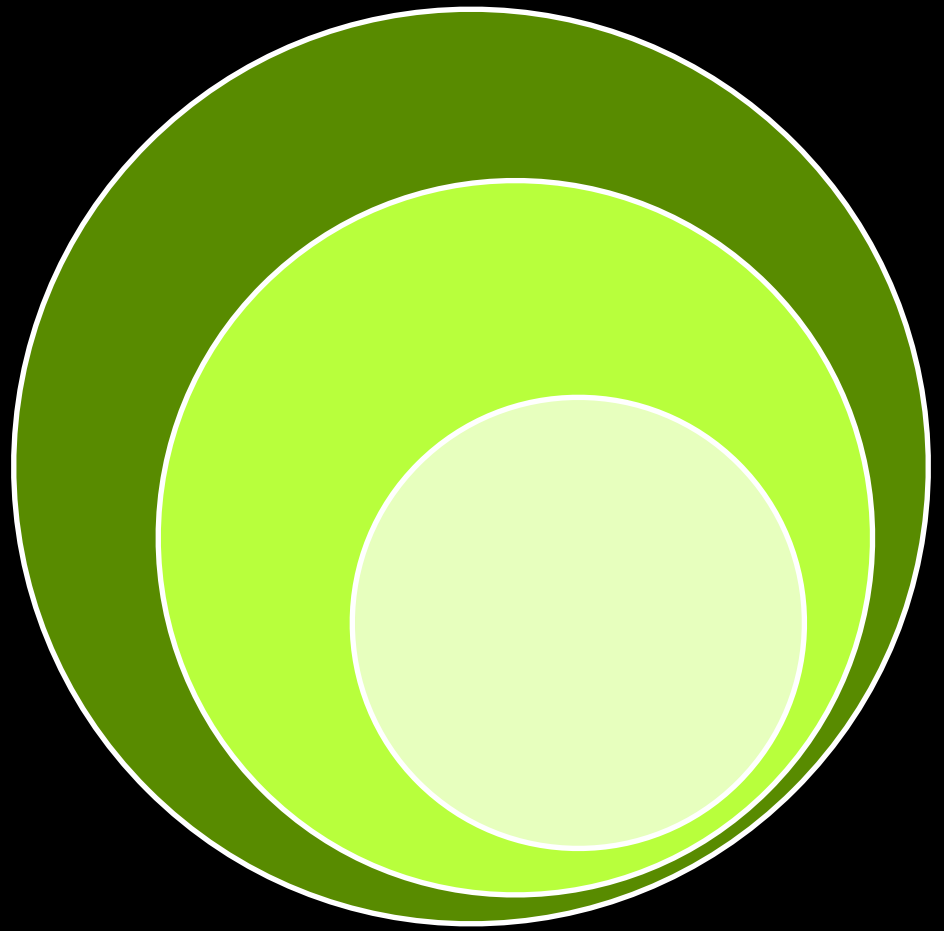
TRAINING: FAST AND SLOW

Pablo Ribalta Lorenzo, 21.5.2019



Artificial intelligence

Grand project to build
non-human intelligence

Machine Learning

Machines that learn to be smarter

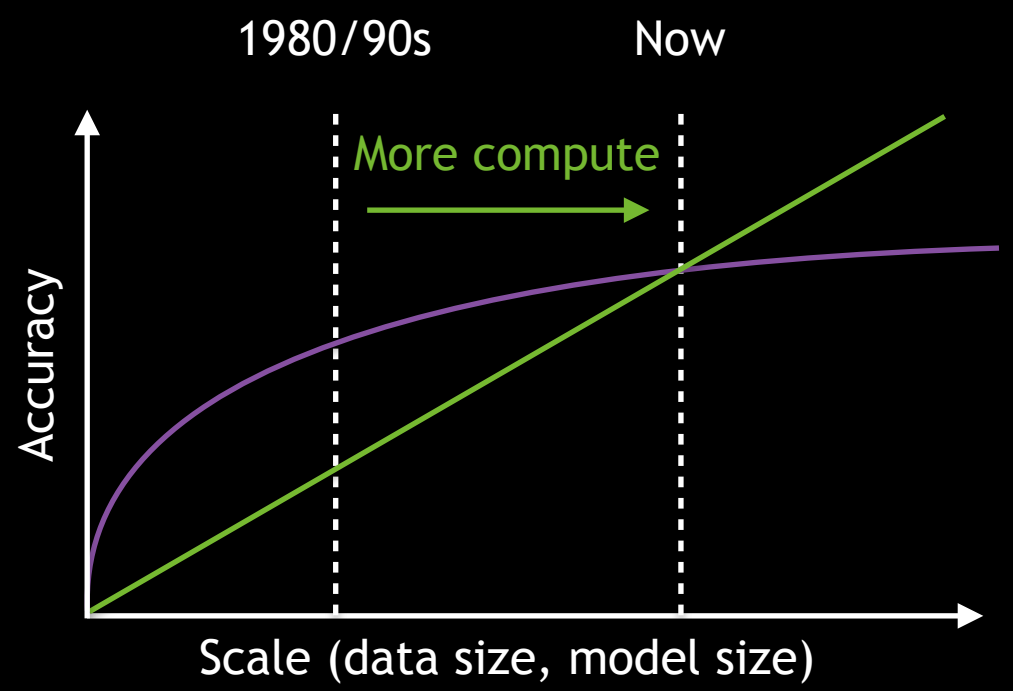
Deep Learning

Particular kind of machine learning

WHAT IS DEEP LEARNING?

Modern Reincarnation of Artificial Neural Networks

Collection of simple **trainable** mathematical units, organized in layers, that work together to solve complicated tasks



Neural Networks

Other Approaches

What's New

Layered network architecture,
New training math,
Scale

Key benefit

Highly accurate
Learns features from raw data
No feature engineering required

TRAINING DEEP NEURAL NETWORKS

Practice makes perfect

The process of training a deep neural network is **iterative.** We set out to achieve our target goal, and let the network figure out how to get there

Loss function

Provides distance to target
Maps directly to accuracy
Our goal is to reduce it over time

