- Building Brains -

Parallel training strategies for large-scale deep learning

IBM Center for Open Source Data and AI Technologies (CODAIT)

Romeo Kienzler

IBM Disclaimer

THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION.

NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, OR SHALL HAVE THE EFFECT OF:

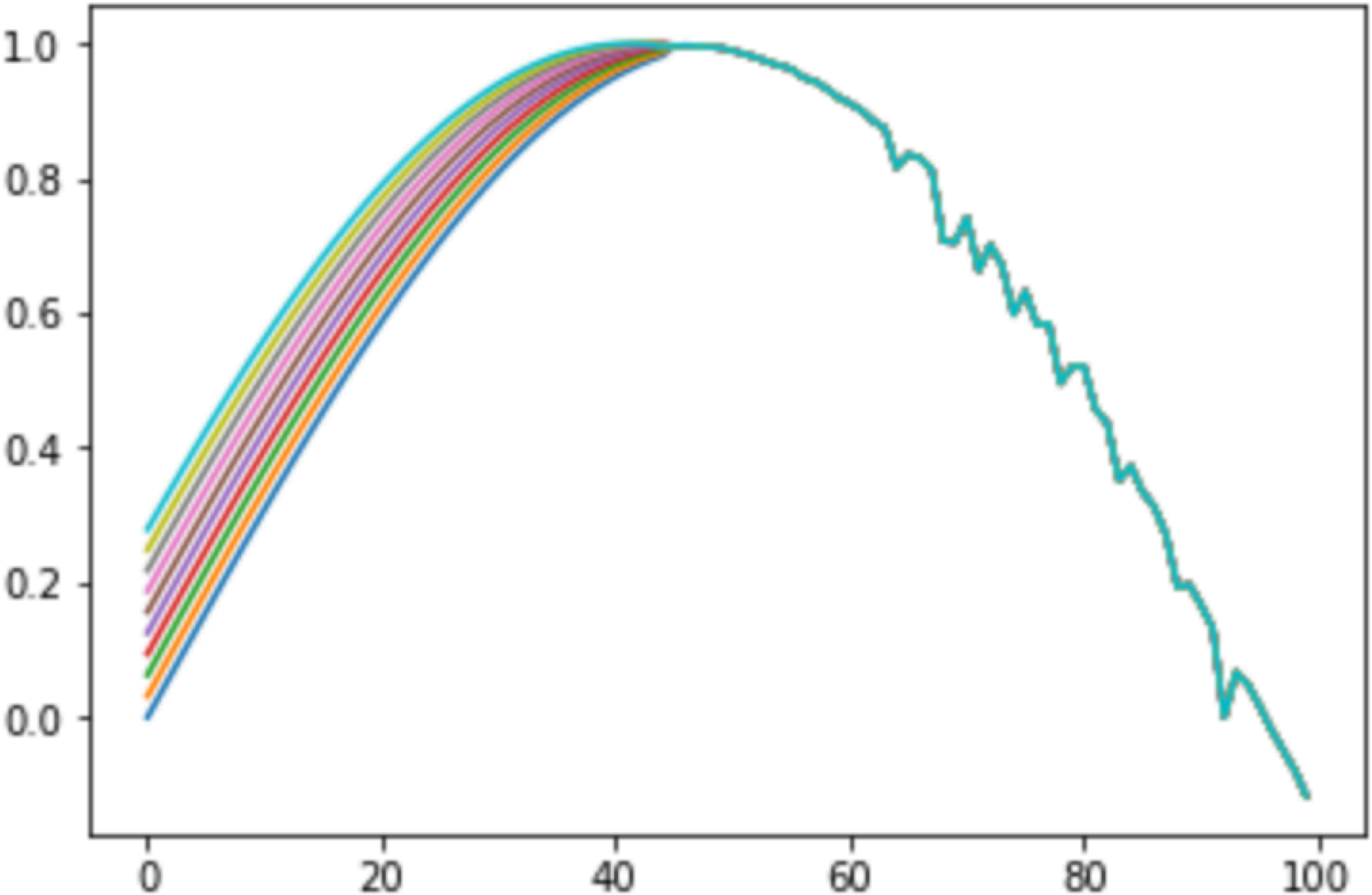
CREATING ANY WARRANTY OR REPRESENTATION FROM IBM (OR ITS AFFILIATES OR ITS OR THEIR SUPPLIERS AND/OR LICENSORS); OR

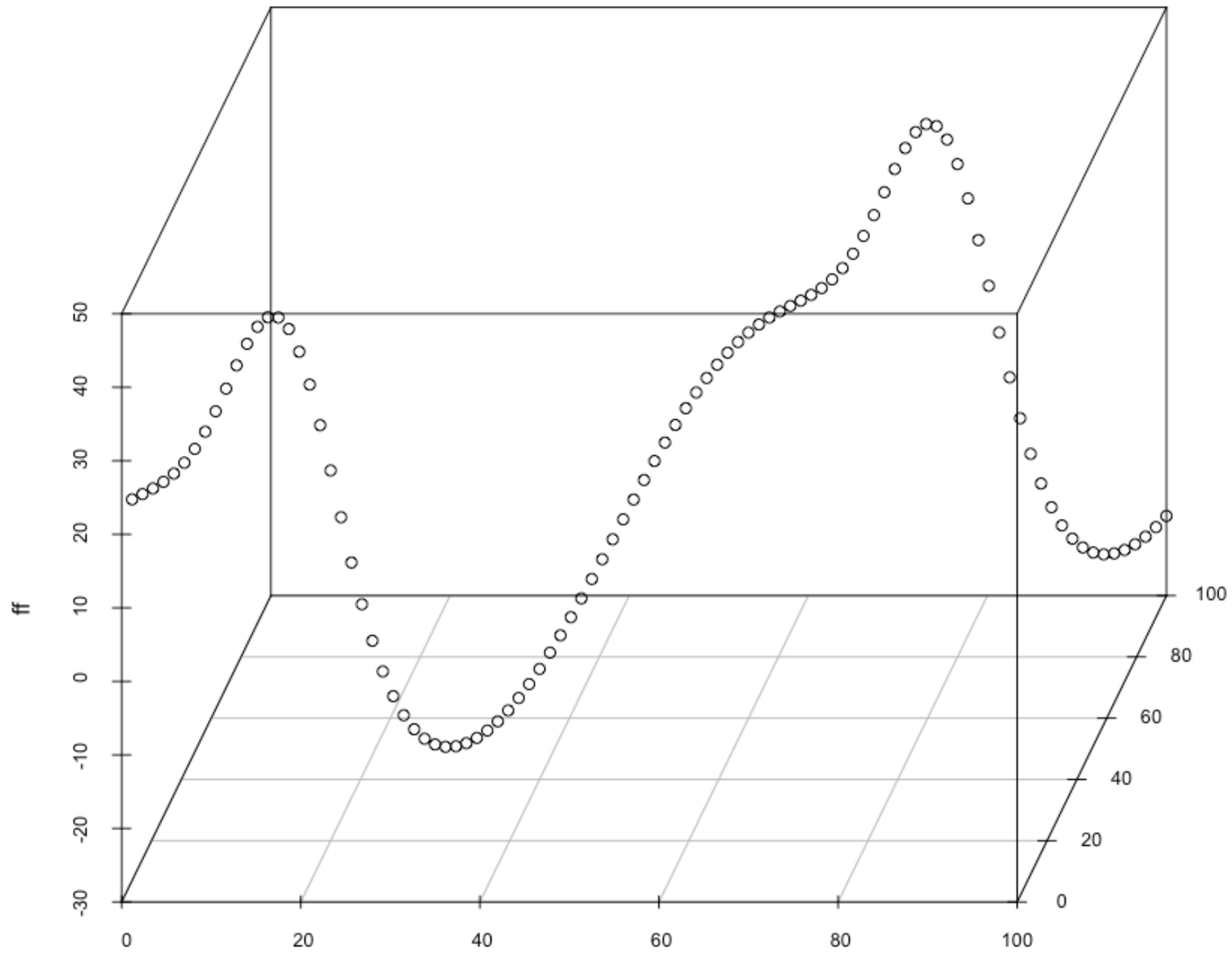
ALTERING THE TERMS AND CONDITIONS OF THE APPLICABLE LICENSE AGREEMENT GOVERNING THE USE OF IBM SOFTWARE.

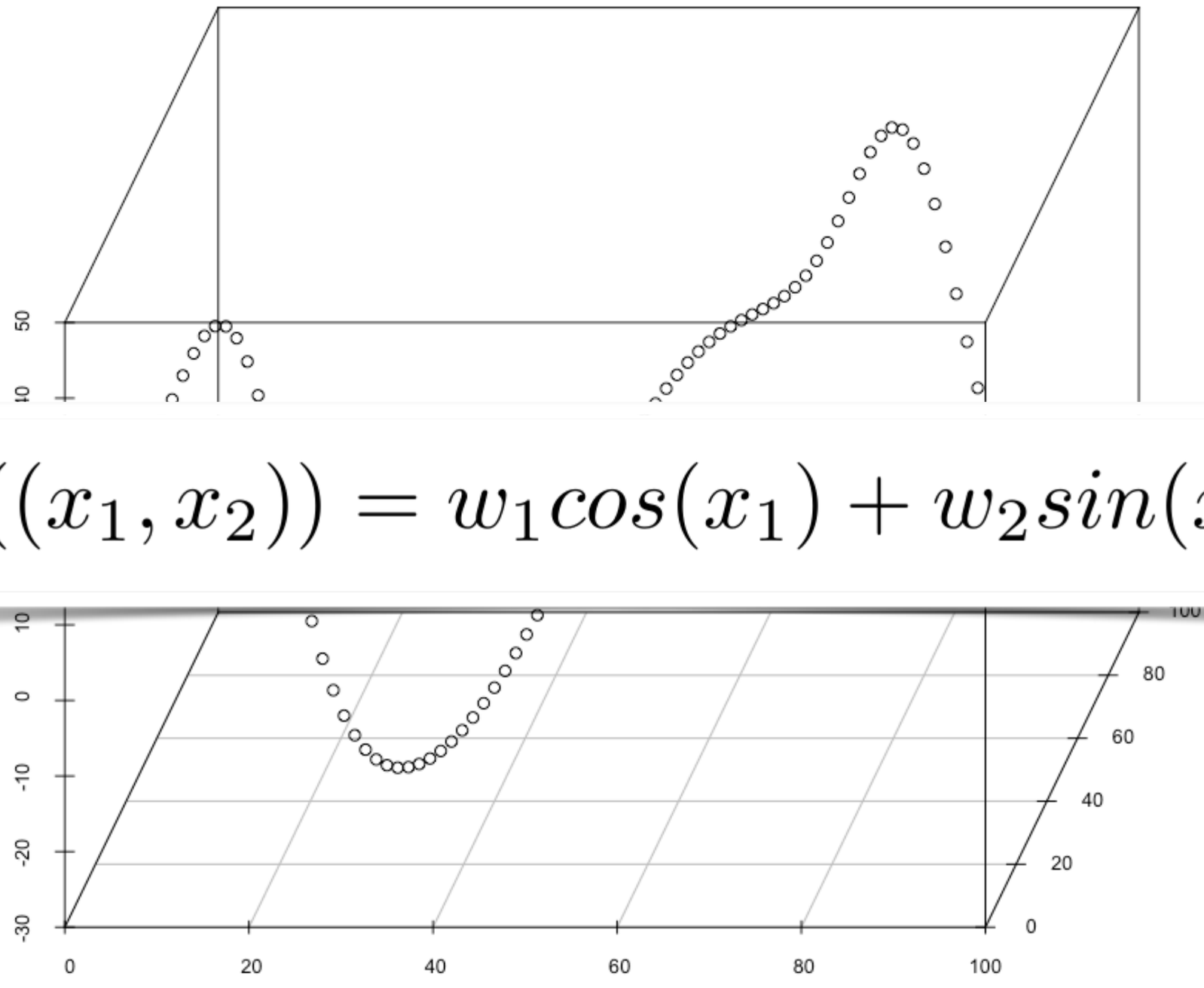
IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

We are looking forward to hearing your suggestions and feedback so that we can improve our solutions and products.

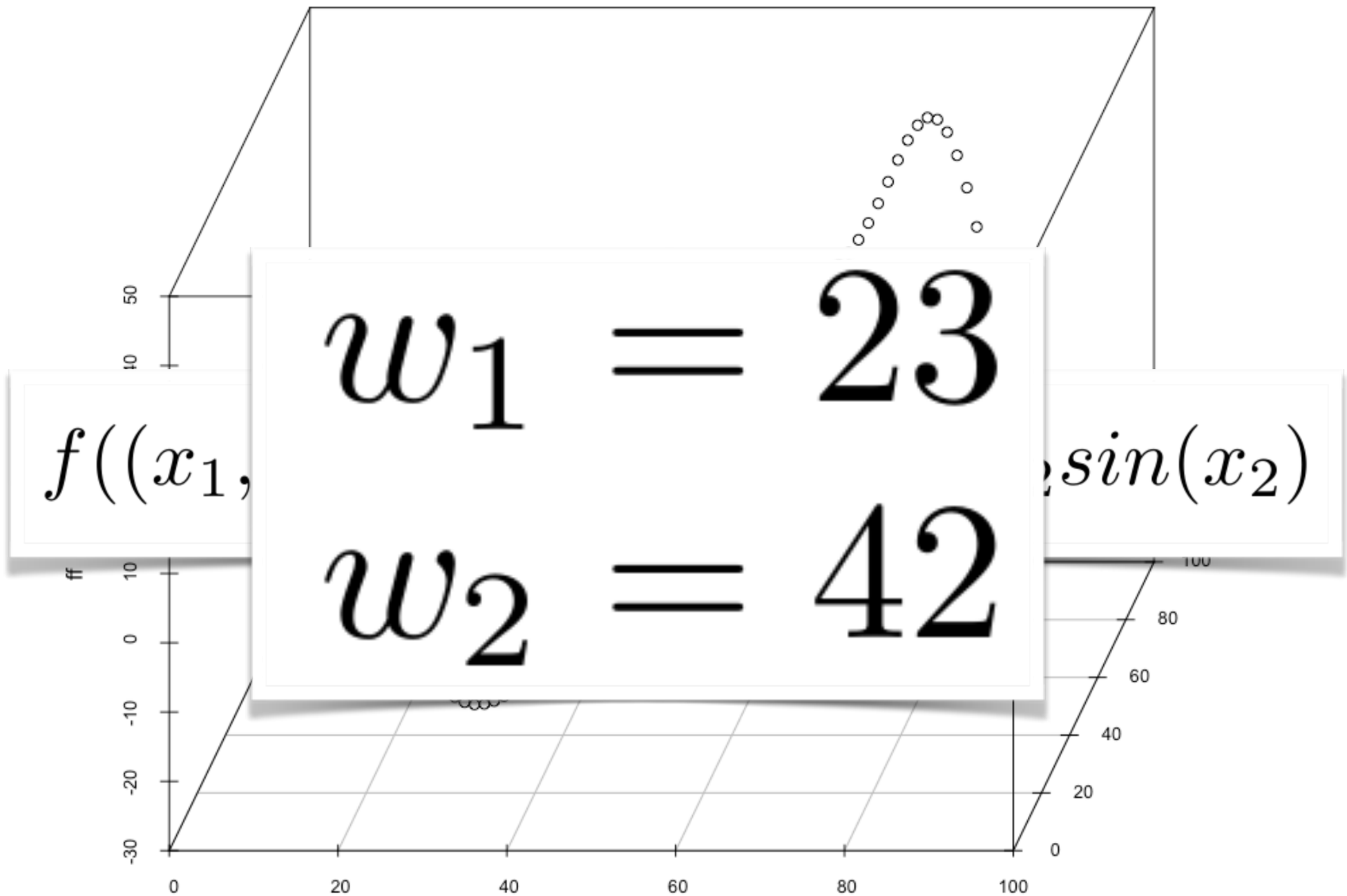
IBM shall be free to use for any purpose and without restriction any oral or written suggestions or feedback that you provide to IBM. By providing IBM with any information or material, you grant IBM an unrestricted, irrevocable license to copy, reproduce, publish, upload, post, transmit, distribute, publicly display, perform, modify, create derivative works from, and otherwise freely use, those materials or information. You also agree that IBM is free to use any ideas, concepts, know-how, or techniques that you provide us for any purpose.



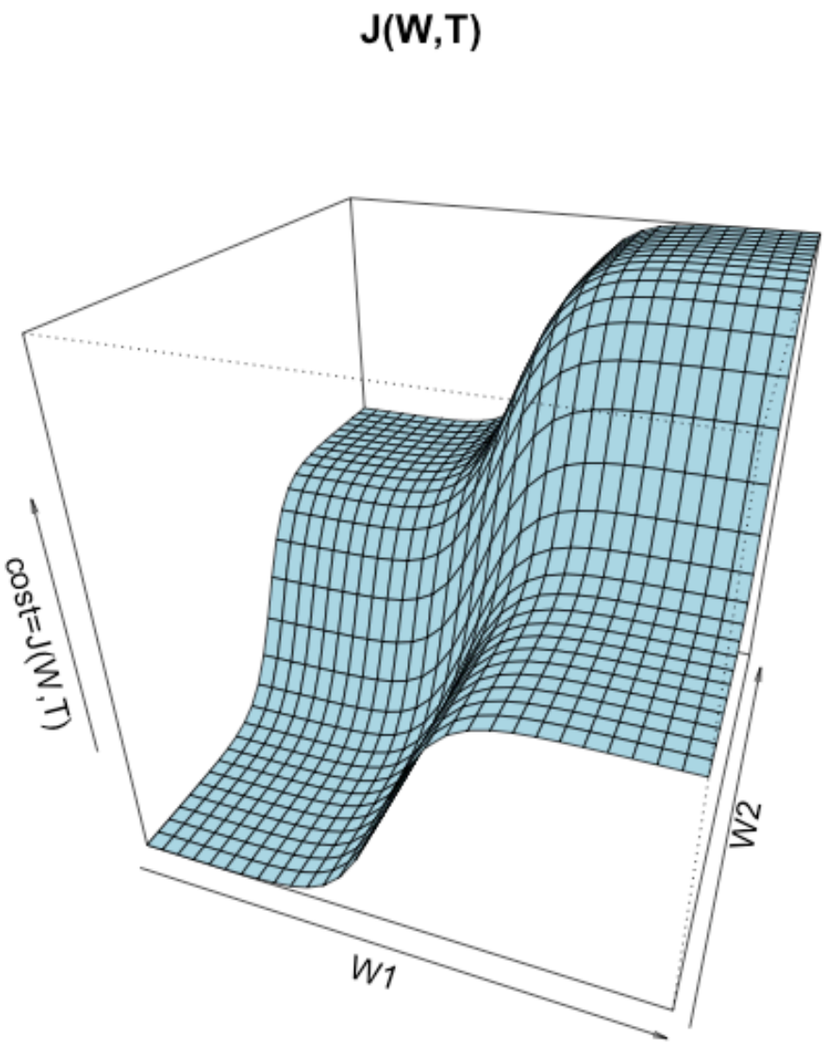




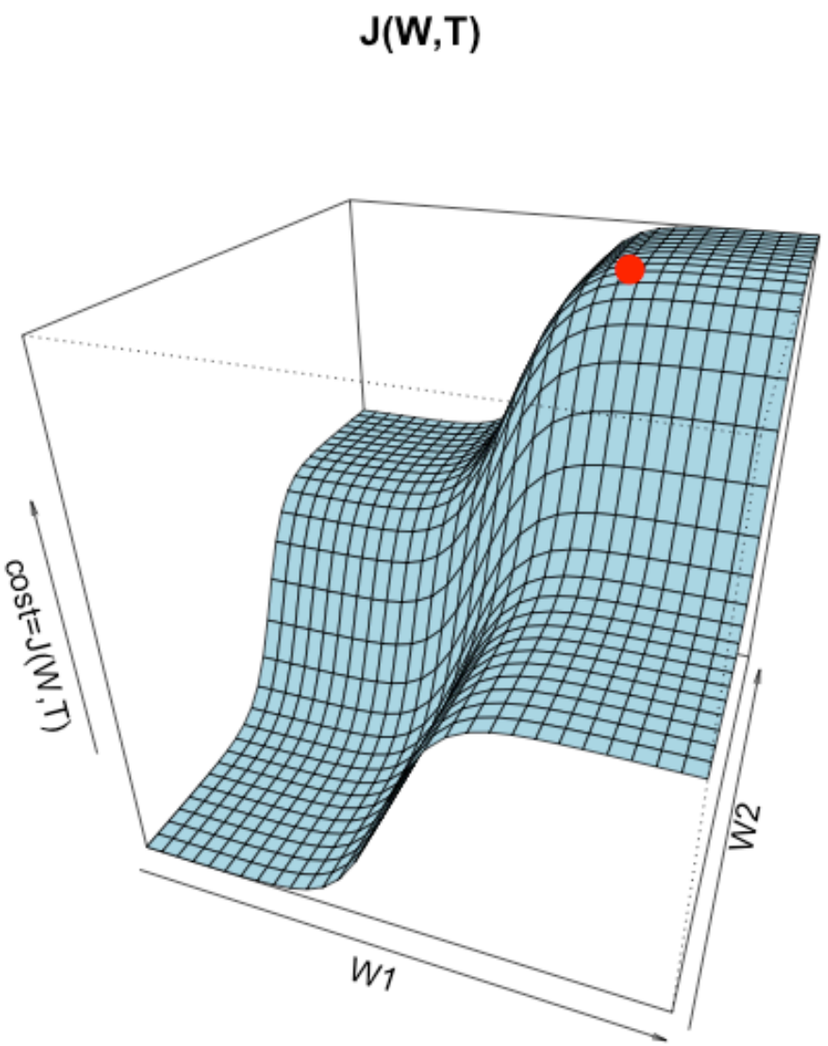
$$f((x_1, x_2)) = w_1 cos(x_1) + w_2 sin(x_2)$$



