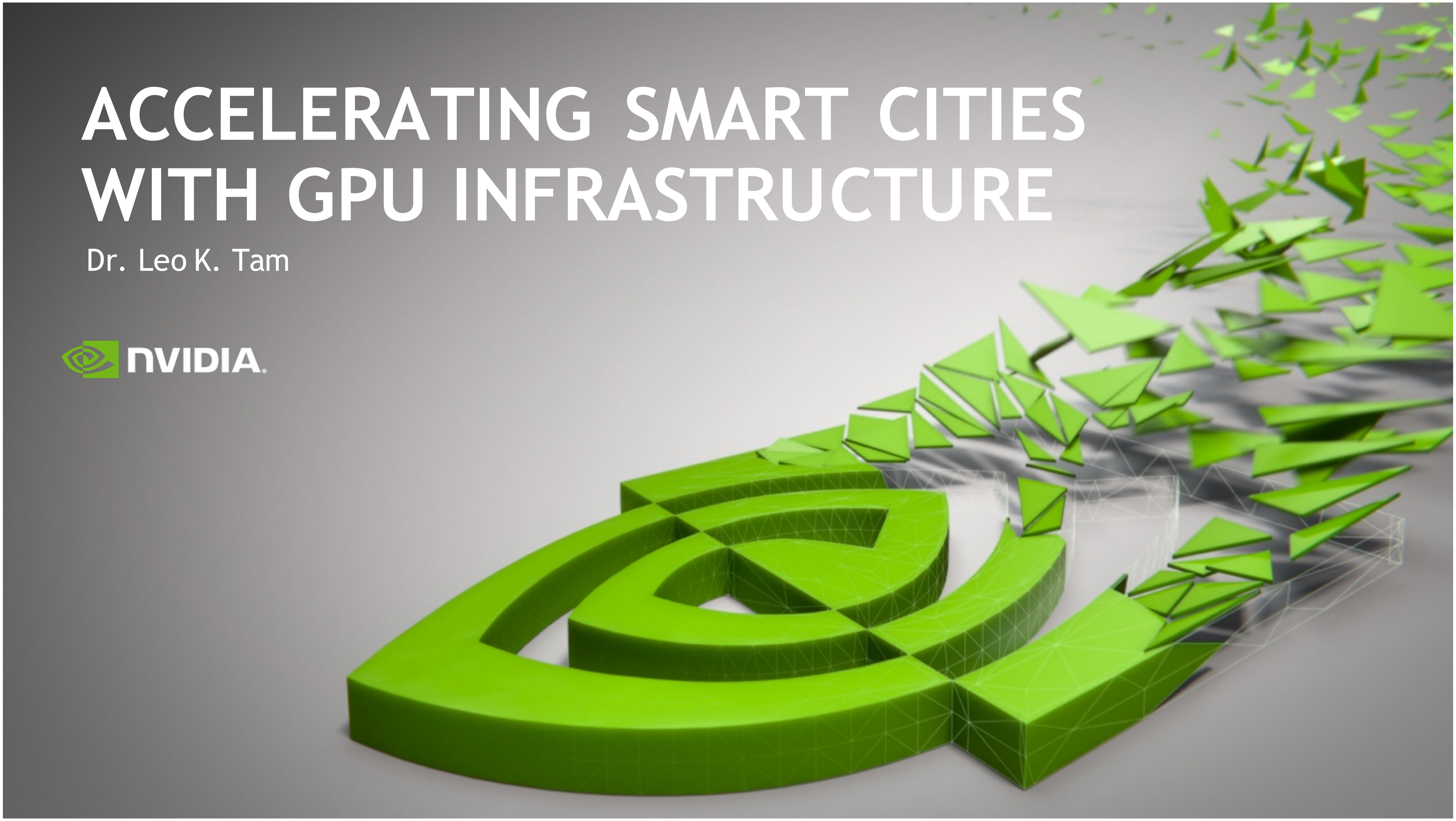
ACCELERATING SMART CITIES WITH GPU INFRASTRUCTURE

Dr. Leo K. Tam

NVIDIA



MEGATRENDS ARE DRIVING WORLD CITIES

URBANIZATION

↑2.5Bn

Urban pop. growth

United Nations DESA

DIGITIZATION

50Bn

Connected things by 2020

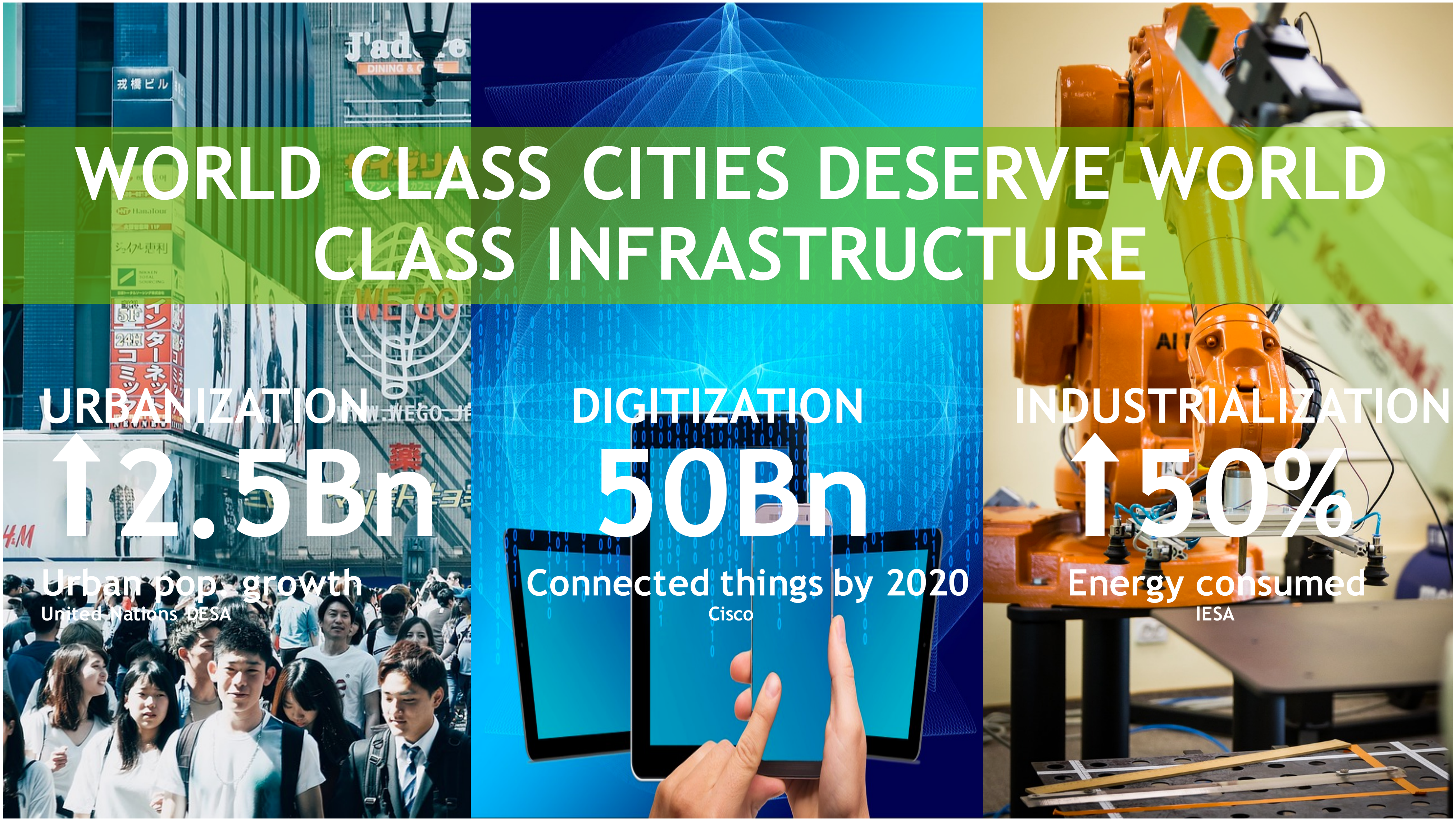
Cisco

INDUSTRIALIZATION

↑50%

Energy consumed

IESA



WORLD CLASS CITIES DESERVE WORLD CLASS INFRASTRUCTURE

URBANIZATION

↑2.5Bn

Urban pop. growth

United Nations DESA

DIGITIZATION

50Bn

Connected things by 2020

Cisco

INDUSTRIALIZATION

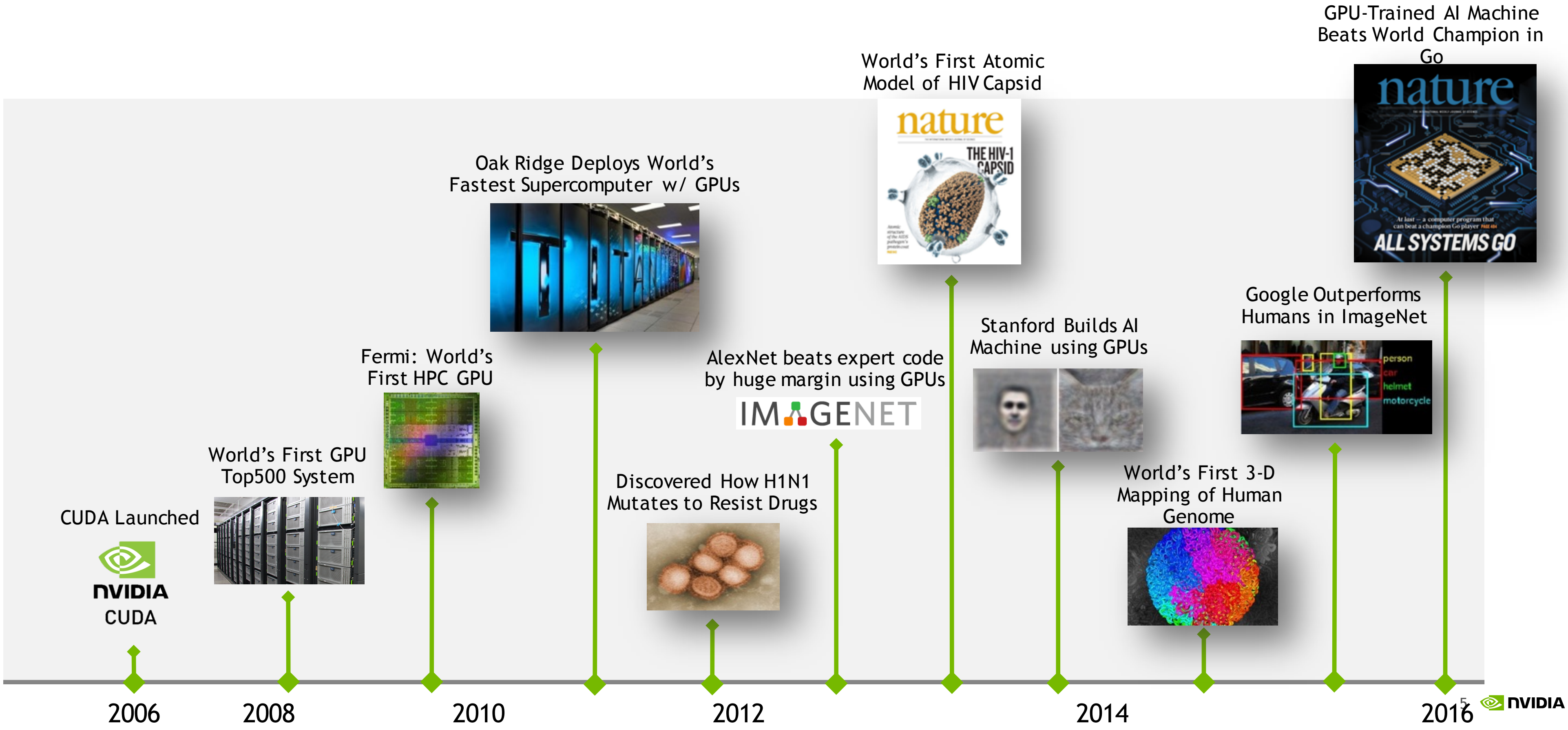
↑50%

Energy consumed

IESA

[redacted]

TEN YEARS OF GPU COMPUTING



CUDA Launched

NVIDIA CUDA

World's First GPU Top500 System

Fermi: World's First HPC GPU

Oak Ridge Deploys World's Fastest Supercomputer w/ GPUs

Discovered How H1N1 Mutates to Resist Drugs

AlexNet beats expert code by huge margin using GPUs

IMAGENET

World's First Atomic Model of HIV Capsid

Stanford Builds AI Machine using GPUs

World's First 3-D Mapping of Human Genome

Google Outperforms Humans in ImageNet

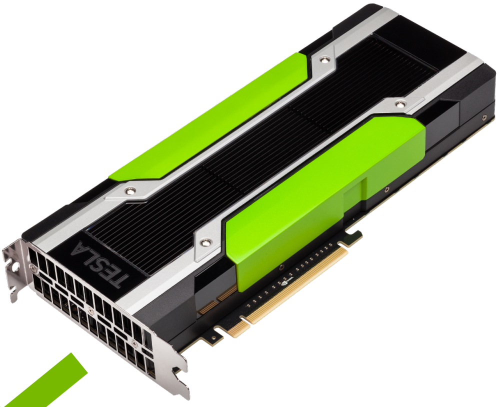
