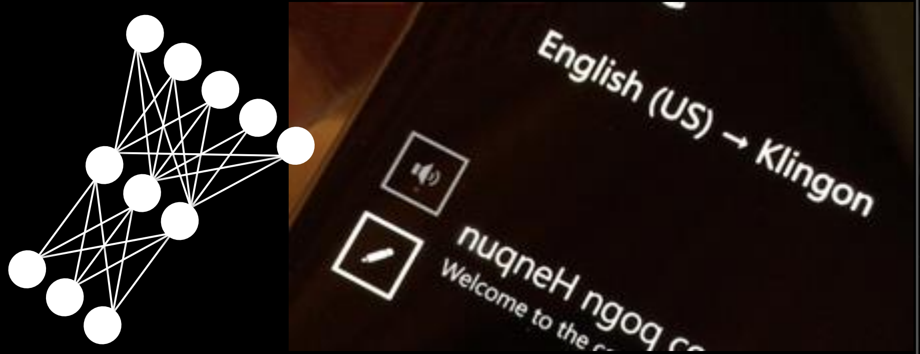
HPC + AI

Mike Houston

PRACTICAL DEEP LEARNING EXAMPLES

Image Classification, Object Detection, Localization, Action Recognition, Scene Understanding

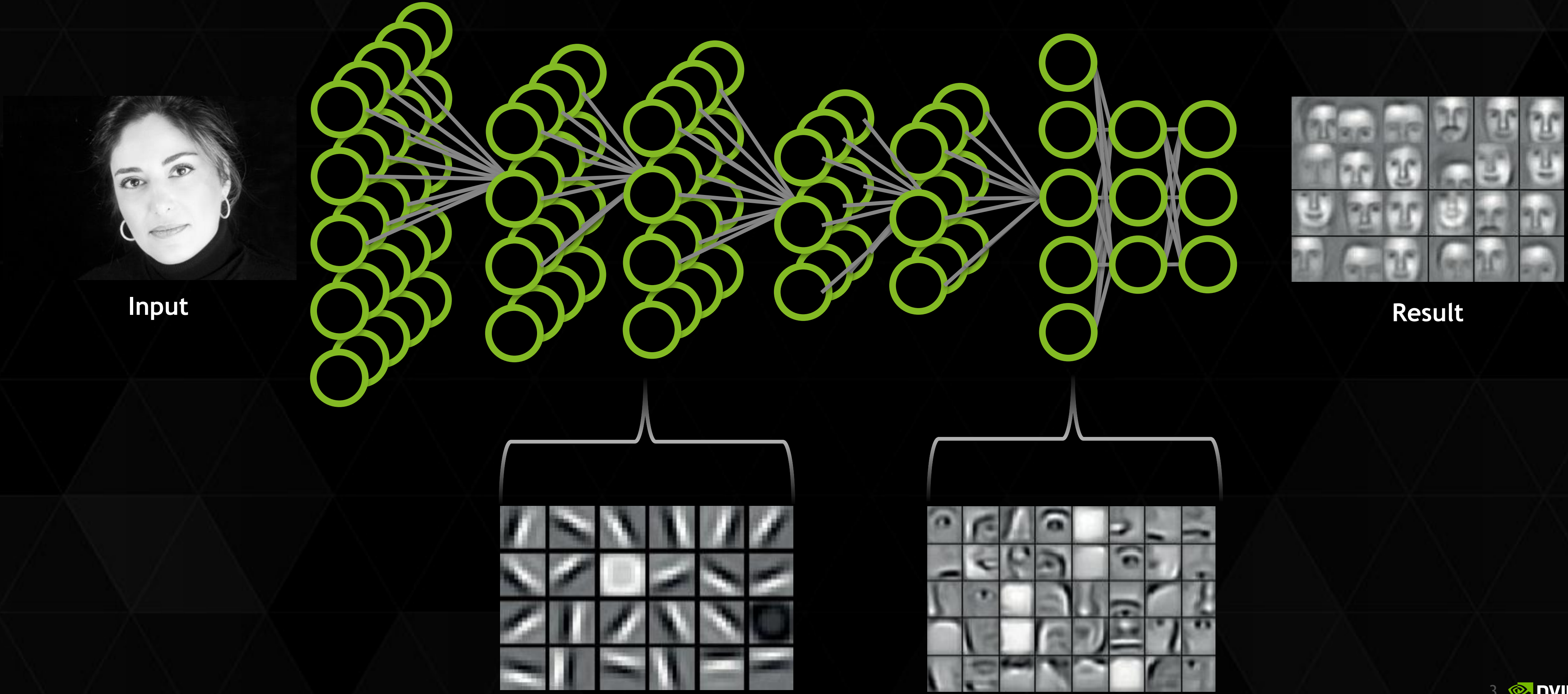


Speech Recognition, Speech Translation, Natural Language Processing

Pedestrian Detection, Traffic Sign Recognition

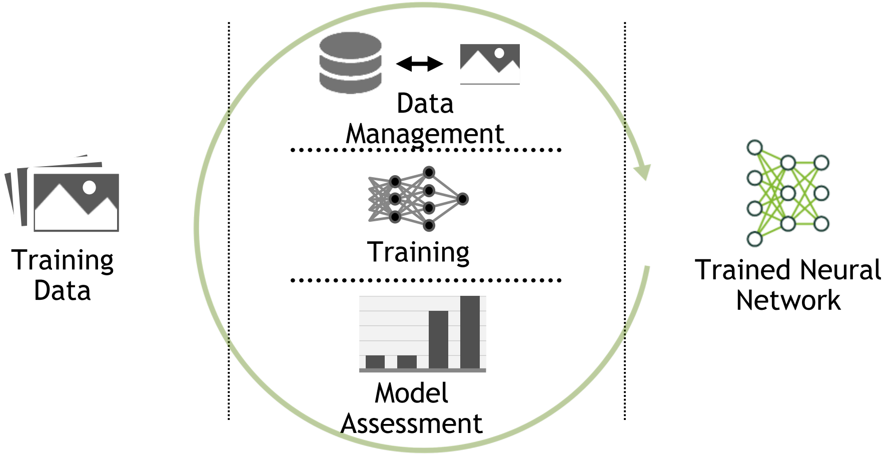
Breast Cancer Cell Mitosis Detection, Volumetric Brain Image Segmentation

WHAT IS DEEP LEARNING?



