Technology Spotlight

Intelligent Simulation Exploits AI to Improve HPC Results

Sponsored by IBM

Steve Conway, Bob Sorensen, Alex Norton, and Earl Joseph
April 2019

EXECUTIVE SUMMARY

HPC-enabled simulation, widely recognized as the third branch of the scientific method, has contributed enormously to national and regional security, scientific and engineering innovation, and the quality of human life. Simulation is primarily responsible for expanding the global HPC market from $2 billion in 1990 to $28 billion in 2018, en route to a Hyperion Research-projected $38 billion in 2022.

One of the fastest-growing components of that forecast (14.9% CAGR) is high performance data analysis (HPDA) – using HPC systems for data-intensive simulation and analytics. Projected to grow even faster (26.3% CAGR) is the important HPDA subcategory of artificial intelligence (AI), primarily meaning the use of HPC systems for R&D at the forefront of AI. Hyperion Research forecasts that by 2022, HPDA (including AI) will contribute $4.7 billion to the HPC server market and more than $10 billion to the $38 billion overall HPC market for servers, storage, software and technical support.

AI is adding a fourth branch to the scientific method: *inferencing* to complement theory, experimentation and established simulation methods. An essential requirement for AI learning models (machine and deep learning aka cognitive computing) is the ability to *infer,* to make intelligent guesses about present or future circumstances based on incomplete evidence. AI inferencing is ultimately about computers that can infer, i.e., repeatedly guess at answers to home in on useful solutions much faster than humans can.

When inferencing is applied to simulation tasks, the result is ***intelligent simulation*** – the use of machine learning, deep learning or other AI methods to reduce the number of simulations needed while providing higher resolution and trustworthiness than would otherwise occur. Intelligent simulation creates the potential for reviewing much more data and exploring more of the problem space in a given time period, especially by predicting which data and exploratory runs are likely to be useful and omitting the rest. Intelligent simulation can be applied not only to scientific and engineering workloads, but also to analytics-driven business operations such as fraud detection, business intelligence, affinity marketing, ERP and sales planning.

This paper discusses the nature and beneficial uses of intelligent simulation, then looks at IBM's integrated software solutions designed to address challenging AI workloads. The world's two most powerful contemporary supercomputers, the Summit and Sierra systems, are both built on IBM technology and are expected to advance the state of the art in AI and HPC.

Note: this page is intentionally blank.

Rapid Growth of HPC Simulation and Analytics

The number of HPC systems sold annually, mainly for simulation, skyrocketed from under a thousand in 1990 to 90,980 in 2018, expanding the worldwide HPC market from about $2 billion to $28 billion, en route to our forecast $38 billion in 2022. HPC simulation began in government and academic research organizations, to tackle daunting problems in the "hard sciences": physics, chemistry, biology, astronomy/cosmology and geology. HPC soon spread to a broad spectrum of private sector firms, from large global enterprises to 25-person SMBs. Even within academia, the use of HPC simulation now extends to disciplines including cultural anthropology and archeology, historical linguistics and the social sciences.

The use of HPC systems primarily for integer-based, data-intensive computing, as opposed to floating point-based simulation, began in the intelligence/defense community in the 1960s, at the start of the supercomputer era, and spread to large investment banks in the financial services industry in the 1980s. Today, HPDA and AI methods are being added to the computing mix of traditional HPC users, most of whom pursue upstream R&D using dedicated HPC data centers. These methods are also motivating a growing number of commercial firms to integrate HPC systems into enterprise data centers to advance business operations, especially fraud detection and cyber security, business intelligence, affinity marketing, ERP and sales planning. Commercial firms are being driven to do this by competitive forces, especially the need to direct more complex questions at their data structures in near-real time.

Intelligent Simulation

Simulation remains by far the most popular problem-solving method associated with HPC, but simulation and HPDA-AI data analytics will increasingly join forces to tackle problems faster and more thoroughly than either approach could alone. Intelligent simulation is the favored term for this potent combination.

