DEEP LEARNING AND CHALLENGES OF SCALE

François Courteille <fcourteille@nvidia.com>, October, 2018

DL IS A HPC WORKLOAD

HPC expertise is important for success

It makes sense to build an AI team and a separate systems/HPC team and have the two teams sit next to each other.

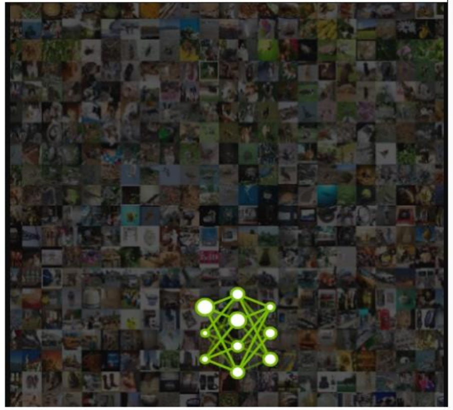
That is because solving some of the problems discussed in the lecture requires very specialised systems/HPC knowledge. It is incredibly difficult for any single human to acquire both the AI and systems/HPC knowledge.

For detailed discussion see Andrew Ng’s “Nuts and Bolts of Applying Deep Learning ”:
https://www.youtube.com/watch?v=F1ka6a13S9I&t=120s

NEURAL NETWORK COMPLEXITY IS EXPLODING

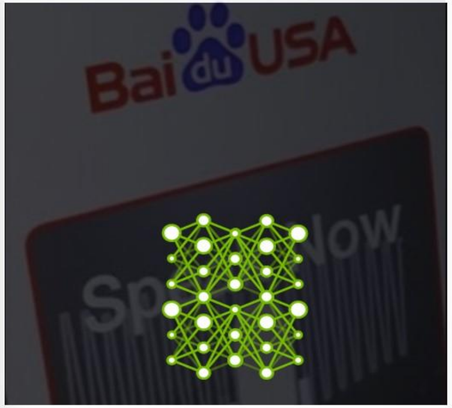
To Tackle Increasingly Complex Challenges

7 ExaFLOPS
60 Million Parameters



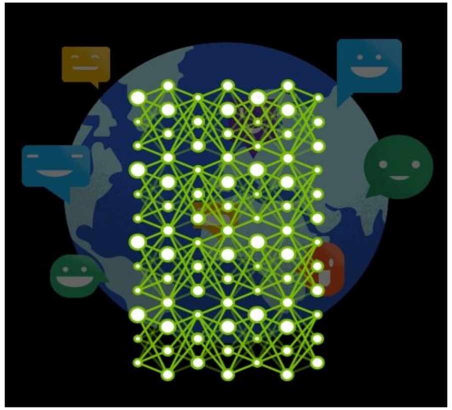
2015 - Microsoft ResNet
Superhuman Image Recognition

20 ExaFLOPS
300 Million Parameters



2016 - Baidu Deep Speech 2
Superhuman Voice Recognition

100 ExaFLOPS
8700 Million Parameters

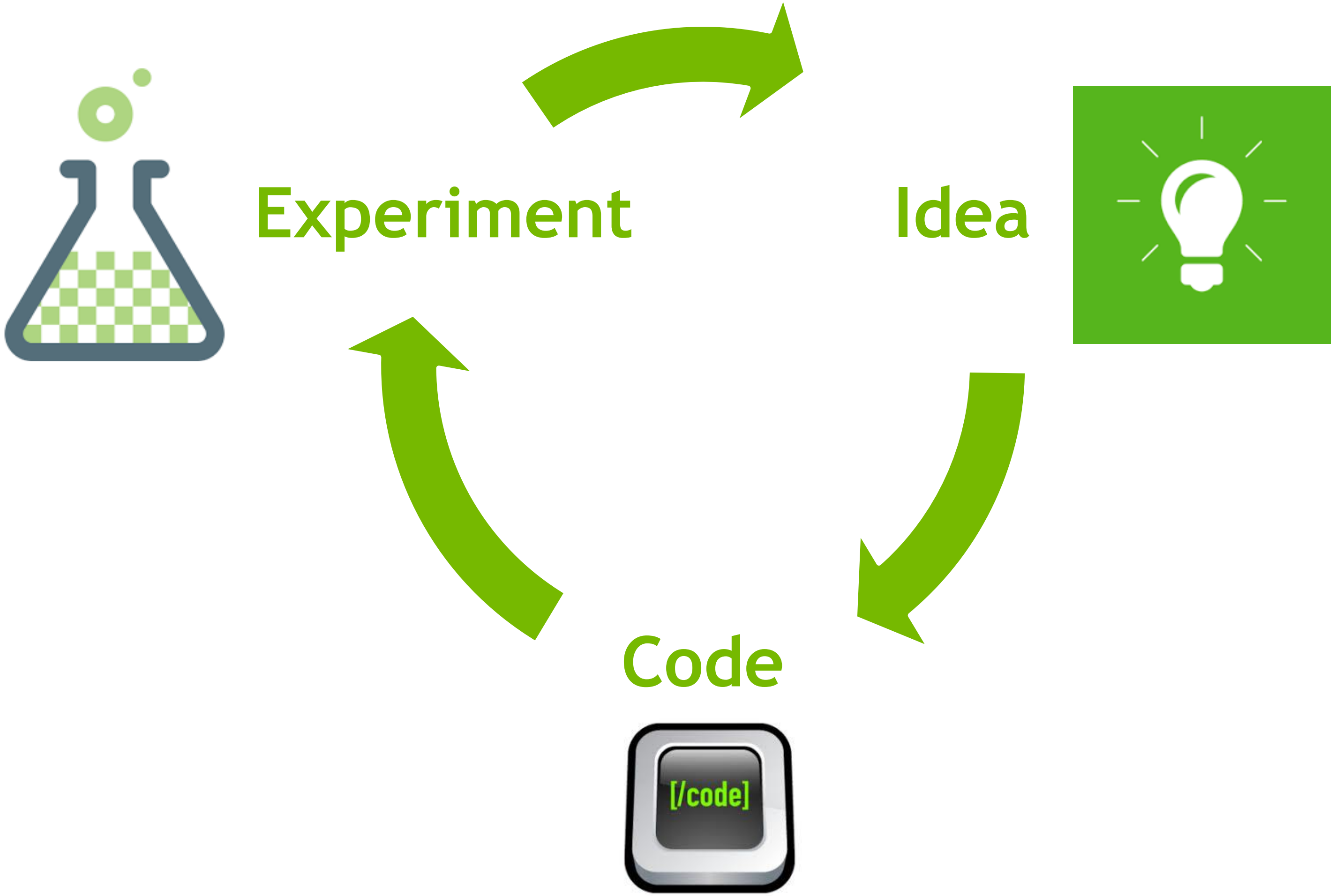


2017 - Google Neural Machine Translation
Near Human Language Translation

100 EXAFLOPS
=
2 YEARS ON A DUAL CPU SERVER

IMPLICATIONS

Experimental Nature of Deep Learning - Unacceptable training time

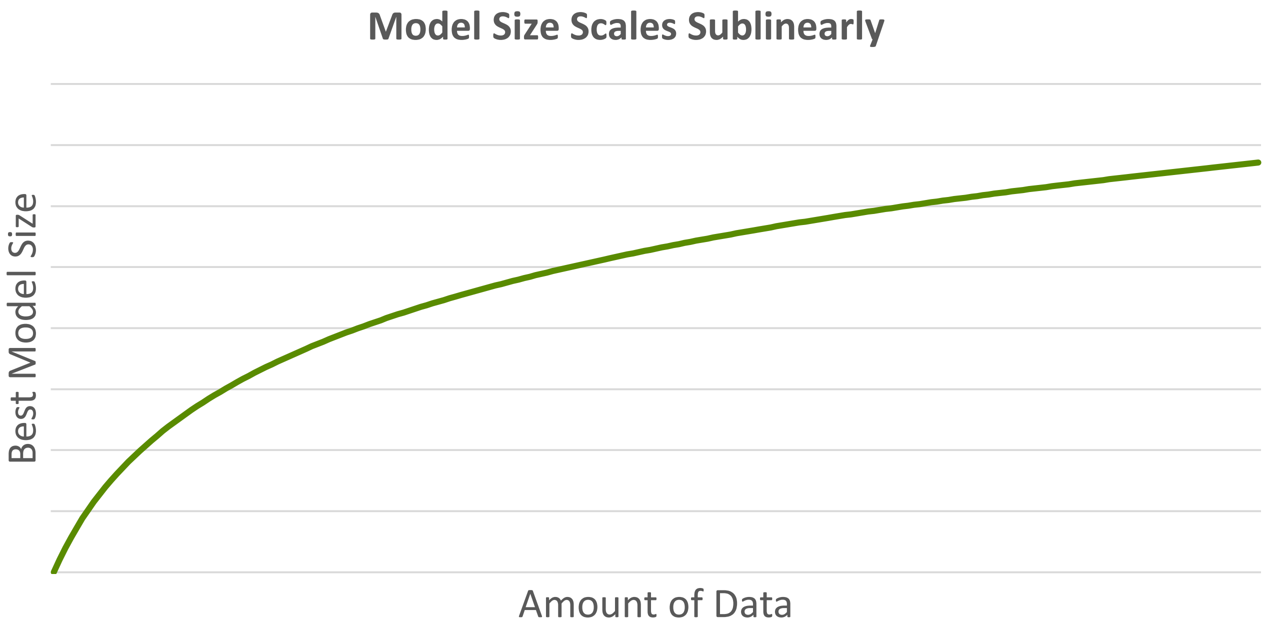


AGENDA

- Introduction
- Data & Models
- Algorithms & hyperparameters
- Software architecture & environment
- Infrastructure
- People
- Conclusion

