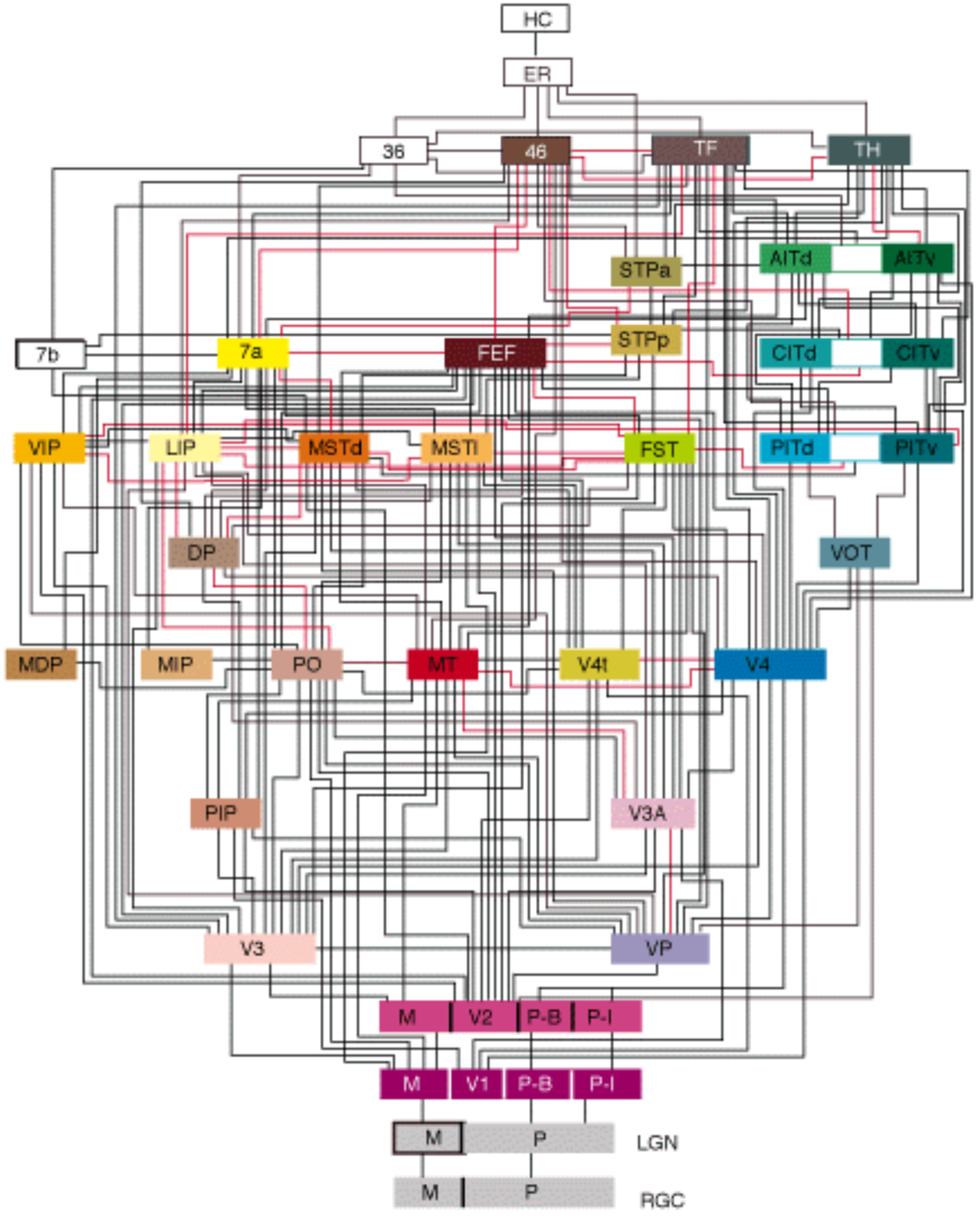
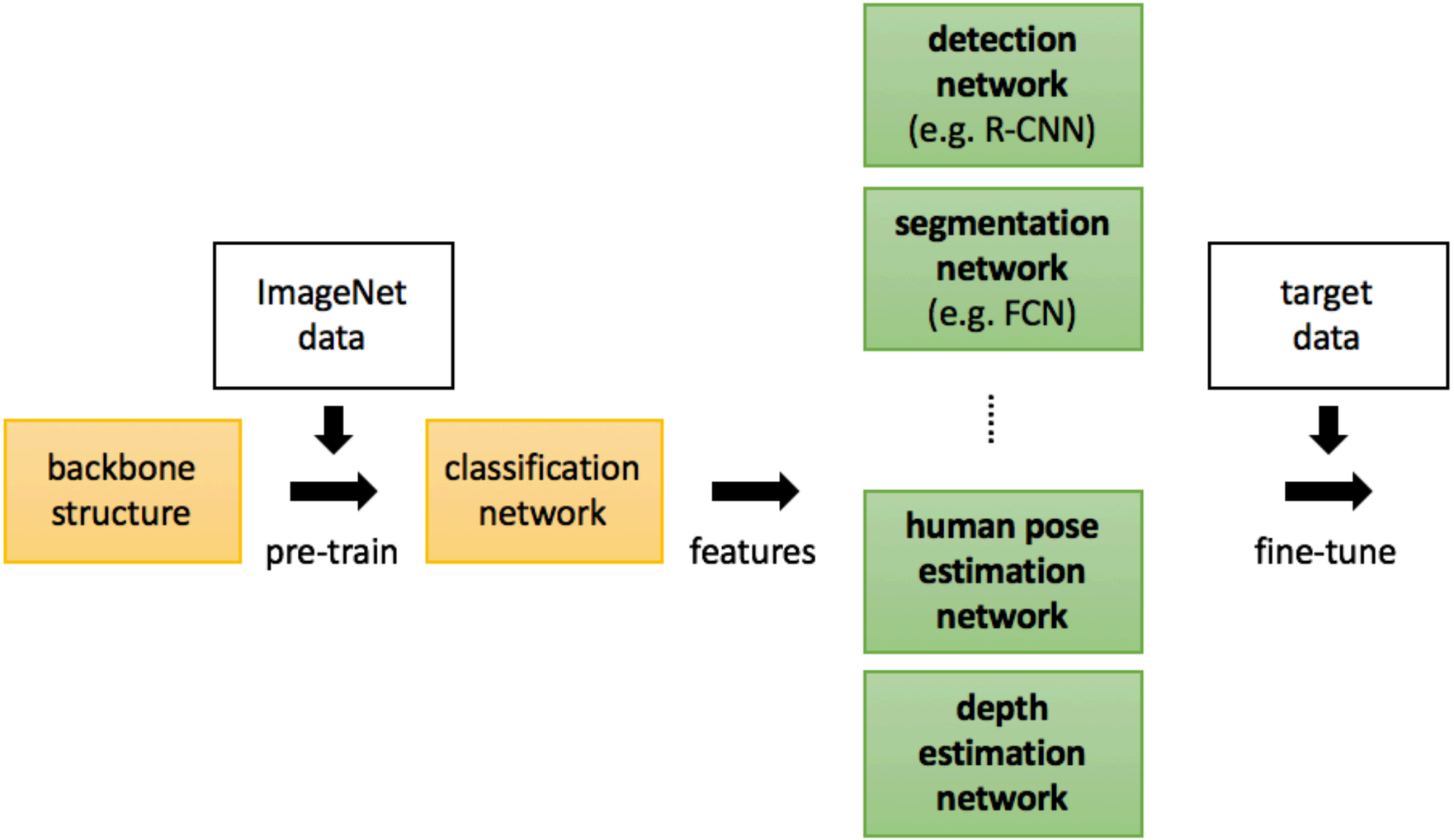
Recent Trends in Computer Vision and Deep Learning Systems

