SYNTHESIS MODELING: THE INTERSECTION OF HPC AND MACHINE LEARNING

NEW APPROACH CAN SOLVE PREVIOUSLY INTRACTABLE PROBLEMS & REDUCE COSTS

INTRODUCTION

Historically, numerical analysis has formed the backbone of supercomputing for decades by applying mathematical models of first-principle physics to simulate the behavior of systems from subatomic to galactic scale. Recently, scientists have begun experimenting with a relatively new approach to understand complex systems using machine learning (ML) predictive models, primarily Deep Neural Networks (DNN), trained by the virtually unlimited data sets produced from traditional analysis and direct observation. Early results indicate that these "synthesis models," combining ML and traditional simulation, can improve accuracy, accelerate time to solution and significantly reduce costs.

This paper examines how machine learning is being used as a new tool for scientific discovery, augmenting traditional techniques to improve our understanding of the universe. As scientists become more comfortable with this new approach, and as the methodologies become more robust, we believe that machine learning has the potential to emerge as a mainstream tool for many areas of scientific computing.

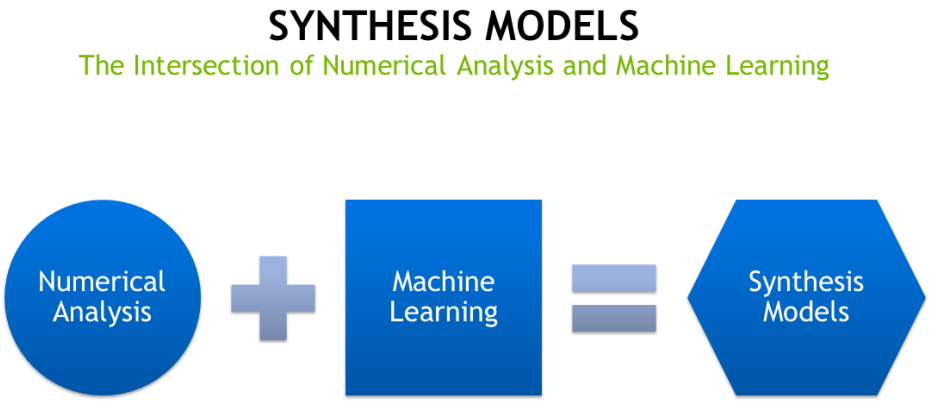
APPLYING MACHINE LEARNING IN HPC

Machine learning enables one to predict or approximate simulation outcomes instead of calculating a full simulation using traditional numerical analysis. Data from simulation and observation are used to train a deep neural network, which is then used for inference analysis to efficiently simulate the system being studied. While numerical models are programs that algorithmically embody the known science, ML models are learned from vast stores of data. While traditional approaches require competent, specialized domain scientist to create a good numerical model, any talented data scientist can train a good ML model if they have access to sufficient quantities of tagged data, without knowing anything about the particular scientific field of study.

Note that numerical analysis typically requires a lot of expensive 64-bit floating-point precision calculations, while a trained neural network can typically be used to predict outcomes or classify a data sample using far more efficient 8-bit integer operations.

Consequently, early research projects have shown that ML often requires orders of magnitude fewer resources to unlock problems that have often been beyond the grasp of traditional techniques. Since performance increases in traditional high-performance computing (HPC) have been highly dependent on Moore's Law, this approach presents a promising avenue to explore.