EXPLODING MODEL COMPLEXITY

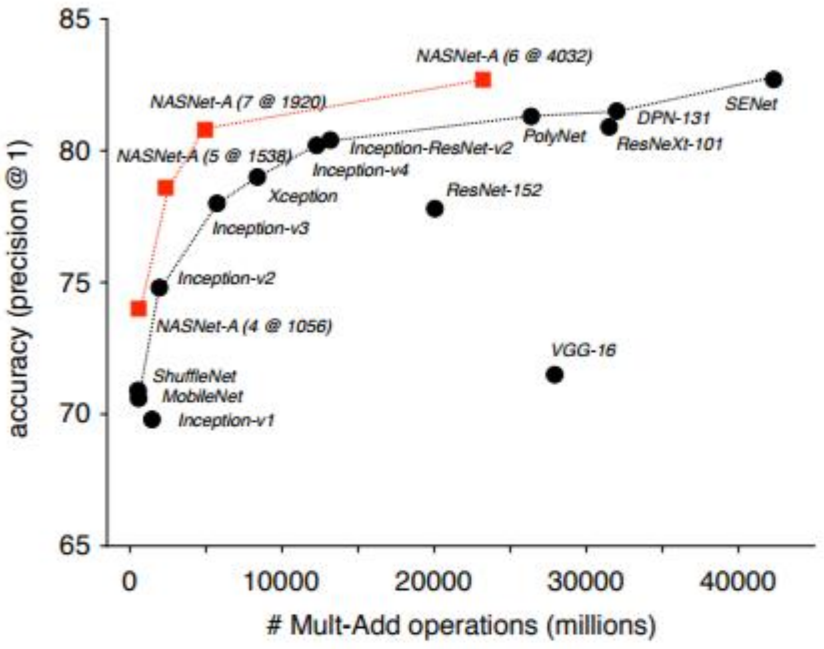
Good news – model size scales sublinearly

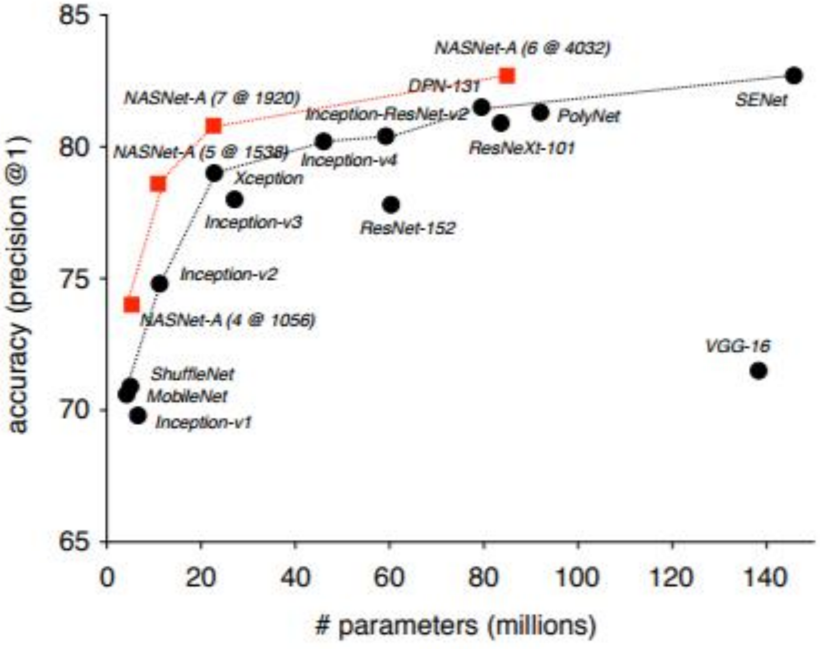


Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv preprint arXiv:1712.00409.

EVIDENCE FROM IMAGE PROCESSING

Good news – model size scales sublinearly

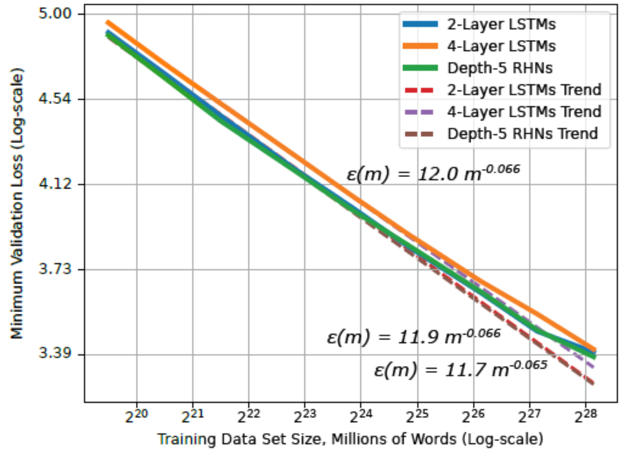




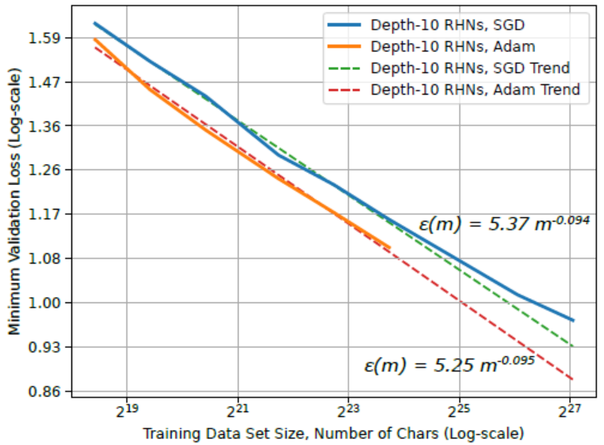
Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." arXiv preprint arXiv:1707.07012 (2017).

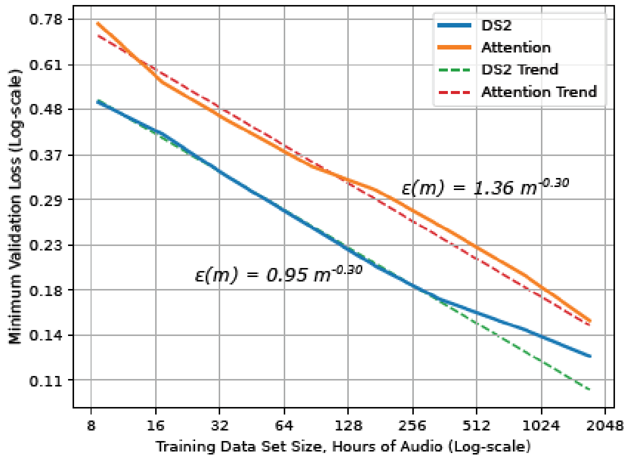
EXPLODING DATASETS

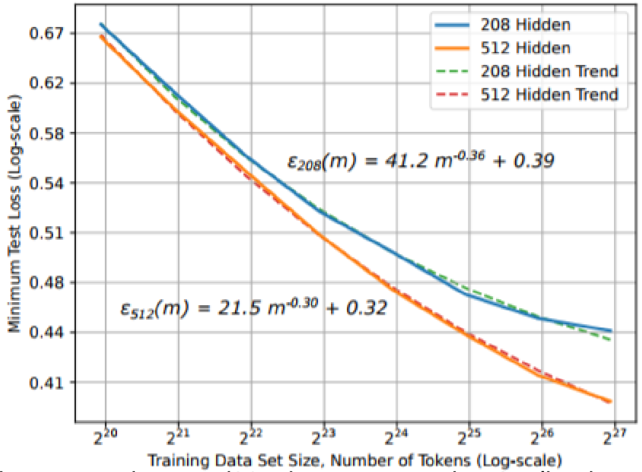
Logarithmic relationship between the dataset size and accuracy

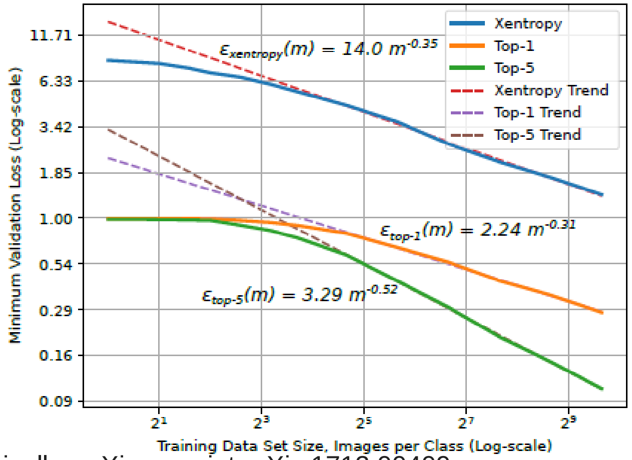


- Translation
- Language Models
- Character Language Models
- Image Classification
- Attention Speech Models





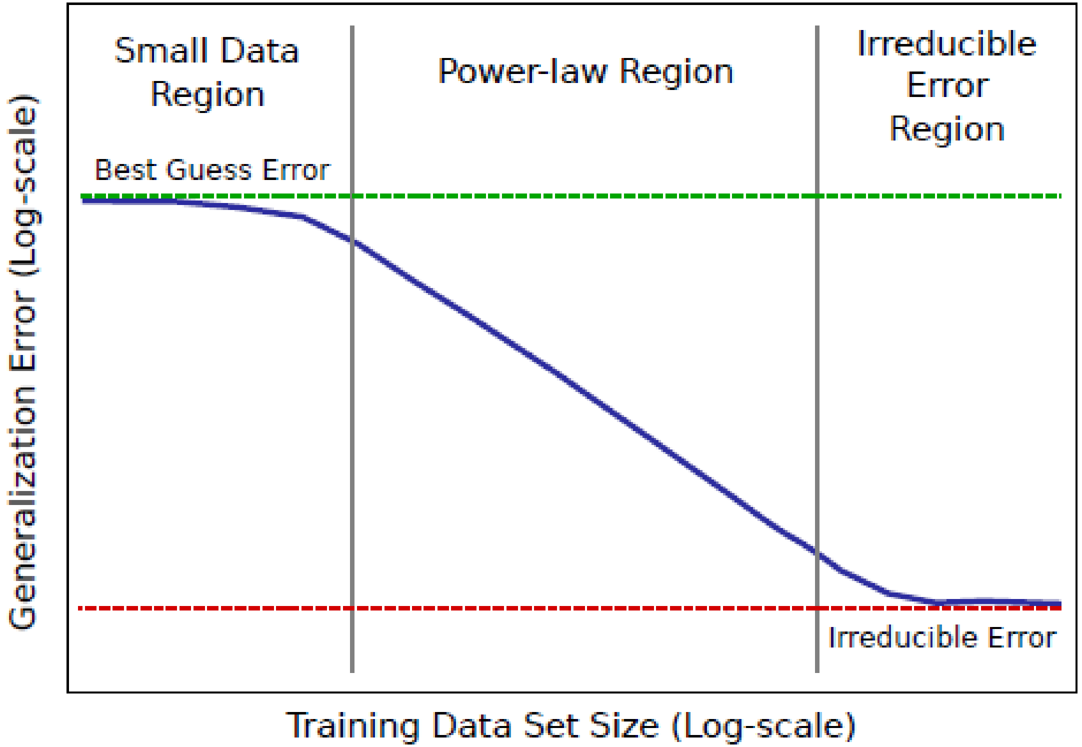




Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv preprint arXiv:1712.00409.