Yangqing Jia

Lead Researcher and Manager of AI Platform, Facebook

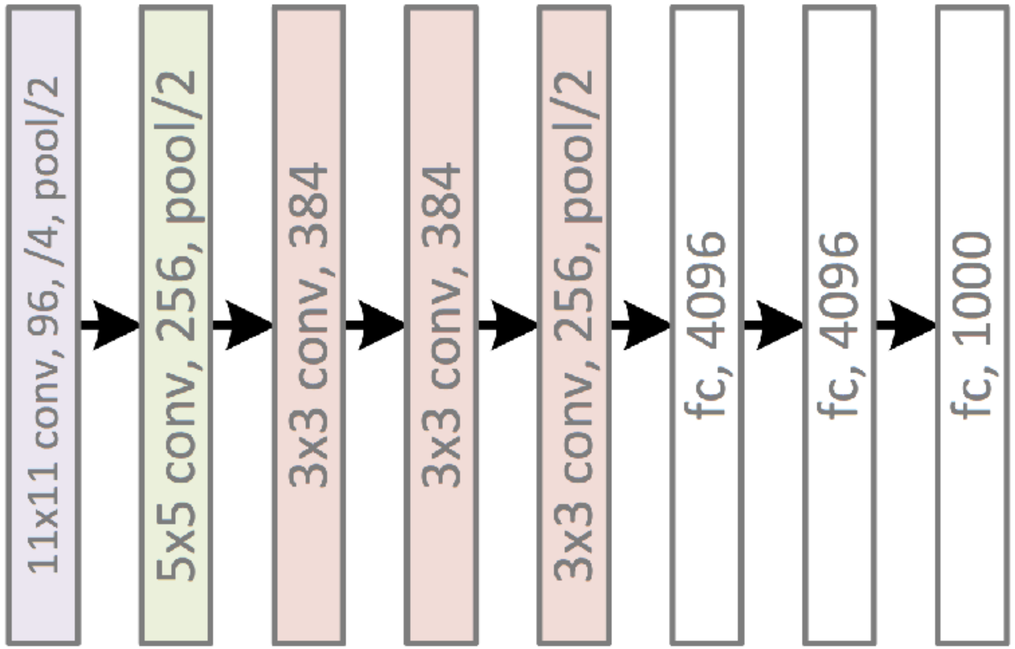

HC
ER
36
46
TF
TH
STPa
AITd
AITv
7b
7a
FEF
STPp
CITd
CITv
VIP
LIP
MSTd
MSTl
FST
PITd
PITv
DP
VOT
MDP
MIP
PO
MT
V4t
V4
PIP
V3A
V3
VP
M
V2
P-B
P-I
M
V1
P-B
P-I
M
P
LGN
M
P
RGC

Computer Vision


ImageNet data
backbone structure
pre-train
classification network
features
detection network (e.g. R-CNN)
segmentation network (e.g. FCN)
human pose estimation network
depth estimation network
target data
fine-tune

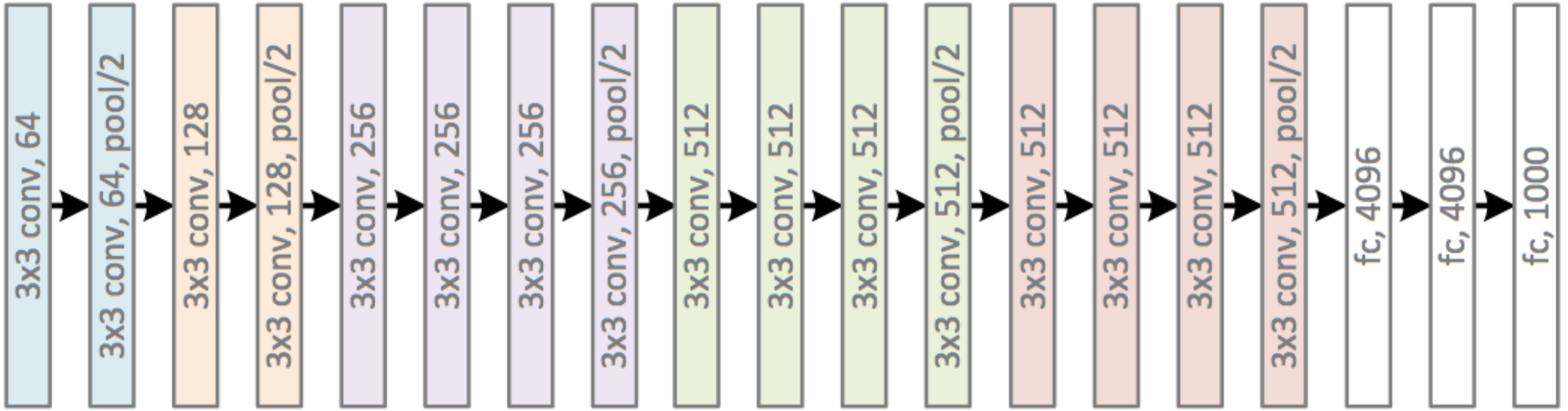
AlexNet

So it begins.



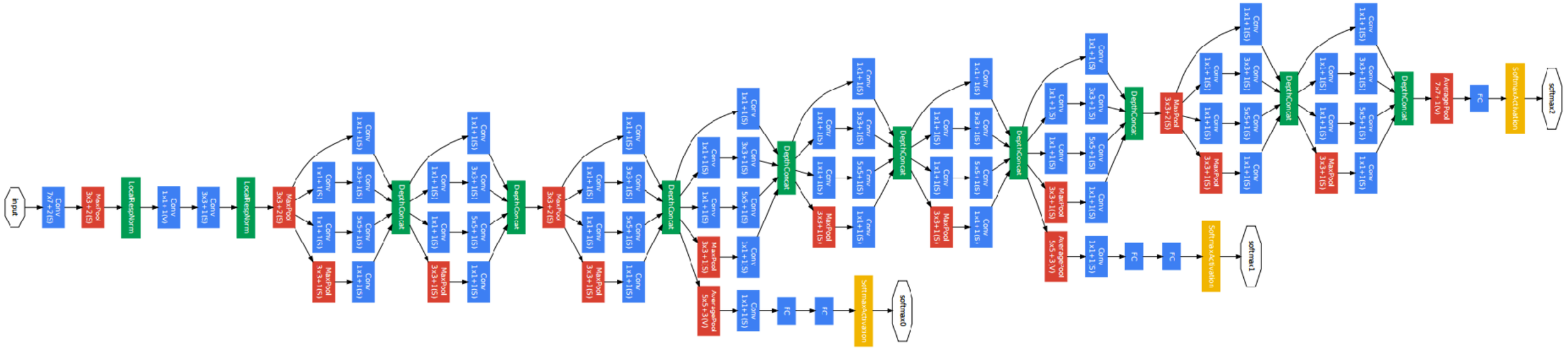
VGGNet

Punch it.



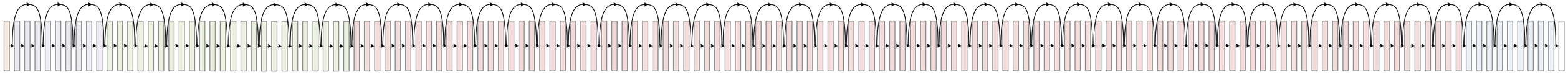
GoogLeNet

We must go deeper.



