MIXED PRECISION TRAINING

Michael O'Connor

NVIDIA

MIXED PRECISION

What is the benefit?

Using **mixed precision** and **Volta** your networks can be:

1. 3-4x **faster**
2. Reduce memory consumption and bandwidth pressure
3. just as **powerful**

with **no architecture change.**

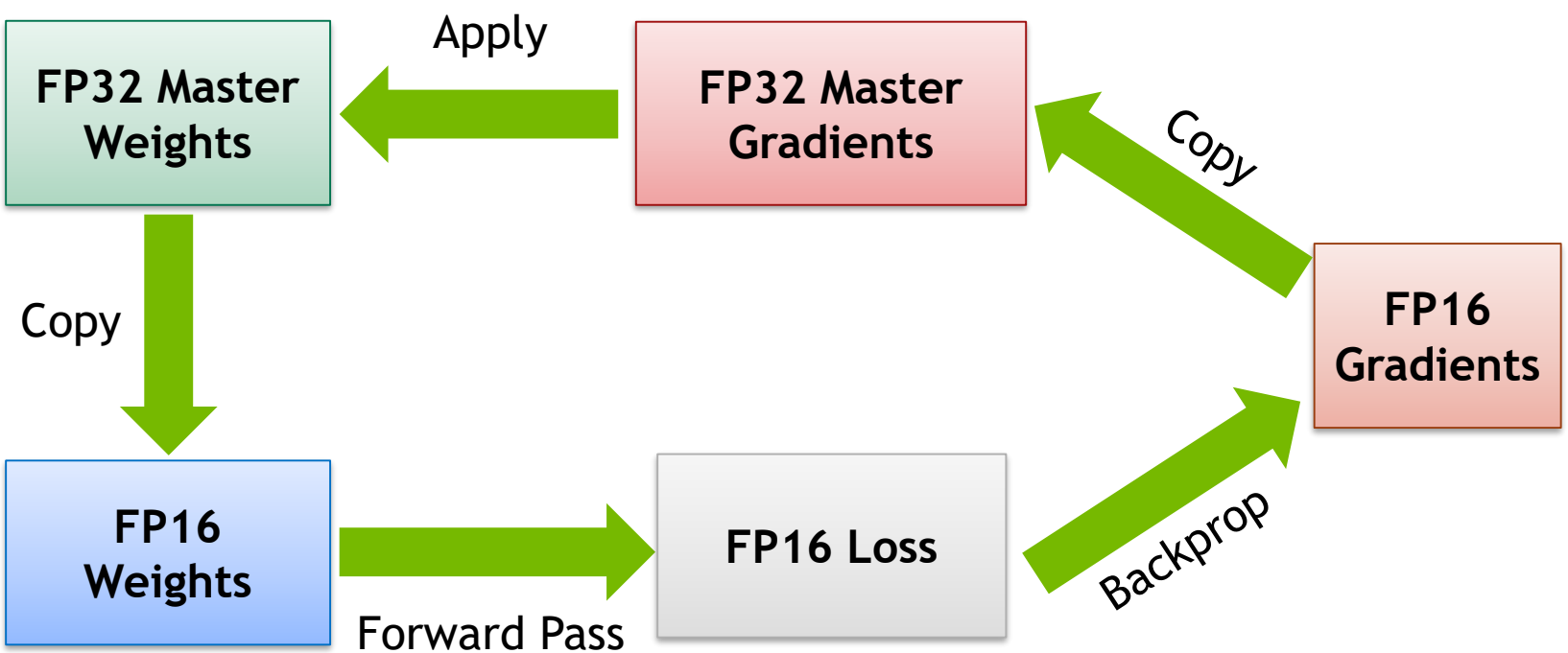
A MIXED PRECISION SOLUTION

Imprecise weight updates → "Master" weights in FP32

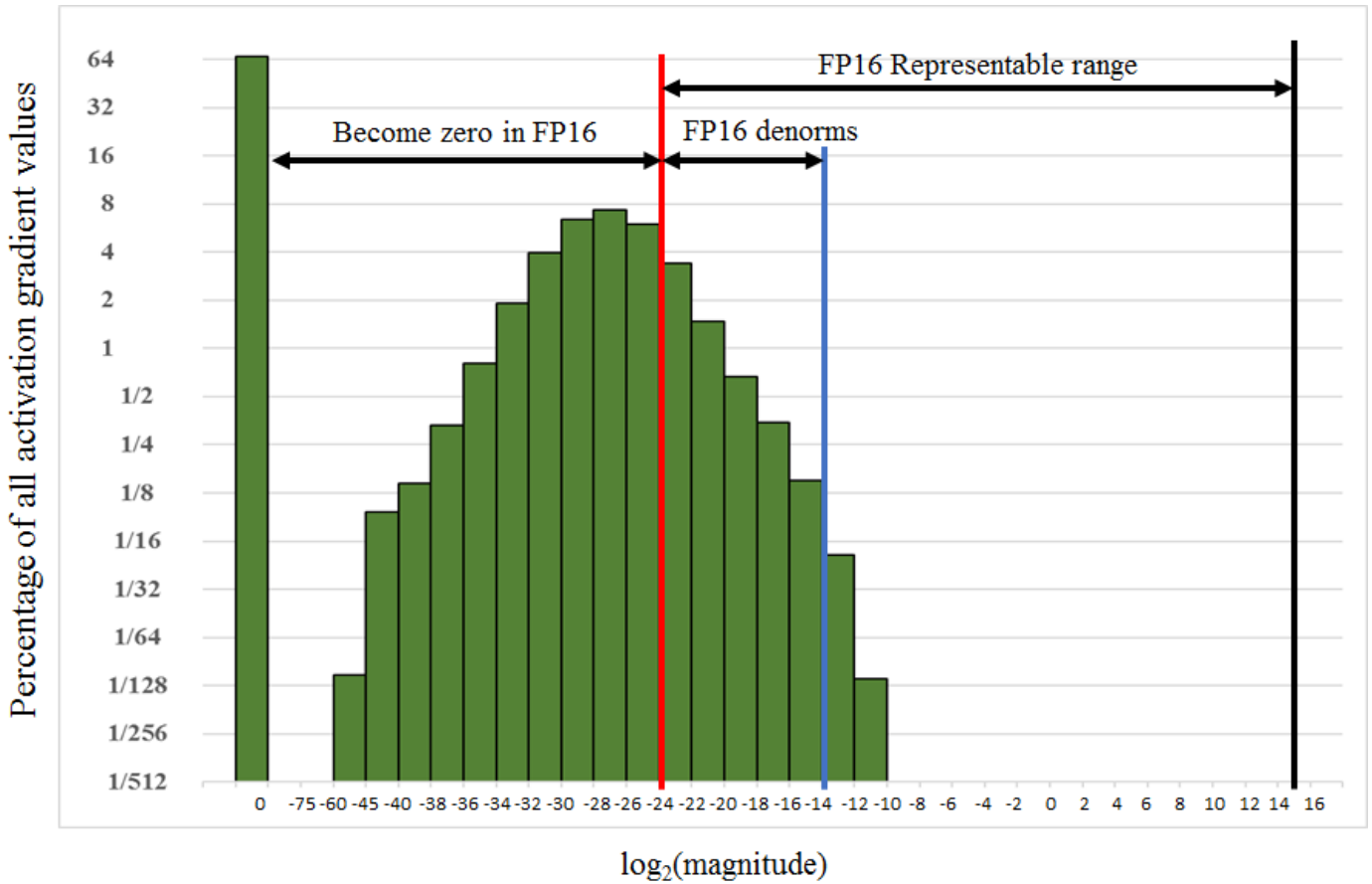
Gradients underflow → Loss (Gradient) Scaling

Maintain precision → Accumulate to FP32 (Tensor Cores)

