful GPU instances in the cloud

- Up to eight NVIDIA Tesla V100 GPUs
- 1 PetaFLOPs of computational performance
  – Up to 14x better than P2
- 300 GB/s GPU-to-GPU communication (NVLink) *– 9X better than P2*
- 16GB GPU memory with *900 GB/sec peak GPU memory bandwidth*

aws

Airbnb

“At Airbnb, we’re using machine learning to optimize search recommendations and improve dynamic pricing guidance for hosts, both of which translate to increased booking conversions. These use-cases are highly specific to our industry and require machine learning models that use several different types of data sources, such as guest and host preferences, listing location and condition, seasonality, and price.

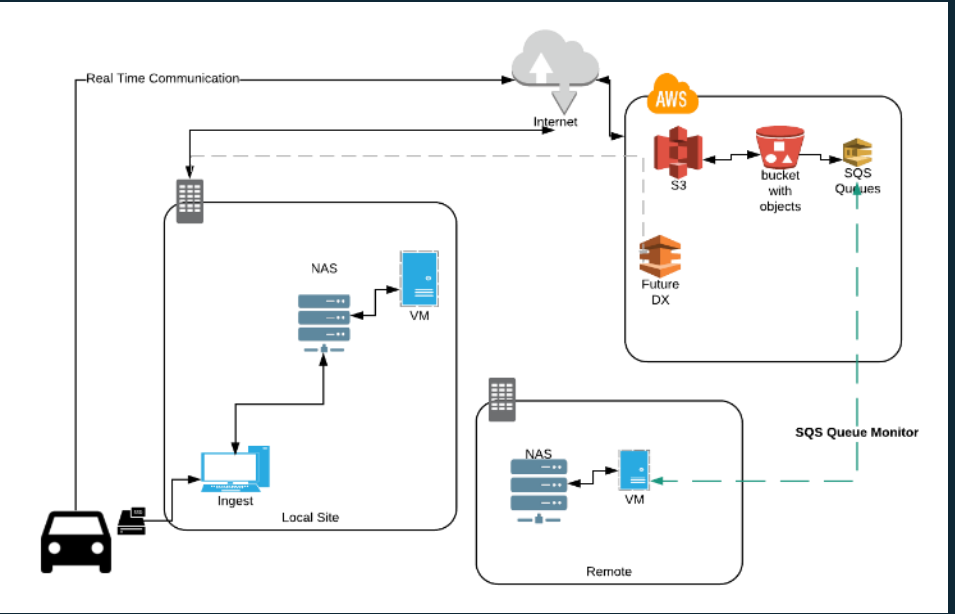


With Amazon EC2 P3 instances, we have the ability to run training workloads faster, enabling us to iterate more, build better machine learning models and reduce cost.”

Toyota Research Institute

Toyota Research Institute's (TRI) mission is to improve the quality of human life through advances in artificial intelligence, automated driving, and robotics. The core compute capability needed by TRI to accelerate the training of their machine learning models is powered by multiple Amazon EC2 P3 instances.

With P3 instances, TRI is seeing a 4X faster time-to-train than the P2 instances they had used previously. This gives them significant agility to optimize and retrain their models quickly and deploy them in their test cars or simulations environment for further testing. Furthermore, the significant performance improvement in P3s over P2s, coupled with pay-as-you-go model translates to lower operating costs.



Machine Learning 101

Recommender System

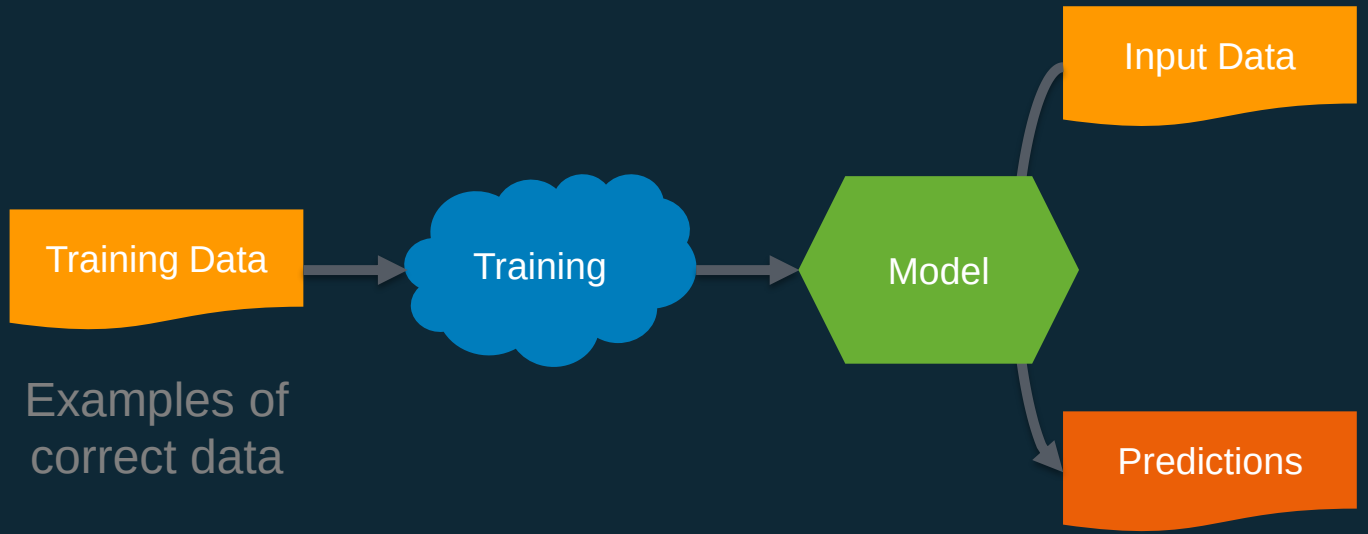
```
/* Predict what Star Rating will user A give movie
M*/

Float predictRating(User A, Movie M)
{
   //How???
}
```

Recommender System

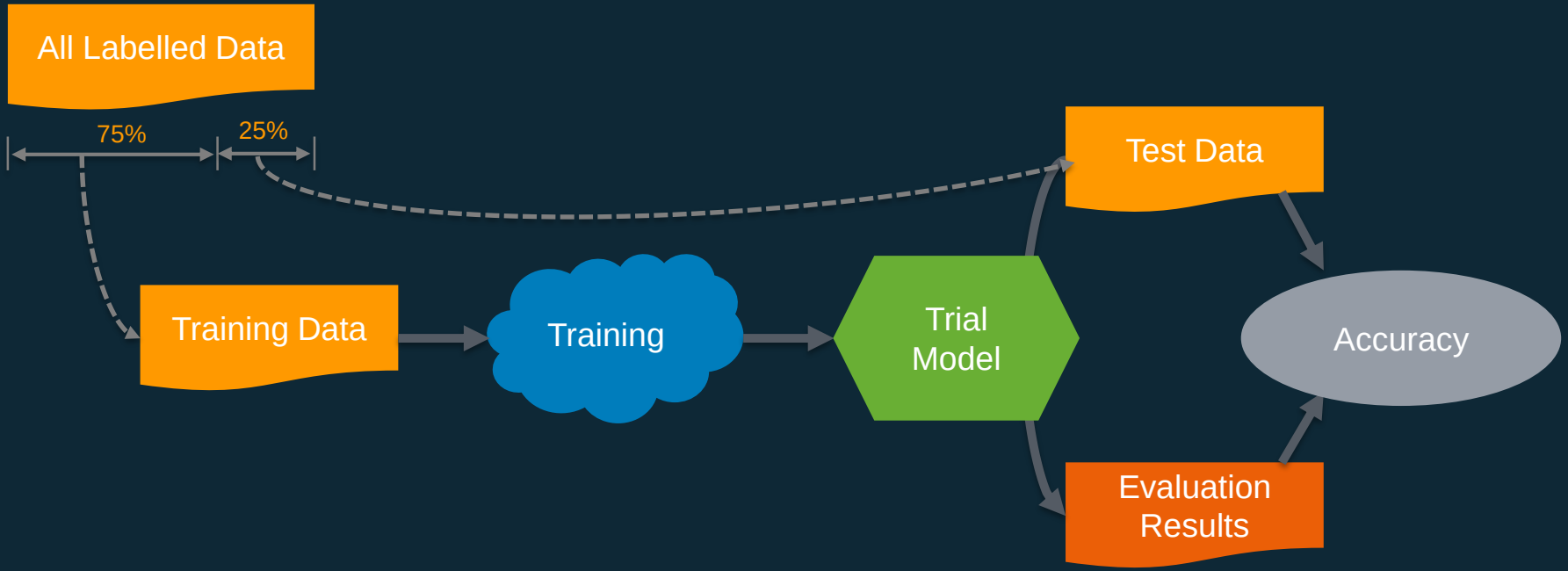
```
/* Predict what Star Rating will user A give movie
M*/

Float predictRating(User A, Movie M)
{
   return mlModel.run(A,M);
}
```


