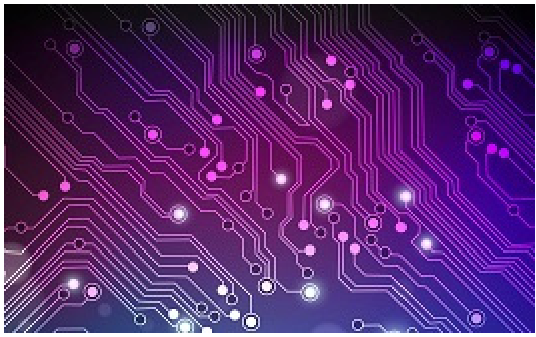




WHITE PAPER

Use Cases for Artificial Intelligence in High-Performance Computing

Published 2Q 2017

COMMISSIONED BY:

ADITYA KAUL
Research Director

CLINT WHEELOCK
Managing Director

SECTION 1

INTRODUCTION

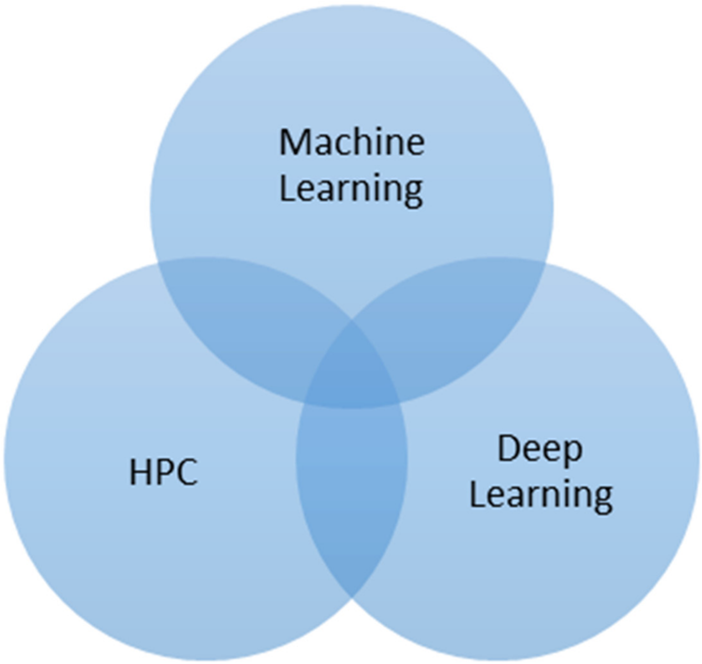
Artificial intelligence (AI) has shown major advances over the past few years, especially in the areas of image, speech, and text recognition. Apart from vision and language, AI techniques are also getting better at applications like algorithmic trading, anomaly detection, or patient data processing that feed on the volume, velocity, and variety of Big Data. The performance of AI algorithms is directly proportional to the size of the training data set. Deep learning, which uses multi-layered neural networks fashioned after the neurons in our brains, has taken advantage of large datasets, with its performance exceeding humans in the areas of image recognition and speech recognition.

The increasing digitization of society and the mass availability of data mean that, in theory, we can continue to drive the performance of these algorithms and grow it at an exponential rate. Within science, this is referred to as the "Fourth Paradigm," as described by Jim Gray in his 2009 book in which he predicted that the first three paradigms of empirical observation and experimentation, analytical or theoretical approaches, and computational science or simulation are beginning to give way to a data-driven approach. AI is one of the technologies driving this shift into the fourth paradigm, but with the AI and deep learning models seeing an exponential increase of data, the challenge lies in scaling computing resources with rapid data growth. This is where high-performance computing (HPC) can play a role as it helps AI algorithms scale performance with data. Standard techniques in HPC like low-latency and high-bandwidth interconnects, sparse matrix multiplication techniques, low-precision compute, and low-power footprint can all be applied to speeding up AI, expanding its opportunities.

Traditionally, HPC has been a staple technique for scientific modeling and computation and, therefore, it is no surprise that the coming together of AI and HPC is first happening in the scientific domain. While AI and scientific computing are distinct branches of computing, there is beginning to be a realization in the scientific community that AI, in particular deep learning, could be used to unravel and discover new knowledge. Where mathematical models fall short in representing physical phenomenon, data-driven AI approaches have been shown to be successful. Recent papers have shown that the accuracy of deep learning models exceed or come very close to traditional scientific modeling methods. With some scientific problems, such as neutron scattering and microscopy, producing terabytes (TB) of data within minutes, deep learning models can be trained to automatically classify streaming images as signal versus noise. Deep learning with its powerful image classification techniques can also be used for visual querying of simulation results.

At the heart of machine learning, deep learning, and HPC are some key commonalities. Under the hood, the software code used in HPC, which includes C++ and MPI, map well to deep learning and machine learning algorithms because they allow for better control, speedup, and improved performance through the manipulation of low-level system features, such as pointer manipulation, matrix multiplication, or memory management. HPC and AI methods like deep learning scale performance with large datasets, with model and data parallelism techniques being common across both the domains. The intersection of machine learning, deep learning, and HPC can be viewed as the sweet spot toward which the future development of AI will gravitate.

