Towards Better DL Frameworks

Yangqing Jia

Research Lead on AI Platforms, Facebook

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik
UC Berkeley
{rbg,jdonahue,trevor,malik}@eecs.berkeley.edu

Abstract

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012—achieving a mAP of 53.3%. Our approach combines two key insights: (1) one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. Since we combine region proposals with CNNs, we call our method R-CNN: *Regions with CNN features. Source code for the complete system is available at* http://www.cs.berkeley.edu/~rbg/rcnn.

1. Introduction

Features matter. The last decade of progress on various visual recognition tasks has been based considerably on the use of SIFT [27] and HOG [7]. But if we look at performance on the canonical visual recognition task, PASCAL VOC object detection [13], it is generally acknowledged that progress has been slow during 2010-2012, with small gains obtained by building ensemble systems and employing minor variants of successful methods.

SIFT and HOG are blockwise orientation histograms, a representation we could associate roughly with complex cells in V1, the first cortical area in the primate visual pathway. But we also know that recognition occurs several stages downstream, which suggests that there might be hierarchical, multi-stage processes for computing features that are even more informative for visual recognition.

Fukushima's "neocognitron" [17], a biologically-inspired hierarchical and shift-invariant model for pattern recognition, was an early attempt at just such a process. The neocognitron, however, lacked a supervised training algorithm. Building on Rumelhart *et al.* [30], LeCun *et al.* [24] showed that stochastic gradient descent via backpropagation was effective for training convolutional neural networks (CNNs), a class of models that extend the neocognitron.

R-CNN: *Regions with CNN features*

warped region
aeroplane? no.
person? yes.
tvmonitor? no.
CNN
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

Figure 1: Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean average precision (mAP) of **53.7% on PASCAL VOC 2010**. For comparison, [34] reports 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part models perform at 33.4%.

CNNs saw heavy use in the 1990s (*e.g.*, [25]), but then fell out of fashion with the rise of support vector machines. In 2012, Krizhevsky *et al.* [23] rekindled interest in CNNs by showing substantially higher image classification accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [9, 10]. Their success resulted from training a large CNN on 1.2 million labeled images, together with a few twists on LeCun's CNN (*e.g.*, $\max(x, 0)$ rectifying non-linearities and "dropout" regularization).

The significance of the ImageNet result was vigorously debated during the ILSVRC 2012 workshop. The central issue can be distilled to the following: To what extent do the CNN classification results on ImageNet generalize to object detection results on the PASCAL VOC Challenge?

We answer this question by bridging the gap between image classification and object detection. This paper is the first to show that a CNN can lead to dramatically higher object detection performance on PASCAL VOC as compared to systems based on simpler HOG-like features. To achieve this result, we focused on two problems: localizing objects

1

Source: XKCD, [Girshick et al. CVPR 2014]

The Needs

Two sides of the same coin

- Researchers: "I will need to reproduce the ResNet paper."
- Companies: "I need to apply DL to drive cars."

Democratizing Deep Learning w/ Caffe

Getting AlexNet running in 10 mins

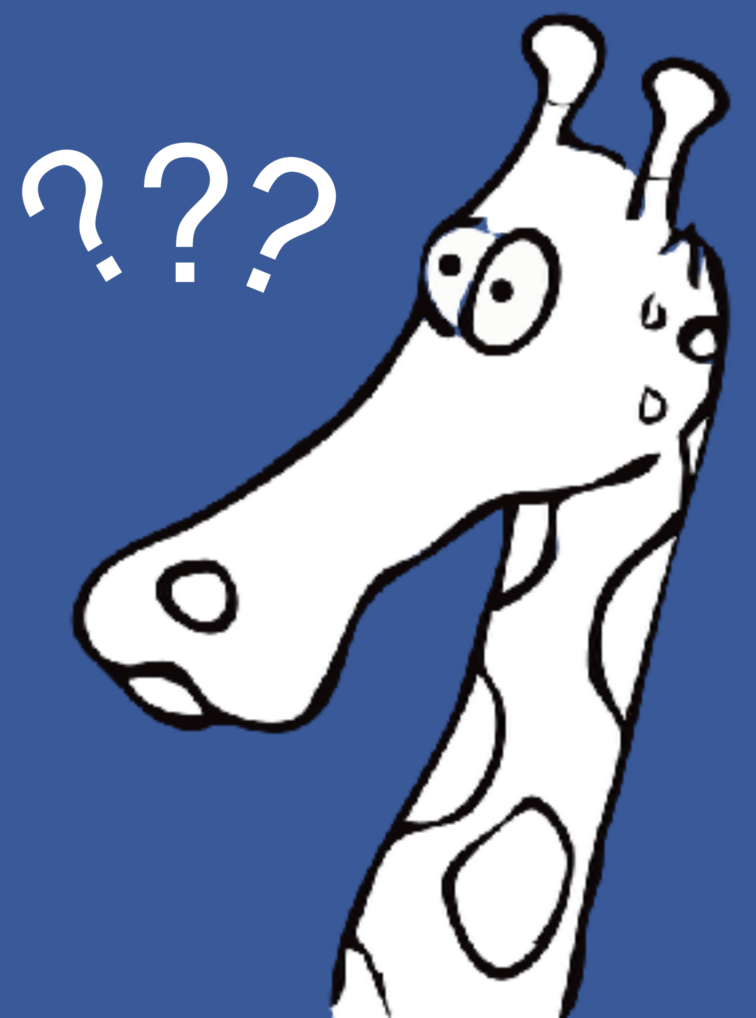
- A grad student driven project
- Started by doing one job really well: image classification
- Adopted by industry participants
- Popular deep learning framework run by a non-profit.