FIGURE 1: THE SYNTHESIS OF NUMERICAL ANALYSIS AND MACHINE LEARNING CAN CREATE NEW PREDICTIVE SIMULATION MODELS



Source: Moor Insights & Strategy

THREE APPROACHES TO APPLYING MACHINE LEARNING IN HPC

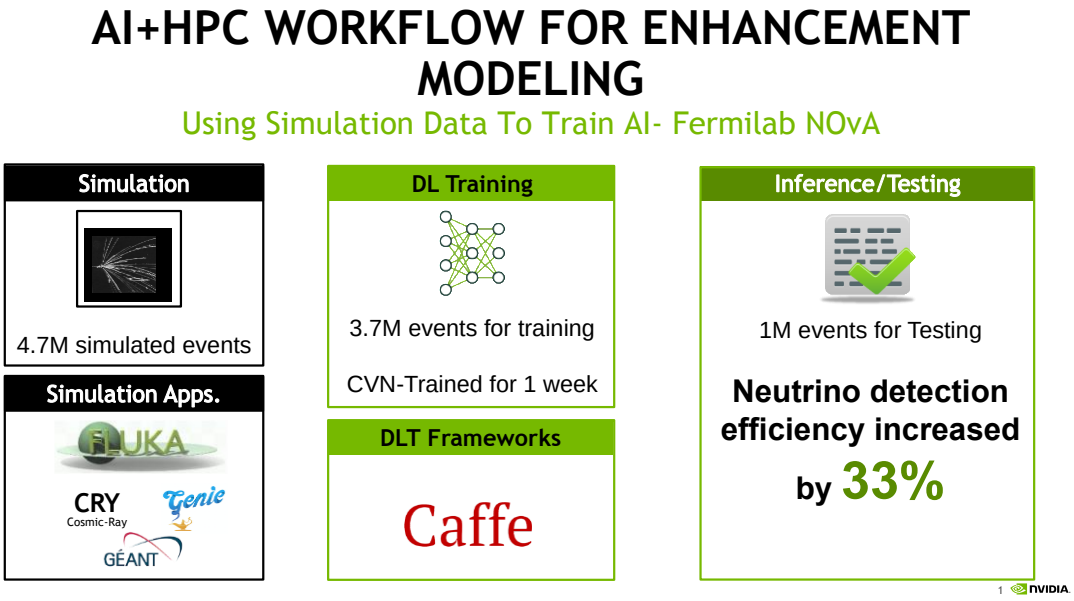
While Machine learning is a relatively new feature on the HPC landscape, scientists are already realizing impressive results by applying synthesis modeling in research. There are three primary approaches being explored today: simulation enhancement, modulation and approximation, described below. Fortunately, researchers can deploy a unified platform to accelerate synthesis modeling while avoiding redundant data storage and movement, using the same hardware (CPUs with fast GPUs) for simulation, model development, training and inference.

1. Enhancement

Pioneering researchers are using machine learning to enhance and augment existing simulations to improve accuracy and reduce latencies by as much as three orders of

magnitude. Simulation provides both a starting point for the DNN, and training data for the neural networks to refine the output of the numerical model. With this approach, machine learning has been shown to improve fidelity over existing methods. Researchers working to detect neutrinos at Fermilab NOvA have realized a 33% improvement in detection using a convolutional neural network that was trained by data produced by four different HPC simulations.

FIGURE 2: MACHINE LEARNING CAN IMPROVE NEUTRINO DETECTION BY COMBINING SIMULATION RESULTS FROM DIFFERENT MODELS TO PRODUCE A SUPERIOR MODEL.

