CONVERGENCE: HPC + AI

Introducing a new era for innovation in the Exascale timeframe
FACTORS DRIVING INNOVATION IN CONVENTIONAL HPC SYSTEM ARCHITECTURE

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 (High Performance Data Analytics) 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

15 TB/Day

Exabyte/Day

10X Increase in Data Volume

30X Increase in power

Personal Genomics
THE POTENTIAL OF EXASCALE HPC + AI

**HPC**
- +40 years of Algorithms based on first principles theory
- Proven statistical models for accurate results in multiple science domains

**AI**
- 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
CONVERGED EXASCALE ERA SYSTEM

- **Modeling**
- **Preparation**
- **Data sets**
- **Training**
- **Inferencing**
- **New data**

**Theories**
- Formulation of Hypothesis as Mathematical Model
- Generation of Explanatory Hypothesis
- Discovery of Patterns and Anomalies

**DEDUCTION** (Drawing necessary conclusions)
- Execution of Model to Generate Predictions

**ABDUCTION** (Making guesses)
- Induction (Inferring generalizations from sampling)
- Assimilation of Predictions & Data to Evaluate Hypothesis

**Predictions**

**Observations**
- From Instruments, Sensors, Records, Visualizations etc.

BIG DATA AND EXTREME-SCALE COMPUTING:
PATHWAYS TO CONVERGENCE
Tech Report No. ICL-UT-17-08
**Usage Taxonomy**

Organizing HPC + AI Convergence

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

Experimental/simulated data is used to train a NN, where resulting inference engaging is used to for real-time detection/control of an experiment or clinical delivery system.
The NN is improved 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
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 a step to the ultimate goal to combine multiple observational data sources that not only hear but also see to the complete spectrum of data in real time.

Challenge
The initial aLIGO 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. The latency is far outside the range needed to activate resources, such as optical, infrared or radio telescopes which observe phenomena in the electromagnetic spectrum in time to “see” what aLIGO can “hear”.

Solution
A DNN was developed and trained with simulated data and verified using from the CACTUS/Einstein Toolkit. The DNN was shown to produce better accuracy with latencies 4500x faster than the original CPU based pattern matching waveform detection.

Impact
Faster and more accurate detection of gravitational waves with the potential to steer other observational data sources.
Background
The “Grand Challenge” of fusion energy would offer the humankind changing opportunity to provide clean, safe energy for millions of years.
ITER is a $25B international experiment to develop the prototype to demonstrate commercially viable fusion reactor.

Challenge
The plasma in a fusion reactor is highly turbulent at the edges of the flow, and disruptions can occur that break the magnetic confinement, which can cause damage to the physical reactor.
It is critical to predict when a disruption will occur to prevent damage and maintain safe operation.
Traditional simulation and ML approaches were 65% to 85% accurate with 5% false alarm rate.

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. Current level of accuracy is 95% prediction with 5% false alarm rate.

Impact
Vision is to operate ITER with FRNN, operating and steering experiments in real-time to minimize damage and down-time.
Background
It takes 14 years and $2.5 Billion to develop 1 drug
Higher than 99.5% failure rate after the drug discovery phase

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

Solution
Researchers at the University of Florida and the University of North Carolina leveraged GPU deep learning to develop a custom framework ANAKIN-ME, to reproduce molecular energy surfaces with super speed (microseconds versus several minutes), extremely high (DFT) accuracy, and at up to 6 orders of magnitude improvement in speed.

Impact
Speed and accuracy could start a revolution in computational chemistry – and forever change the way we discover the medicines of the future
Converged HPC Accelerates Drug Discovery

The “drug discovery” phase of the development process involves exploring all the different possible combinations of protein molecules (targets) and drug chemical compounds to ensure the drug will do what it’s designed to do. Classic Molecular Dynamics simulations are very time-consuming and expensive. Machine Learning models have been designed to help predict probability of the target molecules interacting with the drug chemical compounds, but still require significantly more performance to deliver improved accuracy.