NVIDIA DEEP LEARNING SOFTWARE PLATFORM

TRAINING



Training Data

Data Management

Training

Model Assessment

Trained Neural Network

Caffe2 · Chainer · Microsoft Cognitive Toolkit · mxnet · TensorFlow · PYTORCH · theano

INFERENCE

Data center

GRE + TensorRT

Embedded

JETPACK SDK

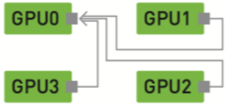
Automotive

DriveWorks SDK

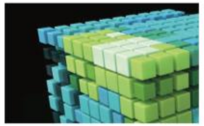
NVIDIA DEEP LEARNING SDK and CUDA

cuDNN

NCCL



cuBLAS



cuSPARSE

TensorRT

developer.nvidia.com/deep-learning-software

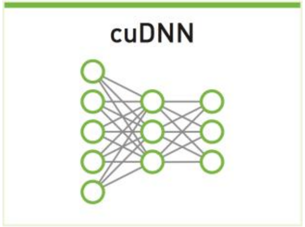
NVIDIA DEEP LEARNING SDK

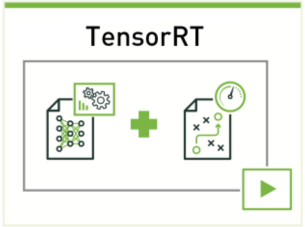
High performance GPU-acceleration for deep learning

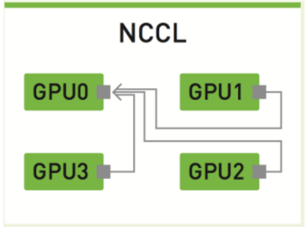
Powerful tools and libraries for designing and deploying GPU-accelerated deep learning applications

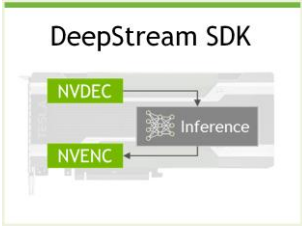
- High performance building blocks for training and deploying deep neural networks on NVIDIA GPUs
- Industry vetted deep learning algorithms and linear algebra subroutines for developing novel deep neural networks
- Multi-GPU and multi-node scaling that accelerates training to hundred of GPUs

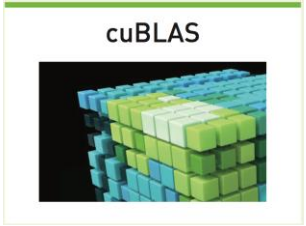
developer.nvidia.com/deep-learning-software

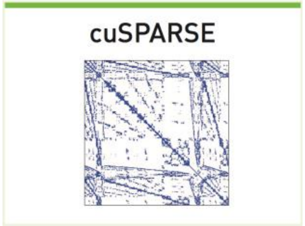

cuDNN


TensorRT


NCCL


DeepStream SDK


cuBLAS


cuSPARSE

> “ 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*

NVIDIA Collective Communications Library (NCCL) 2

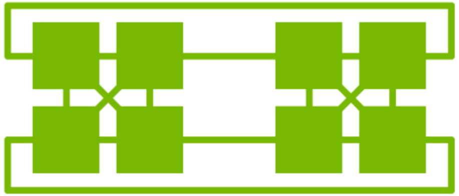
Multi-GPU and multi-node collective communication primitives

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

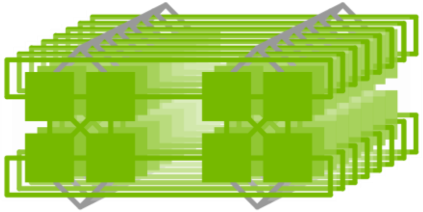
Fast routines for multi-GPU multi-node acceleration that maximizes inter-GPU bandwidth utilization

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

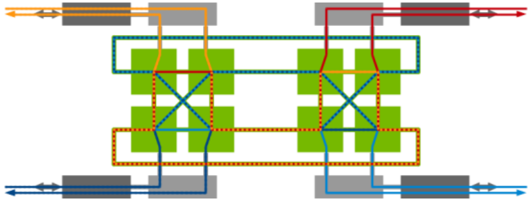
Accelerates leading deep learning frameworks such as Caffe2, Microsoft Cognitive Toolkit, MXNet, PyTorch and more



