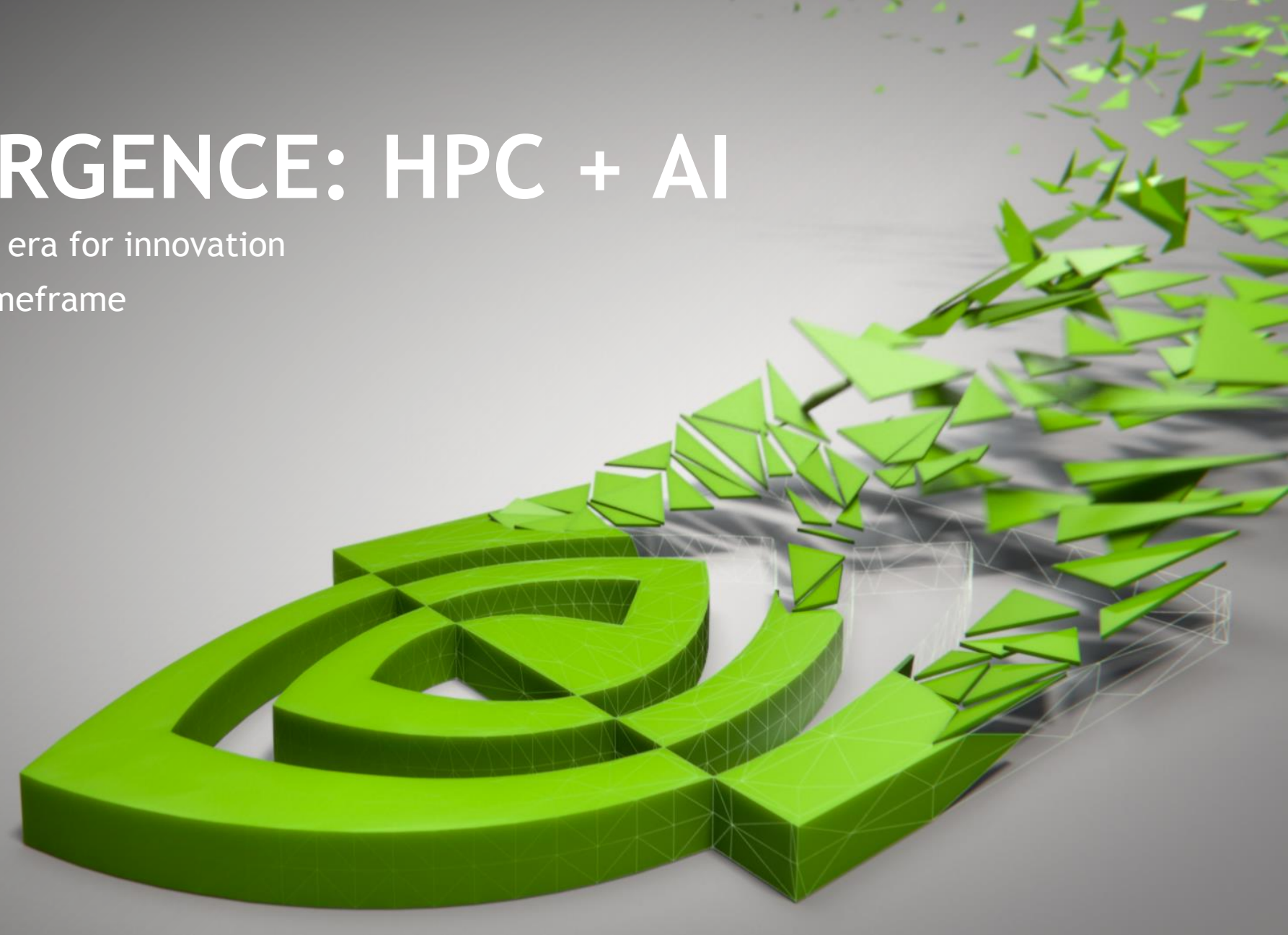
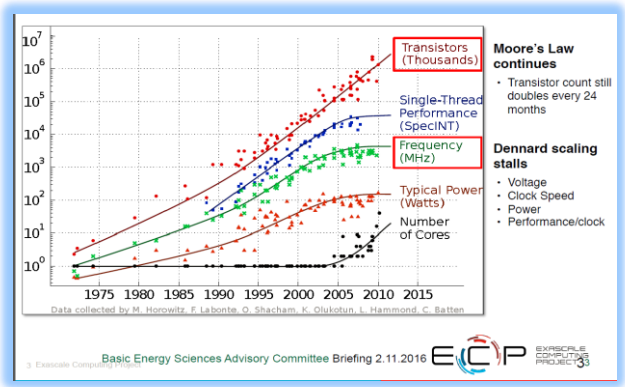
CONVERGENCE: HPC + AI

Introducing a new era for innovation
in the Exascale timeframe

NVIDIA



FACTORS DRIVING INNOVATION IN HPC



End of Dennard Scaling places a cap on single threaded performance

Increasing application performance will require fine grain parallel code with significant computational intensity

AI and Data Science emerging as important new components of scientific discovery

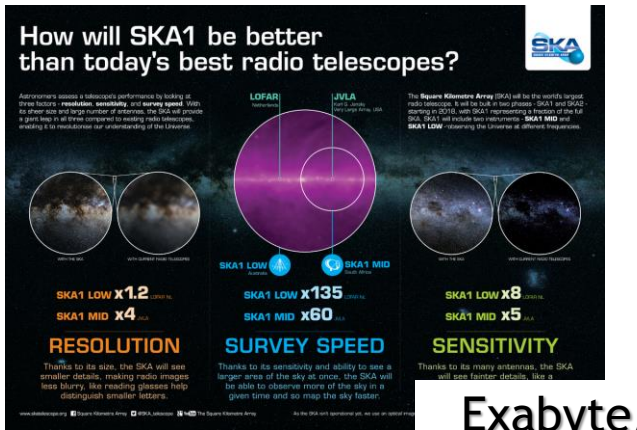
Dramatic improvements in accuracy, completeness and response time yield increased insight from huge volumes of data

Cloud based usage models, in-situ execution and visualization emerging as new workflows critical to the science process and productivity

Tight coupling of interactive simulation, visualization, data analysis/AI

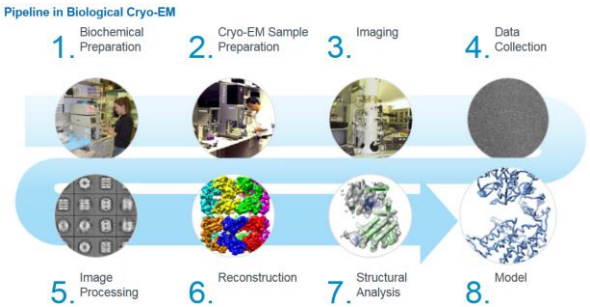
THE EX FACTOR IN THE EXASCALE ERA

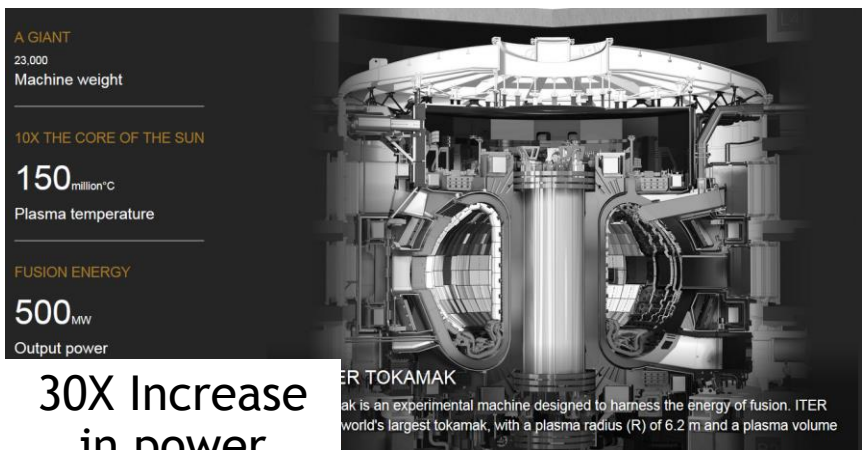
Multiple EXperiments Coming or Upgrading In the Next 10 Years



