Serendipity: How Supercomputing Technology is Enabling a Revolution in Artificial Intelligence

José Moreira
IBM Research

HPML/SBAC-PAD 2018 – Lyon, France, September 2018
## TOP500 – June 2018

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>System</th>
<th>Cores</th>
<th>Rmax [TFlop/s]</th>
<th>Rpeak [TFlop/s]</th>
<th>Power [kW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/SC/Oak Ridge National Laboratory, United States</td>
<td>Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM</td>
<td>2,282,544</td>
<td>122,300.0</td>
<td>187,659.3</td>
<td>8,806</td>
</tr>
<tr>
<td>2</td>
<td>National Supercomputing Center in Wuxi, China</td>
<td>Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC</td>
<td>10,649,600</td>
<td>93,014.6</td>
<td>125,435.9</td>
<td>15,371</td>
</tr>
<tr>
<td>3</td>
<td>DOE/NNSA/LLNL, United States</td>
<td>Sierra - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM</td>
<td>1,572,480</td>
<td>71,610.0</td>
<td>119,193.6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>National Super Computer Center in Guangzhou, China</td>
<td>Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692v4 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT</td>
<td>4,981,760</td>
<td>61,444.5</td>
<td>100,678.7</td>
<td>18,482</td>
</tr>
<tr>
<td>5</td>
<td>National Institute of Advanced Industrial Science and Technology (AIST), Japan</td>
<td>AI Bridging Cloud Infrastructure (ABCI) - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR Fujitsu</td>
<td>391,680</td>
<td>19,880.0</td>
<td>32,576.6</td>
<td>1,649</td>
</tr>
</tbody>
</table>
Summit’s Home: DOE ORNL

Oak Ridge National Laboratory is the largest US Department of Energy (DOE) open science laboratory.
New 200-Petaflops System is World’s Most Powerful and World’s Smartest Supercomputer for Science
<table>
<thead>
<tr>
<th>Feature</th>
<th>Summit</th>
<th>Sierra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Performance</td>
<td>200 Petaflops (DP)</td>
<td>125 Petaflops (DP)</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>4,608</td>
<td>4,320</td>
</tr>
<tr>
<td>Node Performance</td>
<td>43/86/700 Teraflops (DP/SP/HP)</td>
<td>29/58/450 Teraflops (DP/SP/HP)</td>
</tr>
<tr>
<td>Memory per node</td>
<td>512 GB DDR4, 96 GB HBM2</td>
<td>256 GB DDR4, 64 GB HBM2</td>
</tr>
<tr>
<td>NV Memory per node</td>
<td>1600 GB</td>
<td>1600 GB</td>
</tr>
<tr>
<td>Total System Memory</td>
<td>11.1PB DDR4+HBM2+NV</td>
<td>9PB DDR4+HBM2+NV</td>
</tr>
<tr>
<td>Compute nodes</td>
<td>9,216 IBM POWER9™ CPUs</td>
<td>8,640 IBM POWER9™ CPUs</td>
</tr>
<tr>
<td>File System</td>
<td>27,648 NVIDIA Volta™ GPUs</td>
<td>17,280 NVIDIA Volta™ GPUs</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>250 PB, 2.5 TB/s, GPFS™</td>
<td>156 PB, 1.5 TB/s, GPFS™</td>
</tr>
<tr>
<td>Interconnect</td>
<td>15 MW</td>
<td>12 MW</td>
</tr>
<tr>
<td>Operating System</td>
<td>Mellanox EDR 100G InfiniBand</td>
<td>Red Hat Enterprise Linux (RHEL) version 7.4</td>
</tr>
</tbody>
</table>
Key components of Summit/Sierra

**GPU Brawn:** Summit and Sierra link more than 27,000 and 17,000, respectively, deep-learning optimized NVIDIA GPUs with the potential to deliver exascale-level performance (a billion-billion calculations per second) for AI applications.

**High-speed Data Movement:** High speed Mellanox interconnect and NVLink high-bandwidth technology built into all of compute processors supply the next-generation information superhighways.

**CPU Muscle:** IBM Power9 processors to rapidly execute serial code, run storage and I/O services, and manage data so the compute is done in the right place.