aws

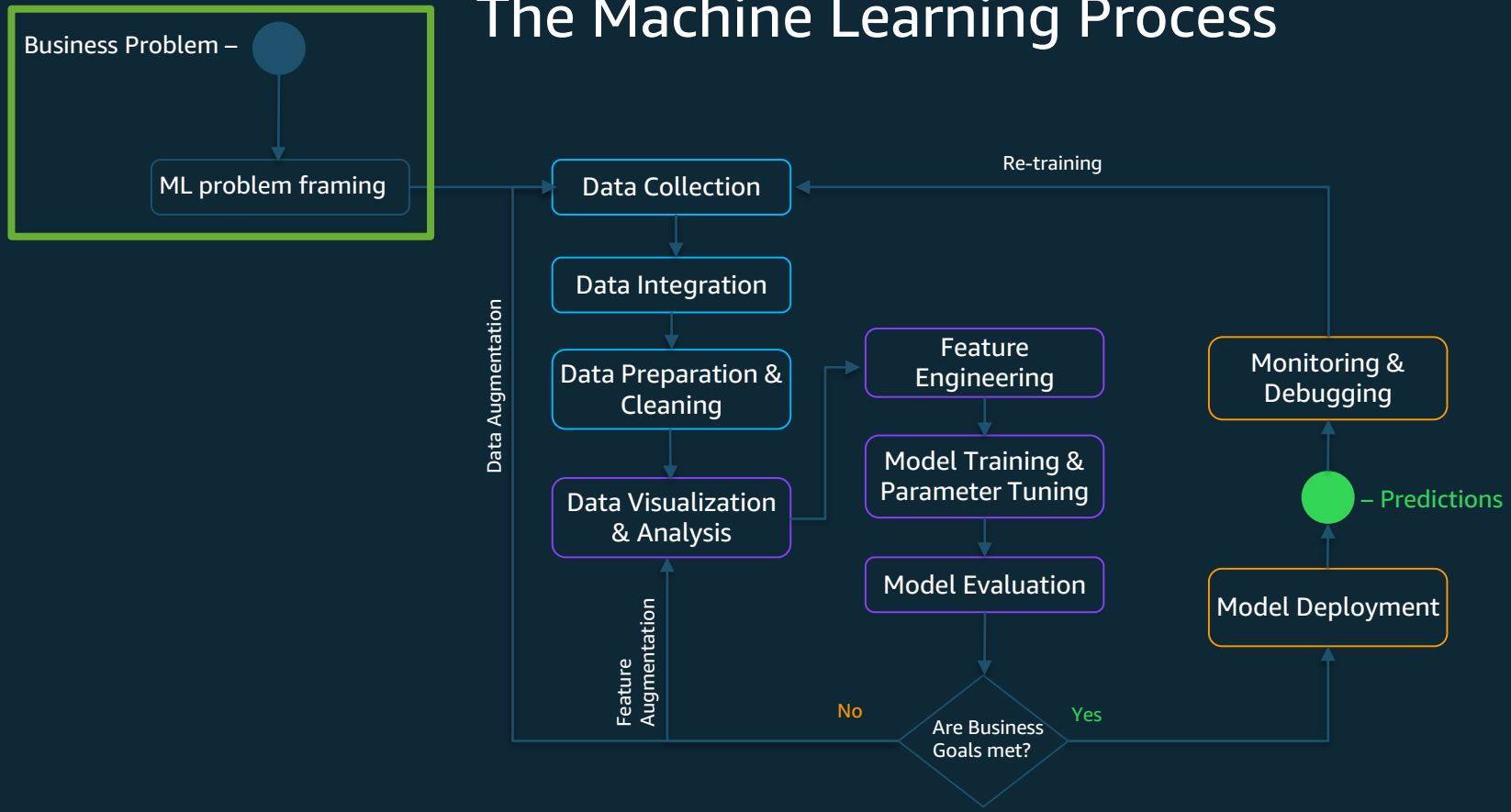
Training a Machine Learning Model – Supervised Learning



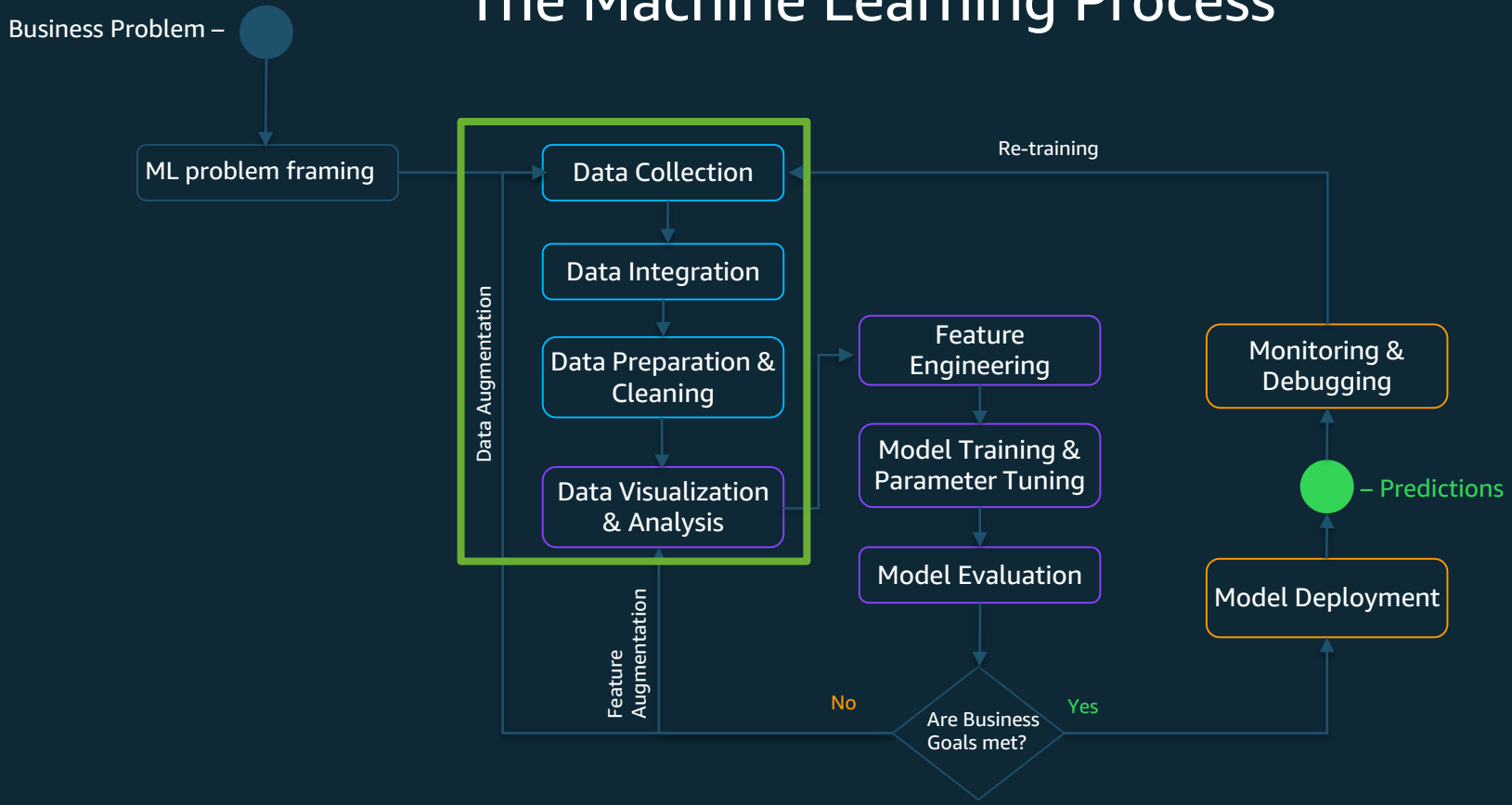
Training a Machine Learning Model – Supervised Learning