Exabyte/Day

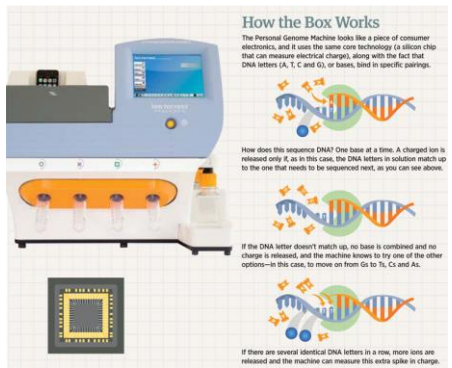
15 TB/Day





30X Increase in power

10X Increase in Data Volume



Personal Genomics

THE POTENTIAL OF EXASCALE HPC + AI

HPC	AI
+40 years of Algorithms based on first principles theory Proven statistical models for accurate results in multiple science domains	New methods to improve predictive accuracy, insight into new phenomena and response time with previously unmanageable data sets

Commercially viable fusion energy

Understanding the Origins of the Universe

Clinically Viable Precision Medicine

Improve/validate the Standard Model of Physics

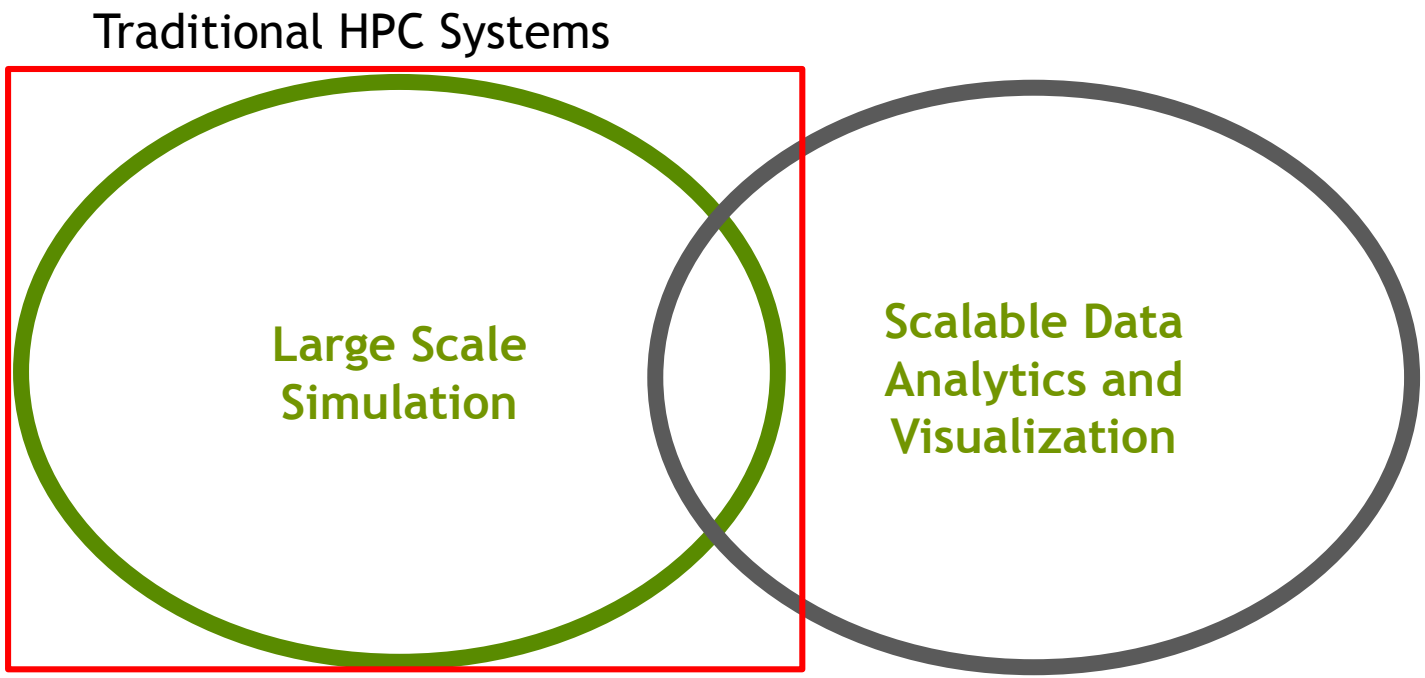
Climate/Weather forecasts with ultra high fidelity