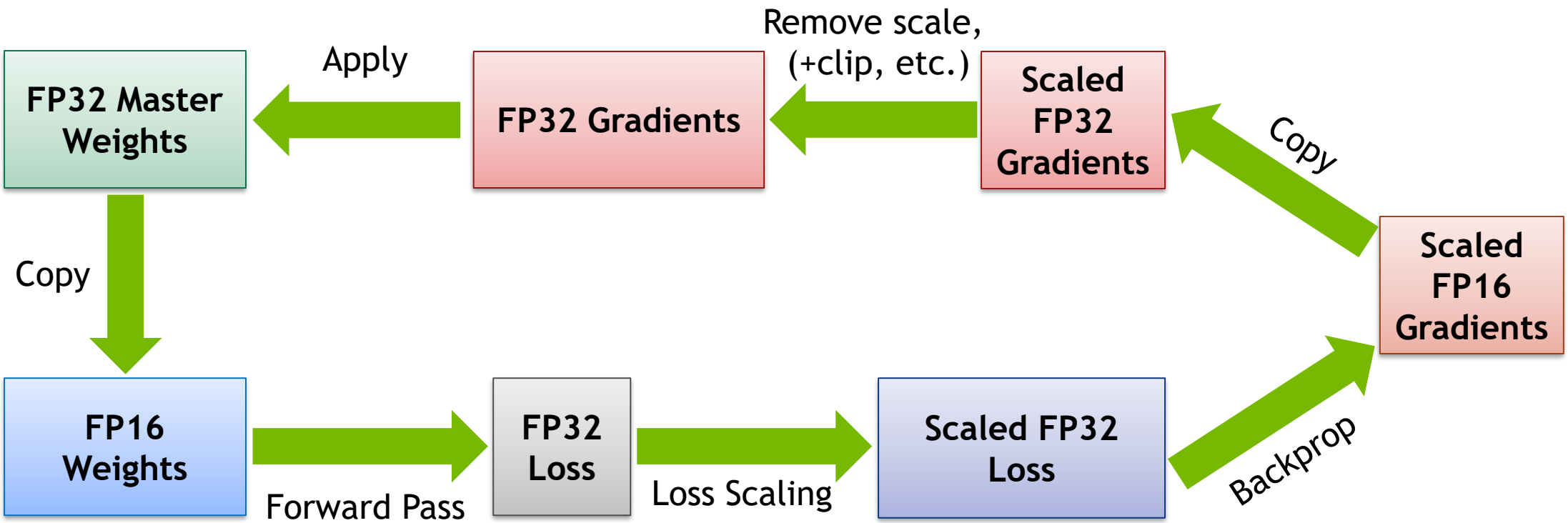
MIXED SOLUTION: FP32 MASTER WEIGHTS



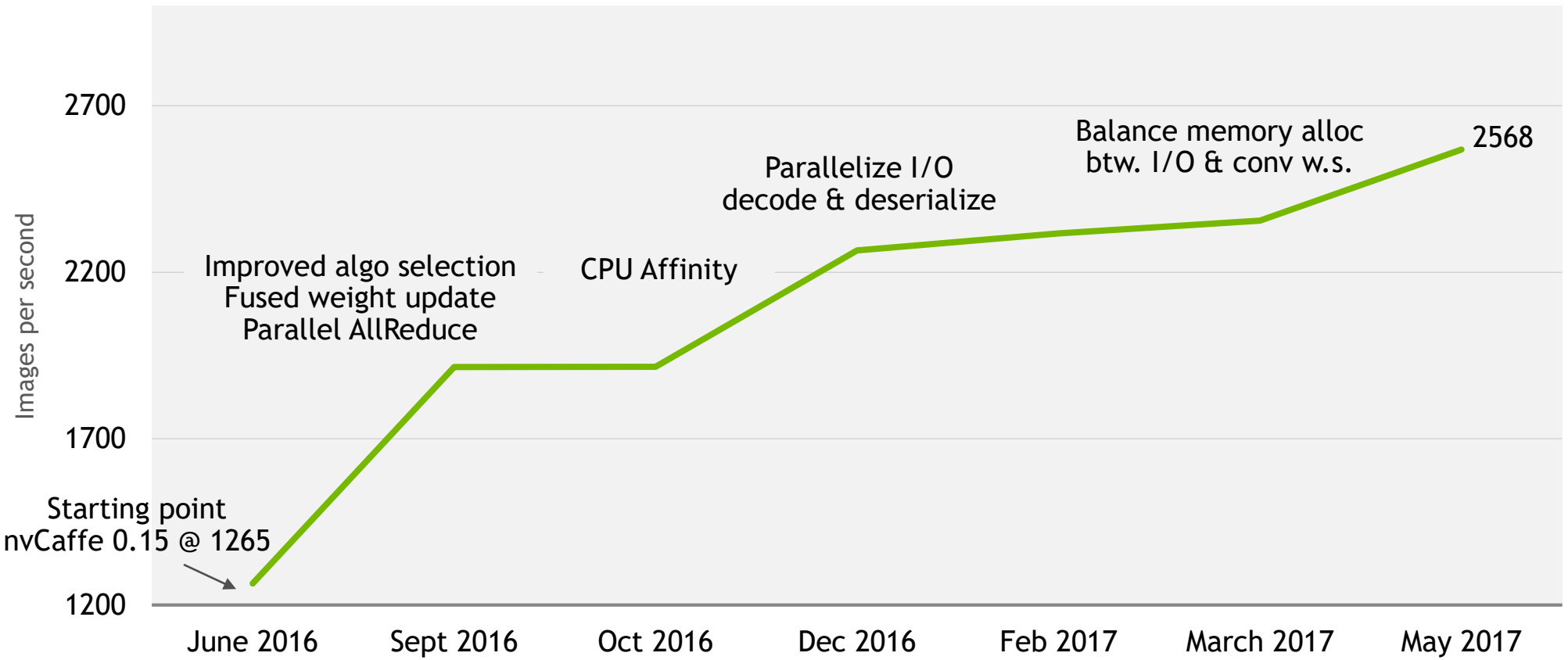
GRADIENTS RANGE OFFSET



MIXED PRECISION TRAINING

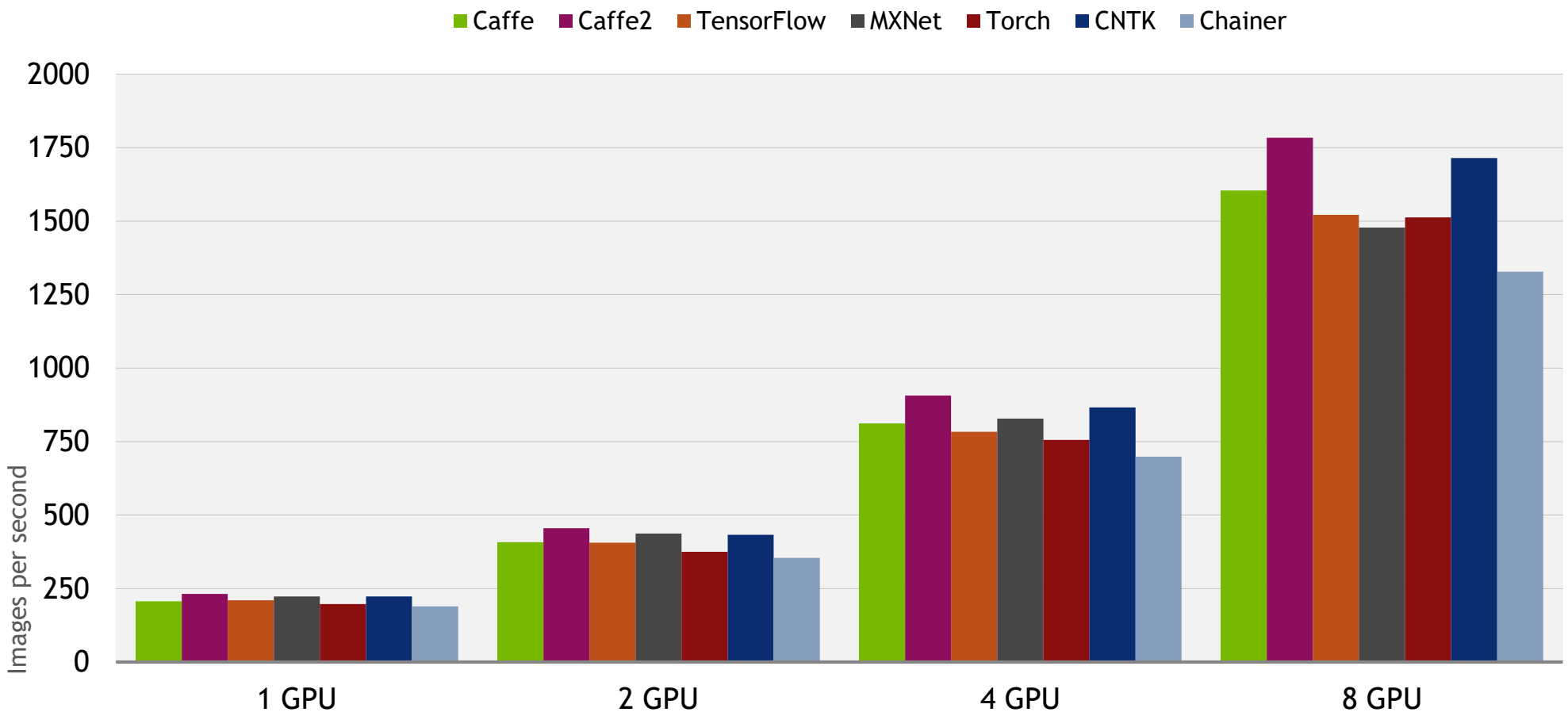