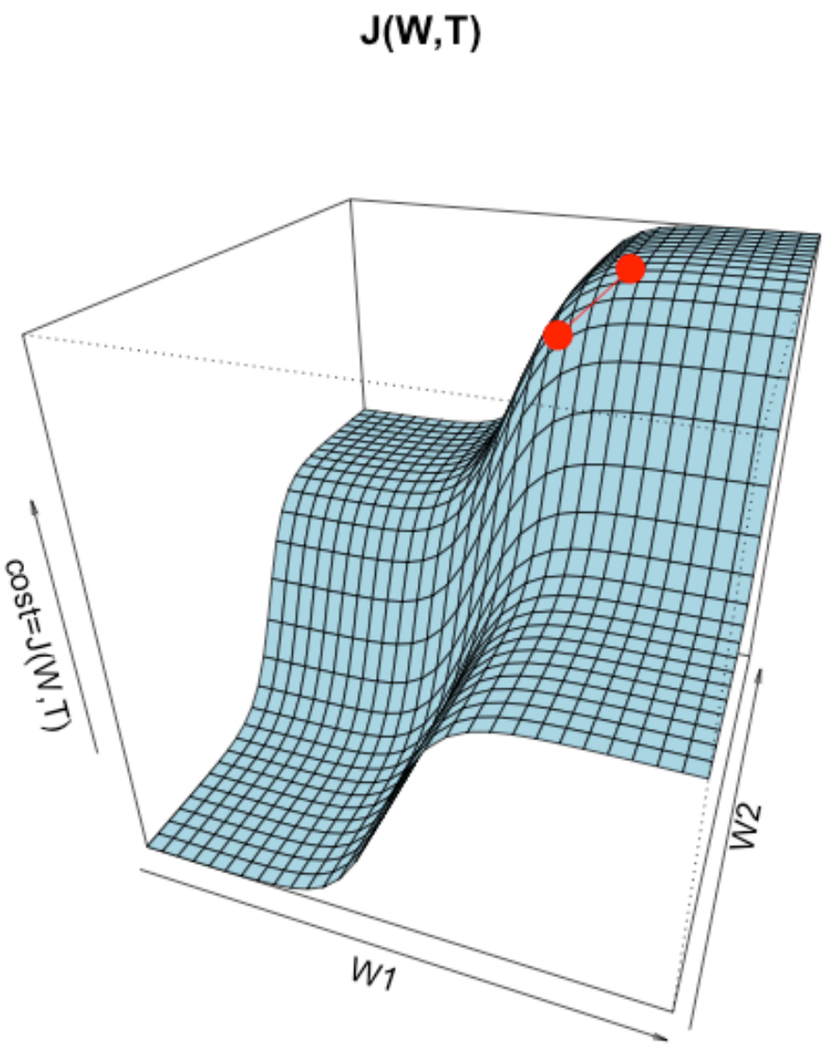
gradient descent



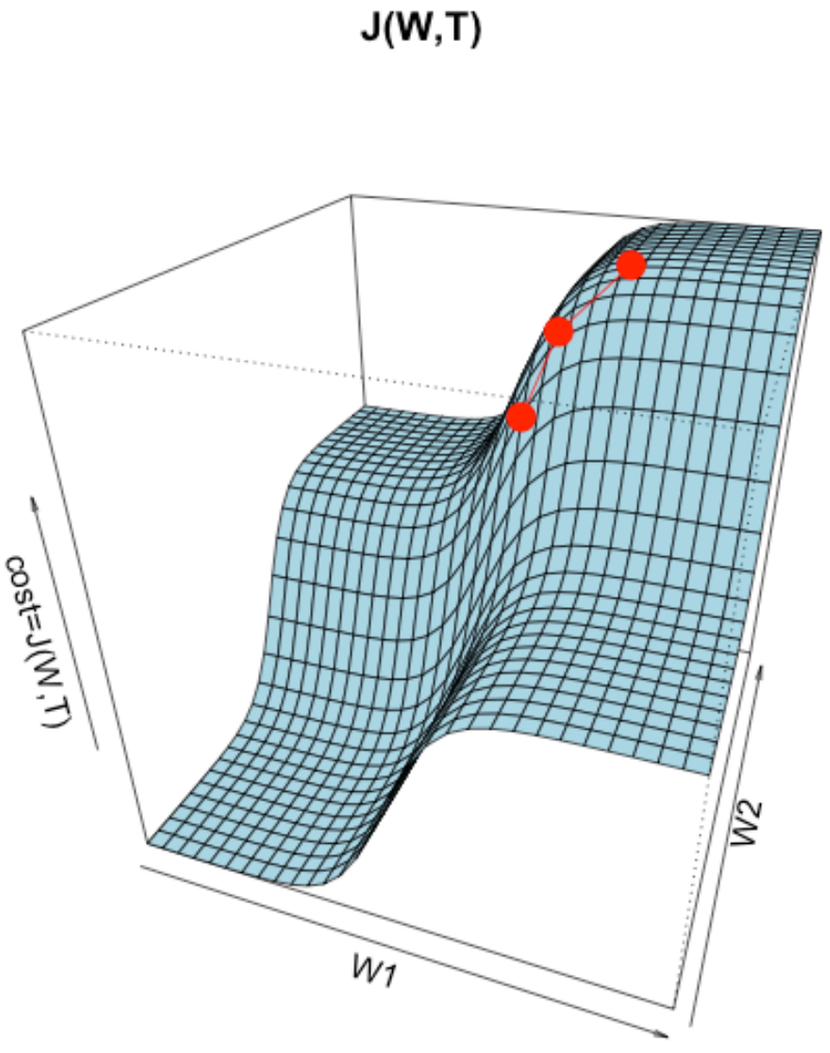
gradient descent



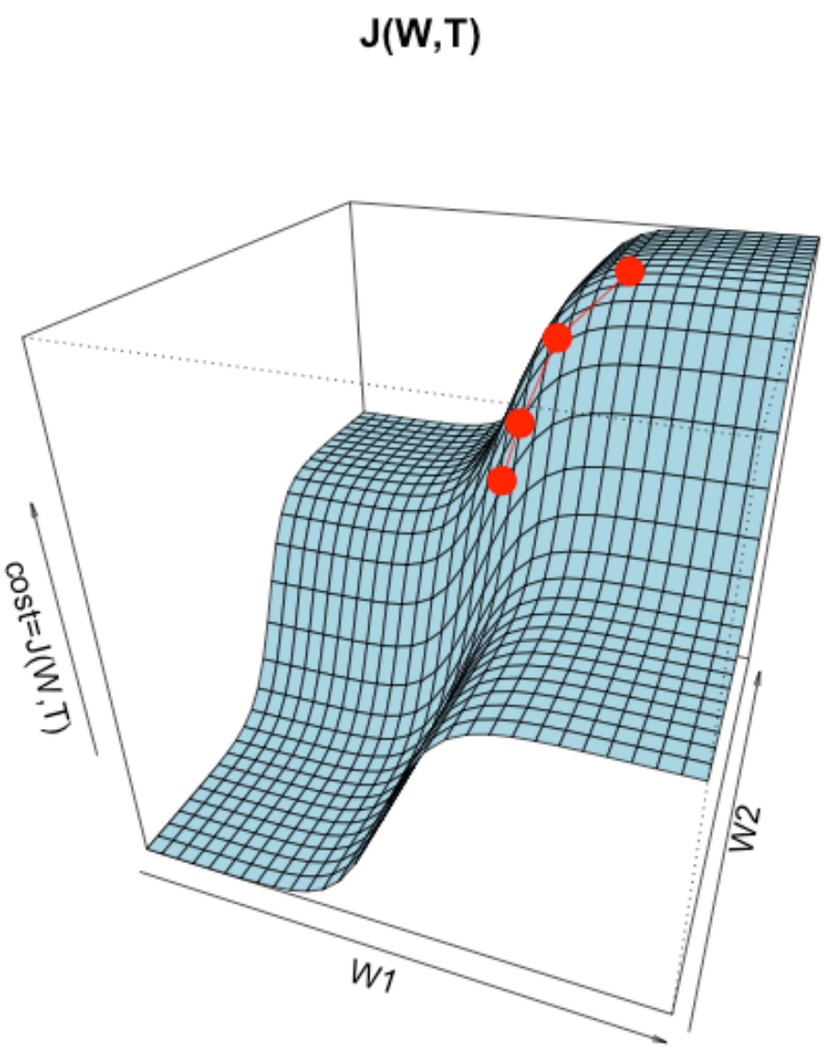
gradient descent



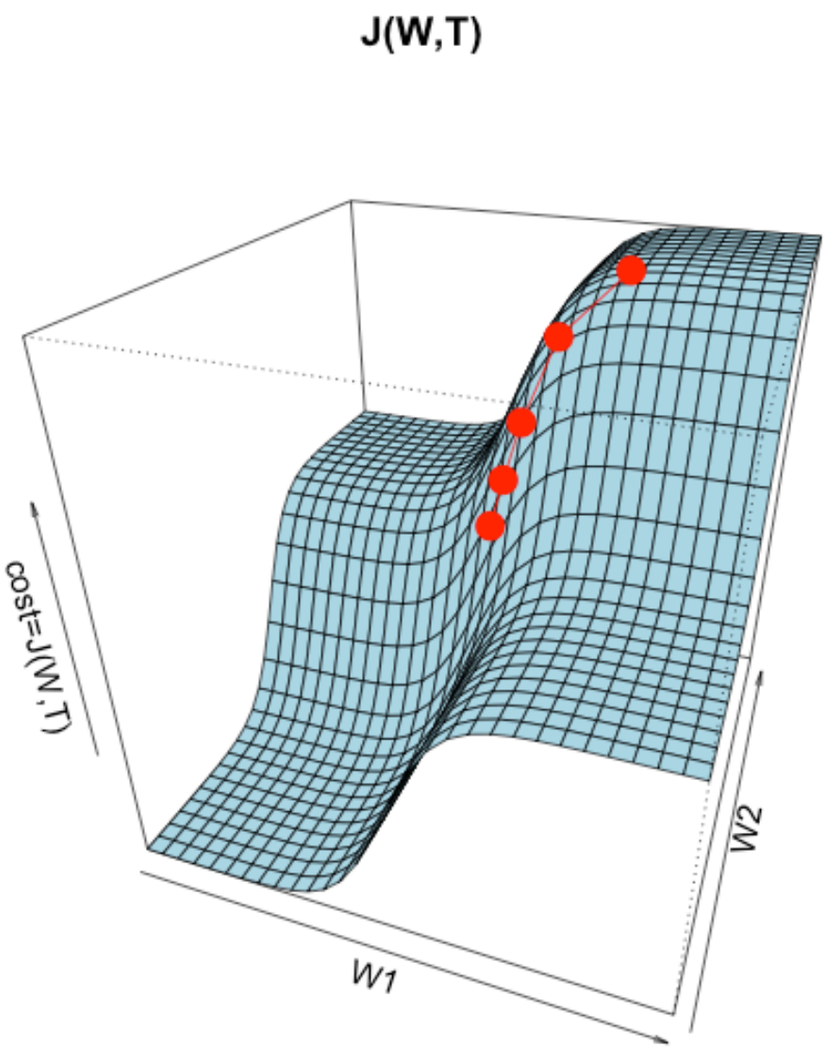
gradient descent



gradient descent



gradient descent

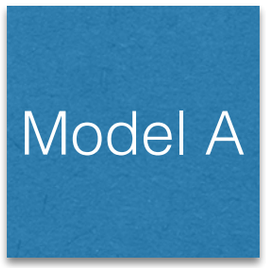


parallelisation

- inter-model parallelism
- data parallelism
- intra-model parallelism
- pipelined parallelism