Figure 1.1 Intersection of Machine Learning, Deep Learning, and High-Performance Computing

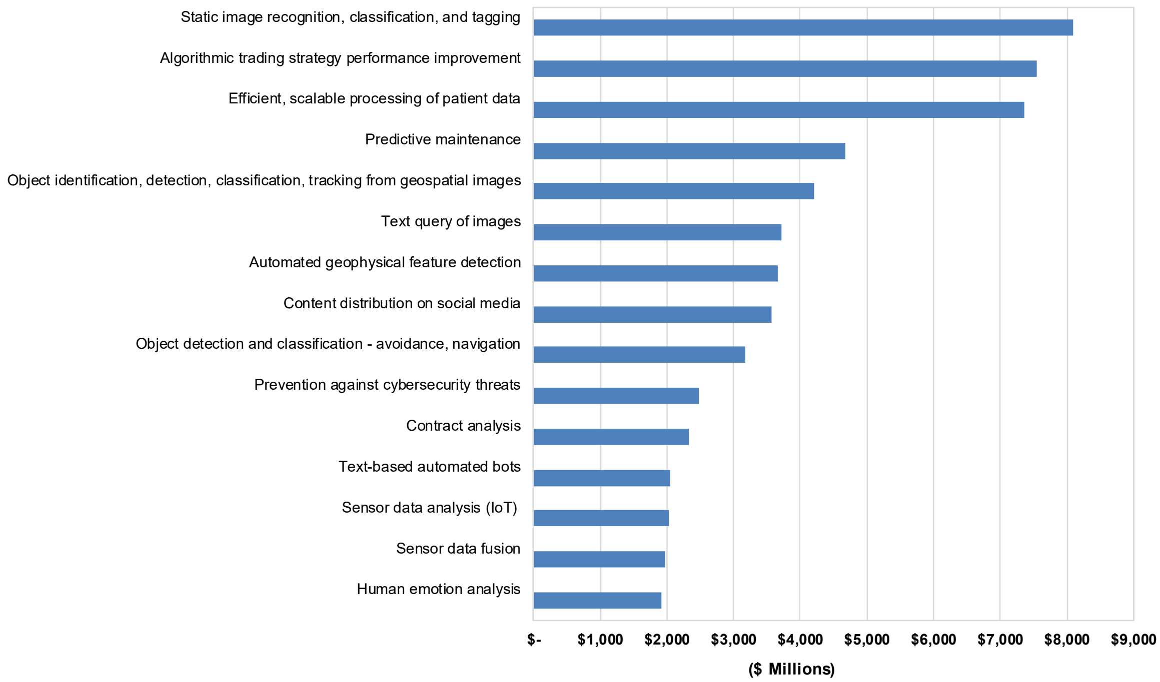


(Sources: Tractica)

As a result, there is not a more appropriate time to bring together two largely disparate areas: HPC and AI. Andrew Ng's keynote speech at the International Supercomputing Conference (ISC) 2016 could not have been timed better, where he made a case that for AI algorithms to raise the bar on performance and robustness of compute and data-centric workloads, there is an increasing need to leverage HPC, which provides one of the best and most familiar frameworks that has been around for decades. Ng was categorical in stating that HPC will be critical for accelerating advancements of AI in the future. Ng has also been a proponent of leading AI companies needing to invest in hiring HPC specialists apart from AI experts, something that Ng drove internally at Baidu until his recent departure.

Applications that lend themselves to Big Data and large workloads are ideally suited for using HPC techniques. Applications such as image, speech, or text recognition can see immediate speedup with HPC. Tractica's own analysis shows that the top 15 applications for AI include a mix of Big Data, vision, and language applications, all of which could see major benefits from HPC.

Chart 1.1 Cumulative Artificial Intelligence Revenue, Top 15 Use Cases, World Markets: 2016-2025



(Source: Tractica)

Also, research institutions and businesses that have been using HPC to run simulations can start to look into machine learning and deep learning techniques to enhance or replace parts of the HPC process. For example, using machine learning or deep learning for front-end or pre-processing steps for data simulations is one area that is already gaining traction. While many HPC platforms already use a graphics processing unit (GPU), which is the traditional workhorse for cutting-edge deep learning and machine learning, porting simulation data into deep learning frameworks like Caffe or TensorFlow is not straightforward. AI workloads also need to be split into training and inference, with separate GPU clusters optimized and scaled for each task, although next-generation hardware can put both training and inference capabilities on the same platform. Despite some of the challenges in the implementation of HPC and AI, it is clear that the future points toward these two rapidly advancing areas of computing coming together. For one, it is becoming easier to port traditional AI toolkits into HPC environments, despite traditional AI toolkits not having been built with massively parallel computing in mind. The recent showcase of the Swiss National Supercomputing Centre (CSCS) running Microsoft's Cognitive Toolkit on a Cray XC50 supercomputer, cutting training time from weeks or months to a few hours or minutes, is an encouraging development.