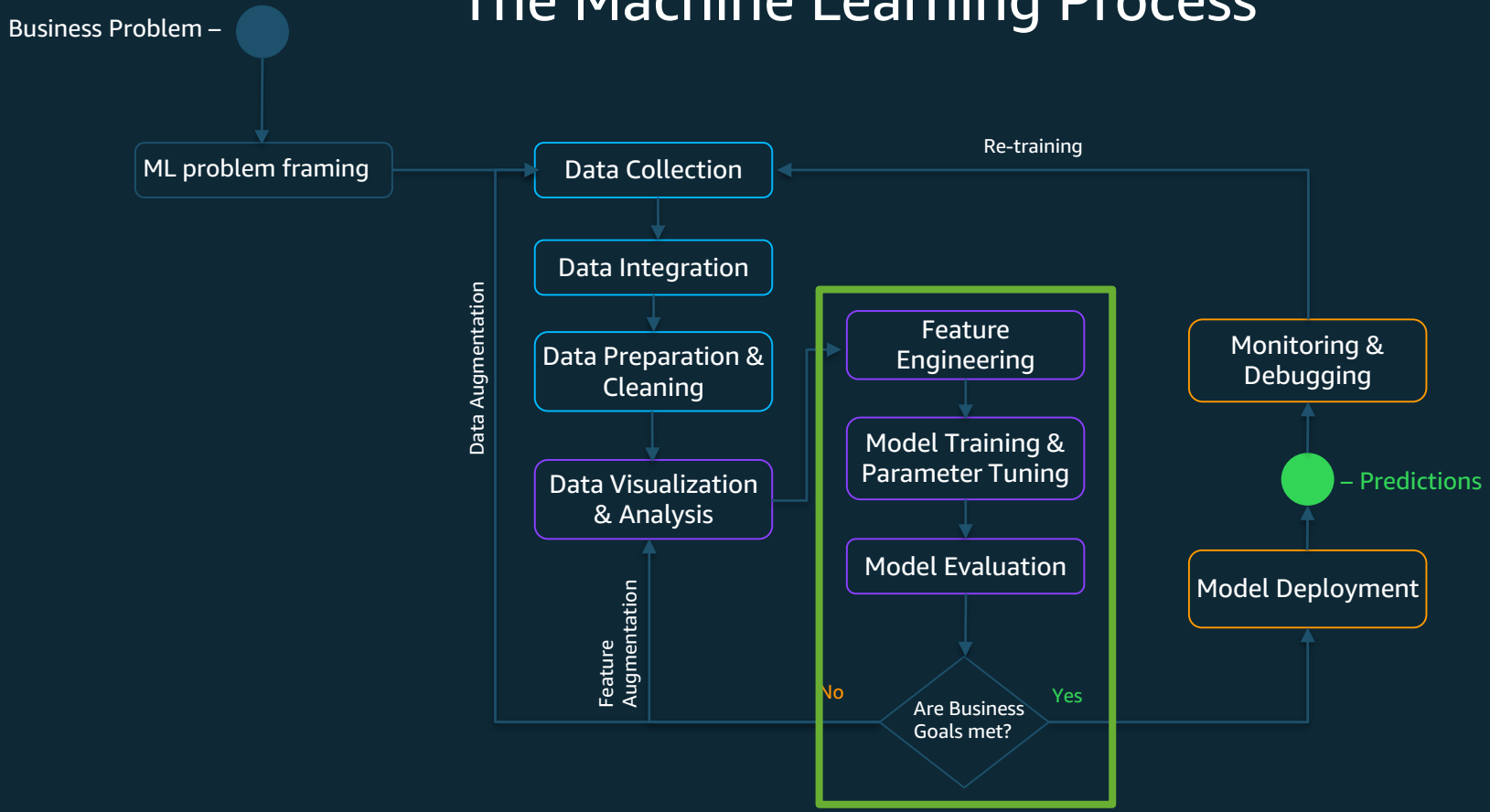
The Machine Learning Process



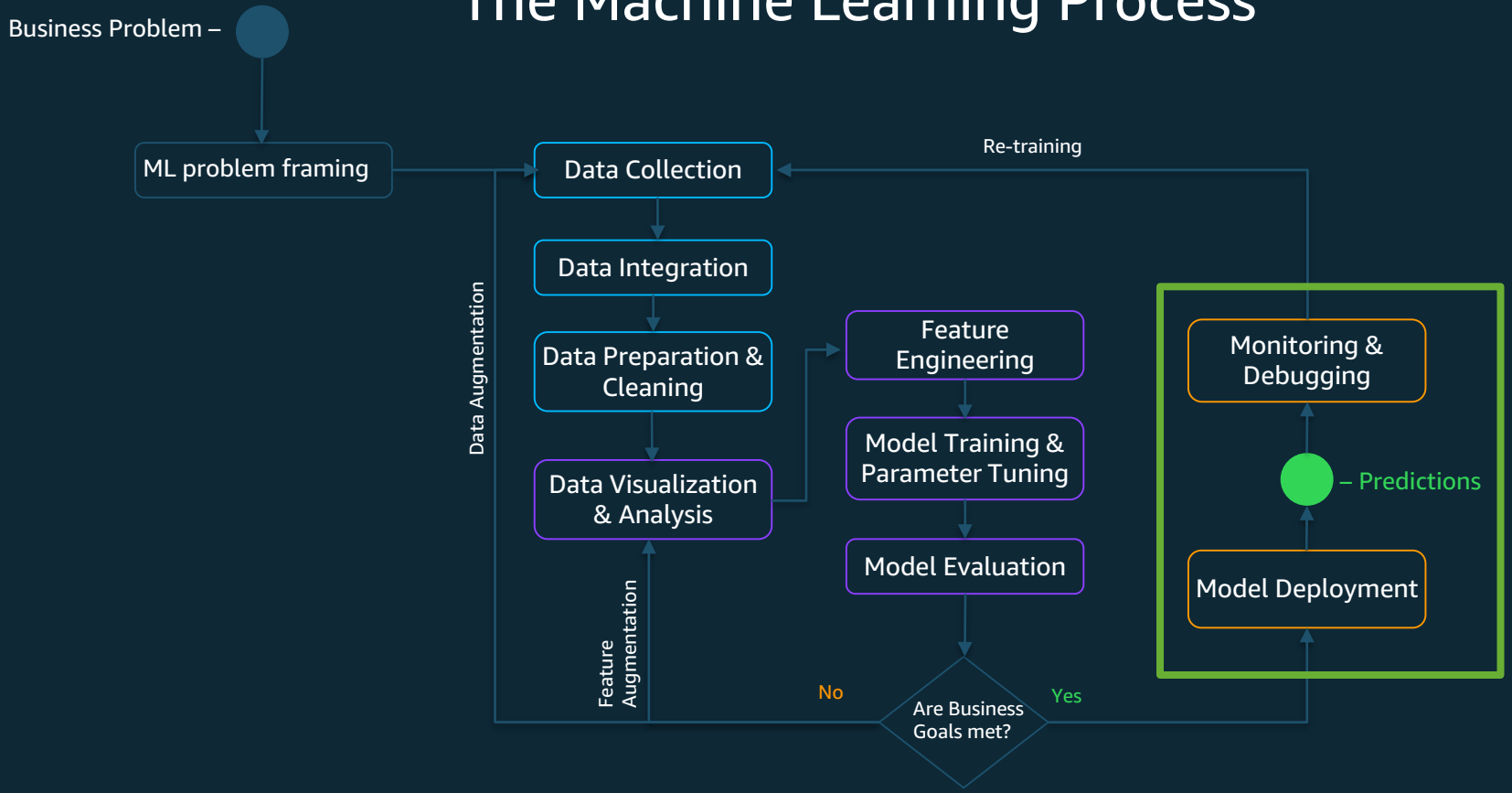
The Machine Learning Process



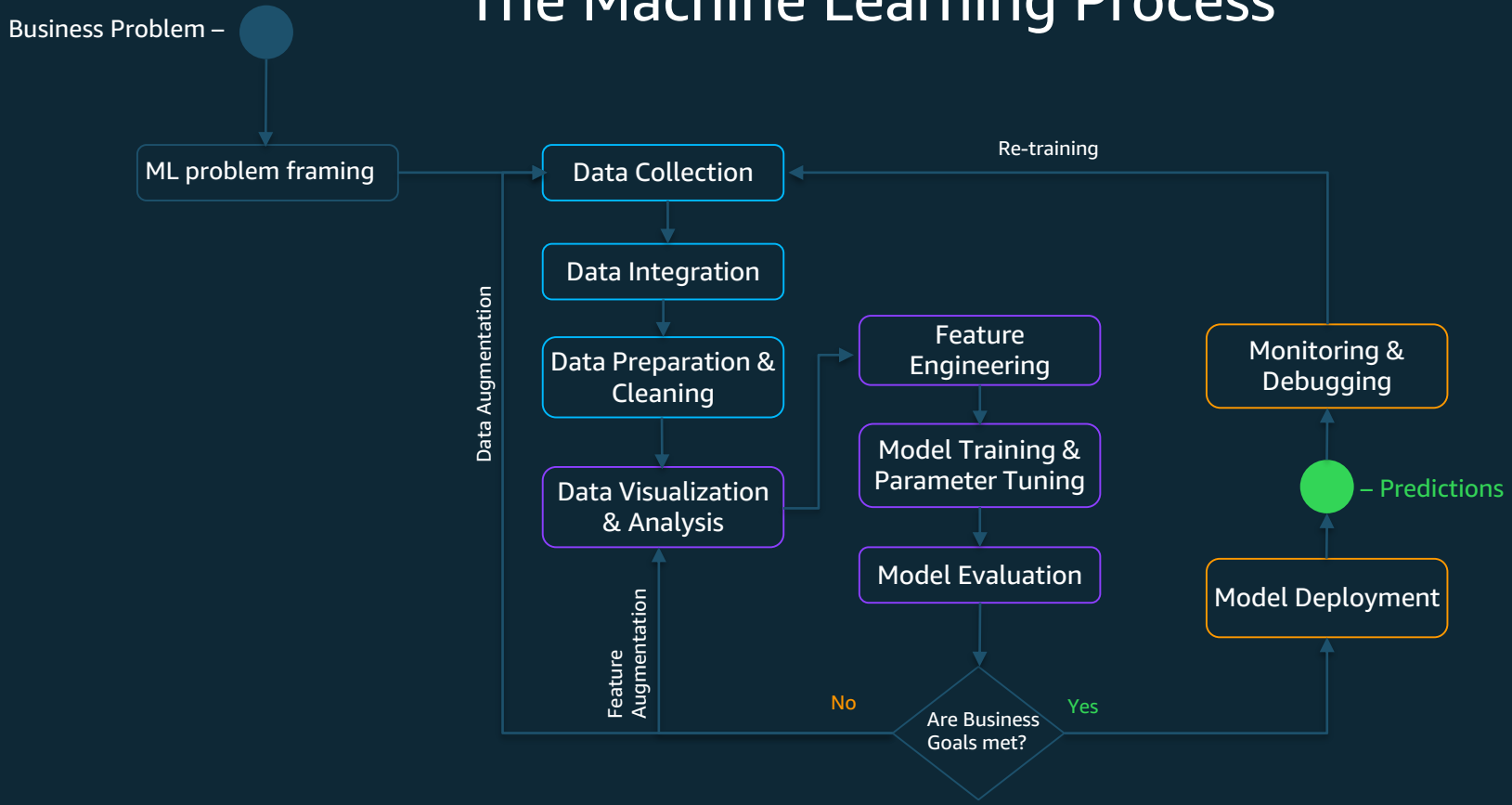
The Machine Learning Process



The Machine Learning Process



The Machine Learning Process



Training Machine Learning Models

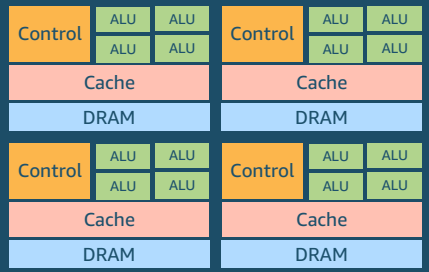
AlexNet, 2012

- A large, deep convolutional neural network with 5 convolutional layer, 60 million parameters and 650,000 neurons
- Created by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton
- Won the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge)
- Used two NVIDIA GTX 580 GPUs
- Took nearly a week to train!

Source - https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