- **Analytics Helping Simulation.** Climate research historically has been one of the most daunting HPC simulation problems, especially with the expansion of ensemble models, but for the past decade researchers have been advancing analytics-based "climate knowledge discovery algorithms" to provide additional insight. The first IEEE workshop on this topic, held in 2008, was called "Data Mining for Climate Change and Impacts." (https://ieeexplore.ieee.org/document/4733959).
- **Simulation Helping Analytics.** The development of automated driving systems, on the other hand, is primarily an analytics problem, but experts say fully automated vehicles will need to be test-driven for several billion miles to establish consumer trust. Completing that many physical test-miles is impractical, so HPC simulation, already a mainstay at major automakers, will come to the rescue. In this case, a virtual vehicle that looks and behaves like an identical "digital twin" of its physical counterpart, is guided by an AI model through millions of test drives representing a comprehensive range of real-world situations. The output from these simulations continually refines the analytics-based AI model and vehicle design.

Intelligent simulation promises to create competitive advantages for HPC users, with the following benefits:

- *Increased accuracy*, by making it possible to review more data and explore more of the problem space in the given timeframe

- *Solution Identification.* Domain-specific cognitive models and knowledge graphs can curate huge volumes of information from disparate sources, allowing you to draw on a vast pool of expertise to quickly focus on the most promising lines of investigation and dismiss the less promising ones
- *Faster solutions*, by using intelligent algorithms to zero in quickly on the input data that is most valuable for insights and innovation
- *Improved cost-efficiency and TCO*, by boosting the productivity of the HPC system and leaving less of the data unanalyzed
- *Greater trustworthiness*, by analyzing more possibilities and increasing the transparency of machine learning/deep learning operations (e.g., the automated driving system example, above)

Intelligent simulation promises to create competitive advantages for HPC users who employ it and competitive disadvantages for those who don't.

IBM, HPC and AI

IBM has long been an HPC leader and underscored that leadership by capturing the top two spots on the November 2018 list of the world's Top500 supercomputers (www.top500.org). Both the Summit and Sierra supercomputers will support breakthrough scientific and engineering research projects under the U.S. Department of Energy's INCITE program, among other tasks. Summit, the world's number one supercomputer, located at Oak Ridge National Laboratory, will use HPC and HPC-enabled AI to tackle a long list of grand challenge problems, including:

- Combatting cancer
- Predicting fusion energy
- Deciphering high-energy physics data
- Identifying next-generation materials

Running AI in An HPC Environment

For organizations exploring the use of AI to augment HPC, by improving the fidelity of results and decreasing solution times for modeling and simulation, infrastructure is a key consideration. Organizations must have a way to prepare and curate massive volumes of data on the front end of the AI training process, from a variety of possible sources, such as the IoT and sensors. Looking more closely at the AI workflow itself, effort is required to build, optimize and validate open source frameworks for machine learning and deep learning, and the accuracy of results depends on intelligent and fast model optimization and training. These tasks are complex, and AI success depends on accelerated and optimized infrastructure. Computing capability and storage performance need to be tempered with software tools facilitating the creation and running of AI workflows alongside existing HPC workloads, in a distributed fashion, both on-premise and in a hybrid cloud.

IBM Spectrum LSF, a ubiquitous workload scheduler, has been available for more than 25 years. IBM reports that today IBM Spectrum LSF provides a number of advanced capabilities that enable organizations to create a flexible infrastructure to run machine learning and deep learning workloads alongside traditional HPC workloads. These include:

- Support for containerized workloads (Docker, Shifter, Singularity)
- Advanced support for NVIDIA GPU workloads
- Processor, core and GPU core affinity

- Connectors for running Apache Spark, Hadoop workloads
- Advanced parallel scheduling capabilities

Organizations that have standardized on specific AI frameworks can run AI/ML/DL training workloads through IBM Spectrum LSF while benefiting from fine-grained GPU access control and the parallel scheduling it provides. IBM demonstrated running Horovod and Distributed Tensorflow through IBM Spectrum LSF in the following videos:

- Tensorflow Deep Learning examples running in an IBM Spectrum LSF Suites v10.2 cluster https://www.youtube.com/watch?v=wxeiPBEItJ4
- Horovod Tensorflow Benchmarks running in an IBM Spectrum LSF Suites v10.2 cluster https://www.youtube.com/watch?v=5kBOVdK_h78

Introducing AI workloads into an existing HPC infrastructure can eliminate compute silos and drive more effective utilization of compute resources. Furthermore, IBM Spectrum LSF is designed to enable the creation of dynamic, hybrid-cloud environments which ebb and flow compute resources in the cloud as needed, irrespective of the workload type.