*

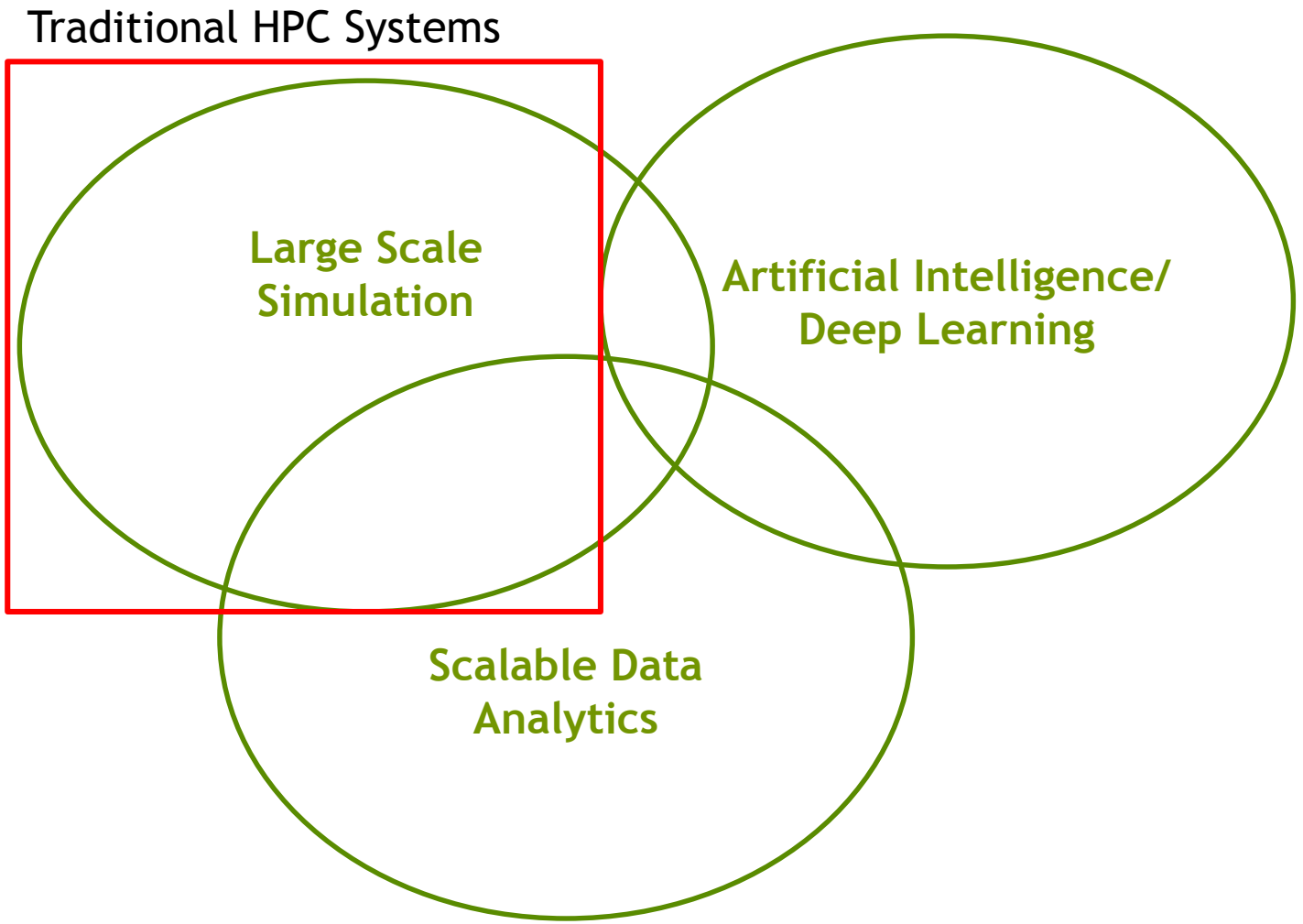
*

*

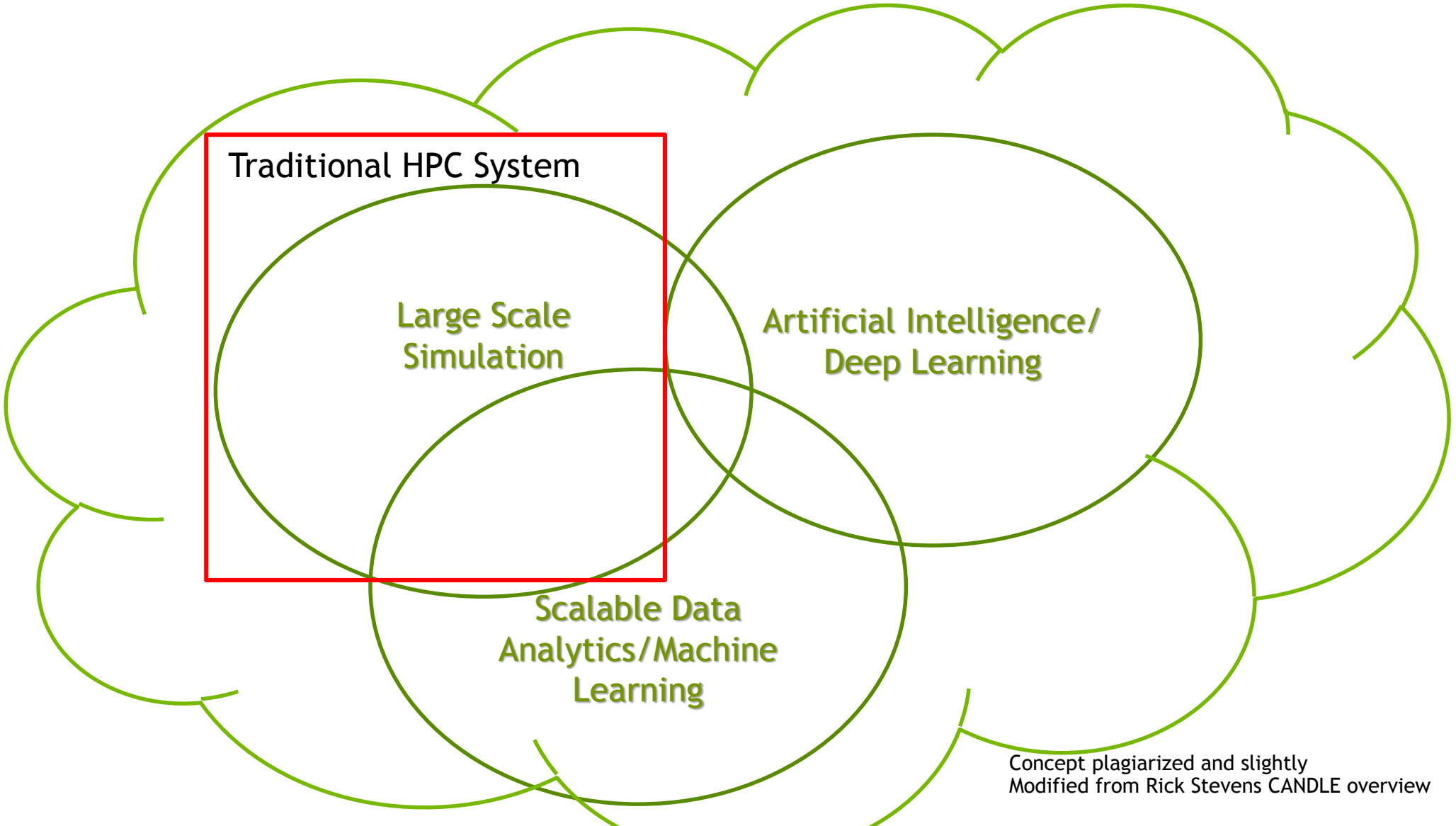
TRADITIONAL HPC METHOD



EVOLUTION OF HPC METHOD



CONVERGED EXASCALE ERA SYSTEM



Concept plagiarized and slightly
Modified from Rick Stevens CANDLE overview

TAXONOMY

Organizing HPC + AI Convergence

Transformation

HPC + AI couple simulation with live data in real time detection/control system

Experimental/simulated data is used to train a NN that is used to for detection/control of an experiment or clinical delivery system in real time.
The NN is improved continuously as new simulated / live data is acquired

Augmentation

HPC + AI combined to improve simulation time to science > orders of magnitude

Experimental/simulated data is used to train a NN that is used to replace all or significant runtime portions of a conventional simulation.
The NN is improved continuously as new simulated / live data is acquired

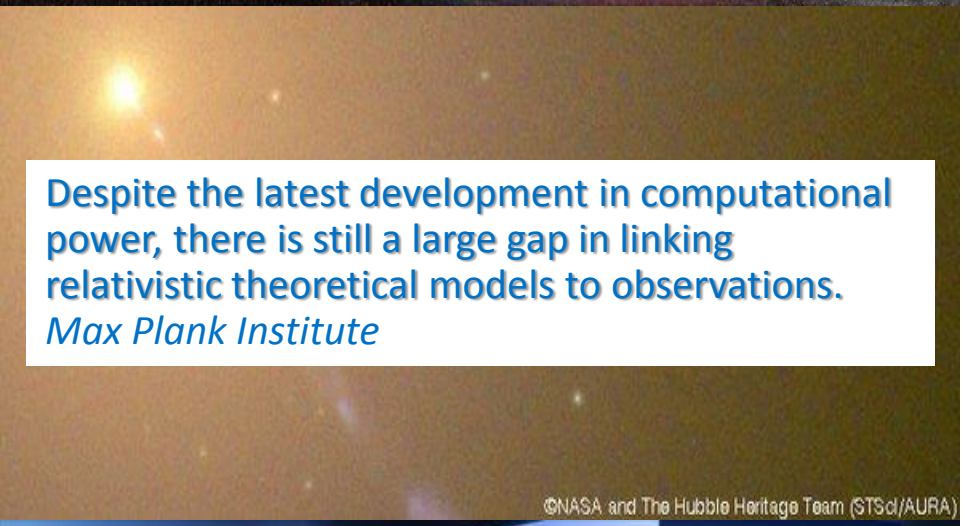
Modulation

HPC + AI combined to reduce the number of runs needed for a parameter sweep

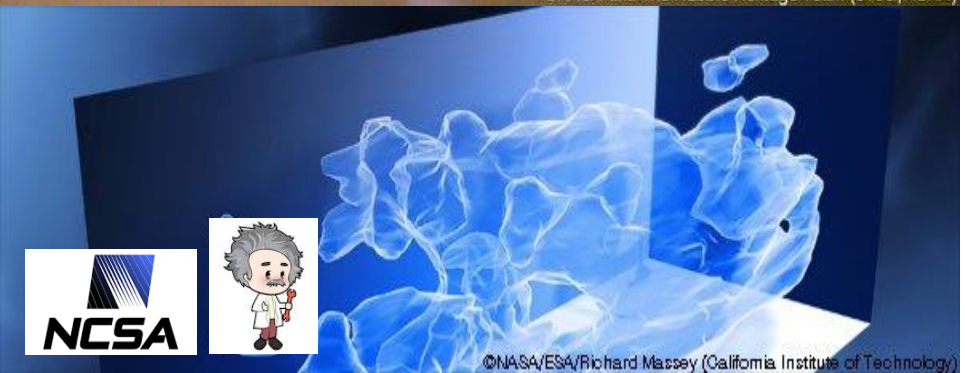
Experimental/simulated data used to train a NN which steers simulation/experiment btwn runs

The steering NN can be trained continuously as new simulated / live data is acquired

Potential for Breakthroughs in Scientific Insight



Despite the latest development in computational power, there is still a large gap in linking relativistic theoretical models to observations.
Max Plank Institute