GPU-Trained AI Machine Beats World Champion in Go

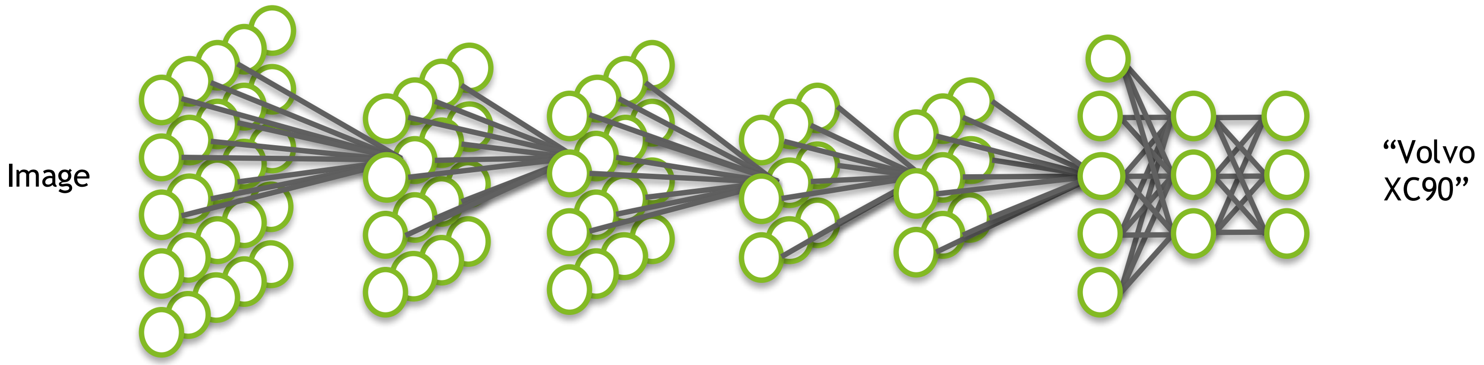
HARDWARE AND DATA DRIVES DEEP LEARNING

facebook | 350 millions images uploaded per day

Walmart | 2.5 Petabytes of customer data hourly

YouTube | 300 hours of video uploaded every minute





MOST PERVASIVE HPC PLATFORM EVER BUILT

ACCESS ANYWHERE

Deskto
p

Server

阿里云计算
Alibaba Cloud Computing

amazon
web services™

Microsoft Azure

IBM
SOFTLAYER

Cloud

BUY ANYWHERE

amazon

Alibaba Group
阿里巴巴集团

Bull
atos technologies

CRAY

DELL

IBM

inspur
浪潮

Hewlett Packard
Enterprise

中科曙光
Sugon

+ 240 Resellers
Worldwide

LEARN EVERYWHERE

78
Countries

1000
Universities Teaching CUDA

400K
CUDA Developers

SCALING DL

ALPHAGO

Training DNNs: 3 weeks, 340 million training steps on 50 GPUs

Play: Asynchronous multi-threaded search

Simulations on CPUs, policy and value DNNs in parallel on GPUs

Single machine: 40 search threads, 48 CPUs, and 8 GPUs

Distributed version: 40 search threads, 1202 CPUs and 176 GPUs

Outcome: Beat World Go champion in best of 5 matches

http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
http://deepmind.com/alpha-go.html

