Research Findings: HPC-enabled AI

Steve Conway
sconway@hyperionres.com
September 2019
Recent Worldwide Studies/Projects for U.S. Federal Agencies

- The Evolution of AI Hardware and Software Ecosystems
- The Evolution of Field Competencies in Machine/Deep Learning and Resultant Industries
- AI Primer for Senior Decision-Makers
- AI Hardware Technology, Vendor Status and Trends
Converging HPC-Enterprise Market

**HPDA Data-Intensive Simulation**
- **Existing HPC users**
  - Larger problem sizes
  - Higher resolution
  - Iterative methods
  - EP jobs to the cloud (Novartis)
- **New commercial users**
  - E.g., SMEs

**HPDA Data-Intensive Analytics**
- **Existing HPC users**
  - Intelligence community, FSI
  - Data-driven science/engineering (e.g., biology)
  - Knowledge discovery
  - ML/DL, cognitive, AI
- **New commercial users**
  - Fraud/anomaly detection
  - Business intelligence
  - Affinity marketing
  - Personalized medicine
  - Smart cities
  - IoT

©Hyperion Research 2019
Most Economically Important HPDA-AI Use Cases

- Precision Medicine
- Automated Driving Systems
- Fraud and anomaly detection
- Affinity Marketing
- Business Intelligence
- Cyber Security
- Edge/IoT/Smart Cities

© Hyperion Research 2019
High Growth Areas: HPDA-AI (May 2019)

- HPDA is growing faster than overall HPC market.
- AI subset is growing faster than all HPDA.

Table 1
Forecast: Worldwide HPC-Based AI Revenues vs Total HPDA Revenues ($ Millions)

<table>
<thead>
<tr>
<th></th>
<th>2018</th>
<th>2019</th>
<th>2020</th>
<th>2021</th>
<th>2022</th>
<th>2023</th>
<th>CAGR 18-23</th>
</tr>
</thead>
<tbody>
<tr>
<td>WW HPC Server Revenues</td>
<td>13,706</td>
<td>14,495.000</td>
<td>15,780</td>
<td>17,376</td>
<td>18,983</td>
<td>19,947</td>
<td>7.8%</td>
</tr>
<tr>
<td>Total WW HPDA Server Revenues</td>
<td>3,153</td>
<td>3,598</td>
<td>3,932</td>
<td>4,737</td>
<td>5,467</td>
<td>6,450</td>
<td>15.4%</td>
</tr>
<tr>
<td>Total HPC-Based AI (ML, DL, and Other)</td>
<td>747</td>
<td>938</td>
<td>1,094</td>
<td>1,399</td>
<td>1,810</td>
<td>2,725</td>
<td>29.5%</td>
</tr>
</tbody>
</table>

Source: Hyperion Research 2019

Table 2
Forecast: Worldwide ML, DL & Other AI HPC-Based Revenues ($ Millions)

<table>
<thead>
<tr>
<th></th>
<th>2018</th>
<th>2019</th>
<th>2020</th>
<th>2021</th>
<th>2022</th>
<th>2023</th>
<th>CAGR 18-23</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML in HPC</td>
<td>532</td>
<td>675</td>
<td>875</td>
<td>1,130</td>
<td>1,479</td>
<td>1,940</td>
<td>29.5%</td>
</tr>
<tr>
<td>DL in HPC</td>
<td>177</td>
<td>216</td>
<td>301</td>
<td>392</td>
<td>510</td>
<td>665</td>
<td>30.3%</td>
</tr>
<tr>
<td>Other AI in HPC</td>
<td>38</td>
<td>47</td>
<td>66</td>
<td>80</td>
<td>95</td>
<td>120</td>
<td>25.9%</td>
</tr>
<tr>
<td>Total</td>
<td>747</td>
<td>938</td>
<td>1,094</td>
<td>1,399</td>
<td>1,810</td>
<td>2,725</td>
<td>29.5%</td>
</tr>
</tbody>
</table>

Source: Hyperion Research 2019
Where Do You Run HPC workloads? (Choose ALL that apply)

<table>
<thead>
<tr>
<th>Option</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-premise HPC data center</td>
<td>67.2%</td>
</tr>
<tr>
<td>On-premise enterprise data center (business operations)</td>
<td>36.2%</td>
</tr>
<tr>
<td>More than one external cloud (e.g., Amazon, Google, Microsoft)</td>
<td>29.3%</td>
</tr>
<tr>
<td>On-premise private or hybrid cloud</td>
<td>19.0%</td>
</tr>
<tr>
<td>One external cloud (e.g., Amazon, Google, Microsoft)</td>
<td>19.0%</td>
</tr>
<tr>
<td>Not sure/don’t know</td>
<td>1.7%</td>
</tr>
<tr>
<td>Other</td>
<td>3.5%</td>
</tr>
</tbody>
</table>
### Machine Learning Goes Back At Least to the 1950s