With the intersection of AI and HPC becoming clearer, and with the size of AI workloads becoming similar to those that have been traditionally the domain of HPC, it is useful to take a deeper dive into some of the real-world applications that are likely to benefit from this intersection. This white paper covers four specific use cases in which HPC and AI can be used together. The use cases include weather and climate modeling (scientific research), precision medicine for cancer (healthcare), fraud detection (financial), and real-time threat analysis (cybersecurity). The goal is to highlight the diversity of applications and industries that are using AI with HPC, and the unique characteristics of each use case in terms of the type and scale of data workloads on offer. The analysis aims to provide a glimpse of how the marriage of HPC and AI is already taking place, and the opportunities that this is likely to unleash for both scientific research and businesses.

SECTION 2

USE CASES

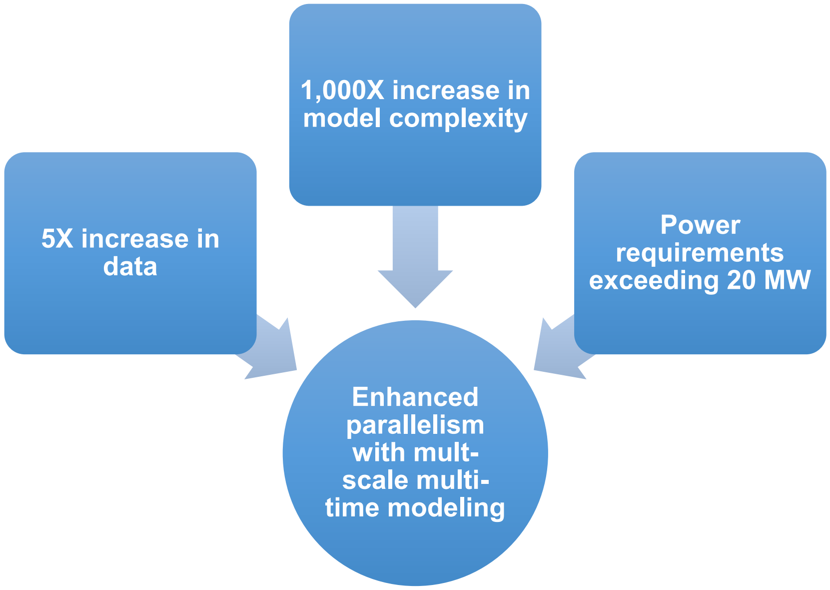
2.1 WEATHER AND CLIMATE MODELING

Climate modeling and weather prediction has been a key adopter of HPC systems. Weather prediction systems take in data from multiple sources from land, air and ocean sensors, aircraft, radar, and satellites. Climate models typically rely on historical observations that can go back by more than a hundred years. In both instances, the datasets are very large. For example, the European Centre for Medium-Range Weather Forecasts (ECMWF) uses approximately 40 million observations daily in their models. For climate modeling, a single 30-year run from a 25 km resolution model produces on the order of 10 TB of multivariate data. For numerical weather prediction (NWP) or climate modeling, the data is then fitted to fill a three-dimensional grid over which multiple simulations are run over time. With constantly changing weather patterns and a warming climate, weather forecasters need to have improved prediction capabilities that extend across time and that provide higher resolutions going down to a few kilometers. Supercomputers have played a vital role due to their massive compute resources. Extreme weather conditions like tornadoes and cyclones can lead to billions of dollars' worth of economic damage and extensive loss of life. Improved weather prediction capabilities and the ability to truly understand how climate patterns are changing and what impact they will have on human life is critical for the future effectiveness of global weather centers. HPC has provided the necessary muscle to bring us this far, but it might be losing steam if it does not adapt going forward.

The weather prediction community is eagerly awaiting the next jump in supercomputing power, exascale computing (a billion billion calculations per second), but simply relying on processors with more brawn is not enough. While there is a 5X increase in data that is expected to happen by 2020, a 1,000X increase in the model complexity is expected, which can only be dealt with by increasing parallelism and improving scalability. At the same time, power requirements need to be in check, with exascale architectures expected to exceed 20 MW, which becomes cost prohibitive.

The U.K. Met Office has recently released a report for its own exascale computing requirements. The requirements include alternative architectures to the central processing unit (CPU), including a greater usage of GPU, 64-bit ARM, field programmable gate arrays (FPGAs), and the D-Wave Quantum computer. The report states that the main performance gains of the future are not just faster processors, which is a factor of Moore's law, but that of parallelism, which is dependent on multi-scale and multi-time-scale model coupling, programming models, input/output (I/O), and workflows. The report also refers to the use of machine learning techniques that could help improve parameterization development, and calls for crossovers and mergers of other fields like AI with HPC to enhance parallelism.