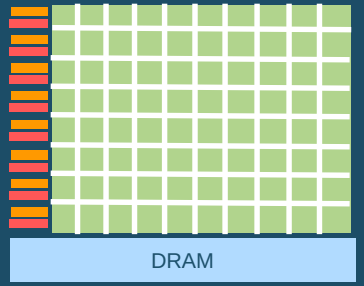
CPUs vs GPUs vs FPGA for Compute

CPU



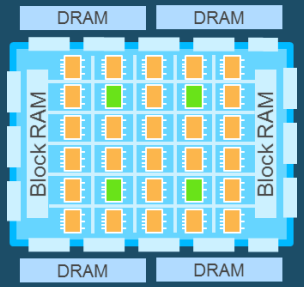
- 10s-100s of processing cores
- Pre-defined instruction set & datapath widths
- Optimized for general-purpose computing

GPU



- 1,000s of processing cores
- Pre-defined instruction set and datapath widths
- Highly effective at parallel execution

FPGA

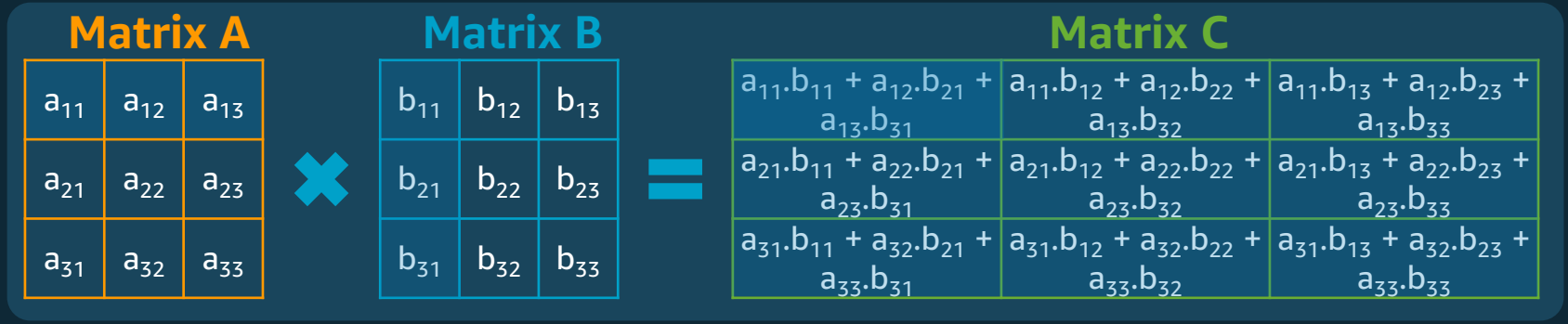


- Millions of programmable digital logic cells
- No predefined instruction set or datapath widths
- Hardware timed execution

aws

GPUs for Machine Learning Training

- Routines for training machine learning models fundamentally map to matrix multiplications
- This coupled with extremely high memory bandwidth makes GPUs ideal for training