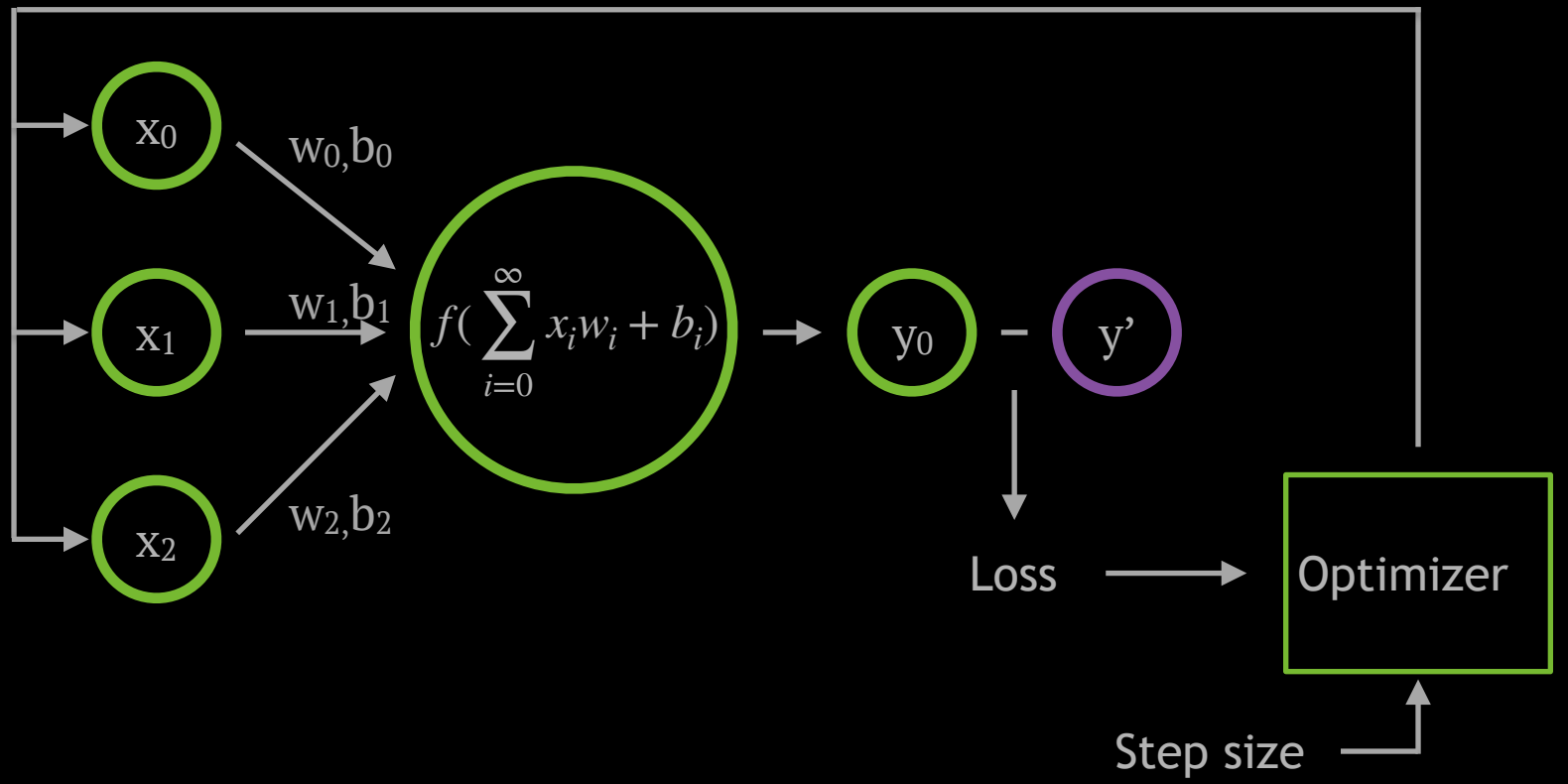
Optimizer

Minimizes loss by adjusting model
Exploits gradient descent
Decides the rate of improvement

TRAINING DEEP NEURAL NETWORKS

Deep neural networks as function approximations

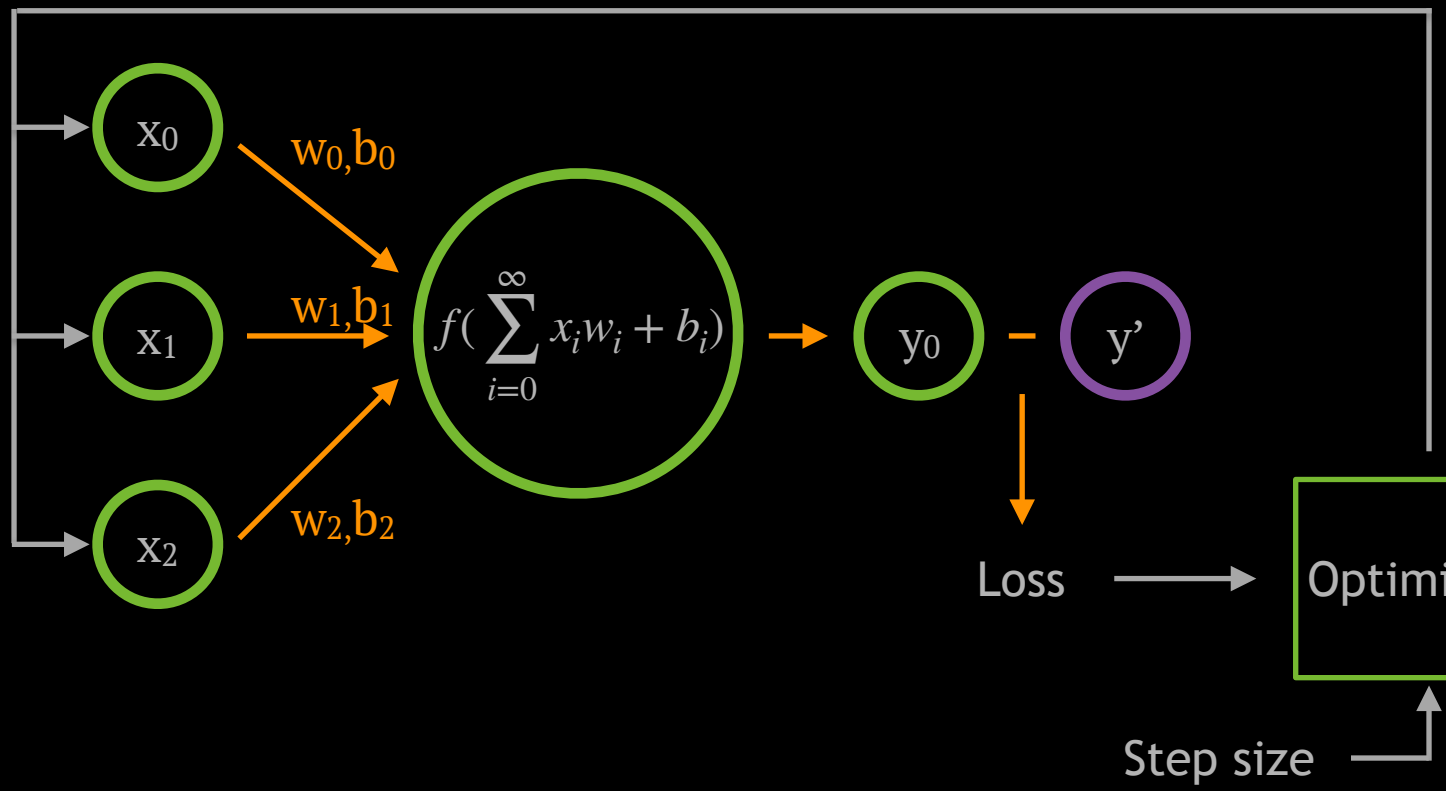
Backpropagation is the algorithm that helps neural networks learn



