

ImageNet is the new MNIST

Chris Ying

Research SWE @ Google Brain
g.co/brain

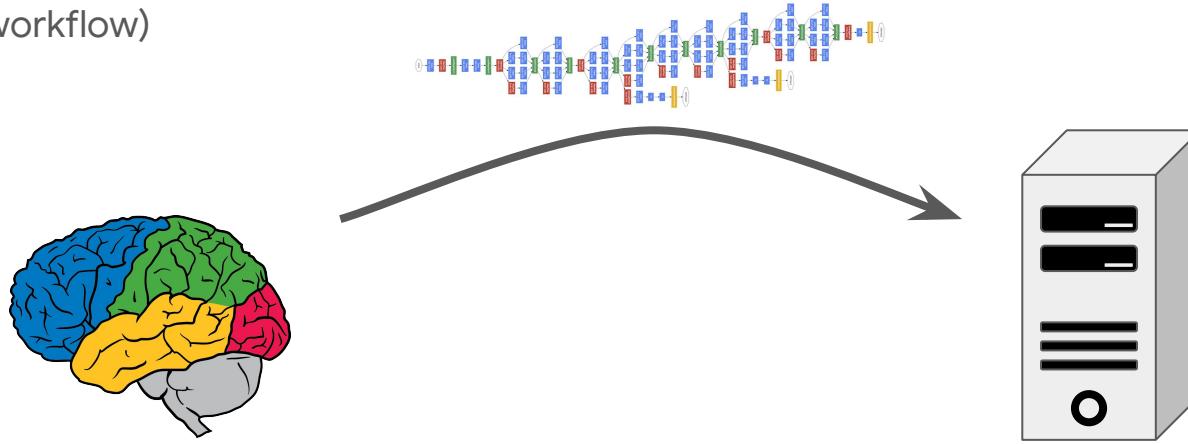
on behalf of **many** people across Google

Goal: “Interactive ML supercomputing”

- Hardware
 - Cloud TPUs
 - TPU pods
- Software
 - TensorFlow Datasets, Layers, and Estimator APIs (open-source)
 - XLA compiler (open-source) with TPU backend
- Research
 - Understanding of generalization gap
 - Large-batch training advances

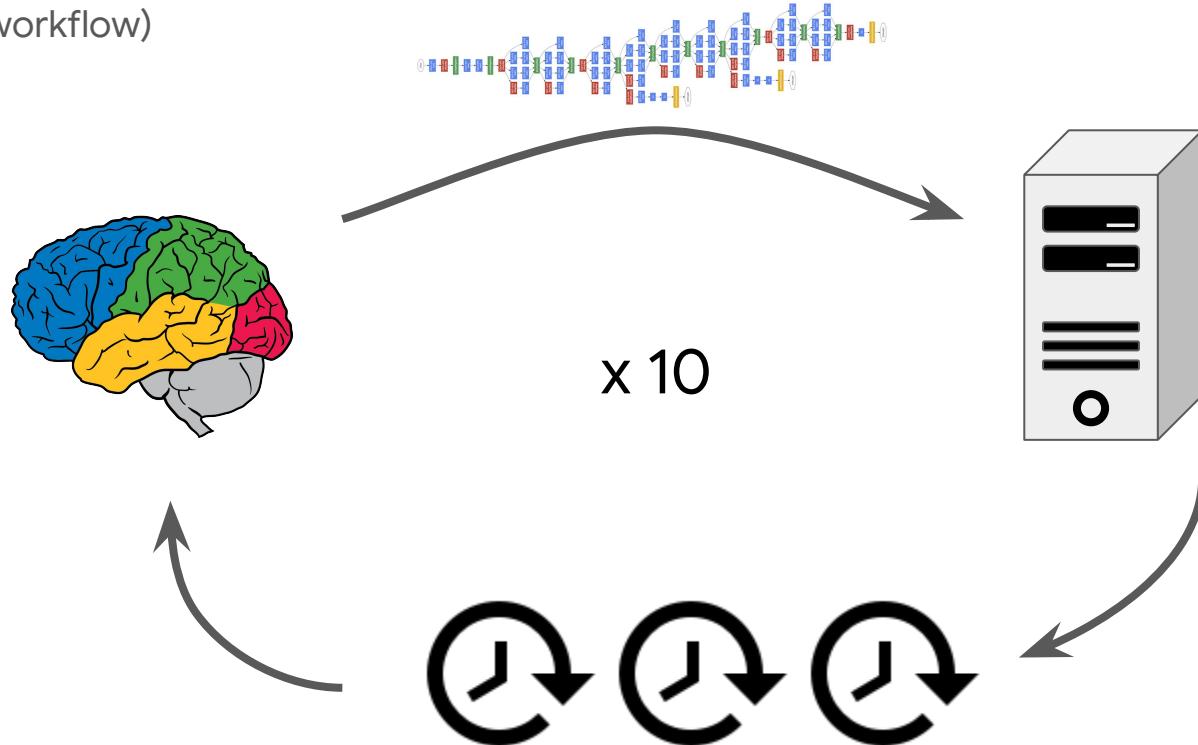
Motivation

(classical workflow)



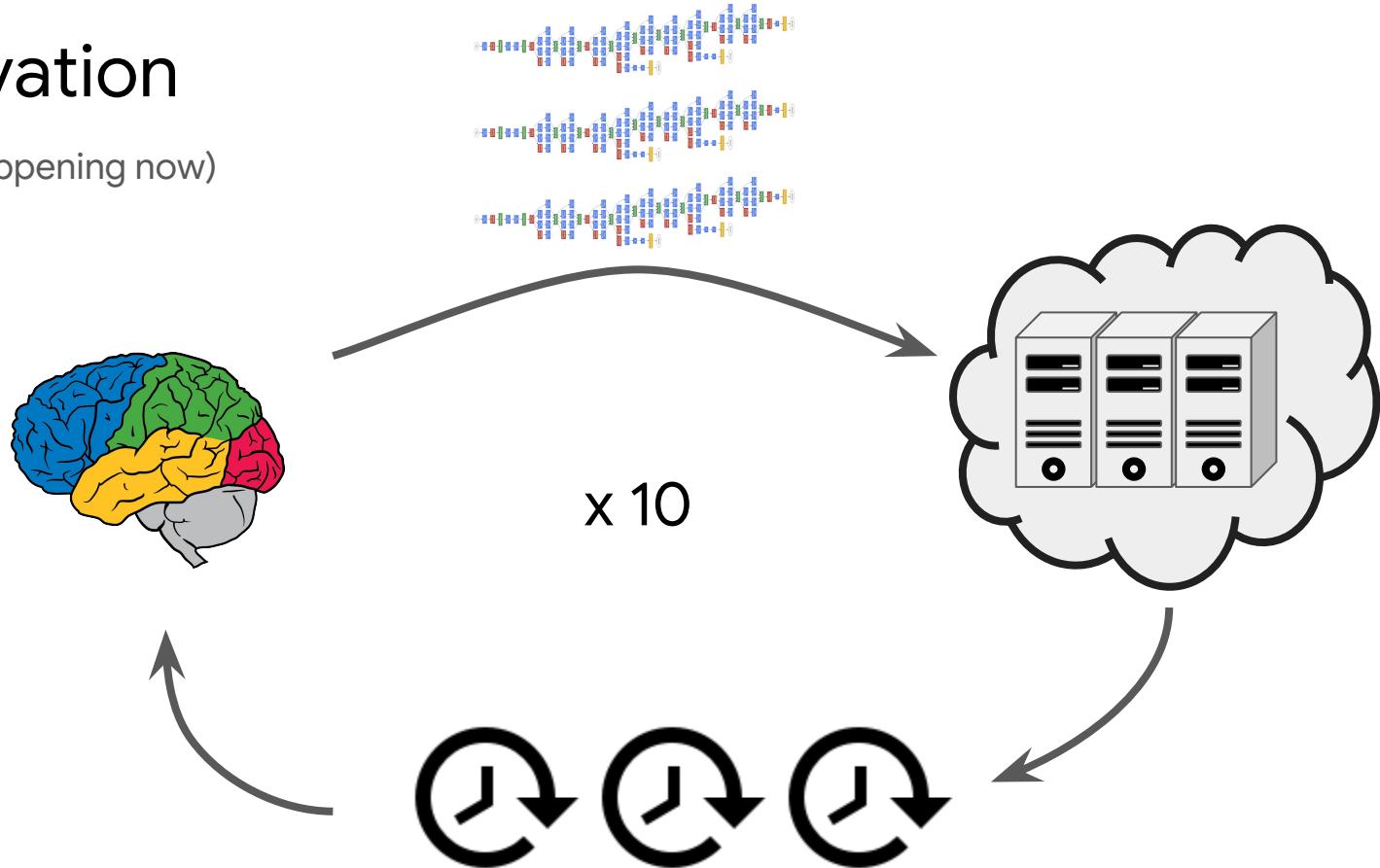
Motivation

(classical workflow)