Figure 2.1 Challenges for High-Performance Computing and Artificial Intelligence in Weather and Climate Monitoring



(Sources: European Centre for Medium Range Weather Forecasts, U.K. Met Office, Tractica)

Most weather centers like the U.K. Met Office extensively use CPU-only systems, which makes the codes difficult to extend across to GPUs or other hardware architectures. While this can be a challenge, this should not stop national weather centers from considering the use of GPUs for improving the performance of their weather models. MeteoSwiss, the Swiss meteorological office, is one of the first weather centers to port all of its code onto GPUs and has successfully run machine learning and deep learning techniques. It took approximately 2 years to port its Fortran code to GPUs, but the results more than make up for the effort involved, as MeteoSwiss has seen a 40X performance boost and a 3X power consumption reduction over its CPU-only hardware. Upgrading to a GPU-based cluster has also allowed for finer 1 kilometer (km) resolution and a forecast that can be updated every 3 hours.

With weather centers like MeteoSwiss switching to a GPU-based HPC environment, they can also start to take advantage of deep learning and AI techniques to enhance the performance of their existing climate and weather prediction models. For example, AI can be used to perform weather pattern detection, such as cyclonic activity or other extreme weather events. The U.S. National Energy Research Computing Center (NERSC) has used convolutional neural networks (CNNs) to classify threatening climate events like cyclones. This work was performed on a CPU-only Cray XC30 supercomputer, with both the training and inference run on the same platform, although there was some effort involved in adapting the CNN algorithm to the climate data. The main goal for NERSC was to have a model learn the characteristics of a cyclone and classify it, an area where human decision-making variance is an issue. With the algorithm having between 80% and 90% accuracy in identifying extreme weather events, this is only the start and shows that AI techniques can be used for classification and identification of more complex weather systems and events.

A number of other applications of AI in weather applications include:

- Nowcasting using classifiers on radar and observations, which allows for very short-range forecasts for rainfall, snowfall, or severe weather events
- Using deep learning-based climate prediction and pattern recognition models trained on simulation data to predict longer-term extreme weather events based on various carbon emission scenarios
- Optimized observation selection allowing for climate models to use a smaller and more selective set of data, allowing for improved speed and performance
- Using AI techniques to find alternative approaches to parameterization where small-scale weather processes, such as clouds, are simplified using certain parameters
- Infilling or smoothing weather or climate model outputs

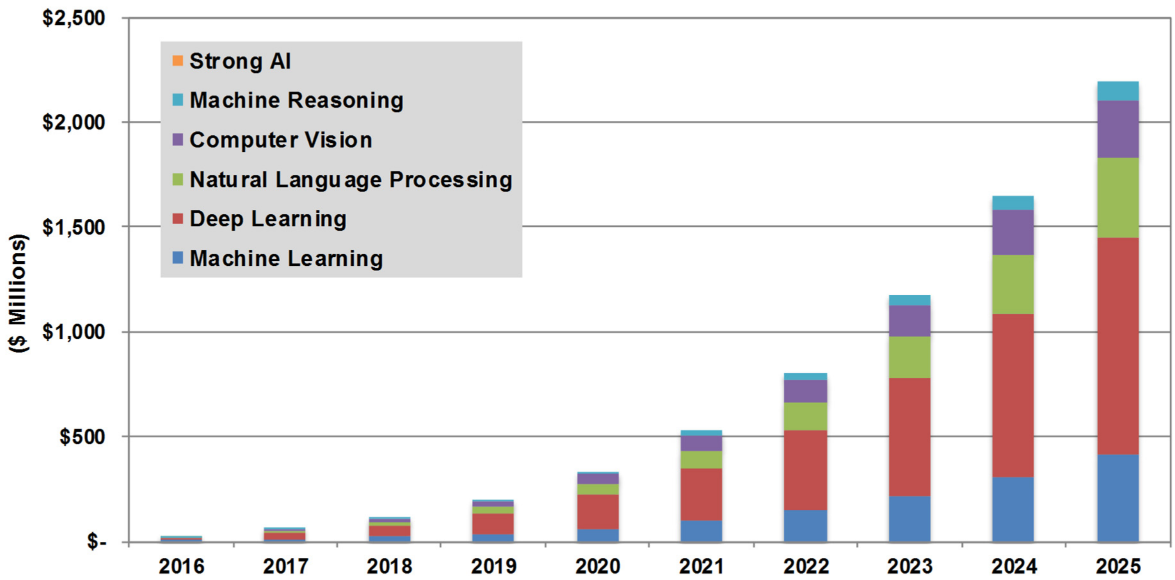
2.2 PRECISION MEDICINE FOR CANCER

HPC is a vital part of cancer research today, but it is at a crossroads with requirements shifting from simply increasing compute capability in flops to integrating multiple databases, improving memory bandwidth, using heterogeneous hardware, and improving software efficiency. The advent of rapid and relatively inexpensive genome sequencing has provided scientists and caregivers the ability to study the links between genes and cancer at the level of individual patients, and to use massive genomics and other biological databases to better understand patient genotypes and their implications for disease and treatment. Cancer has been called a disease of the genome, highlighting the important role that individual genetic variation plays in cancer etiology and treatment.