**Memory Where it Matters:** Summit and Sierra’s sizable memory gives researchers a convenient launching point for data-intensive tasks, an asset that allows for greatly improved application performance and algorithmic accuracy as well as AI training.
IBM POWER9 Processor

• Up to 24 cores
  • Summit’s P9s have 22 cores for yield optimization on first processors

• PCI-Express 4.0
  • Twice as fast as PCIe 3.0

• NVLink 2.0
  • Coherent, high-bandwidth links to GPUs

• 14nm FinFET SOI technology
  • 8 billion transistors

• Cache
  • L1I: 32 KiB per core, 8-way set associative
  • L1D: 32 KiB per core, 8-way
  • L2: 256 KiB per core
  • L3: 120 MiB eDRAM, 20-way
## NVIDIA Volta Details

<table>
<thead>
<tr>
<th></th>
<th>Tesla V100 for NVLink</th>
<th>Tesla V100 for PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PERFORMANCE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with NVIDIA GPU Boost</td>
<td>DOUBLE-PRECISION</td>
<td>DOUBLE-PRECISION</td>
</tr>
<tr>
<td></td>
<td>7.8 TeraFLOPS</td>
<td>7 TeraFLOPS</td>
</tr>
<tr>
<td></td>
<td>SINGLE-PRECISION</td>
<td>SINGLE-PRECISION</td>
</tr>
<tr>
<td></td>
<td>15.7 TeraFLOPS</td>
<td>14 TeraFLOPS</td>
</tr>
<tr>
<td></td>
<td>DEEP LEARNING</td>
<td>DEEP LEARNING</td>
</tr>
<tr>
<td></td>
<td>125 TeraFLOPS</td>
<td>112 TeraFLOPS</td>
</tr>
<tr>
<td><strong>INTERCONNECT BANDWIDTH</strong></td>
<td>NVLINK</td>
<td>PCIe</td>
</tr>
<tr>
<td>Bi-Directional</td>
<td>300 GB/s</td>
<td>32 GB/s</td>
</tr>
<tr>
<td><strong>MEMORY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoWoS Stacked HBM2</td>
<td>CAPACITY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16 GB HBM2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BANDWIDTH</td>
<td></td>
</tr>
<tr>
<td></td>
<td>900 GB/s</td>
<td></td>
</tr>
</tbody>
</table>
Summit 200 PetaFlops System – June, 2018

**POWER9**: 2 Socket Server
- 2 P9 + 6 Volta GPU
- 512 GiB SMP Memory (32 GiB DDR4 RDIMMs)
- 96 GiB GPU Memory (HBM stacks)
- 1.6 TB NVMe

**Volta**: 7.0 DP TF/s
- 16GB @ 0.9 TB/s

**POWER9** 2 Socket Server
- Standard 2U 19in. Rack mount Chassis

**System**
- Compute Rack: 18 nodes
  - 775 TF/s
  - 10.7 TiB
  - 59 KW max
- ESS Building Block
  - SSC (4x IBM ESS GL4):
    - 8 servers, 16 JBOD
    - 16.8 PB (gross)
    - 38 KW max
- Mellanox IB4X EDR 648p Directors Full bisection
- Mellanox IB4X EDR Switch IB-2

**ESS Building**
- Block

**Mellanox**
- IB4X EDR Switch IB-2

**Floor plan rack concept**
- Compute (256)
- Switch (18)
- Storage (40)
- Infrastructure (4)

09/24/2018
HPML/SBAC-PAD 2018 - Lyon, France
1. Uncovering hidden networks of genes by comparing genetic variations within a population (ORNL) – 2.36 exaops!

2. Applying artificial intelligence and mixed-precision arithmetic to the simulation of earthquake physics in urban environments (University of Tokyo)

3. Using deep neural networks to identify extreme weather patterns from high-resolution climate simulations (LBNL) – 1.13 exaops!

4. Intelligent software that can automatically identify materials’ atomic-level information from electron microscopy data (ORNL) – 152.5 petaflops

5. Improved lattice quantum chromodynamics algorithms to help scientists predict the lifetime of neutrons and answer fundamental questions about the universe (LBNL/LLNL) - 15× over Titan

Cognitive Discovery depends on the ingestion and representation of knowledge, be that from the literature, experiment, or from previously run simulations to develop a deep knowledge space.

Intelligent Simulation uses machine learning techniques to optimize ensemble analysis, speed convergence, improve fidelity, increase numerical stability, and actively steer simulations to speed up time to solution.