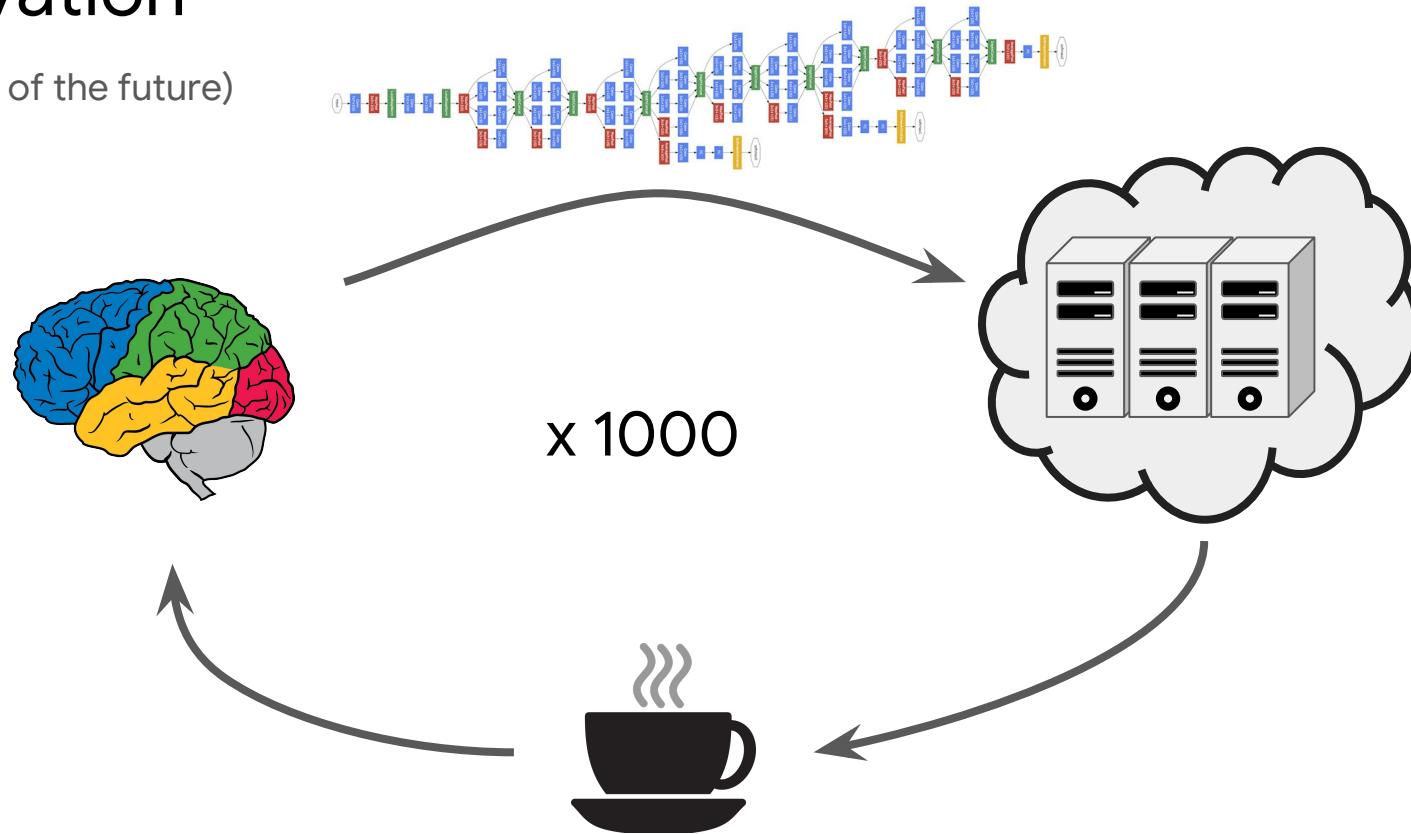
Motivation

(what's happening now)



Motivation

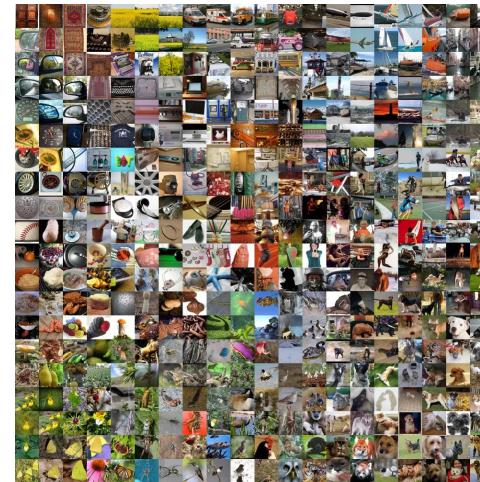
(our vision of the future)



ImageNet is the new MNIST

3 4 2 1 9 5 6 2 / 8
8 9 1 2 5 0 0 6 6 4
6 7 0 1 6 3 6 3 7 0
3 7 7 9 4 6 6 1 8 2
2 9 3 4 3 9 8 7 2 5
1 5 9 8 3 6 5 7 2 3
9 3 1 9 1 5 8 0 8 4
5 6 2 6 8 5 8 8 9 9
3 7 7 0 9 4 8 5 4 3
7 9 6 4 7 0 6 9 2 3

MNIST: 60,000 B&W images



ImageNet: 1,281,167 color images

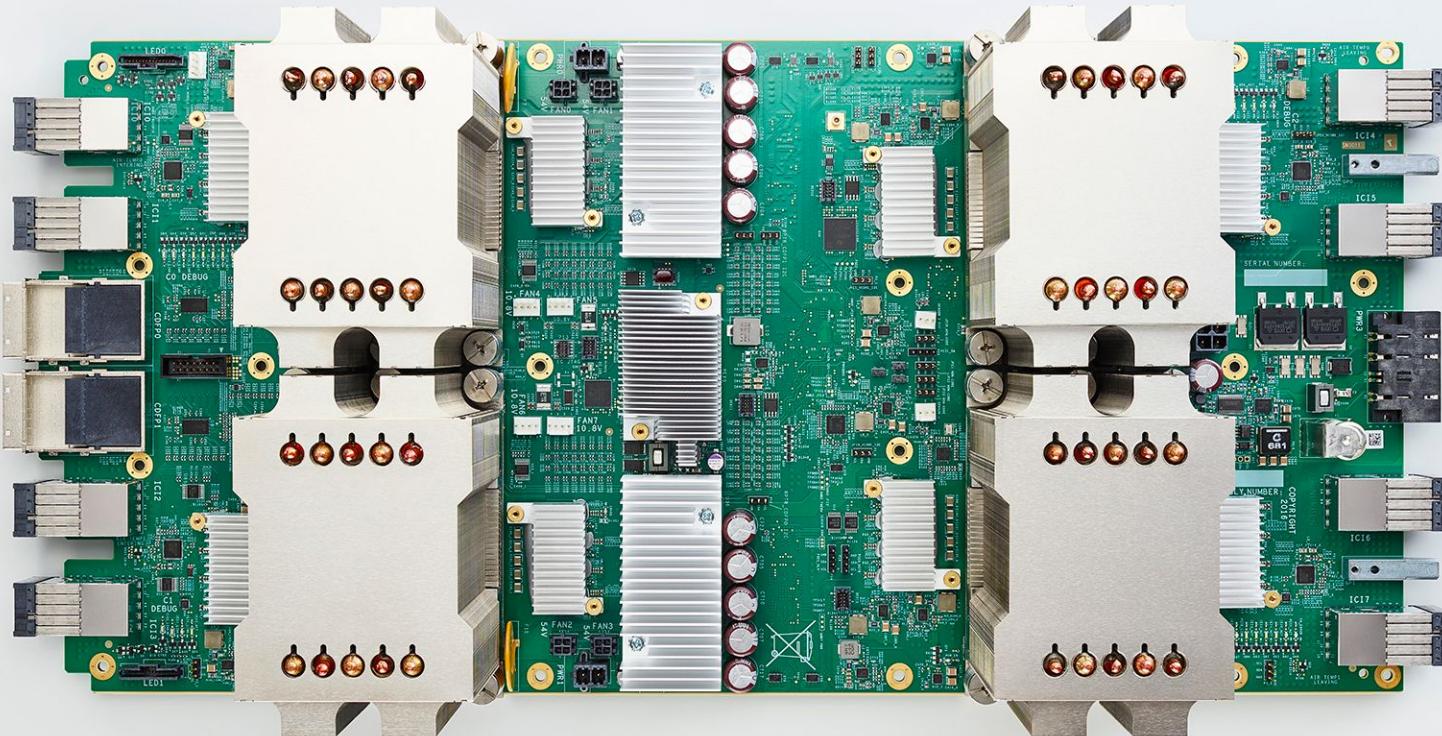
Motivating results

ResNet-50-v2 on ImageNet

# of TPU devices	Batch size	Time to 90 epochs	Accuracy
1	256	23 hours 22 minutes	76.6%
4	1024	5 hours 48 minutes	76.3%
16	4096	1 hour 30 minutes	76.5%
32	8192	45 minutes	76.1%
64	16384	22 minutes	75.0%