EXPLODING DATASETS

Logarithmic relationship between the dataset size and accuracy

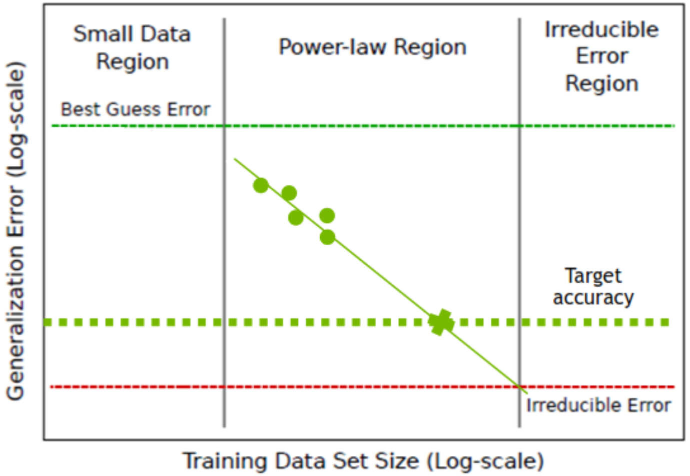


Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv preprint arXiv:1712.00409.

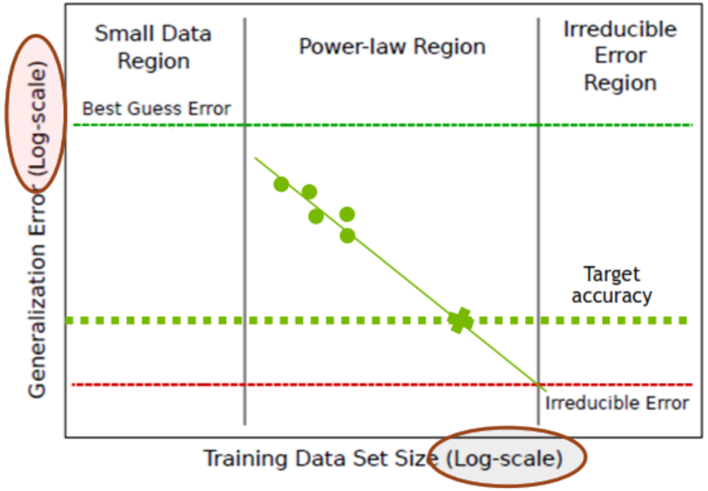
EXPLODING DATASETS

The good news – you can calculate how much data you need

Step 3: Interpolate



The bad news – log-scale



IMPLICATIONS

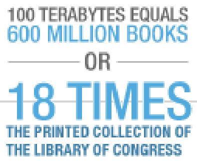
Good and bad news

- The good news: Requirements are predictable.
  - We can predict how much data we will need
  - We can predict how much computing power we will need
- The bad news: The values can be significant.

IMPLICATIONS

Automotive example

Majority of useful problems are too complex for a single GPU training

	VERY CONSERVATIVE	CONSERVATIVE
Fleet size (data capture per hour)	100 cars / 1TB/hour	125 cars / 1.5TB/hour
Duration of data collection	260 days * 8 hours	325 days * 10 hours
Data Compression factor	0.0005	0.0008
Total training set	104 TB 	 487.5 TB
InceptionV3 training time (with 1 Pascal GPU)	9.1 years	42.6 years 
AlexNet training time (with 1 Pascal GPU)	 1.1 years 	 5.4 years 

ALGORITHMS & HYPERPARAMETERS

ALGORITHMIC IMPROVEMENTS

Stochastic Gradient Descent Variants : Asynchronous SGDs

Gradient Compression

- Lin, Yujun, et al. "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training." *arXiv preprint arXiv:1712.01887* (2017).

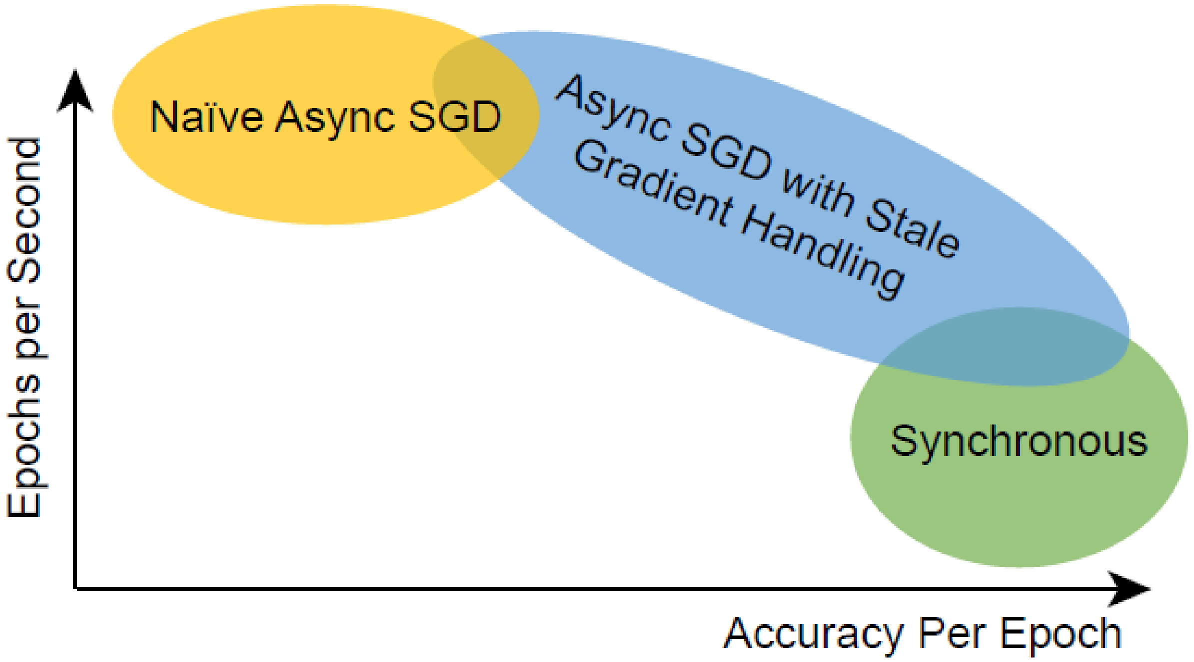
- Wei, Bingzhen, et al. "Minimal effort back propagation for convolutional neural networks." arXiv preprint arXiv:1709.05804 (2017).

Improved model architecture : Communication efficient design

- Iandola, Forrest N., et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size." *arXiv preprint arXiv:1602.07360* (2016).

STOCHASTIC GRADIENT DESCENT VARIANTS

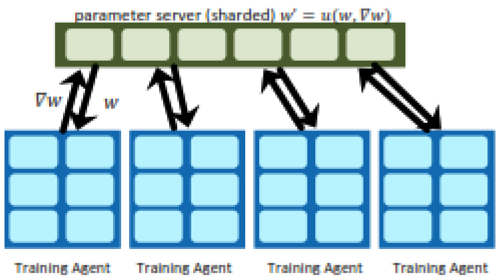
Continuous space

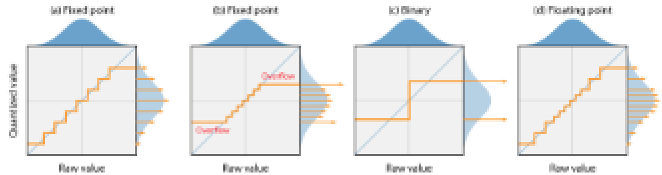


COMMUNICATION OPTIMIZATION

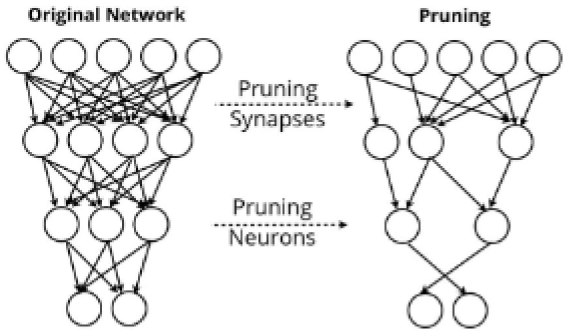
Communication optimizations

- **Different options how to optimize updates**
  - Send ∇w, receive w
  - Send FC factors (o_{l-1}, o_l), compute ∇w on parameter server
    Broadcast factors to not receive full w
  - Use lossy compression when sending, accumulate error locally!
- **Quantization**
  - Quantize weight updates and potentially weights
  - Main trick is stochastic rounding [1] – expectation is more accurate
    Enables low precision (half, quarter) to become standard
  - TernGrad - ternary weights [2], 1-bit SGD [3], …
- **Sparsification**
  - Do not send small weight updates **or** only send top-k [4]
    Accumulate them locally





source: ai.intel.com



[1] S. Gupta et al. Deep Learning with Limited Numerical Precision, ICML'15
[2] F. Li and B. Liu. Ternary Weight Networks, arXiv 2016
[3] F. Seide et al. 1-Bit Stochastic Gradient Descent and Application to Data-Parallel Distributed Training of Speech DNNs, In Interspeech 2014
[4] C. Renggli et al. SparCML: High-Performance Sparse Communication for Machine Learning, arXiv 2018