TESLA BUILT FOR THE DATA CENTER

24/7 Uptime

Maximize reliability

Scalable Performance

Boost data center throughput

Data Center Ready

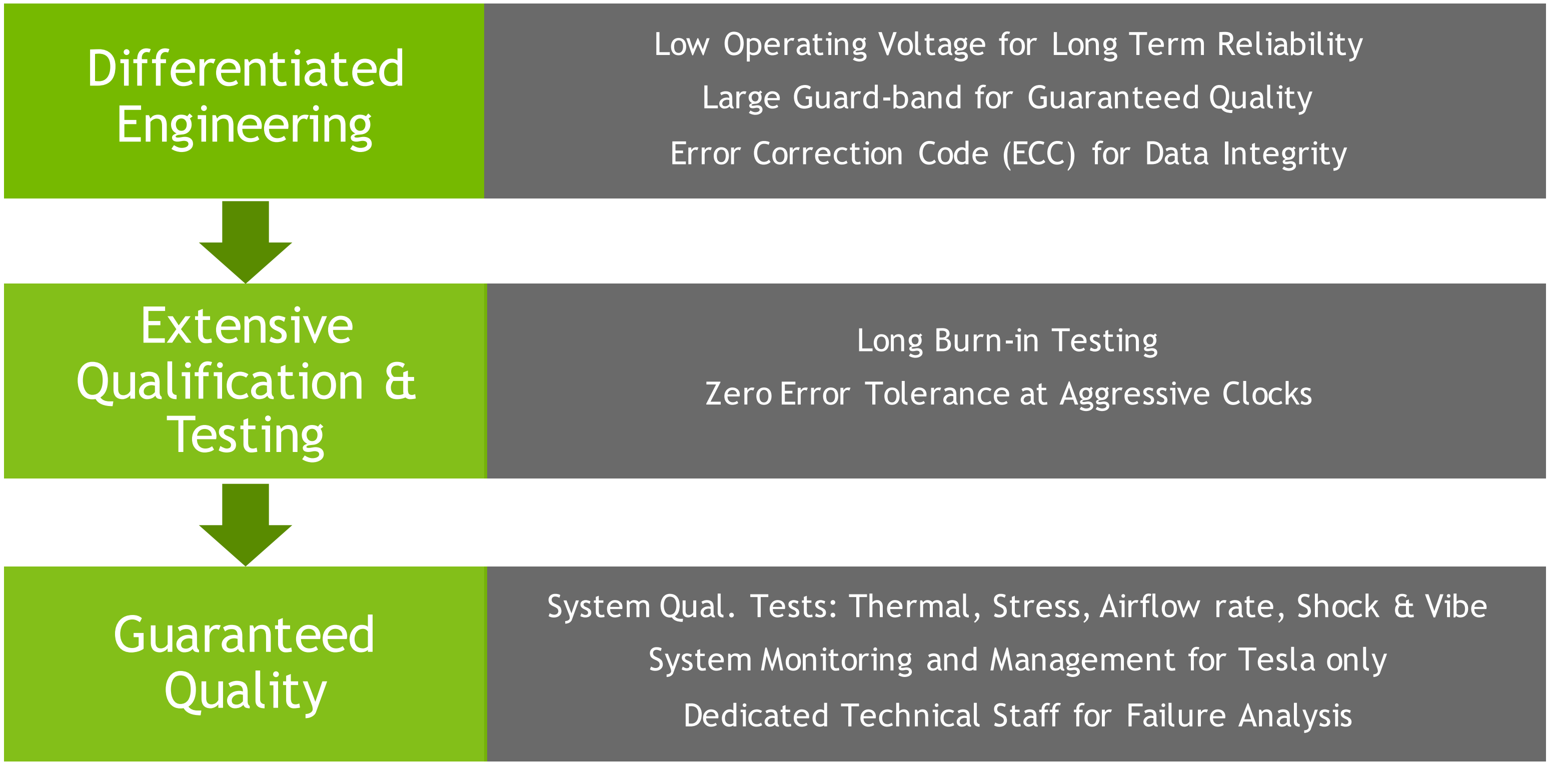
Simplify system operations

END-TO-END DESIGN FOR SYSTEM UPTIME

24/7 Uptime

Scalable Performance

Data Center Ready



Differentiated Engineering

- Low Operating Voltage for Long Term Reliability
- Large Guard-band for Guaranteed Quality
- Error Correction Code (ECC) for Data Integrity

Extensive Qualification & Testing

- Long Burn-in Testing
- Zero Error Tolerance at Aggressive Clocks

Guaranteed Quality

- System Qual. Tests: Thermal, Stress, Airflow rate, Shock & Vibe
- System Monitoring and Management for Tesla only
- Dedicated Technical Staff for Failure Analysis