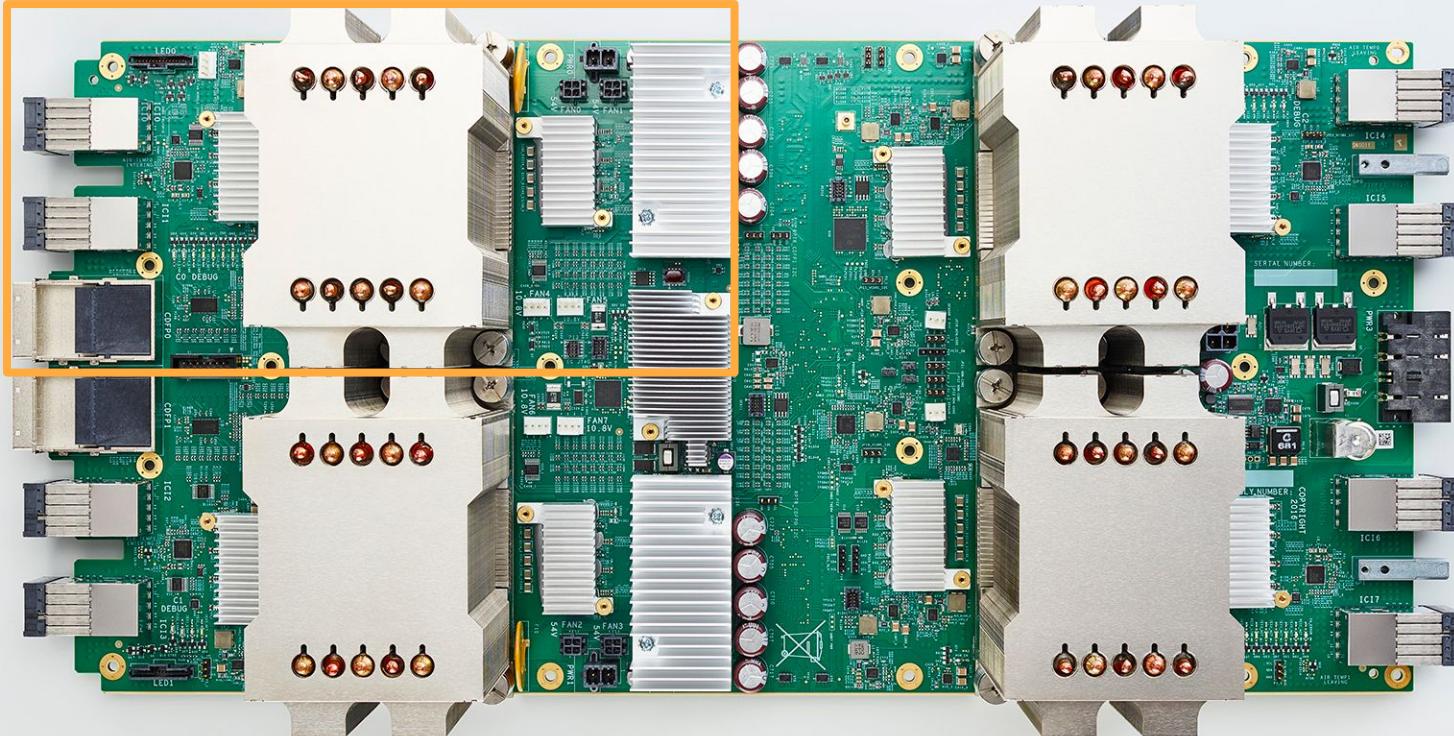
Only change between different runs is batch size (linearly scale LR) and hardware,
no model changes or hyperparameter re-tuning!

Cloud TPU

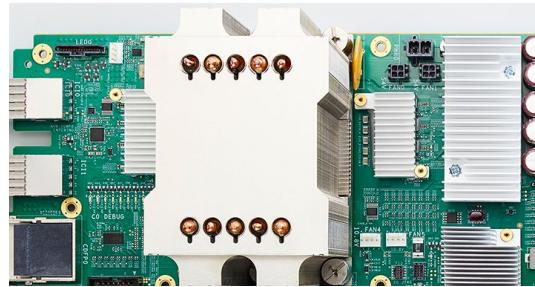


180 TFLOPS of computation, 64 GB of HBM memory, 2400 GB/s mem BW

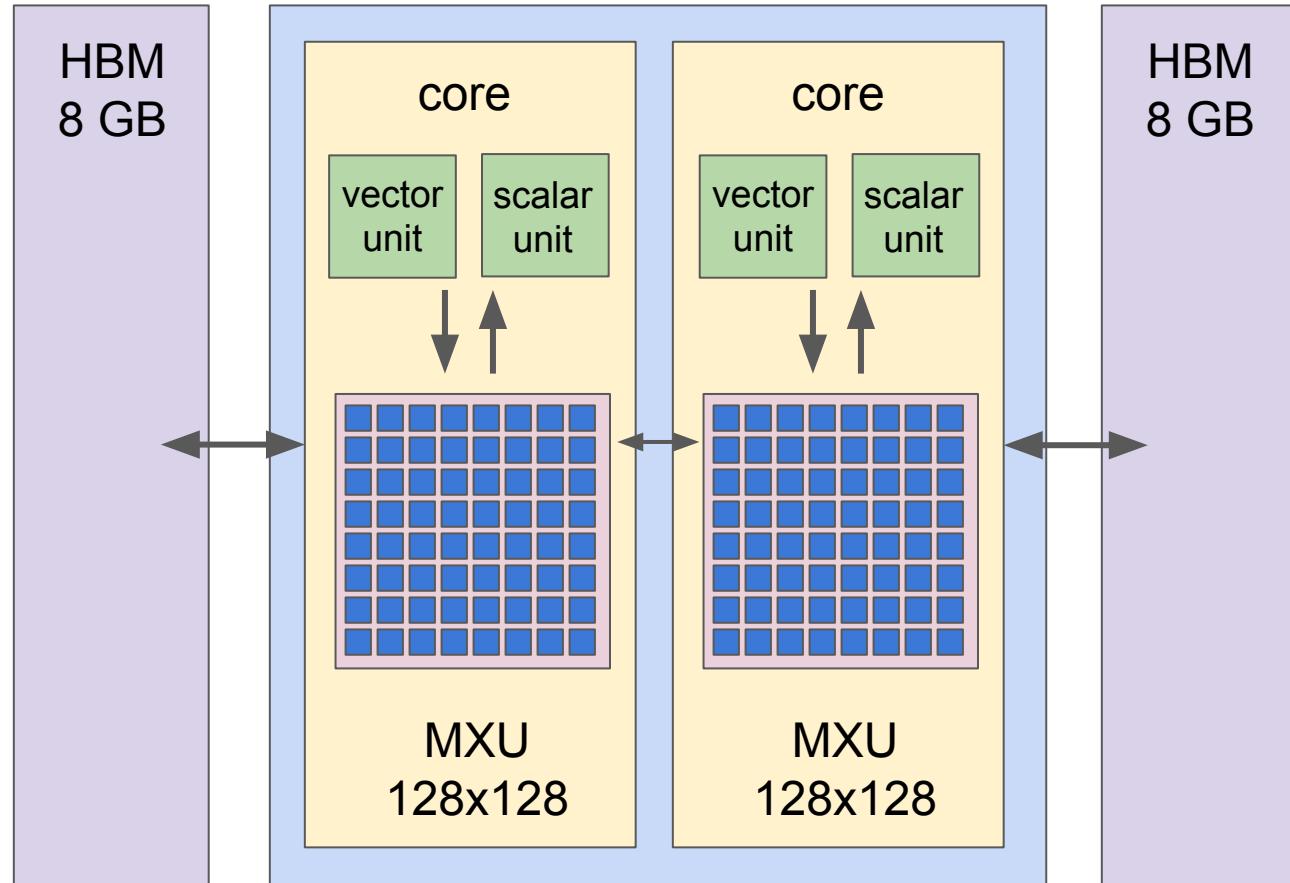
Cloud TPU



TPUv2 Chip



- 45 TFLOPS
- 16 GB of HBM
- 600 GB/s mem BW
- Vector unit: float32
- Scalar unit: float32
- Matrix unit (MXU):
float32 input/output,
reduced precision
multiplication