inter-model parallelism

aka. hyper parameter space exploration / tuning

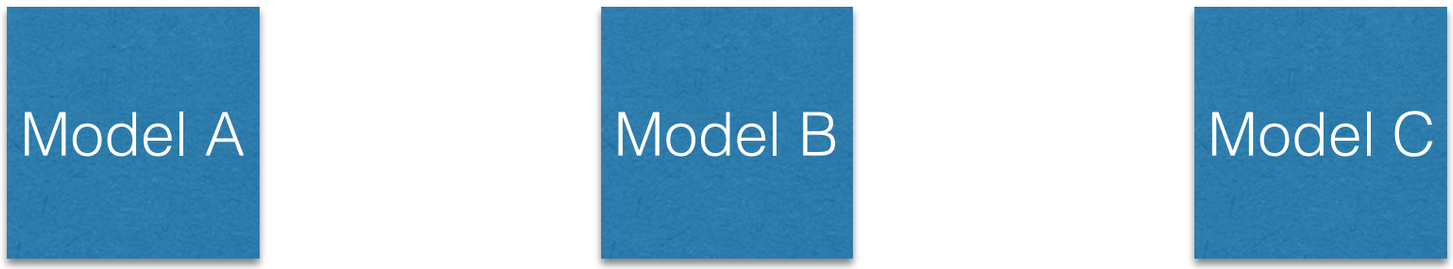


inter-model parallelism

aka. hyper parameter space exploration / tuning

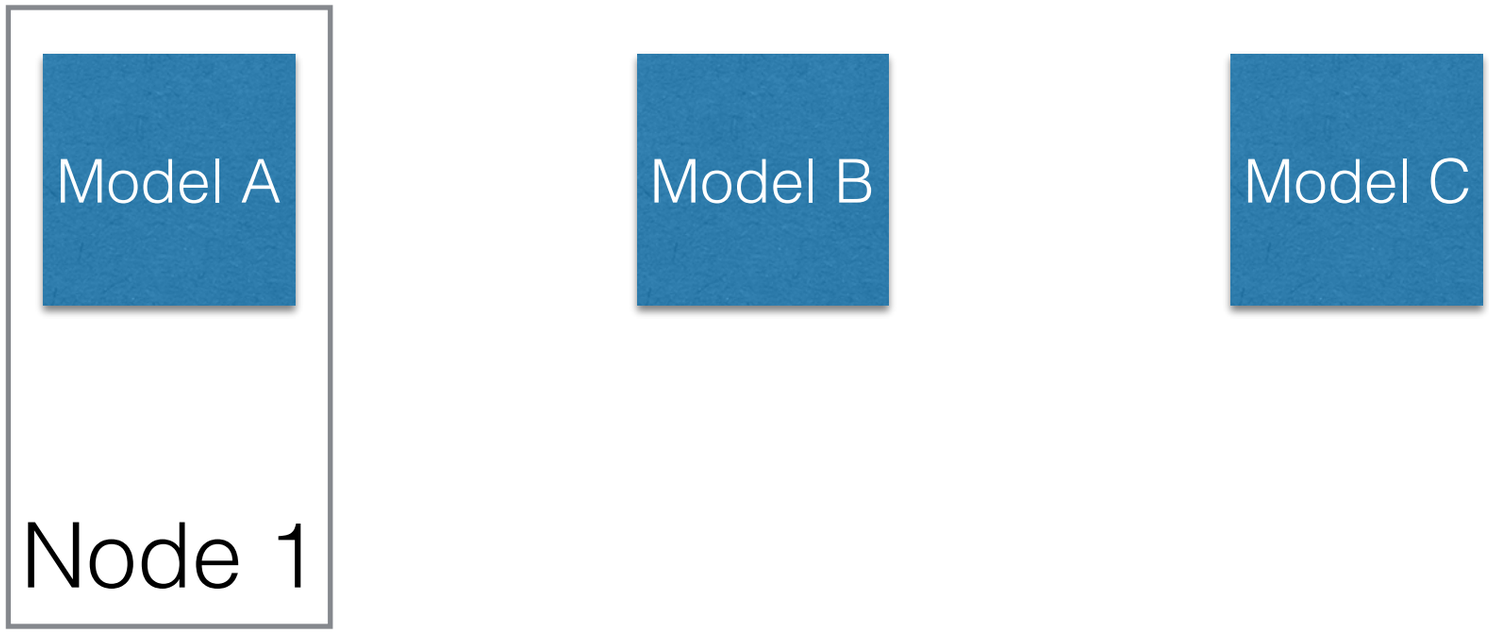
inter-model parallelism

aka. hyper parameter space exploration / tuning



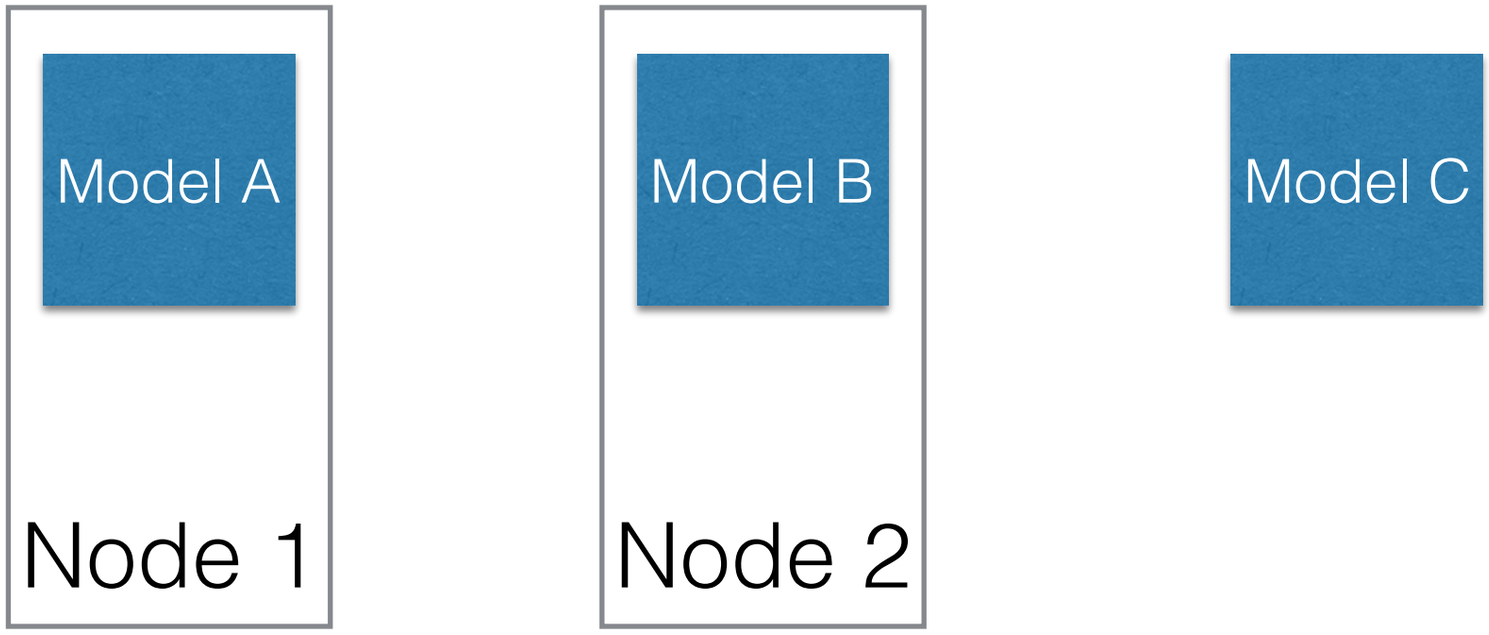
inter-model parallelism

aka. hyper parameter space exploration / tuning



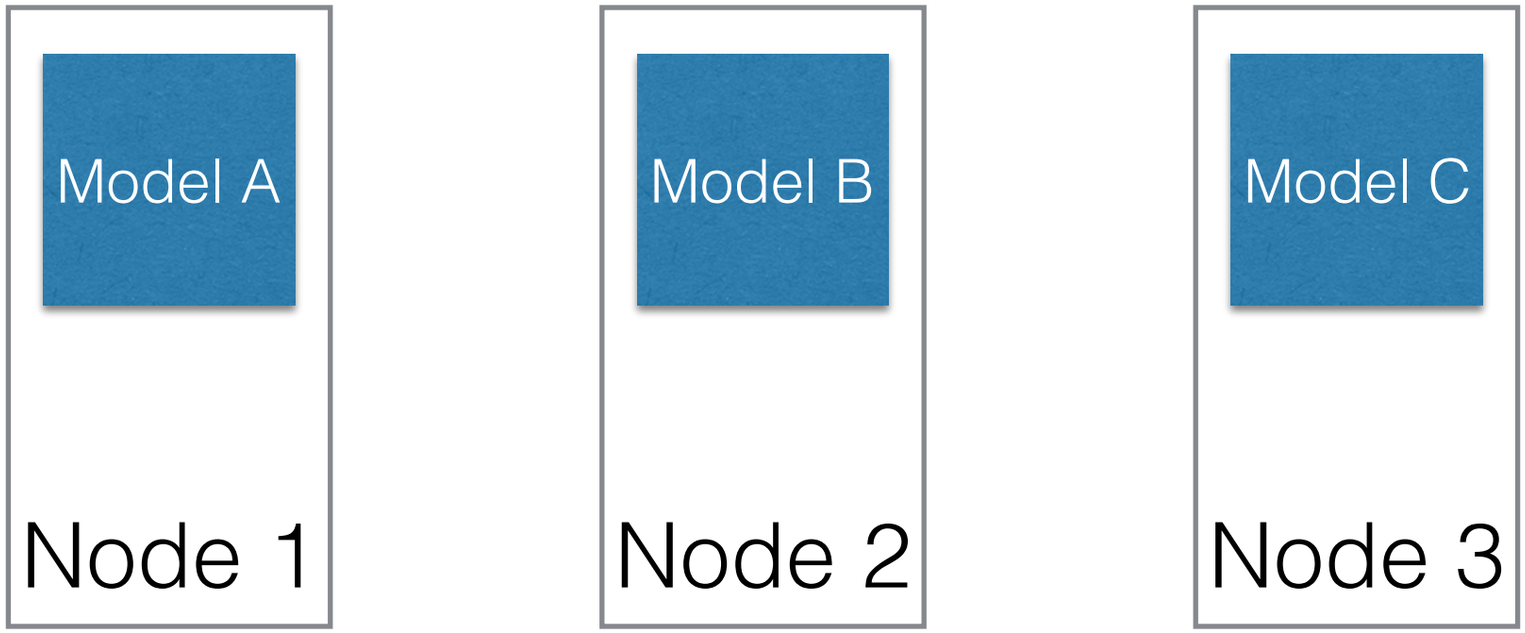
inter-model parallelism

aka. hyper parameter space exploration / tuning



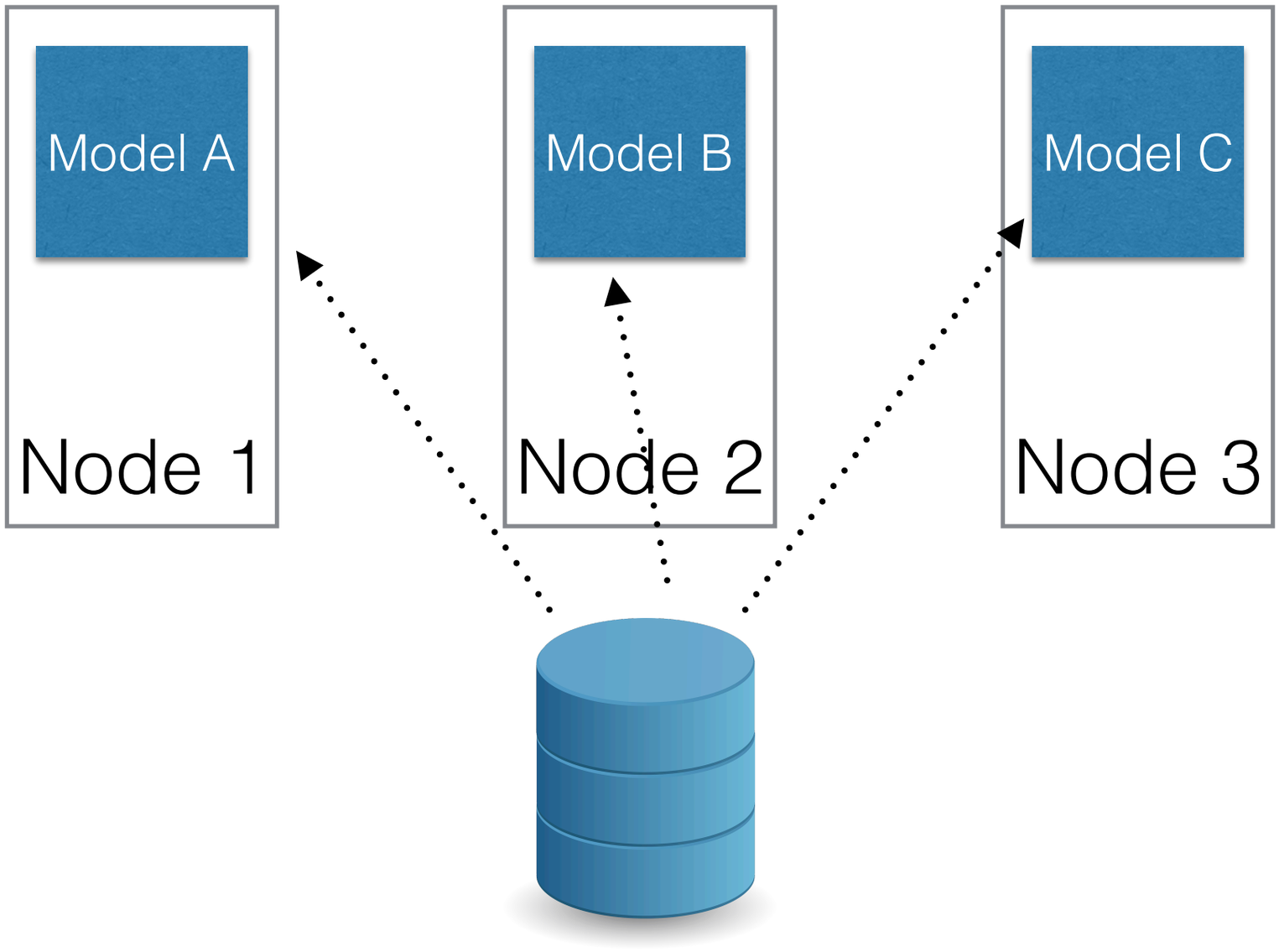
inter-model parallelism

aka. hyper parameter space exploration / tuning



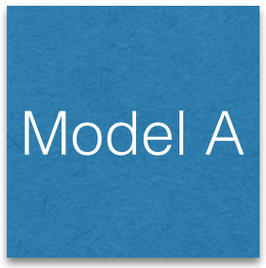
inter-model parallelism

aka. hyper parameter space exploration / tuning



data parallelism

aka. “Jeff Dean style” parameter averaging

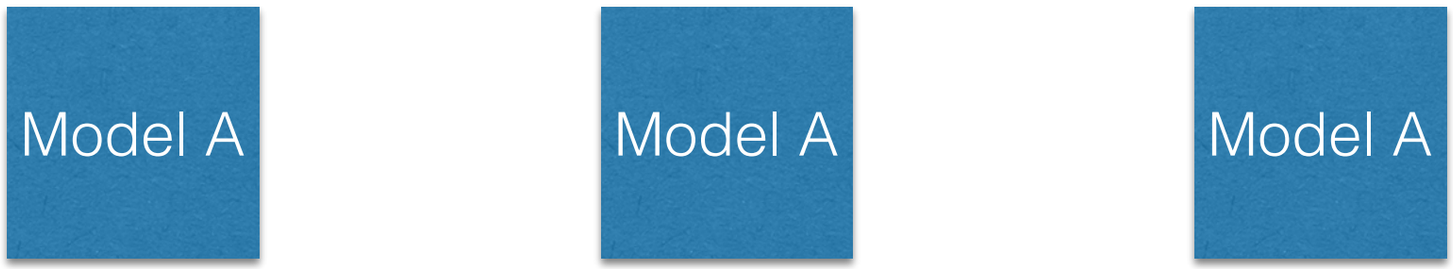


data parallelism

aka. “Jeff Dean style” parameter averaging

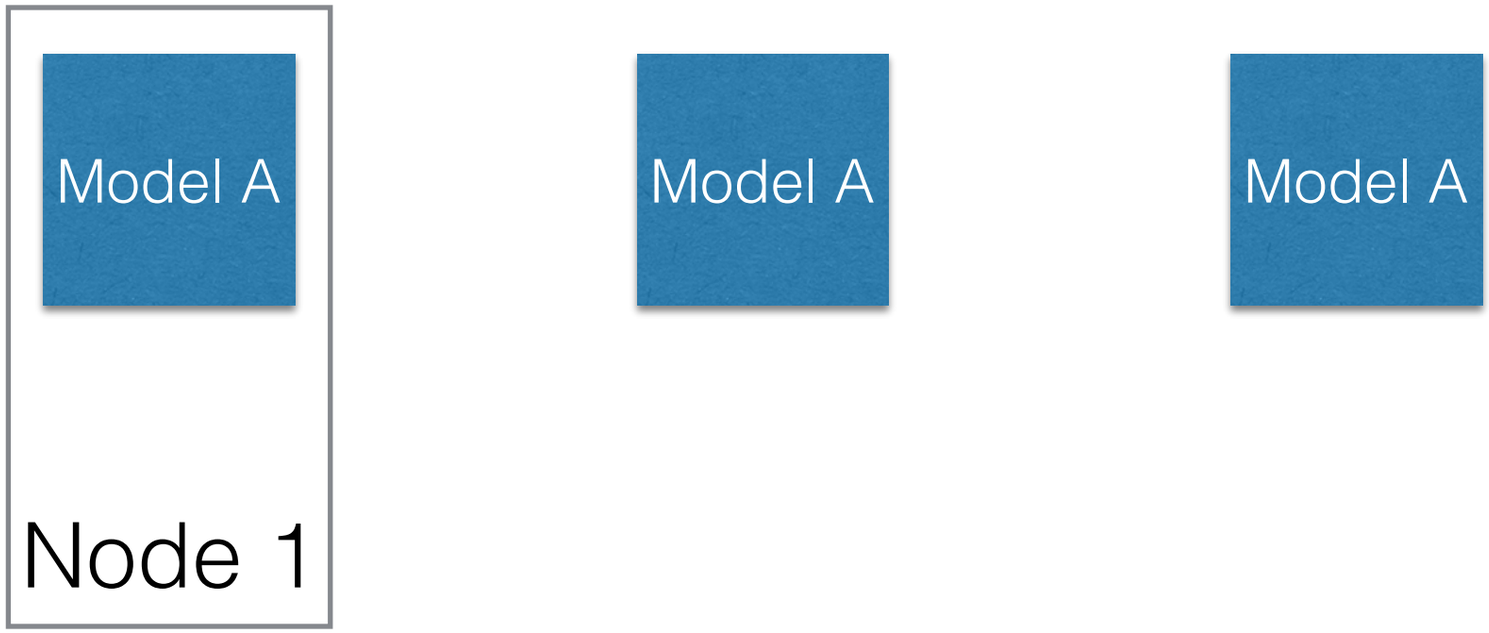
data parallelism

aka. “Jeff Dean style” parameter averaging



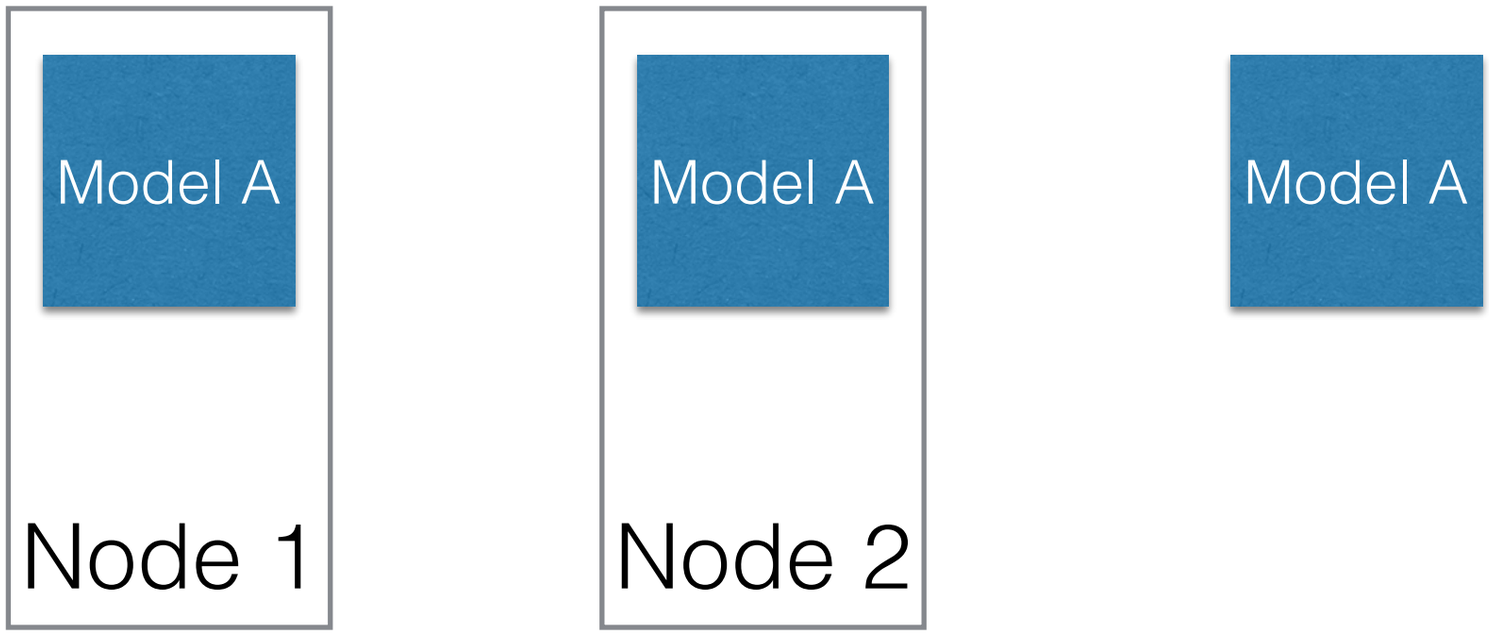
data parallelism

aka. “Jeff Dean style” parameter averaging



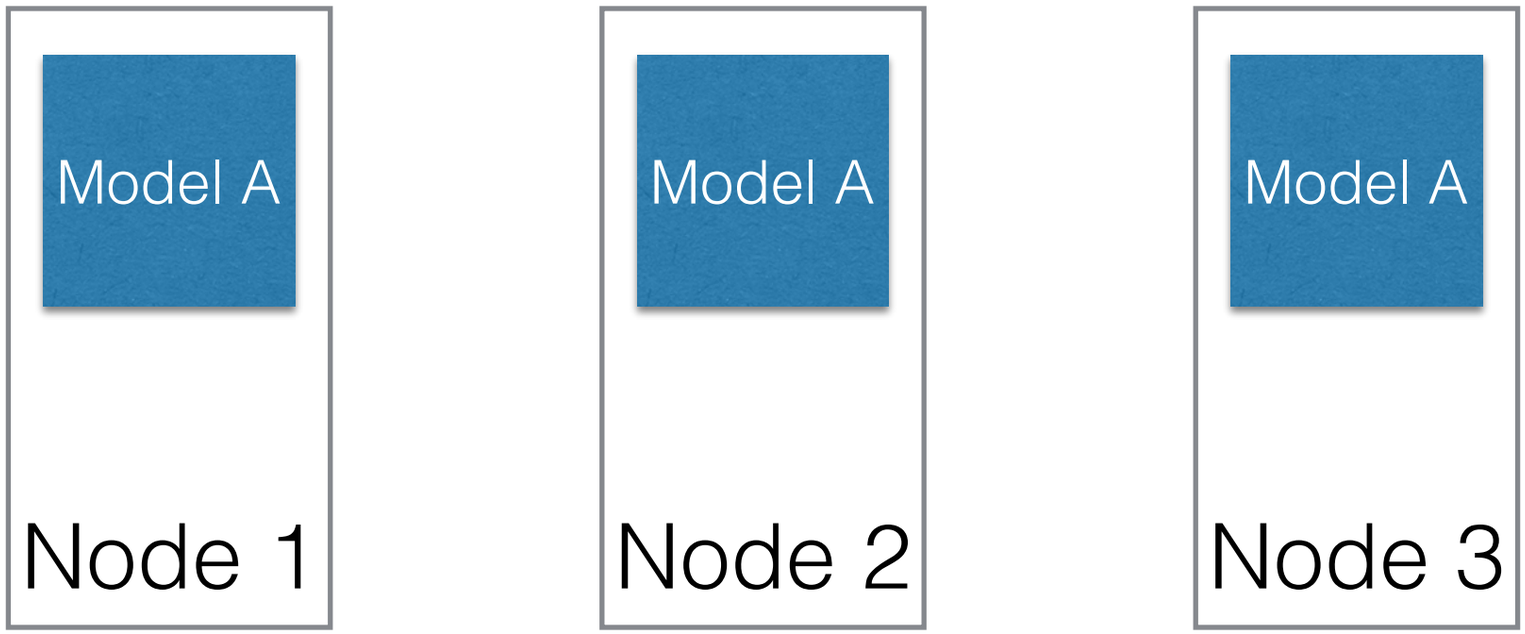
data parallelism

aka. “Jeff Dean style” parameter averaging



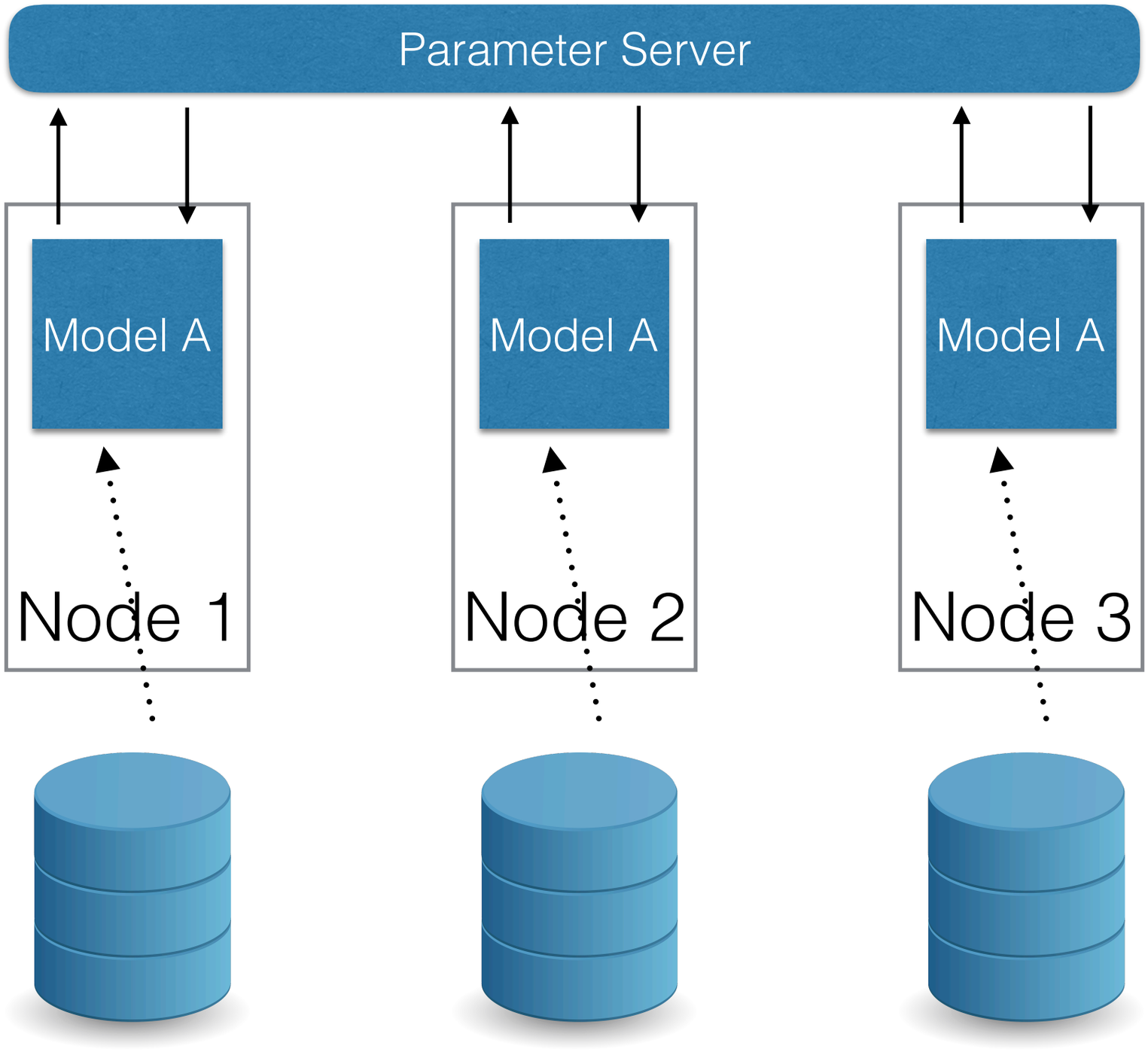
data parallelism

aka. “Jeff Dean style” parameter averaging