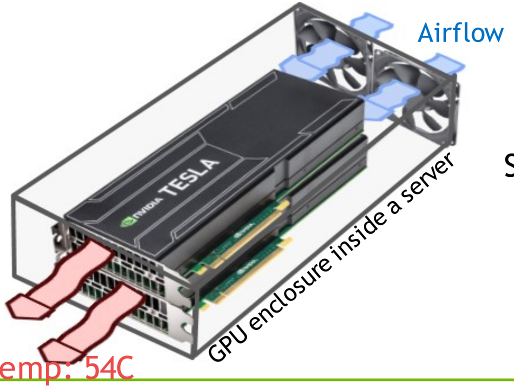
DATA CENTER QUALIFIED BY SERVER OEMS

24/7 Uptime

Scalable Performance

Data Center Ready

Server with Tesla GPU



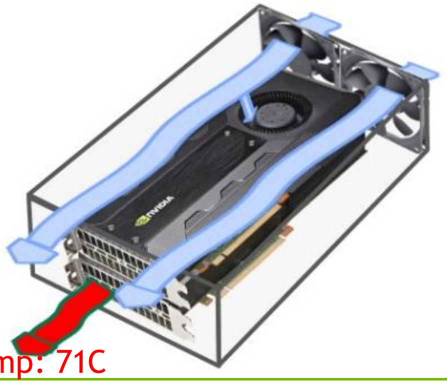
Designed for max airflow through GPU

Supports airflow front-to-back & back-to-front

Lower power consumption

GPU Temp Running Linpack: 54C

Server with Unqualified GPU



Works against server airflow

Higher power consumption

Lower reliability

GPU Temp Running Linpack: 71C

SCALE-OUT PERFORMANCE IN THE DATA CENTER

24/7 Uptime

Scalable Performance

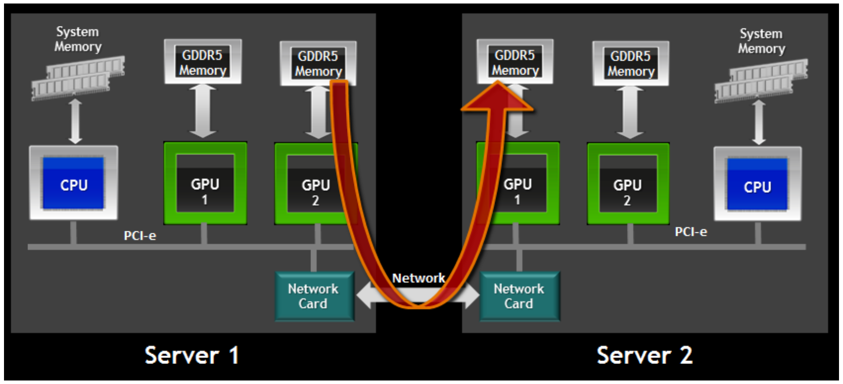
Data Center Ready

GPUDirect RDMA

Direct transfers between GPUs

67% Lower GPU-to-GPU Latency

5x Higher GPU-to-GPU MPI Bandwidth

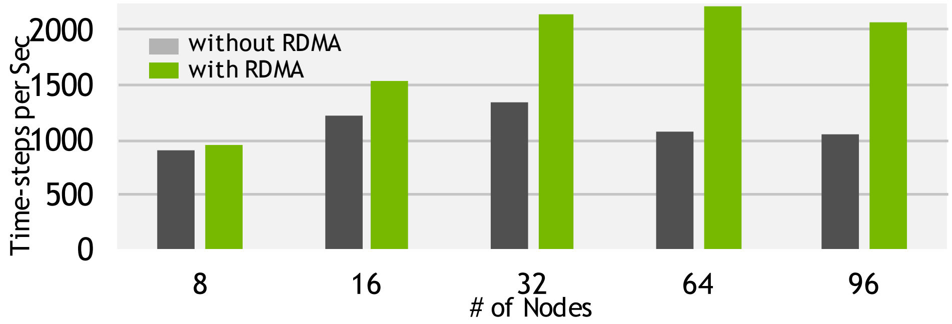


Up to 2x Faster

Application Performance at Scale with GPUDirect RDMA

Hoomd-Blue Application

LJ Liquid Benchmark, 256K Particles

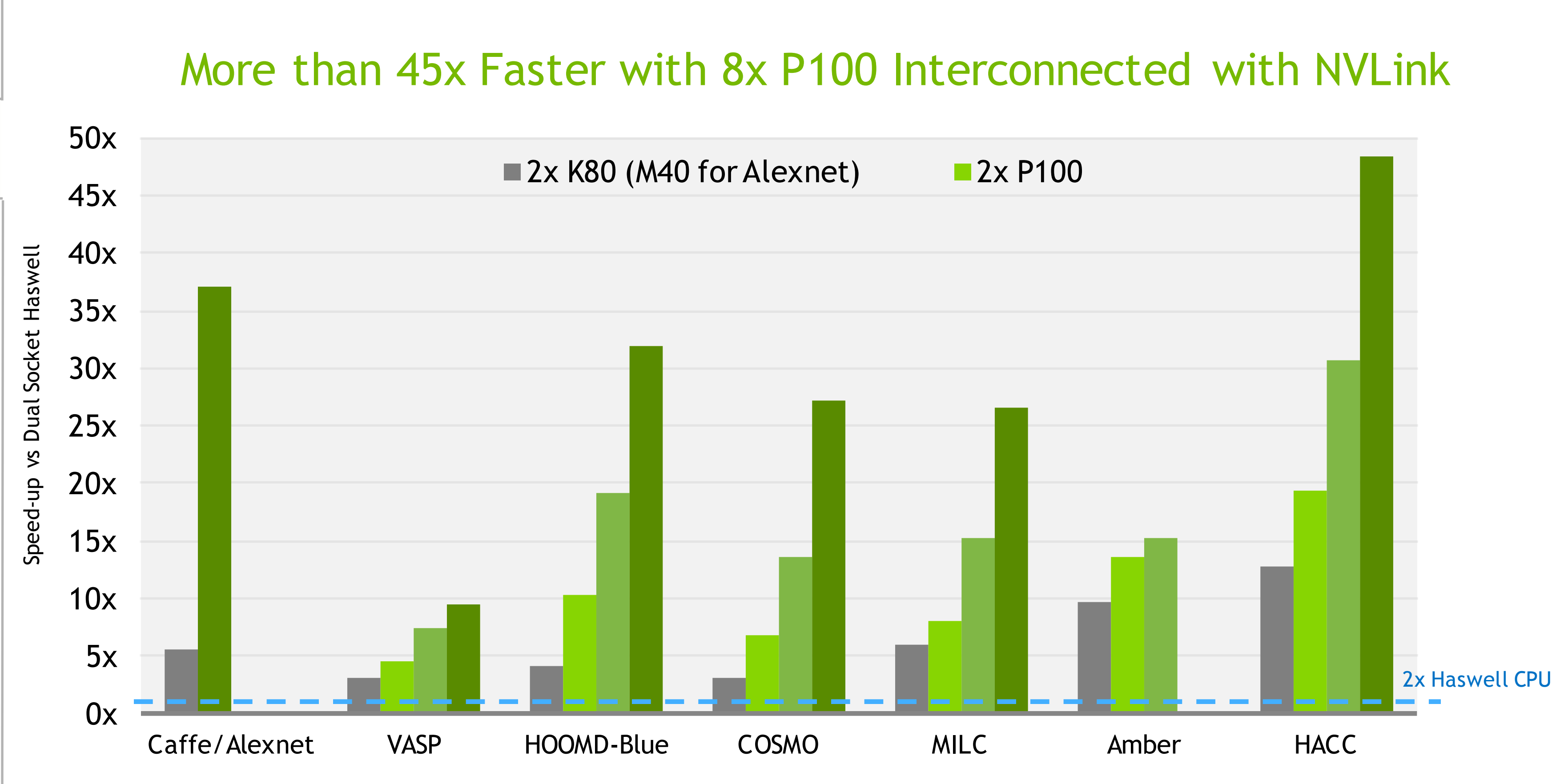


NVLINK DELIVERS SCALABLE PERFORMANCE

24/7 Uptime

Scalable Performance

Data Center Ready



DATA CENTER GPU MANAGEMENT

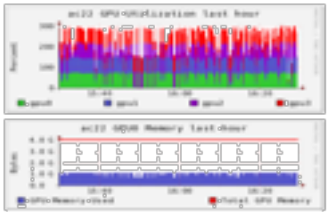
Enterprise-Grade Management Tool for Operating the Data Center

24/7 Uptime

Scalable Performance

Data Center Ready

Device Management



Per GPU Configuration & Monitoring

- Device Identification
- Board Monitoring
- Clock Management

All GPUs Supported

Data Center GPU Manager

Active Health Monitoring

Runtime Health Checks

Prologue Checks

Epilogue Checks

Diagnostics & System Validation

Deep HW Diagnostics

System Validation Tests

Policy & Group Config Management

Pre-configured policies

Job level accounting

Stateful configuration

Power & Clock Mgmt.

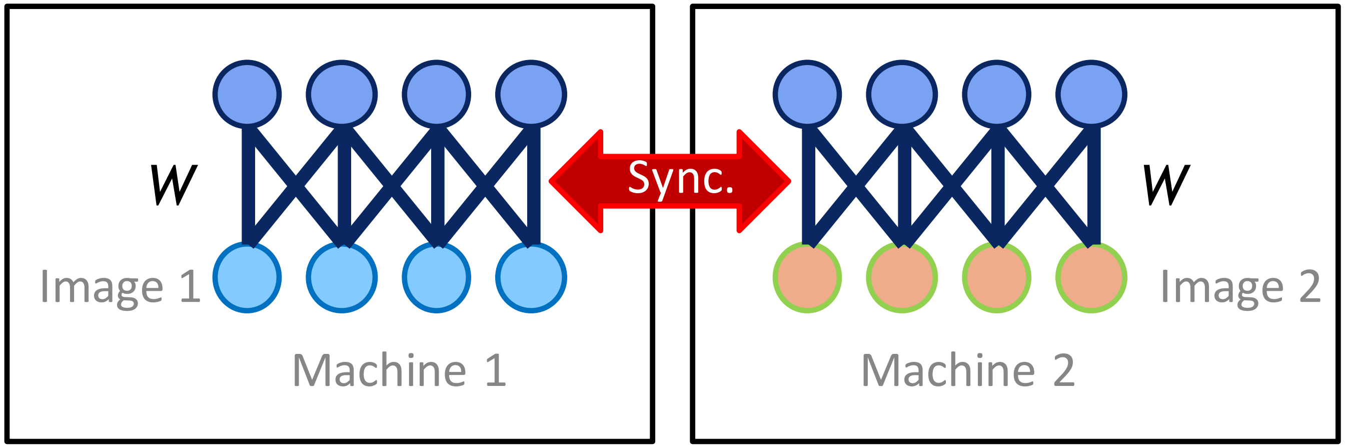
Dynamic Power Capping

Synchronous Clock Boost

*Prerelease Now, GA Q3

SCALING NEURAL NETWORKS

Data Parallelism



Notes:

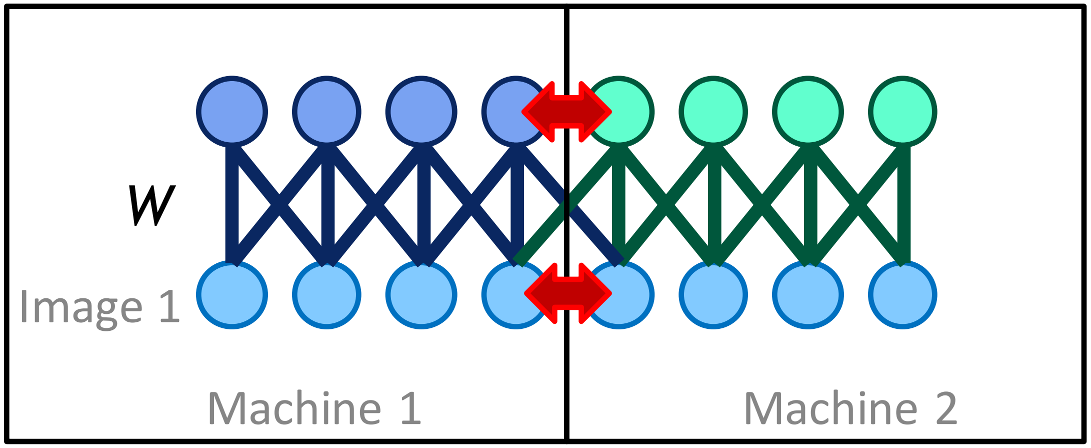
Need to sync model across machines

- Requires P-fold larger batch size
- Works across many nodes – parameter server approach – linear speedup

Adam Coates, Brody Huval, Tao Wang, David J. Wu, and Andrew Ng

SCALING NEURAL NETWORKS

Model Parallelism



Notes:

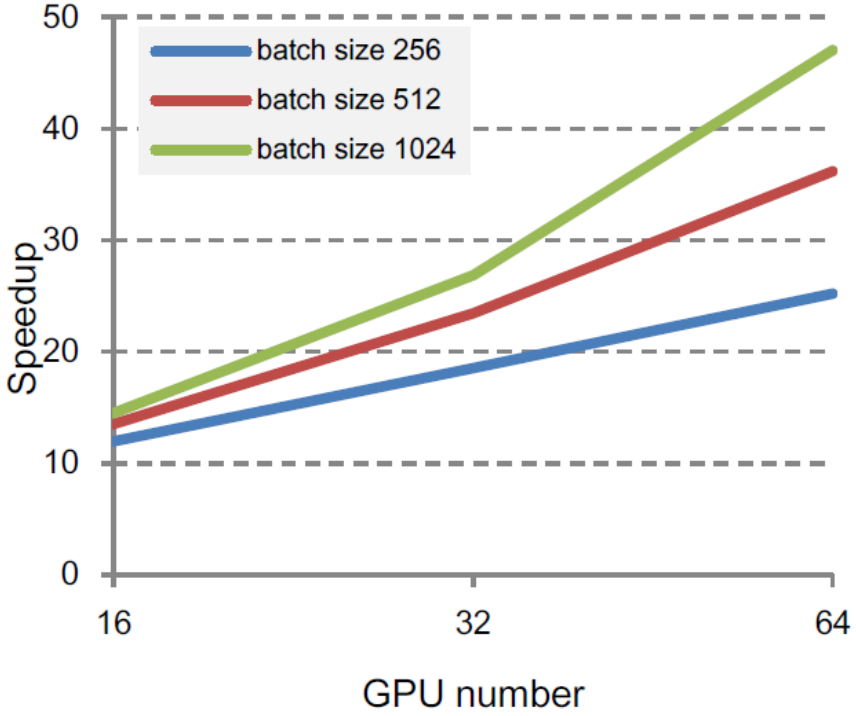
Allows for larger models than fit on one GPU

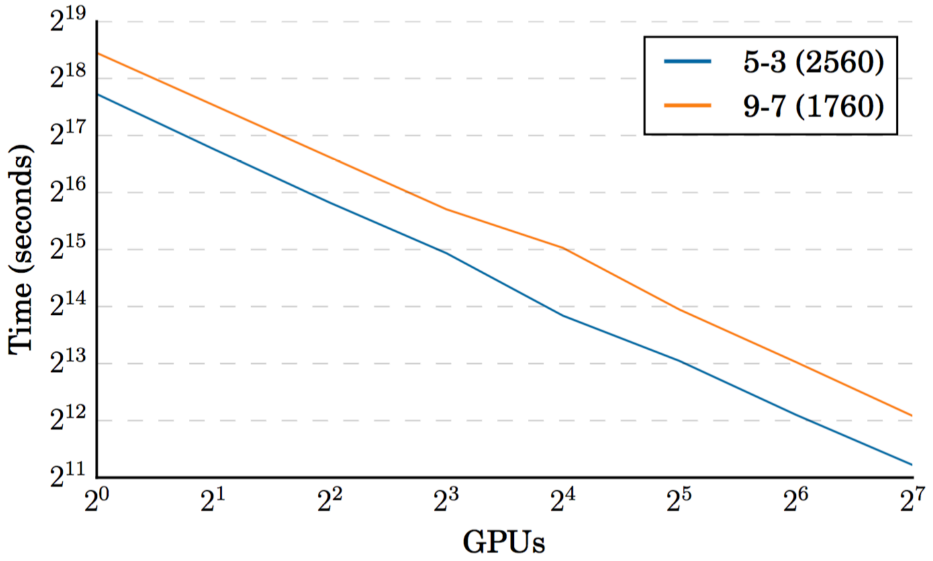
- Most commonly used within a node – GPU P2P
- Effective for the fully connected layers
- Requires much more frequent communication between GPUs

Adam Coates, Brody Huval, Tao Wang, David J. Wu, and Andrew Ng

PARTNER RESULTS – BAIDU

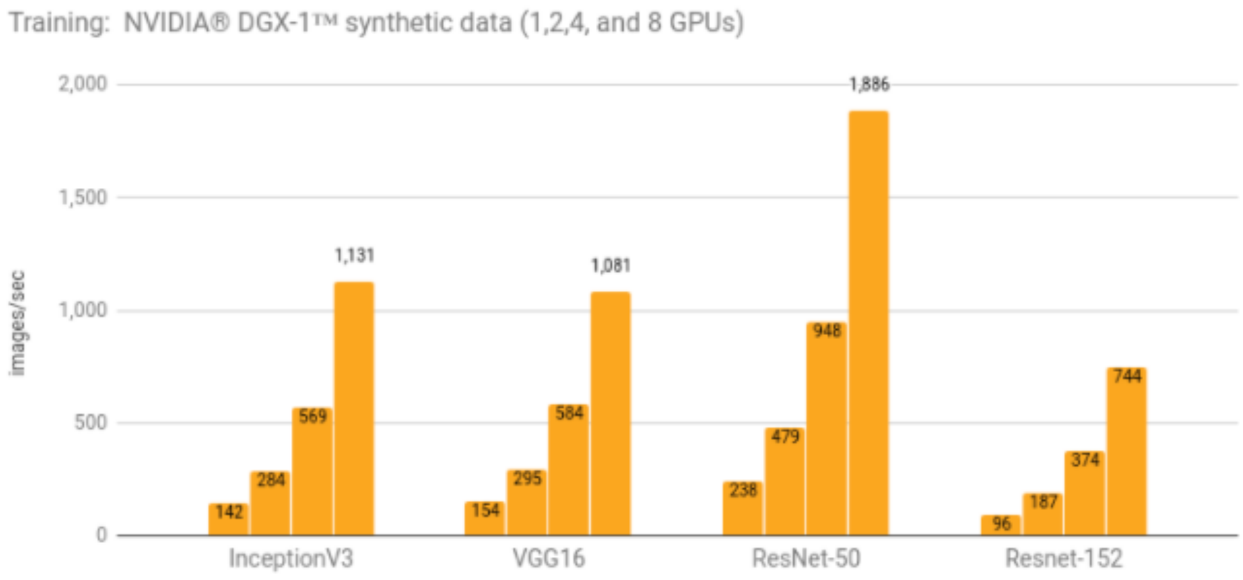
Near linear scaling – synchronous training