IBM PowerAI combines popular optimized deep learning frameworks, supporting libraries and development tools that according to IBM help to dramatically reduce time to AI environment readiness, while providing greater performance for deep learning and machine learning training. IBM reports that numerous clients are currently running IBM PowerAI workloads alongside traditional HPC modeling and simulation workloads through IBM Spectrum LSF.

AI model training is the most computationally intensive process in an overall AI workflow today. Distributed approaches to AI model training are therefore becoming commonplace to accelerate these tasks. However, a number of key steps in the overall AI workflow bear consideration. For example, data preparation can represent up to 80% of the work of data scientists[1] and hyperparameter optimization can also be a very time-consuming, necessary process for creating optimal AI models.

For a solution that treats the AI workflow with the data scientist in mind, IBM offers IBM Watson Machine Learning Accelerator (formerly IBM PowerAI Enterprise). According to IBM, it provides an end-to-end solution for the creation of a distributed deep learning environment, with capabilities to streamline the training, tuning and deployment of AI models. IBM indicates that it is also possible to run IBM Watson Machine Learning Accelerator instances on top of an existing IBM Spectrum LSF cluster.

Features of IBM Watson Machine Learning Accelerator

- **Elastic distributed training** dynamically assigns GPUs to models while training, speeding time to results. GPUs can be added and removed without needing to stop the training. This feature enables jobs to start and complete faster and leads to greater model accuracy.
- **Faster model development** with hyper-parameter search and optimization aims to improve accuracy with suggestion-based logic while the training is running.

[1] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#926554a6f637

- **Resource sharing** enables intelligent workload scheduling to maximize utilization, while providing seamless access to heterogeneous computing environments, including IBM Power Systems and x86 servers with GPU, on-premise or in a hybrid cloud.
- **Storage resource connectors** aim to simplify and maintain ongoing connections to multiple data sources.
- **Multitenancy** reduces wasted time, cost and administrative overhead and provides access to a larger cluster by supporting multiple data scientists, frameworks and applications on a common shared compute cluster.

Real-World Examples of IBM-Supported Intelligent Simulation

IBM is working with clients across a broad range of industries who are running AI modeling and inferencing on the same infrastructure as their traditional HPC workloads. More and more organizations with an HPC pedigree are trying to break down silos by using an HPC infrastructure to support their entire businesses, from upstream R&D to daily business operations. It's becoming more common to run AI workloads alongside existing HPC workloads, and to apply both simulation and advanced analytics to the same problem in order to deepen insights and innovation. IBM software solutions aim to enable organizations to compose resources based on the nature and priority of incoming workloads.

Following are just a few examples of intelligent simulation that the IBM solutions for AI and HPC are intended to address.

Electronic Design Automation (EDA)

A semiconductor manufacturing company is running wafer inspection models on the same infrastructure as their physical design work, not only reducing their overall costs but providing increased flexibility because, theoretically, if there were a major issue with wafer quality, the priority of wafer inspection workloads could be increased so that it occupied the majority of the compute resource.

Automotive Design

- Many vehicle and component designs evolve from previous designs. There are millions of potential changes that might improve a vehicle's aerodynamics, structural integrity, environmental friendliness, and passenger comfort. Exhaustively simulating each of these possibilities at every speed is impractical, hence some worthwhile optimization may often be missed. But we already have a significant body of knowledge about the existing design. Can this body of knowledge be used to train a model to evaluate vast numbers of potential changes and identify those that promise to provide benefits? This model would explore the design space comprehensively and eliminate dead ends, leaving a tractable number of promising possibilities to be simulated in the traditional way. A number of automotive companies are already investing in this area. For example, Ford spoke about this at the HPC User Forum in Dearborn, Michigan, in September 2018. They reported that for 2D models, this approach was more than 95% accurate and 300 times faster than traditional simulation alone.

Computational Chemistry

- Bayesian optimization is being used in a number of areas. For example, IBM and the UK's Hartree Centre have published a use case in chemical formulation, running on top of IBM Spectrum LSF. Small changes to the proportions of chemicals in, say, a lubricant in shampoo or toothpaste can make significant changes to its physical properties, some good, some bad. Testing all possible combinations in the wet lab in steps of 1 milliliter might require 500,000 experiments. Even doing this computationally would be extremely time-consuming, Applying

Bayesian optimization and deep learning allows users to quickly develop a model and predict how chemicals will mix in a fraction of the time, resulting in major cost savings.