TRAINING DEEP NEURAL NETWORKS

Deep neural networks as function approximations

Backpropagation is the algorithm that helps neural networks learn



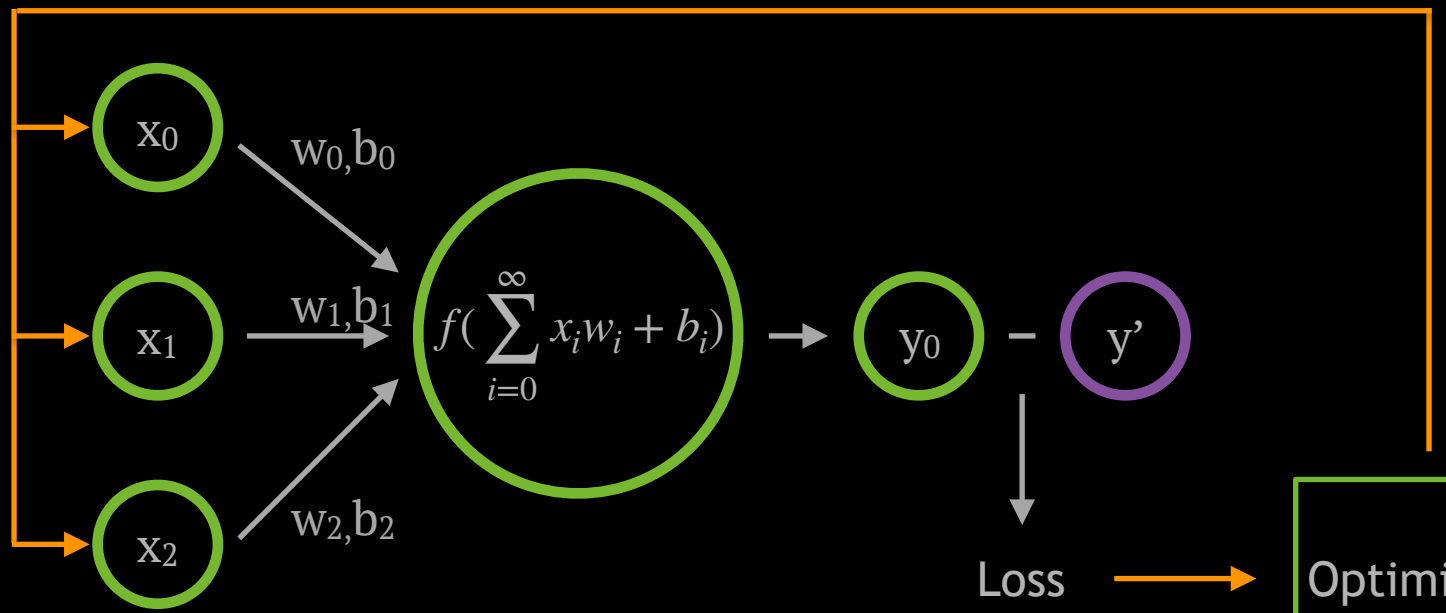
Forward pass

Takes an input vector
Computes output with weights
Uses ground truth to get loss

TRAINING DEEP NEURAL NETWORKS

Deep neural networks as function approximations

Backpropagation is the algorithm that helps neural networks learn



Forward pass

Takes an input vector
Computes output with weights
Uses ground truth to get loss

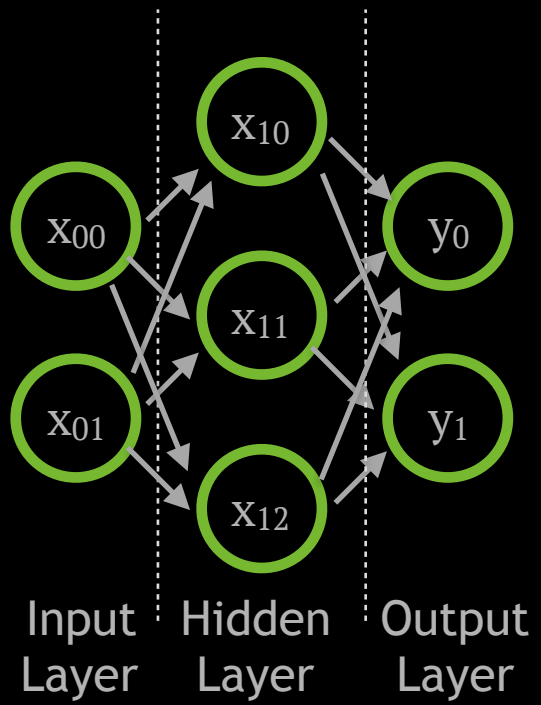
Backward pass

Calculate gradient of each weight
Update weights using step size

TRAINING DEEP NEURAL NETWORKS

Backpropagation at scale

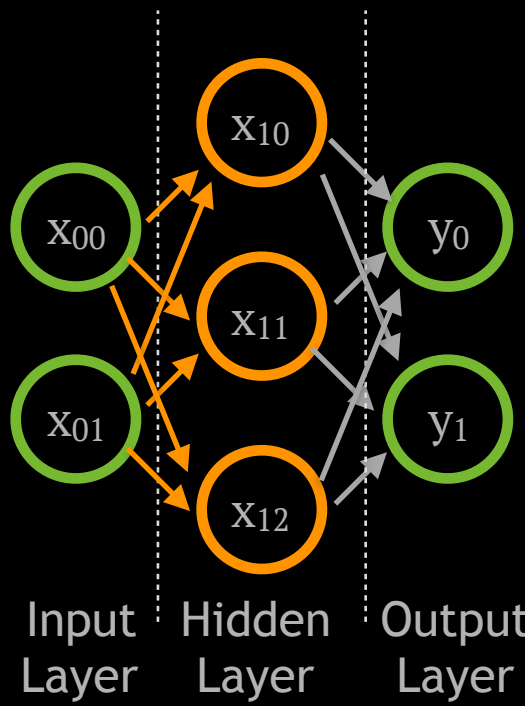
Neurons are grouped in layers for efficiency, and forward and backward passes are computed through **matrix multiplication**, which is very GPU friendly



TRAINING DEEP NEURAL NETWORKS

Backpropagation at scale

Neurons are grouped in layers for efficiency, and forward and backward passes are computed through **matrix multiplication**, which is very GPU friendly



$$\begin{bmatrix} w_{00} & w_{01} & w_{02} \\ w_{10} & w_{11} & w_{12} \\ w_{20} & w_{21} & w_{22} \end{bmatrix} \times \begin{bmatrix} x_{00} \\ x_{01} \\ x_{02} \end{bmatrix} + \begin{bmatrix} b_{00} \\ b_{01} \\ b_{02} \end{bmatrix} = \begin{bmatrix} x_{10} \\ x_{11} \\ x_{12} \end{bmatrix}$$

Deep neural networks

Thousands of neurons per layer
Hundreds of layers
Millions of weights to compute
Multiple domains

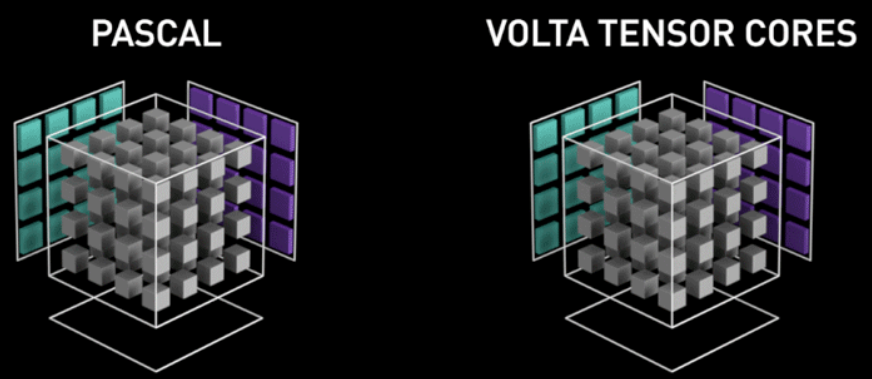
Matrix multiply

Lots of options
Many known optimizations
Very well suited for GPU

MIXED PRECISION TRAINING

Hardware-optimized deep network training

By operating in lower **numerical precision**, it is possible to increase throughput and reduce latency at no expense in accuracy



NVIDIA Tensor Core

New CUDA TensorOp instructions & data format
4x4 matrix processing array
D[FP32] = A[FP16] x B[FP16] + C[FP32]
Optimized for deep learning

Benefits

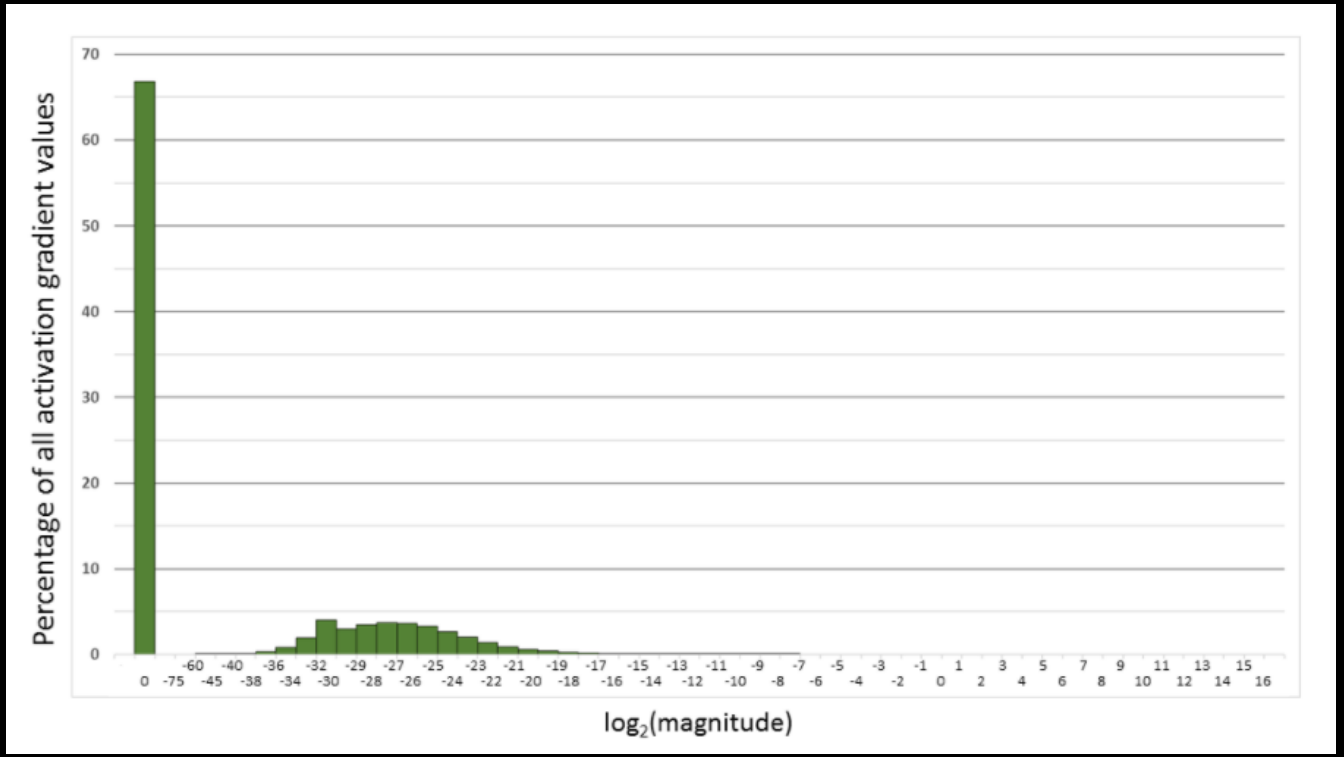
3x4 times **faster**
Reduce memory consumption and bandwidth
Just as **powerful**
No changes in architecture