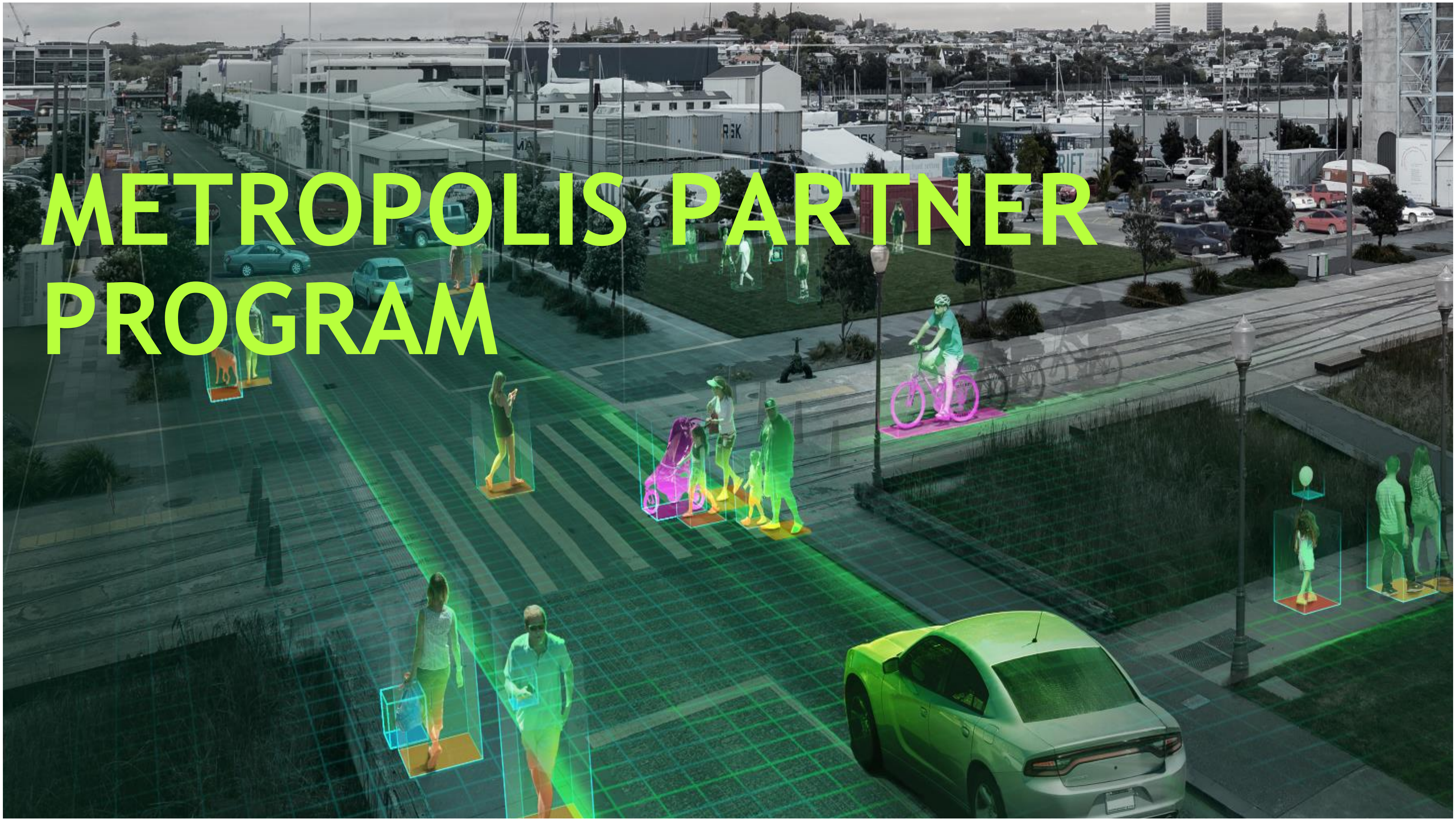
Ren Wu et al, Baidu, “Deep Image: Scaling up Image Recognition.” arXiv 2015
Dario Amodei, et. al. Baidu, “Deep Speech 2” arXiv 2015

PARTNER RESULTS – DGX-1 TENSORFLOW



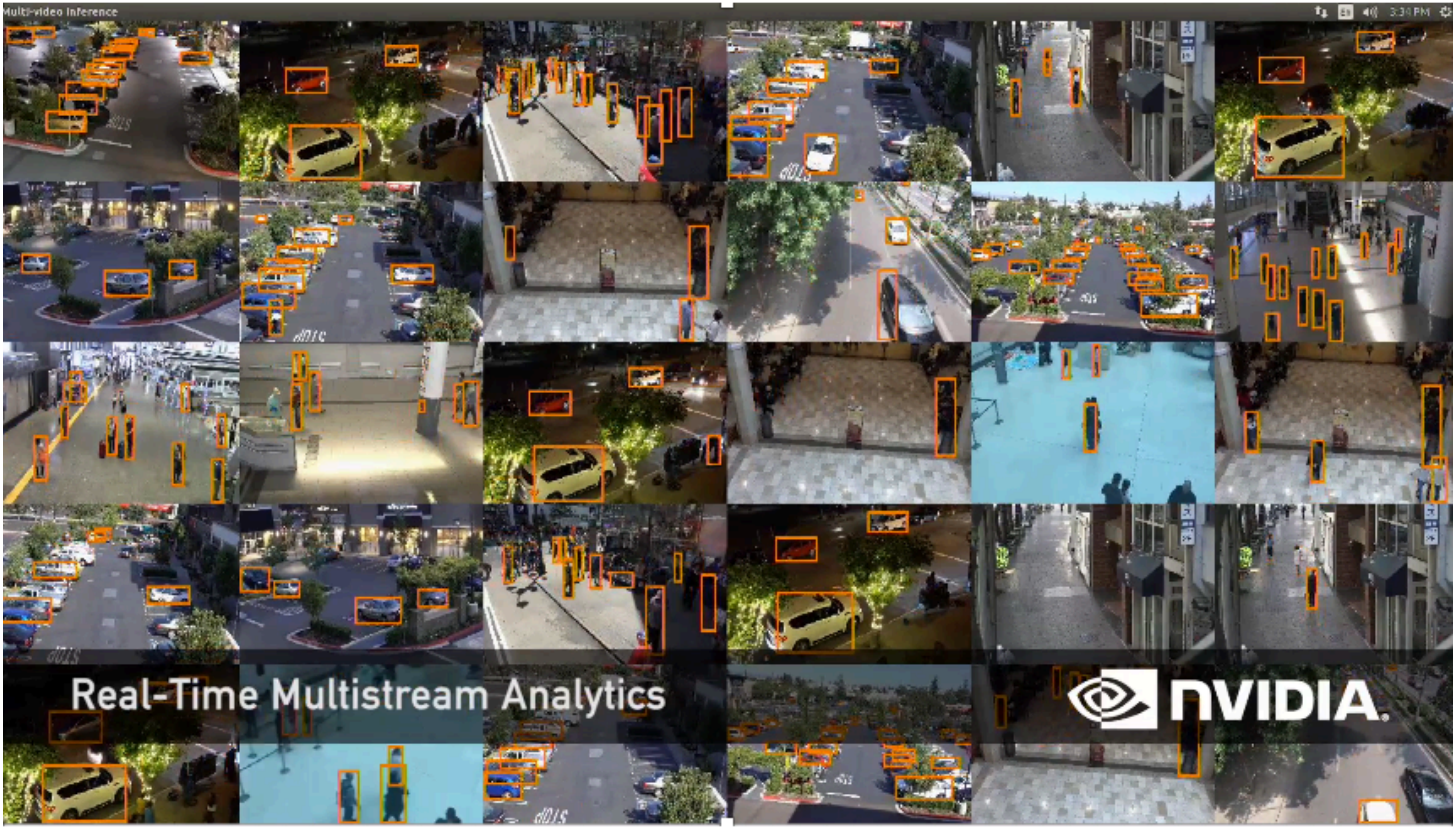
https://www.tensorflow.org/performance/benchmarks#methodology