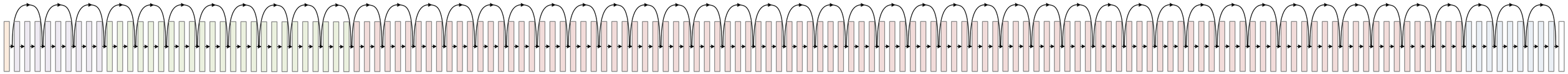
ResNet

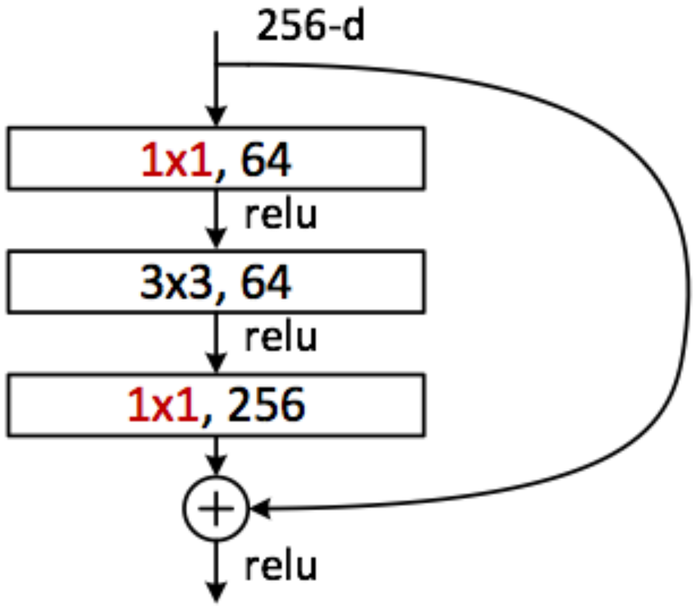
And we took the word seriously



ResNet

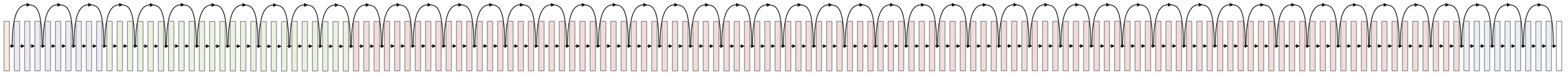
And we took the word seriously

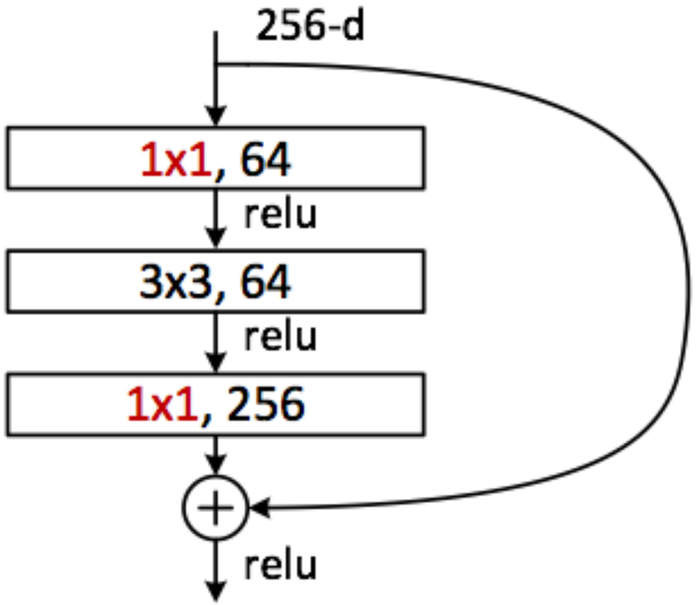