Activation inputs Weight inputs Output results

MIXED PRECISION TRAINING

Adding support for mixed precision training

Mixed precision requires changes to accommodate the new arithmetic and ensure that numerical problems do not arise



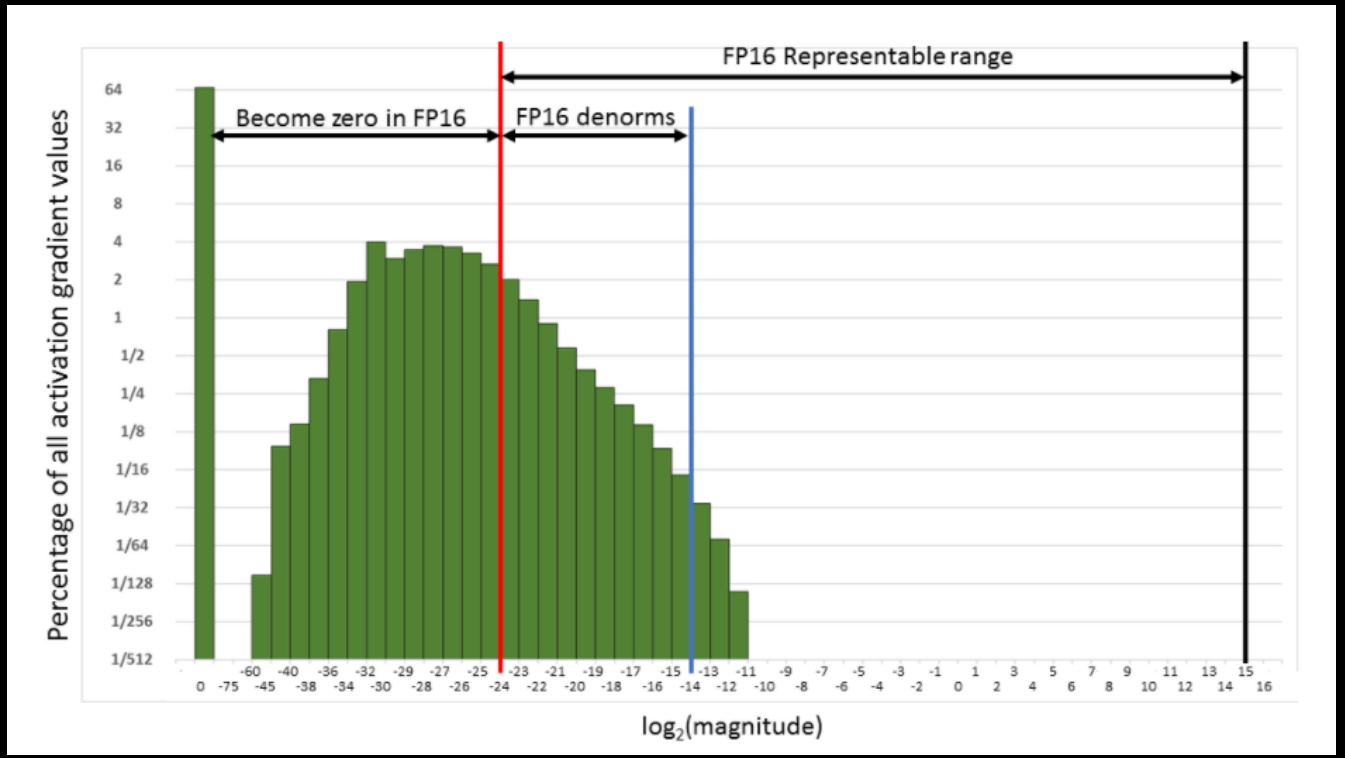
Pitfalls

Imprecise weight updates
Gradients underflow
Maintain precision

MIXED PRECISION TRAINING

Adding support for mixed precision training

Mixed precision requires changes to accommodate the new arithmetic and ensure that numerical problems do not arise



Pitfalls

Imprecise weight updates
Gradients underflow
Maintain precision

IEEE 754 standard

Normalized values = 2^{-14} to 2^{15}
Denormal values = 2^{-24} to 2^{-15}
Normal maximum = 65,504
Normalized minimum = ~6.10e-5
Minimum denormal = ~5.96e-8

MIXED PRECISION TRAINING

A mixed precision solution

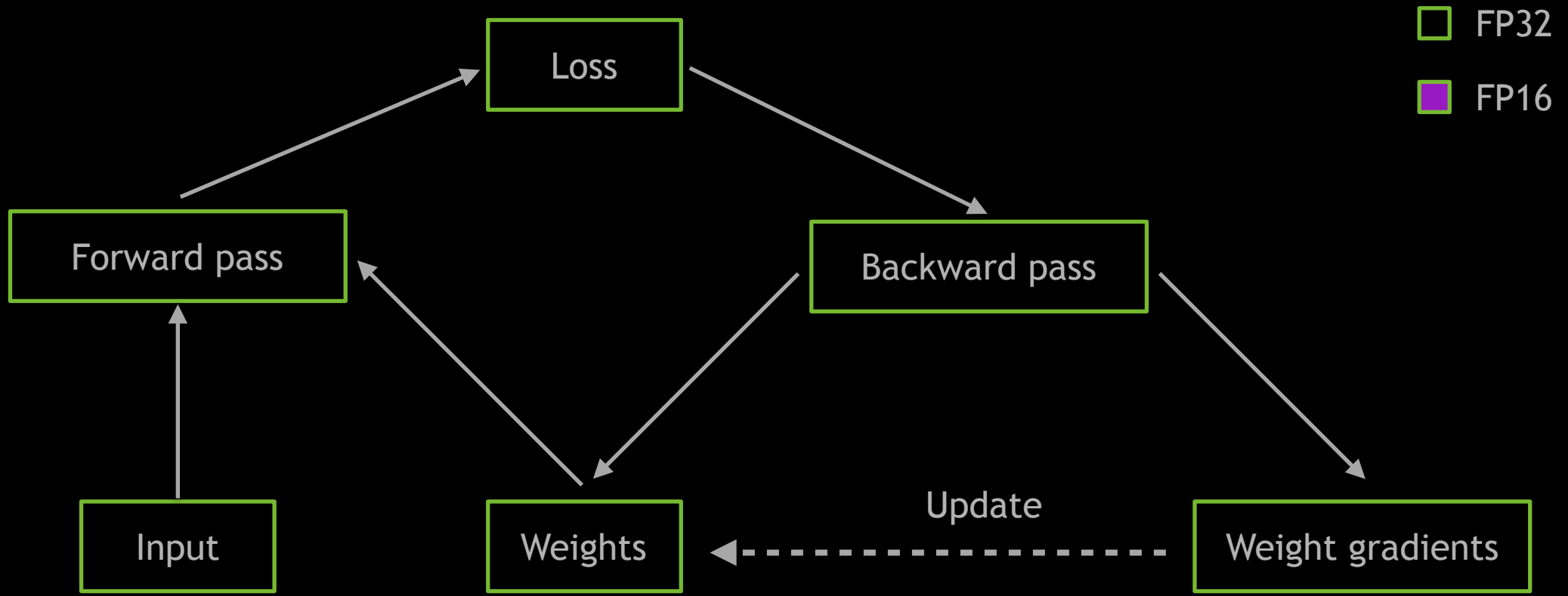
Imprecise weight updates → Master weights in FP32