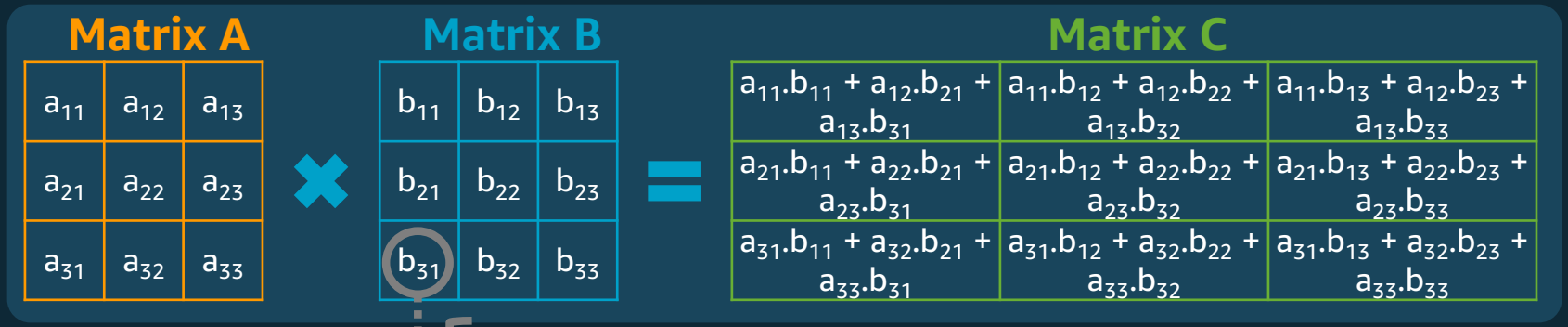


These multiply and accumulate operations can be parallelized across 1,000s of core available in a typical GPU

aws

GPUs for Machine Learning Training

- Routines for training machine learning models fundamentally map to matrix multiplications
- This coupled with extremely high memory bandwidth makes GPUs ideal for training



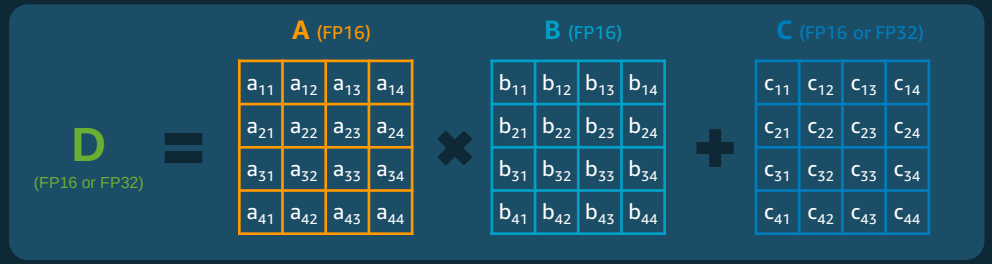
Numeric Precision:

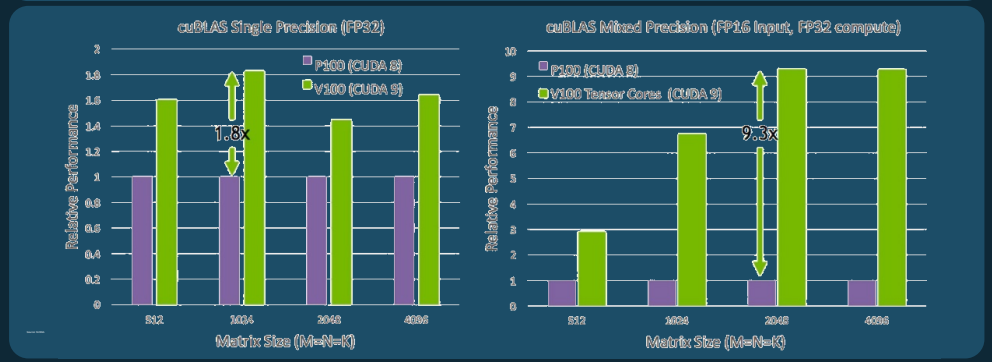
- Half-Precision (FP16) – 16 bit Floating Point
- Single-Precision (FP32) – 32 bit Floating Point
- Double-Precision (FP64) – 64 bit Floating Point

aws

New Tensor Cores in NVIDIA V100 GPUs

- In addition to 5,120 CUDA Cores, each Tesla V100 has 640 Tensor Cores with 125 TFLOPs of mixed-precision performance per GPU
- Each Tensor Core provides a 4x4x4 matrix processing array, which performs the operation D = A X B + C.





AWS EC2 P3 Instances for Accelerating Development of ML Applications

Amazon EC2 P3 Instances (October 2017)

One of the fastest, most powerful GPU instances in the cloud

- Up to eight NVIDIA Tesla V100 GPUs
- 1 PetaFLOPs of computational performance
  – Up to 14x better than P2
- 300 GB/s GPU-to-GPU communication (NVLink) *– 9X better than P2*
- 16GB GPU memory with *900 GB/sec peak GPU memory bandwidth*

aws

P3 Instances Details

Instance Size	GPUs	GPU Peer to Peer	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth	On-Demand Price/hr*	1-yr RI Effective Hourly*	3-yr RI Effective Hourly*
P3.2xlarge	1	No	8	61	Up to 10Gbps	1.7Gbps	$3.06	$1.99 (35% Disc.)	$1.23 (60% Disc.)
P3.8xlarge	4	NVLink	32	244	10Gbps	7Gbps	$12.24	$7.96 (35% Disc.)	$4.93 (60% Disc.)
P3.16xlarge	8	NVLink	64	488	25Gbps	14Gbps	$24.48	$15.91 (35% Disc.)	$9.87 (60% Disc.)
Note: P3 instances are also supported via Spot Instances with up to 70% discount									

Regional Availability

P3 instances are generally available in AWS US East (Northern Virginia), US East (Ohio), US West (Oregon), EU (Ireland), Asia Pacific (Seoul), Asia Pacific (Tokyo), AWS GovCloud (US) and China (Beijing) Regions.

Framework Support

P3 instances and their V100 GPUs supported across all major frameworks (such as TensorFlow, MXNet, PyTorch, Caffe2 and CNTK)

aws

EC2 Pricing and Usage Options

On-Demand

Pay for compute capacity per second. No longer-term commitments or upfront payments are needed. You can increase or decrease your compute capacity depending on the demands of your application and only pay the hourly rate.

Recommended for:

- Users that prefer the low cost and flexibility of Amazon EC2 without any up-front payment or long-term commitment
- Applications with short-term, spiky, or unpredictable workloads that cannot be interrupted
- Applications being developed or tested on Amazon EC2 for the first time

Reserved Instances

Reserved Instances provide you with a significant discount (up to 60%) compared to On-Demand instance pricing. In addition, when Reserved Instances are assigned to a specific Availability Zone, they provide a capacity reservation, giving you additional confidence in your ability to launch instances when you need them.

Recommended for:

- Applications with steady state usage
- Applications that may require reserved capacity
- Customers that can commit to using EC2 over a 1 or 3 year term to reduce their total computing costs

Spot Instances

Spot instances allow you to request spare Amazon EC2 computing capacity for up to 70% off the On-Demand price. The only difference between On-Demand instances and Spot Instances is that Spot instances can be interrupted by EC2 with two minutes of notification when EC2 needs the capacity back.

Recommended for:

- Applications that have flexible start and end times
- Applications that are only feasible at very low compute prices
- Users with urgent computing needs for large amounts of additional capacity

aws

P3 Instances Details

Instance Size	GPUs	GPU Peer to Peer	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth	On-Demand Price/hr*	1-yr RI Effective Hourly*	3-yr RI Effective Hourly*
P3.2xlarge	1	No	8	61	Up to 10Gbps	1.7Gbps	$3.06	$1.99 (35% Disc.)	$1.23 (60% Disc.)
P3.8xlarge	4	NVLink	32	244	10Gbps	7Gbps	$12.24	$7.96 (35% Disc.)	$4.93 (60% Disc.)
P3.16xlarge	8	NVLink	64	488	25Gbps	14Gbps	$24.48	$15.91 (35% Disc.)	$9.87 (60% Disc.)