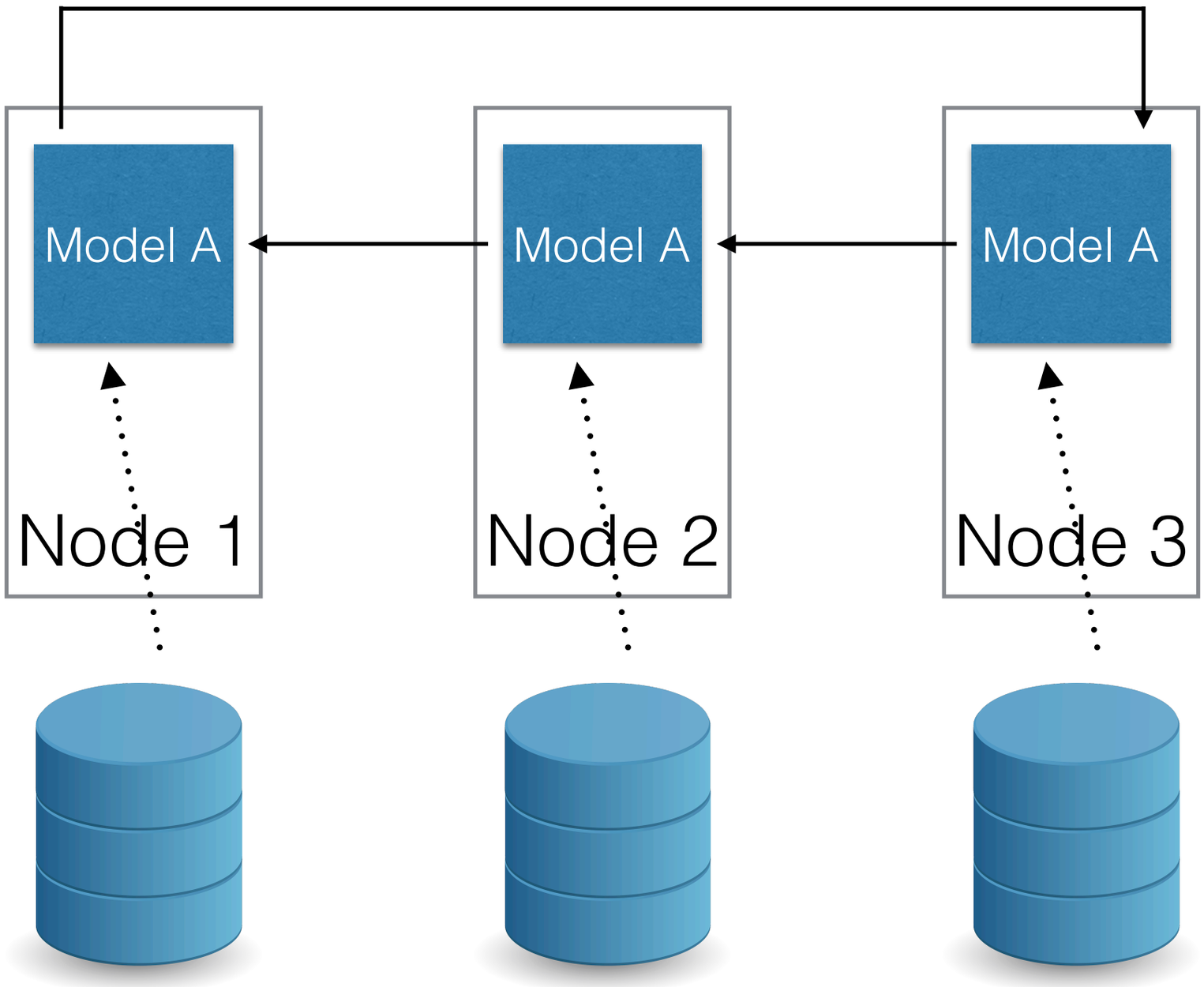
data parallelism

aka. “Jeff Dean style” parameter averaging

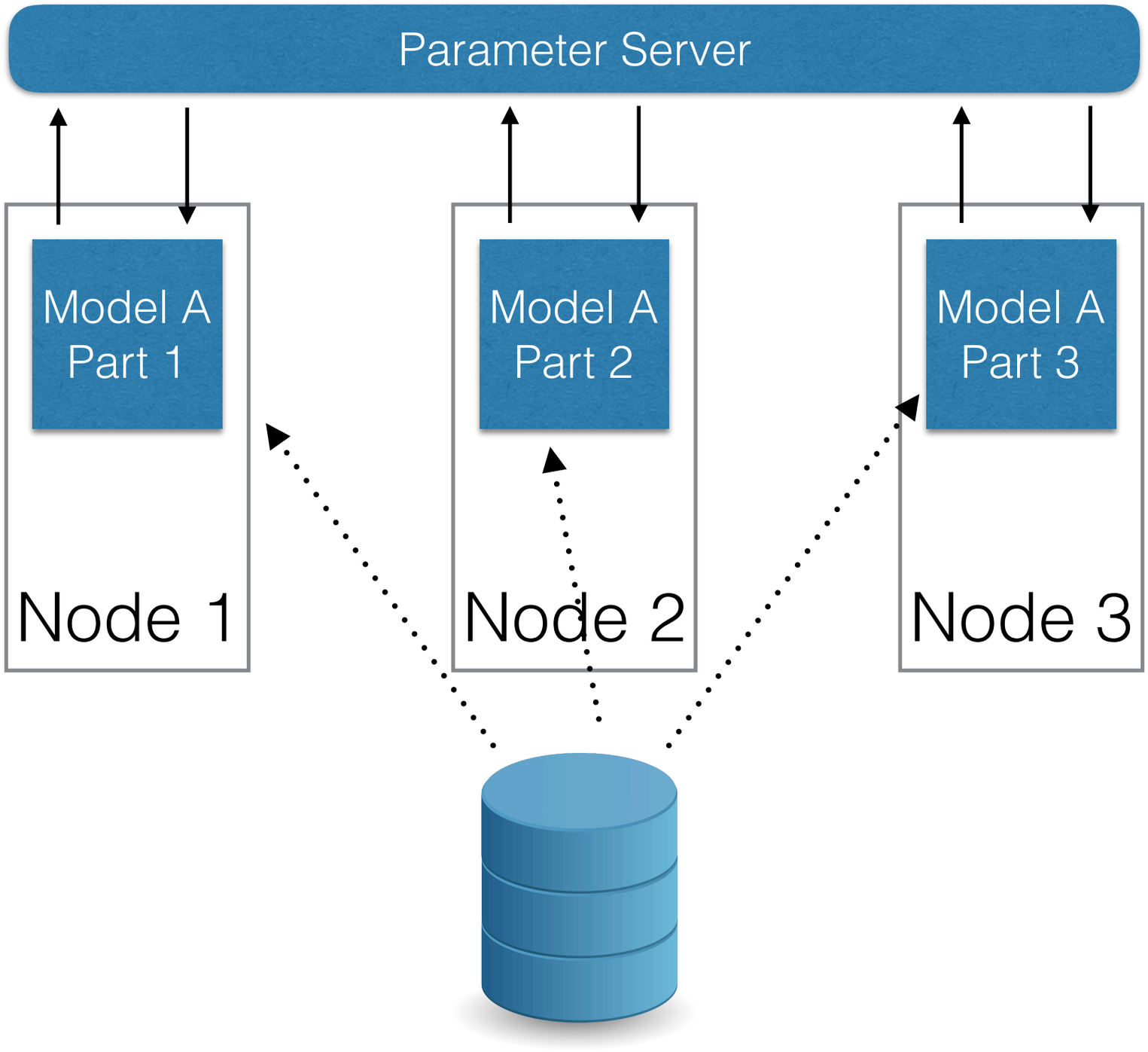


data parallelism

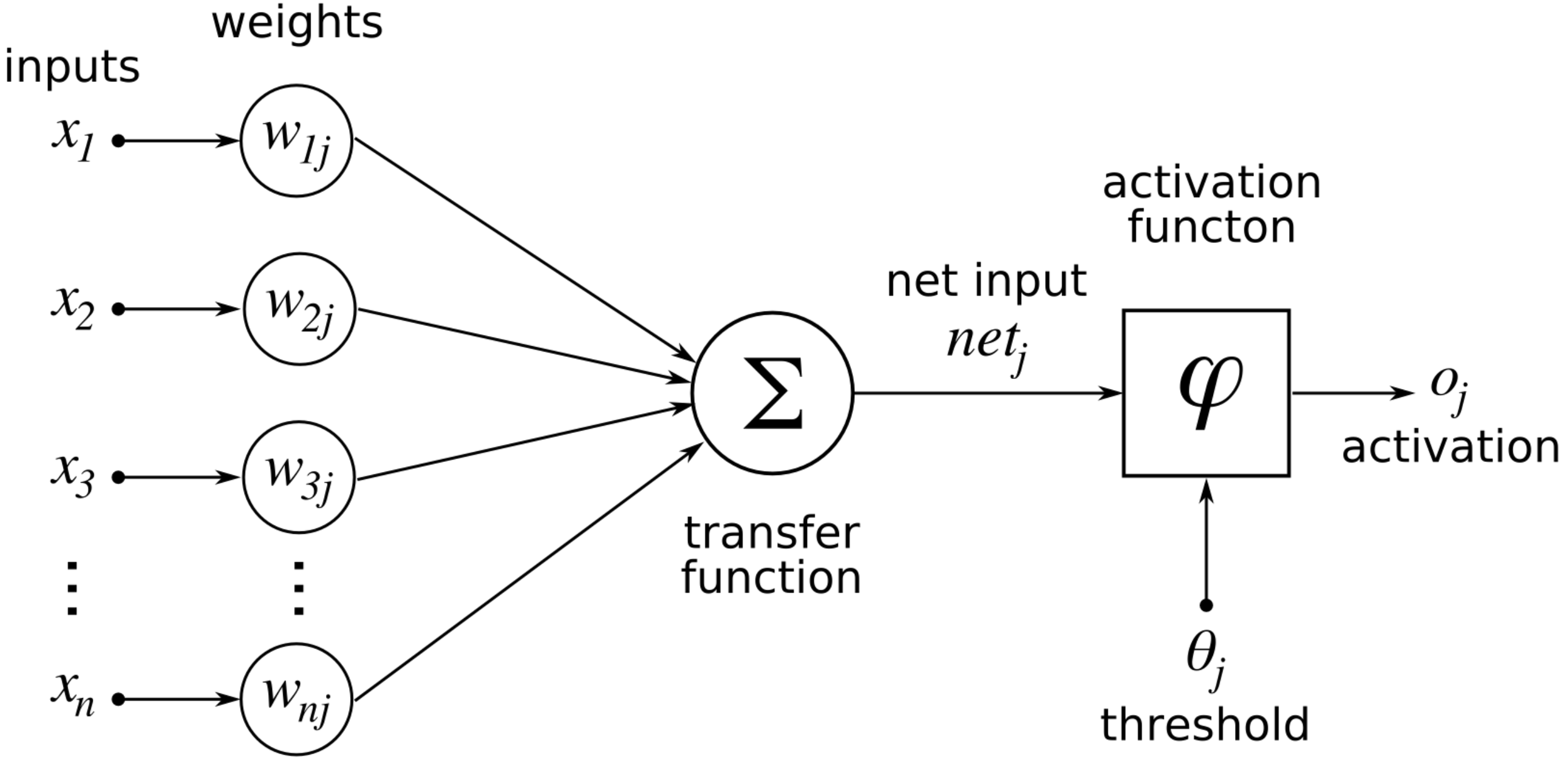
aka. “ReduceAll” parameter averaging



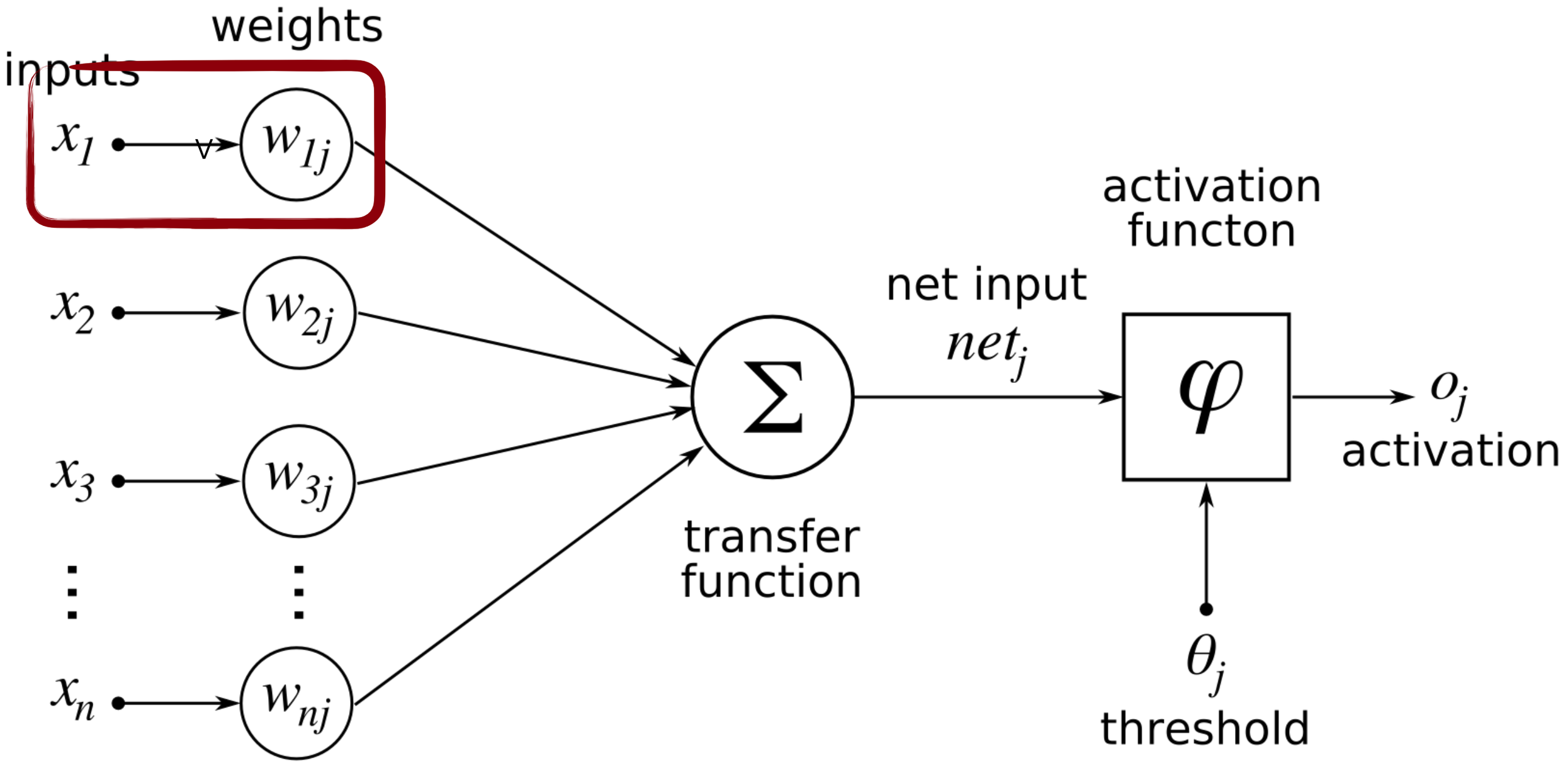
intra-model parallelism



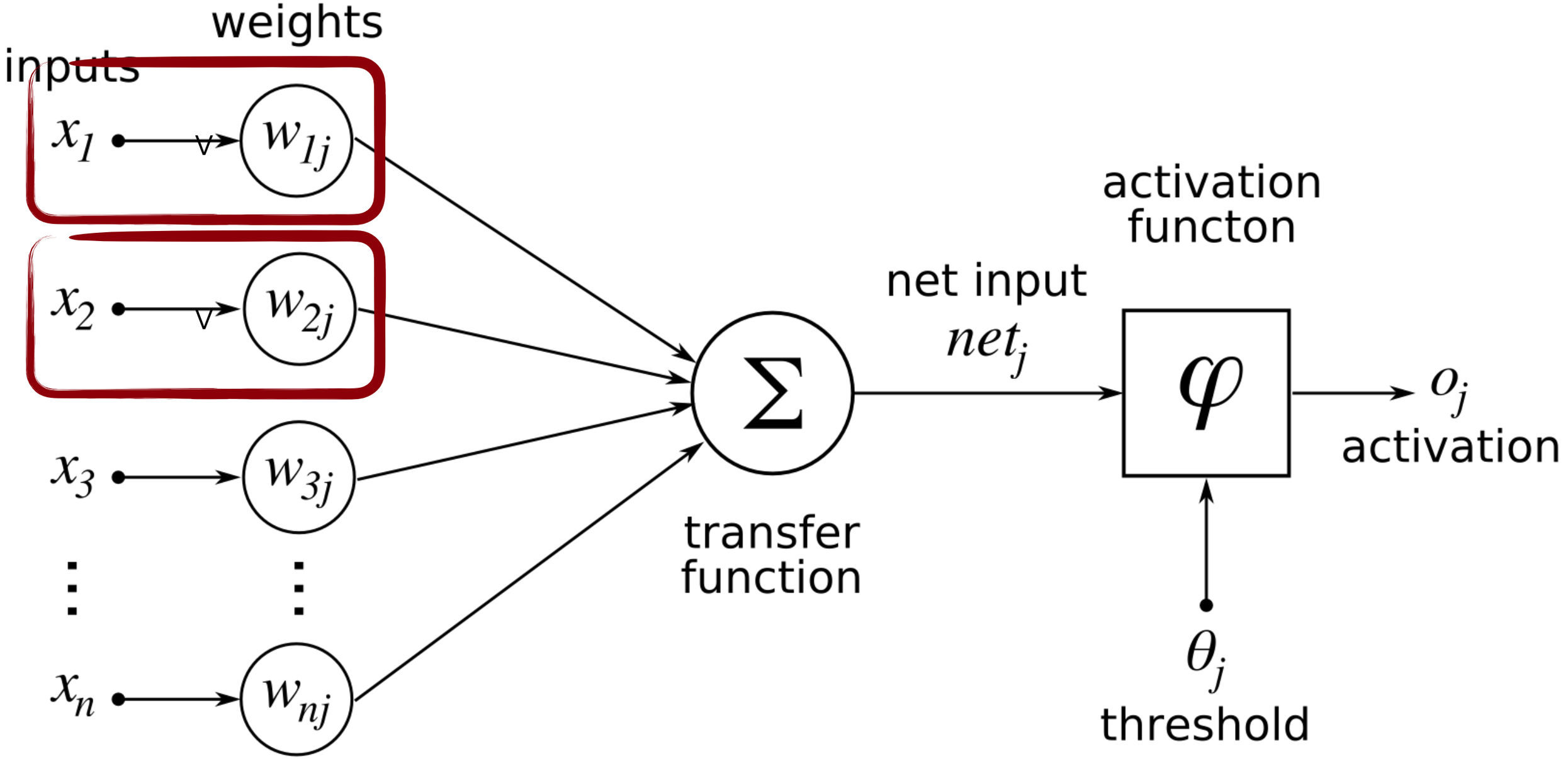
intra-model parallelism



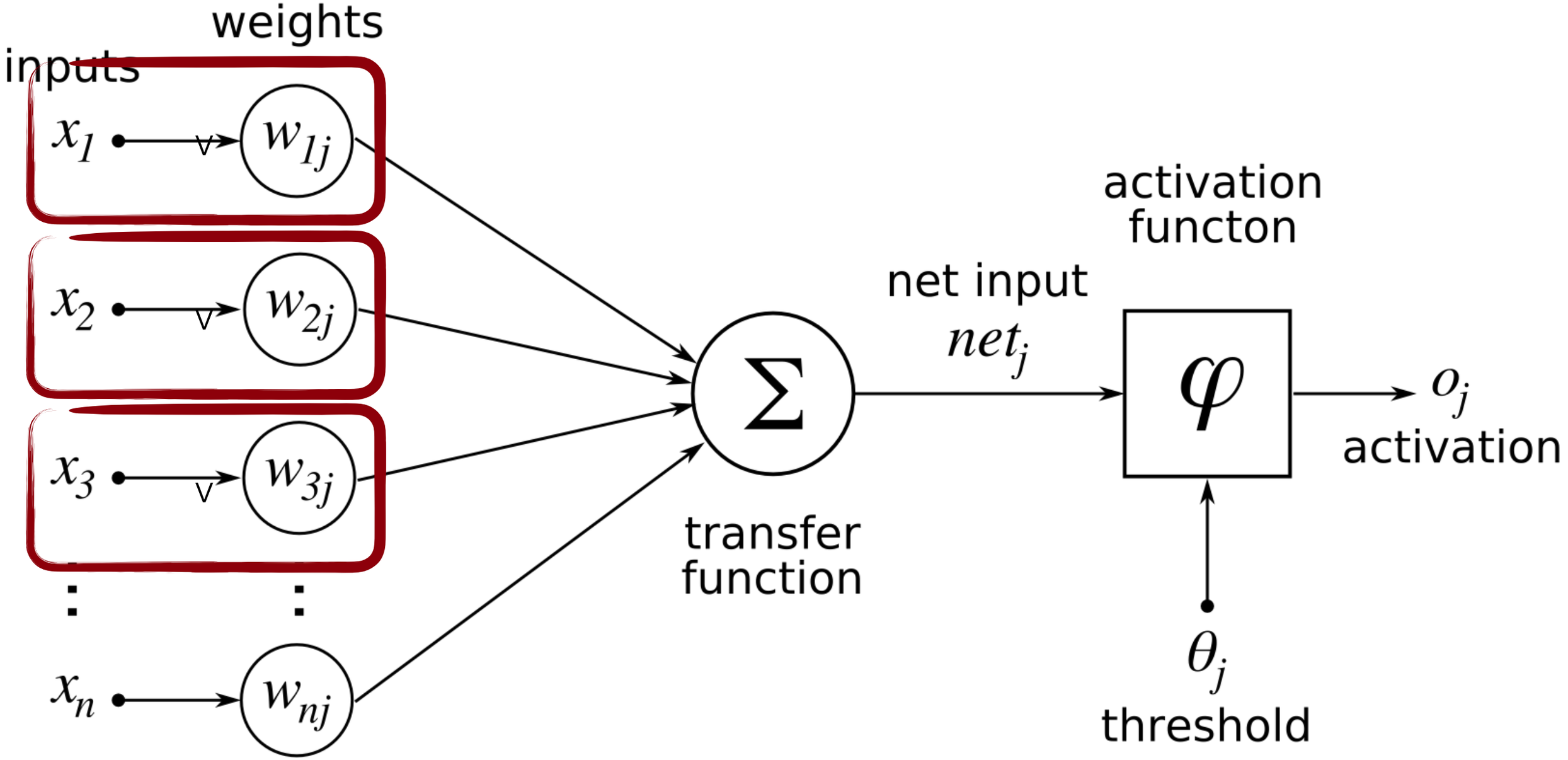
intra-model parallelism



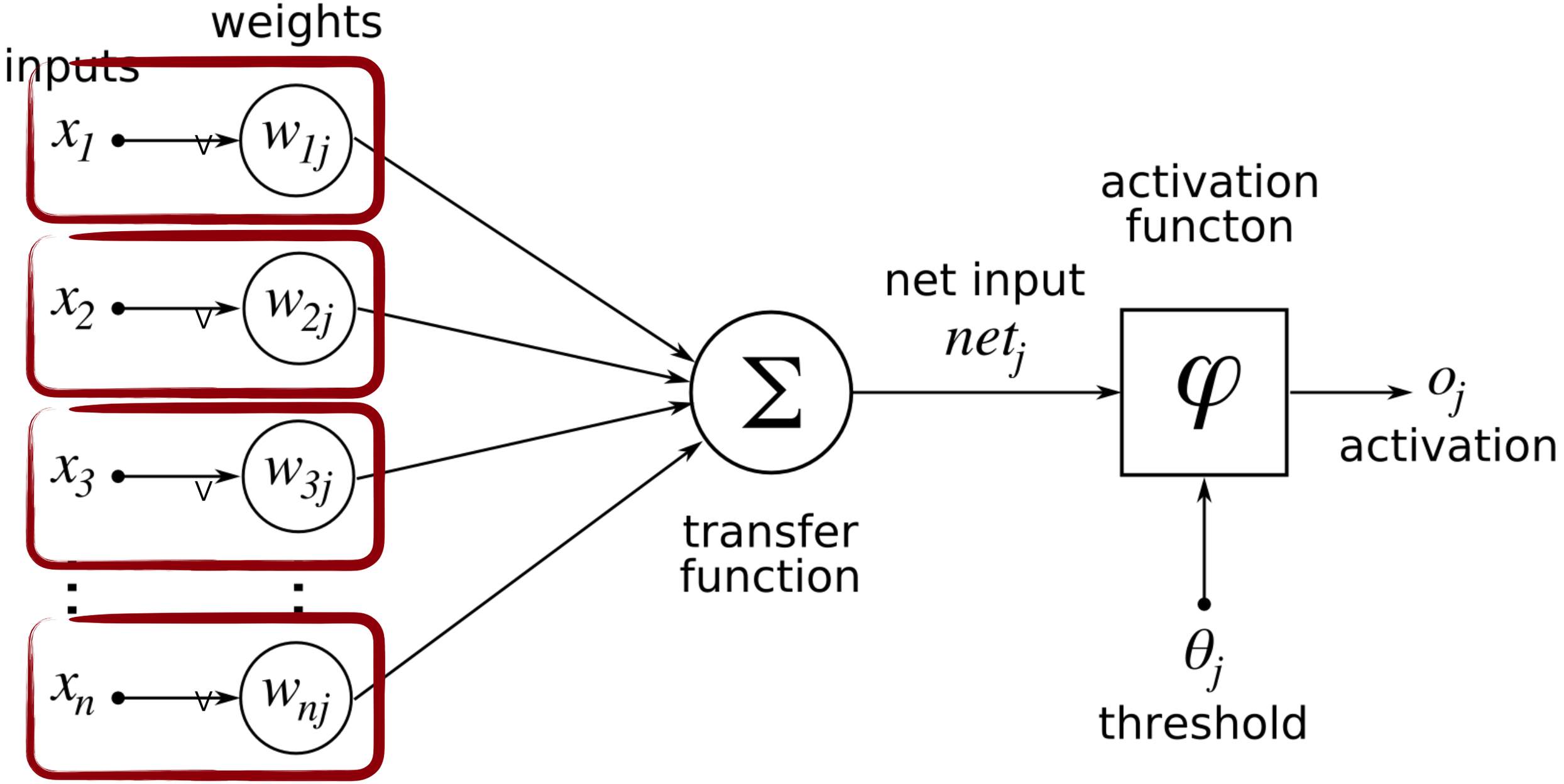
intra-model parallelism



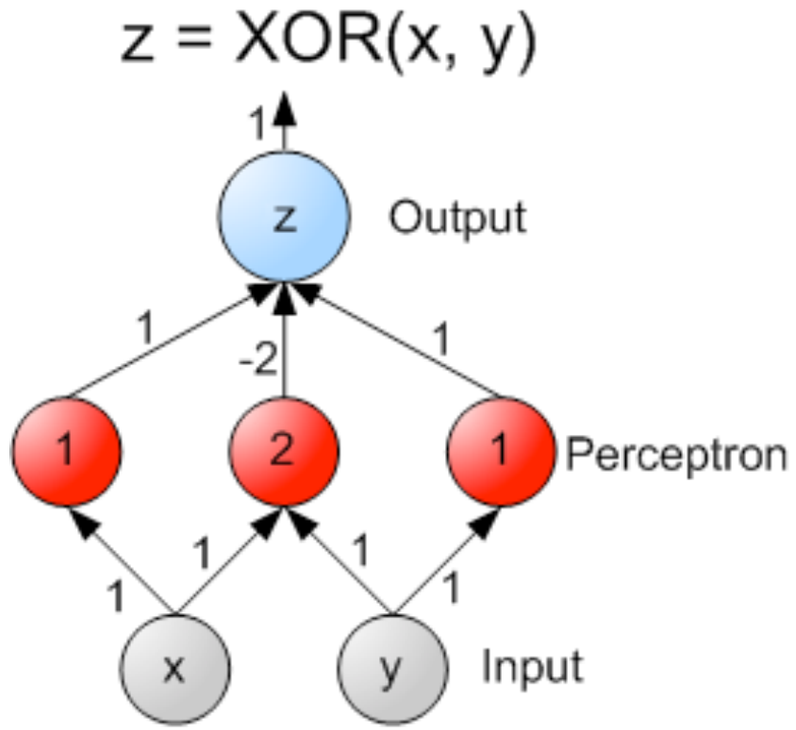
intra-model parallelism



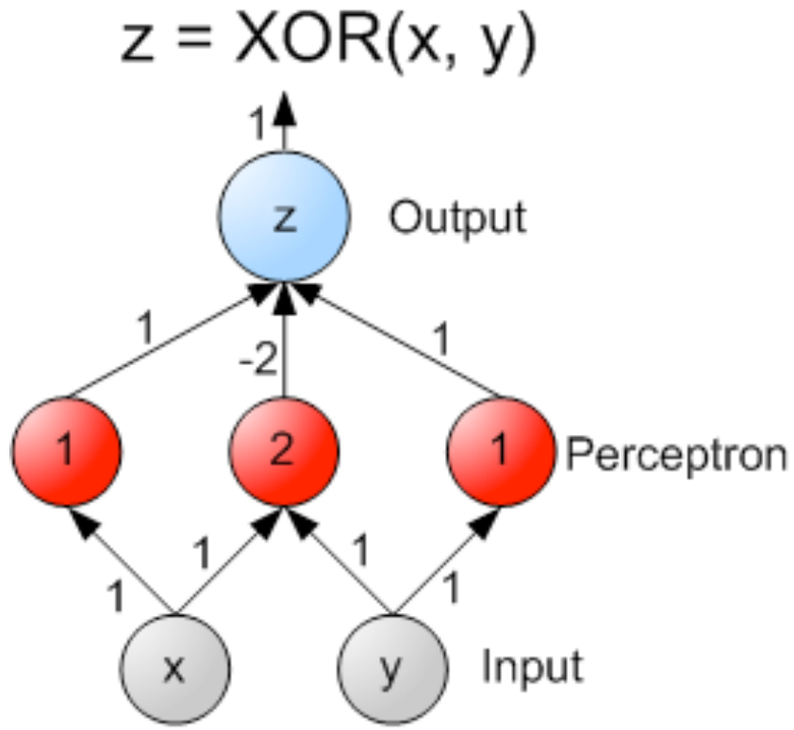
intra-model parallelism



pipelined parallelism

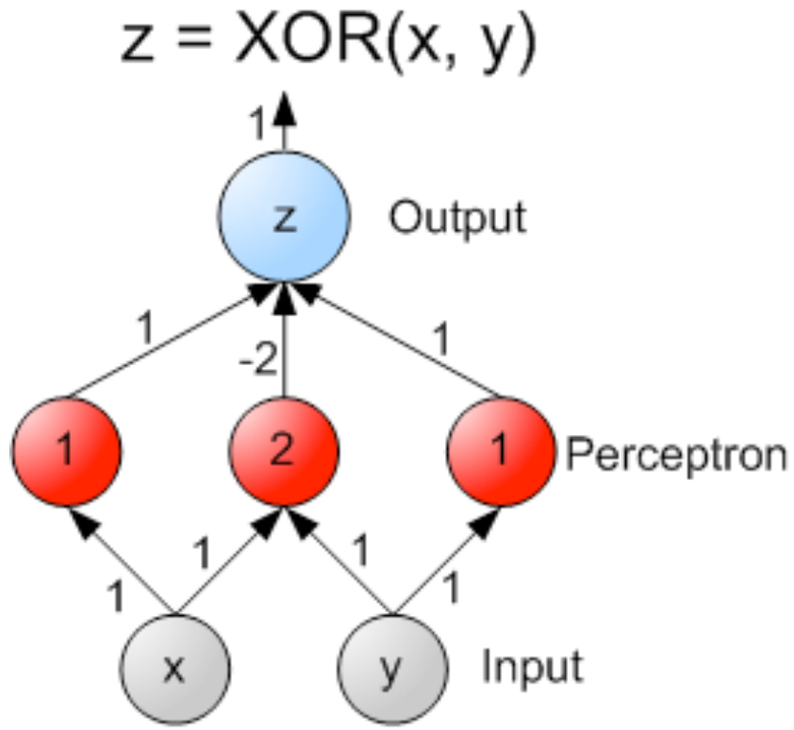


pipelined parallelism



0 0
0 1
1 0
1 1

pipelined parallelism



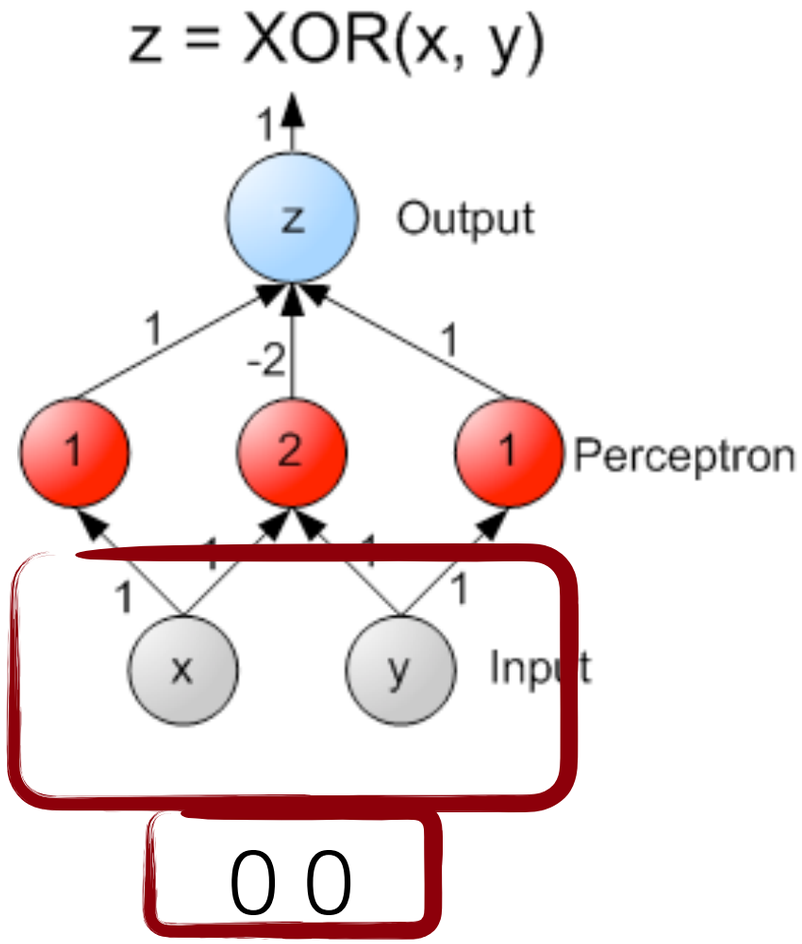
0 0

0 1

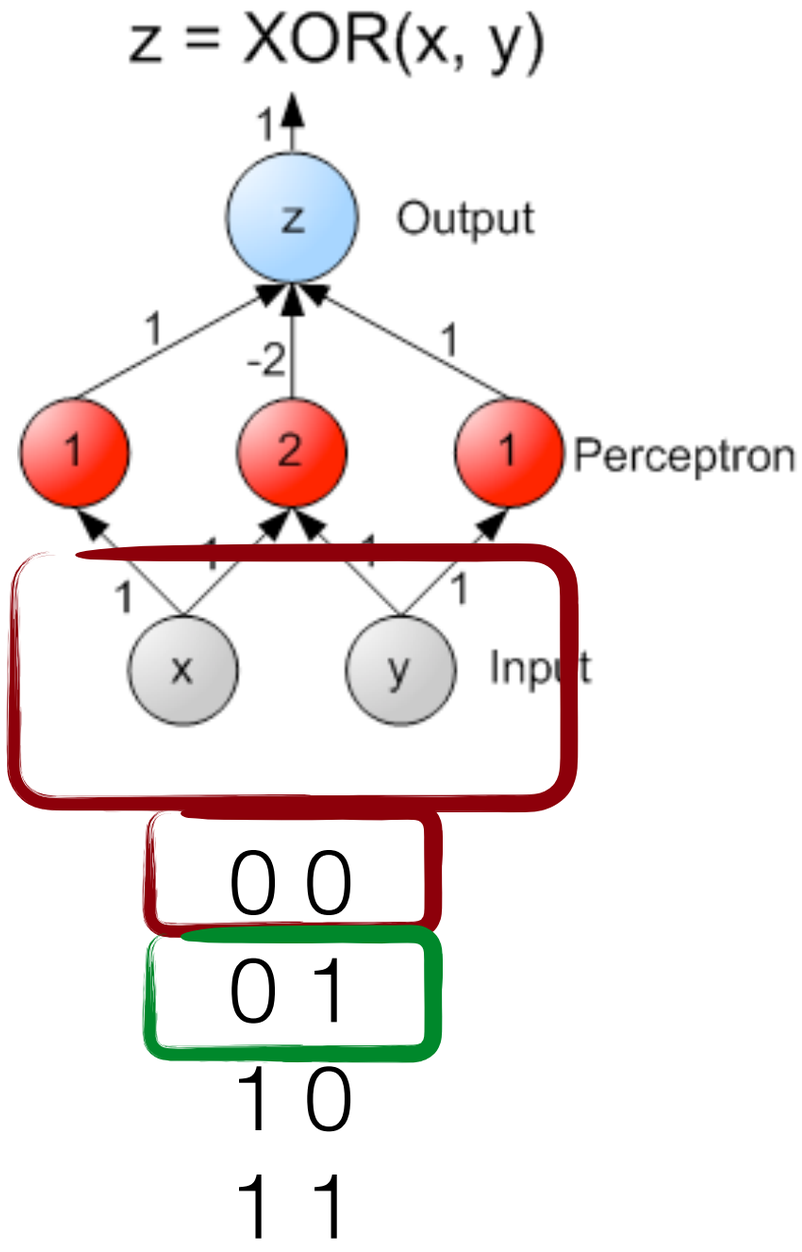
1 0

1 1

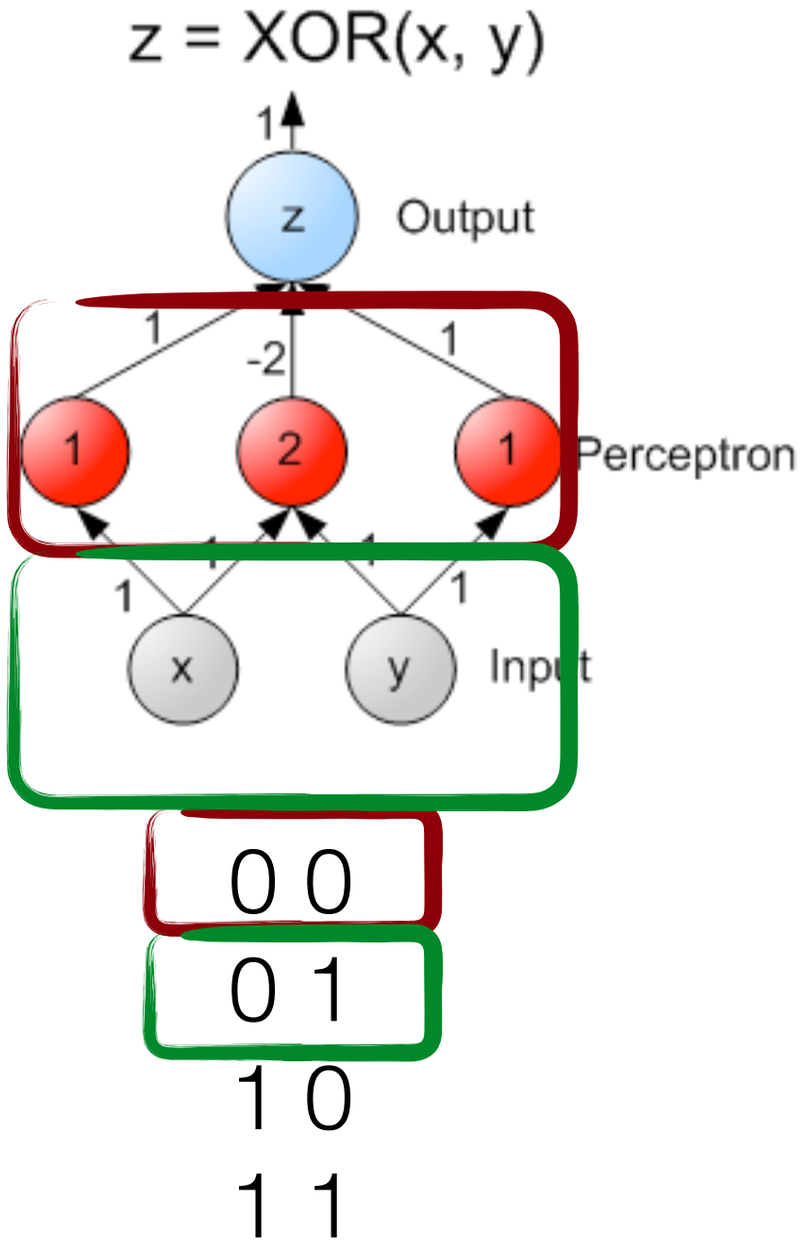
pipelined parallelism

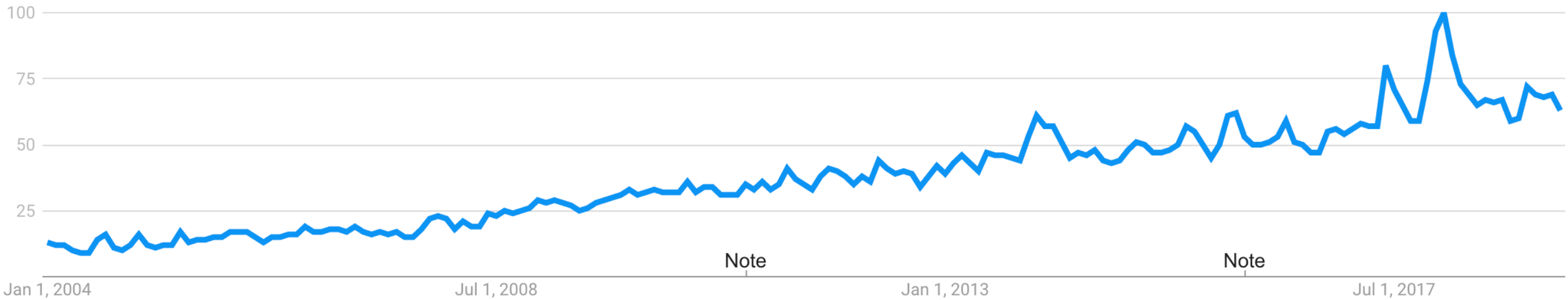


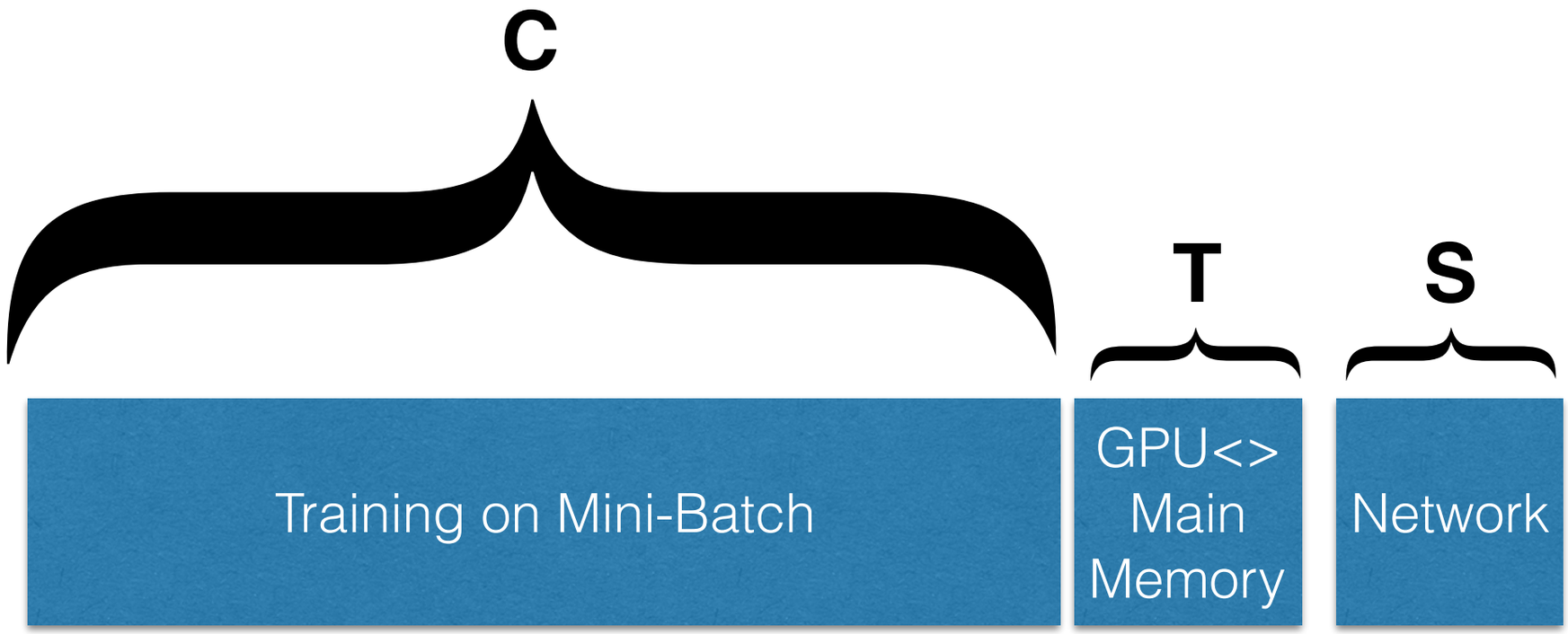
pipelined parallelism



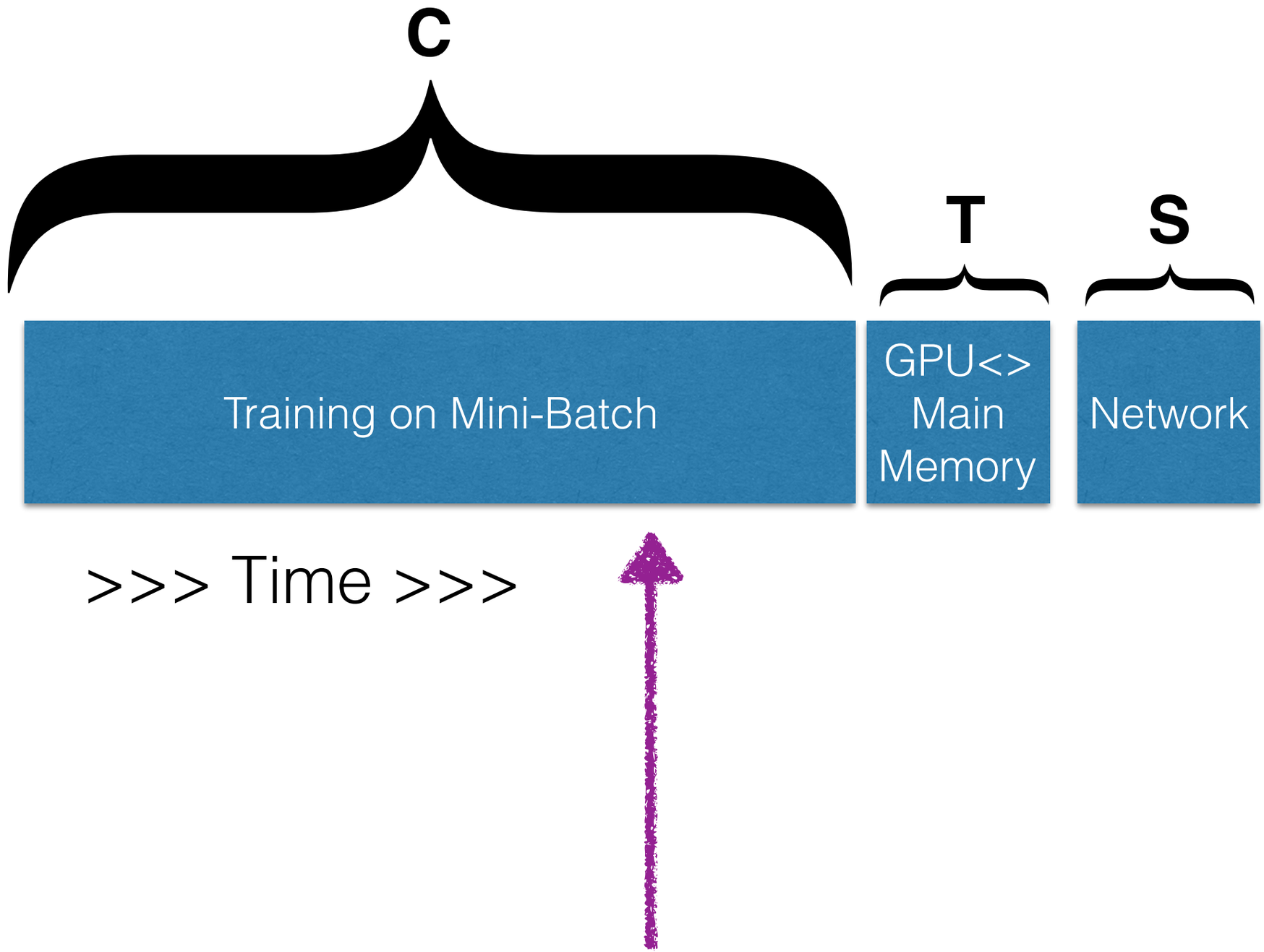
pipelined parallelism

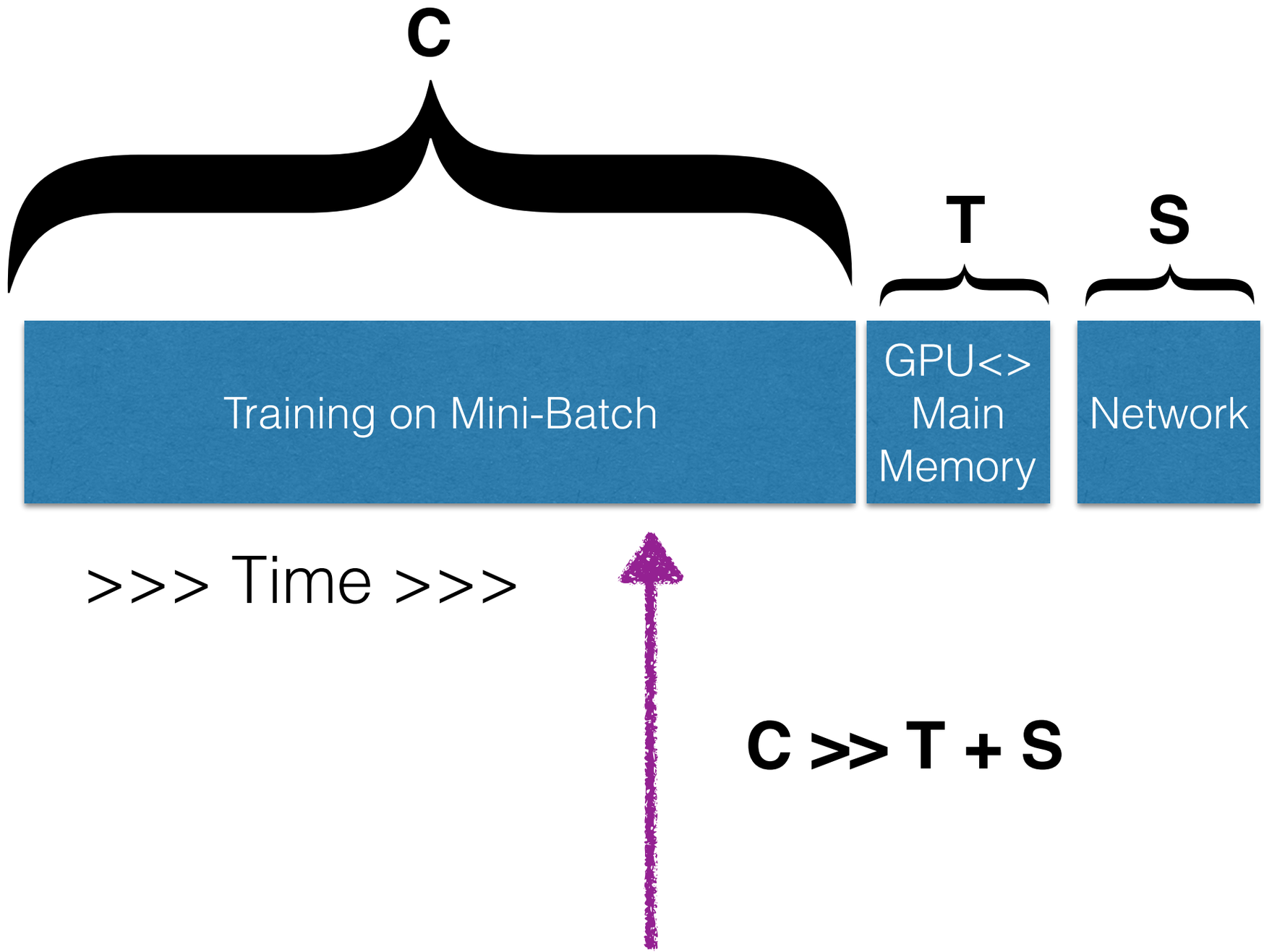


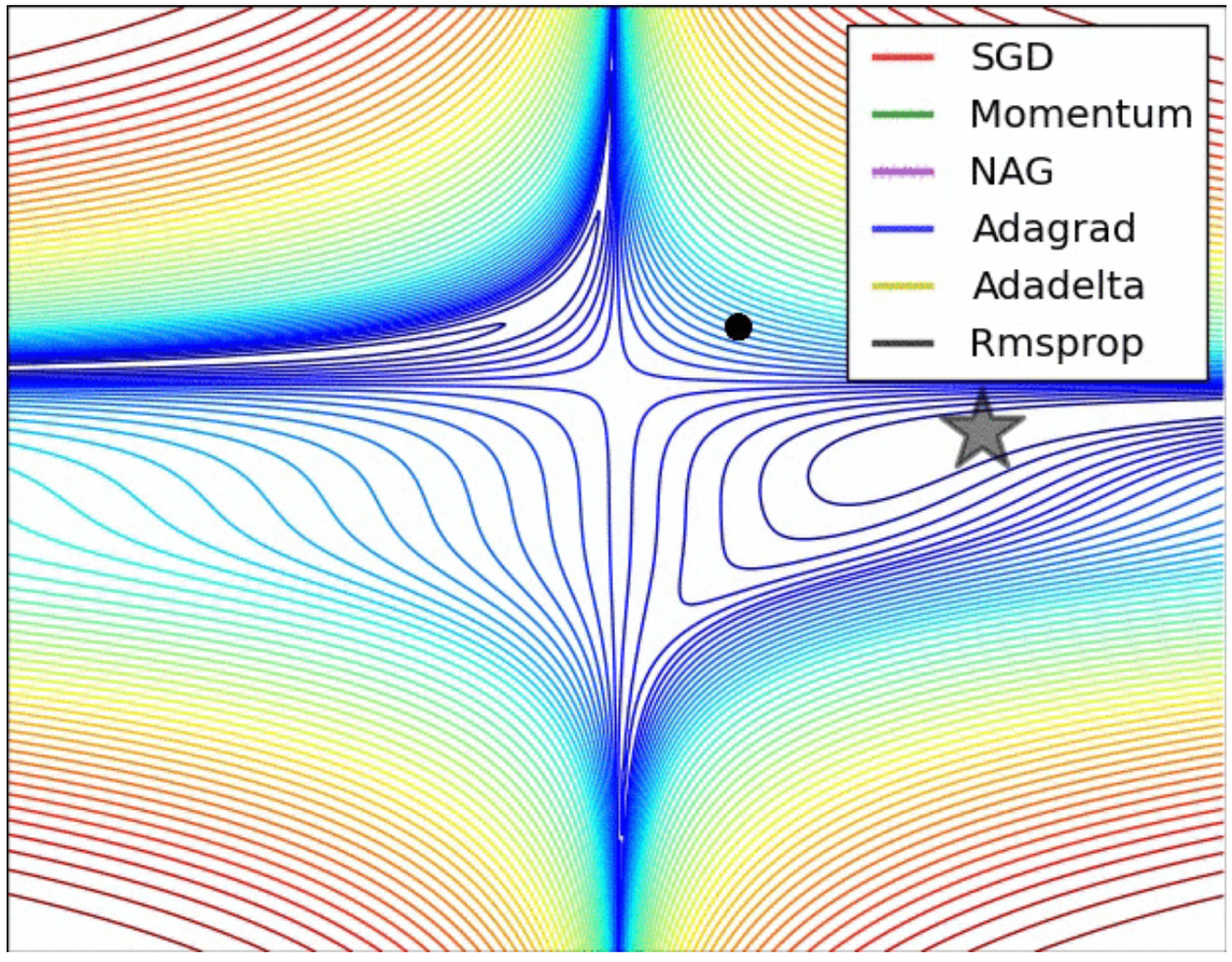

100
75
50
25
Note
Note
Jan 1, 2004
Jul 1, 2008
Jan 1, 2013
Jul 1, 2017

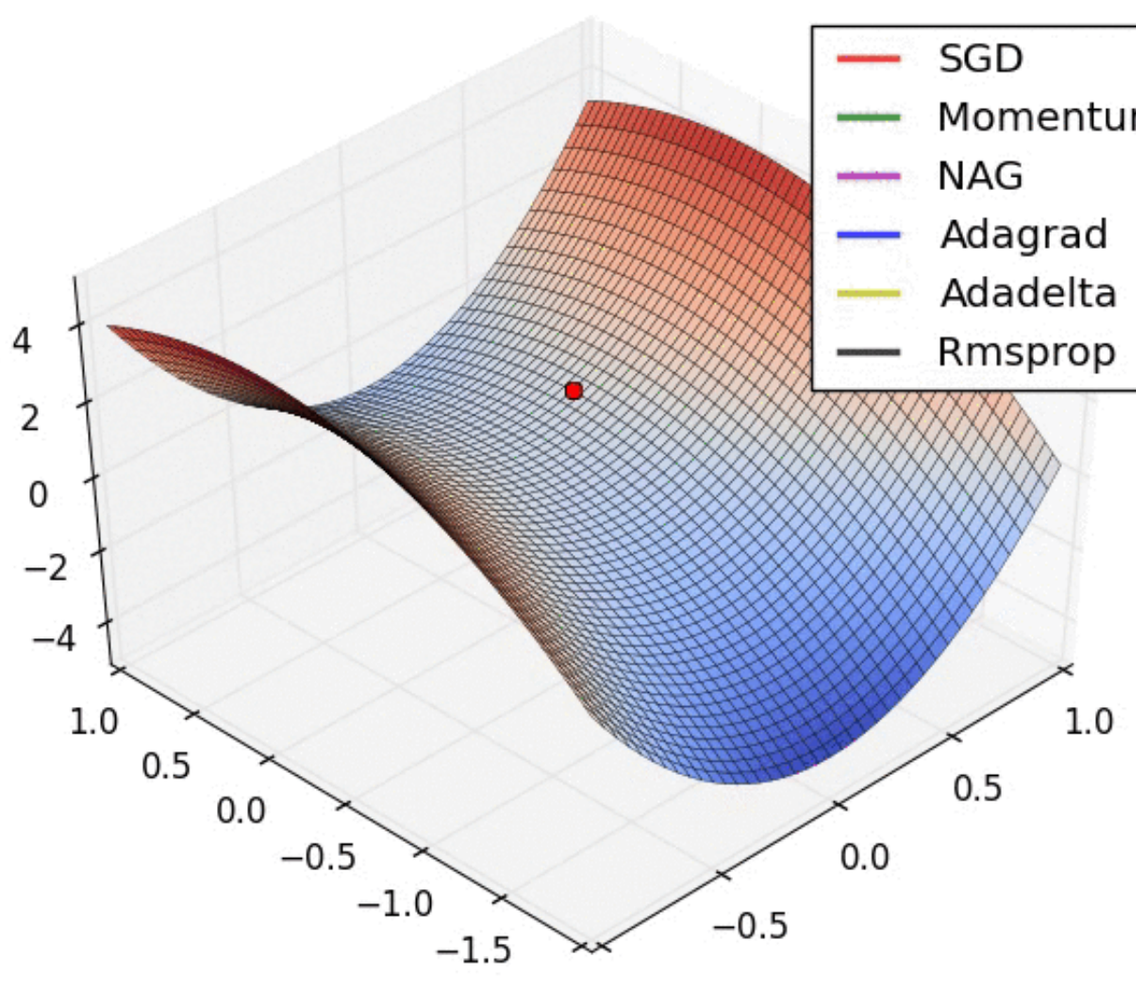


>>> Time >>>