Gradients underflow → Loss (Gradient) scaling

Maintain precision → Accumulate to FP32

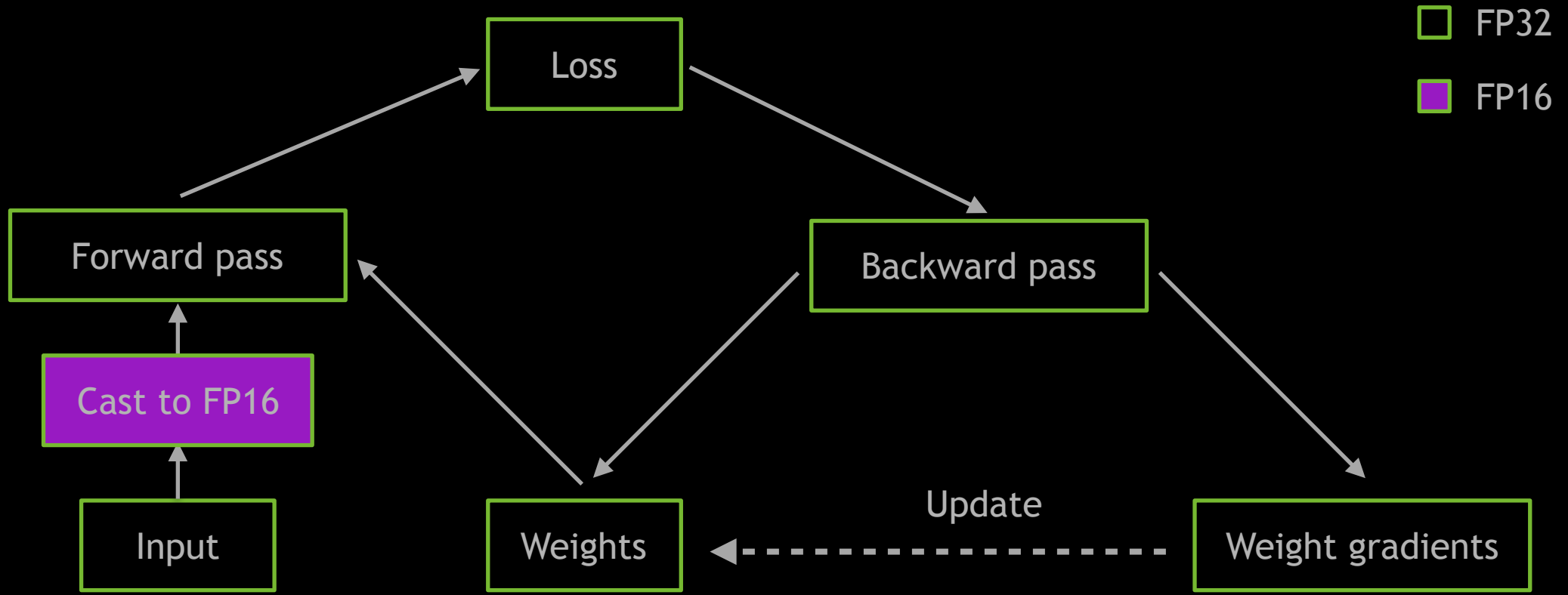
MIXED PRECISION TRAINING

Original graph



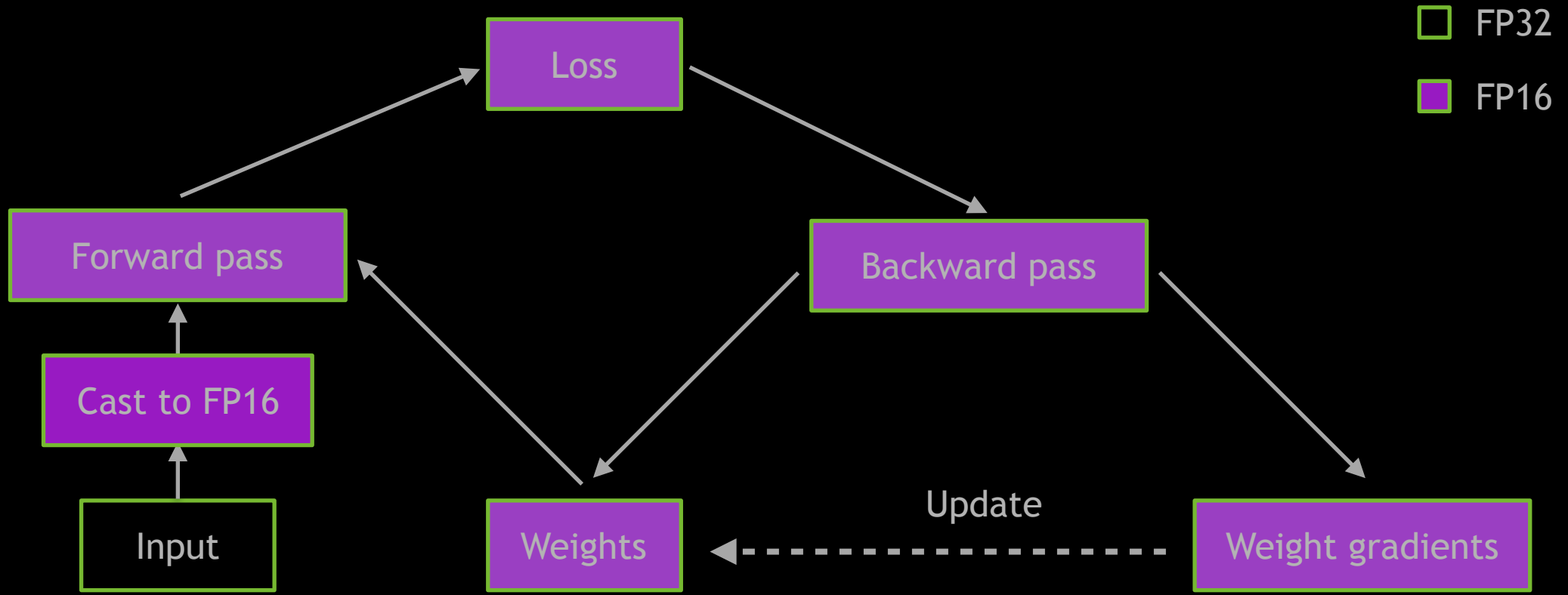
MIXED PRECISION TRAINING

Step 1: Convert to FP16



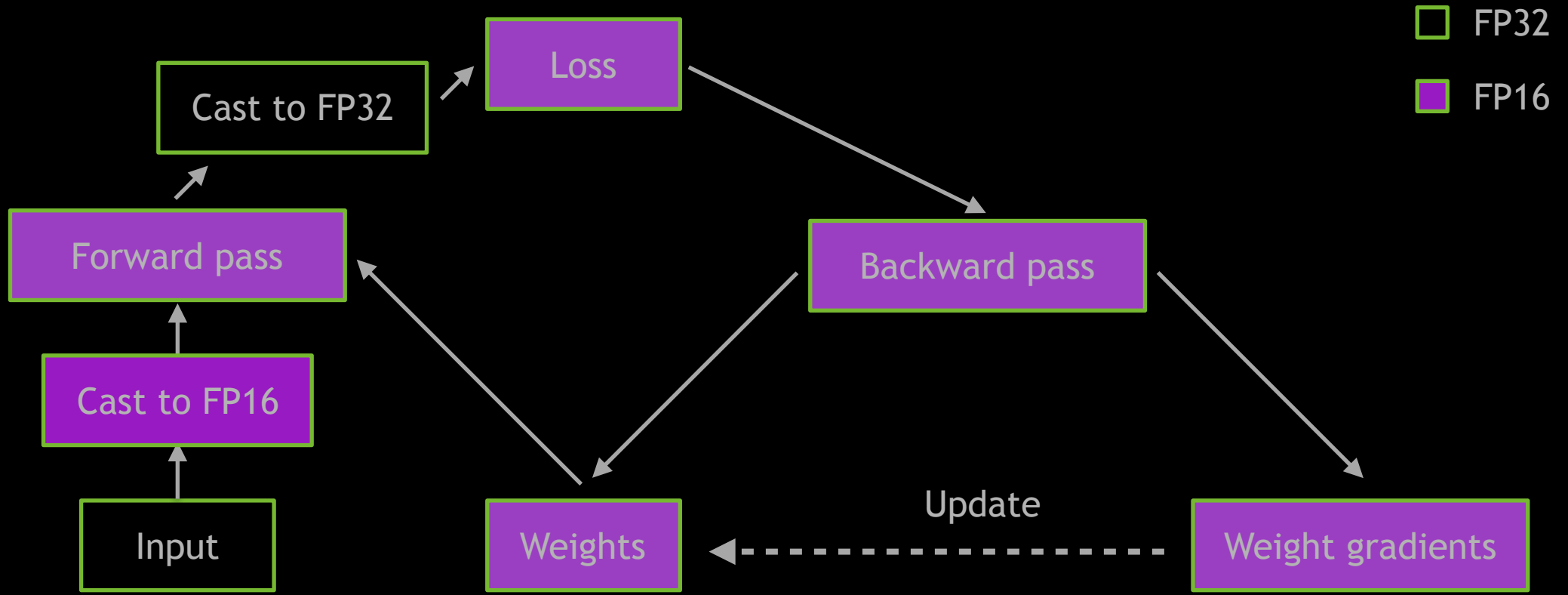
MIXED PRECISION TRAINING

New graph



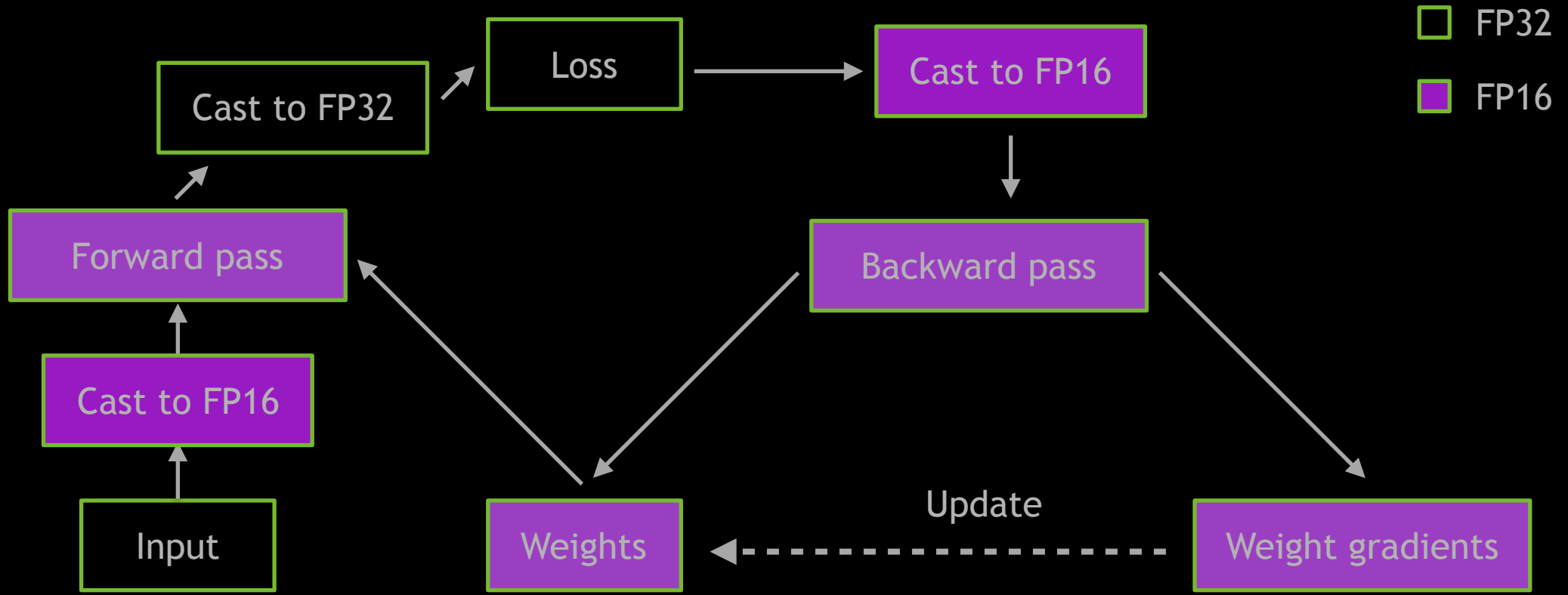
MIXED PRECISION TRAINING

Use FP32 to compute the loss



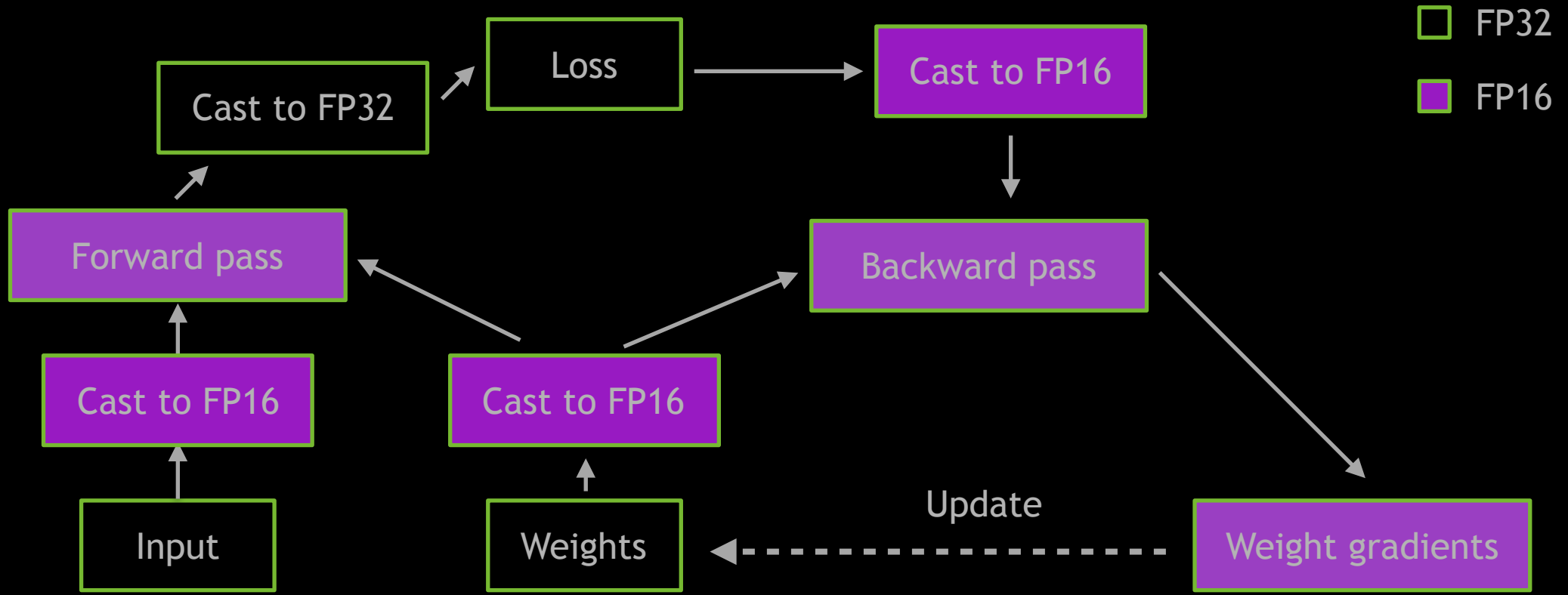
MIXED PRECISION TRAINING

New graph



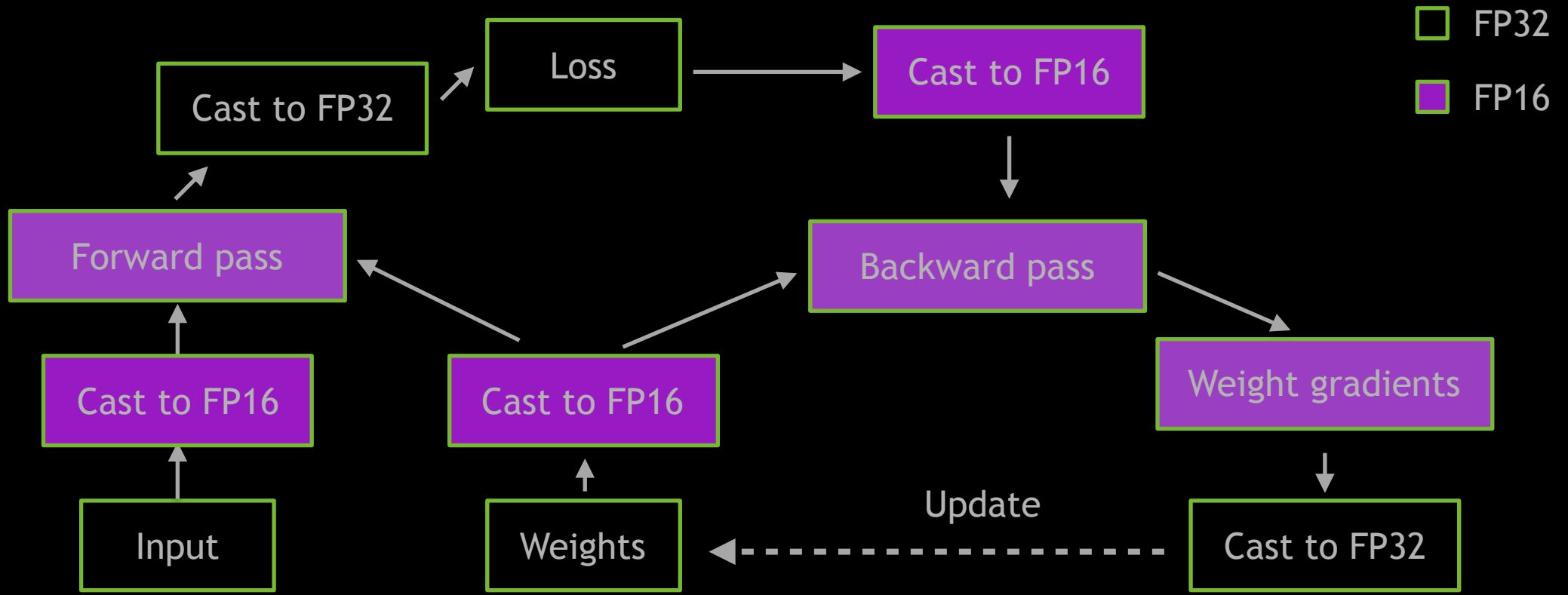
MIXED PRECISION TRAINING

Keep master weights in FP32



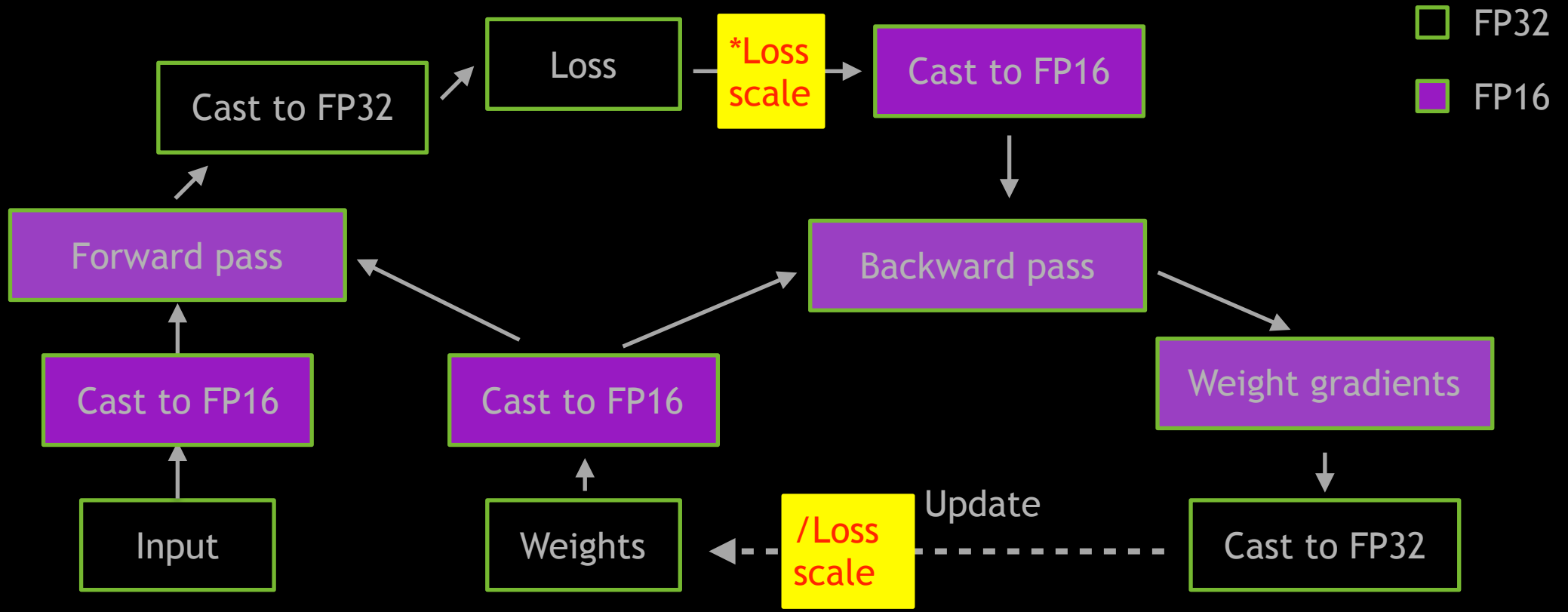
MIXED PRECISION TRAINING

Maintain precision by accumulating in FP32



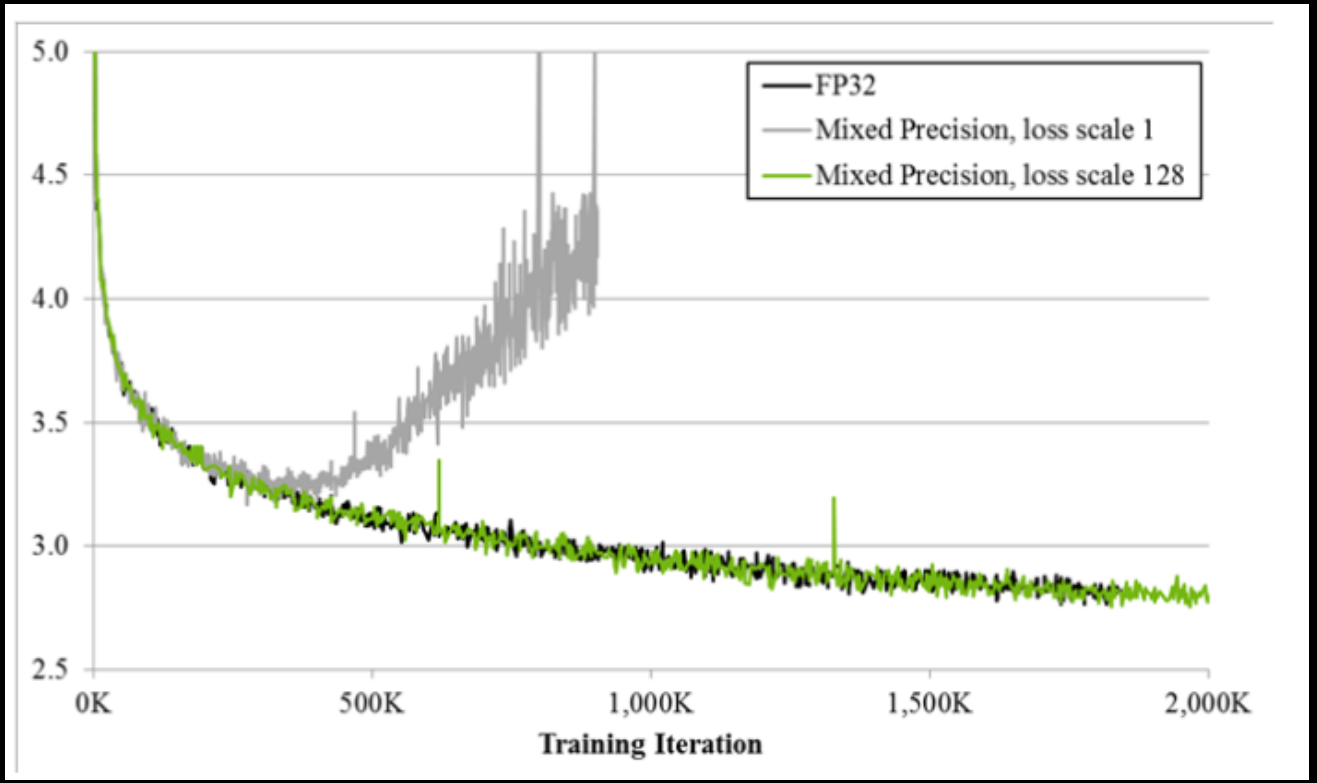