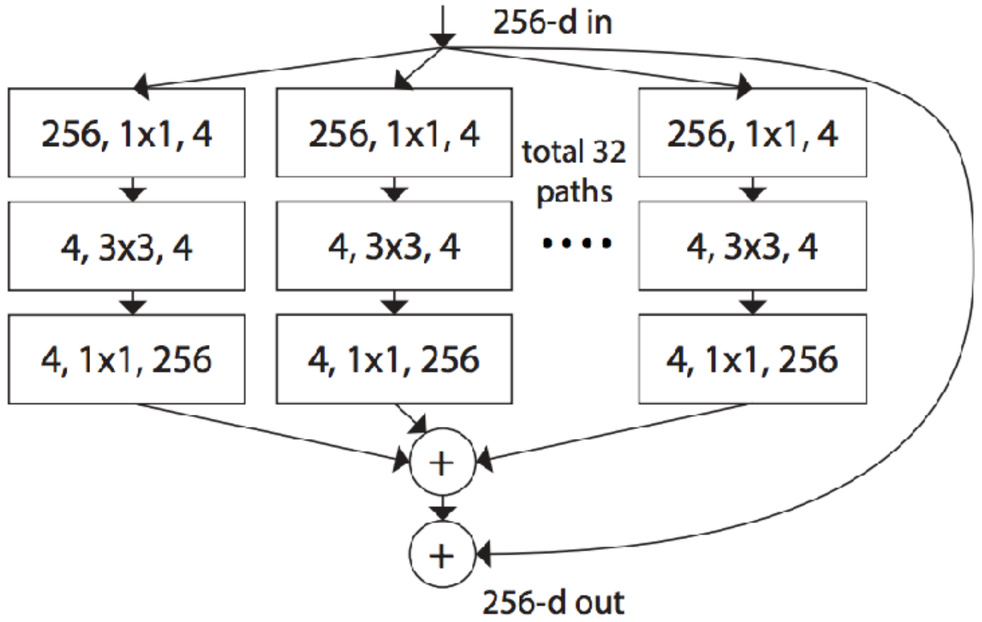


ResNeXT

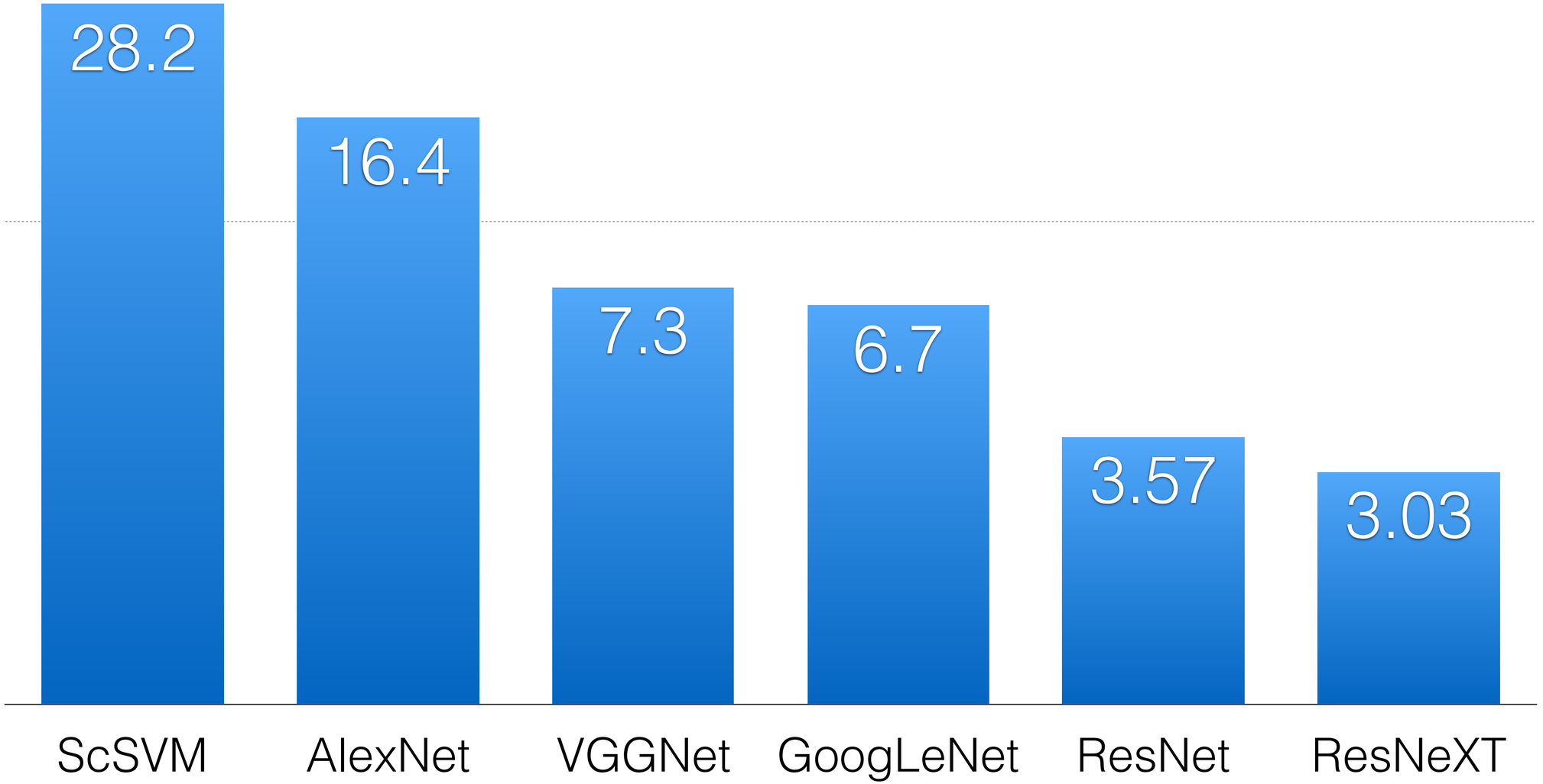
We totally see it coming





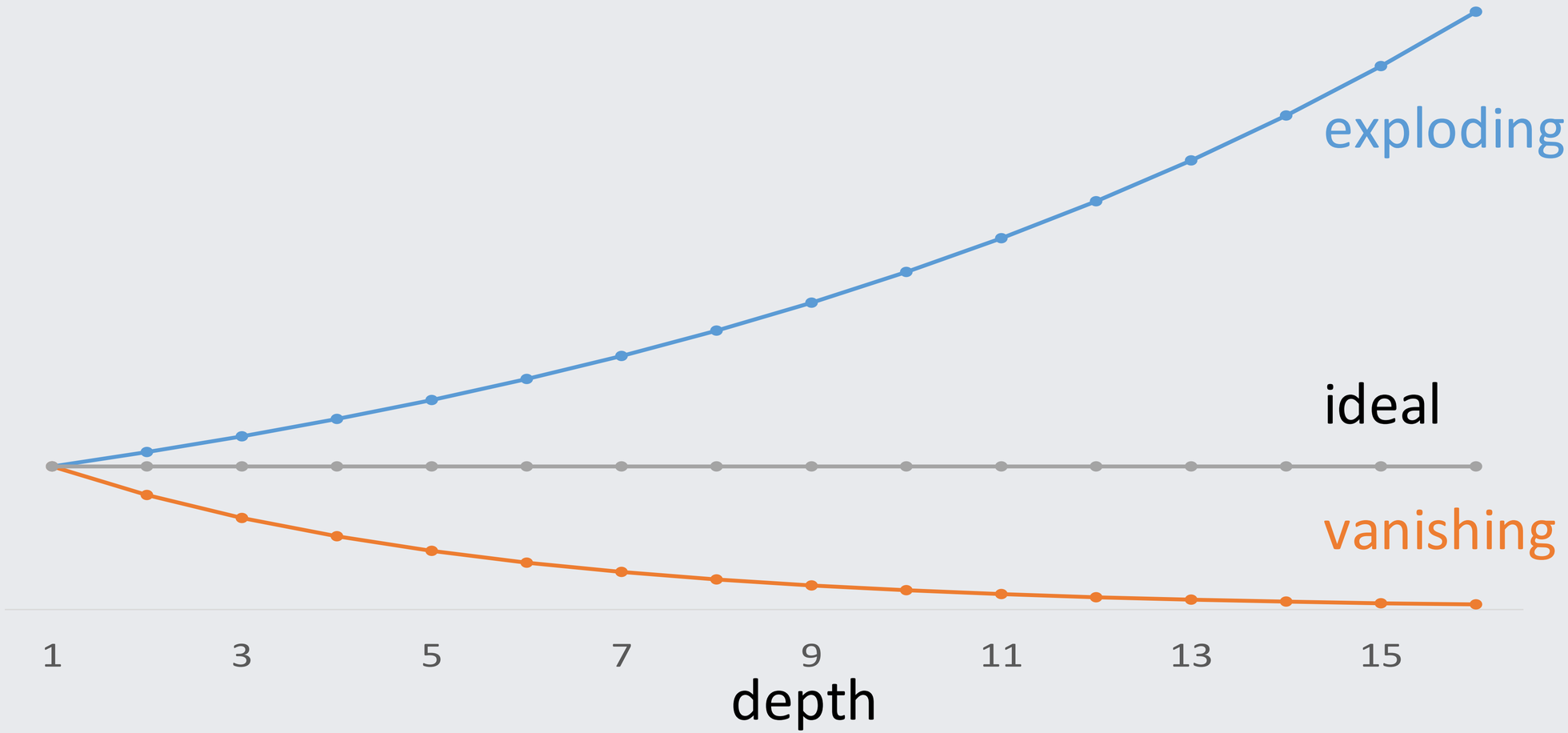


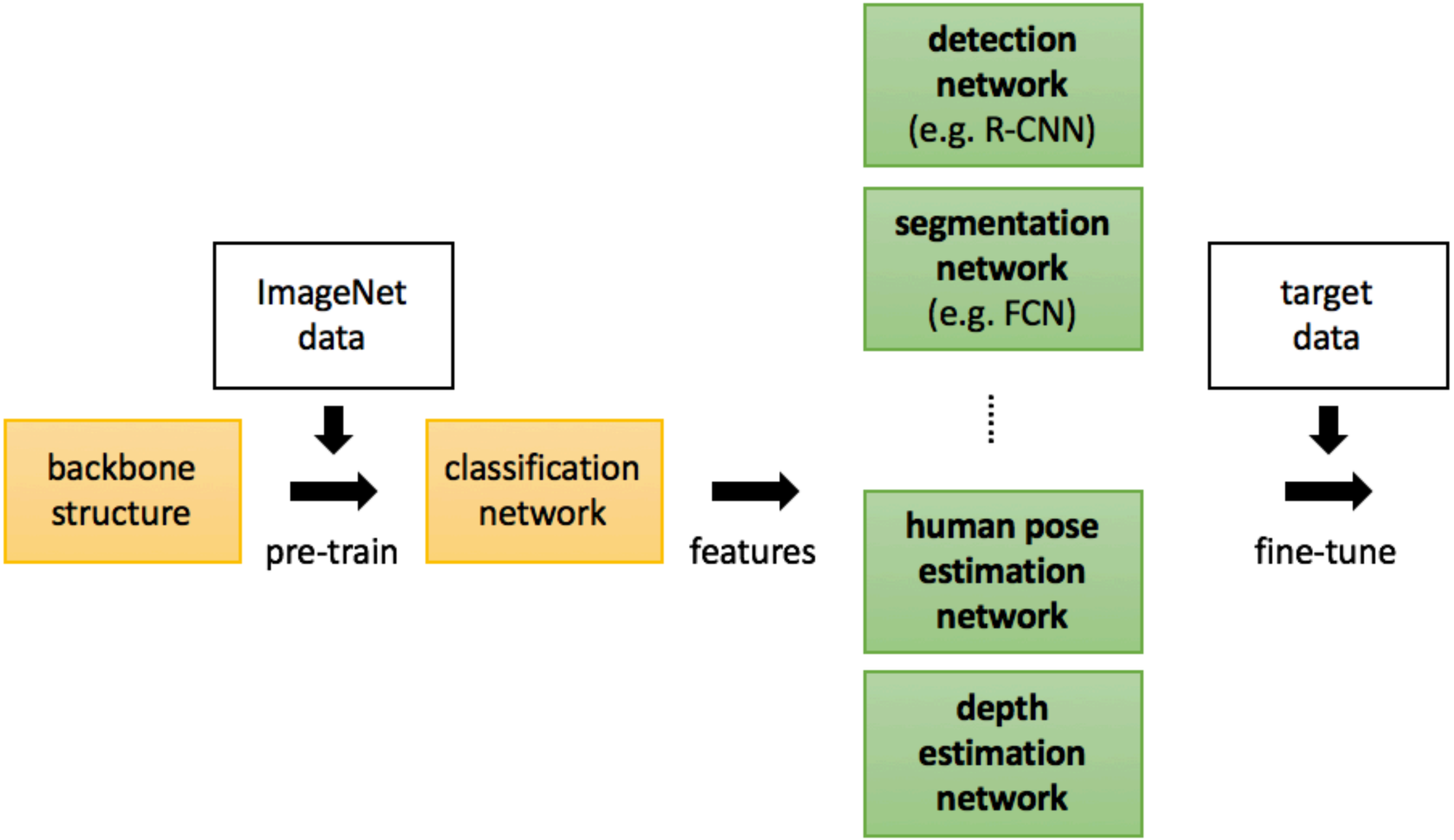
Pushing the Performance



Why is it challenging?