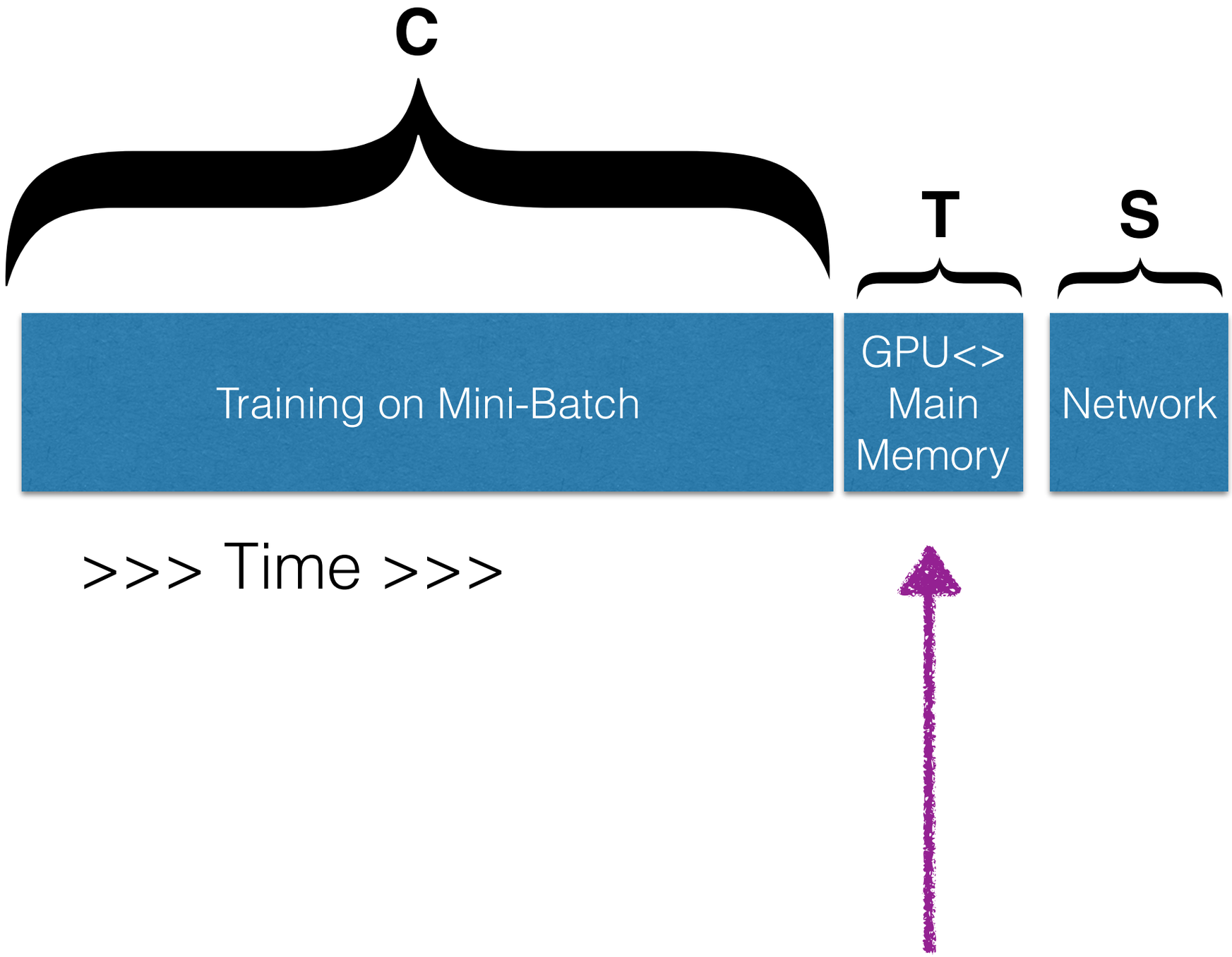




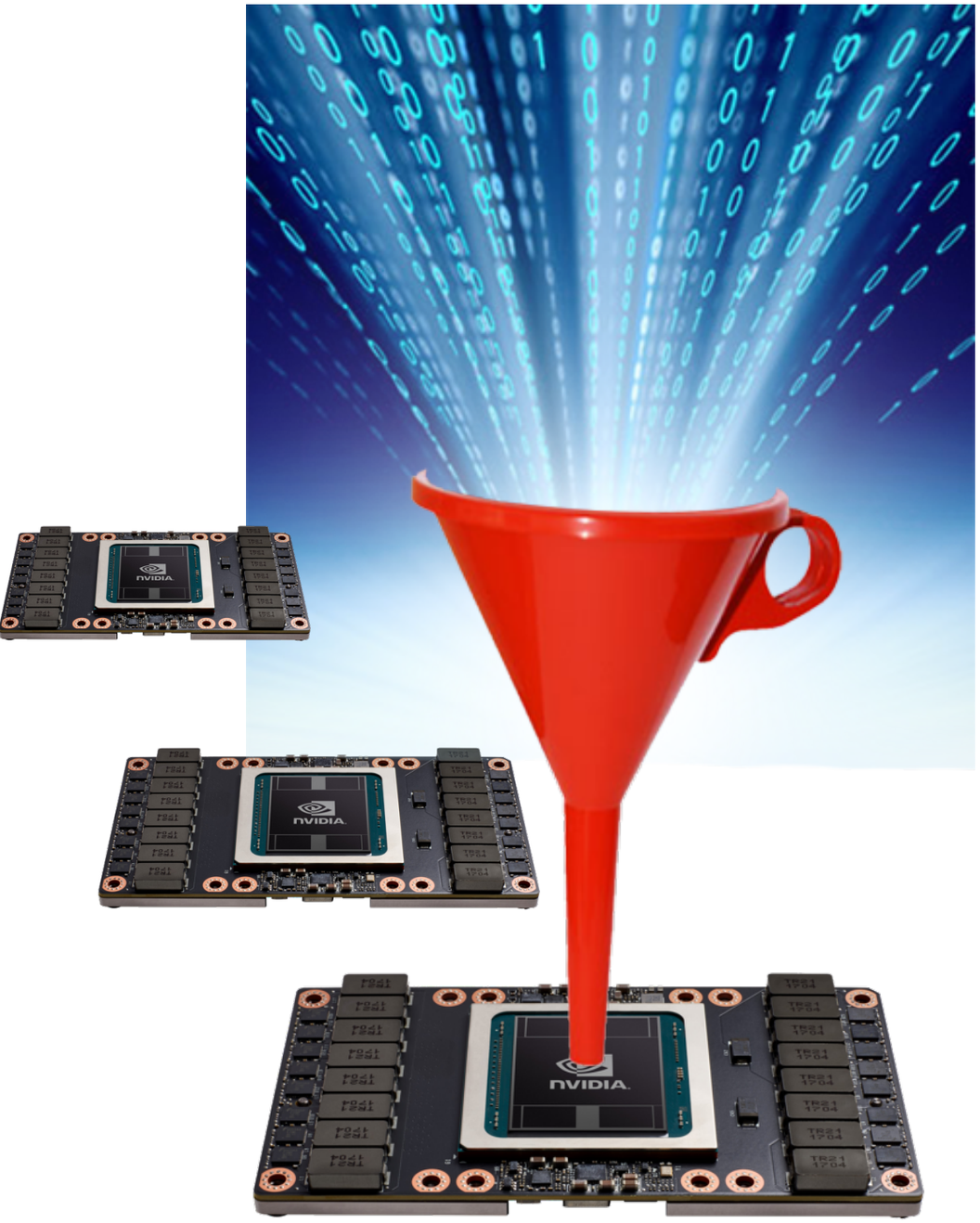




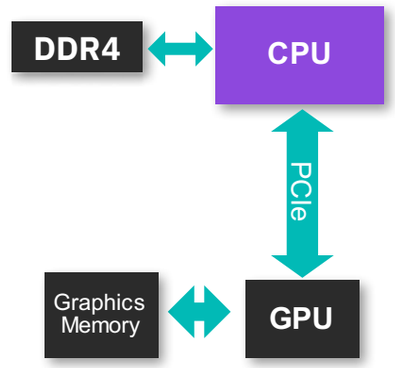
Credits:
http://www.denizyuret.com/2015/03/alec-radfords-animations-for.html



Time(%)	Time	Calls	Avg	Name
49.35%	29.581ms	1	29.581ms	[CUDA memcpy DtoH]
47.48%	28.462ms	1	28.462ms	[CUDA memcpy HtoD]
3.17%	1.9000ms	1	1.9000ms	naiveTransposeKernel



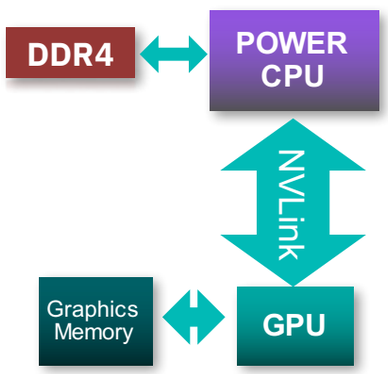
```
Time(%)   Time     Calls   Avg      Name
49.35%    29.581ms    1    29.581ms    [CUDA memcpy DtoH]
47.48%    28.462ms    1    28.462ms    [CUDA memcpy HtoD]
3.17%     1.9000ms    1    1.9000ms    naiveTransposeKernel
```

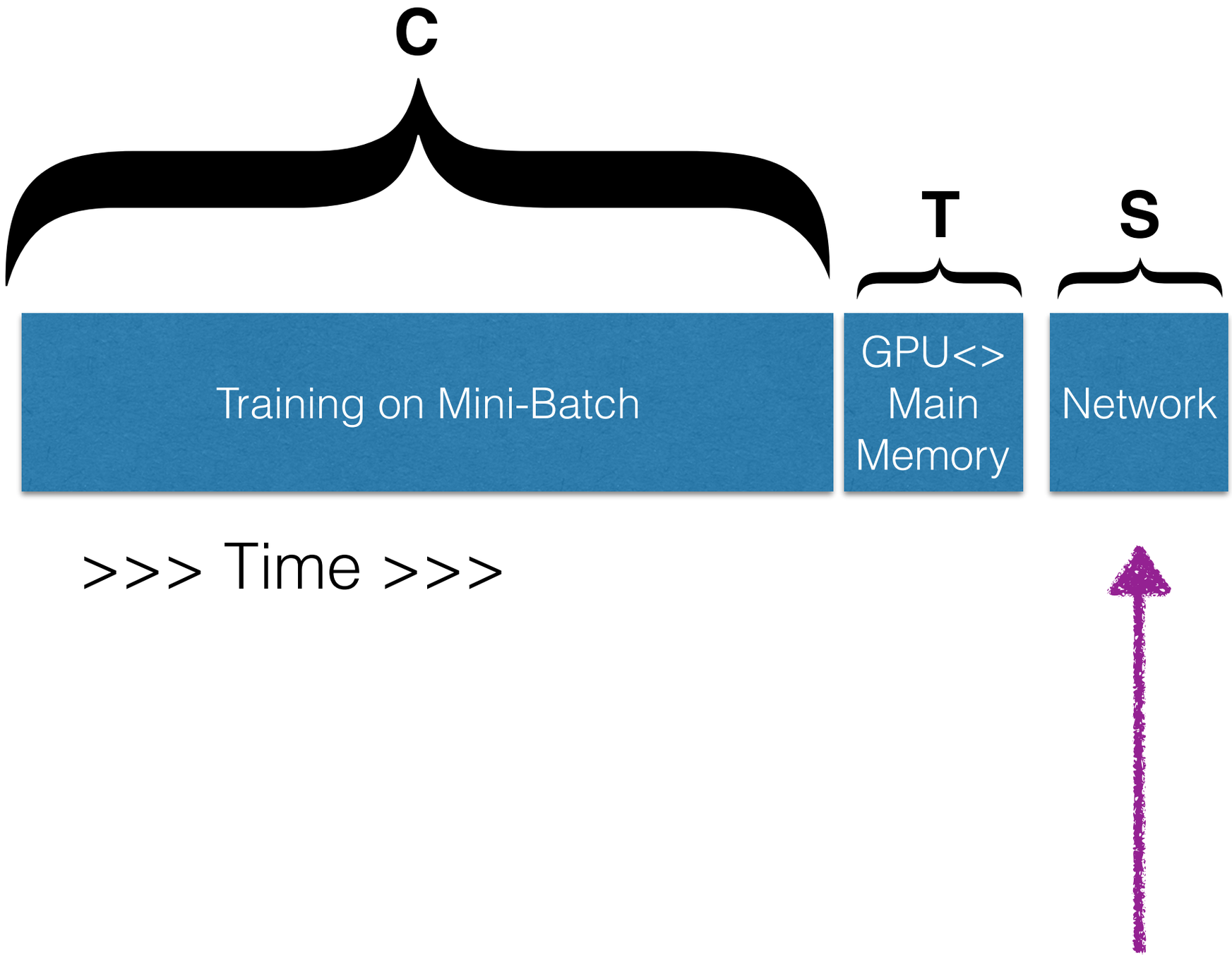


PCIe Gen3 – 32GB/sec

NVLink 1.0 – 80GB/sec

NVLink 2.0 – 150 GB/sec

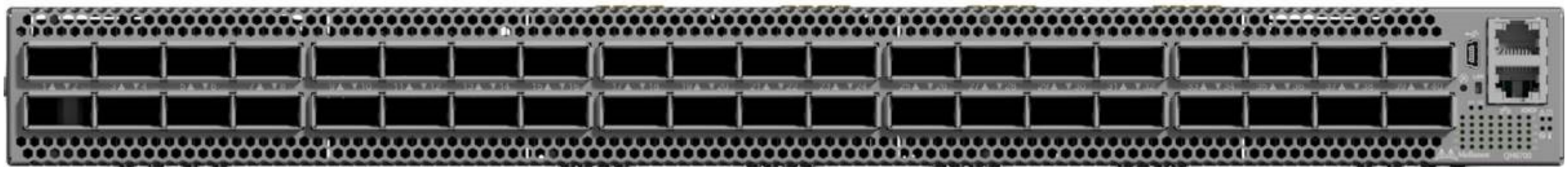




QM8700 Series - Mellanox Quantum™ HDR 200Gb/s InfiniBand Smart Switches

40-port Non-blocking HDR 200Gb/s InfiniBand Smart Switch

Mellanox provides the world's smartest switches, enabling in-network computing through the Co-Design Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ technology. The QM8700 series has the highest fabric performance available in the market with up to 16Tb/s of non-blocking bandwidth with sub 90ns port-to-port latency.



Open MPI:
Open Source High Performance Computing

IBM Spectrum MPI