- P3 instances provide GPU-to-GPU data transfer over **NVLink**
- P2 instanced provided GPU-to-GPU data transfer over **PCI Express**

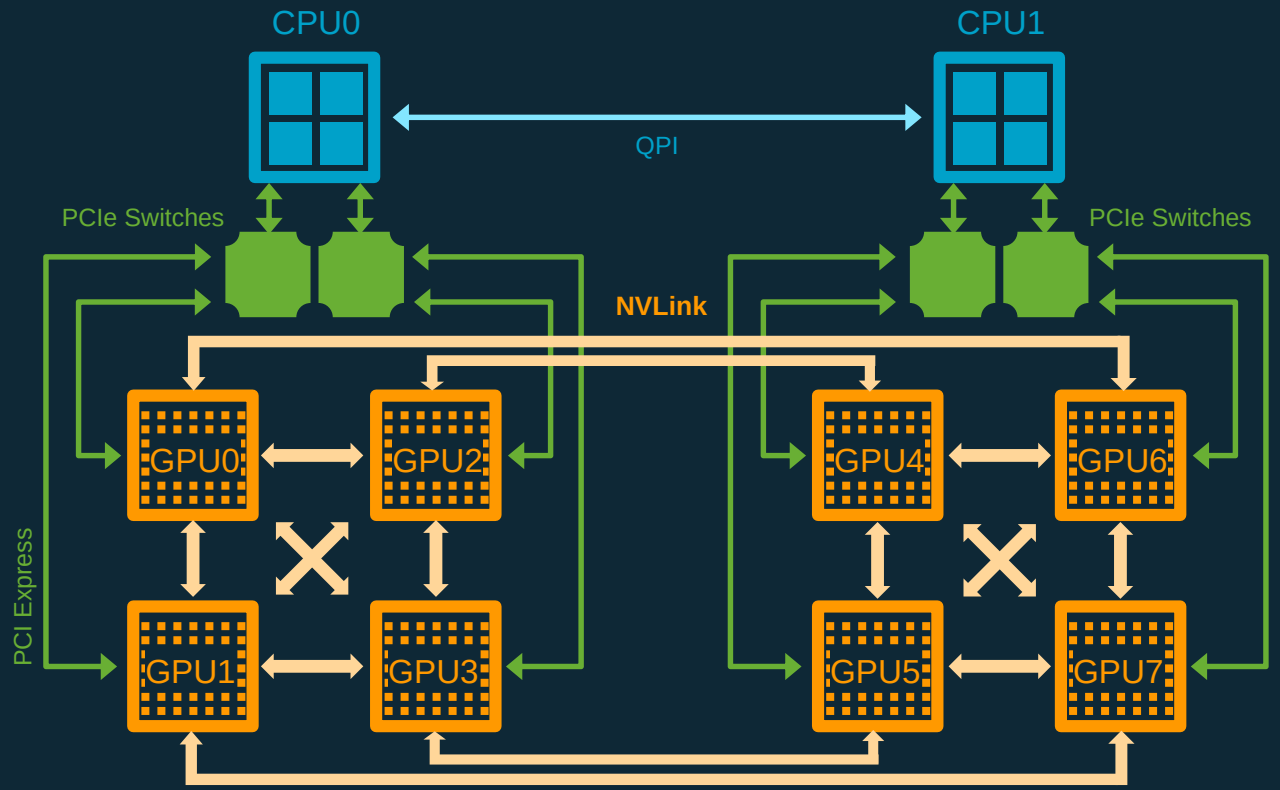
aws

P3 vs P2 Peer-to-Peer Configurations

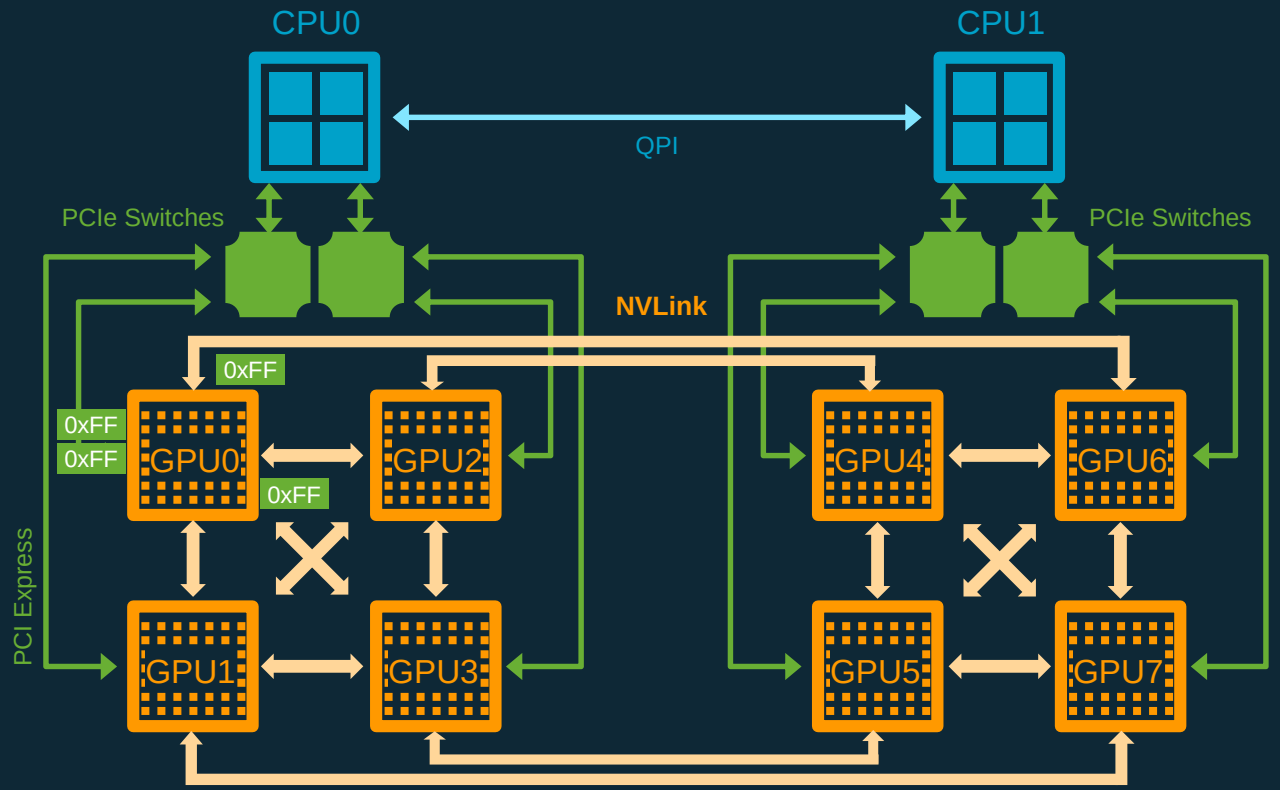
Description	P3.16xlarge	P2.16xlarge	P3 GPU Performance Improvement
Number of GPUs	8	16	-
Number of Accelerators	8 (V100)	8 (K80)	
GPU – Peer to Peer	NVLink – 300 GB/s	PCI-Express - 32 GB/s	9.4X
CPU to GPU Throughput PCIe throughput per GPUs	8 GB/s	1 GB/s	8X
CPU to GPU Throughput Total instance PCIe throughput	64 GB/s (Four x16 Gen3)	16 GB/s (One x16 Gen3)	4X

aws

P3 PCIe and NVLink Configurations



P3 PCIe and NVLink Configurations



AWS Storage Options

EFS

Highly available, multi-AZ, fully managed network-attached elastic file system.

For near-line, highly-available storage of files in a traditional NFS format (NFSv4).

Use for read-often, temporary working storage

EC2+EBS

Create a single-AZ shared file system using EC2 and EBS, with third-party or open source software (e.g., ZFS, Intel Lustre, etc).

For near-line storage of files optimized for high I/O performance.

Use for high-IOPs, temporary working storage

Amazon S3

Secure, durable, highly-scalable object storage. Fast access, low cost.

For long-term durable storage of data, in a readily accessible get/put access format.

Primary durable and scalable storage for data

Amazon Glacier

Secure, durable, long term, highly cost-effective object storage.

For long-term storage and archival of data that is infrequently accessed.

