OrionX.net

The OrionX Constellation™ Reports
Artificial Intelligence

Evolution

Is Artificial Intelligence the Future of High Performance Computing?

Stephen Perrenod, Peter ffoulkes, Shahin Khan

The OrionX Position

The intersection of HPC and AI has created a vibrant new market: "High Performance Artificial Intelligence" (HPAI).

After decades of slow progress, HPC has given AI the sharp boost it needed to be taken seriously. Enabled by supercomputing technologies, HPC techniques such as Deep Learning are transforming AI to create a new breed of "thinking machines".

HPAI combines HPC (numerically intensive statistical analysis and optimization) with traditional AI (search algorithms and expert systems) to profoundly impact the IT industry and customer investment priorities, to influence every aspect of human life, and to pose its own grand challenges.

HPAI: High Performance Computing meets Artificial Intelligence

The world of Artificial Intelligence (AI) and quasi-sentient computer systems has long been the domain of science fiction. Decades after it first gained prominence and proceeded to underwhelm, AI has received the boost it needed to become a realistic and attainable goal. The necessary ingredients:

- Big data, generated by digitized processes, sensors, and instruments
- Massive computational power, often in the form of cloud computing, and
- Economically attractive use cases

are coming together to create a new breed of "Thinking Machines" that can automate complex tasks and decision processes, augmenting or replacing mechanical and electrical machines and people.

The intersection of HPC and AI is showing that cognition can be computable in a practical way (see, for example, this 1978 paper titled "Computability and Cognition"). It represents a blend of logic processing with numerically intensive computation. It is an area of intense activity in academic, commercial, industrial, and government settings.

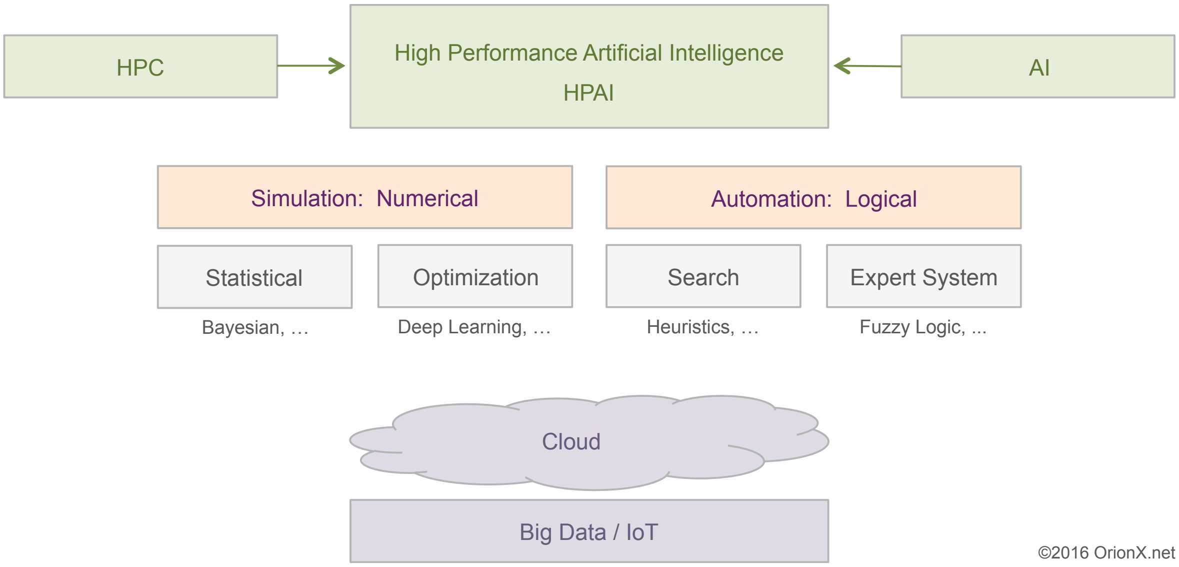
OrionX Constellation™ reports cover 5 Es: industry milestones (Events), progress in a technology segment (Evolution), vendors and forces shaping the market (Environment), customer decision criteria (Evaluation), and how vendors in a segment score (Excellence). The OrionX methodology considers market presence and trends, customer needs and readiness, and product capabilities and roadmap.