Gradients, as one example




ImageNet data
backbone structure
pre-train
classification network
features
detection network (e.g. R-CNN)
segmentation network (e.g. FCN)
human pose estimation network
depth estimation network
target data
fine-tune


CTRL
ALT
DEL

Deep Learning Systems

"SAP"
-
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

1.2 million =

(# of images in ImageNet1K)

(# of new images @FB every 5 mins in 2013)

(# of AI jobs per month @FB)

Scalability

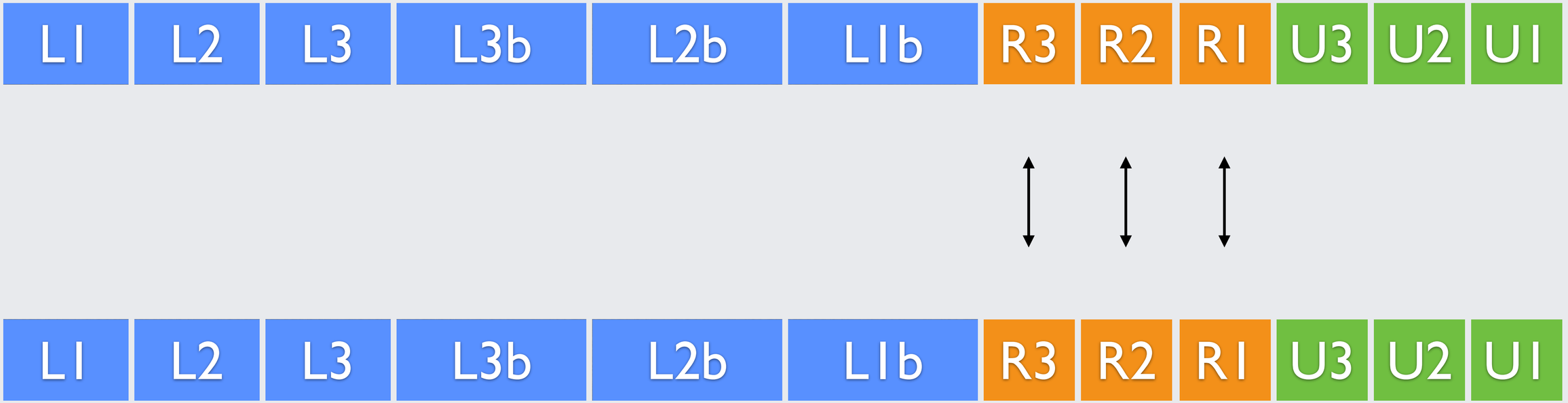
Run fast, run far

Scalability

Run fast, run far

Scalability

Run fast, run far



Scalability

Run fast, run far

The Return of MPI

"I'm your father", said Allreduce.

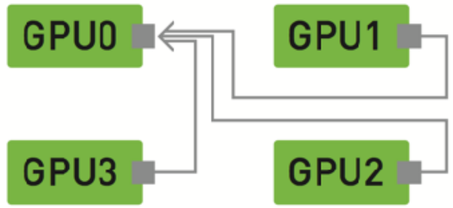
Allreduce

Tree based - O(MlogN)

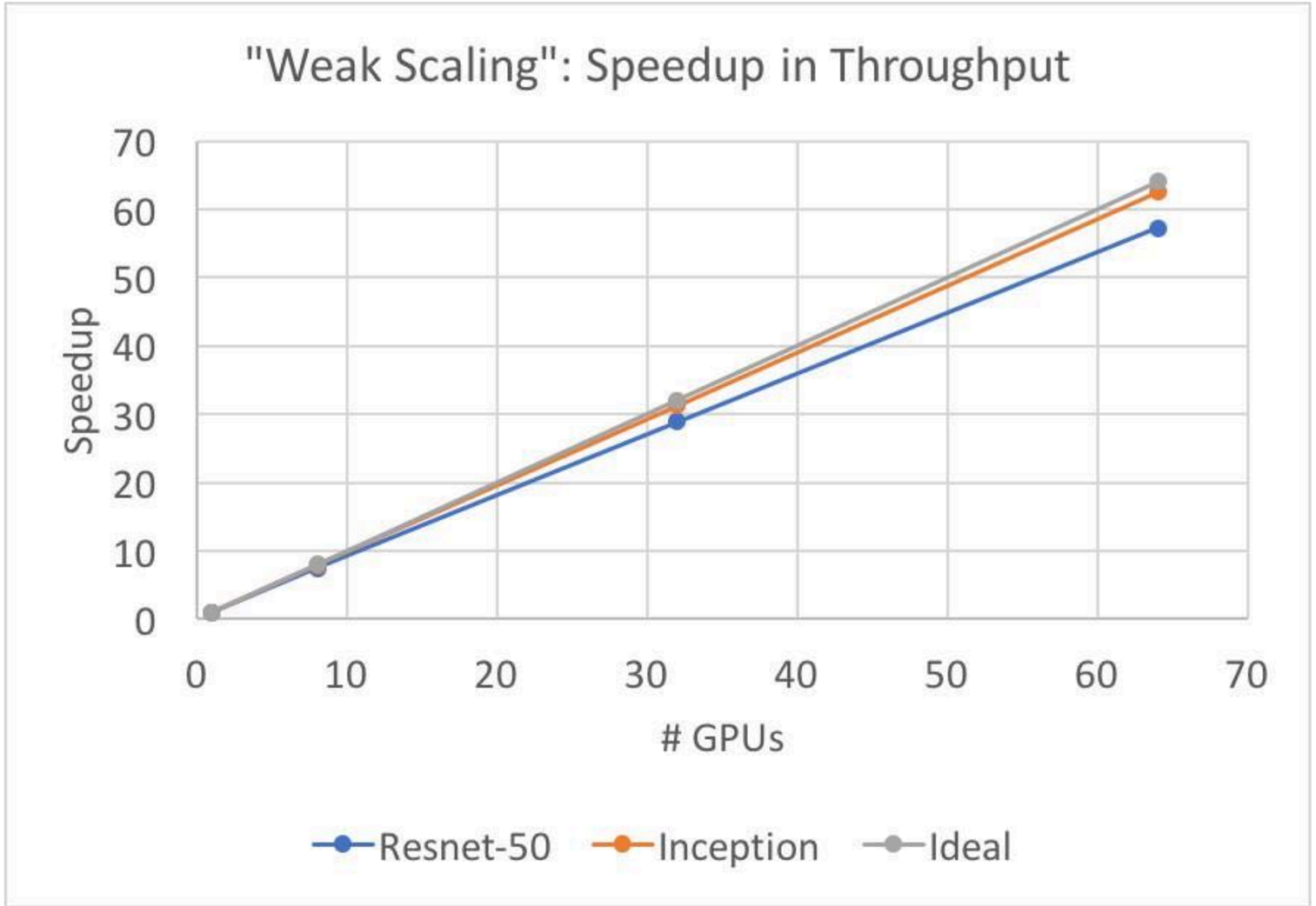
Ring based - O(M)

etc.