Accelerating high-performance application parallelization

Romeo Kienzler

Levels

Benefits

Current Members

How to Join

Current Members

Platinum Level

Google

IBM

Mellanox TECHNOLOGIES

Micron

NVIDIA

redhat.

ubuntu Supported by Canonical

XILINX

PowerAI Vision

Auto-ML for Images & Video

Label | Train | Deploy

Watson Machine Learning

PowerAI: Open Source Frameworks

TensorFlow™ PYTORCH Chainer SnapML

Large Model Support (LMS)

Watson Machine Learning Accelerator

Distributed Deep Learning (DDL)

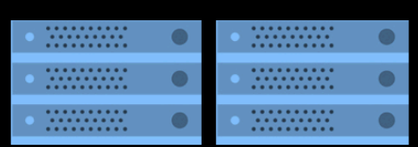
Elastic Distributed Inference (Preview)

IBM Spectrum Conductor
Cluster Virtualization, Elastic Training
Auto Hyper-Parameter Optimization

Deep Learning Impact (DLI) Module

Data & Model Management, ETL, Visualize, Advise

Accelerated Infrastructure



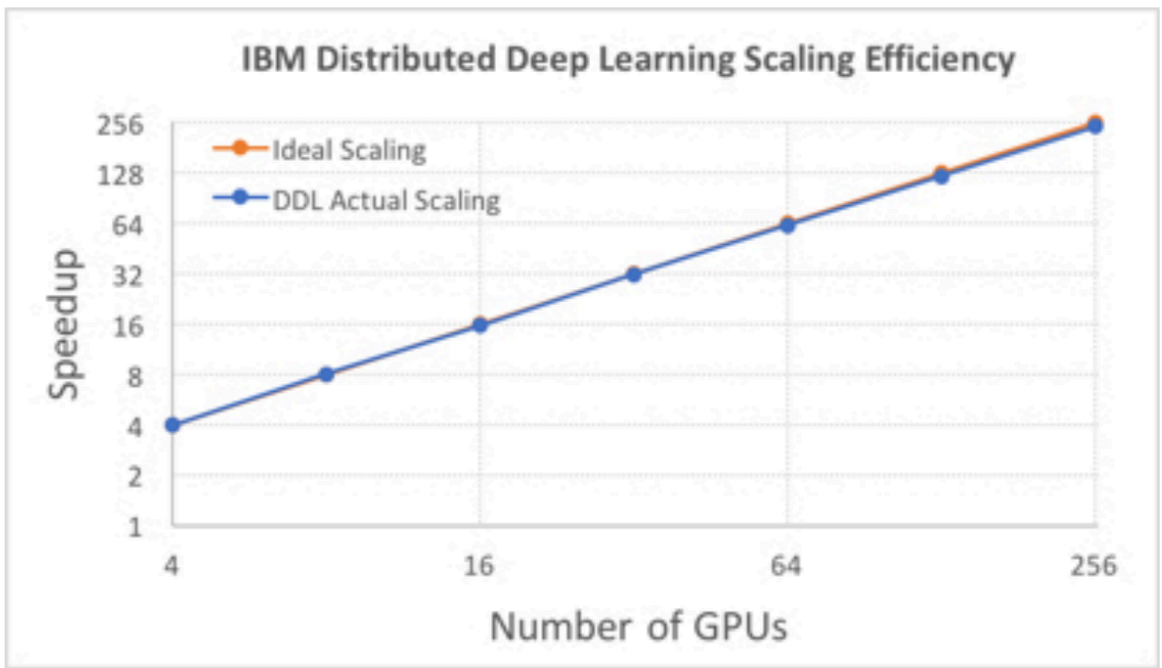
Accelerated Servers AC922

Storage (Spectrum Scale ESS)

August 8, 2017
Posted in: AI, Cognitive Computing

IBM Research achieves record deep learning performance with new software technology

Summary: IBM Research publishes in arXiv close to ideal scaling with new distributed deep learning software which achieved record communication overhead and 95% scaling efficiency on the Caffe deep learning framework over 256 NVIDIA GPUs in 64 IBM Power systems. Previous best scaling was demonstrated by Facebook AI Research of 89% for