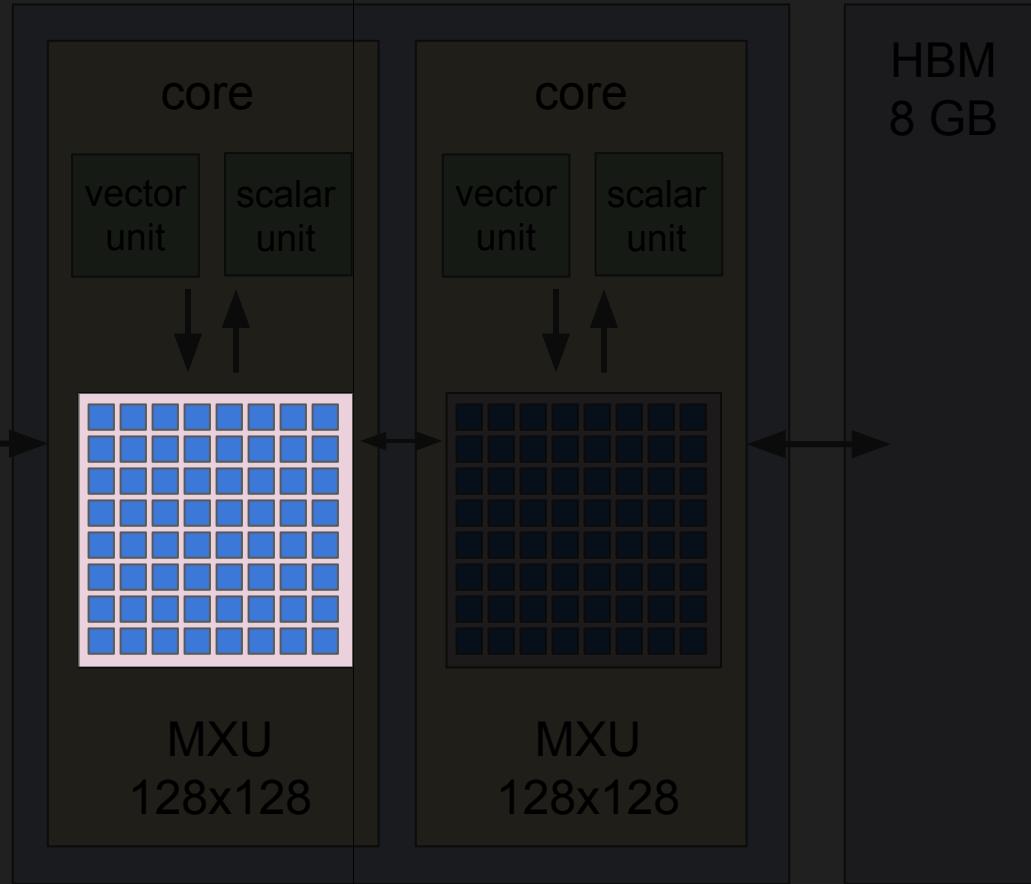
TPUv2 Chip



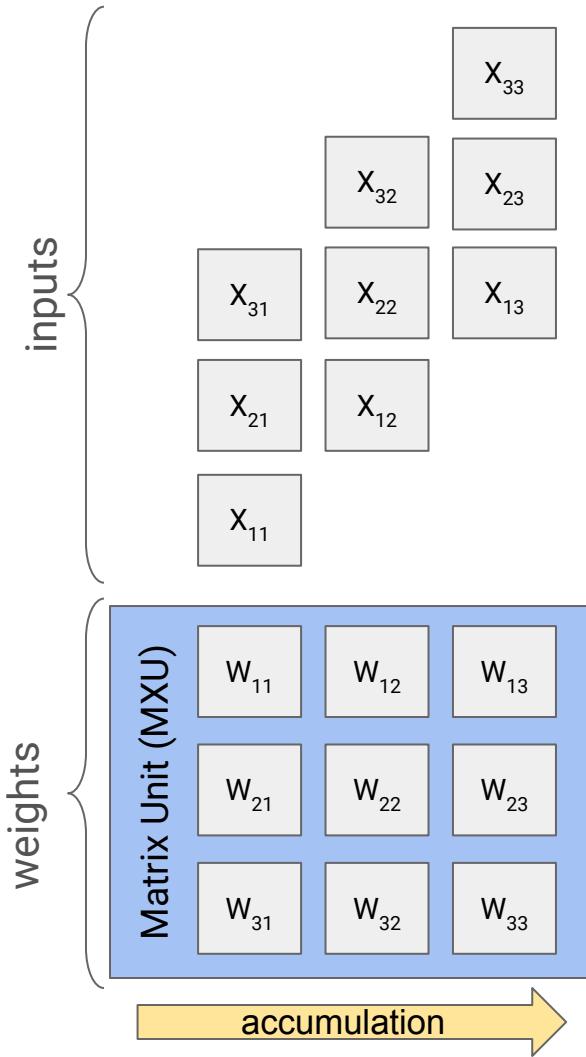
HBM
8 GB

Matrix Unit

- 128x128 systolic array
- 600 GFB/s mem BW
- Scalar unit: 32b float
- MXU: 32b float accumulation but
 - * reduced precision multiplication
 - reduced precision for multipliers
- 45 TFLOPS