Multi-GPU:
NVLink
PCIe



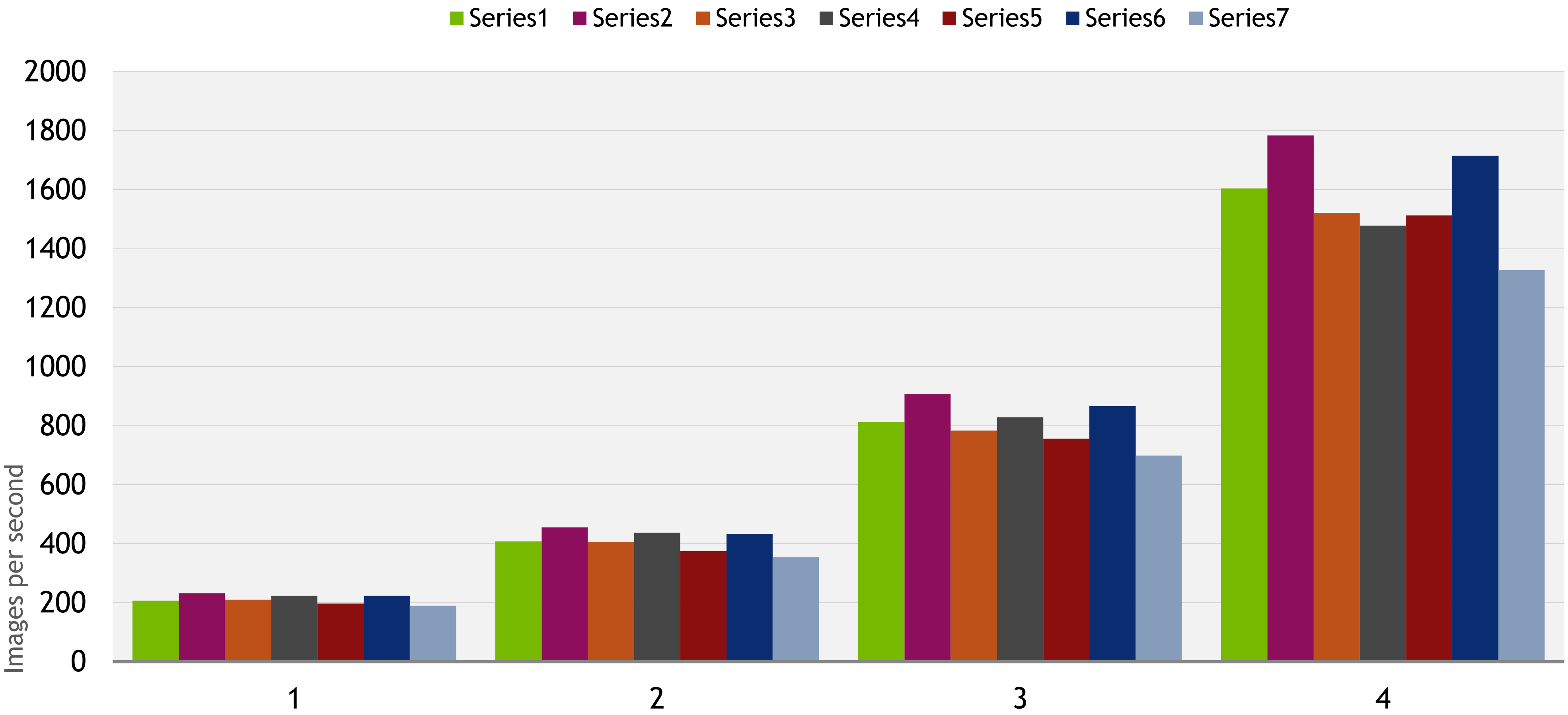
Multi-Node:
InfiniBand verbs
IP Sockets



Automatic
Topology
Detection

developer.nvidia.com/nccl

RESNET-50 FP32 PERFORMANCE



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

NVIDIA TensorRT

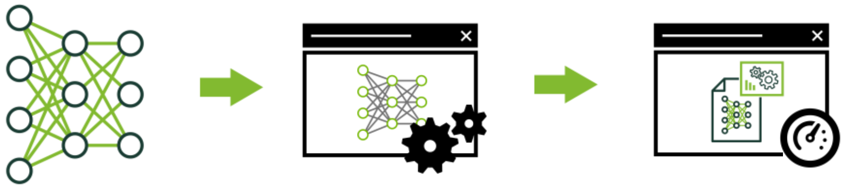
Deep Learning Inference Optimizer and Runtime

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

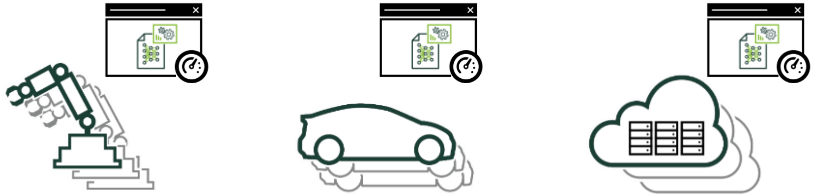
Deploy faster, more responsive and memory efficient deep learning applications with INT8 and FP16 optimized precision support