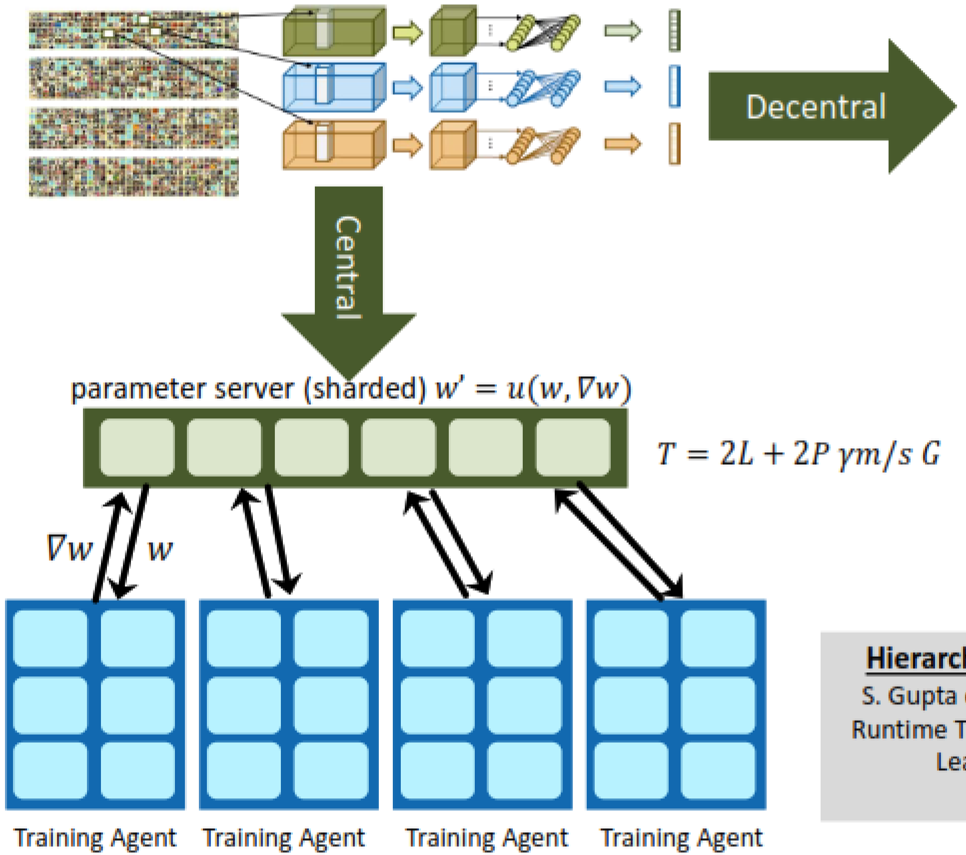
15

Source T. Hoefler

What every HPC researcher should know about (scalable) parallel deep learning

COMMUNICATION OPTIMIZATION

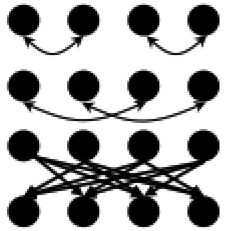
Updating parameters in distributed data parallelism



$T = 2L + 2P\ \gamma m/s\ G$

MPI

- Collective operations
- Topologies
- Neighborhood collectives
- RMA?



$T = 2L\log_2 P + 2\gamma m G(P-1)/P$

Hierarchical Parameter Server

S. Gupta et al.: Model Accuracy and Runtime Tradeoff in Distributed Deep Learning: A Systematic Study. ICDM'16

Adaptive Minibatch Size

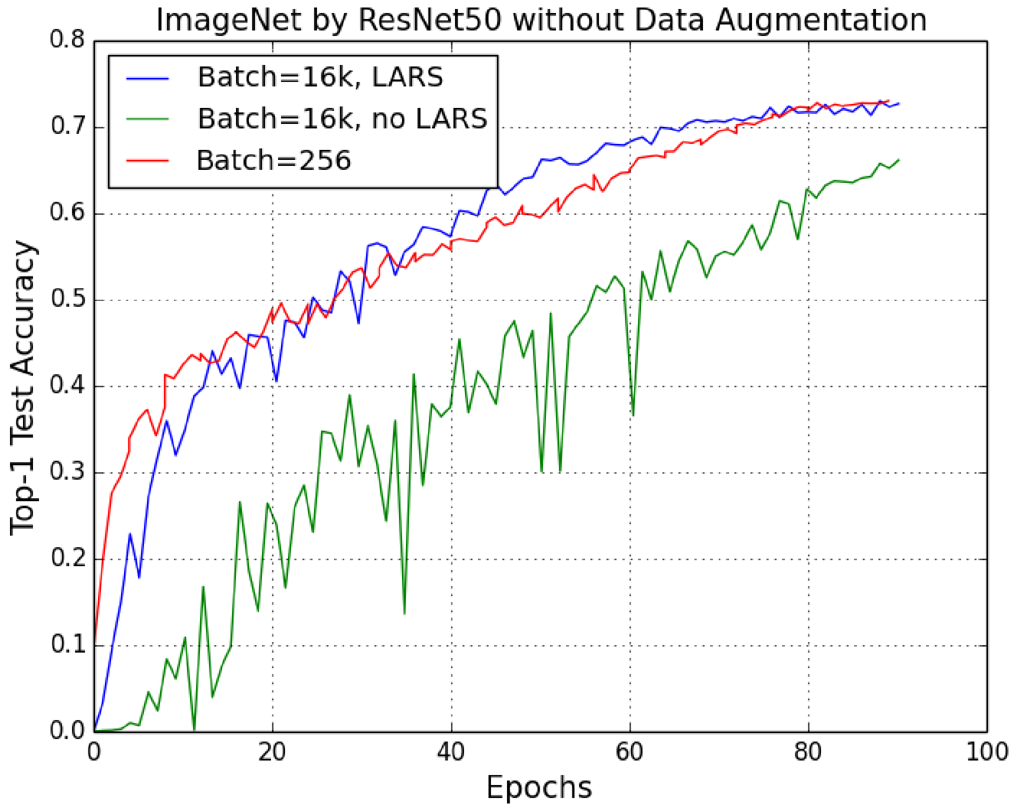
S. L. Smith et al.: Don't Decay the Learning Rate, Increase the Batch Size, arXiv 2017