http://caffe.berkeleyvision.org/

Maximally accurate | Maximally specific

cat	1.79559
feline	1.74239
domestic cat	1.71551
tabby	0.95449
domestic animal	0.77145

What makes a better DL library?



"MAPS"

"MAPS"

-

Scalability

Scalability

Run fast, run far

"How do I train on
multiple GPUs and machines?"

- Probably the most question we got from Caffe users

Scalability

Run fast, run far

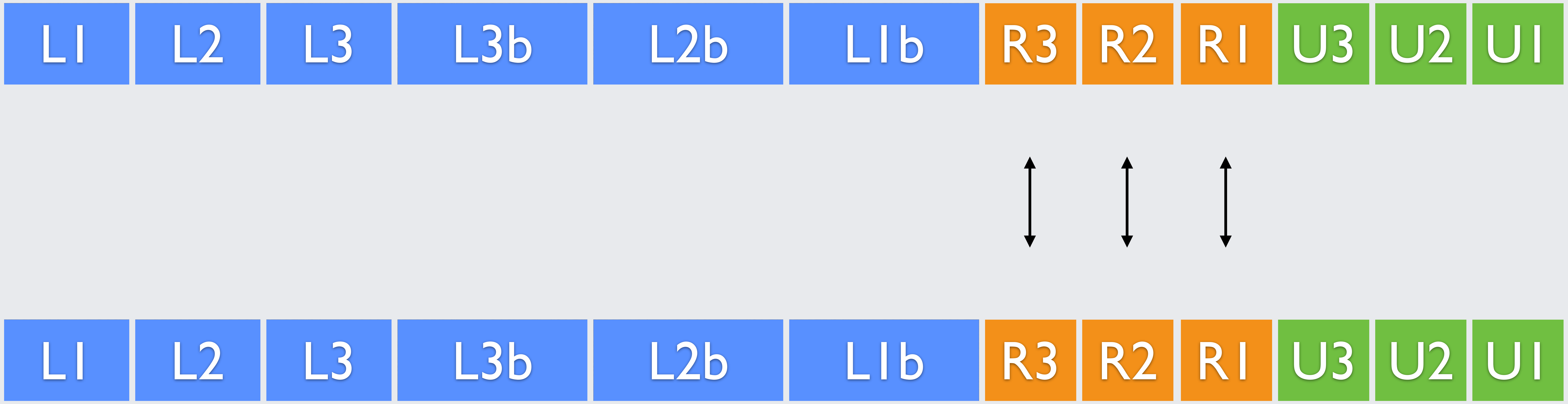
Scalability

Run fast, run far

L1	L2	L3	L3b	L2b	L1b	R3	R2	R1	U3	U2	U1

Scalability

Run fast, run far



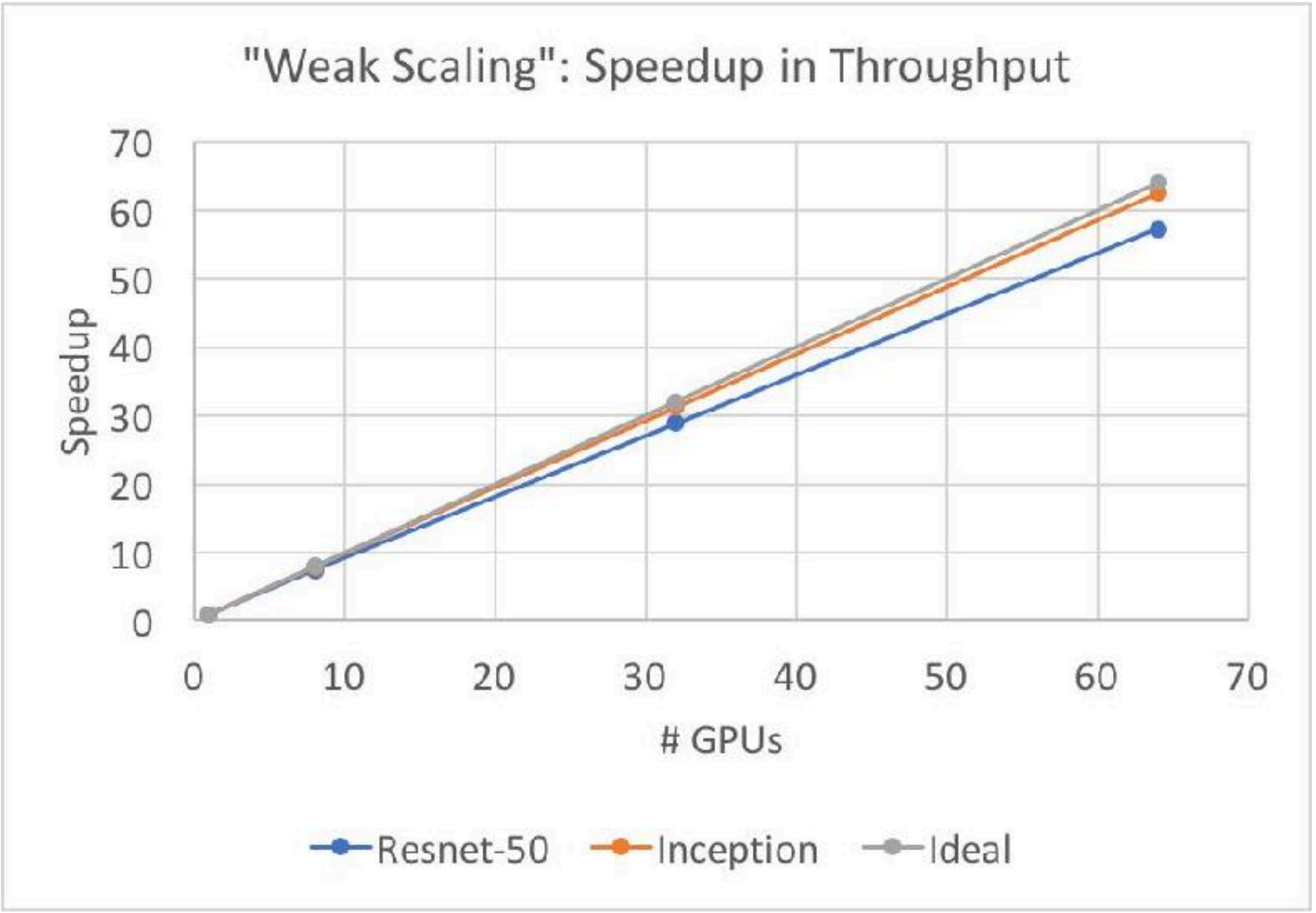
Scalability

Run fast, run far

The Return of MPI

"I'm your father", said Allreduce.

Allreduce
Tree based - O(MlogN)
Ring based - O(M)
etc.