NCCL



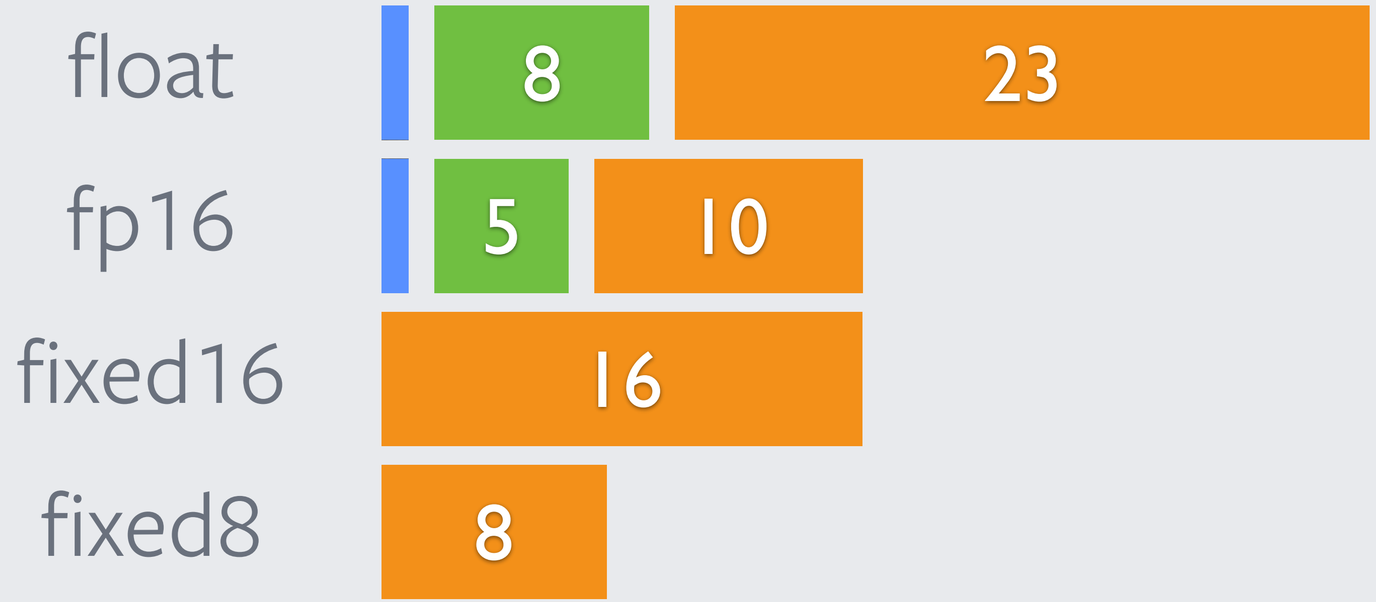
And so we scale



"SAP"
-
Arithmetics

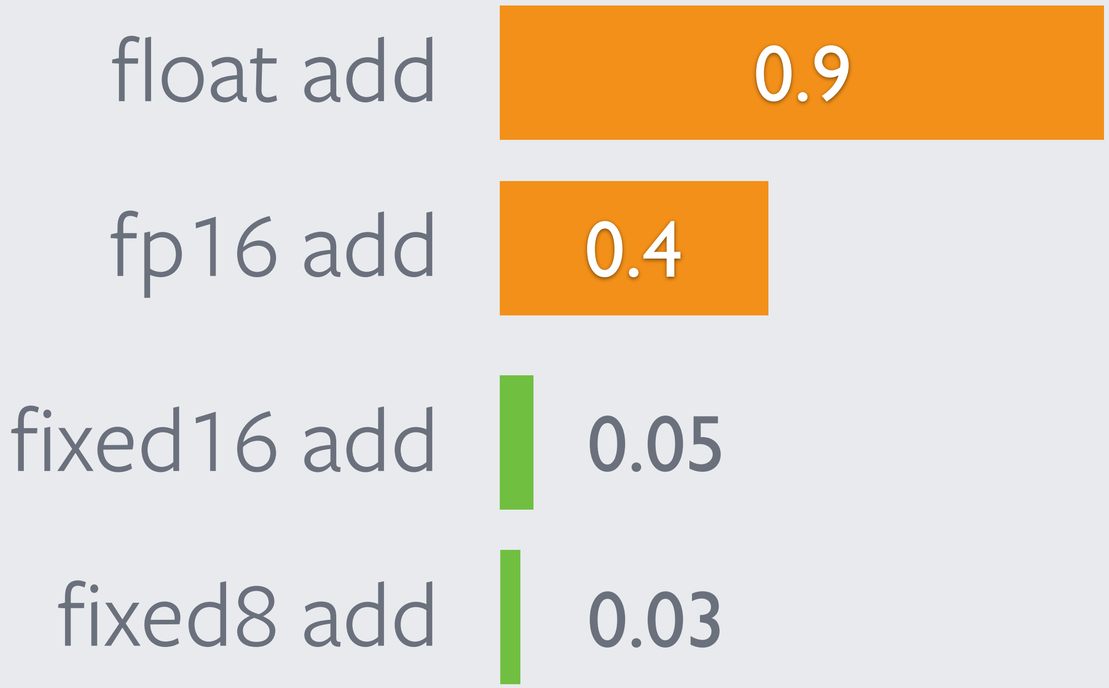
Quantized Computation

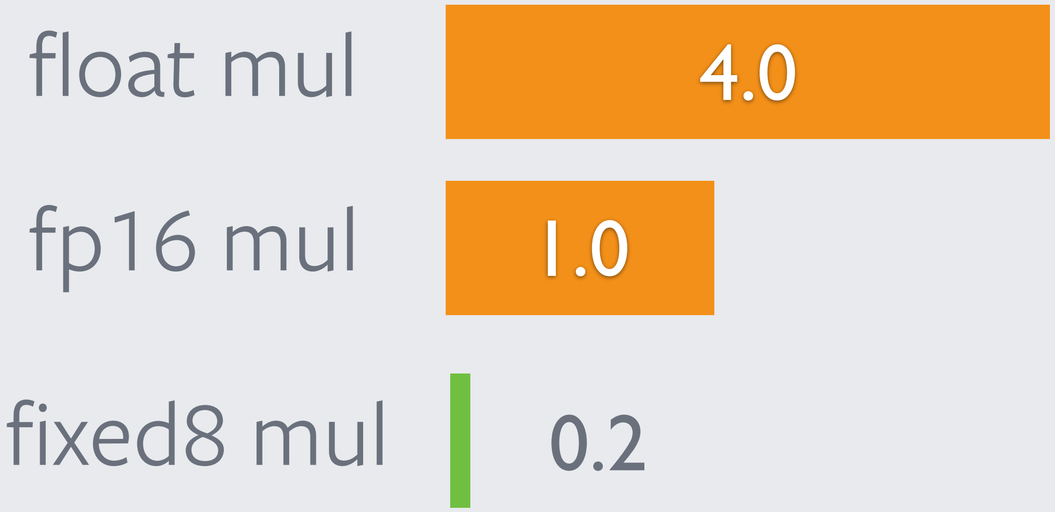
Forget about float, the world is bigger



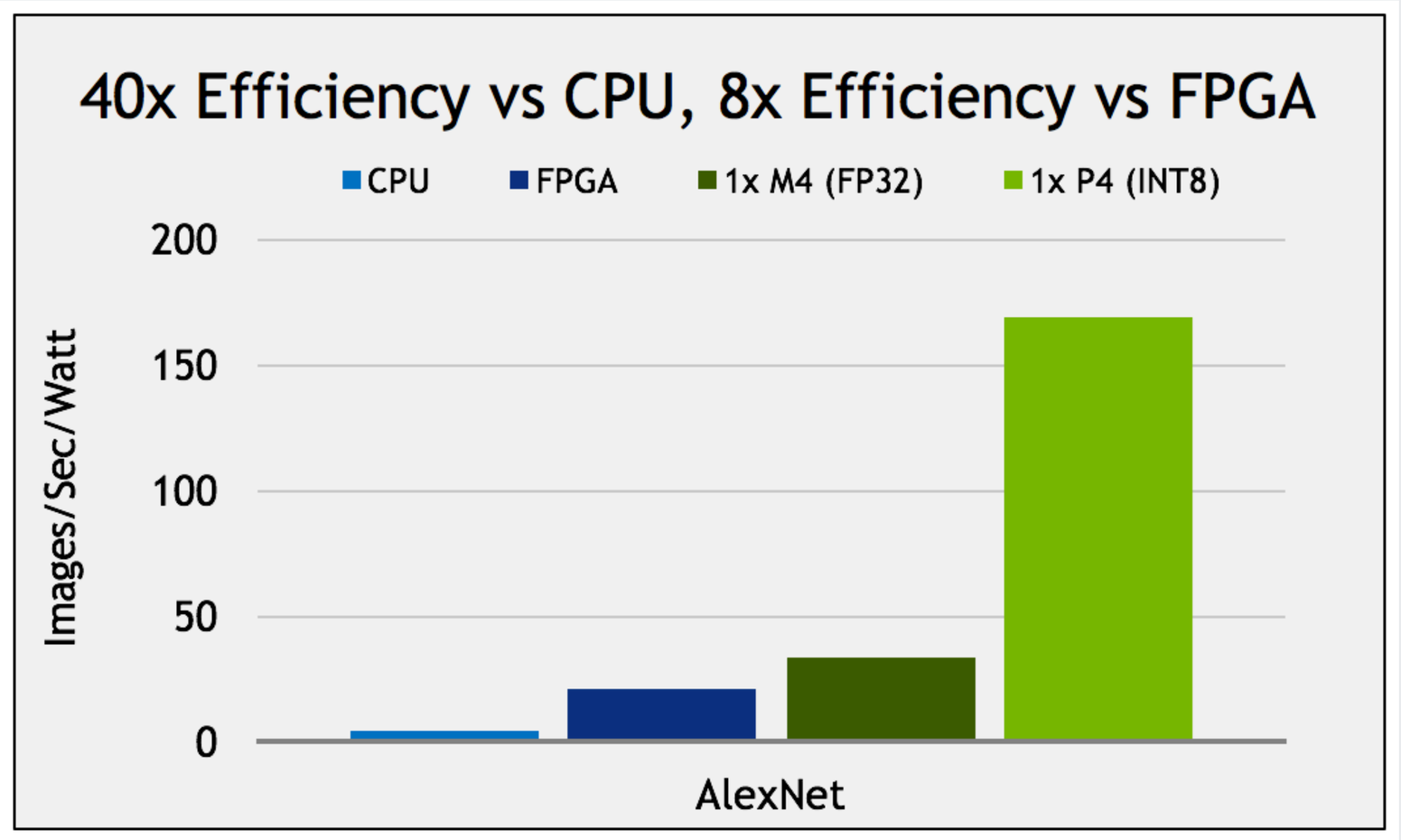
Why do we care?