HPAI combines HPC (numerically intensive statistical analysis and optimization) with traditional AI (search algorithms and expert systems) to profoundly impact the IT industry and customer investment priorities, to influence every aspect of human life, and to pose its own grand challenges.

Just as with humans, mastery of tasks is a time consuming and iterative process for computers. HPC expertise are required to make this tractable:

- Advanced mathematical algorithms,
- Highly scalable system architecture, and
- Software and hardware optimization techniques

The intersection of HPC and AI represents a blend of traditional logic processing with numerically intensive computation. It is an area of intense activity in academic, commercial, industrial, and government settings. OrionX refers to this rapidly evolving and growing market as HPAI.



©2016 OrionX.net

HPAI Tecniques

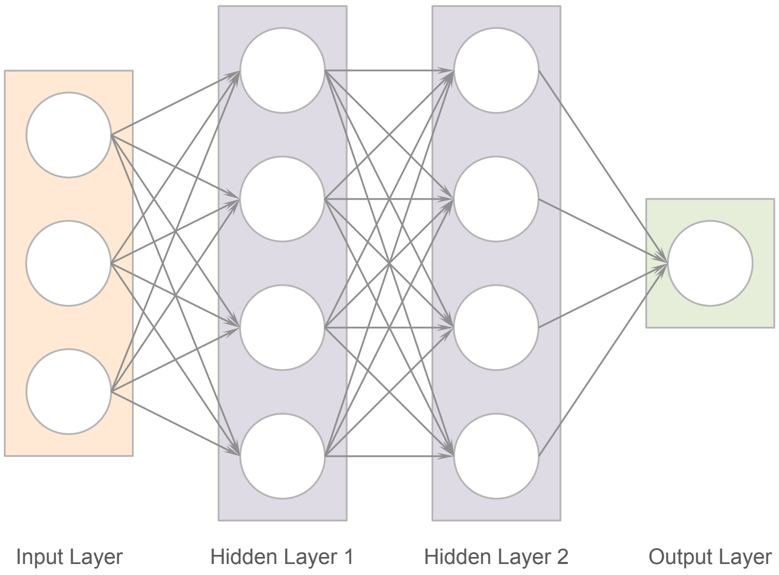
From simple origins in techniques such as regression analysis, ever more complex pattern matching techniques have formed the foundation of machine learning which provides the basis for *categorization*, *cognition*, *comprehension* and *conclusion* that can inform a decision or action.

Machine Learning

Machine learning has evolved from the study of pattern recognition and computational learning. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. Deep learning extends early machine learning systems through the use of neural networks – high performance systems designed to mimic the functions of the mammalian brain – which unlike conventional systems gain capability through training rather than explicitly programmed logic (as is the case for expert systems).

Neural Networks

Multi-layer neural networks process numerous input data sources through multiple ‘hidden layers’ of decision nodes that progressively filter and refine the results passed to each successive layer, ultimately providing the conclusion delivered by the final output layer.



The complexity of the neural network in terms of its number of nodes and layers clearly affects the complexity of real world problems that can be addressed, but the accuracy of the delivered result is the paramount consideration.

Deep Learning

A subset of machine learning, computer-based deep learning systems operate an iterative process, which accepts the inputs and delivers an output that is assessed for accuracy. A weighting value can then be applied or adjusted for any individual node in any given hidden layer in the neural network, the process rerun, passed through the network, assessed for changes and readjustments made iteratively to optimize the accuracy of the result. A 'back propagation' process is used to highlight any problems in the weighting applied to earlier layers in the decision matrix and to adjust and guide further iterations of the training regime until acceptable results are derived. This can clearly be a complex and time consuming exercise.

Building a sufficiently complex technology framework that is capable of being trained, as opposed to simply programmed, is the first critical step, followed by training with sufficient data sets to enable 'supervised machine learning' in a controlled environment. This involves both a training set of examples that deliver the desired outcomes combined with a larger test set of data to assess the accuracy and effectiveness of the training and application to real world situations.

More important than the infrastructure is the software toolkit that enables a deep learning or artificial intelligence system to be developed, trained and deployed in a fast and cost effective manner if deep learning is to move beyond research and non-critical hyperscale environments to widely deployed commercial enterprise applications.