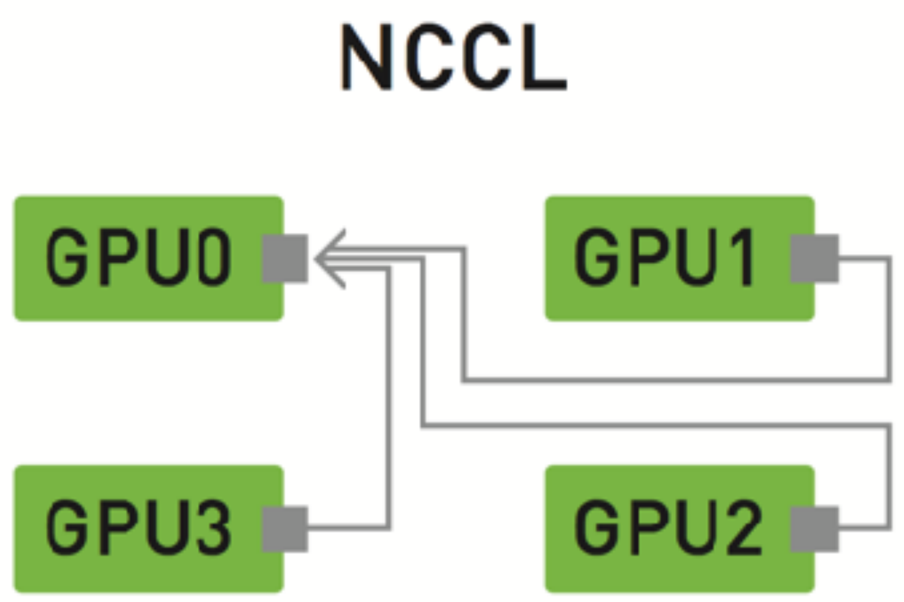


Scalability

Sitting on top of giants

OPEN MPI

MVAPICH



ØMQ

... and many more

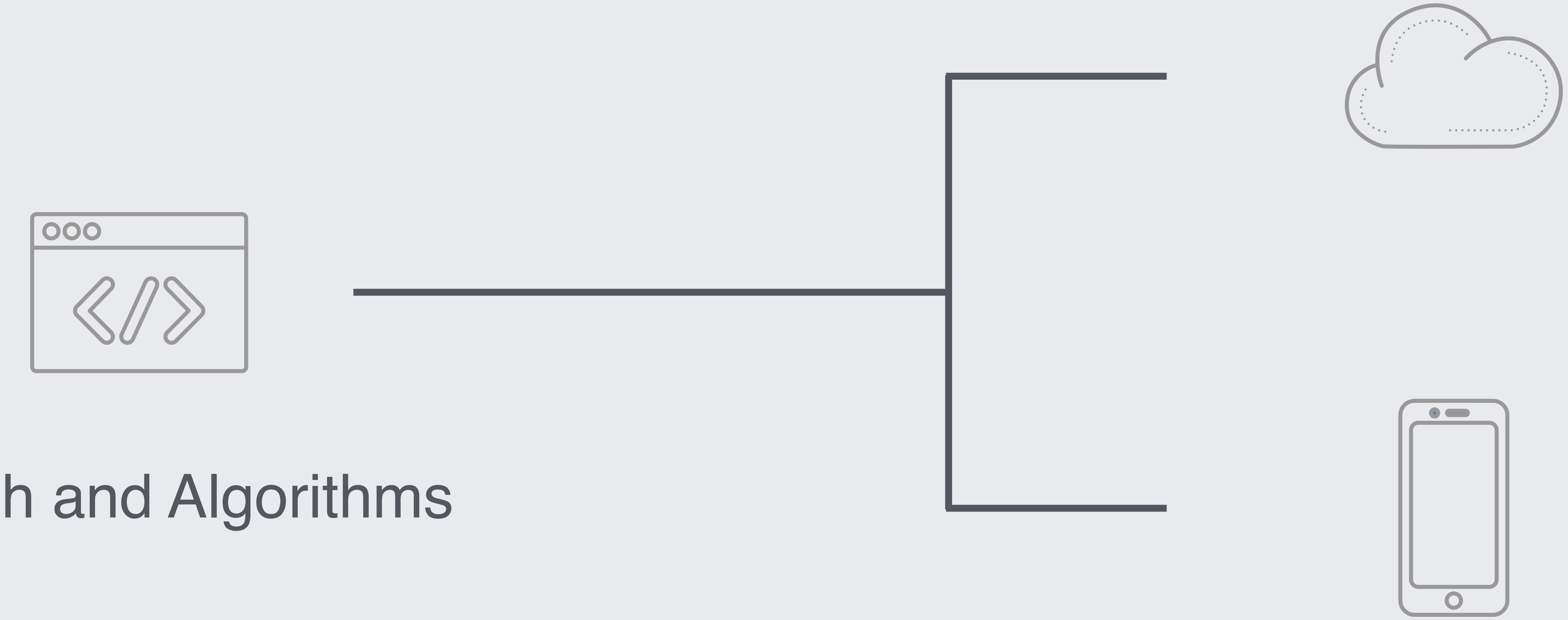
"MAPS"

-

Portability

Portable System

Cloud, Mobile, IoT, Cars, Drones, Coffee makers

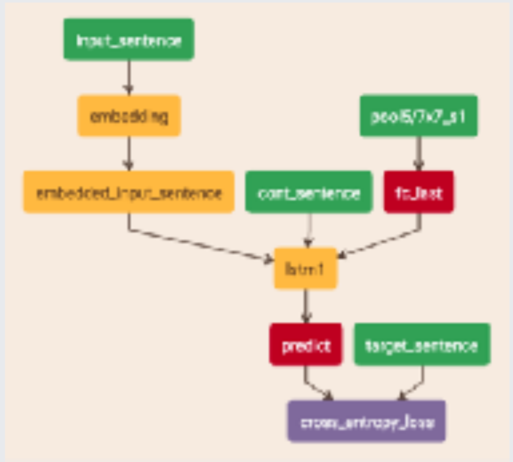


AI Math and Algorithms

Deployment Platforms

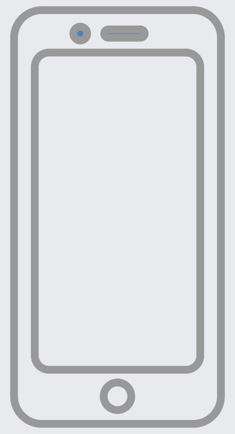
Portable System

Cloud, Mobile, IoT, Cars, Drones, Coffee makers



Model

```
auto predictor =
    caffe2::Predictor(model_file)
```



```
public class Predictor implements
    Caffe2ModelInterface;
```