NVCAFFE V0.16 TRAINING ALEXNET



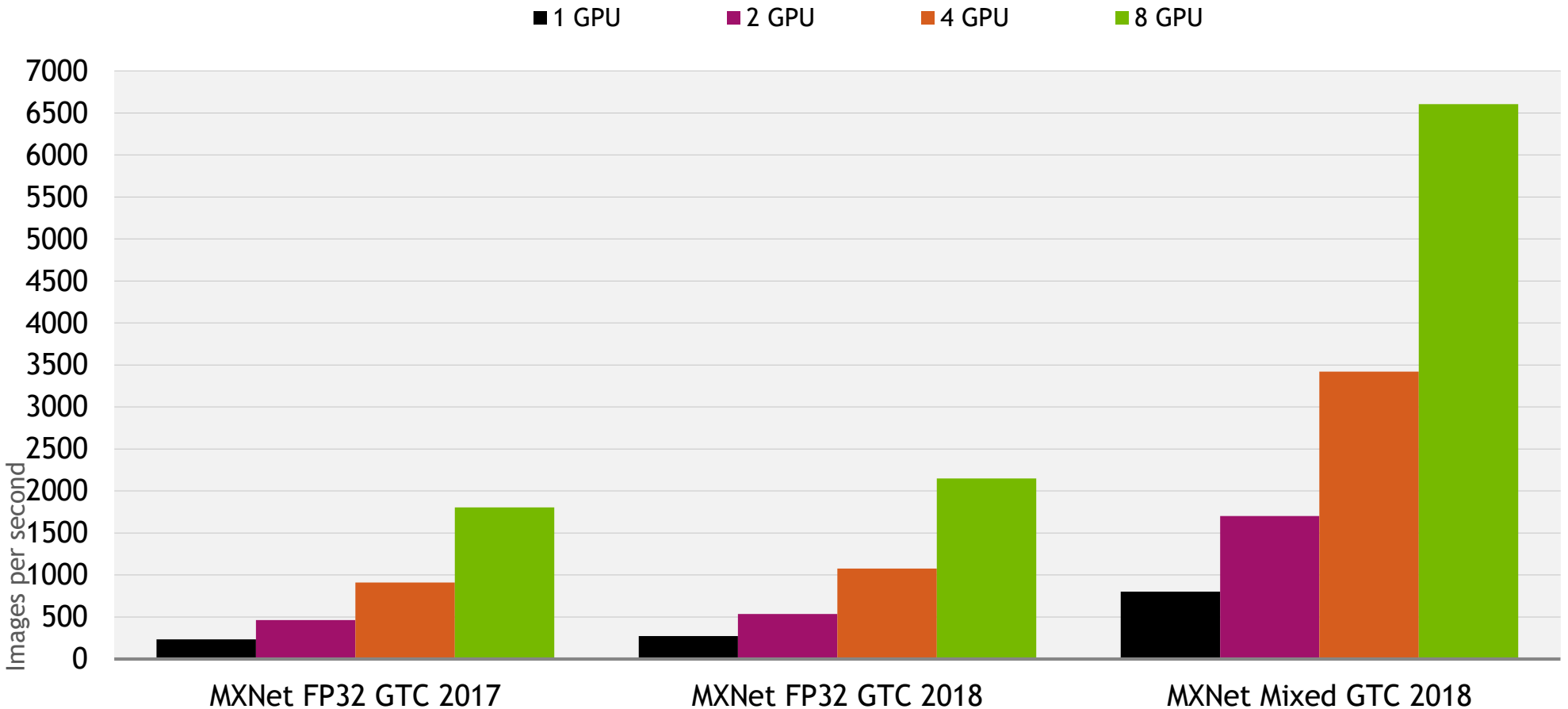
Single P100 GPU, Batch Size=128

RESNET-50 FP32 PERFORMANCE



4/30/2017 : *DGX-1 with Batch Size=64 per GPU. Chainer numbers are preliminary.*

RESNET-50 MIXED PRECISION AND FP32



INFORMATION SOURCES

Where to learn about mixed precision training

CE8130 - Connect with the Experts: Deep Learning Training for Volta Tensor Cores Tu 2PM