Precision medicine is being used as a tool to customize cancer treatment for individuals, based on their genetic makeup and the type of cancer that is affecting them. The Cancer Moonshot announced earlier in 2016 by the Obama administration was launched expressly to accelerate cancer research, and is expected to rely heavily on the use of HPC to solve some of the complex problems inherent to cancer prevention, detection, and treatment. Part of the Cancer Moonshot is the Cancer Distributed Learning Environment (CANDLE) project, which is a framework that will exploit HPC, machine learning, and data analytics to advance precision medicine for treating cancer.

Genome databases that are hundreds of TBs or bigger is just the beginning of the Big Data problem in precision medicine. The 31,000 base pairs in each genome that produce a unique individual in the human population, who constantly interacts with a few thousand environmental variables – and could suffer from one (or more) of the possible ~30,000 diseases and symptoms. Better predictive understanding of the genotype-phenotype mapping to treat every individual is bigger than the biggest deep learning problem to date. Scientists are looking to leverage HPC environments to overcome the following limitations: (1) State-of-the-art deep learning models are based on a simple layered network topology, i.e., highly connected layers, without intra-layer connections; (2) the networks are manually configured to achieve optimal results, and (3) the implementation of neural networks with increasing model complexity is expensive in both cost and power.

Chart 2.1 Healthcare Artificial Intelligence Revenue by Technology, World Markets: 2016-2025



(Source: Tractica)

Current HPC techniques allow researchers to simulate protein interactions at the molecular level to understand how cancer develops. Deep learning and machine learning techniques can be used to complement HPC, to allow these simulations to be rolled forward to reveal a bigger picture, revealing larger-scale interactions between the system and the surrounding environment. AI is essentially used in combination with HPC to provide more efficient simulations, and improve productivity by 10X every year, which is one of the goals of the Cancer Moonshot initiative. The CANDLE framework is also being used to study the deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) signatures of common cancer types to help with precision cancer treatment, as well as using deep learning and large-scale data analytics to automate information extraction and analysis of millions of patient records.

Natural language processing (NLP) techniques are used to regularize the data, while machine learning is used to extract useful information from the data. The automation of the analysis of patient records will replace the human effort required in cancer surveillance where biomarkers of cancer progression and outcomes are extracted manually from clinical reports. The National Cancer Institute (NCI) is also providing more than 20 types of data sources that feed into a neural network model, which then establishes a relationship between drug dosages and patient response. The data sources include information on genetic sequence, gene expression profiles, proteomics, and metabolomics. The NCI also has a large historical database of medical images, for which the neural network can build a relationship to molecular profiles data of cancer, which go back 10 to 15 years. The goal is to predict a molecular profile using medical image data, upon which specific drugs can be prescribed to produce a positive patient outcome.

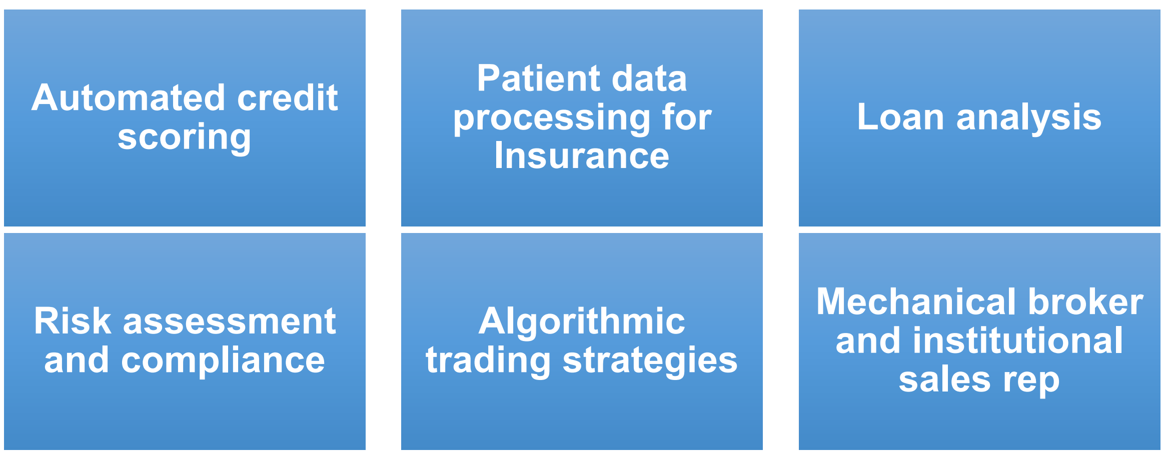
2.3 Fraud Detection in Financial Services

Among enterprise users of HPC, financial services is a key industry with applications in fraud detection, risk analysis, econometric modeling, algorithmic back-testing, high-frequency trading, derivatives pricing, or providing front-office real-time trading analytics. Data sets in financial services run into multiple petabytes, often including real-time statistics, such as economic indicators and news feeds, with execution time generally in less than a millisecond. Data I/O is one of the key bottlenecks due to accessing datasets that could be as large as 50 TB taking a few days to load. HPC is able to provide high-bandwidth, shared interconnects, and fast workload options, which can be optimized for multi-TB-scale problems, as seen in finance.