Trained Neural Network

TensorRT Optimizer

TensorRT Runtime Engine



Embedded

Automotive

Data center

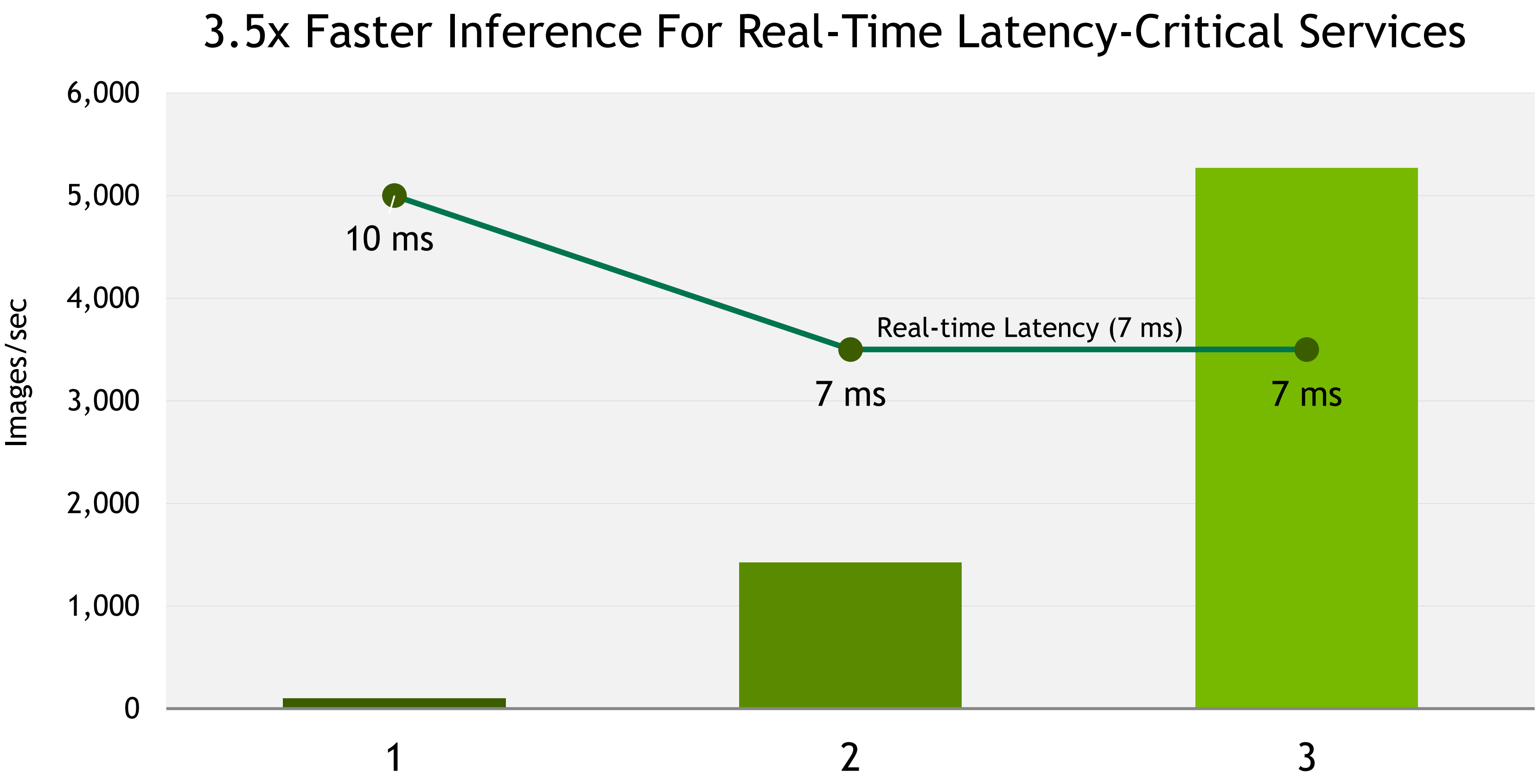
Jetson

Drive PX

Tesla

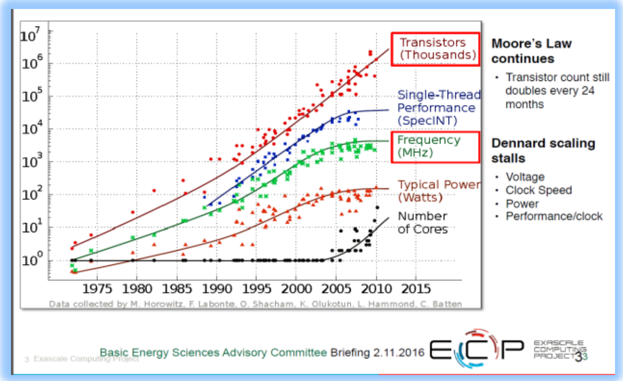
developer.nvidia.com/tensorrt

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,

FACTORS DRIVING HISTORIC CHANGES IN HPC



Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten

Moore's Law continues

- Transistor count still doubles every 24 months

Dennard scaling stalls

- Voltage
- Clock Speed
- Power
- Performance/clock

3 Exascale Computing Project

Basic Energy Sciences Advisory Committee Briefing 2.11.2016

3

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

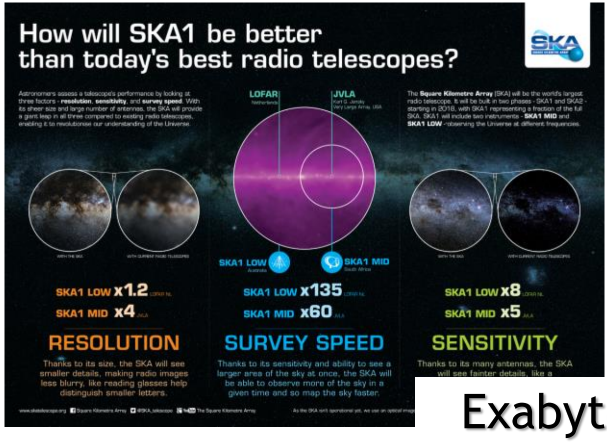
Dramatic improvements in accuracy, completeness and response time yield increased insight from huge volumes of data

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



How will SKA1 be better than today's best radio telescopes?

Astronomers assess a telescope's performance by looking at three factors - **resolution**, **sensitivity**, and **survey speed**. With its sheer size and large number of antennas, the SKA will provide a giant leap in all three compared to existing radio telescopes, enabling it to revolutionise our understanding of the Universe.

The **Square Kilometre Array** (SKA) will be the world's largest radio telescope. It will be built in two phases - SKA1 and SKA2 - starting in 2018, with SKA1 representing a fraction of the full SKA. SKA1 will include two instruments - **SKA1 MID** and **SKA1 LOW** - observing the Universe at different frequencies.

RESOLUTION

Thanks to its size, the SKA will see smaller details, making radio images less blurry, like reading glasses help distinguish smaller letters.

SURVEY SPEED

Thanks to its sensitivity and ability to see a larger area of the sky at once, the SKA will be able to observe more of the sky in a given time and so map the sky faster.