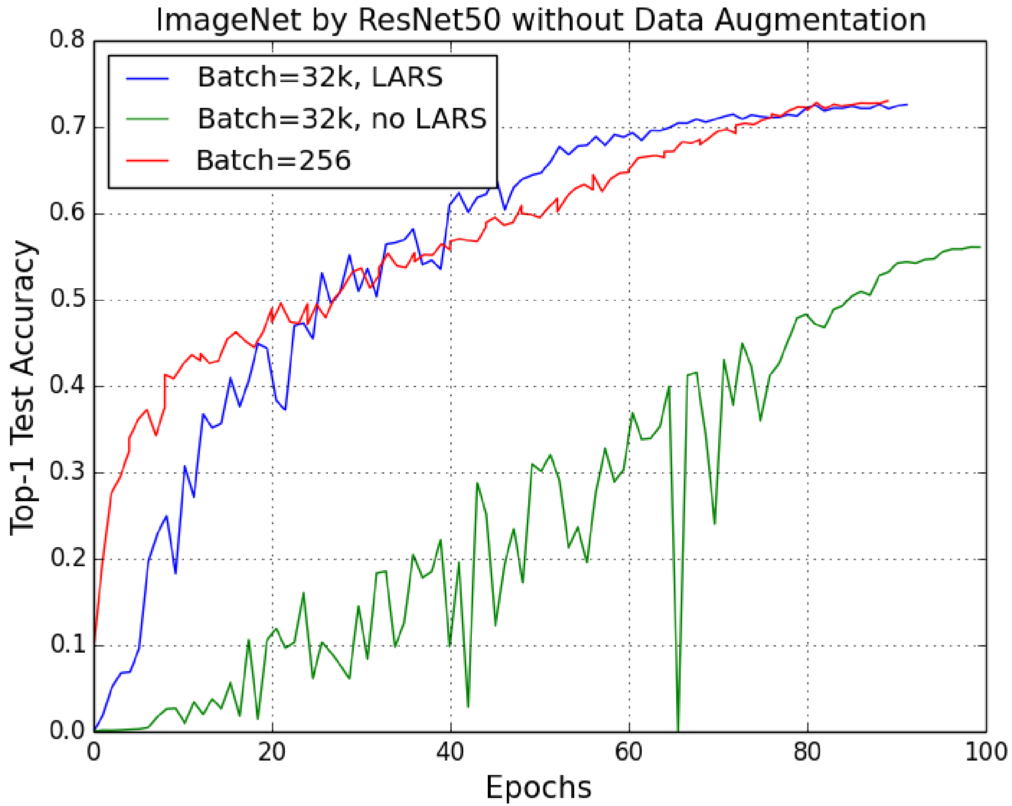
Source T. Hoefler

What every HPC researcher should know about (scalable) parallel deep learning

LARGE MINIBATCH - IMPACT ON ACCURACY

Naïve approaches lead to degraded accuracy

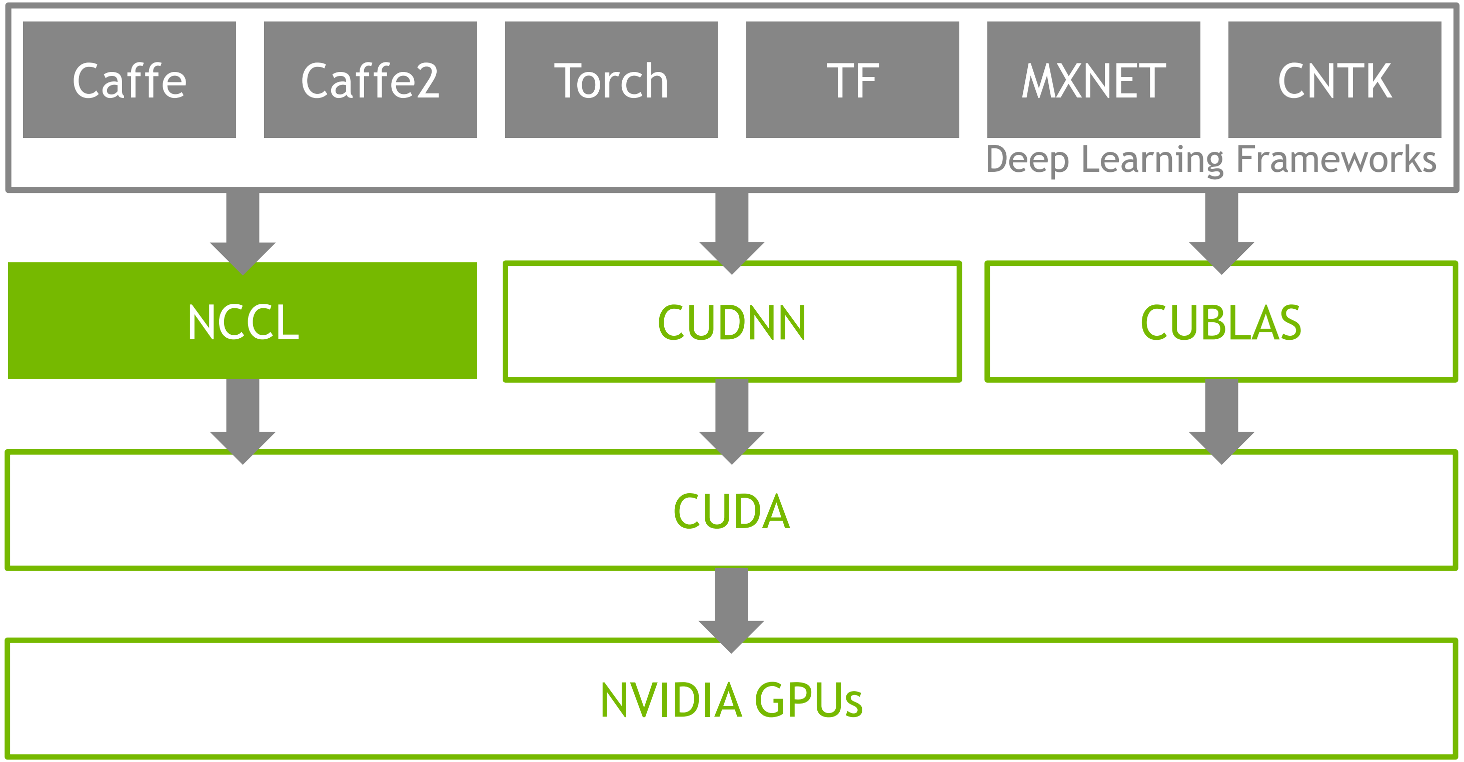




You, Y., Zhang, Z., Hsieh, C., Demmel, J., & Keutzer, K. (2017). ImageNet training in minutes. *CoRR, abs/1709.05011*.

SOFTWARE CONSIDERATIONS

SOFTWARE ARCHITECTURE



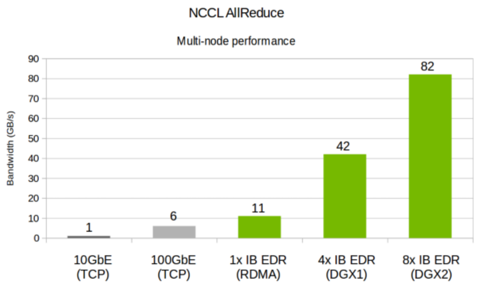
NVIDIA Collective Communications Library (NCCL)

Multi-GPU and multi-node collective communication primitives

High-performance multi-GPU and multi-node collective communication primitives optimized for NVIDIA GPUs

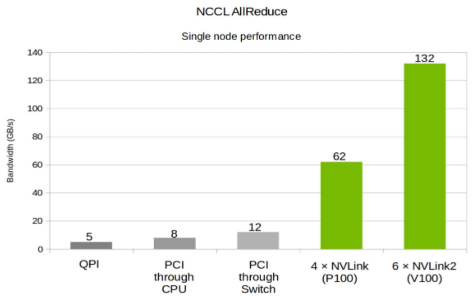
- Fast routines for multi-GPU multi-node acceleration that maximizes inter-GPU bandwidth utilization
- Easy to integrate and MPI compatible.
- Accelerates leading deep learning frameworks

NCCL INTER-NODE PERFORMANCE

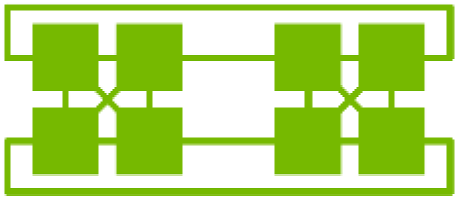


Bandwidth scaling with different topologies over multiple nodes

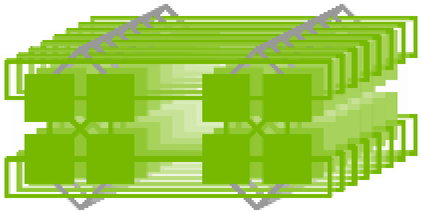
NCCL INTRA-NODE PERFORMANCE



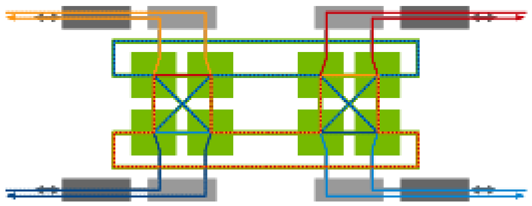
NCCL performance scaling on a single GPU with different topologies



Multi-GPU:
NVLink
PCIe



Multi-Node:
InfiniBand
IP Sockets
RoCE

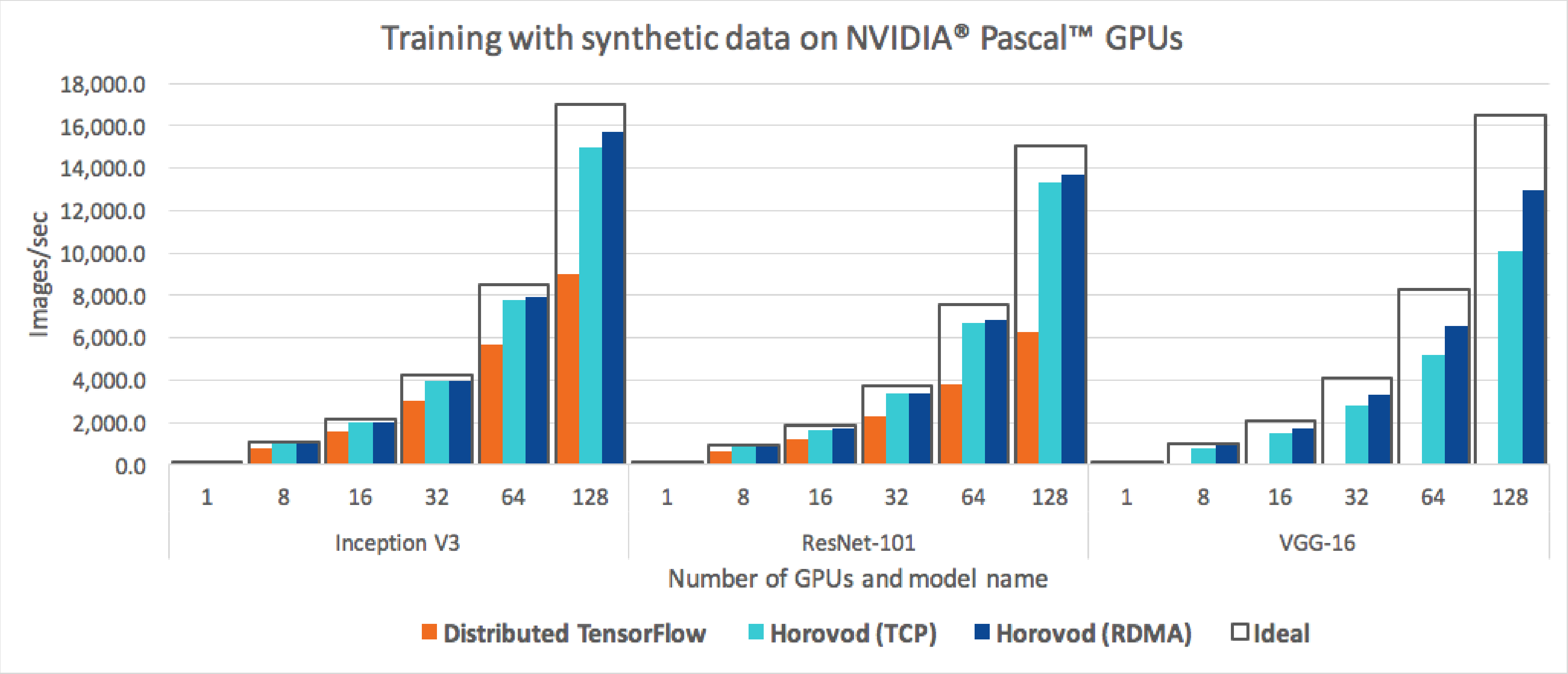


Automatic Topology Detection

developer.nvidia.com/nccl   https://github.com/NVIDIA/nccl

TRAINING AT SCALE

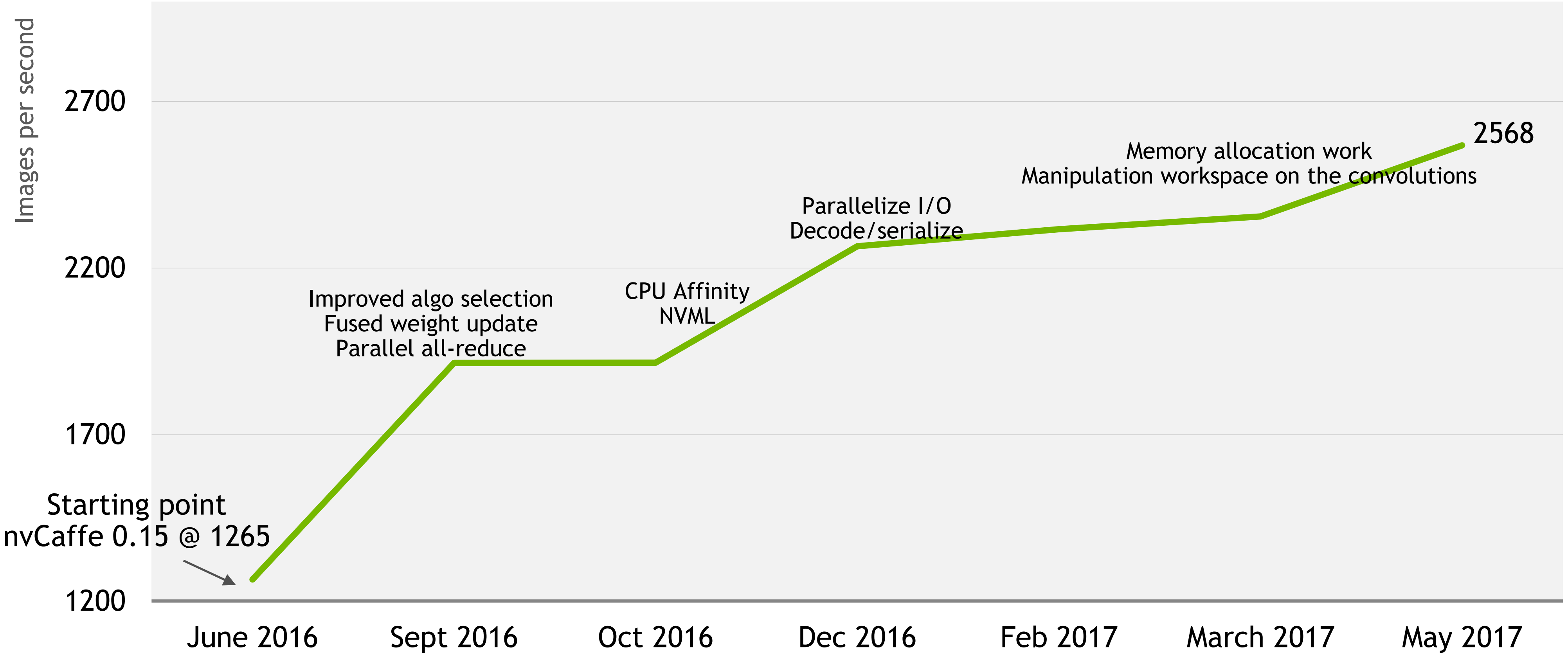
How do you train efficiently on larger systems