Monte Carlo simulations are typically used in financial markets to estimate risk where HPC systems can bring their massively parallel nodes and shared data interconnects to speed up computation, as the problem can be split into multiple pieces and the simulation run simultaneously. Graph technologies are also used within the financial industry in conjunction with HPC to create ontologies and relationships between regulations and policies, with transactions scanned against these deep-layered ontologies to ensure regulatory compliance.

Another area where HPC has been used extensively is in performing real-time fraud detection, again using graph analytics to create relationships between transactions and fraudulent activity. PayPal, for example, has been using HPC for real-time fraud detection to operate at Big Data scales, with 192 million users and 13 million financial transactions per day, amounting to more than $700 million in payment volumes per day.

Figure 2.2 Finance and Investment Artificial Intelligence Use Cases



(Source: Tractica)

While finance is considered to be one of the largest commercial segments for HPC, the gradual adoption of machine learning and deep learning techniques in areas like algorithmic trading, credit scoring, loan analysis, mechanical brokerage, risk assessment, and compliance is occurring. Companies like Sentient Technologies have been using deep learning techniques and evolutionary algorithms to find the best trading strategies. However, Sentient is using a network of distributed computers, rather than an HPC platform to run the software. While Sentient claims to be able to extract performance from a bunch of idle computers, running the same problem over millions of cores, the bottleneck is likely to end up being the interconnect, with the computers relying on consumer-grade Internet connections, rather than a high-bandwidth HPC interconnect.

Techniques like Monte Carlo or Graph Analysis can be combined with deep learning neural networks to increase the accuracy of prediction. For example, a Monte Carlo Tree Search (MCTS) can be used to generate additional data, which can then be fed into a neural network for training, with a larger training set leading to better predictions. Deep learning and machine learning techniques are also powerful at identifying correlations between unrelated data points, which could include economic data, geopolitical events, etc. In literature, deep learning and machine learning techniques have also been shown to perform a text-based analysis of news and classify events into buy or sell decisions, or predict stock returns. In combination with HPC, deep learning techniques can find hidden value in data, transactions, and customer activity; in the financial case, this is live time-series, multi-modal data. This data can be classified and used to make predictions on stock movements, pricing, volatility, or fraudulent activity.

2.4 Real-Time Cyber Threat Analysis

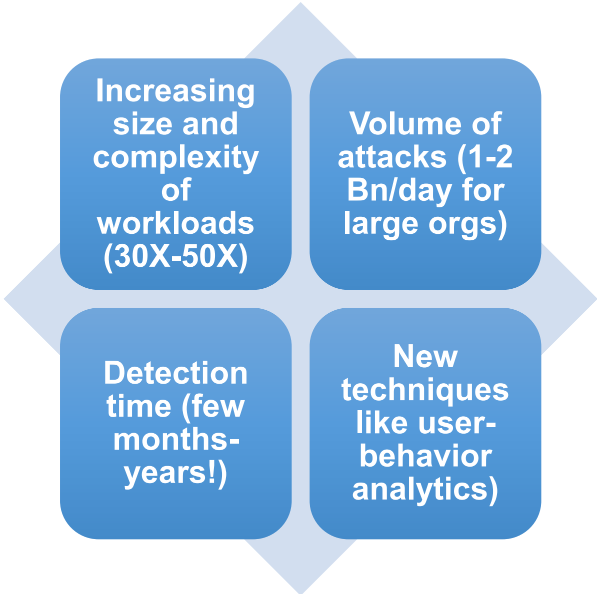
Organizations are experiencing a rapid increase in the volume and complexity of cyber threats and attacks. Traditional solutions in cybersecurity are unable to cope with the emerging nature of threats like advanced persistent threats (APTs), botnets, and zero day vulnerabilities. Signature-based malware detection, which drives most anti-virus solutions today, compares the signature of an unidentified code to known malwares. For new, unidentified malwares, manual intervention is needed for which malware databases can take weeks or months to update. Heuristic techniques that study the behavior of malware at runtime have limited predictive capabilities, while sandbox techniques that run malware in a virtual runtime environment fall short in providing real-time threat analysis.