Researchers developed and trained a convolutional neural network accelerated with NVIDIA GPU’s to improve the model performance and prediction accuracy. Ultimately, they improved prediction accuracy from approximately 52% to 70% compared to other machine learning-based models (Vina Docking). (35% relative improvement)
AN AI MONITOR OF EARTH’S VITALS

The Earth’s climate has changed throughout history, but in recent years there have been record increases in temperature, glacial retreat and rising sea levels. NASA Ames is using satellite imagery to measure the effects of carbon and greenhouse gas emissions on the planet. To do so, they developed DeepSat—a deep learning framework for satellite image classification trained on a GPU-powered supercomputer. The enhanced satellite imagery will help scientists plan to protect ecosystems and farmers improve crop production.
AI Accelerates the Production of Ultra Cold Gases

Bose-Einstein Condensate (BEC) is a state of matter formed by cooling a gas to near-zero absolute temperature. BEC is achieved by controlling the intensity of the lasers to trap only the ultra-cold atoms and allowing other atoms to escape. BECs are super sensitive to external disturbances. This makes them suitable for very precise measurements of things like tiny changes in Earth’s magnetic field or gravity.

Researchers at the University of North South Whales used AI to create a BEC gas 14 times faster than conventional methods.
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
Multiple Examples of AI for earthquake prediction are underway

Shaazam for Earthquakes

Can Artificial Intelligence Predict Earthquakes?

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

SC/EC

an NSF + USGS center

© Warner Bros.
# WORLD’S ONLY UNIFIED AI + HPC PLATFORM

Nvidia Tesla Platform for Accelerating Data Centers

<table>
<thead>
<tr>
<th>APPLICATIONS</th>
<th>INTERNET SERVICES</th>
<th>ENTERPRISE APPLICATIONS</th>
<th>HPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alibaba</td>
<td>Amazon</td>
<td>Facebook</td>
<td>ANSYS</td>
</tr>
<tr>
<td>Amazon</td>
<td>Bing</td>
<td>eBay</td>
<td>SIMULIA</td>
</tr>
<tr>
<td>Facebook</td>
<td>Flickr</td>
<td>Google</td>
<td>GROMACS</td>
</tr>
<tr>
<td>Pinterest</td>
<td>Microsoft</td>
<td>Pinterest</td>
<td>c+450 Applications</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>Skype</td>
<td>Twitter</td>
<td>AWS</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>Twitter</td>
<td>Yahoo</td>
<td>Intel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INDUSTRY FRAMEWORKS &amp; TOOLS</th>
<th>FRAMEWORKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe</td>
<td>mxnet</td>
</tr>
<tr>
<td>mxnet</td>
<td>PyTorch</td>
</tr>
<tr>
<td>PyTorch</td>
<td>TensorFlow</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>Theano</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NVIDIA SDK</th>
<th>DEEP LEARNING SDK</th>
</tr>
</thead>
<tbody>
<tr>
<td>cuDNN</td>
<td>cuBLAS</td>
</tr>
<tr>
<td>TensorRT</td>
<td>cuSPARSE</td>
</tr>
<tr>
<td>NCCL</td>
<td>DeepStream SDK</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TESLA GPU &amp; SYSTEMS</th>
<th>TESLA GPU</th>
<th>NVIDIA DGX-1</th>
<th>NVIDIA HGX-1</th>
<th>SYSTEM OEM</th>
<th>CLOUD</th>
</tr>
</thead>
</table>
LEVERAGE NVIDIA PROGRAMS
Free download, Learn DL, or Enhance your code

NVIDIA GPU Cloud (Optimized SW containers)

20,000+ Registered Organizations
30 Containers

DLI (Deep Learning Institute)

Developer Program

Join the NVIDIA Developer Program
Access everything you need to develop with NVIDIA products.
Register Now

NVIDIA SDKs for Developer
NVIDIA AI TECHNOLOGY CENTRE (NVAITC)
A team of AI computing experts to conduct applied solution research

- Collaborate in Training, Evangelism, Research Project, or Strategic Lab

OUR MISSION

- Make NVIDIA trusted advisor and partner in AI
- Support groundbreaking research, learn state of art through collaboration, and develop talents for NVIDIA
- Share best practices and use cases among NVAITC network