Use for long-term, lower-cost archival of data

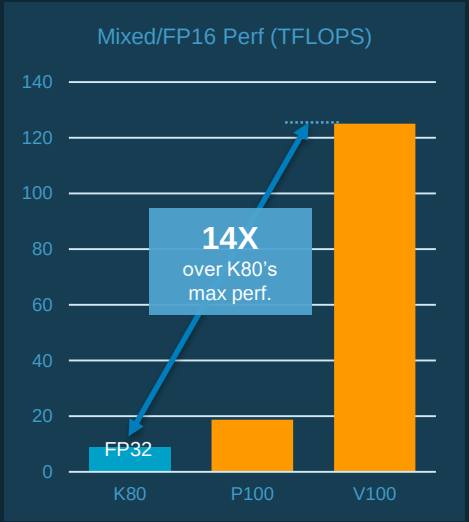
aws

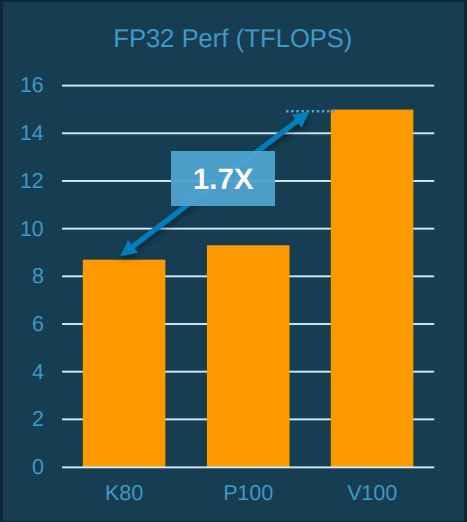
Lowering Training Times with AWS EC2 P3 Instances

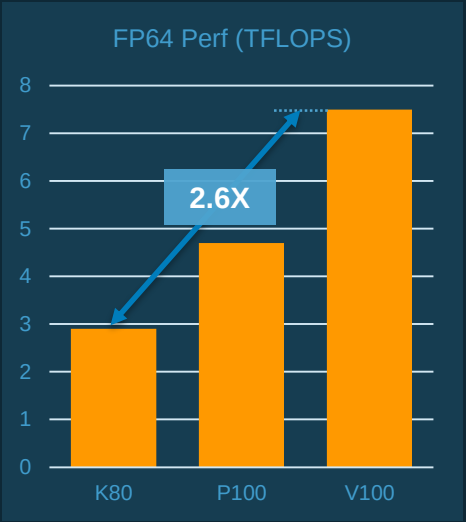
AWS P3 vs P2 Instance

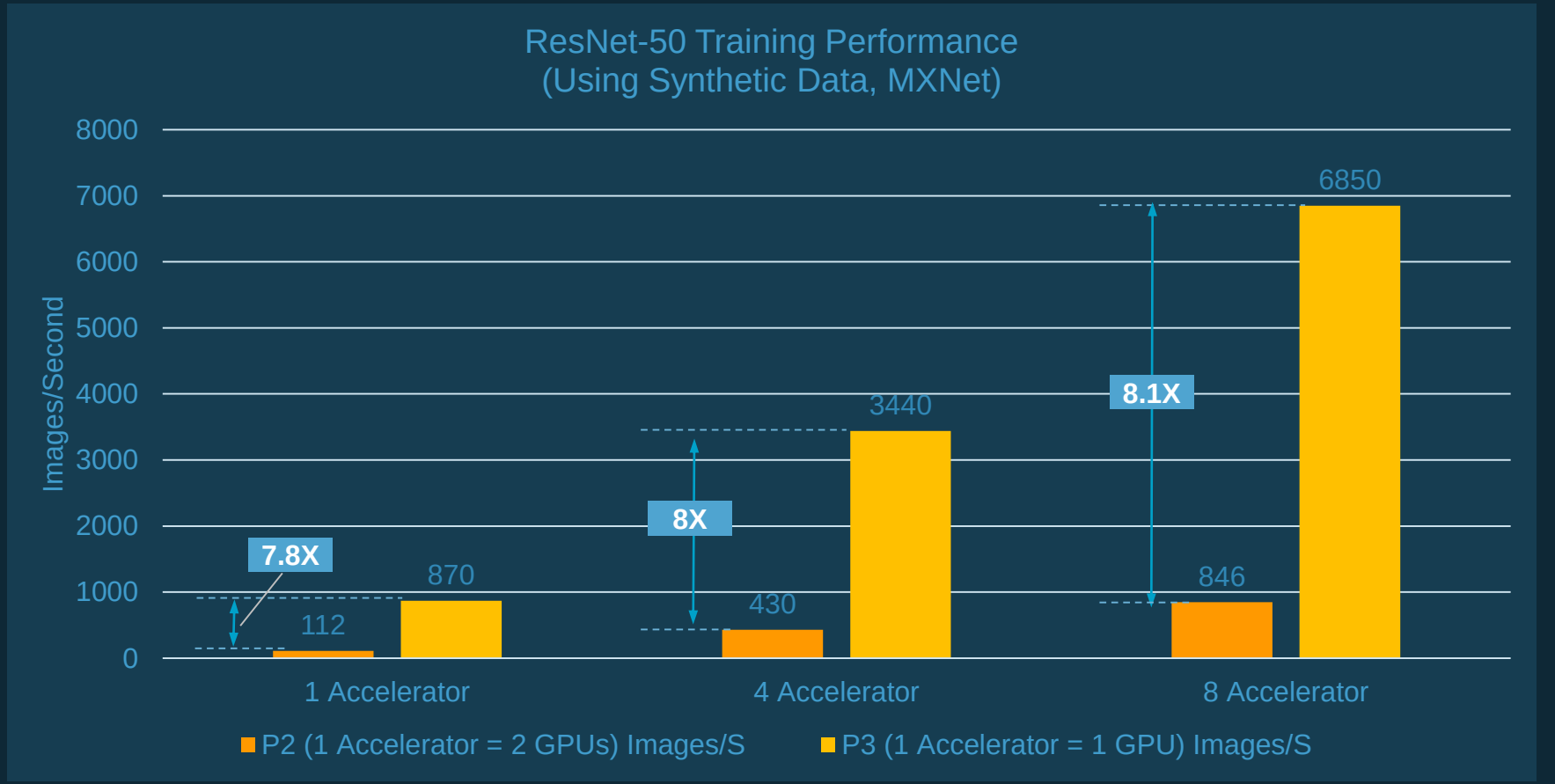
GPU Performance Comparison

- P2 Instances use K80 Accelerator (Kepler Architecture)
- P3 Instances use V100 Accelerator (Volta Architecture)



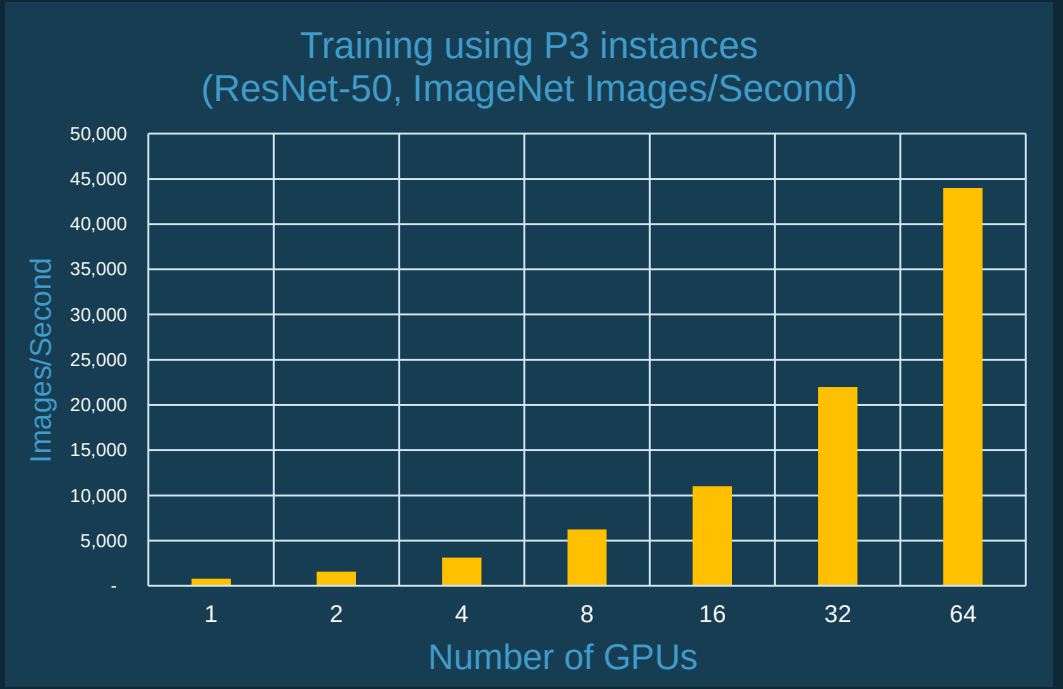






Scaling Performance using Distributed Training

- Using single P3 instances, with Volta GPUs, customers can cut down training times of their machine learning models from days to a few hours.
- Using distributed training via multiple P3 instances, high performance networking and storage solutions, customers can further cut down their time-to-train from hours in to minutes.
- Example – We been able to train ResNet-50 to Top1 validation accuracy of 76% in 47 mins using 8 P3.16xlarge instances.