The success of deep learning in image classification and speech recognition has been extended to the cybersecurity space where algorithms learn about threats in real time and are especially good at detecting first-time malware, or changing attack surfaces without the need for laborious feature engineering to take place. Graph analysis has also shown a lot of promise and is well suited to represent cybersecurity data, which can be mapped by nodes, and is especially good at detecting botnet threats, port scans, or lateral movement in networks. However, graph analytics is highly compute-intensive and incapable of doing real-time analytics on a CPU cluster when running in large organizations, which can see as many as 1 to 2 billion events per day.

Most large institutions have cybersecurity solutions deployed, but in most cases, they fall short and are underprepared for the rapidly changing threat scenarios. There are varying degrees of preparedness for large companies, with financial services firms being the best equipped, whereas public utilities and hospitals are the worst equipped. For example, breach detection within financial services can take an average of 3 to 4 months, while in retail, it can take 7 months, but in public institutions it can take 2 years!

Cybersecurity solution vendors that use machine learning and deep learning to process threats expect to see a rapid increase in the number of malicious software samples being used to train algorithms. Today's workloads involve a few million samples of malicious code, but the samples are expected to increase by 30X to 50X in the next 5 years. At that scale, using an enterprise-grade server would fall short. Using GPU-based HPC solutions will need to be adopted. Vendors are also adopting new approaches like studying user behavior analytics, server logs, director entries, and VPN logs, all of which can be fed into AI algorithms to spot suspicious patterns.

Figure 2.3 Challenges for High-Performance Computing and Artificial Intelligence in Cybersecurity



(Source: Tractica)

By combining deep learning, machine learning, and graph analytics with HPC, companies can avail themselves of the best of both worlds, taking advantage of high-bandwidth compute and data analytics, along with state-of-the-art neural networks that provide top performance and high accuracy for threat analysis and detection. Large companies and cybersecurity vendors can reduce false positives, detect threats faster, and improve the efficiency of cyber threat analysts.

2.5 Conclusions

The intersection of AI and HPC is the result of a number of trends that are coming together. On one hand, AI techniques like deep learning are starting to hit performance bottlenecks as data workloads grow exponentially. Scientific computation, which has been one of the largest application areas for HPC, is seeing an increasing use of data-driven deep learning methods to improve and enhance mathematical, simulation-based models, as we enter the fourth paradigm of scientific research. Both AI and HPC look very similar at the code level, with HPC bringing a number of standard techniques like high-bandwidth interconnects, low-precision compute, sparse matrix multiplication, and an optimized power footprint. HPC is built from the ground up to scale with data and compute, and for AI algorithms to raise the bar on performance of compute and data-centric workloads; there is an increasing need to leverage the familiar and well-honed methods of HPC.

The coming together of HPC and AI is a wakeup call for companies and institutions familiar with HPC and those that are not. Clearly, AI will transform the technology landscape and touch almost every industry over the next 10 years. Within scientific computation, which extensively uses HPC, there is a need to focus on approaches and models that are naturally inclined toward deep learning and machine learning. Companies that come at it from a pure AI standpoint and are beginning to hit performance bottlenecks as their workloads reach HPC levels need to reconsider their computation strategy and embrace HPC techniques.

This transition is still in the early stages, as AI and HPC come together; therefore, it is important to pay careful consideration to vendors that can provide a comprehensive set of tools and solutions across software, compute, storage, and networking. Access to validated deep learning and machine learning libraries and toolkits is very important, as is a vendor's industry experience in helping customers move their AI workloads to HPC or transitioning HPC customers toward AI. While the vendor landscape is still maturing around AI and HPC, with vendors bringing on board a range of capabilities, a handful of vendors today can provide the necessary expertise and solutions to take advantage of this market opportunity. Cray Inc. is one such vendor that provides both a comprehensive toolkit around HPC, but also has real-world experience helping customers both on the HPC and AI ends of the market to bridge the divide.

SECTION 3

TABLE OF CONTENTS

SECTION 4

TABLE OF CHARTS AND FIGURES

SECTION 5

SCOPE OF STUDY

This white paper provides an overview of how the two areas of HPC and AI are coming together. The paper highlights four different use cases and the industries that are likely to see major benefits by using AI and HPC techniques in tandem. The four use cases include weather and climate modeling (scientific research), precision medicine for cancer (healthcare), fraud detection (financial), and real-time threat analysis (cybersecurity).