a training run on Caffe2, at higher communication overhead. IBM Research also beat Facebook's time by training the model in 50 minutes, versus the 1 hour Facebook took. Using this software, IBM Research achieved a new image recognition accuracy of 33.8% for a neural network trained on a very large data set (7.5M images). The previous record published by Microsoft demonstrated 29.8% accuracy.

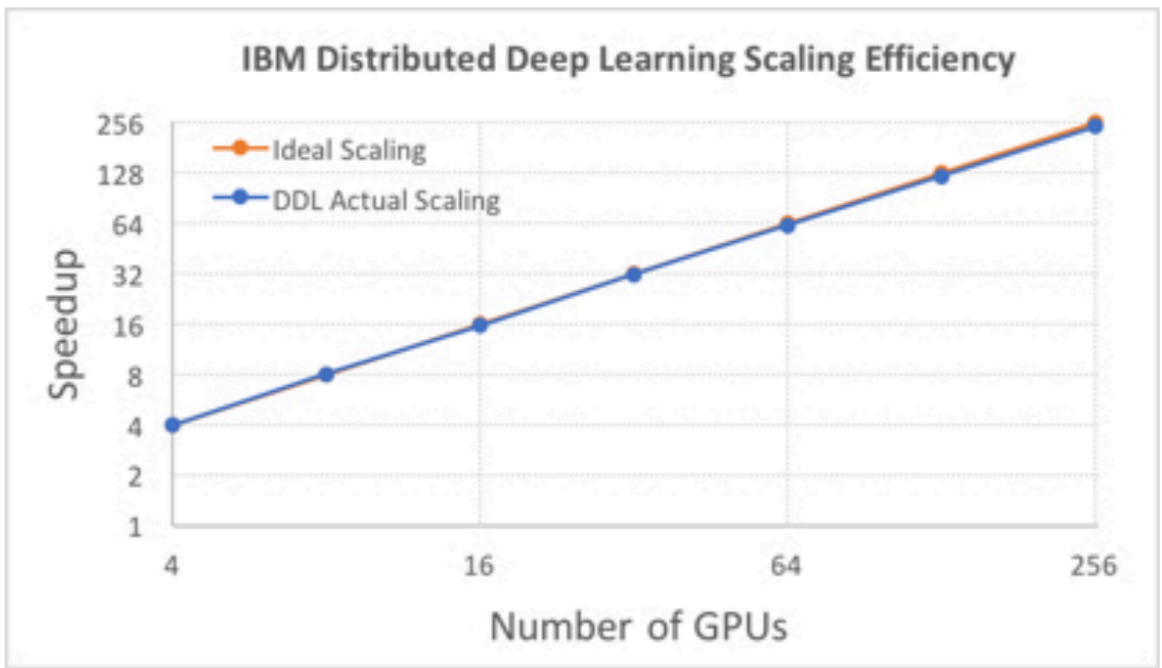


August 8, 2017
Posted in: AI, Cognitive Computing

IBM Research achieves record deep learning performance with new software technology

Scaling Efficiency
Facebook: 89%
IBM: 95%

Summary: IBM Research publishes in arXiv close to ideal scaling with new distributed deep learning software which achieved record communication overhead and 95% scaling efficiency on the Caffe deep learning framework over 256 NVIDIA GPUs in 64 IBM Power systems. Previous best scaling was demonstrated by Facebook AI Research of 89% for a training run on Caffe2, at higher communication overhead. IBM Research also beat Facebook's time by training the model in 50 minutes, versus the 1 hour Facebook took. Using this software, IBM Research achieved a new image recognition accuracy of 33.8% for a neural network trained on a very large data set (7.5M images). The previous record published by Microsoft demonstrated 29.8% accuracy.