https://aws.amazon.com/blogs/machine-learning/scalable-multi-node-deep-learning-training-using-gpus-in-the-aws-cloud/

aws

Getting Started

AWS Deep Learning AMI

- Get started quickly with easy-to-launch tutorials
- Hassle-free setup and configuration
- Pay only for what you use – no additional charge for the AMI
- Accelerate your model training and deployment
- Support for popular deep learning frameworks

aws

Demo

Launching an EC2 GPU Instance & running a Synthetic Benchmark

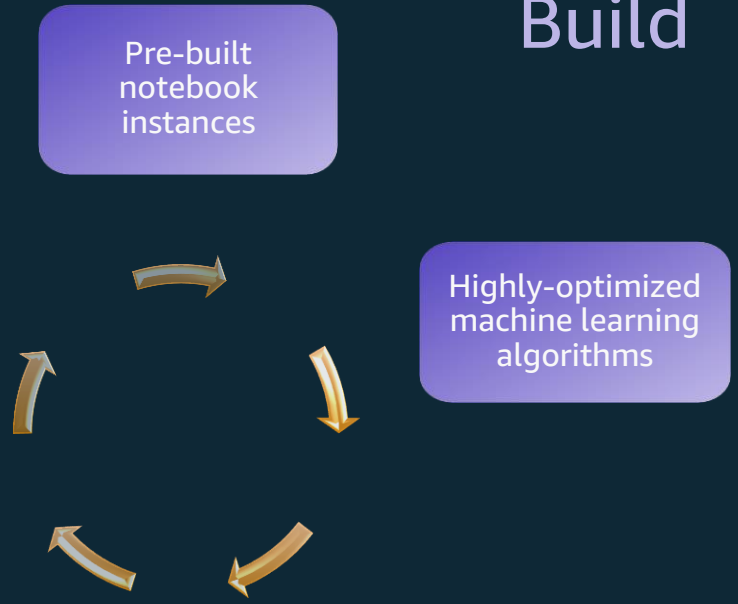
Amazon SageMaker

A **fully managed service** that enables **data scientists** and **developers** to quickly and easily **build** machine-learning based models **into production** smart applications.

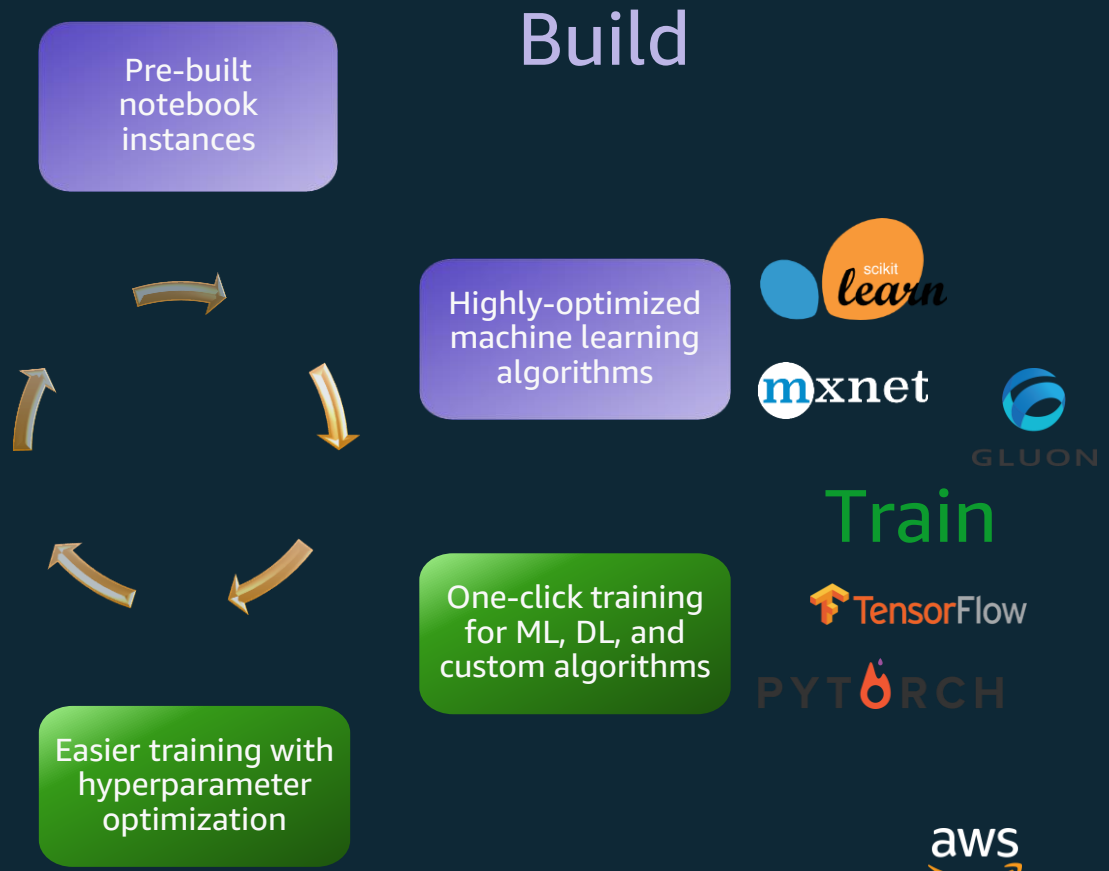
aws

Amazon SageMaker

Build



Amazon SageMaker



aws

Amazon SageMaker

Build

Pre-built notebook instances

Highly-optimized machine learning algorithms

Train

One-click training for ML, DL, and custom algorithms

Easier training with hyperparameter optimization

Deploy

Deployment without engineering effort

Fully-managed hosting at scale

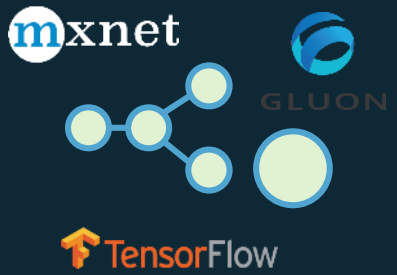
aws

Amazon SageMaker

Build, train, and deploy machine learning models at scale

End-to-End Machine Learning Platform

Zero setup



Flexible Model Training

Pay by the second

aws

Conclusion

Amazon EC2 P3 Instances

One of the fastest, most powerful GPU instances in the cloud

Performance

- Up to 1 PetaFLOPs on computations performance per instances
- Train models in hours instead of days
- Scaling out with multiple P3 reduces training time further and enhances team productivity.

Cost

- Offer flexible pricing options including On-Demand, Reserved and Spot instances
- Up to 70% discount from On-Demand
- Can save 50% for training ML models

Versatile

- Support for all ML frameworks including TensorFlow, MXNet, PyTorch, Caffe2, and CNTK
- Can be used to train different model types (CNN, RNN & GANs)

Ease of Use

- Integrated with Amazon SageMaker – End to end workflow management service
- Alternatively, quickly get Started with AWS Deep Learning AMIs

Broad Availability

- Available at-scale across 8 regions worldwide
- Significant adoption by the ML Community
- Most popular platform for running TensorFlow

aws

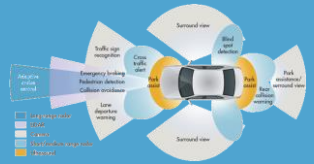
Use-Cases for P3 Instances

Machine Learning/AI

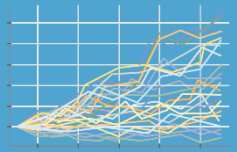
Natural Language Processing

Image and Video recognition

Autonomous vehicle systems

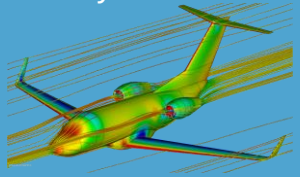


Recommendation Systems



High Performance Computing

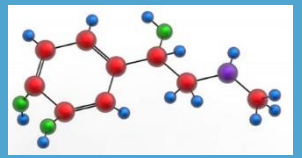
Computational Fluid Dynamics



Financial and Data Analytics

Weather Simulation

Computational Chemistry



aws

Thank You!