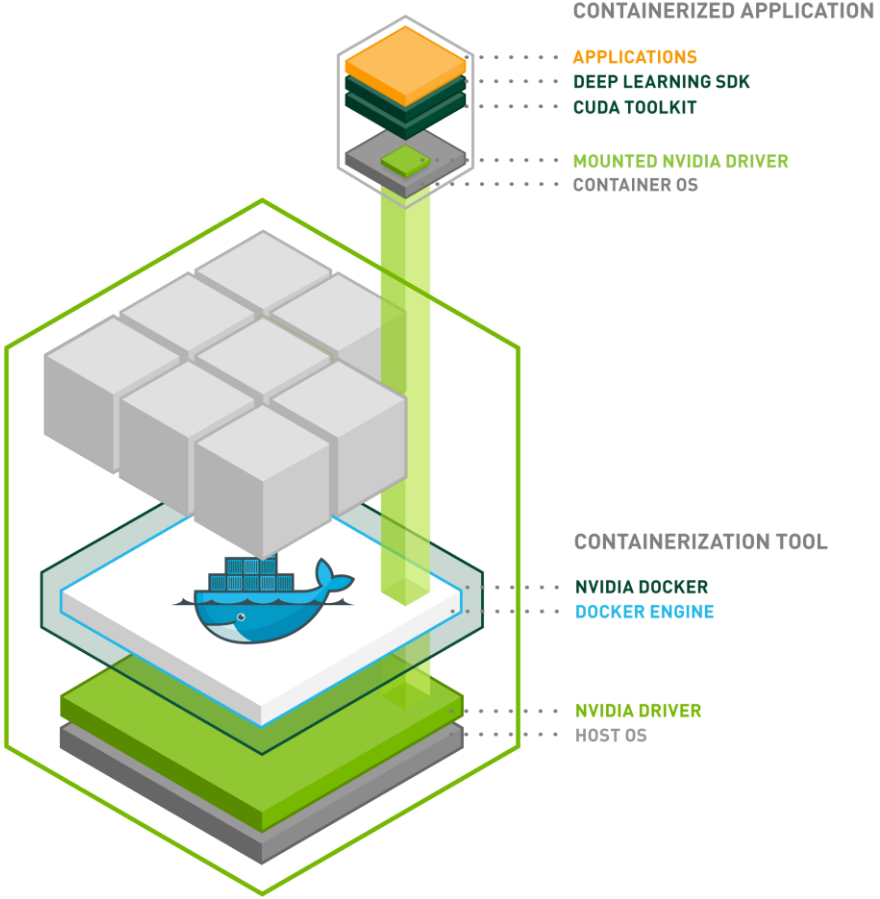
DEEP LEARNING FRAMEWORKS

FW	Productivity layer	NV Optimized	Multi-GPU	Multi-Node with NCCL	Mixed Precision
Caffe	DIGITS	Yes	NCCL		Yes
Caffe 2		Yes	NCCL	WIP	WIP
Chainer		WIP	proprietary		No
Cognitive Toolkit		Yes	NCCL/MPI	Yes	No
DL4J		No	N/A		No
Matlab	Own	No	N/A		No
MXNet		Yes	NCCL/ peer to peer		WIP
PyTorch		Yes	NCCL	WIP	WIP
TensorFlow	Keras/DIGITS	Yes	NCCL/ Proprietary	Sockets/ GRPC	WIP
Theano	Keras	Yes	Limited		No
Torch	DIGITS	Yes	Yes		No

NVCAFFE V0.16 TRAINING ALEXNET



Single P100 GPU, Batch Size=128



NVIDIA GPU CLOUD (NGC)

INNOVATE IN MINUTES, NOT WEEKS WITH DEEP LEARNING CONTAINERS

Benefits of Containers: Monthly updates

Simplify deployment of GPU-accelerated applications, eliminating time-consuming software integration work

Isolate individual frameworks or applications

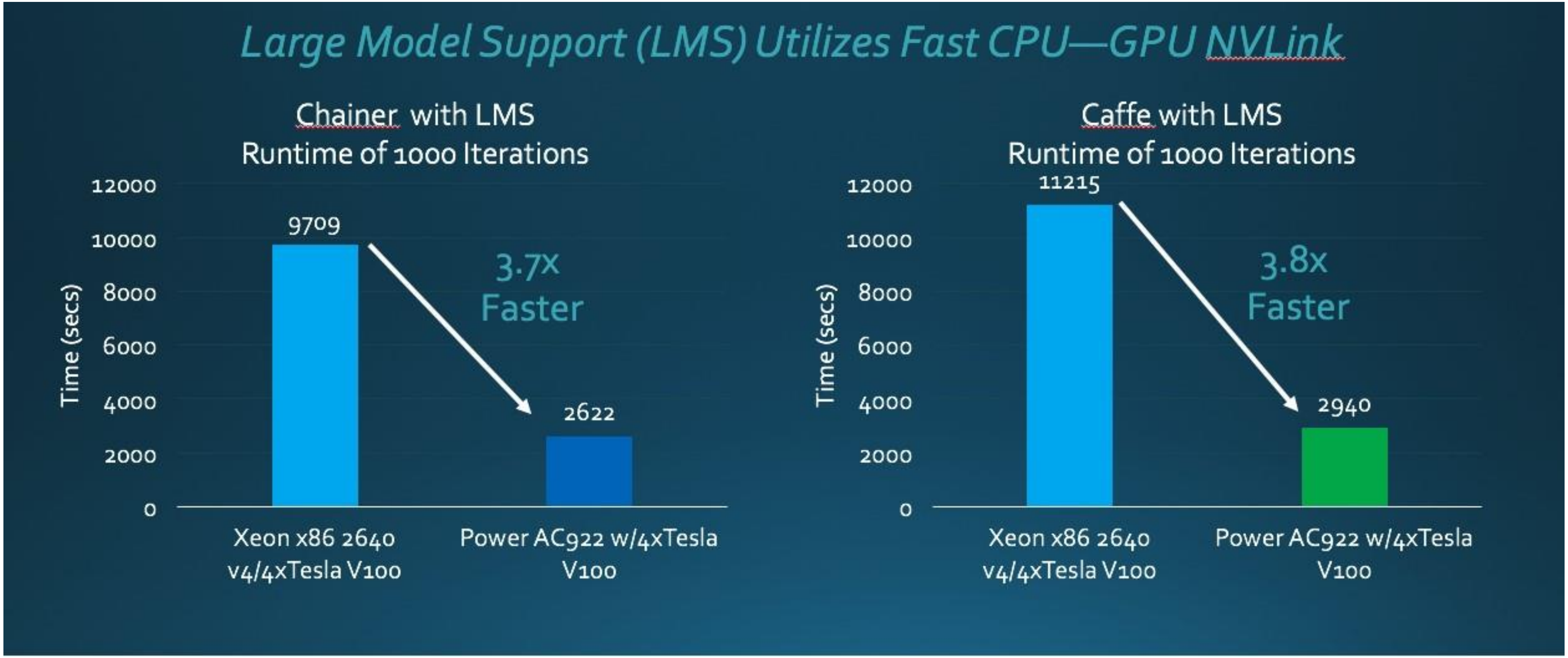
Share, collaborate, and test applications across different environments

BALANCED SYSTEM DESIGNED FOR SCALE

ENGINEERING CHALLENGES

- Data Input Pipeline
  - Storage
  - Networking
  - Augmentation
- Communication
- Reference Architecture
- Other

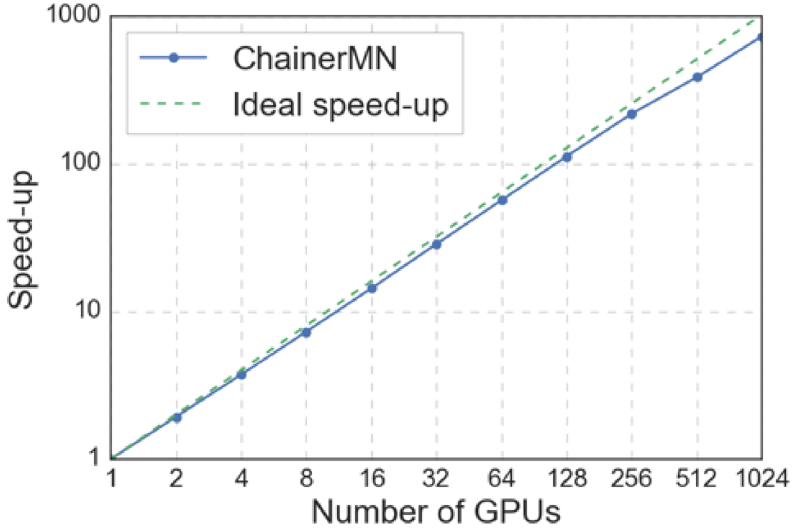
IBM POWER9 ARCHITECTURE
CPU-GPU FAST CONNECTION WITH NVLink2