Matrix Unit Systolic Array



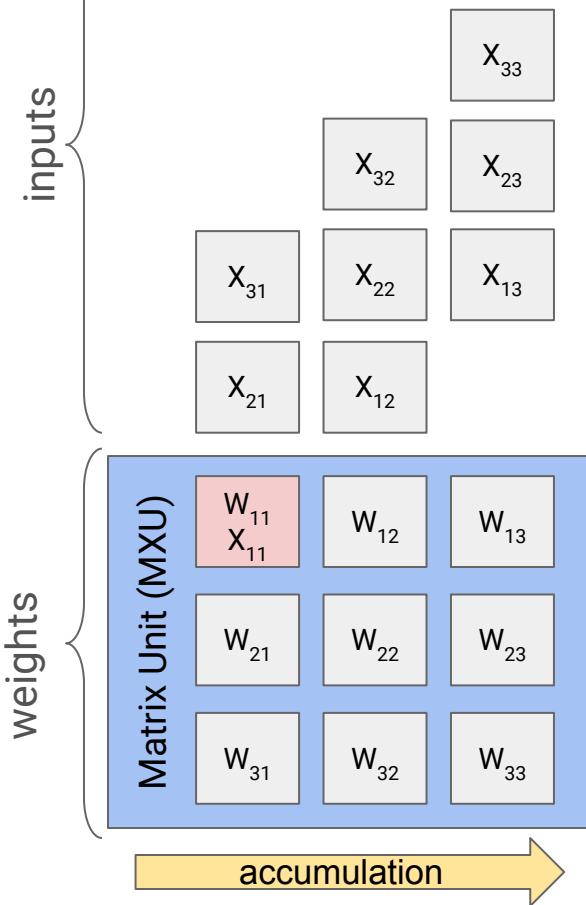
Computing $y = Wx$

Toy example: 3x3 systolic
array

$W = 3 \times 3$ matrix

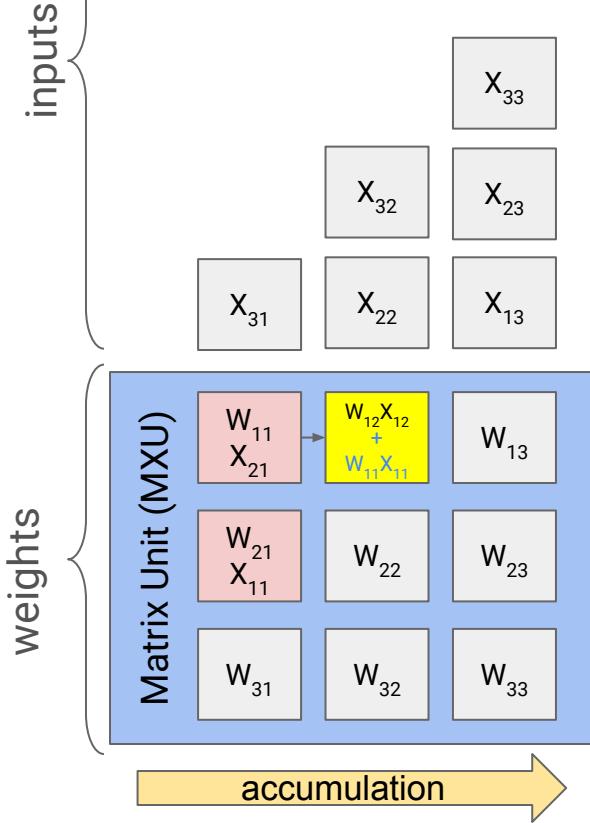
`batch_size(x) = 3`

Matrix Unit Systolic Array



Computing $y = Wx$
with $W = 3 \times 3$, $\text{batch_size}(x) = 3$

Matrix Unit Systolic Array

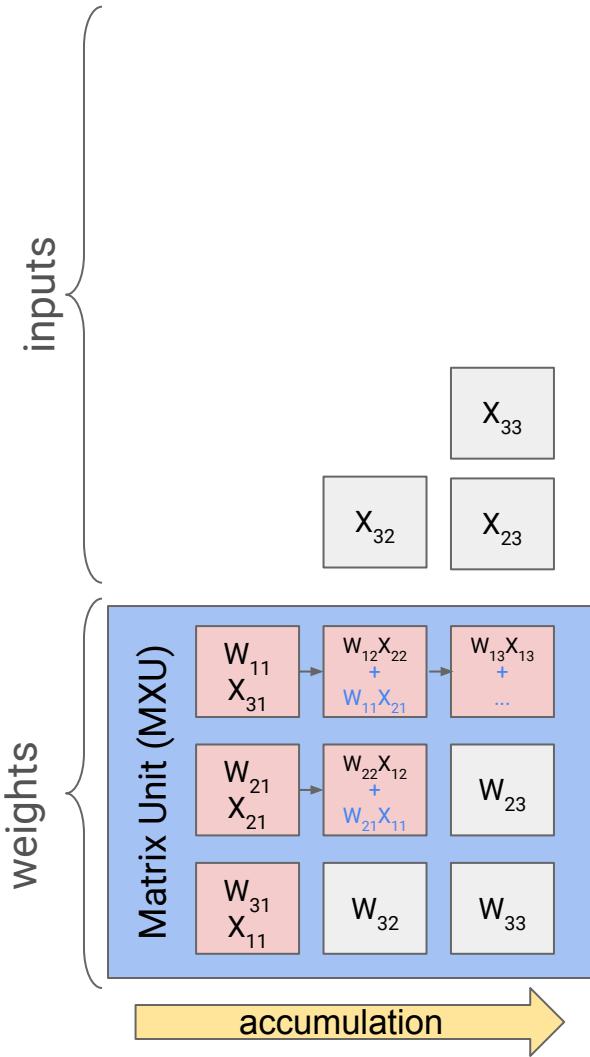


Computing $y = Wx$
with $W = 3 \times 3$, $\text{batch_size}(x) = 3$

Matrix Unit Systolic Array

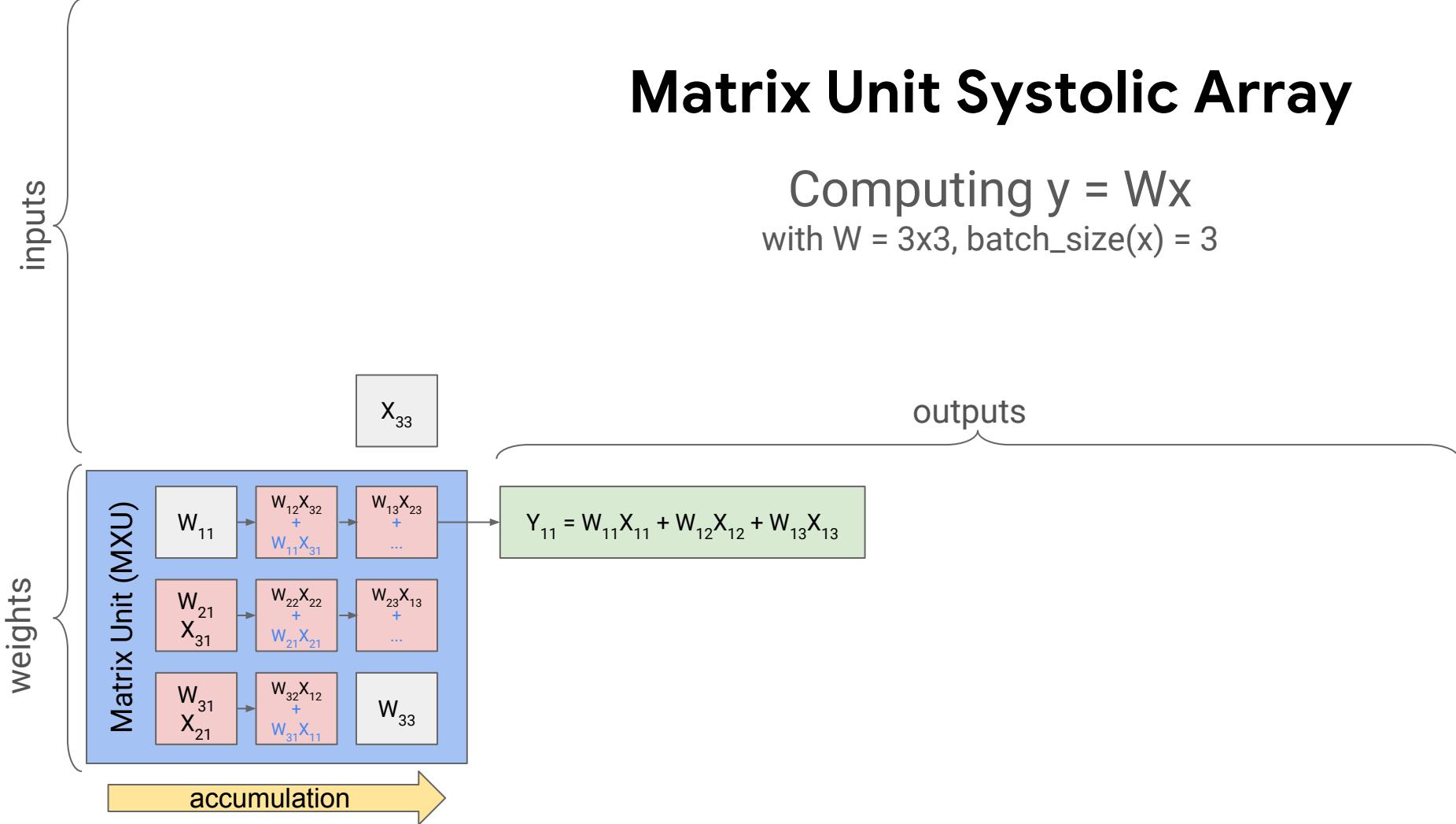
Computing $y = Wx$

with $W = 3 \times 3$, $\text{batch_size}(x) = 3$