Beyond primary processor capabilities, math accelerators are now increasingly critical for deep convolutional neural networks. They accelerate backpropagation functions that optimize learned weights, and can significantly reduce training times

Technology Drivers of Deep Learning

Deep learning has progressed rapidly during the past decade due to:

Big Data

We are increasingly deluged with data, a trend that will only accelerate. The potential for machines to filter, categorize, and recognize patterns is a huge advantage in most areas of human activity. After all, pattern recognition is largely what we do to support our decision making processes in daily life.

An increased availability of large data sets for training and deployment has also driven the need for deeper nets in order to address a greater number of parameters of interest. The simultaneous rapid growth in Internet of Things (IoT) technologies accelerates this process.

Deeper nets

Deep neural nets (DNN) have many layers, and often possess higher-order architecture (width) within a given layer.

Clever training

It was discovered that a large dose of unsupervised learning in the earlier stages of training allowed for the net to do its own automated, lower level, feature recognition and extraction, and pass those features on to the next stage for higher level feature recognition.

HPC systems

Clustered systems, enhanced with accelerator technology, have become essential to training large deep nets.

Deep learning is, at its heart, a computational and data intensive problem as we often associate with high performance computing (HPC). While the core principles for machine learning have been understood for decades, the technology and available datasets to implement it cost-effectively have not been available, limiting its usefulness in practical applications.

The Core Technologies of Deep Learning

The compute-intensive and data-intensive nature of deep learning has significant overlaps with the needs of the high performance computing market. Therefore,the TOP500 list provides a good proxy of current market dynamics and trends.

Processing and system architecture in the HPAI market is the subject of active innovation globally. Major requirements are high compute performance, high memory bandwidth, low power consumption, and excellent short arithmetic performance. The requisite computational resources are clusters whose nodes are populated with a sufficient number of accelerators. These provide the needed performance while keeping power consumption low.

Hardware

From the central computation perspective, today’s multicore processor architectures dominate the TOP500 with 91% based on Intel processors. However, further developments may include alternative CPU architectures such as OpenPOWER and ARM. In addition System on a Chip approaches that combine general purpose processors with technologies such as field programmable gate arrays (FPGAs), digital signal processors (DSPs), and special purpose application-specific integrated circuit (ASIC) designs are playing an increasing role in deep learning applications.

Beyond primary processor capabilities, math accelerators are now increasingly critical to the requirements of deep convolutional neural networks. They accelerate backpropagation functions that optimize learned weights, and can significantly reduce training times.

Nvidia GPUs are the most popular acceleration technology in deep learning today while other accelerator technologies are also popular. Other alternatives beyond GPUs are usually based on ASICs or FPGAs. From Intel we have Altera FPGAs, Nervana Engines (being acquired), and Movidius VPUs (being acquired), as well as the Knight's Mill (the next-generation 2017 version of Phi). From other companies, solutions include Alphabet's Google TPUs, Wave Computing DPUs, DeePhi Tech DPUs, and IBM's True North neuromorphic chips. Notably, all of these technologies have enhanced performance for reduced precision arithmetic.

Driven by digitization and the dawn of the Information Age, HPAI relies on the presence of large bodies of data, advanced mathematical algorithms, and high performance hardware and software

Software

Software frameworks and toolkits are perhaps the most significant elements in delivering an effective deep learning environment, and this is a rapidly developing area.

There are number of popular open source deep learning toolkits gaining traction including:

- Caffe from Berkeley Vision and Learning Center
- Torch supported by Facebook, Twitter, Google and others
- Theano from the Université de Montréal
- Tensorflow from Google
- The NVIDIA Deep Learning SDK
- The HPE Cognitive Computing Toolkit from Hewlett Packard Labs

In addition, accelerated computing technologies are typically dependent on the use of associated parallel programming environments and application programming interfaces (APIs) such as CUDA from NVIDIA, Parallel Studio XE 2016 from Intel, or heterogeneous environments such as OpenCL.