SENSITIVITY

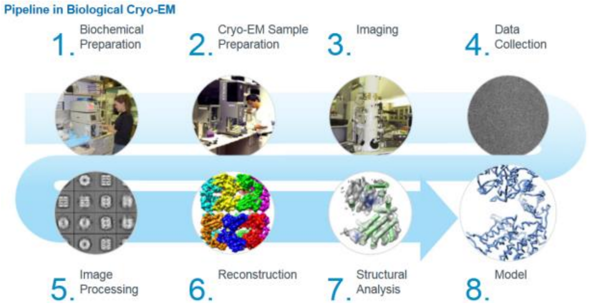
Thanks to its many antennas, the SKA will see fainter details, like a

www.skatelescope.org Square Kilometre Array @SKA_telescope The Square Kilometre Array

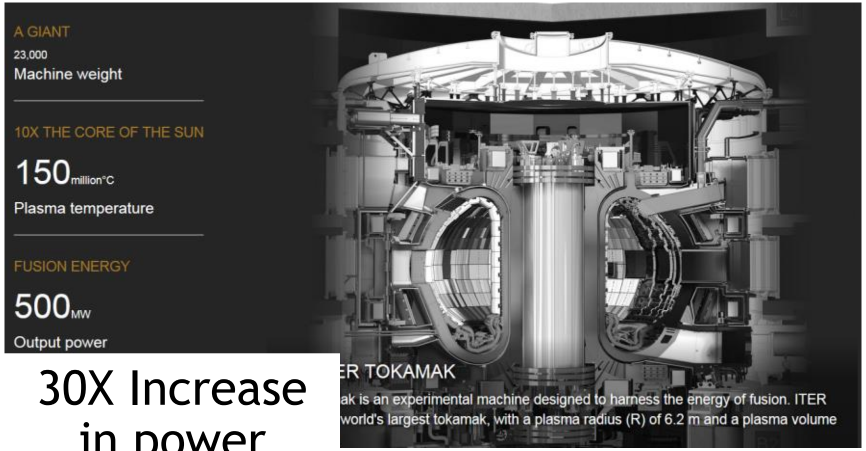
As the SKA isn't operational yet, we use an optical imag

Exabyte/Day

15 TB/Day



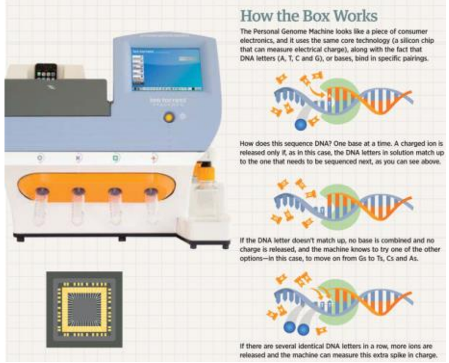
Pipeline in Biological Cryo-EM



ak is an experimental machine designed to harness the energy of fusion. ITER world's largest tokamak, with a plasma radius (R) of 6.2 m and a plasma volume

30X Increase in power

10X Increase in Data Volume



How the Box Works

The Personal Genome Machine looks like a piece of consumer electronics, and it uses the same core technology (a silicon chip that can measure electrical charge), along with the fact that DNA letters (A, T, C and G), or bases, bind in specific pairings.

How does this sequence DNA? One base at a time. A charged ion is released only if, as in this case, the DNA letters in solution match up to the one that needs to be sequenced next, as you can see above.

If the DNA letter doesn't match up, no base is combined and no charge is released, and the machine knows to try one of the other options—in this case, to move on from Gs to Ts, Cs and As.

If there are several identical DNA letters in a row, more ions are released and the machine can measure this extra spike in charge.