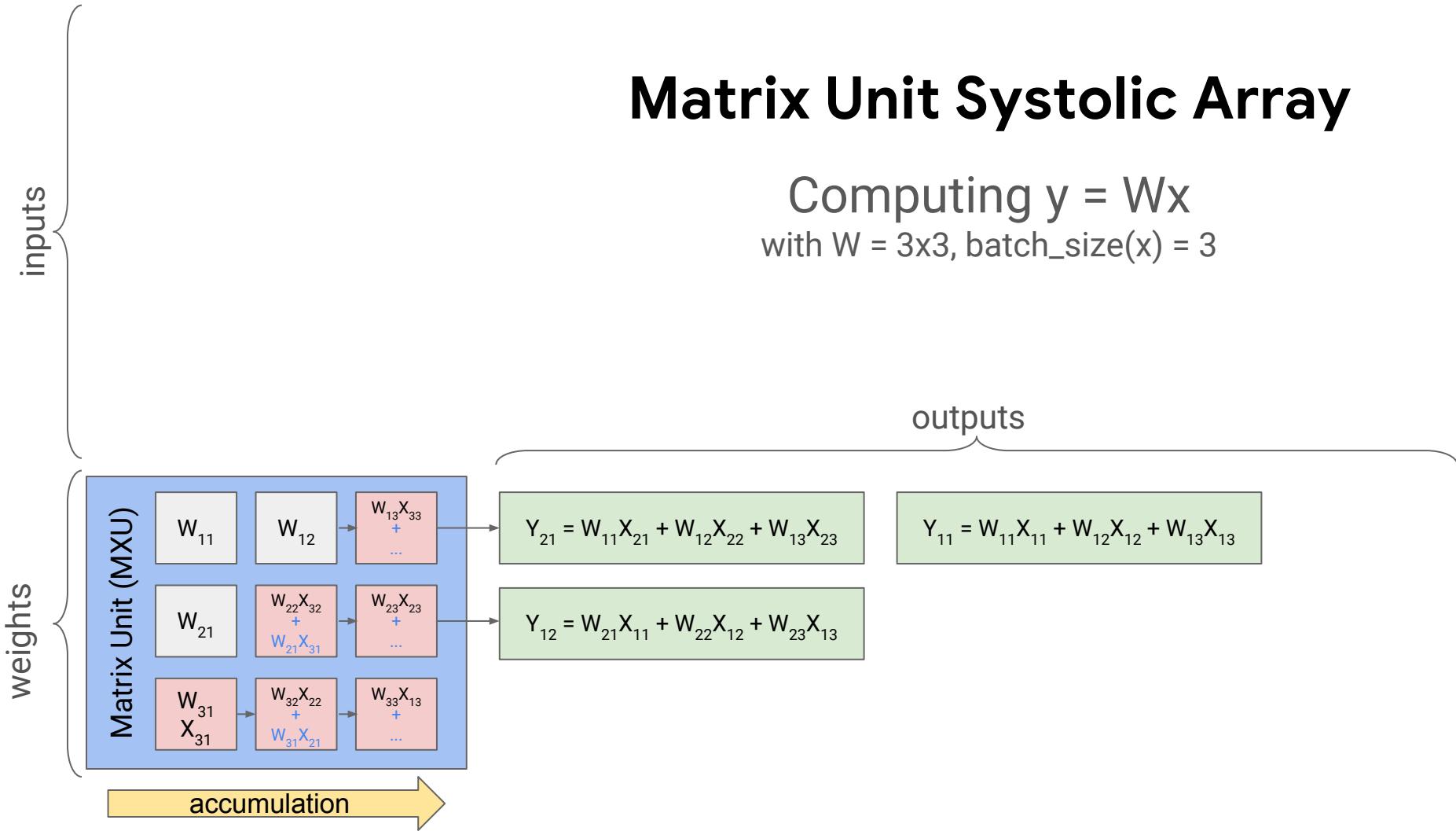
Matrix Unit Systolic Array

Computing $y = Wx$
with $W = 3 \times 3$, $\text{batch_size}(x) = 3$



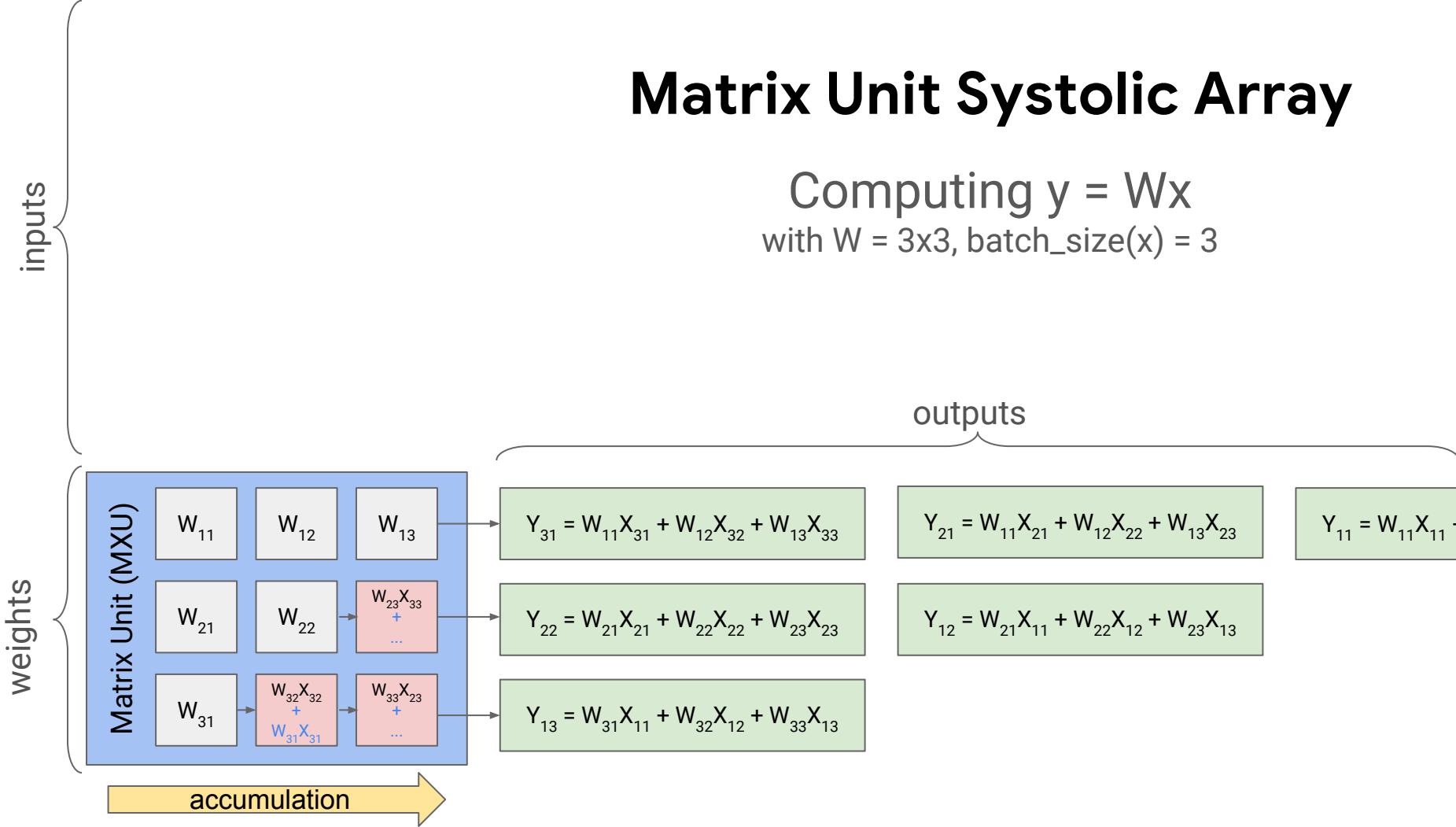
Matrix Unit Systolic Array

Computing $y = Wx$
with $W = 3 \times 3$, $\text{batch_size}(x) = 3$



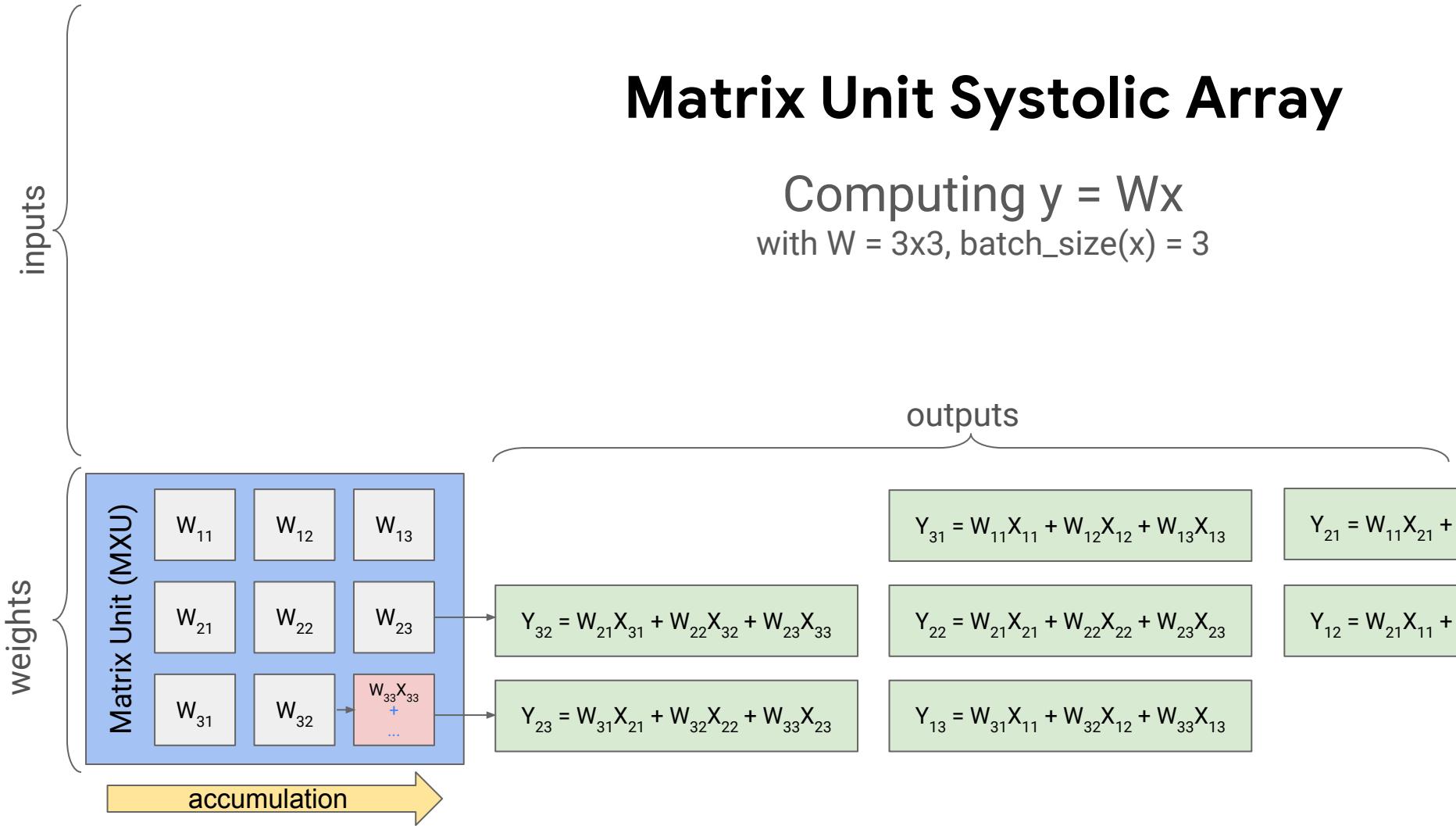
Matrix Unit Systolic Array

Computing $y = Wx$
with $W = 3 \times 3$, $\text{batch_size}(x) = 3$



Matrix Unit Systolic Array

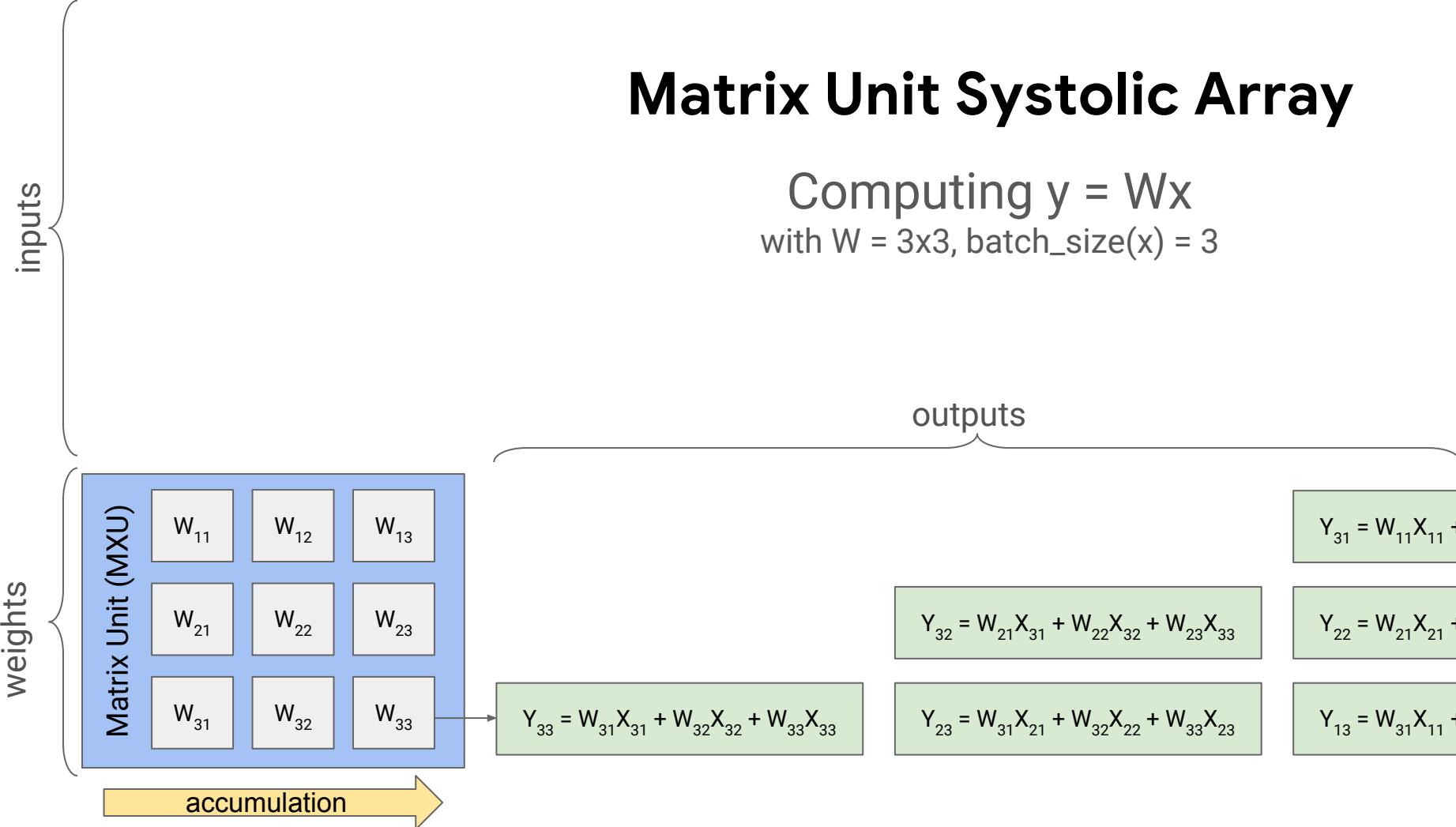
Computing $y = Wx$
with $W = 3 \times 3$, $\text{batch_size}(x) = 3$