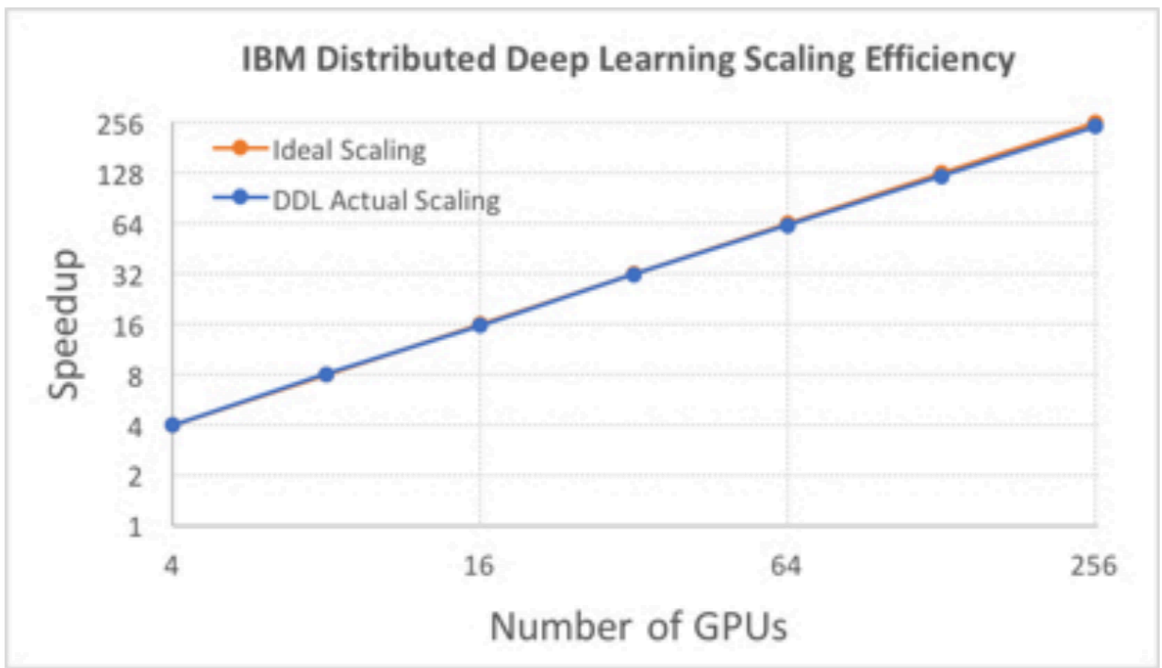
August 8, 2017
Posted in: AI, Cognitive Computing

IBM Research achieves record deep learning performance with new software technology

Scaling Efficiency
Facebook: 89%
IBM: 95%

Runtime
Facebook: 1h
IBM: 50 minutes

Summary: IBM Research publishes in arXiv close to ideal scaling with new distributed deep learning software which achieved record communication overhead and 95% scaling efficiency on the Caffe deep learning framework over 256 NVIDIA GPUs in 64 IBM Power systems. Previous best scaling was demonstrated by Facebook AI Research of 89% for a training run on Caffe2, at higher communication overhead. IBM Research also beat Facebook's time by training the model in 50 minutes, versus the 1 hour Facebook took. Using this software, IBM Research achieved a new image recognition accuracy of 33.8% for a neural network trained on a very large data set (7.5M images). The previous record published by Microsoft demonstrated 29.8% accuracy.



August 8, 2017
Posted in: AI, Cognitive Computing

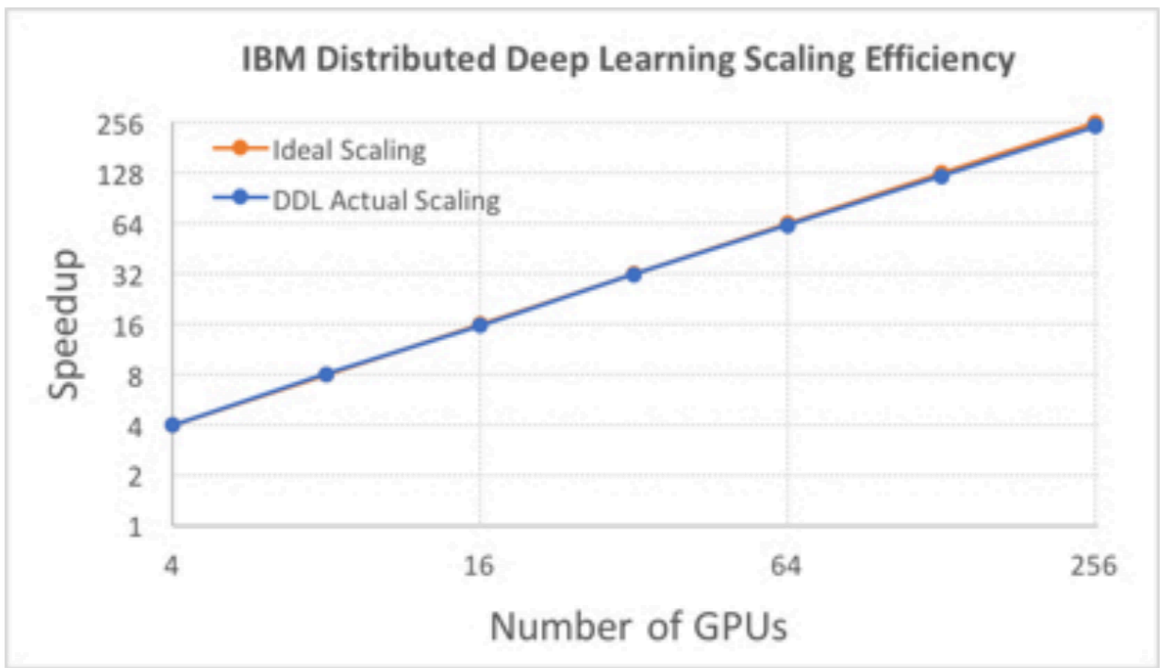
IBM Research achieves record deep learning performance with new software technology

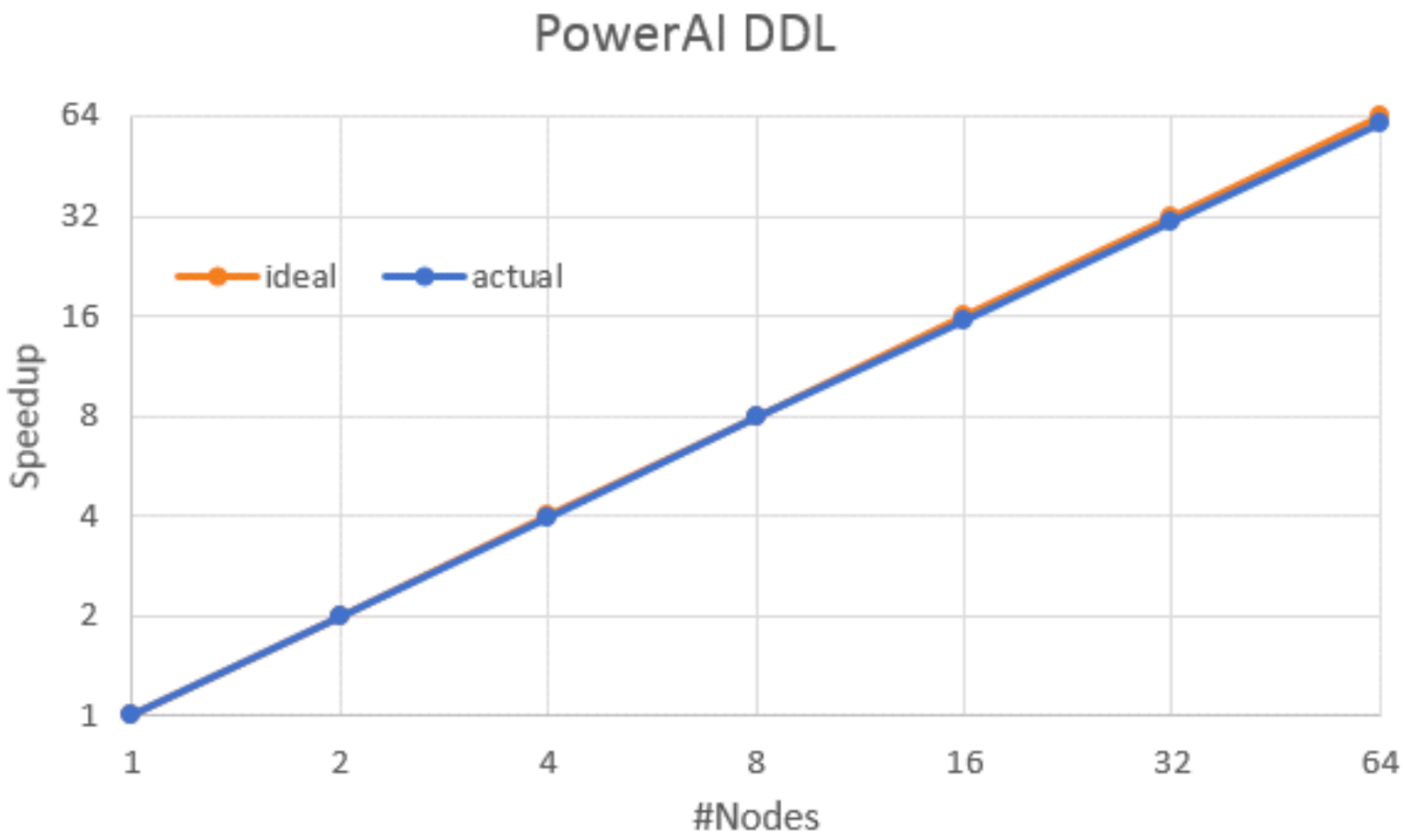
Scaling Efficiency
Facebook: 89%
IBM: 95%

Runtime
Facebook: 1h
IBM: 50 minutes

Summary: IBM Research publishes in arXiv close to ideal scaling with new distributed deep learning software which achieved record communication overhead and 95% scaling efficiency on the Caffe deep learning framework over 256 NVIDIA GPUs in 64 IBM Power systems. Previous best scaling was demonstrated by Facebook AI Research of 89% for a training run on Caffe2, at higher communication overhead. IBM Research also beat Facebook's time by training the model in 50 minutes, versus the 1 hour Facebook took. Using this software, IBM Research achieved a new image recognition accuracy of 33.8% for a neural network trained on a very large data set (7.5M images). The previous record published by Microsoft demonstrated 29.8% accuracy.

Accuracy
Microsoft: 29.8%
IBM: 33.8%





#GPUs	4	8	16	32	64	128	256
#Nodes	1	2	4	8	16	32	64
Speedup	1.0	2.0	3.9	7.9	15.5	30.5	60.6
Scaling efficiency	1.00	1.00	0.98	0.99	0.97	0.95	0.95

Figure 2: Resnet-50 for 1K classes using up to 256 GPUs with Caffe.

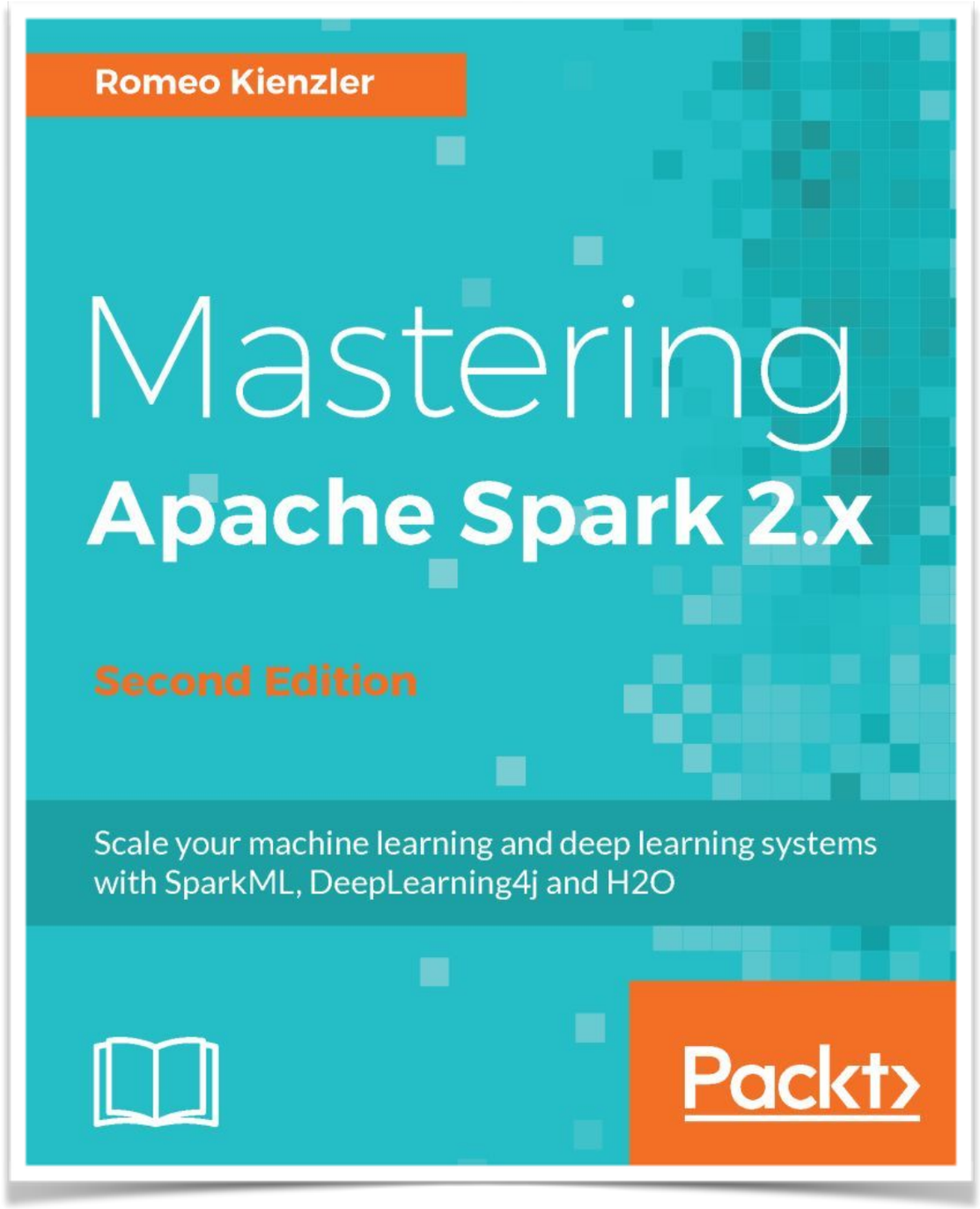
U.S. DEPARTMENT OF ENERGY

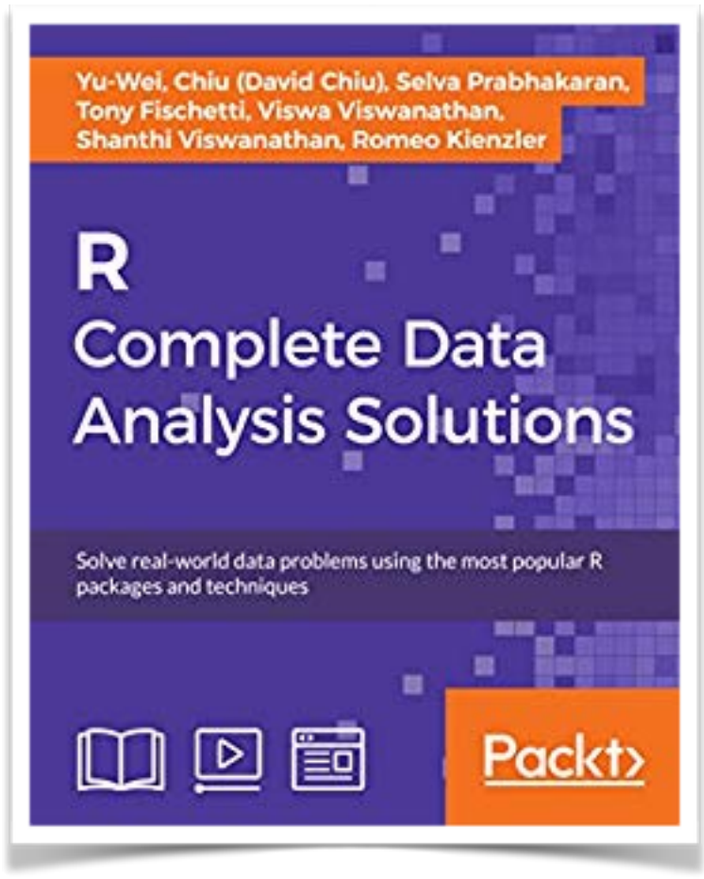
Office of Science

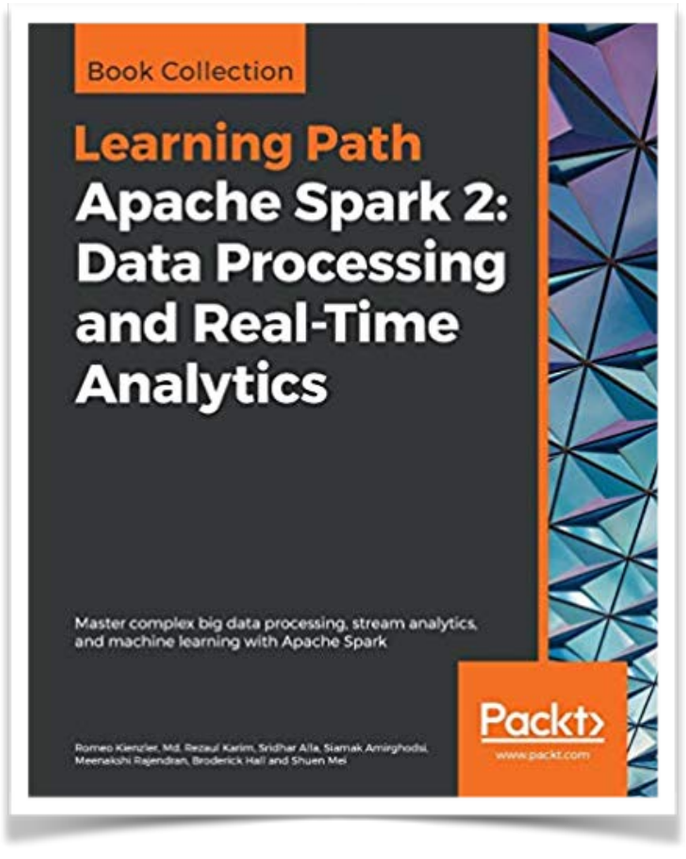
National Nuclear Security Administration

- 200 petaflops
- 4600 nodes
- 9200 Power9 CPU’s
- 220800 cores
- 1766400 hyper threads
- 27600 nVidia V100 GPUs
- 10 PB memory
- Power 15 MW
- 1% of Ethereum’s hash rate

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,282,544	122,300.0	187,659.3	8,806
2	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
3	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/NNSA/LLNL United States	1,572,480	71,610.0	119,193.6	


Romeo Kienzler
Mastering
Apache Spark 2.x
Second Edition
Scale your machine learning and deep learning systems
with SparkML, DeepLearning4j and H2O
Packt>


Yu-Wei, Chiu (David Chiu), Selva Prabhakaran,
Tony Fischetti, Viswa Viswanathan,
Shanthi Viswanathan, Romeo Kienzler
R
Complete Data
Analysis Solutions
Solve real-world data problems using the most popular R
packages and techniques
Packt>


Book Collection
Learning Path
Apache Spark 2:
Data Processing
and Real-Time
Analytics
Master complex big data processing, stream analytics,
and machine learning with Apache Spark
Packt>

IBM Advanced Data Science Specialization Certificate on Coursera

Fundamentals of Scalable Data Science
www.coursera.org/learn/ds

Advanced Machine Learning and Signal Processing
www.coursera.org/learn/advanced-machine-learning-signal-processing

Applied AI with DeepLearning
www.coursera.ai/learn/ai

Advanced Data Science Capstone Project
https://www.coursera.org/learn/advanced-data-science-capstone

?