Portable System Challenges

Still, a lot of thoughts needed

- Limited computation
- Battery life is a thing
- Our models may be luxurious
- Ecosystem less developed

"MAPS"

-

Augmented Comp Patterns

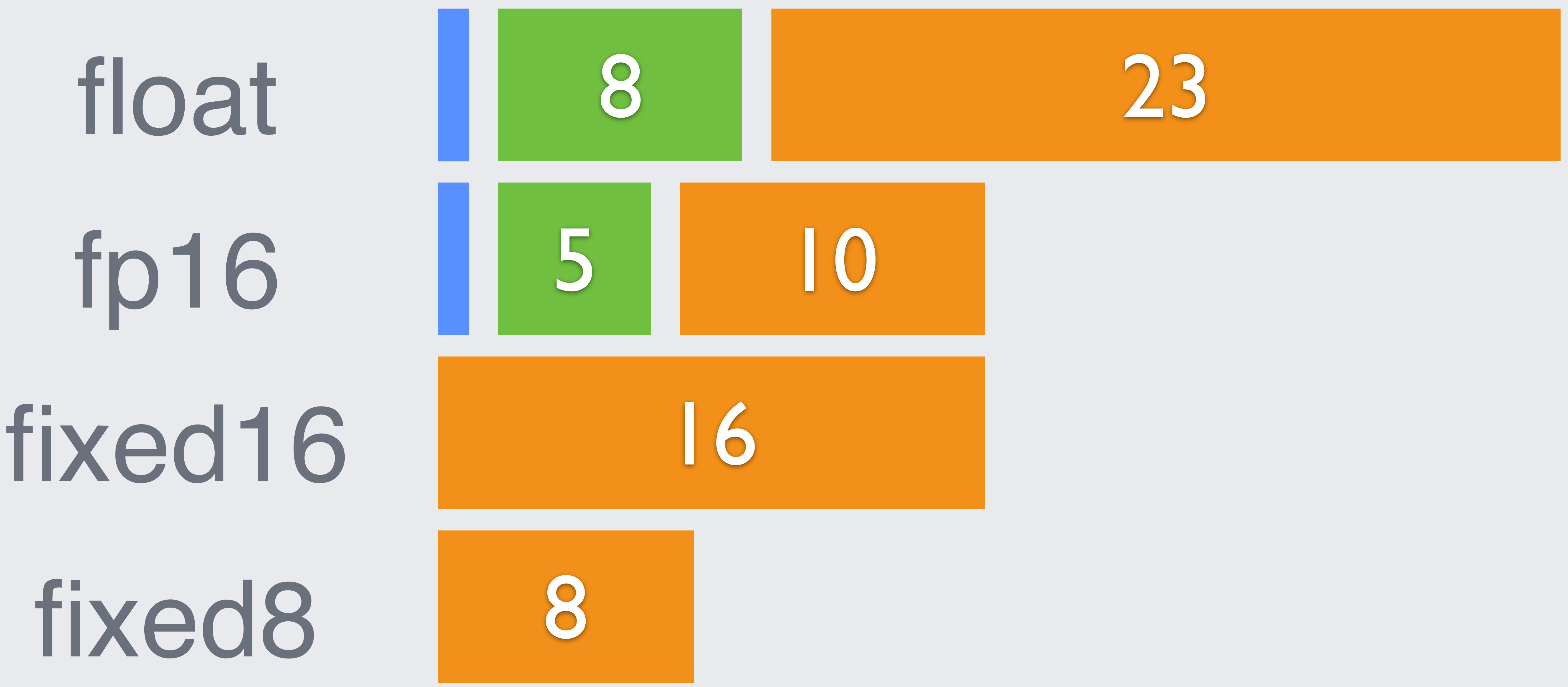
Augmented Comp Patterns

Forget about float dense math, the world is bigger

- Quantized Computation
- Sparse Math Libraries
- Model Compression
- Rethinking Existing Operations