With all of these technology components coming together we are on the cusp of a new era of cognitive computing that can elevate deep learning and artificial intelligence systems beyond the research level and high profile media events. Examples such as IBM's Watson system appearing on and winning the TV quiz program Jeopardy or Google's DeepMInd AlphaGo system convincingly beating one of the world's leading Go champions in a game that is several orders of magnitude more complex than chess are extremely impressive. These milestones point to the opportunities for the mainstream adoption of deep learning systems. When combined with HPC systems and techniques, they herald the dawn of a new market of high performance artificial intelligence.

Characteristics of HPAI: Deep, Wide, Short and High

The application of HPC to AI has had three notable impacts on the processing and workflow of machine learning applications:

Deep and Wide

The key computational kernels involve linear algebra, including matrix and tensor arithmetic. A **deep** neural net can have millions or even billions of parameters due to their rich connectivity. While depth refers to the number of layers, the layers can also be quite **wide**, with hundreds to thousands of neurons in a given layer. The weights of these connections must be adjusted iteratively until a solution is reached in a space of very high dimensionality.

Short

Because of the large number of parameters and the generally modest accuracy required for the final output (examples include, say: "*is this image a cat?"* or, in a different field, *"is this a fraudulent application?"*), low precision arithmetic typically suffices. Training can be successful with floating point half precision (16 bits) or with fixed point or integers (as low as 8 bits in some cases). This is the **short** aspect.

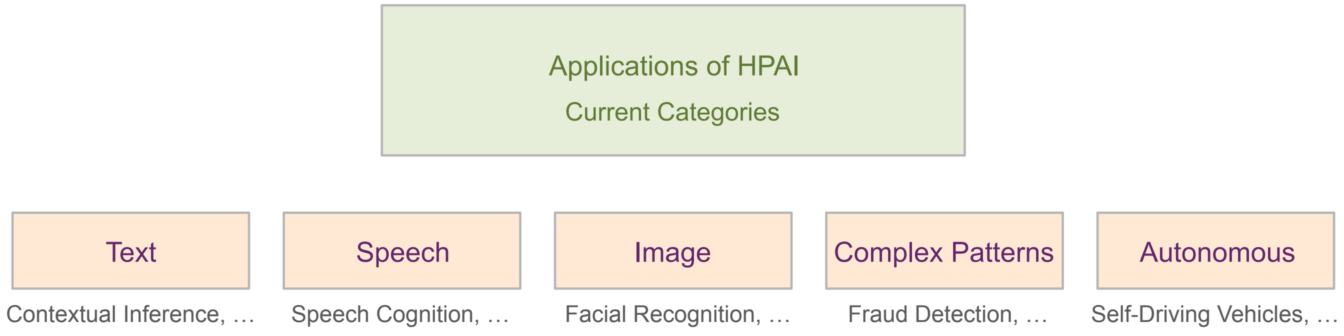
Yann LeCun, one of the pioneers of deep learning has noted, "Getting excellent results on ImageNet is easily achieved with a convolutional net with something like 8 to 16 bit precision on the neuron states and weights." (source: Facebook post regarding IBM's True North chip).

High

The dominance of linear algebra kernels plus short precision indicates that accelerator hardware is extremely useful in deep learning. Overall the class of problems being addressed is that of very high order optimization problems with very large input data sets; it is thus natural that deep learning has entered the realm of **high** performance computing.

Practical Applications of HPAI

It is much harder to try and think of areas where HPAI systems can't add value than the areas where they do. Whether it is threat assessment, managing drones or robots, or optimally managing someone's calendar, the list is endless. Currently, we identify five broad categories:



©2016 OrionX.net

The Future of HPAI

AI has been evolving for decades. Initial inference-based expert systems laid the foundation, and taught us how to formulate and solve AI problems. With deep learning and HPC technologies, AI is taking an evolutionary leap into a new phase. HPAI will include the following challenges and advances.

Advances

Advanced Algorithms

Current algorithms make simplifying assumptions that will be relaxed in the future. In addition to the depth and breadth of layers, there will be cross-links connecting various layers, and dynamically created mini-layers, to provide more flexibility for deep neural networks. Furthermore, while current algorithms iteratively approach an optimum set of parameters, future algorithms will pursue many paths in parallel.

More Realistic Neurons

Current implementations of neuron models are simplistic, with S-curve like or other simple transfer functions. Real-world neurons have much richer connectivity, and often exhibit very spiky signaling behavior. The frequency of spikes can transmit information as well. Future neural nets will incorporate such additional complexity for higher accuracy and to achieve similar results with fewer neurons in the model. Computational complexity will increase, however.