PREFERRED NETWORKS

Training ImageNet in 15 minutes

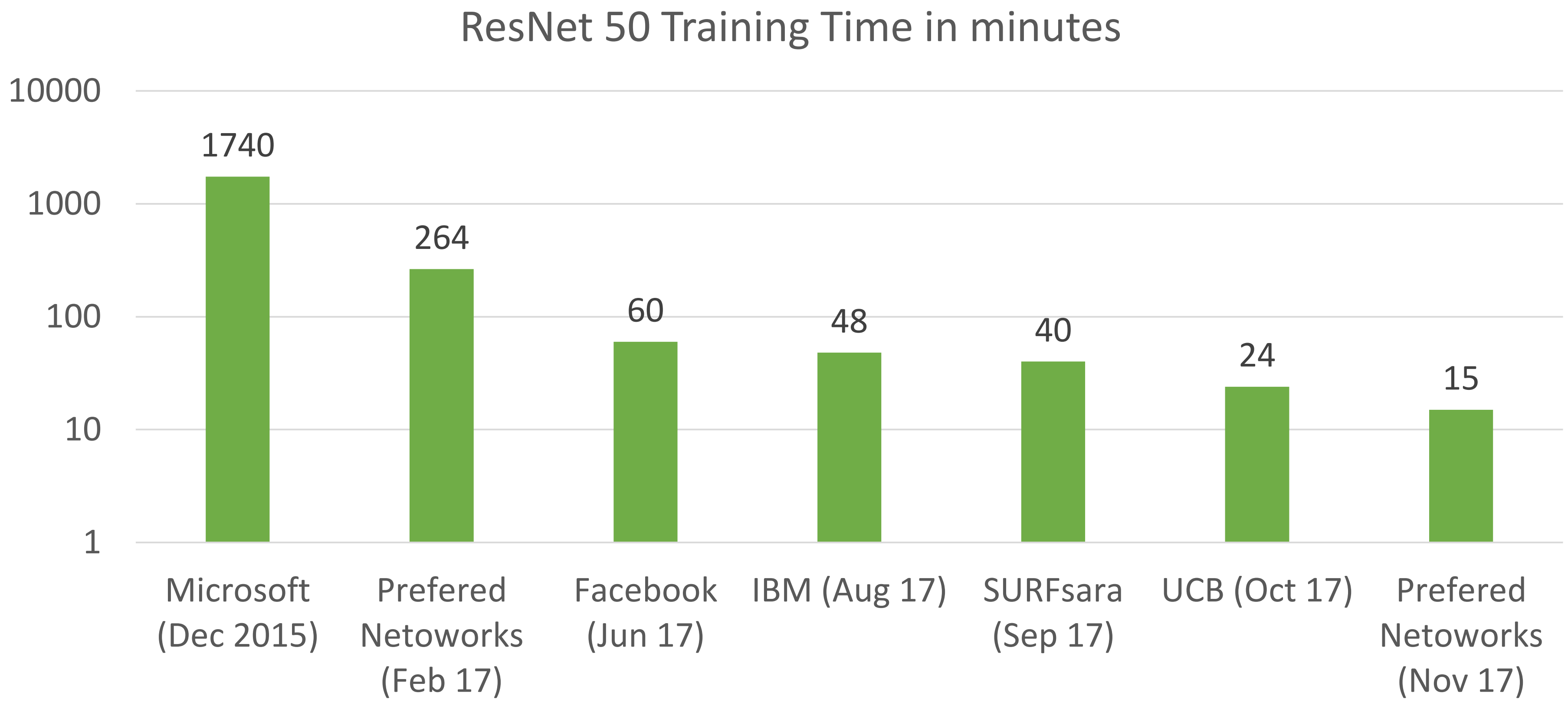
- It consists of 128 nodes with 8 NVIDIA P100 GPUs each, for **1024 GPUs in total**.
- The nodes are connected with two FDR Infiniband links (56Gbps x 2).



Akiba, T., Suzuki, S., & Fukuda, K. (2017). Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes. *arXiv preprint arXiv:1711.04325*.

ITERATION TIME

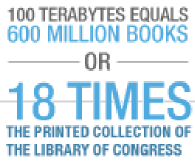
Short iteration time is fundamental for success



IMPLICATIONS

Automotive example

Majority of useful problems are too complex for a single GPU training

	VERY CONSERVATIVE	CONSERVATIVE
Fleet size (data capture per hour)	100 cars / 1TB/hour	125 cars / 1.5TB/hour
Duration of data collection	260 days * 8 hours	325 days * 10 hours
Data Compression factor	0.0005	0.0008
Total training set	104 TB 	 487.5 TB
InceptionV3 training time (with 1 Pascal GPU)	9.1 years	42.6 years 
AlexNet training time (with 1 Pascal GPU)	 1.1 years 	 5.4 years 

CONCLUSIONS

Need to scale the training process for a single job

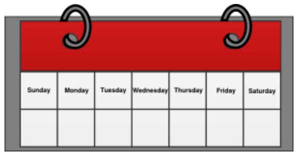
1 NVIDIA DGX-1

10 NVIDIA DGX-1's

	VERY CONSERVATIVE	CONSERVATIVE
Total training set	104 TB	487.5 TB
InceptionV3 (one DGX-1V)	166 days (5+ **months**)	778 days (2+ **years**)
AlexNet (one DGX-1V)	21 days (3 **weeks**)	98 days (3 **months**)
InceptionV3 (10 DGX-1V's)	16 days (2+ **weeks**)	77 days (11 **weeks**)
AlexNet (10 DGX-1V's)	2.1 days	9.8 days