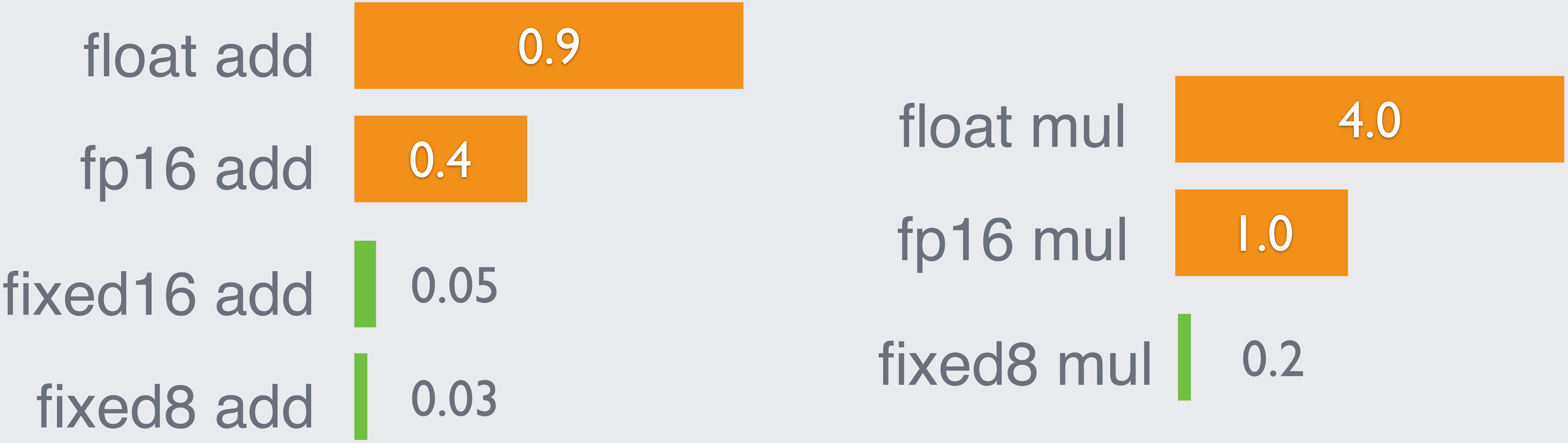
Quantized Computation

Forget about float, the world is bigger

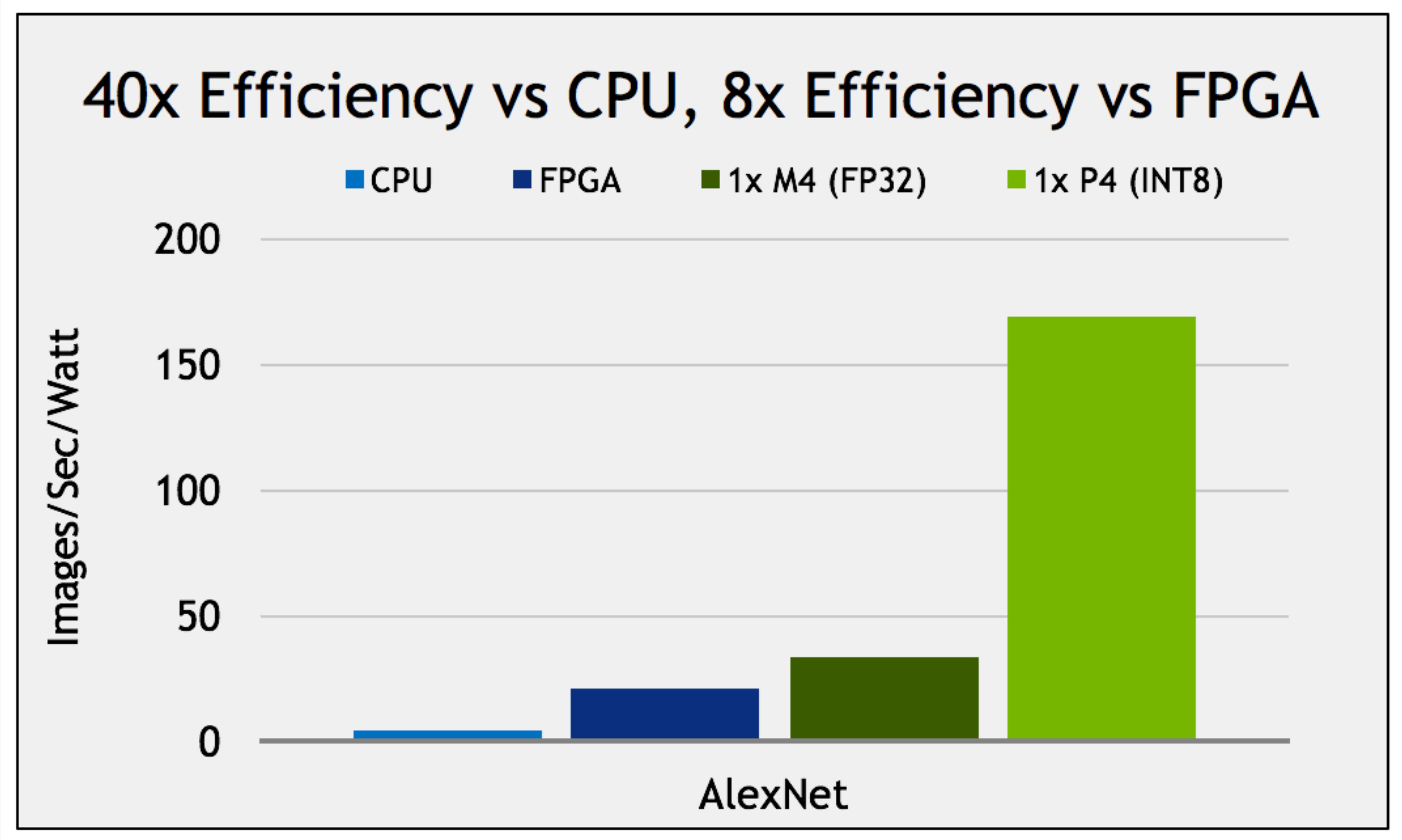