<table>
<thead>
<tr>
<th>Decade</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1950s</td>
<td>Statistical methods are discovered and refined.</td>
</tr>
<tr>
<td>1950s</td>
<td>Pioneering <strong>machine learning</strong> research is conducted using simple algorithms.</td>
</tr>
<tr>
<td>1960s</td>
<td><strong>Bayesian methods</strong> are introduced for <strong>probabilistic inference</strong> in machine learning.¹</td>
</tr>
<tr>
<td>1970s</td>
<td>'AI Winter' caused by pessimism about machine learning effectiveness.</td>
</tr>
<tr>
<td>1980s</td>
<td>Rediscovery of <strong>backpropagation</strong> causes a resurgence in machine learning research.</td>
</tr>
<tr>
<td>1990s</td>
<td>Work on machine learning shifts from a knowledge-driven approach to a data-driven approach. Scientists begin creating programs for computers to analyze large amounts of data and draw conclusions – or &quot;learn&quot; – from the results.² Support vector machines (SVMs) and [³] recurrent neural networks (RNNs) become popular. The fields of [⁴] computational complexity via neural networks and super-Turing computation started.</td>
</tr>
<tr>
<td>2000s</td>
<td>Support Vector Clustering [⁵] and other <strong>Kernel methods</strong> [⁶] and unsupervised machine learning methods become widespread.⁷</td>
</tr>
<tr>
<td>2010s</td>
<td><strong>Deep learning</strong> becomes feasible, which leads to machine learning becoming integral to many widely used software services and applications.</td>
</tr>
</tbody>
</table>

By the early 1990s, George Washington University Hospital (Washington, DC) was routinely using a Cray supercomputer to help detect breast cancer after training it to identify early indicators, called microcalcifications, on X-ray films with better-than-human ability.
Coupled Environments

- **Automated Driving Systems**
  - Embedded processor for local control (car-car, car-environment)
  - Private cloud for citywide and beyond (“air traffic control”)

- **Healthcare/Precision Medicine**
  - Healthcare systems are already private cloud-based.
  - Future: couple in-office HPC decision-support engine to private cloud.

- **5G Will Reduce Local-Cloud Latency Issue**
AI Is Still Near the Start

Today: Bounded Problems
- Many observations but few choices to make
- "One trick dogs": 10 AI solutions in a box to solve 10 problems
- Already very useful:
  - Image & voice recognition
  - Advanced driver assistance
  - Reading an MRI

Future: Unbounded, Too
- Many observations, many choices to make
- Versatile decision-makers capable of serious experiential learning
- Examples:
  - Discerning human motivation
  - Fully automated driving
  - Diagnosing/"curing" a cancer

©Hyperion Research 2019
<table>
<thead>
<tr>
<th>Importance</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely important</td>
<td>47</td>
<td>87</td>
</tr>
<tr>
<td>Somewhat Important</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Not very important</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Not sure/don't know</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>54</td>
<td>100</td>
</tr>
</tbody>
</table>
Intelligent Simulation

- AI adds fourth branch to the scientific method, *inferencing*. Complements theory, experiments & established simulation methods.
- Inferencing is the ability to guess, based on incomplete information.
- Simulation is becoming much more data-intensive—esp. iterative methods.
- When inferencing is applied to data-intensive simulation, the result is *intelligent simulation*. 
Simulation Helping Analytics

Building Consumer Trust in ADS

- RAND Corp. estimates 8.8 billion miles of physical testing would be needed to attain 95% consumer trust in self-driving vehicles. This would take 400 years.
- Adding HPC simulation can reduce time frame to 5-10 years.
Climate research is inherently data-intensive.

Ensemble models & adding new factors (e.g., carbon cycle) have made it much more so.

Climate knowledge discovery algorithms add a data analytics approach.

The first IEEE workshop on this topic (2008), was called "Data Mining for Climate Change and Impacts."
Future HPC System Architectures

One Swim Lane in the Future

- Today: Mostly simulation and analytics on same compute-centric HPC system (budget reasons)
- In 2 years: More buyers will acquire separate system for analytics-AI.
  - But orthogonal findings of simulation & analytics runs will still need to be combined in the researcher’s brain.
- Farther ahead: same system efficiently performs concurrent simulation & analytics runs – and integrates the orthogonal results.
  - Finally, economically efficient!
An Exploding Ecosystem

- The AI ecosystem has been growing quickly
- Targeted problems are more complex
- One result: new, AI-specific hardware
  - From companies large and small
  - China is very active
- The categories of processors and technologies continue to grow as well, and now include:
  - GPUs
  - TPUs
  - FPGAs
  - ASICs & eASICs
  - Neuromorphic Chips
  - IPUs
  - Inference Chips
  - Training Chips
  - Dataflow processors
  - Vector processors
  - 3D stacking
  - Optical interconnects
Co-Design

- AI chips will be centered on co-design, with specific tasks in mind. Examples:
  - Low-power ASICs at the edge
  - Custom AI chips in hyperscale data centers or the cloud
- GPUs will remain important but not for all AI workloads.
- Software and model-designed hardware is the direction forward.
Power

- Power consumption is critical to chip design for AI workloads.
  - Low power chips can be placed closer to the edge.
  - Latency for near real time training and inference require the compute to be next to the stored data.
- Processing and memory also need to be closer together.
  - With faster interconnect/fabric speeds
Cloud Companies Joining the Processor Development Party

- Google uses tensor cores to accelerate machine learning workloads.
  - Only available on Google cloud for now
  - Google announced the third generation TPU last year.
- Amazon, at their re:Invent conference in November of 2018, announced their inference chip, Inferentia.
  - Designed to accelerate machine learning, especially inferencing.