METROPOLIS PARTNER PROGRAM



NVIDIA METROPOLIS PARTNERS

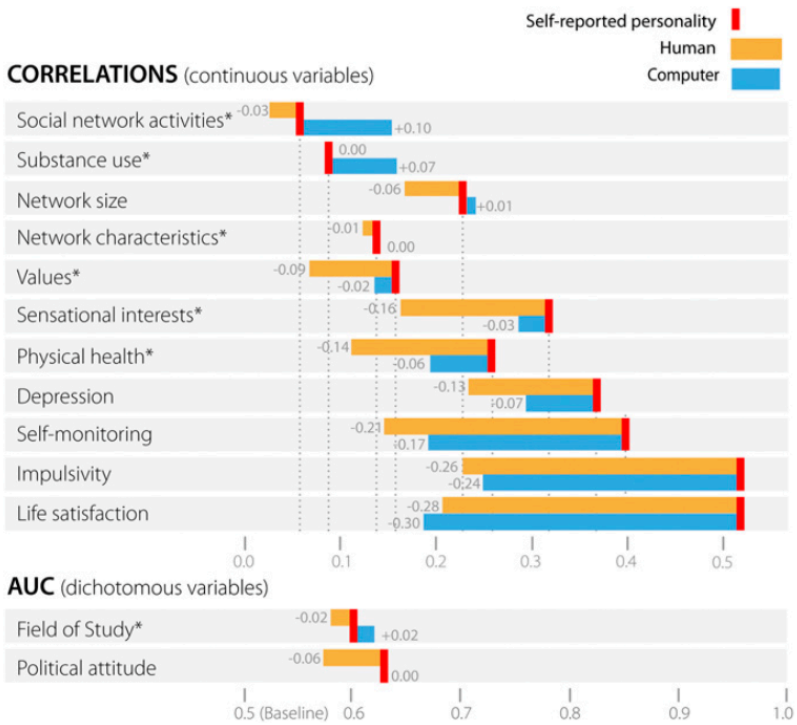
AllGoVision see. sense. secure	anyvision. A BETTER TOMORROW	BriefCam The Video Synopsis Company
herta	iCetana	MotionDSP
OMNI Ai	OpenALPR	SENSETIME
VisionLabs MACHINES CAN SEE	VOCORD PROFESSIONAL VIDEO SURVEILLANCE AND AUDIO RECORDING SYSTEMS	XJERA LABS


Multi-video Inference
3:34 PM
Real-Time Multistream Analytics
NVIDIA

ADVANCED MODELS MAY ERODE PRIVACY

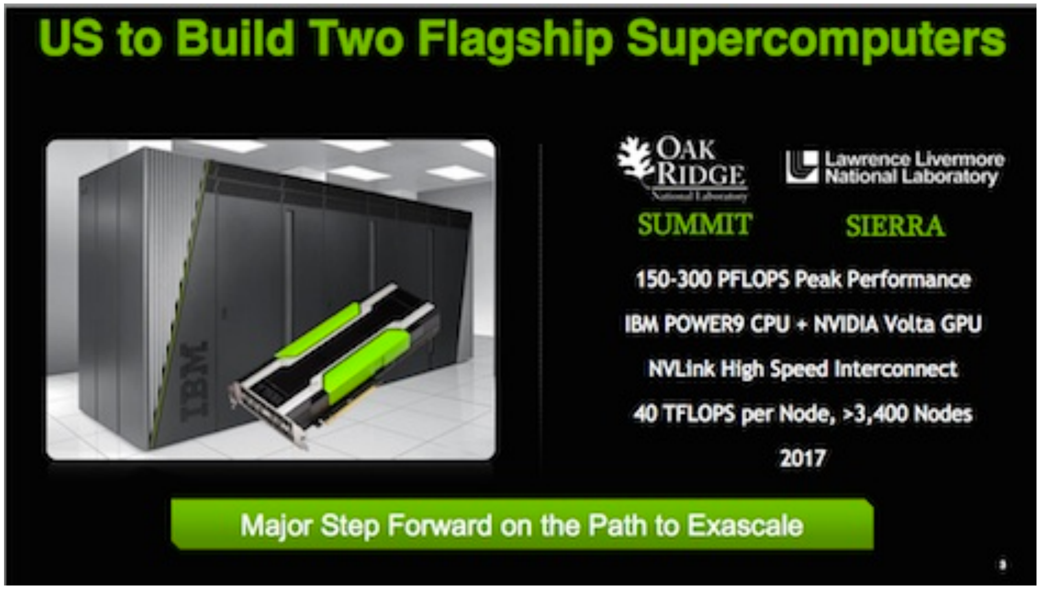
The Target Dilemma

- Using a basket of 25 product features, Target generated classification score
- This resulted in empathetically recommending baby-related promotions
- In the literature, Youyou et. al. worked with Facebook likes