MULTI-MESSENGER ASTROPHYSICS

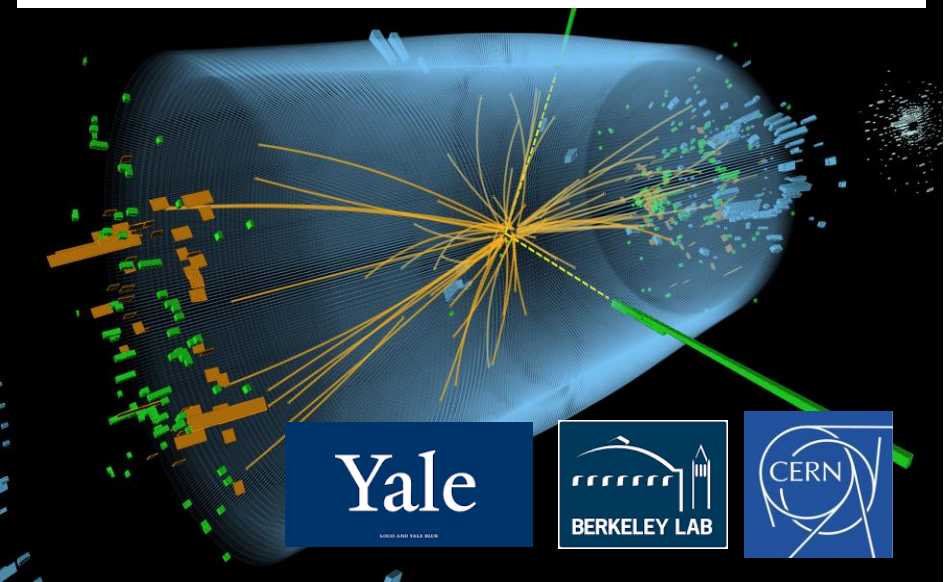
Background
The aLIGO (Advanced Laser Interferometer Gravitational Wave Observatory) experiment successfully discovered signals proving Einstein's theory of General Relativity and the existence of cosmic Gravitational Waves. While this discovery was by itself extraordinary it is seen to be highly desirable to combine multiple observational data sources to obtain a richer understanding of the phenomena.

Challenge
The initial a LIGO discoveries were successfully completed using classic data analytics. The processing pipeline used hundreds of CPU's where the bulk of the detection processing was done offline. Here the latency is far outside the range needed to activate resources, such as the Large Synaptic Space survey Telescope (LSST) which observe phenomena in the electromagnetic spectrum in time to "see" what aLIGO can "hear".

Solution
A DNN was developed and trained using a data set derived from the CACTUS simulation using the Enistein Toolkit. The DNN was shown to produce better accuracy with latencies 1000x better than the original CPU based waveform detection.

Impact
Faster and more accurate detection of gravitational waves with the potential to steer other observational data sources.

For the moment, we can state that the multithreaded version of Geant4 is not yet optimized to compete with a distributed submission of simulations on a farm of CPU clusters…..
Performance Evaluation of Multithreaded Geant4
P. Schweitzer 2015



UNDERSTANDING THE STANDARD MODEL OF PHYSICS

Background

The experiments at the CERN Large Hadron Collider were responsible for the validation/discovery of Higgs Boson and other particles that are critical to the understanding of the Standard Model of Physics. The current data processing is achieved using on-site filtering and off site global grid. A critical part of the scientific workflow is simulating the events that are expected to occur to both prepare the experiment and validate the output

Challenge

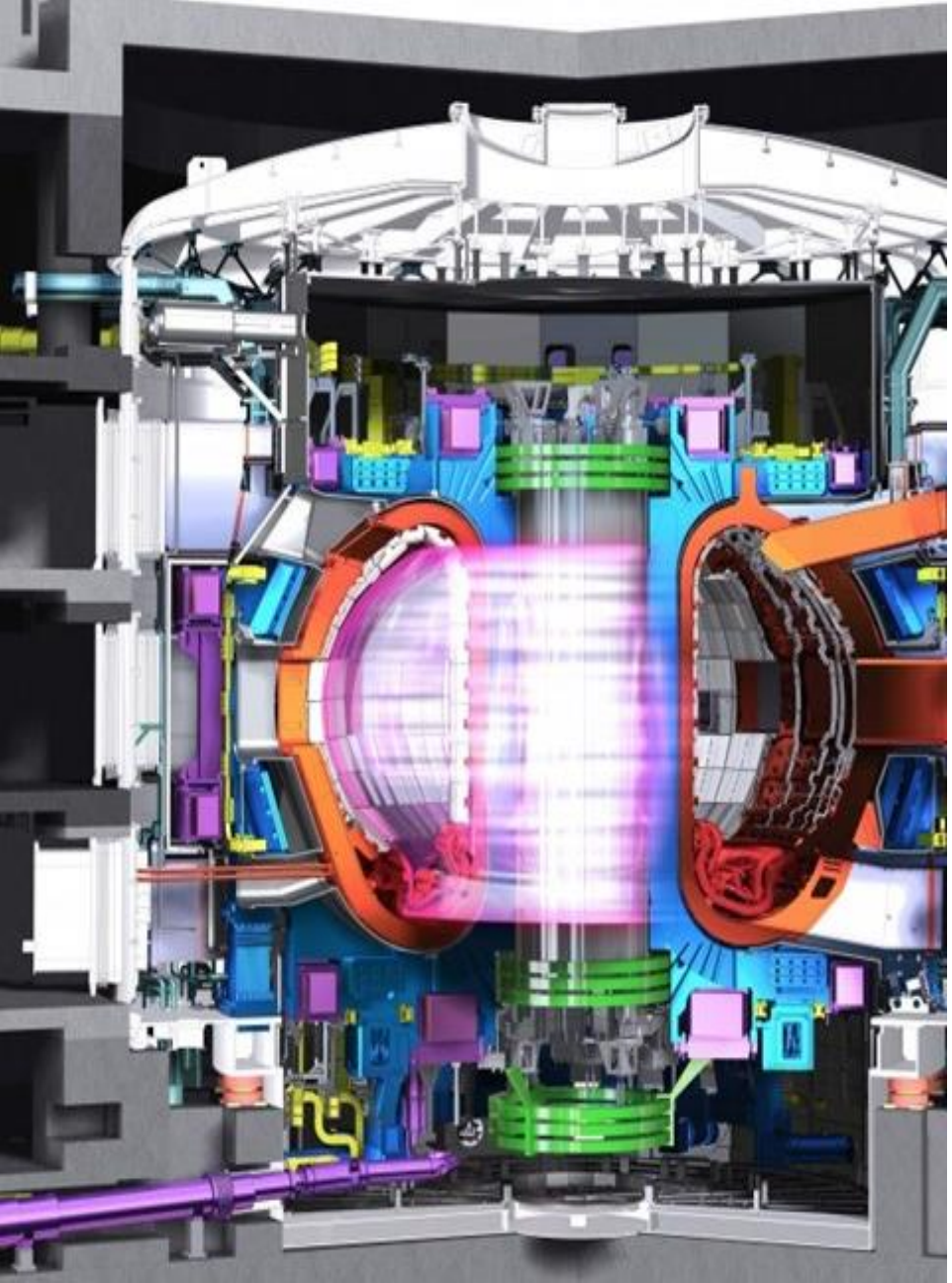
The simulation for the LHC particle collider is known as GEANT and it is numerically intensive, requires significant compute cycles as part of the scientific workflow and is extremely hard to optimize for modern CPU or GPU type architectures. The High Luminosity LHC is expected to generate 10X the volume of data, and the compute load on GEANT is in turn expected to challenge future computing systems within a flat budget profile.

Solution

A GAN (caloGAN) was constructed by a team at Lawrence Berkeley Lab, CERN and Yale that was customized for Calorimiter Experiments similar to ATLAS in the LHC. A training data set was developed using 100,000 GEANT events as input. 50 epochs were then used on 18xK80's to train the GAN with KERAS and TensorFlow.

Impact

The CaloGAN is more accurate and showed up to 5 Orders of Magnitude performance speed up relative to the original simulation on a single K80 GPU



Predicting Disruptions in Fusion Reactor using DL

Background

Grand challenge of fusion energy offers mankind changing opportunity to provide clean, safe energy for millions of years. ITER is a $25B international investment in a fusion reactor.

Challenge

Fusion is highly sensitive, any disruption to conditions can cause reaction to stop suddenly. Challenge is to predict when a disruption will occur to prevent damage to ITER and to steer the reaction to continue to produce power. Traditional simulation and ML approaches don't deliver accurate enough results.