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

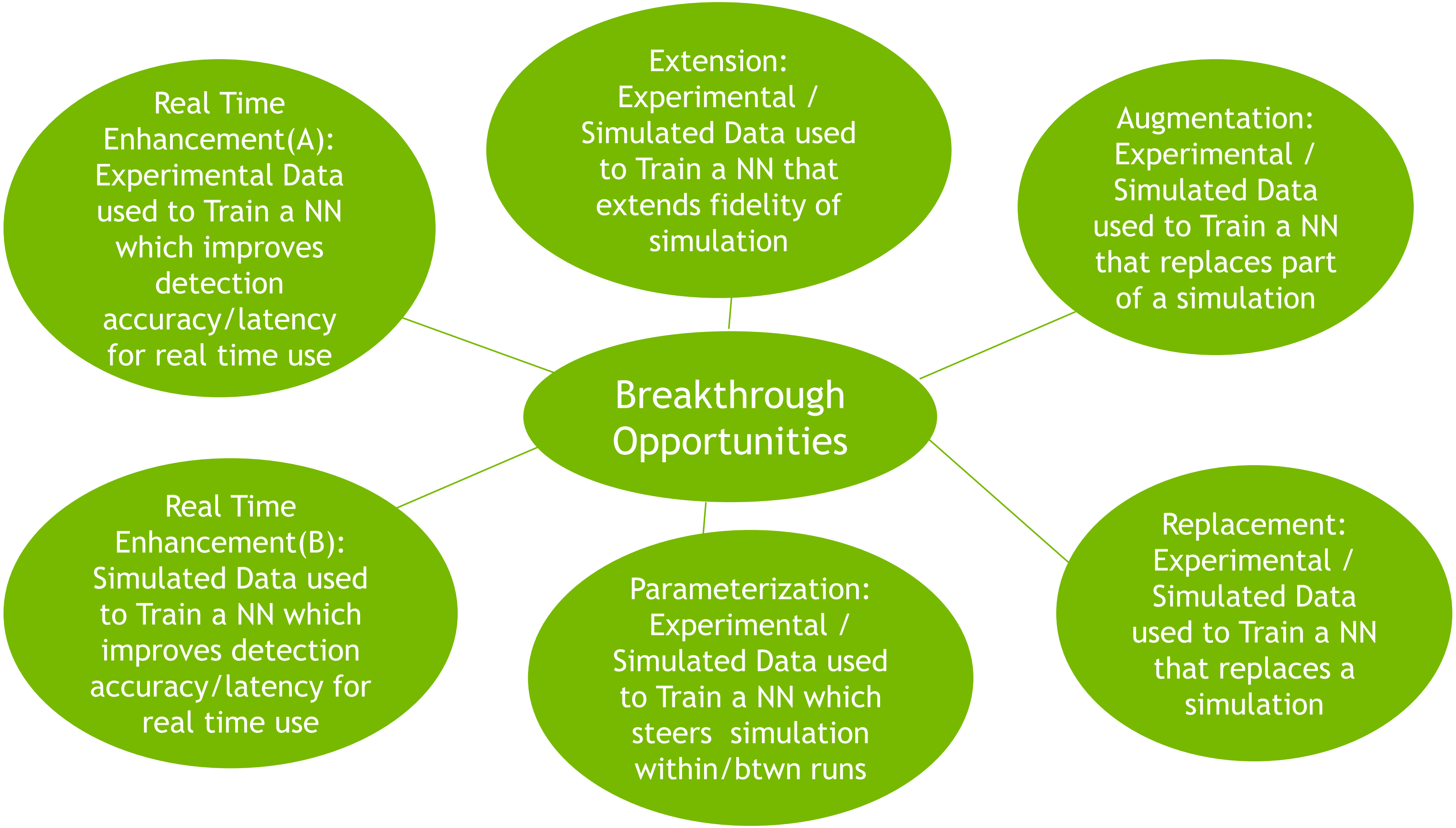
*

*

*

TAXONOMY

Examples of HPC + AI Convergence



Breakthrough Opportunities

Real Time Enhancement(A): Experimental Data used to Train a NN which improves detection accuracy/latency for real time use

Extension: Experimental / Simulated Data used to Train a NN that extends fidelity of simulation

Augmentation: Experimental / Simulated Data used to Train a NN that replaces part of a simulation

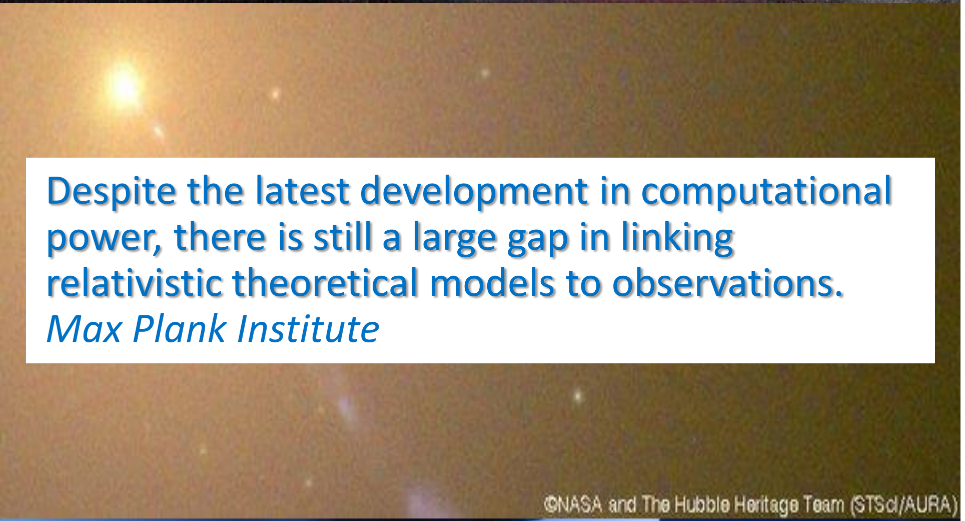
Real Time Enhancement(B): Simulated Data used to Train a NN which improves detection accuracy/latency for real time use

Parameterization: Experimental / Simulated Data used to Train a NN which steers simulation within/btwn runs