Quantized Computation

Forget about float, the world is bigger



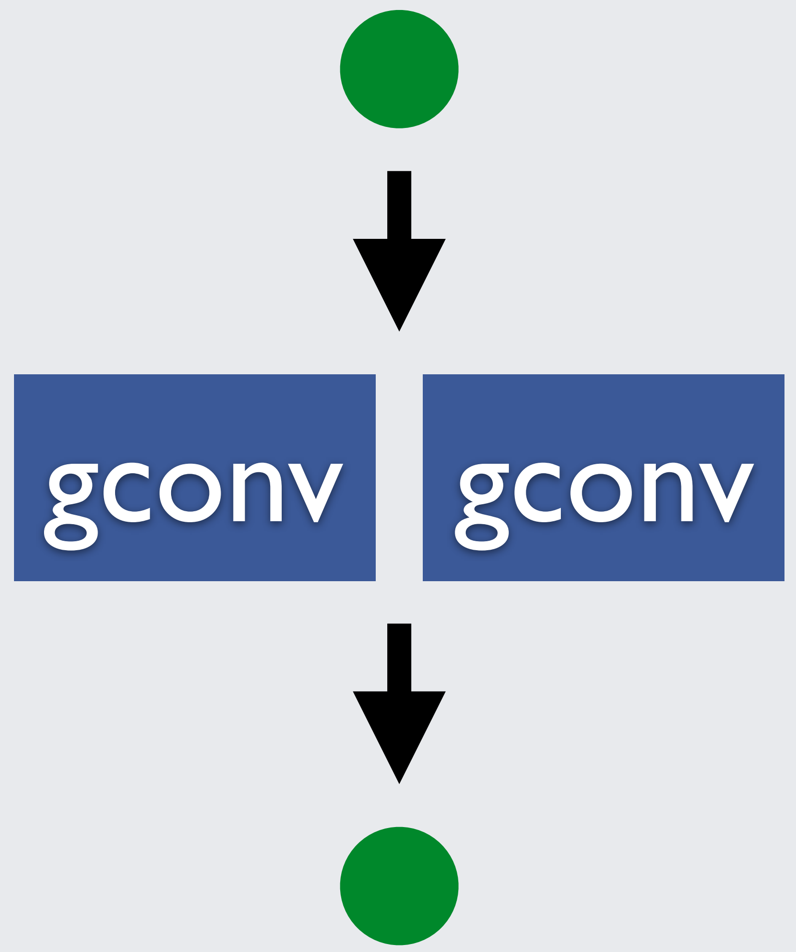
Why?