IT Systems

Deep learning is already accelerating new system architecture and component technologies. We expect a period of blossoming innovation across the board: accelerator technologies, new types of CPUs specifically optimized for new workloads, new data storage and processing models such as In-Situ Processing, and entirely novel approaches such as Quantum Computing. These will all evolve rapidly in the coming years.

HPAI shows that sufficiently complex sets of mathematical equations can make cognition computable

Man-Machine Interactions

Natural language processing, augmented and virtual reality, haptic and gesture systems, and brain wave analysis are examples of new forms of interaction between humans and information machines.

Synergy with IoT and HPC

HPAI relies on large bodies of data, which are often generated by sensors and edge devices. Depending on the use case, this data can feed cognitive processing. At the same time, the quest for more accuracy across more and more fathomable situations will continue to justify the designation HPAI.

Smart and Autonomous Devices

Because learning can be separated from practice, and practice can be computationally cheap, a proliferation of smart devices can be expected. This trend is already visible but will expand to entirely new classes of devices. Edge devices, wearables, artificial limbs and exoskeletons, and near-permanent attachments such as smart contact lenses are examples.

Robots

A special class of autonomous devices, robots aim to mimic humans and animals. As such, they not only perform tasks better than humans and perform tasks that humans are unable to perform. They will also become increasingly social. Turing tests will be passed. Humans are social animals and can easily develop emotional bonds with robots.

Cyborg

This is the ultimate in integration of technology and humans into a single cognitive being. Cyborg technologies will become a permanent part of host humans.

Challenges and Grand Challenges

HPAI can help solve existing grand challenge problems by better integrating theory, simulation, and experiment, but it will create new grand challenges that span multiple disciplines.

Unpredictability

HPAI shows that sufficiently complex sets of equations can make cognition computable. But that same complexity makes them unpredictable.

Consequences of AI systems are not always adequately or widely understood, and advanced applications of AI can be a monumental case of unintended consequences. In short, system complexity can easily exceed human competence.

Ethical Complexity

Like any advanced tool, AI can be used for good or evil. Most often, it is quite straightforward to tell whether the application of a technology is good or bad for its users or the society. With AI, this is not always simple.

Current anxieties about AI include the imminent elimination of large classes of jobs by AI systems. Future concerns are about humans making a so-called Darwinian mistake: creating something that will threaten the survival of its creators.

Counter arguments point to the still-primitive nature AI systems in terms of the breadth of its capabilities or the more nuanced aspects of human intelligence.

An ethical framework, similar to that proposed by Asimov for robots, would allow a more structured discussion. Ethical concerns about AI are valid even as they temper the adoption of AI technologies and require formal efforts to study ethical implications of AI.

Legal Framework

Arguably a more important parameter than technological advances, and in light of its ethical complexities, AI poses significant challenges for legal systems, and requires new norms and legislation. We expect progress in this area will lag actual deployments of technologies and will be more reactive than proactive.

Autonomy

Autonomy will be limited by the precise definition of the tasks that are automated, the environment (exact boundaries) in which they operate, and tolerance for mistakes.

Of course, for some tasks, machines do not have to be perfect, but simply better than humans, or more practically, better than the specific human responsible for a task at a given time and place. In such cases, mistakes will be made. Being at peace with a mistake made by a machine may or may not be easier than that made by a human. Society is far from accepting mistakes made by machines at the same level for which human error is accepted.

Fully autonomous systems are far from imminent.

Epilogue

The intersection of HPC and AI has created the HPAI market, a vibrant and rapidly growing segment with far reaching implications not just for the IT industry but humanity as a whole.

Driven by digitization and the dawn of the Information Age, HPAI relies on the presence of large bodies of data, advanced mathematical algorithms, and high performance hardware and software.

Just as industrial machines ushered in a new phase in human history, new “information machines” will have a profound impact on every aspect of life. No different than industrial machines, information machines can help when the scope of their activity is fully defined: the right tool for the right task.

If it can be defined, it can be automated. Whether, or how well, it can be defined is the crux of the matter.

Can we successfully program in Asimov’s three laws?