Replacement: Experimental / Simulated Data used to Train a NN that replaces a simulation

MULTI-MESSENGER ASTROPHYSICS

©NASA/JPL-Caltech



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

©NASA and The Hubble Heritage Team (STScI/AURA)

©NASA/ESA/Richard Massey (California Institute of Technology)

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

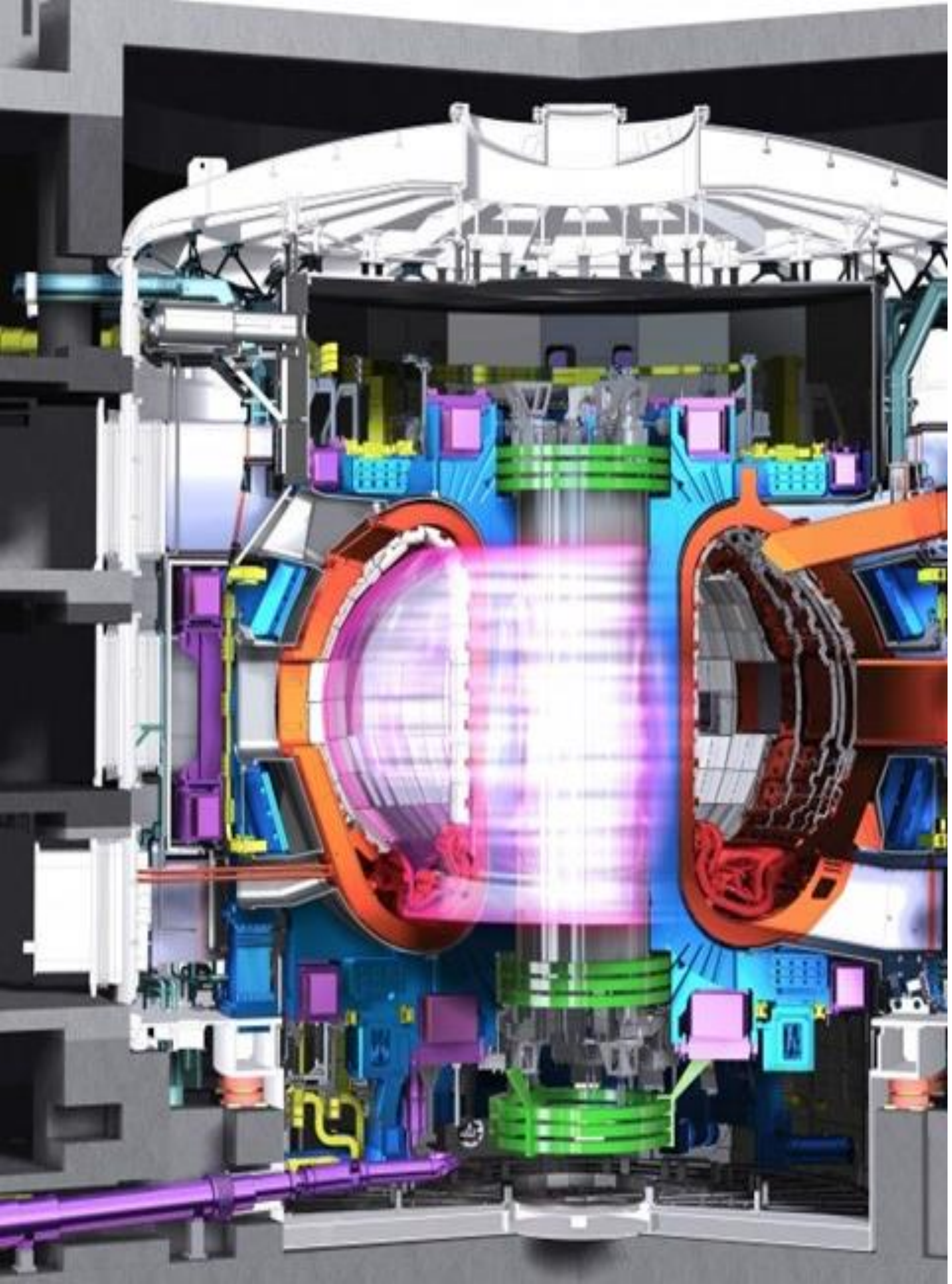
The initial a LIGO discoveries were successfully completed using classic data analytics. The processing pipeline used hundreds of CPU’s where the bulk of the detection processing was done offline. Here the latency is far outside the range needed to activate resources, such as the Large Synaptic Space 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 Einstein 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.