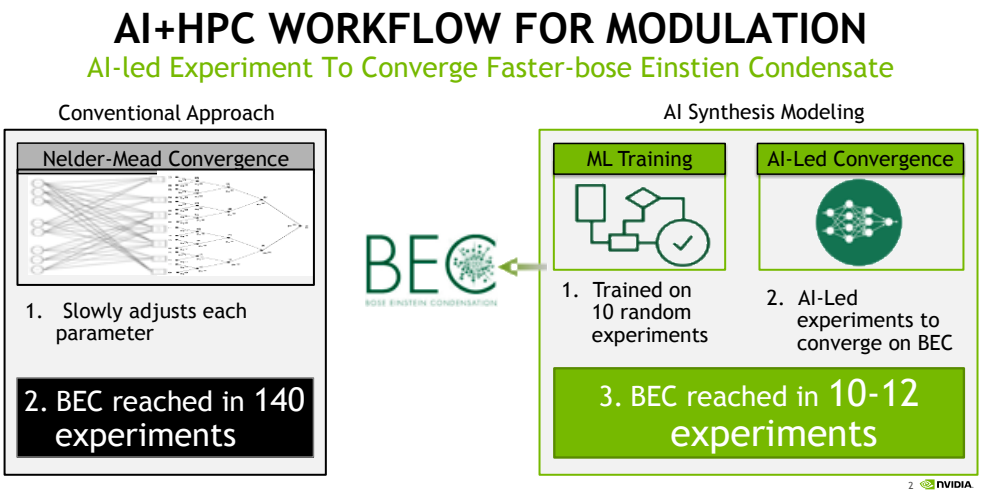


Source: NVIDIA

2. Modulation

ML can also be used to steer simulation or experiments between successive iterations, accelerating convergence to a stable, reliable model. In this case, simulation is used to train the neural network, which then refines the model to create input for the next simulation run or experiment. Researchers at the University of New South Wales experimenting with Bose-Einstein Condensates (BEC) have been able to achieve the BEC state in only 10-12 experiments using synthesis modeling, compared to the 140 experiments that are typically required using traditional models.

FIGURE 3: BOSE-EINSTEIN CONDENSATE ACHIEVED CONVERGENCE AFTER ONLY 10-12 EXPERIMENTS USING MACHINE LEARNING, COMPARED TO 140 EXPERIMENTS USING THE TRADITIONAL APPROACH.



Source: NVIDIA

3. Approximation

Finally, perhaps the most impactful application of machine learning in HPC is to replace production numerical simulation models with ML approximation. This approach has the potential to transform HPC. However, adoption will require scientists to embrace a method that may eventually render obsolete the codes they have spent decades to develop and optimize. Note that numerical simulation models are still required to train and tune the ML models as conditions and science evolve.

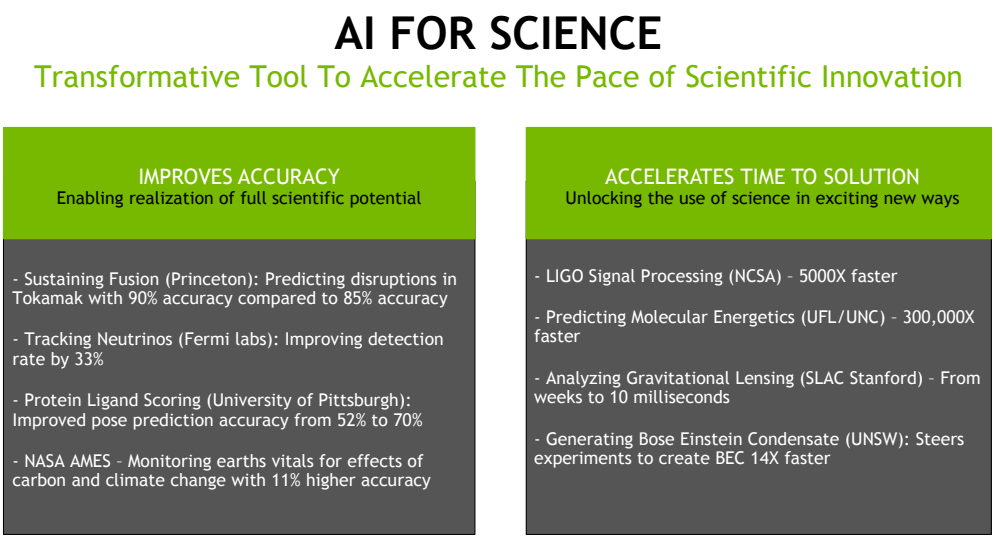
But, there is also some good news here for numerical model application authors, especially domain experts who are not skilled computer scientists and may find parallel optimization onerous and daunting. Existing numerical codes will no longer require the time-consuming rewrites needed to achieve high efficiency on new parallel processors. Scientists can now focus on the underlying science. When using ML for approximation, numerical codes can actually be useful even when relatively inefficient, as these codes would no longer be used to run the application at final scale. The numerical models need only be large enough to create training data to represent what happens at a single point in the problem set domain decomposition. In effect, the numerical model code

becomes the development environment, while the ML model acts as the deployment-at-scale, run-time model.