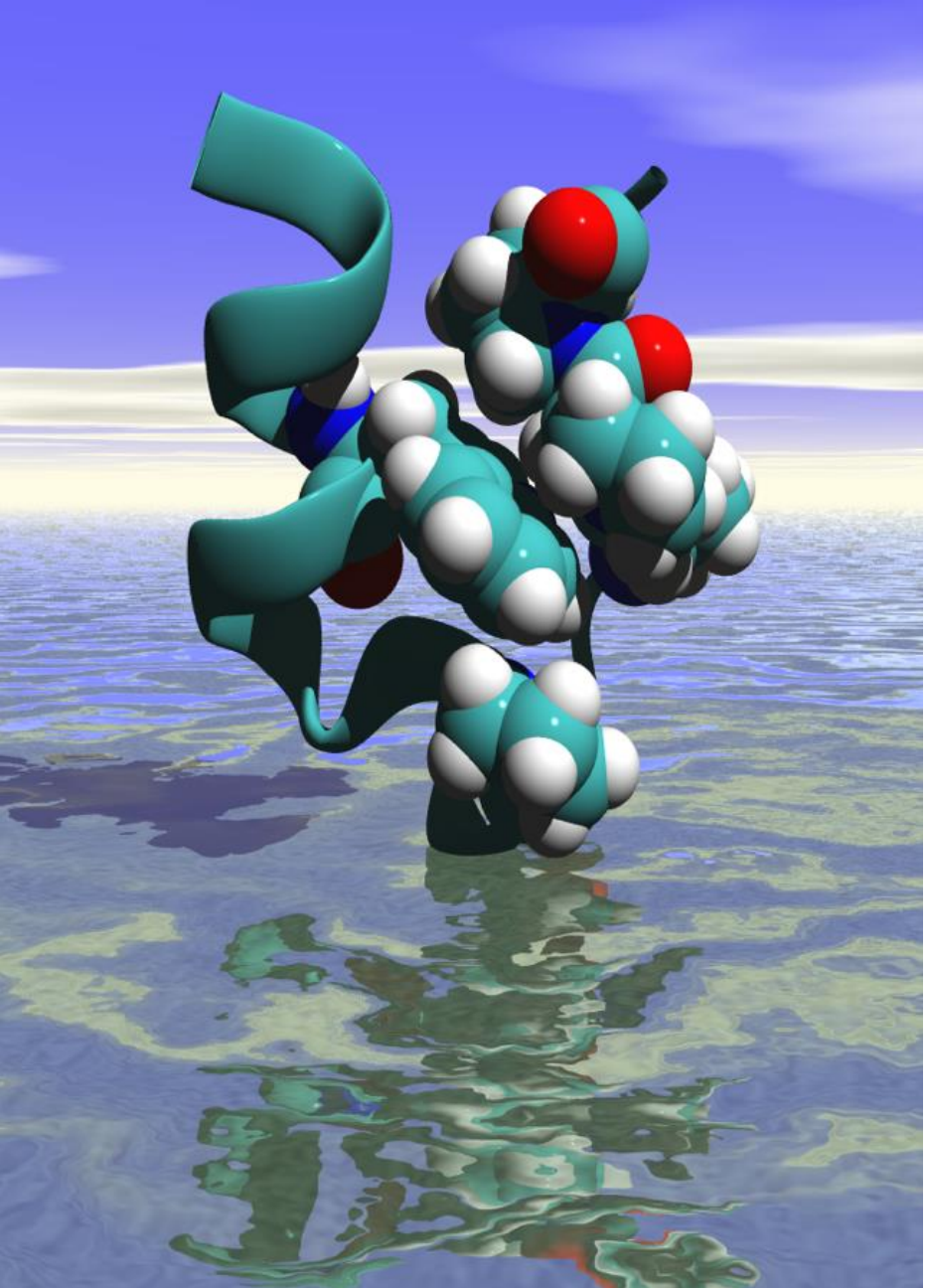
Solution

DL network called FRNN using Theano exceeds today's best accuracy results. It scales to 200 Tesla K20s, and with more GPUs, can deliver higher accuracy. Goal is to reach 95% accuracy.

Impact

Vision is to operate ITER with FRNN, operating and steering experiments in real-time to minimize damage and down-time.

PRINCETON UNIVERSITY



AI Quantum Breakthrough

Background

Developing a new drug costs $2.5B and takes 10-15 years. Quantum chemistry (QC) simulations are important to accurately screen millions of potential drugs to a few most promising drug candidates.

Challenge

QC simulation is computationally expensive so researchers use approximations, compromising on accuracy. To screen 10M drug candidates, it takes 5 years to compute on CPUs.

Solution

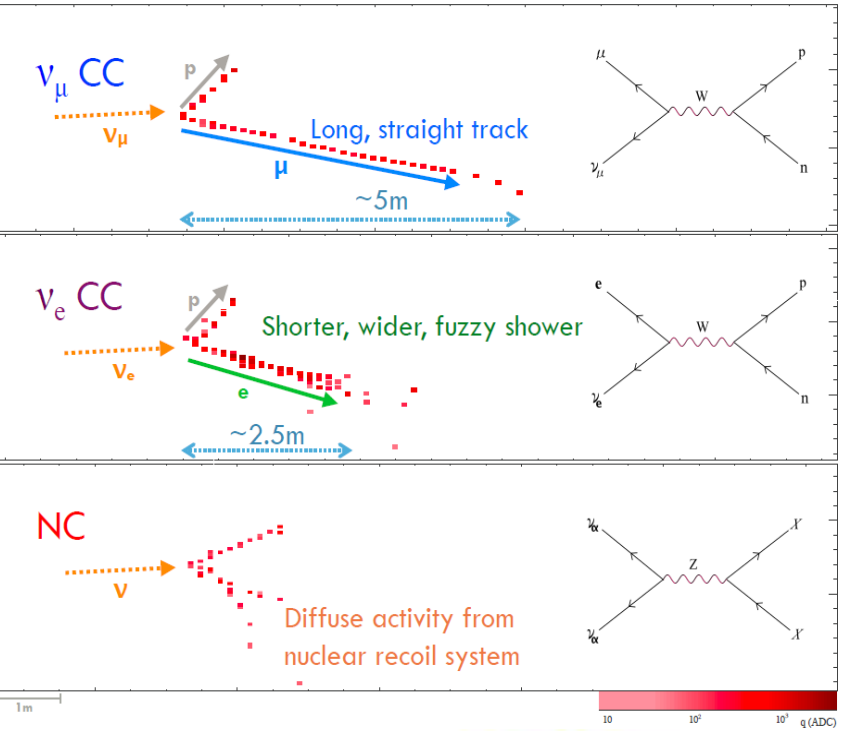
Researchers at the University of Florida and the University of North Carolina leveraged GPU deep learning to develop ANAKIN-ME, to reproduce molecular energy surfaces with super speed (microseconds versus several minutes), extremely high (DFT) accuracy, and at 1-10/millionths of the cost of current computational methods.

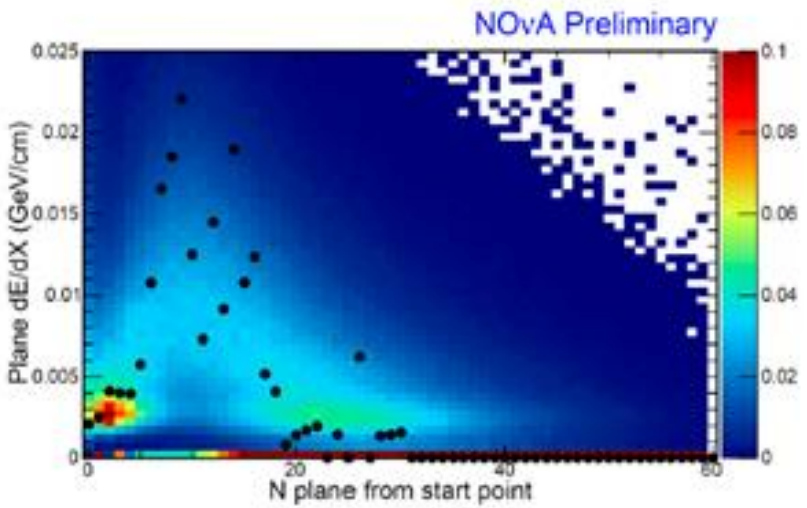
Impact

Faster, more accurate screening at far lower cost

UF UNIVERSITY of FLORIDA

FINDING THE "GHOST PARTICLE" WITH AI





Background

The NoVA experiment managed by Fermi lab comprises 200 scientists at 40 institutions in 7 countries. The goal is to track neutrino's, which are often referred to as the "Ghost Particle", and detect oscillation which is used to better understand how this super abundant, and elusive particle interacts with matter.