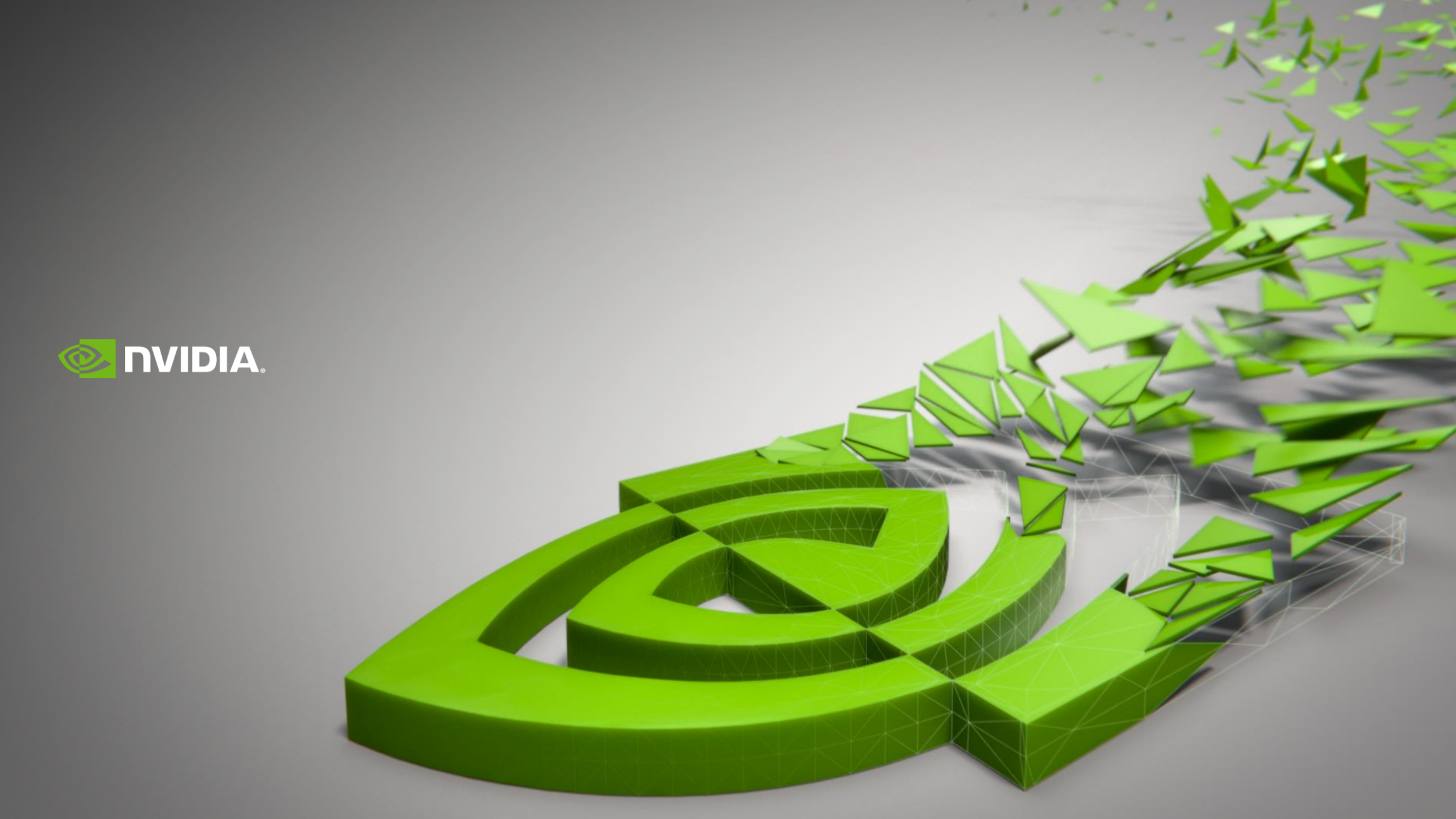
S8923 - Training Neural Networks with Mixed Precision: Theory and Practice Wed 2PM

S81012 - Training Neural Networks with Mixed Precision: Real Examples Th 9 AM

CE8162 - Connect with the Experts: Deep Learning Training for Volta Tensor Cores Th 2PM

Mixed- Precision Training of Deep Neural Networks (NVIDIA Developer Blog)

Training with Mixed Precision (NVIDIA User Guide)


NVIDIA