Machine Learning Use Cases in HPC

While we are still in the early days of experimentation of ML in HPC, NVIDIA has compiled a few references that demonstrate higher accuracy and faster time to solution. We've selected three examples below to highlight the approaches of synthesis modeling: enhancement, modulation, and approximation.

Figure 4: Example use cases for Synthesis Modeling



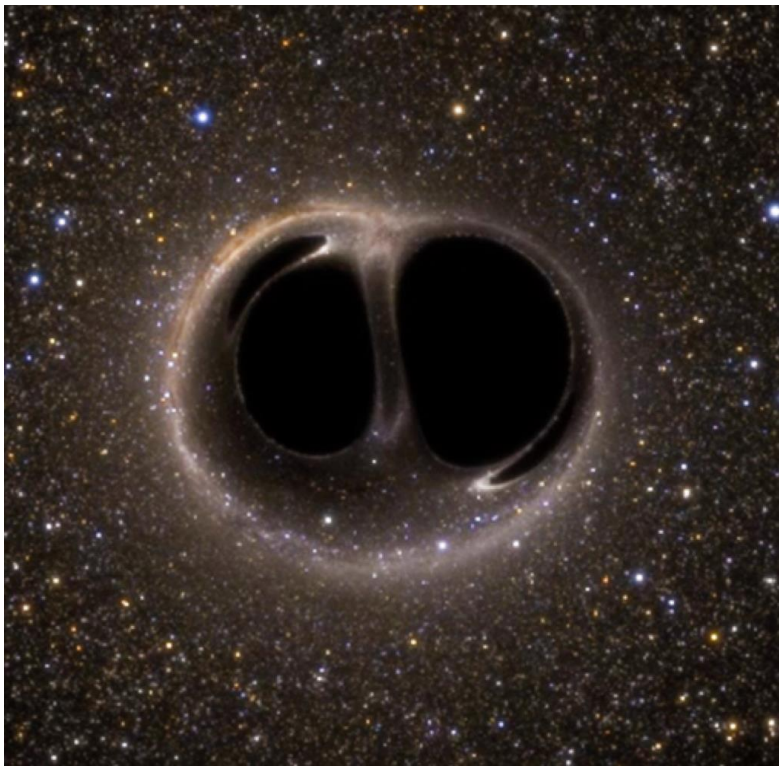
Source: NVIDIA

Application of Model Enhancement at LIGO

Background

The Laser Interferometer Gravitational-wave Observatory (LIGO) experiment successfully discovered the first signals to prove Einstein's theory of General Relativity and the existence of cosmic gravitational waves. While this discovery was extraordinary, the goal now is to combine multiple observational data sources to obtain a richer understanding of the phenomena.

FIGURE 5: SPINNING BLACK HOLES CREATE GRAVITATIONAL WAVES, RIPPLES IN THE FABRIC OF SPACE AND TIME. MACHINE LEARNING IS NOW ENHANCING OUR UNDERSTANDING OF THESE PHENOMENA.



Source: ***Simulating eXtreme Spacetimes (SXS) project and LIGO Caltech***

Challenge

The initial LIGO discoveries were successfully completed using classic data analytics. The processing pipeline used hundreds of CPUs where the bulk of the detection processing was done offline. Thus, the latency is far outside the range needed to activate other resources, such as the Large Synaptic Space Survey Telescope (LSST), which observes phenomena in the electromagnetic spectrum in time to "see" what LIGO can "hear".

Solution

Scientists are now using GPU-powered machine learning to make this computationally intensive approach possible. Using Convolutional Neural Networks, the National Center for Supercomputing Applications (NCSA) trained

its system to process gravitational wave data more than 5000 times faster than the original CPU-based waveform detection, making real-time analysis possible and putting them one step closer to understanding the universe's oldest secrets.