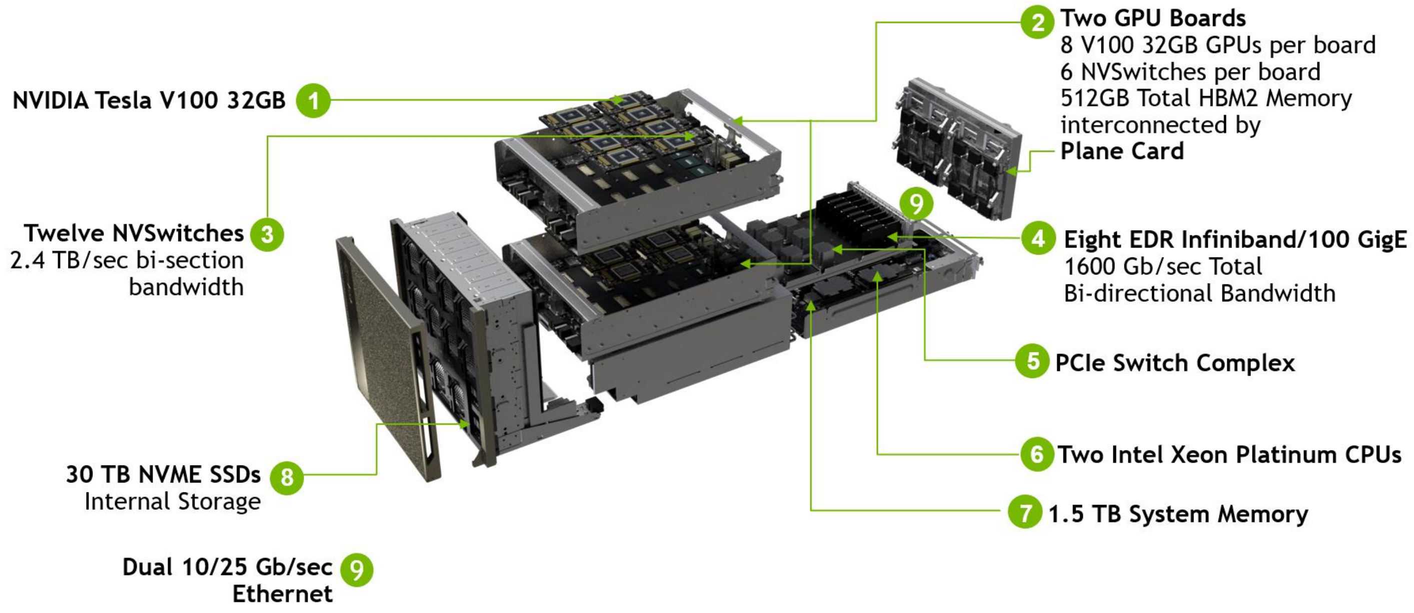
Training From Months or Years

To Weeks or Days



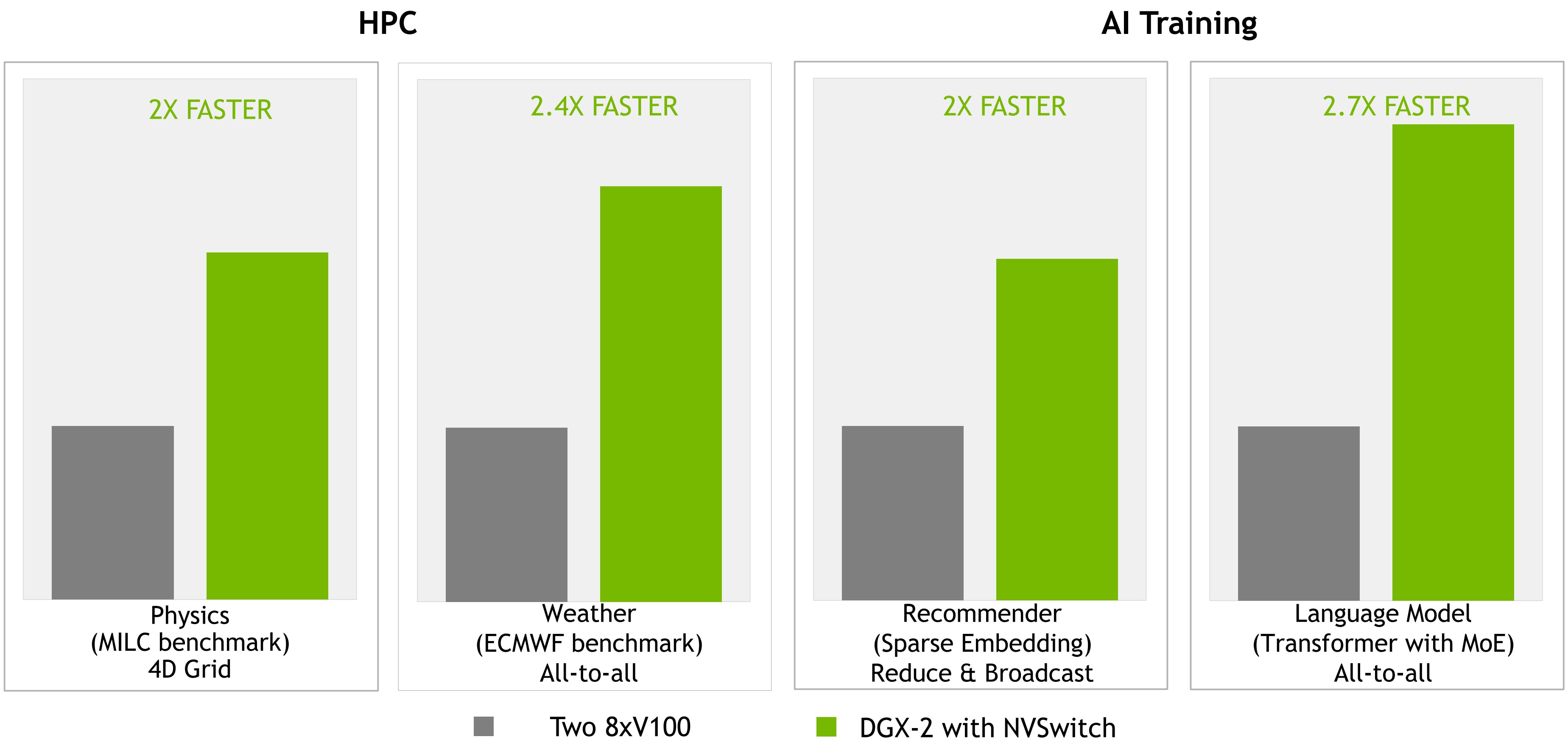
DGX FAMILY

DGX-2 – 2 PetaFLOPs



OVER 2X HIGHER PERFORMANCE WITH NVSWITCH

DGX-2 vs Multi-System Interconnect



2 8xV100 servers have dual socket Xeon E5 2698v4 Processor. 8 x V100 GPUs. Servers connected via 4X 100Gb IB ports
DGX-2 server has dual-socket Xeon Platinum 8168 Processor. 16 V100 GPUs

SATURN V

660 DGX-1 Volta Nodes

- 660 Nodes with a total of 5280 Volta GPUs
- 660 PFLOPs for AI training

MULTI-NODE DGX
"A-HA" MOMENTS IN DL CLUSTER DESIGN

Additional design insights to get you started

Overall Cluster	Rack Design	Networking	Storage	Facilities	Software
• HPC similar to DL • HPC expertise can help in design • Even with HPC, the similarities are limited	• DL drives close to operational limits; • Assume less headroom • Proper airflow is crucial to cluster performance	• Like HPC, InfiniBand is preferred • Require high bandwidth, low latency • Maximize per-node IB connections	• DGX-1 read cache is critical • Datasets range from 10k's to millions objects • Terabyte levels of storage • Large variance	• GPU data center operates at near-max power • Assume higher watts per-rack • Dramatically higher FLOPS/watt = floor space saved	• Scale requires "cluster-aware" software • NCCL2 = GPU/multi-node acceleration • Automatic topology detect • DL framework optimizations

MULTI-NODE SCALING WHITEPAPER

Use this asset to aid the design process, ensuring you develop the optimized architecture for your multi-node cluster, following NVIDIA best practices learned from our customer deployments and our own DGX SATURNV

NIDIA DGX Data Center Reference Design

White Paper

NVIDIA® DGX™ Data Center Reference Design

Easy Deployment of DGX Servers for Deep Learning

2018-08-29

IDIA AI Software

software running on the DGX POD provides a high-performance DL training
ent for large scale multi-user AI software development teams. NVIDIA AI software
he DGX operating system (DGX OS), cluster management and orchestration tools,
raries and frameworks, workload schedulers, and optimized containers from the NGC
registry. To provide additional functionality, the DGX POD management software

y open-source tools recommended by NVIDIA which have been tested to work on DGX
the NVIDIA AI software stack. Support for these tools can be obtained directly through
y support structures.

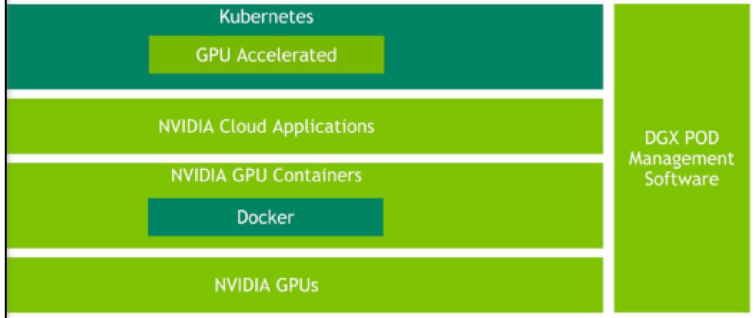


Figure 2. NVIDIA AI software

DGX POD — DGX-1

Reference Architecture in a Single 35 kW High-Density Rack

Fit within a standard-height 42 RU data center rack

- Nine DGX-1 servers (9 x 3 RU = 27 RU)
- Twelve storage servers (12 x 1 RU = 12 RU)
- 10 GbE (min) storage and management switch (1 RU)
- Mellanox 100 Gbps intra-rack high speed network switches (1 or 2 RU)

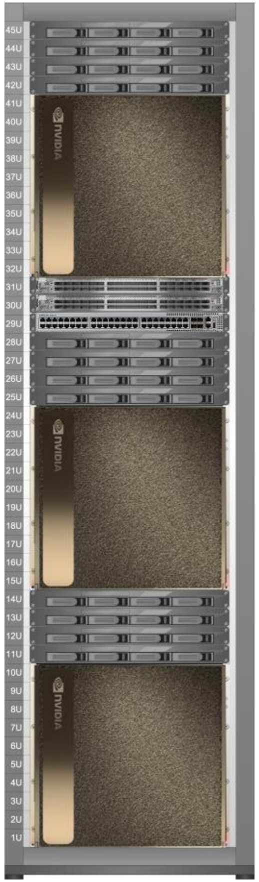
In real-life DL application development, one to two DGX-1 servers per developer are often required
One DGX POD supports five developers (AV workload)
Each developer works on two experiments per day
One DGX-1/developer/experiment/day*
*300,000 0.5M images * 120 epochs @ 480 images/sec Resnet-18 backbone detection network per experiment

DGX POD — DGX-2

Reference Architecture in a Single 35 kW High-Density Rack

Fit within a standard-height 48 RU data center rack

- Three DGX-2 servers (3 x 10 RU = 30 RU)
- Twelve storage servers (12 x 1 RU = 12 RU)
- 10 GbE (min) storage and management switch (1 RU)
- Mellanox 100 Gbps intra-rack high speed network switches (1 or 2 RU)



In real-life DL application development, one DGX-2 per developer minimizes model training time
One DGX POD supports at least three developers (AV workload)
Each developer works on two experiments per day
One DGX-2/developer/2 experiments/day*
*300,000 0.5M images * 120 epochs @ 480 images/sec Resnet-18 backbone detection network per experiment

PEOPLE

DEEP LEARNING INSTITUTE

DLI Mission: Help the world to solve the most challenging problems using AI and deep learning

We help developers, data scientists and engineers to get started in architecting, optimizing, and deploying neural networks to solve real-world problems in diverse industries such as autonomous vehicles, healthcare, robotics, media & entertainment and game development.

APPLYING DEEP LEARNING

Every major DL framework leverages NVIDIA SDKs

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION

SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

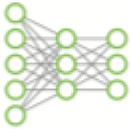
TIME SERIES

ANOMALY DETECTION & CLASSIFICATION

REINFORCEMENT LEARNING

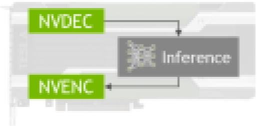
NVIDIA DEEP LEARNING SDK

cuDNN

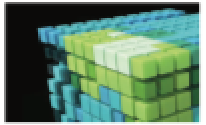


TensorRT

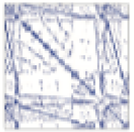
DeepStream SDK



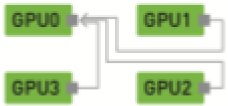
cuBLAS



cuSPARSE



NCCL



MULTI GPU TRAINING

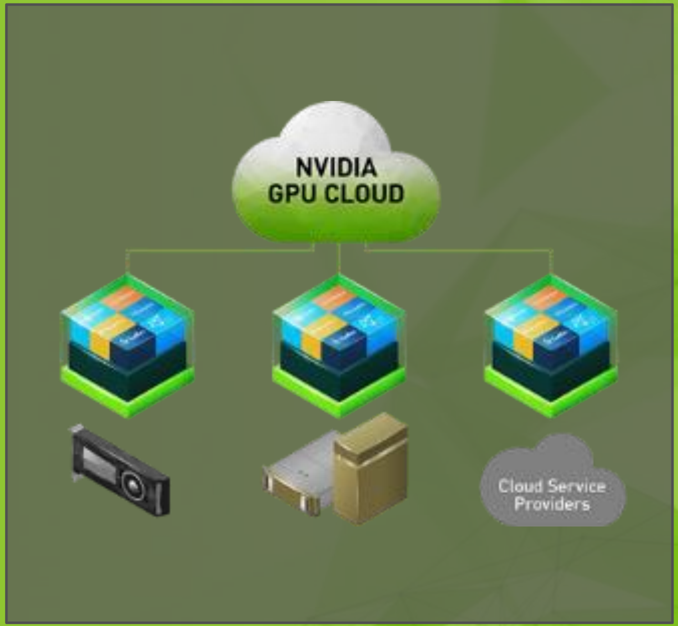
NVIDIA.

DEEP
LEARNING
INSTITUTE

CONCLUSION : NEXT STEPS

GTC Munich | October 9-11 2018
www.nvidia.com/

NVIDIA Deep Learning Institute
www.nvidia.com/en-us/deep-learning-ai/education



NGC
www.nvidia.com/en-us/gpu-cloud