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.

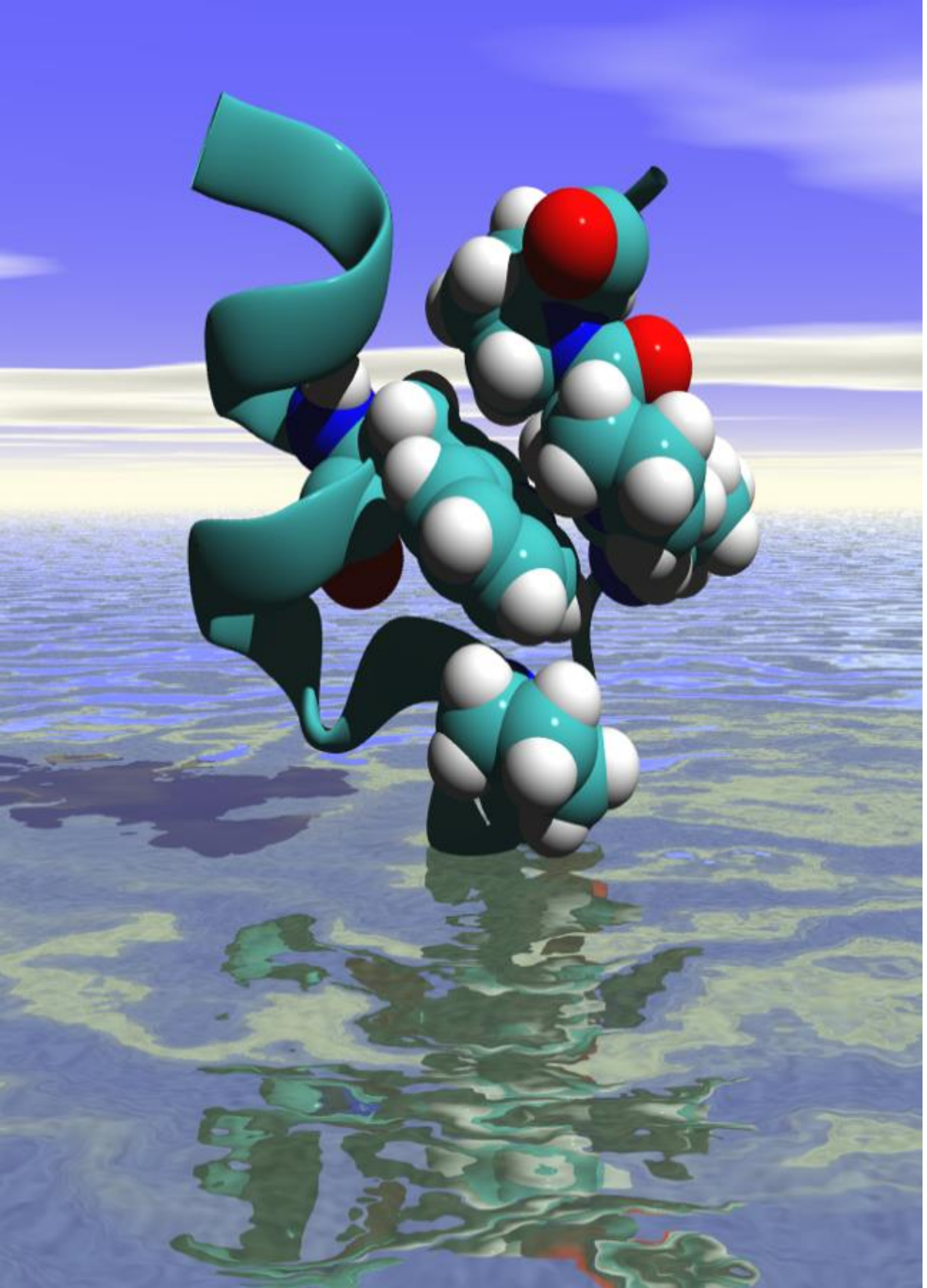
Solution

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

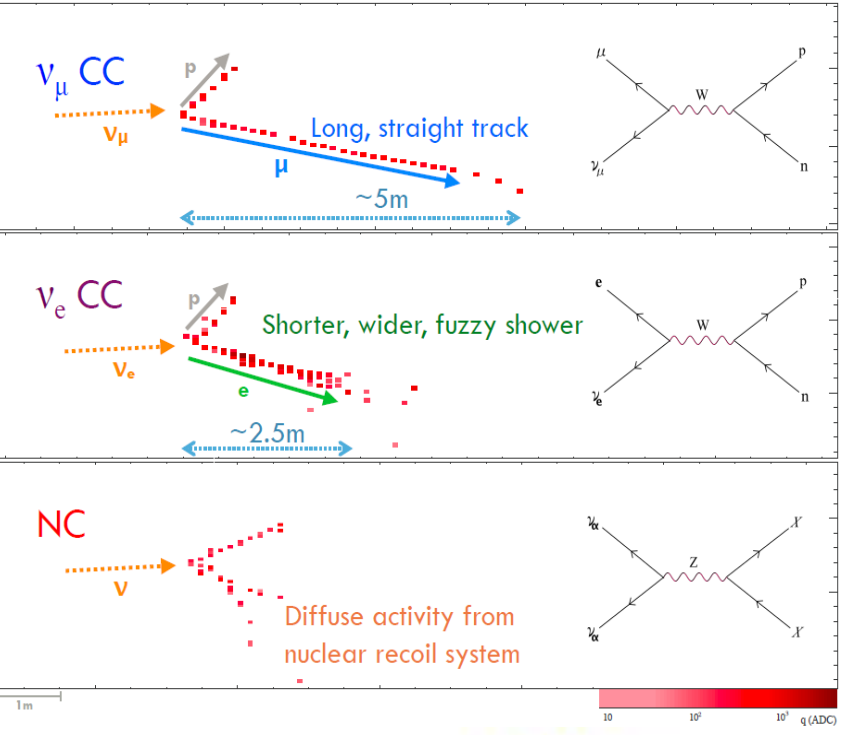
Researchers at the University of Florida and the University of North Carolina leveraged GPU deep learning to develop ANAKIN-ME, to reproduce molecular energy surfaces with super speed (microseconds versus several minutes), extremely high (DFT) accuracy, and at 1-10/millionths of the cost of current computational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

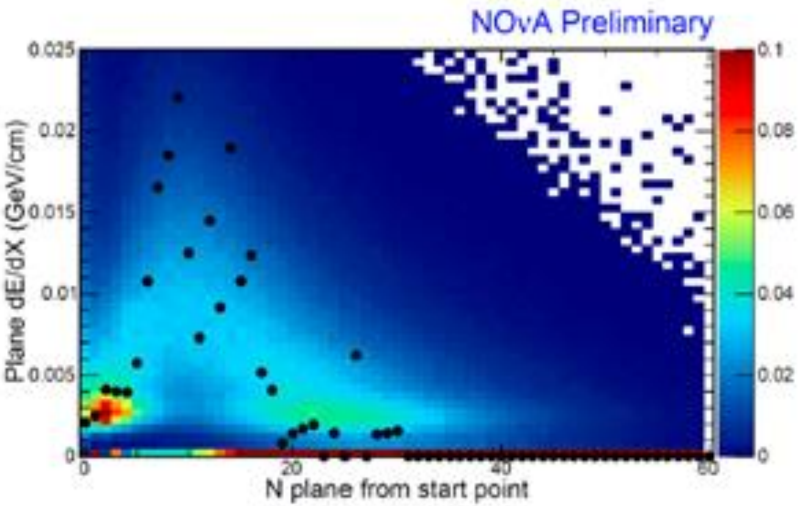
Impact

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 based 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

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.

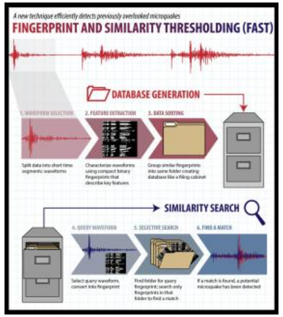
MeteoSwiss

© Warner Bros.

Earthquake Prediction

Multiple Examples of AI for earthquake prediction are underway

Shaazam for Earthquakes



A new technique efficiently detects previously overlooked microquakes

FINGERPRINT AND SIMILARITY THRESHOLDING (FAST)

DATABASE GENERATION

1. WAVEFORM SELECTION: Split data into short time segments: waveforms

2. FEATURE EXTRACTION: Characterize waveforms using compact binary fingerprints that describe key features

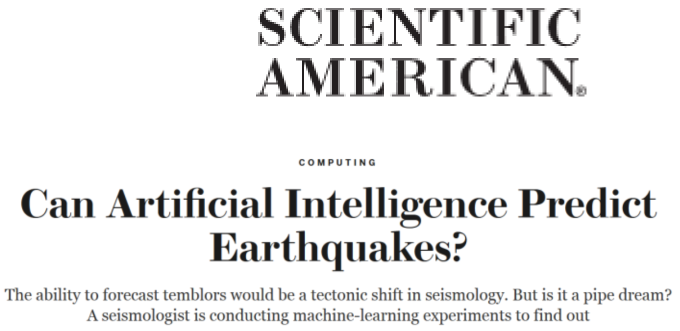
3. DATA SORTING: Group similar fingerprints into same folder creating database like a filing cabinet

SIMILARITY SEARCH

4. QUERY WAVEFORM: Select query waveform, convert into fingerprint

5. SELECTIVE SEARCH: Find folder for query fingerprint; search only fingerprints in that folder to find a match

6. FIND A MATCH: If a match is found, a potential microquake has been detected



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

PRINCETON UNIVERSITY

SCEC an NSF+USGS center

Stanford University

NVIDIA