Challenge

The experiment is built underground and is comprised of a main injector beam and two large detector apparatus located 50 miles apart. The near detector is 215 Tons and the Far detector is 15,000 Tons. The experiment can be thought of as a 30 Mn pound detector that takes 2 Mn pictures per second.

The detectability of the current experiment is proportional to the size of the detectors, so increasing the "visibility" is complex and costly.

Solution

A DNN was developed and trained using a data set derived from multiple HPC simulations including GENIE and GEANT using 2 K40 GPU's. the CVN was basd on convolutional neural networks used for image processing

Impact

The result was an overall improvement of 33%, where the optimized CVN signal-detection-optimized efficiency of 49% is a significant gain over the efficiency of 35% quoted in prior art. This would net to a 10Mn pound increase the physical detector

Forecasting Fog at Zurich Airport

WORK IN PROGRESS

Background

Unexpected fog can cause an airport to cancel or delay flights, sometimes having global effects in flight planning.

Challenge

While the weather forecasting model at MeteoSwiss work at a 2km x 2km resolution, runways at Zurich airport is less than 2km. So human forecasters sift through huge simulated data with 40 parameters, like wind, pressure, temperature, to predict visibility at the airport.

Solution

MeteoSwiss is investigating the use of deep learning to forecast type of fog and visibility at sub-km scale at Zurich airport.

Impact

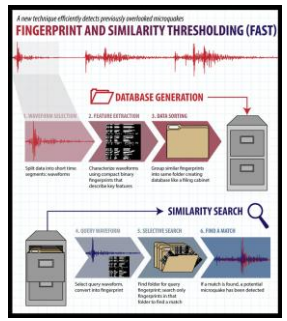
MeteoSwiss

Earthquake Prediction

WORK IN PROGRESS

Multiple Examples of AI for earthquake prediction are underway

Shaazam for Earthquakes



SCIENTIFIC AMERICAN

COMPUTING

Can Artificial Intelligence Predict Earthquakes?

The ability to forecast temblors would be a tectonic shift in seismology. But is it a pipe dream? A seismologist is conducting machine-learning experiments to find out