SOURCES AND METHODOLOGY

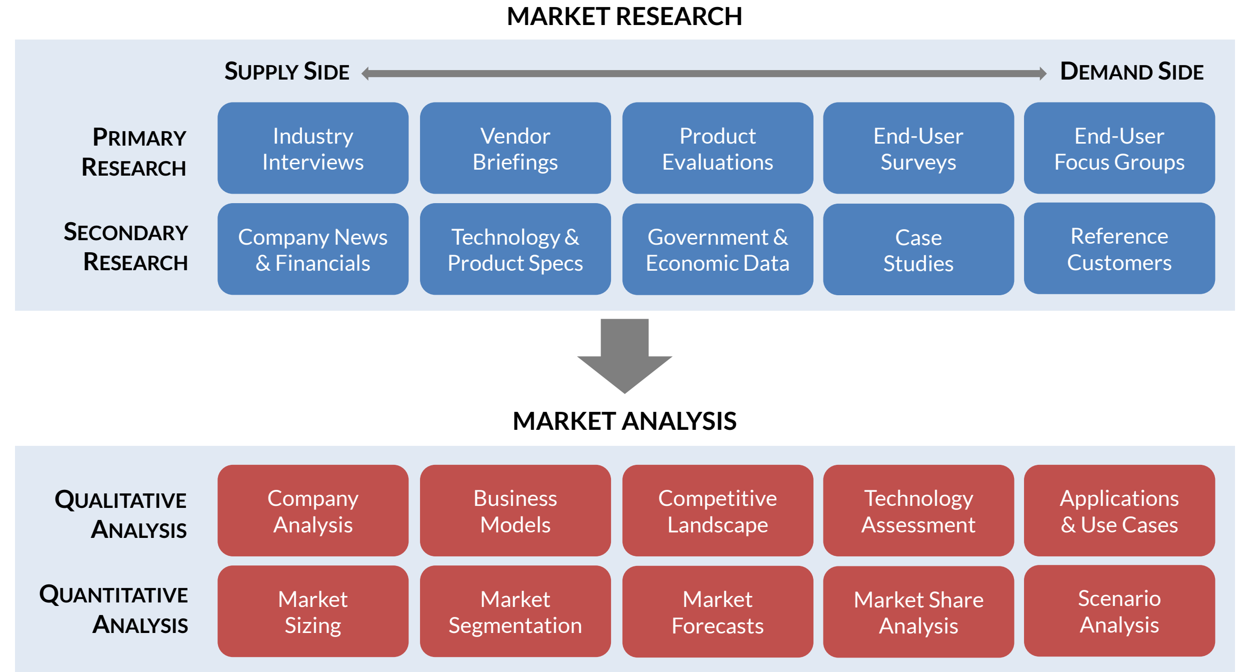
Tractica is an independent market research firm that provides industry participants and stakeholders with an objective, unbiased view of market dynamics and business opportunities within its coverage areas. The firm's industry analysts are dedicated to presenting clear and actionable analysis to support business planning initiatives and go-to-market strategies, utilizing rigorous market research methodologies and without regard for technology hype or special interests including Tractica's own client relationships. Within its market analysis, Tractica strives to offer conclusions and recommendations that reflect the most likely path of industry development, even when those views may be contrarian.

The basis of Tractica's analysis is primary research collected from a variety of sources including industry interviews, vendor briefings, product demonstrations, and quantitative and qualitative market research focused on consumer and business end-users. Industry analysts conduct interviews with representative groups of executives, technology practitioners, sales and marketing professionals, industry association personnel, government representatives, investors, consultants, and other industry stakeholders. Analysts are diligent in pursuing interviews with representatives from every part of the value chain in an effort to gain a comprehensive view of current market activity and future plans. Within the firm's surveys and focus groups, respondent samples are carefully selected to ensure that they provide the most accurate possible view of demand dynamics within consumer and business markets, utilizing balanced and representative samples where appropriate and careful screening and qualification criteria in cases where the research topic requires a more targeted group of respondents.

Tractica's primary research is supplemented by the review and analysis of all secondary information available on the topic being studied, including company news and financial information, technology specifications, product attributes, government and economic data, industry reports and databases from third-party sources, case studies, and reference customers. As applicable, all secondary research sources are appropriately cited within the firm's publications.

All of Tractica's research reports and other publications are carefully reviewed and scrutinized by the firm's senior management team in an effort to ensure that research methodology is sound, all information provided is accurate, analyst assumptions are carefully documented, and conclusions are well-supported by facts. Tractica is highly responsive to feedback from industry participants and, in the event errors in the firm's research are identified and verified, such errors are corrected promptly.

Chart 5.1 Tractica Research Methodology



(Source: Tractica)

NOTES

CAGR refers to compound annual growth rate, using the formula:

CAGR = (End Year Value ÷ Start Year Value)$^{(1/\text{steps})}$ – 1.

CAGRs presented in the tables are for the entire timeframe in the title. Where data for fewer years are given, the CAGR is for the range presented. Where relevant, CAGRs for shorter timeframes may be given as well.

Figures are based on the best estimates available at the time of calculation. Annual revenues, shipments, and sales are based on end-of-year figures unless otherwise noted. All values are expressed in year 2017 U.S. dollars unless otherwise noted. Percentages may not add up to 100 due to rounding.

Published 2Q 2017

1111 Pearl Street, Suite 201
Boulder, CO 80302 USA
Tel: +1.303.248.3000
Email: info@tractica.com
www.tractica.com