Battery life is life.



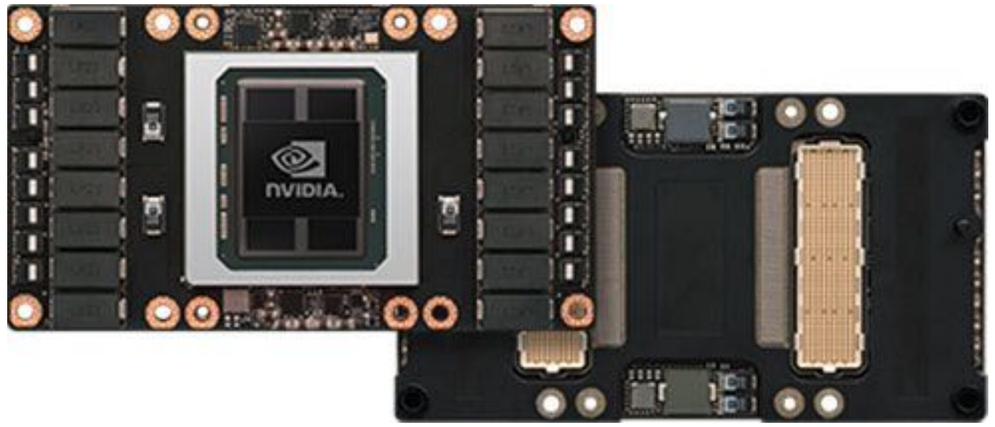


How does it perform?



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

Why does it matter for cars?



250 watts

10 -> 20 TFlops

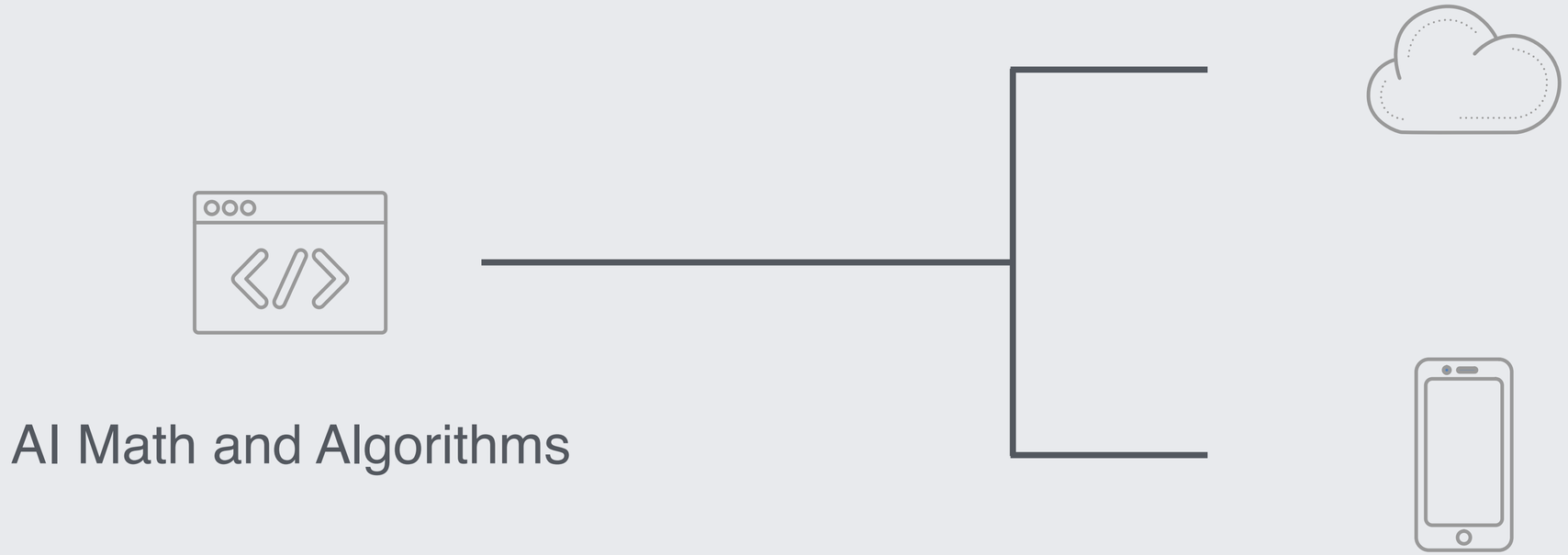
10 watts

0.7 -> 1.5 TFlops

"SAP"
-
Portability

Portable System

One software to rule them all, and...