MIXED PRECISION TRAINING

Maintain precision by accumulating in FP32



MIXED PRECISION TRAINING

Loss scaling as a tool to prevent numerical instability



Lightweight operation

Loss magnitude proportional to grad
Scaling loss -> Descaling gradient

Objectives

Avoid NaN loss during training
Provide a stable gradient

MIXED PRECISION TRAINING

NVIDIA does the hard work and you get to enjoy it

TensorFlow

Enable an environment variable

```
export TF_ENABLE_AUTO_MIXED_PRECISION=1
```

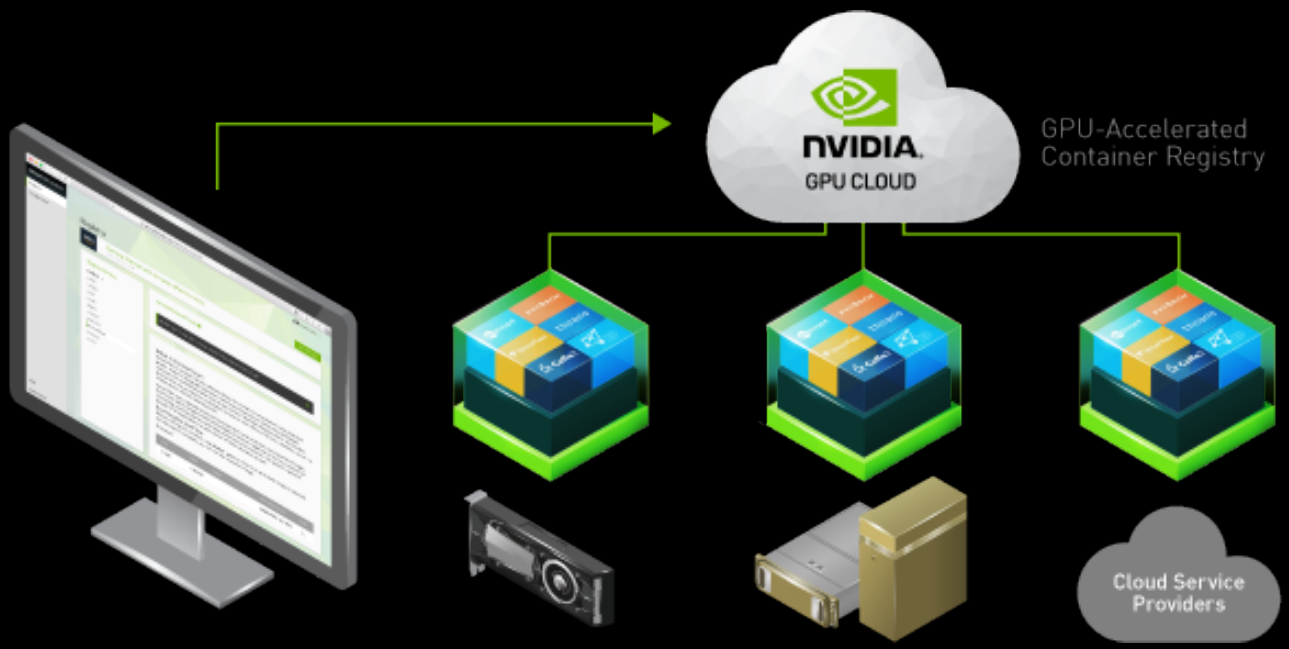
PyTorch

Use NVIDIA Apex

```
model, opt = amp.initialize(model, opt)
with amp.scale_loss(loss, opt) as scaled_loss:
  scaled_loss.backward()
```

ARTIFICIAL INTELLIGENCE CONTAINERS

Ready-to-run deep learning software