Duhigg 2016, New York Times

THE ROLE OF PUBLIC INFRASTRUCTURE



US to Build Two Flagship Supercomputers

IBM
OAK RIDGE National Laboratory
Lawrence Livermore National Laboratory

SUMMIT SIERRA

150-300 PFLOPS Peak Performance

IBM POWER9 CPU + NVIDIA Volta GPU

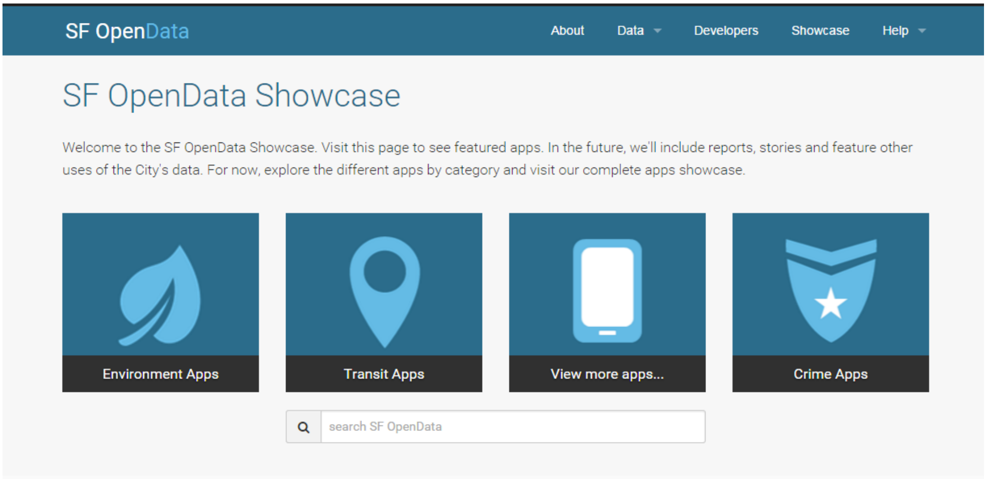
NVLink High Speed Interconnect

40 TFLOPS per Node, >3,400 Nodes

2017

Major Step Forward on the Path to Exascale

InsideHPC



SF OpenData Showcase

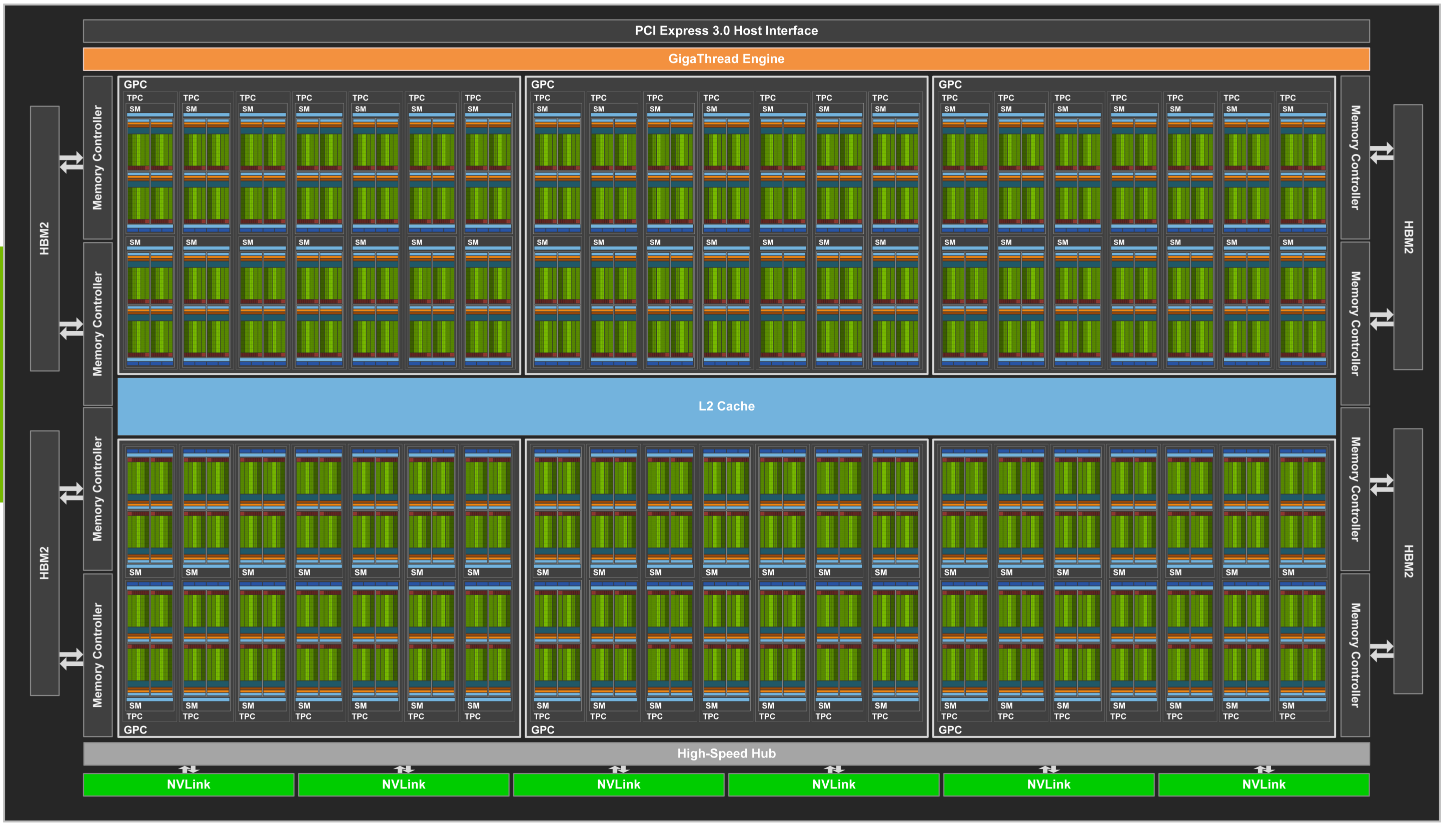
Welcome to the SF OpenData Showcase. Visit this page to see featured apps. In the future, we'll include reports, stories and feature other uses of the City's data. For now, explore the different apps by category and visit our complete apps showcase.

TESLA V100

21B transistors
815 mm^2

80 SM
5120 CUDA Cores
640 Tensor Cores

16 GB HBM2
900 GB/s HBM2
300 GB/s NVLink

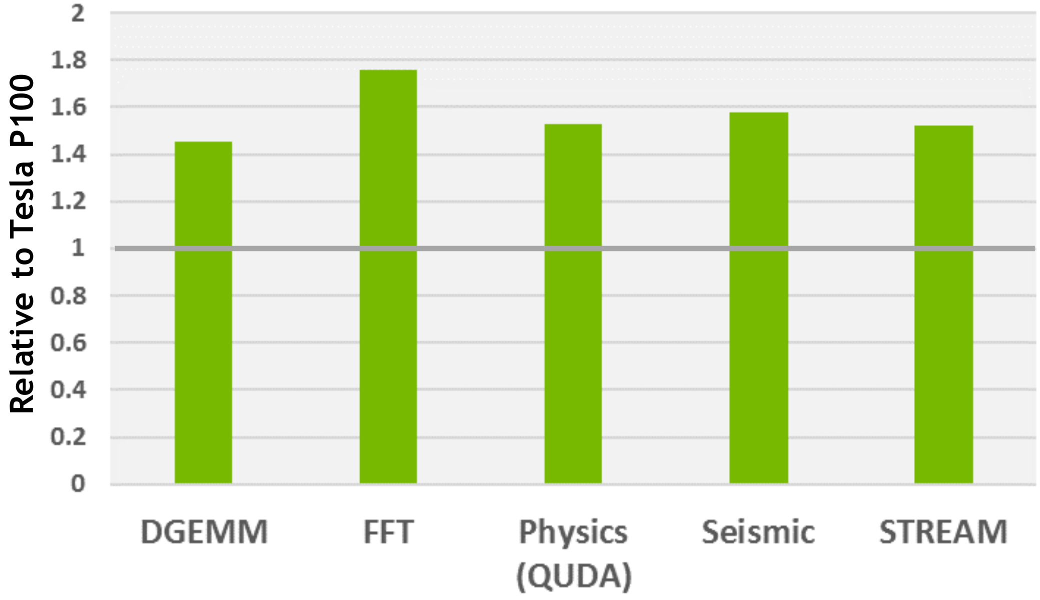


*full GV100 chip contains 84 SMs

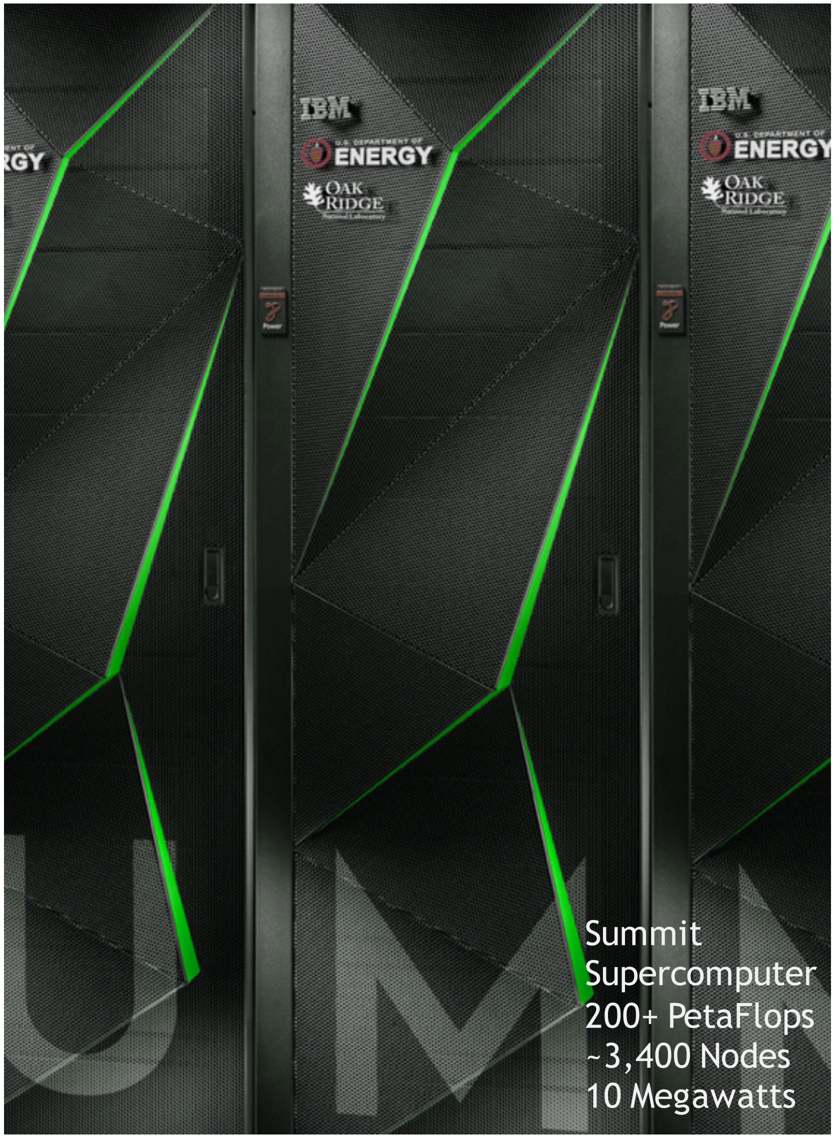
ROAD TO EXASCALE

Volta to Fuel Most Powerful US Supercomputers

Volta HPC Application Performance



System Config Info: 2X Xeon E5-2690 v4, 2.6GHz, w/ 1X Tesla P100 or V100. V100 measured on pre-production hardware.



Summit Supercomputer
200+ PetaFlops
~3,400 Nodes
10 Megawatts

Thank you!
leot@nvidia.com