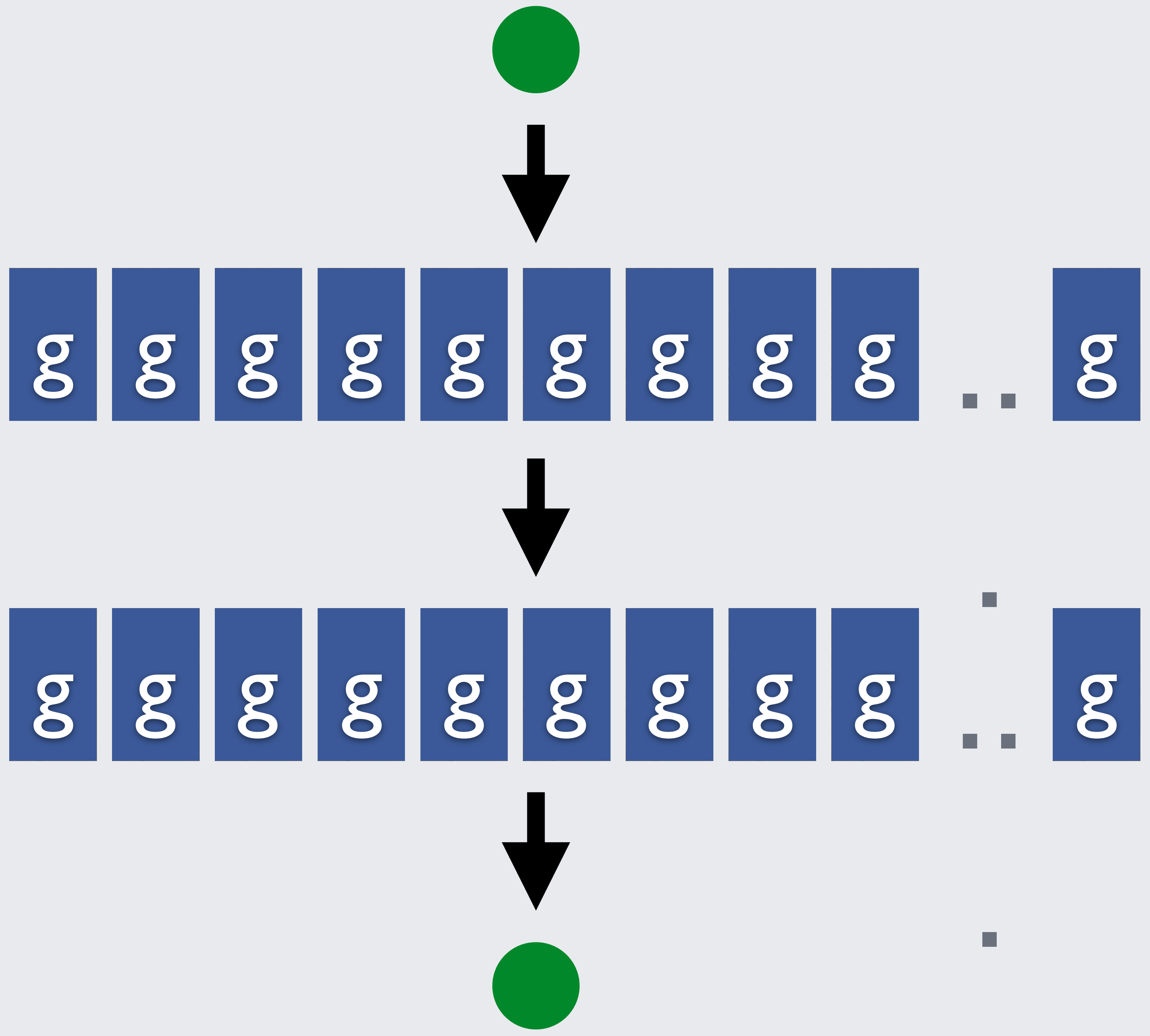
Source: Nvidia https://devblogs.nvidia.com/parallelforall/mixed-precision-programming-cuda-8/

Rethinking Existing Operations

ResNEXT is coming to town



AlexNet Group Conv



ResNext

Augmented Math Challenges

Forget about float, the world is bigger

- Solutions
  - Eigen fp16
  - CuDNN
  - NNPack
  - gemmlowp
- Challenges
  - Seamless conversion?
  - Model training?
  - Performance tuning?
  - ...

"MAPS"

-

Modularity

A Repeated Pattern

Many key components in deep learning
are
reusable
across frameworks.

In 2013 it used to be...

Caffe	Torch	Theano	...

Unix Philosophy?

or, "UnFramework"

Applications

Caffe, Torch, TF, MXNet, etc...

DataBases
LevelDB
RocksDB
Hadoop
Amazon S3
your old disk

Core Math
Eigen
CuDNN
NNPack
THNN
MKL

Comms
NCCL
MPI
ZeroMQ
Redis
...

Low Level
CUDA
OpenGL
OpenCL
Vulkan
...

Compilers

MAPS for a good framework

Modular Designs	Augmented Mathematics	Portable System	Scalability
Interface to Existing Toolkits	Optimized Math Libraries	Efficient Mobile Runtimes	Tuned Collective Primitives

+

Flexible Framework Design

No Silver Bullet?

There is no silver bullet

D4J etc. | TensorFlow | Theano

Caffe | Torch

Industry:
Stability
Scale & speed
Data Integration
Relatively Fixed

Research:
Flexible
Fast Iteration
Debuggable
Relatively bare-

There is no silver bullet

Industry:

Stability

Scale & speed

Data Integration

Relatively Fixed

Research:

Flexible

Fast Iteration

Debuggable

Relatively bare-

"In open source, we feel strongly that
to really do something well,
you have to get a lot of people involved."

— Linus Torvalds

Thank you!

Towards Better Deep Learning Frameworks

Yangqing Jia, Research Lead on AI Platforms, Facebook