Performance

Containers provide developers
With record-setting performance

Access from anywhere

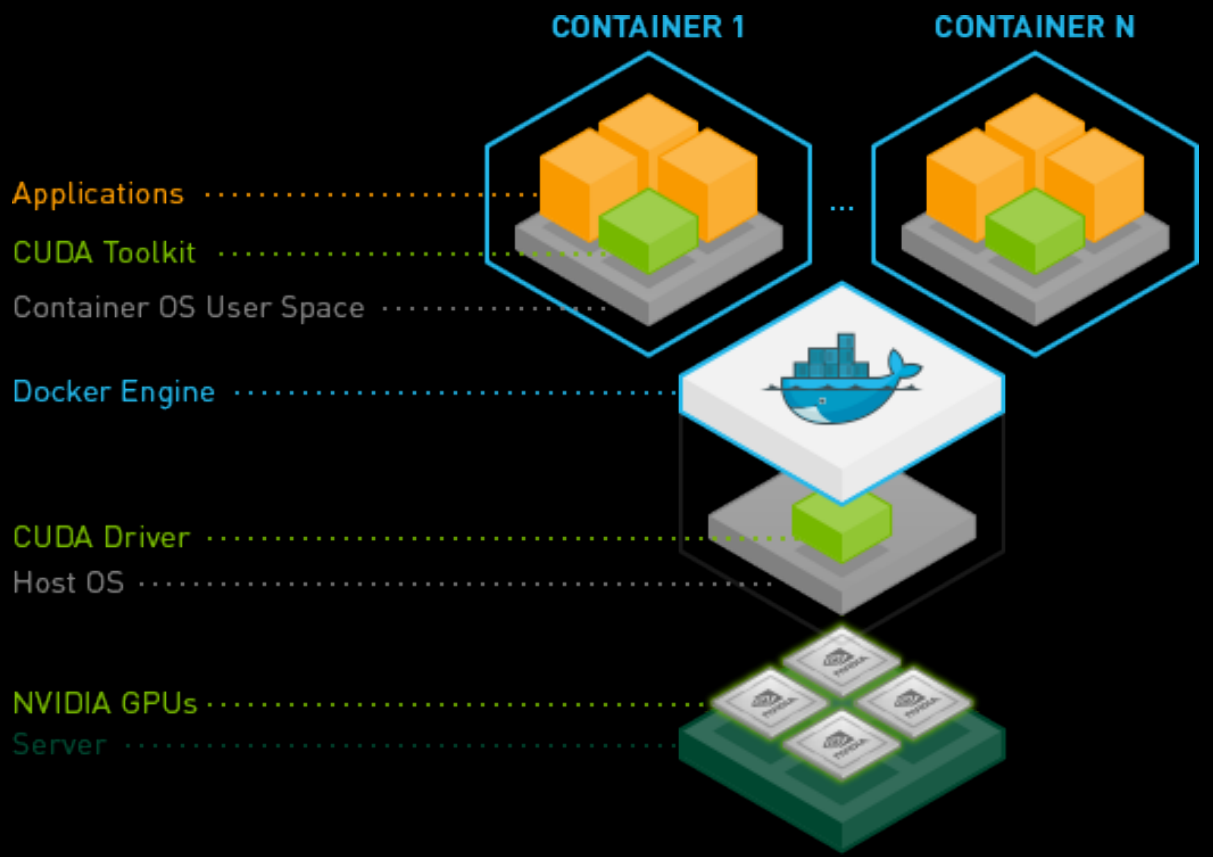
Available to anyone —at no cost—

Innovation for all

Containers ship optimized versions
Of the most popular models and
Architectures ready to use

ARTIFICIAL INTELLIGENCE CONTAINERS

Ready-to-run deep learning software



Platform independent

Enjoy deep learning on your favorite OS thanks to docker

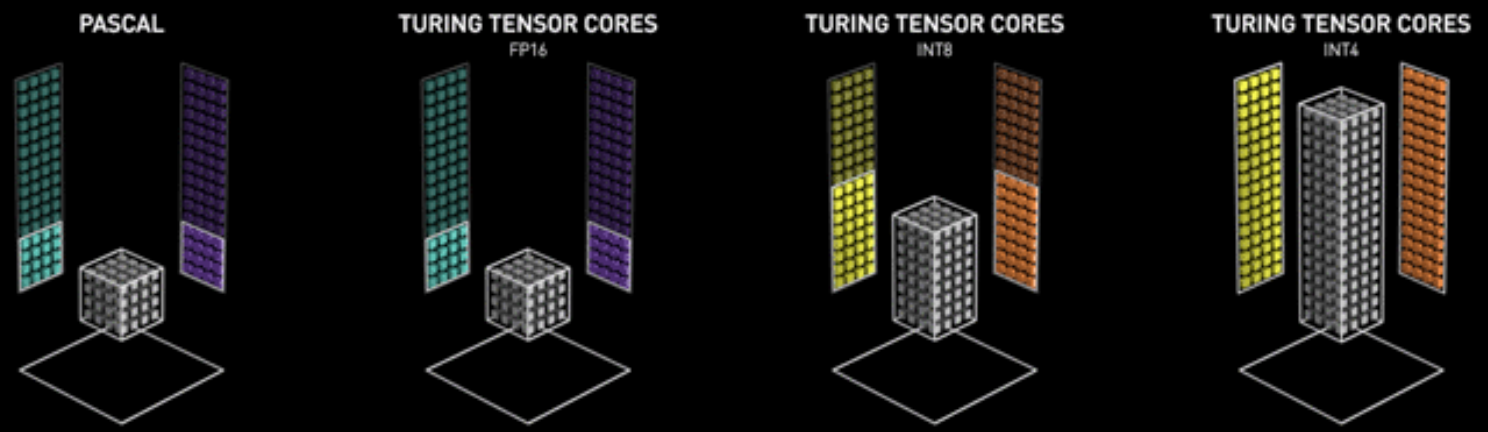
Want to know more?

Try it out, now:

https://github.com/NVIDIA/nvidia-docker

NVIDIA TENSOR CORE

Hardware-optimized deep network training



DOES THIS SOUND INTERESTING?

NVIDIA Poland

Do you want to get started in ML?

Check out our containers. They include examples and support all major frameworks.

Are you a researcher?

We sponsor research institutions doing exciting work and requiring a Tensor Core boost. We recently granted the University of Warsaw 35 Volta GPUs for research

Do you want to be part of our mission?

Just let me know! Send your cv at pribalta@nvidia.com or look for me in LinkedIn

MIXED PRECISION TRAINING

Links and further reading

- Mixed precision training (ICLR 2018): Link
- Mixed precision training of deep neural networks (NVIDIA blog): Link
- Deep learning performance guide: Link
- Mixed precision training of deep neural networks (GTC 2019): Link
- New automated mixed-precision tools for TensorFlow (GTC 2019): Link
- PyTorch Apex repository: Link

TRAINING: FAST AND SLOW

Pablo Ribalta Lorenzo, 21.5.2019