Beyond intelligent simulation, there are other ways in which AI can be applied in HPC environments to drive efficiencies:

Cognitive Discovery

- Data provides the foundation of modern research. Data preparation for these modern approaches, including AI and deep learning, is not a simple task. Data scientists today spend a significant portion of their time in data preparation before training can even begin. Data may be obtained from numerous sources, including IoT sensor data, business data and legacy data. This data must be transformed into a common format suitable for analysis. With this understanding, IBM has demonstrated a tool known as IBM RXN that predicts the outcome of organic chemical reactions. Using a knowledge graph based on volumes of relevant, ingested data, this cognitive discovery method acts as an expert advisor to help predict results. This video provides more details on IBM RXN for Chemistry.
  https://www.youtube.com/watch?v=7DVu9ZKPUws&feature=youtu.be

Uncertainty Quantification

- Uncertainty quantification (UQ) is a method commonly applied to computational fluid dynamics (CFD) to increase the reliability of the simulation results by quantifying the impact of uncertain parameters in the model. Historically, UQ has relied on Monte Carlo analysis, which is computationally intensive, requiring a significant number of simulations to exercise a model with different sets of input parameters. IBM research is exploring how AI methods applied to UQ as part of a UQ "as-a-service" architecture could potentially generate a more optimal set of UQ parameters, with the goal of an overall shorter time to solution. More details regarding this effort can be found in the research paper *Uncertainty quantification-as-a-service:* https://dl.acm.org/citation.cfm?id=3291328

Scheduling Optimization

- Workload management has become a standard fixture today in HPC environments. Workload managers are used to scheduling workloads across increasingly complex, heterogeneous computing environments serving users with diverse backgrounds. Frequently, users of HPC environments ask for more resources than are actually required for their workload. This is to avoid failure of applications due to insufficient resources. However, this can lead to non-optimal utilization of costly resources. Taking EDA as an example, the vast majority of applications are single core, resulting in many jobs running on the same machine. Thus, memory contention is a significant consideration. Furthermore, users often do not have a strong understanding of how much memory a given job will require. Under-requesting memory will result in jobs dying, while over-requesting memory will dramatically lower overall utilization and throughput. IBM is evaluating machine learning methods with the IBM Spectrum LSF Predictor technical preview for predicting memory and runtime needs for workloads.
    - In addition, IBM customer MediaTek developed their own DL/ML model for job memory prediction and reported over 95% accuracy, with significant increases in utilization and throughput.
    - Accenture have also invested in this area: https://insidehpc.com/2018/11/video-accenture-engineering-compute-takes-hpc-enterprise/

FUTURE OUTLOOK

Recent Hyperion Research global surveys of noted AI experts in government, academia and industry confirmed that the worlds of HPC and AI are quickly converging. HPC is indispensable at the forefront of AI R&D for new uses ranging from precision medicine to automated driving systems, affinity marketing, business intelligence, cyber operations and the Internet of Things, among others. At the same time, established HPC users are pushing to add AI methods and tools to their simulations. Hyperion Research forecasts continued robust growth of the worldwide HPC market to $38 billion in 2022, with more than $10 billion of that total coming from HPDA (14.9% CAGR) and AI (26.3% CAGR) revenues as the worlds of HPC and large-scale commercial computing increasingly meld.

Buyers and users on both sides of the fence, HPC and HPDA-AI, tell us they want to move toward one "swim lane," meaning HPC resources that can efficiently support both HPC and advanced analytics workloads, especially the growing number of problems that promise to benefit from both simulation and analytics runs. Eventually, users want HPC systems to be able to integrate simulation and analytics results in near-real time. This is the promise held out by intelligent simulation.

The converged infrastructure for HPC and AI from IBM aims to provide a common, shared platform for running traditional HPC workloads coupled with advanced analytics (machine and deep learning). The IBM Summit and Sierra supercomputers, numbers one and two in the world today, promise to advance the HPC-HPDA-AI convergence by tackling grand challenge problems requiring breakthrough performance in both competencies, i.e., intelligent simulation. Hyperion Research believes that IBM is well positioned to accelerate the HPC-HPDA-AI convergence and to continue benefiting strongly from this projected growth of the converging HPC-HPDA-AI market.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user and vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com or www.hpcuserforum.com

7402497USEN-01