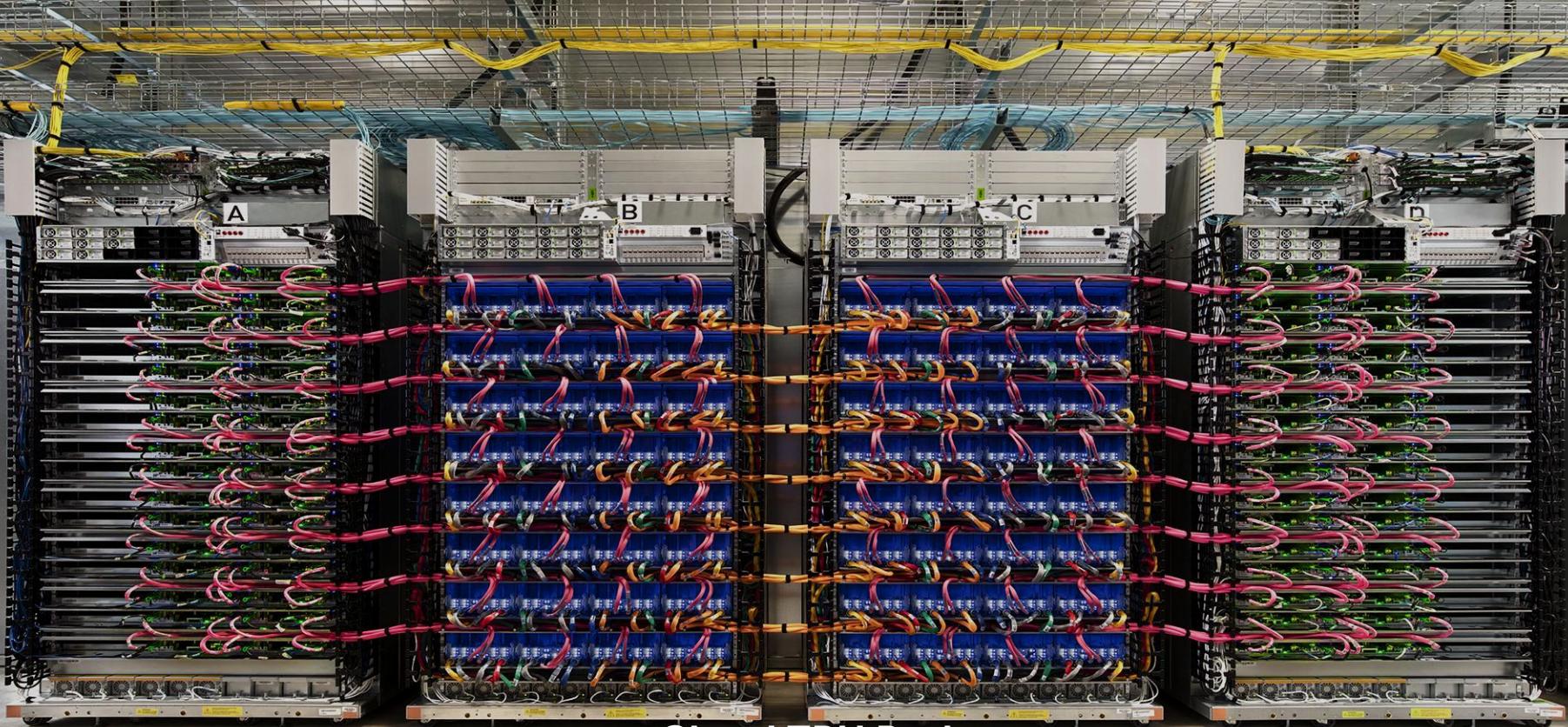


Matrix Unit Systolic Array

Computing $y = Wx$

with $W = 3 \times 3$, $\text{batch_size}(x) = 3$





Cloud TPU Pod
64 Cloud TPUs in 2-D toroidal mesh
11.5 petaFLOPS
4 terabytes of HBM memory

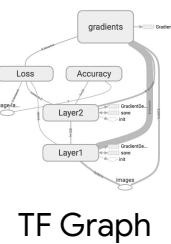
Accelerated Linear Algebra (XLA)

- JIT / AOT compiler for linear algebra
- Targets multiple backends, e.g. CPUs, GPUs, and TPUs
- Compiler, runtime, and accelerator-specific optimizer

The life of a neural network:



TF Estimator code

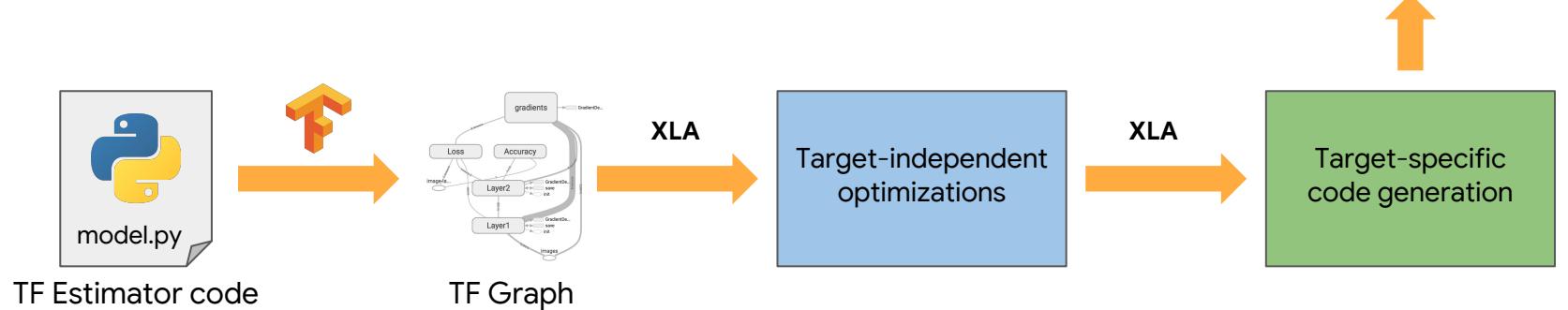


TF Graph

Accelerated Linear Algebra (XLA)

- JIT / AOT compiler for linear algebra
- Targets multiple backends, e.g. CPUs, GPUs, and TPUs
- Compiler, runtime, and accelerator-specific optimizer

The life of a neural network:

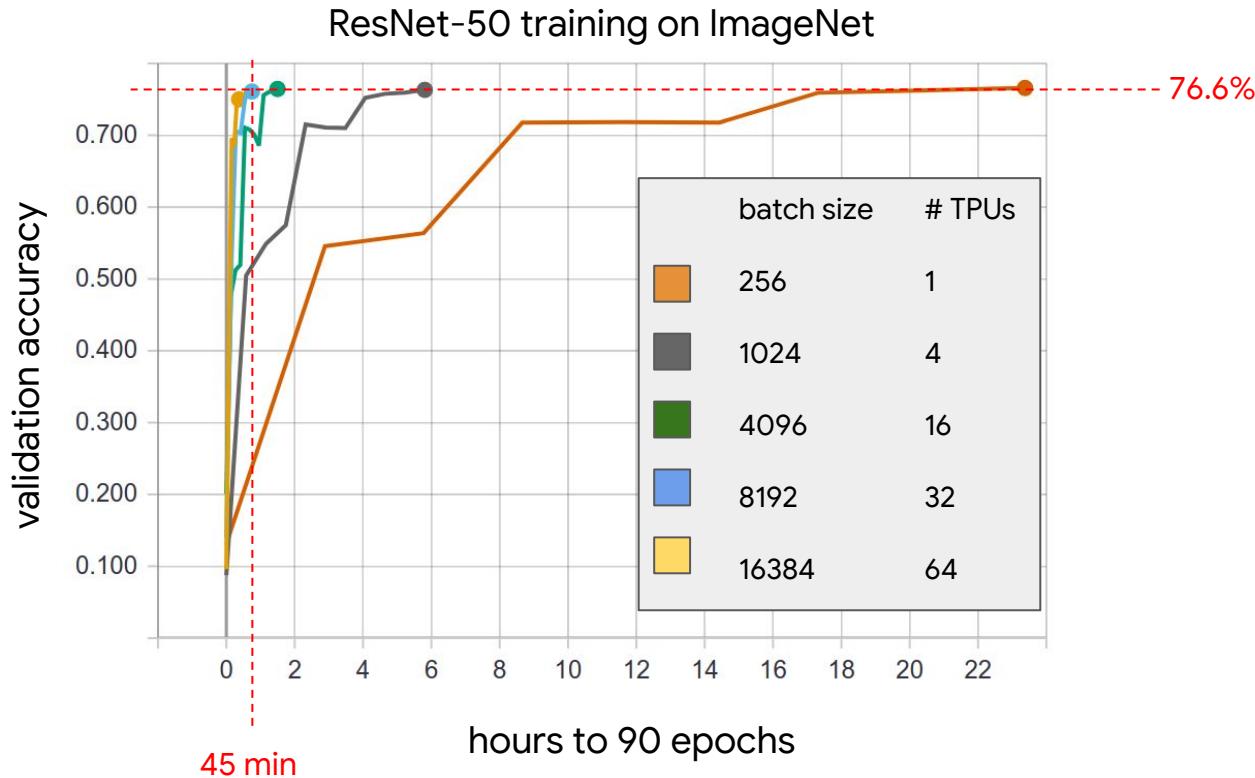


Large batch training

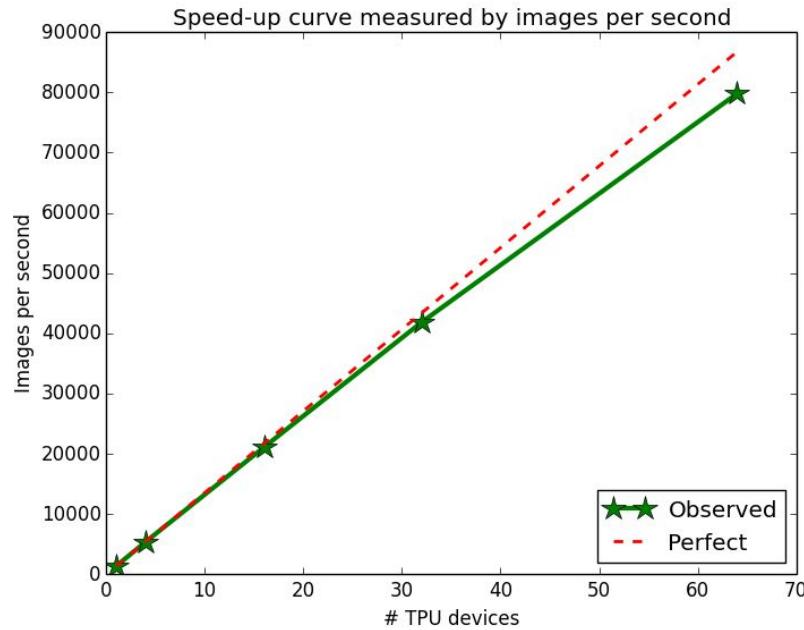
- Understanding generalization gap ([2016 N. Keskar et. al.](#), [2017 E. Hoffer et. al.](#))
- Relationship of batch size and noise scale ([2018 S. Smith et. al.](#))
- Learning rate scaling and schedule ([2017 P. Goyal et. al.](#))
- New optimizers
 - K-FAC*: approximate Fisher information matrix ([2015 J. Martens](#))
 - Neumann*: approximate inverse Hessian ([2018 S. Krishnan et. al.](#))
 - LARS: per-layer learning rate ([2018 Y. You et. al.](#))

* stick around after this talk to hear more about these!

Experiments



Experiments



Experiments

# of TPU devices	Batch size	Time to 90 epochs	Accuracy
32	8192	44.9 minutes	76.1%
64	8192	29.8 minutes	75.7%
64	16384	22.3 minutes	75.0%
64	65536	17.5 minutes	65.4%
64	8192 → 16384 ^[1]	29.5 minutes	76.1%

Only change between different runs is batch size (linearly scale LR) and hardware,
no model changes or hyperparameter re-tuning!

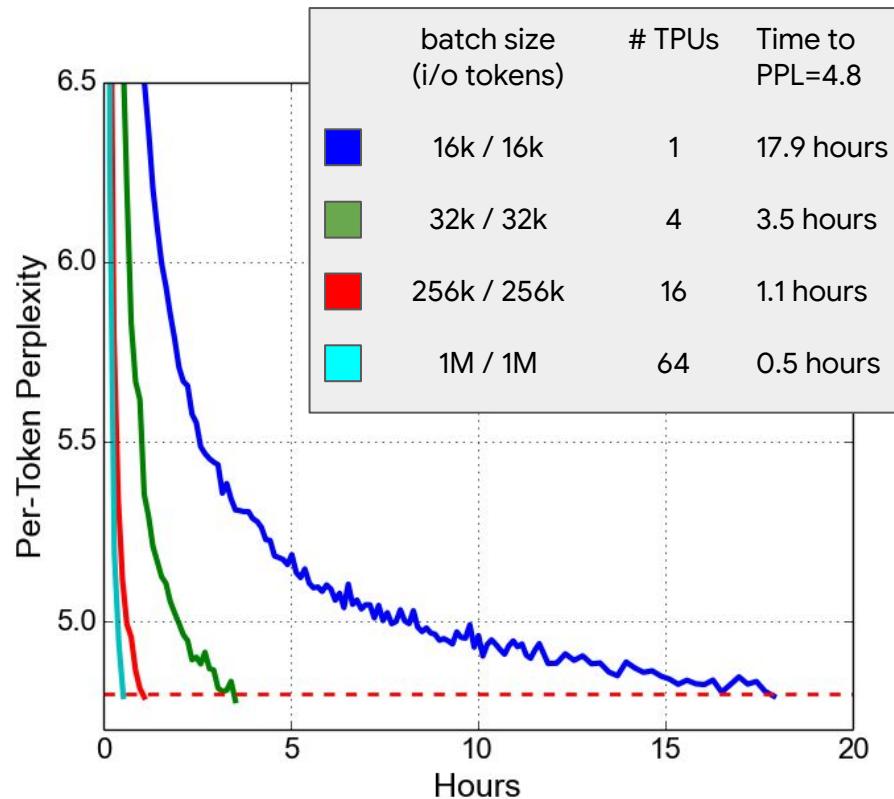
[1] Don't Decay the Learning Rate, Increase the
Batch Size ([2018 S. Smith et. al](#))

More than just ImageNet

Transformer model from "Attention is All You Need" ([2017 A. Vaswani et. al.](#))

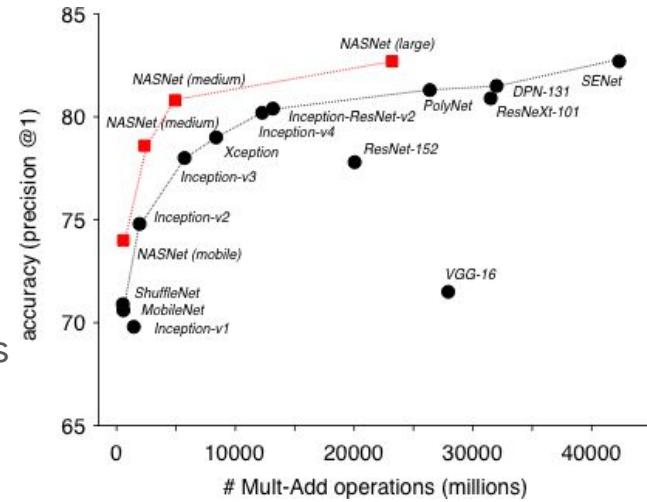
WMT'14 English-German translation task

Adam optimizer - same learning rate schedule across configurations



Implications

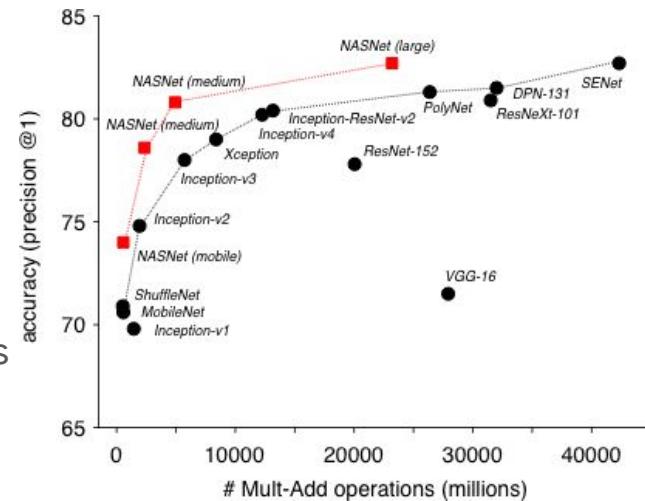
- Faster training enables neural architecture search
 - Reinforcement learning architectures beat existing models in accuracy and cost [1]



[1] Learning Transferable Architectures for Scalable Image Recognition (2017 B. Zoph et. al)

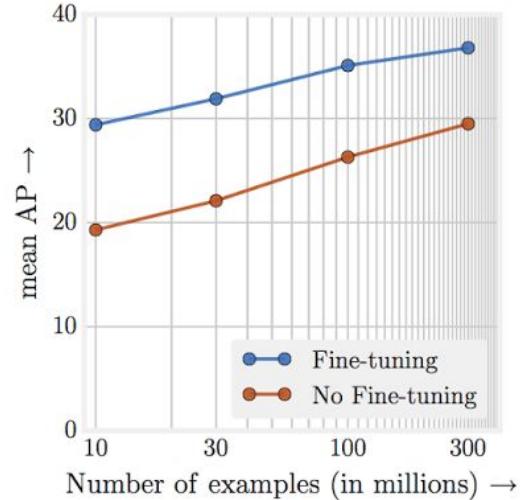
Implications

- Faster training enables neural architecture search
 - Reinforcement learning architectures beat existing models in accuracy and cost [1]
- What's the "new ImageNet"?
 - Full ImageNet (14M), Open Images (9M), YouTube-8M
 - Performance increases logarithmically with data [2]



[1] Learning Transferable Architectures for Scalable Image Recognition (2017 B. Zoph et. al)

[2] Revisiting Unreasonable Effectiveness of Data in Deep Learning Era (2017 C. Sun et. al)





Pieter-jan



Brennan



Sam



Jonathan



Sameer



Zak



Quoc



Bjarke



Noam



Naveen



Chris

Thank you!

chrisying@google.com

g.co/brain

g.co/tpusignup