MIT

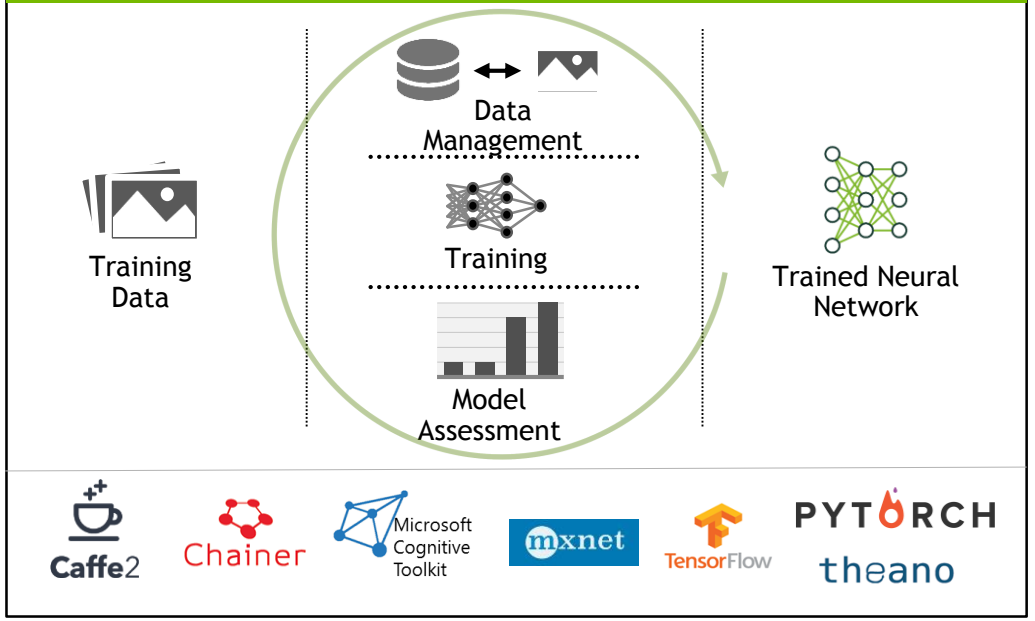
PRINCETON UNIVERSITY

SCEC an NSF+USGS center

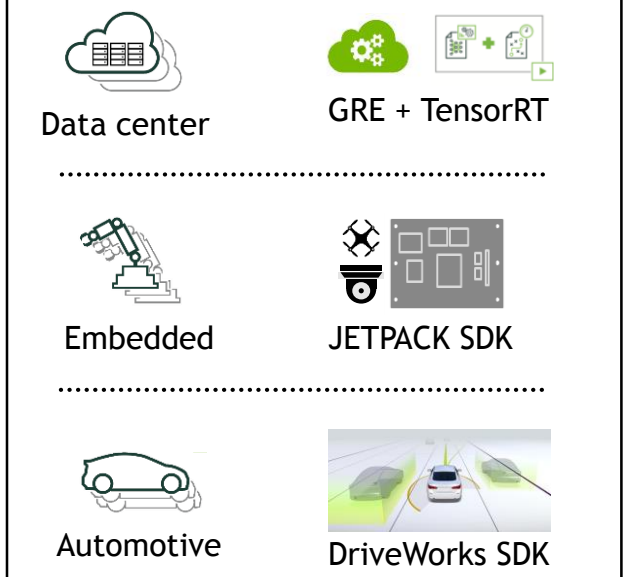
Stanford University

NVIDIA DEEP LEARNING SOFTWARE PLATFORM

TRAINING



INFERENCE

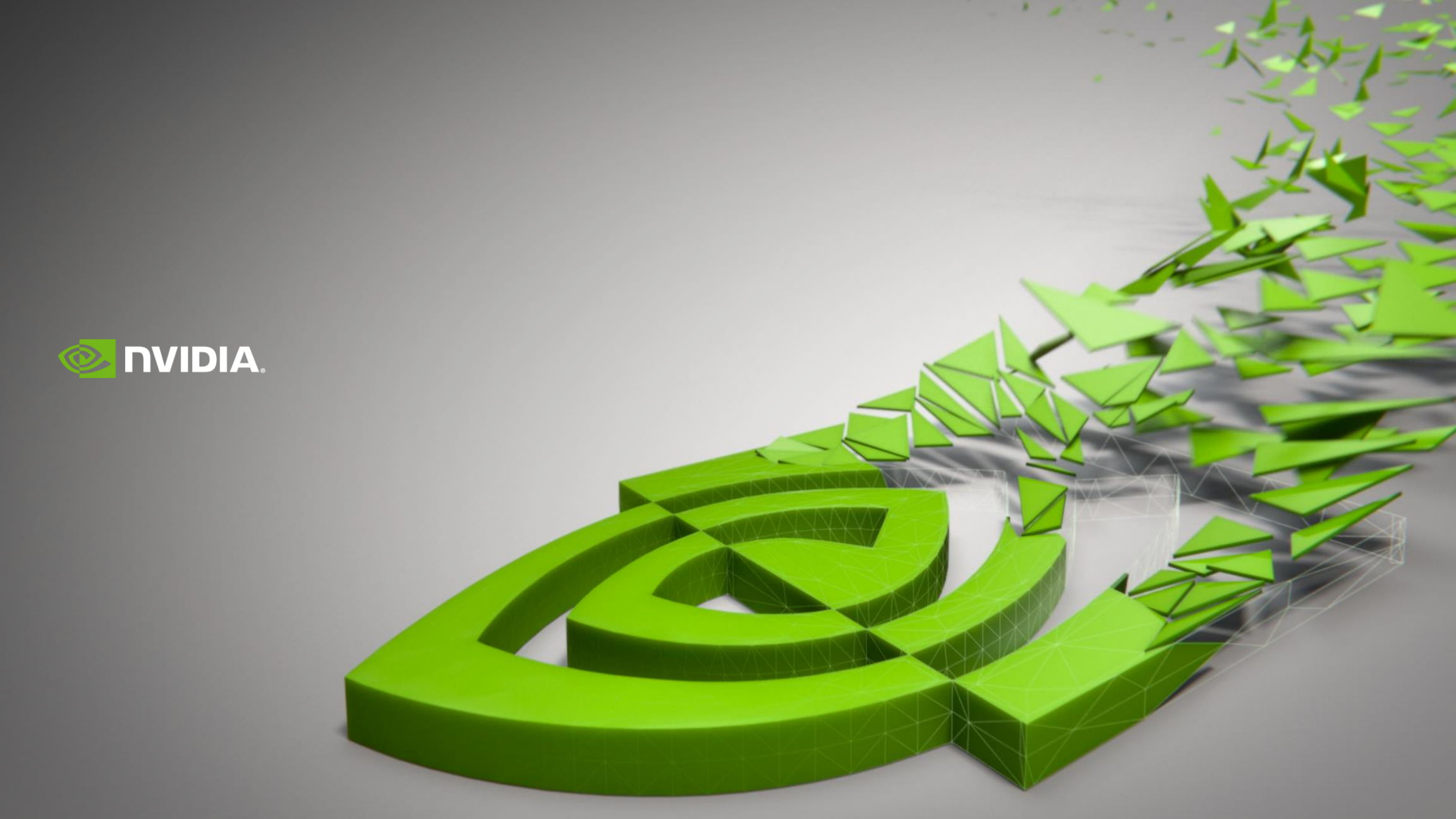


NVIDIA DEEP LEARNING SDK and CUDA

developer.nvidia.com/deep-learning-software

THERE WILL BE NO REASON TO ASK

WHY DOES HPC + AI MATTER?


NVIDIA

NVIDIA DEEP LEARNING SDK

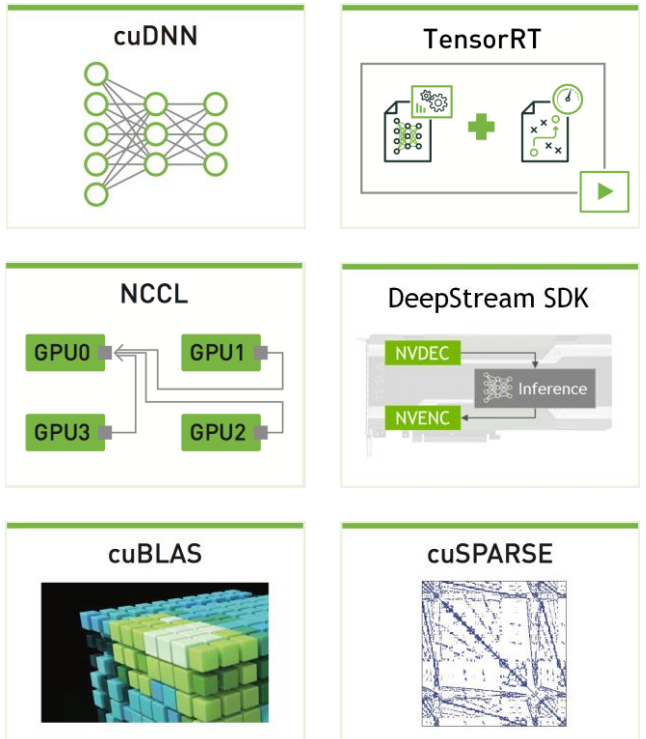
High performance GPU-acceleration for deep learning

Powerful tools and libraries for designing and deploying GPU-accelerated deep learning applications

High performance building blocks for training and deploying deep neural networks on NVIDIA GPUs

Industry vetted deep learning algorithms and linear algebra subroutines for developing novel deep neural networks

Multi-GPU and multi-node scaling that accelerates training to hundred of GPUs



"We are amazed by the steady stream of improvements made to the NVIDIA Deep Learning SDK and the speedups that they deliver."

— Frédéric Bastien, Team Lead (Theano) MILA

developer.nvidia.com/deep-learning-software

NVIDIA Collective Communications Library (NCCL) 2

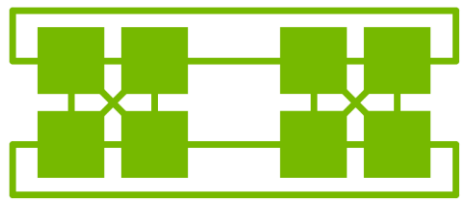
Multi-GPU and multi-node collective communication primitives

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

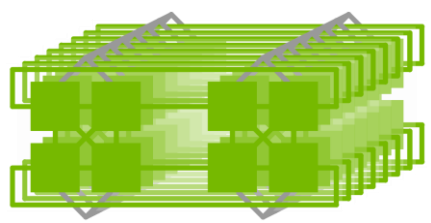
Fast routines for multi-GPU multi-node acceleration that maximizes inter-GPU bandwidth utilization

Easy to integrate and MPI compatible. Uses automatic topology detection to scale HPC and deep learning applications over PCIe and NVink

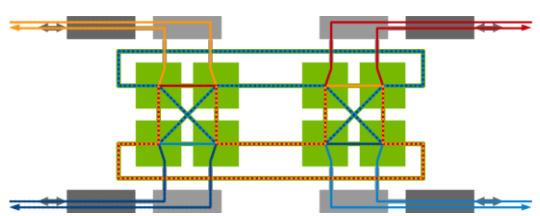
Accelerates leading deep learning frameworks such as Caffe2, Microsoft Cognitive Toolkit, MXNet, PyTorch and more