Impact

The ability to detect a richer mix of signals through machine learning will allow the LIGO researchers to better understand dark matter. Further, the ability to coordinate multiple sensors could have a profound impact on the understanding of transient phenomena, which are almost impossible to detect with one source. Advanced detectors such as LIGO and Virgo (another laser interferometer), working together with space and ground-based electromagnetic telescopes, will offer a greater opportunity to explore the Universe, thanks in part to machine learning.

APPLICATION OF MODEL MODULATION AT ITER

Background

The grand challenge of fusion energy offers the opportunity to provide clean, safe energy for millions of years. To this end, thirty-five nations are collaborating to build the $25 billion ITER (International Thermonuclear Experimental Reactor) Tokomak magnetic fusion facility in southern France.

Challenge

Fusion is highly sensitive; any disruption to conditions can cause reactions to stop suddenly. The challenge is to predict when a disruption will occur in order to prevent damage to ITER and to steer the reaction to continue producing power. Traditional simulation and previous ML approaches don't deliver accurate enough results. The best current models were only 85-percent accurate and incurred a high rate of false positives.

Solution

The team used the open-source TensorFlow ML framework and NVIDIA GPUs to develop a deep learning network, which now exceeds today's best traditional methods for accuracy. It scales to 200 Tesla K20 GPUs, and with faster GPUs is expected to deliver even higher accuracy: the goal is to reach 95 percent.

Impact

The vision is to operate ITER with neural networks, steering experiments in real-time to minimize damage and downtime. The benefit to Fusion Energy Science (FES) of machine-learning approaches to disruption prediction is the potential to significantly shorten the time needed to establish reliable operation in ITER.

Application of Model Approximation in Quantum Chemistry

Background

Developing a new drug costs upwards of $2.5 billion and takes 10 to15 years. Quantum chemistry (QC) simulations are used to accurately screen millions of potential drugs to identify the few most promising drug candidates.

Challenge

QC simulation is computationally expensive so researchers typically use approximations, which compromises accuracy. Using traditional techniques may take as long as five years to screen 10 million drug candidates.

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 in microseconds versus several minutes, with extremely high accuracy.

Impact

Machine Learning is enabling faster and more accurate screening at far lower cost. One interesting note about this example is that the accurate, first-principles numerical solver, Discrete Fourier Transform (DFT), was so slow that people had created faster hand-coded approximation solvers to be more productive, but these tools suffered from poor accuracy. Not only is the DNN — trained by labeled data generated by slow-but-accurate DFT — statistically as accurate as DFT, but it is also about six orders of magnitude faster. As a result, molecule screens looking for promising new drug candidates that would have taken years can now be accomplished in minutes.

CONCLUSIONS

Throughout the history of HPC, simulating nature and the laws of physics with numerical analysis models has been the standard for scientific research. However, for certain problems, the synthesis of simulation and machine learning can provide an alternative which can be faster, more accurate and less expensive to help scientists understand the behavior and characteristics of the physical world.

While this new approach is in its infancy, and is certain to be somewhat controversial as its impact is felt across HPC domains, early Machine Learning research projects indicate this approach can reduce computing resources and energy consumption by orders of magnitude compared to traditional simulation. Machine learning will almost certainly be used in the future to augment first principle physics models, which will still be used to create the massive datasets required to build neural networks and to keep pace with the advancement of the underlying science. Since both simulations and training can be performed on the same GPU-enabled hardware infrastructure, one avoids additional capital expenses as well as the need to transfer massive amounts of data from simulation to training systems.

It is becoming clear that the next big advances in HPC may not have to wait for exascale-class systems, but are being realized today using Machine Learning methodologies. In fact, the convergence of HPC and ML “Synthesis Models” could potentially redefine what an exascale system should even look like. We expect the adoption of ML in HPC to accelerate significantly in the next few years, given the hype around AI, the investment funding available from governments and industry, and the extremely efficient GPU hardware now available.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR

Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES

This paper was commissioned by NVIDIA, Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.