AI Math and Algorithms

Deployment Platforms

Portable System

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

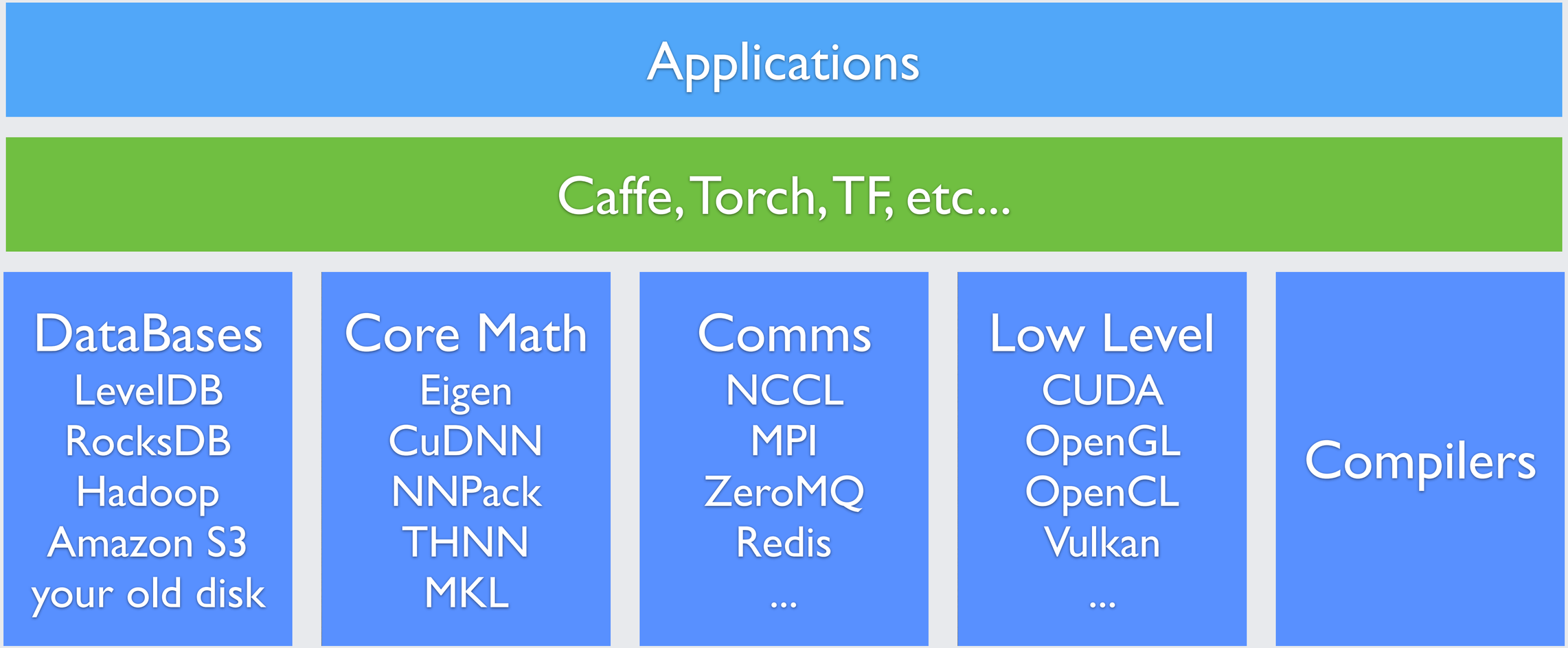
Model

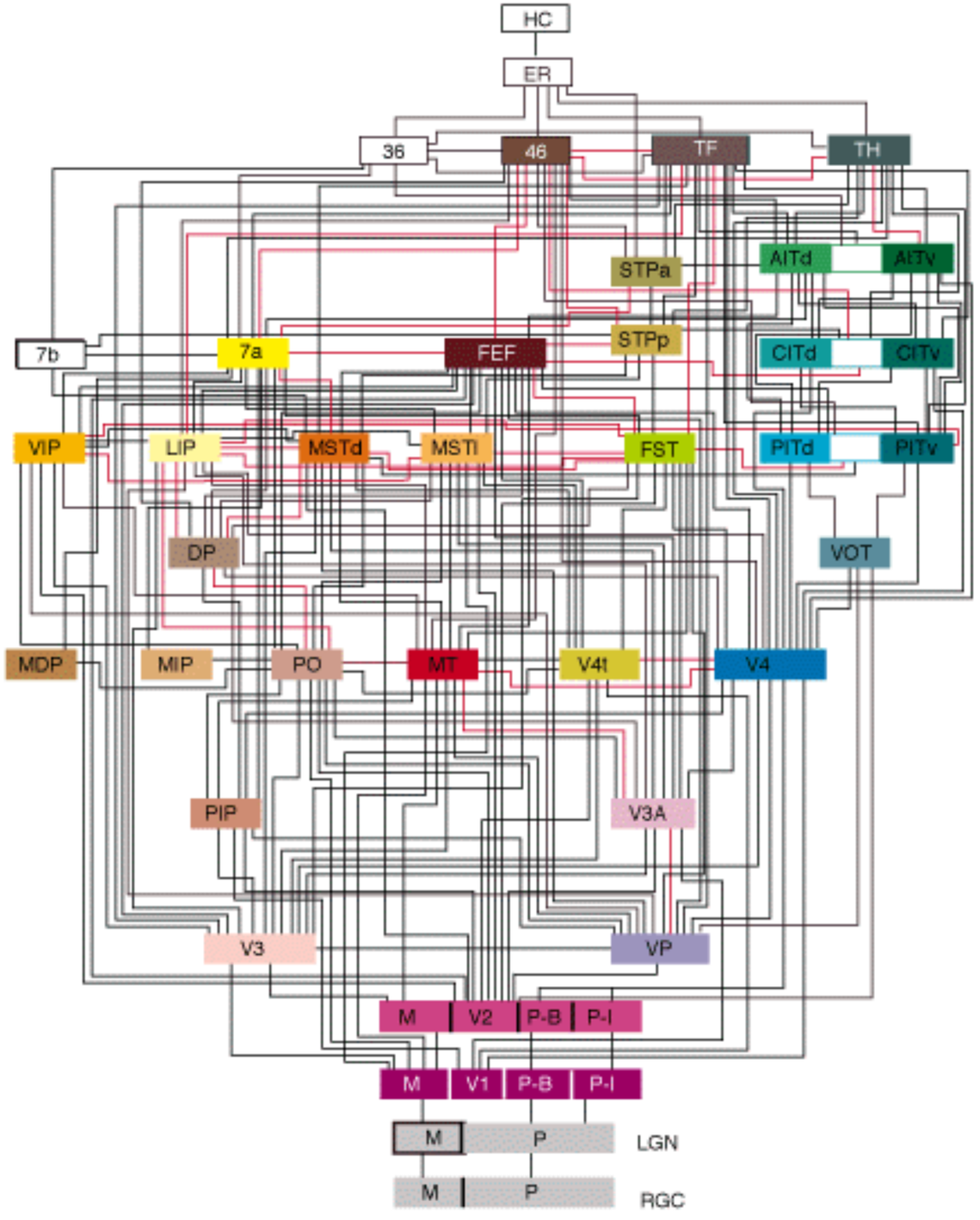
```
auto predictor =
    caffe2::Predictor(model_file)
```

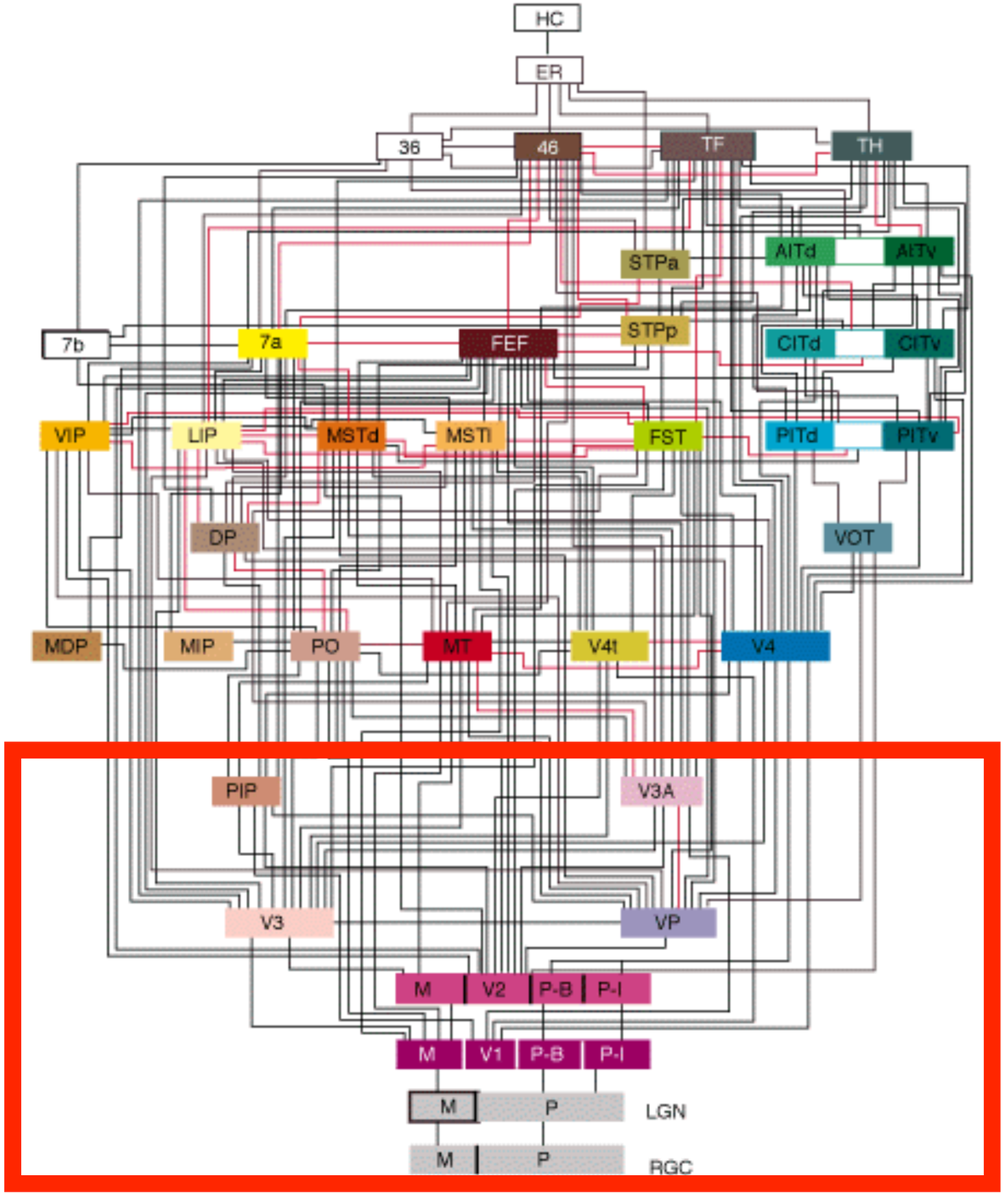
```
public class Predictor implements
    Caffe2ModelInterface;
```

The Land of Deep Learning System

Not as complex as a car, but still.




HC
ER
36
46
TF
TH
STPa
AITd
AITv
7b
7a
FEF
STPp
CITd
CITv
VIP
LIP
MSTd
MSTl
FST
PITd
PITv
DP
VOT
MDP
MIP
PO
MT
V4t
V4
PIP
V3A
V3
VP
M
V2
P-B
P-I
M
V1
P-B
P-I
M
P
LGN
M
P
RGC


HC
ER
36
46
TF
TH
STPa
AITd
AITv
7b
7a
FEF
STPp
CITd
CITv
VIP
LIP
MSTd
MSTl
FST
PITd
PITv
DP
VOT
MDP
MIP
PO
MT
V4t
V4
PIP
V3A
V3
VP
M
V2
P-B
P-I
M
V1
P-B
P-I
M
P
LGN
M
P
RGC

Thank you!

Recent Trends in Computer Vision and Deep Learning Systems

Yangqing Jia