Multi-GPU:
NVLink
PCIe



Multi-Node:
InfiniBand verbs
IP Sockets

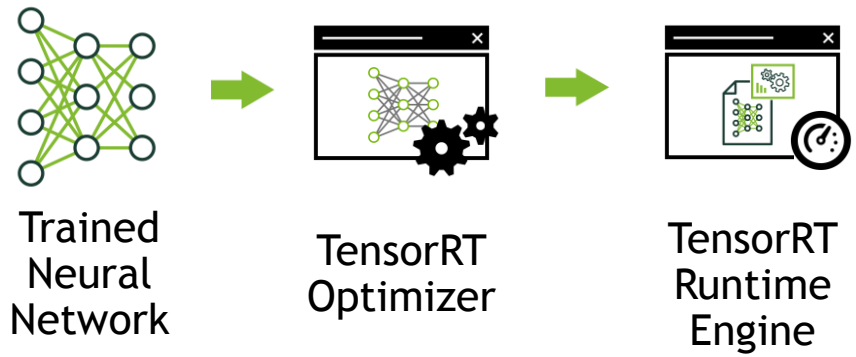


Automatic
Topology
Detection

developer.nvidia.com/nccl

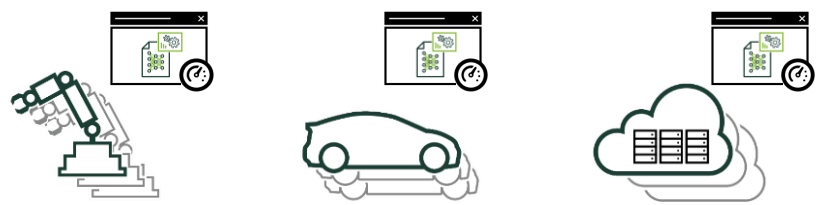
NVIDIA TensorRT

Deep Learning Inference Optimizer and Runtime



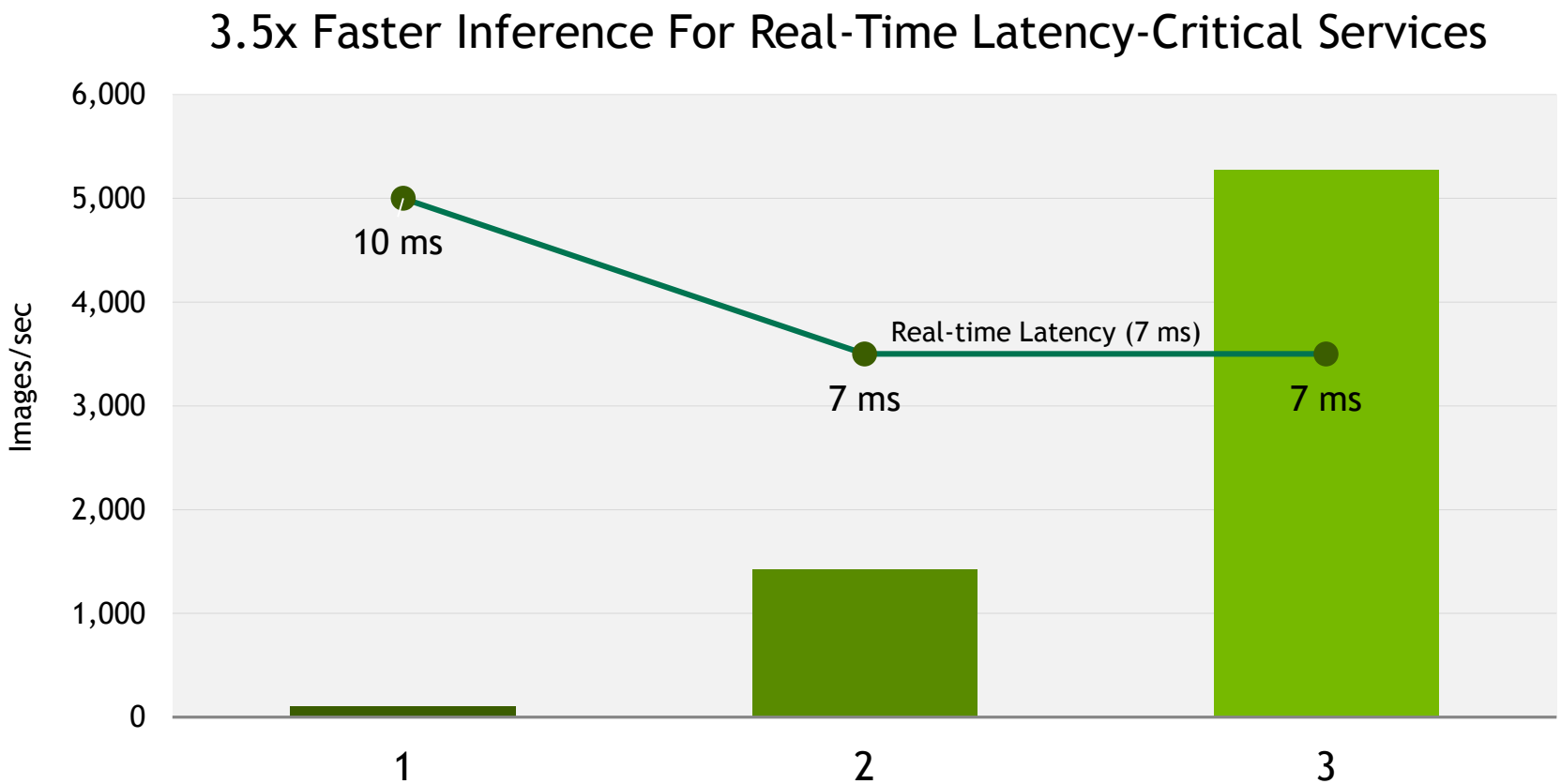
High performance neural network inference optimizer and runtime engine for production deployment

- Maximize inference throughput for latency-critical services in hyperscale datacenters, embedded, and automotive production environments.
- Optimize models trained in TensorFlow or Caffe to generate runtime engines that maximizes inference throughput
- Deploy faster, more responsive and memory efficient deep learning applications with INT8 and FP16 optimized precision support



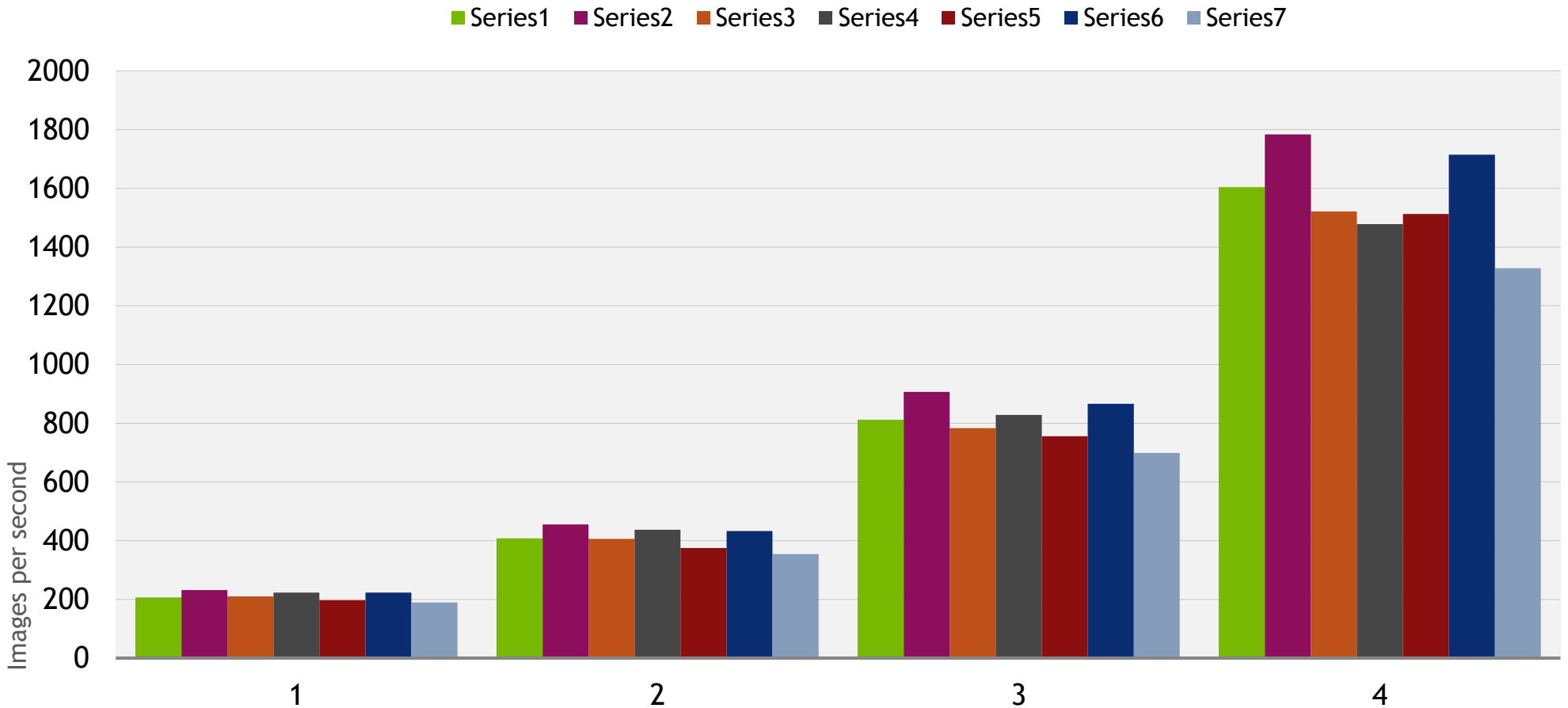
developer.nvidia.com/tensorrt

TensorRT 3: 3.5X FASTER INFERENCE



ResNet50 Inference, TensorRT performance (images/sec), TensorRT + K80: Batch Size =1, Latency = 10 ms
TensorRT + P100 (FP16): Batch Size =9 Latency= 7ms, TensorRT + V100 (FP16): Batch Size =26 Latency= 7ms,

RESNET-50 FP32 PERFORMANCE



4/30/2017 : *DGX-1 with Batch Size=64 per GPU. Chainer numbers are preliminary.*

COMBINING THE STRENGTHS OF HPC AND AI

HPC	AI
+40 years of Algorithms based on first principles theory Proven statistical models for accurate results in multiple science domains	New methods to improve predictive accuracy, insight into new phenomena and response time with previously unnavigable data sets
Develop training data sets using first principal models Apply Bayesian regression methods to expedite/ensure training accuracy Incorporate AI models in semi-empirical style applications to improve throughput Validate new findings from AI	Train inference models to improve accuracy and comprehend more of the physical parameter space Implement inference models with real time interactivity Analyze data sets that are simply intractable with classic statistical models Control and manage complex scientific experiments or apparatus

INNOVATION: A HISTORICAL VIEW