AI supervisor uses deep knowledge space
- knowledge extracted from literature
- data derived from experimentation (e.g., wet labs)
- data from simulations

AI accelerated modelling and simulation uses machine learning techniques to
- accelerate parameterization of models
- guide ensembles of simulations
- replace expensive simulation routines with AI-derived approximations
The Unreasonable Effectiveness of Mathematics in the Natural Sciences

• Eugene Wigner, 1960:

“The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure, even though perhaps also to our bafflement, to wide branches of learning.”

• How about the effectiveness of HPC hardware for ML/AI?

That one is a little easier to explain ... 😊
The serendipity between GPUs and DNNs

*Shallow* neural network

*Deep* neural network

Any questions?
Deep neural networks are hard to train

Gradient instabilities!
Hard ≠ impossible

• As described in a 2010 paper, Cireșan, Meier, Gambardella, and Schmidhuber* have trained a flat, 5-hidden layer DNN (2500, 2000, 1500, 1000, and 500 neurons, respectively – approximately 12 million weights) to classify the MNIST dataset

• This required no techniques beyond what was known in the 1980’s, but consumed a lot of compute power, achieved through GPUs

• They achieved a classification accuracy of 99.65%! (that is, 35 misclassified images)

*Deep, Big, Simple Neural Nets Excel on Handwritten Digit Recognition, by Dan Claudiu Cireșan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber (2010).
Divergence brewing?

### Tesla V100 GPU

<table>
<thead>
<tr>
<th>PERFORMANCE with NVIDIA GPU Boost&lt;sup&gt;*&lt;/sup&gt;</th>
<th>Tesla V100 for NVLink</th>
<th>Tesla V100 for PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOUBLE-PRECISION</td>
<td>7.8 TeraFLOPS</td>
<td>7 TeraFLOPS</td>
</tr>
<tr>
<td>SINGLE-PRECISION</td>
<td>15.7 TeraFLOPS</td>
<td>14 TeraFLOPS</td>
</tr>
<tr>
<td>DEEP LEARNING</td>
<td>125 TeraFLOPS</td>
<td>112 TeraFLOPS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INTERCONNECT BANDWIDTH Bi-Directional</th>
<th>NVLink</th>
<th>PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td>300 GB/s</td>
<td>32 GB/s</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MEMORY CoWoS Stacked HBM2</th>
<th>CAPACITY</th>
<th>16 GB HBM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>900 GB/s</td>
<td>BANDWIDTH</td>
<td>900 GB/s</td>
</tr>
</tbody>
</table>

### NVIDIA Deep Learning Accelerator

- Headless NVDA core
- Convolution buffer → Convolution core
- Activation engine (SDF) → Pooling engine (PDP)
- Local resp. norm (CDP) → Reshape (RUBIK)
- Bridge DMA

### Tesla T4 GPU (int8/int4)
Machine consciousness

• Our machines are approaching the (suspected) processing power of the human brain
• We have been able to train and deploy deep neural networks of a scale unimaginable just a decade ago
• We continue to make progress at better algorithms and better implementation of large machine learning solutions
• Does that mean we are close to achieving General Artificial Intelligence?
• If so, when do we declare such machines thinking/conscious?
Quiz: What is this?

Diagram of a finite automaton with transitions labeled by symbols and directions (L for left, R for right). The states are labeled from 1 to 7, with transitions connecting them. The automaton includes a HALT state labeled "HALT."
Quiz: What is this? (Marvin Minsky)

Answer: A (Minsky’s) universal Turing machine (minus the tape)
Equivalent to any machine we know how to build
Not equivalent to everything our minds can conceive

- Let $T = \{T_i: T_i$ is a Turing machine$\}$
- Let $I = \{I_j: I_j$ is an input sequence to a Turing machine$\}$
- $|T| = \aleph_0$
- $|I| = \aleph_0$
- Let $C = \{C_i(j): C_i(j)$ is the computation performed by $T_i$ on $I_j$$\}$
- $|C| = \aleph_0$
- $|\mathbb{R}| = \aleph_1 = 2^{\aleph_0}$, that is, almost all real numbers are noncomputable
But, is it equivalent to everything that exists?
But, is it equivalent to everything that exists?

Answer: We don’t know!
Hypothesis of Digital Physics

1. The physical world is essentially informational
2. The physical world is essentially computable
3. The physical world can be described digitally
4. The physical world is in essence digital
5. The physical world is itself a computer
6. The physical world is the output of a simulated reality exercise

**Computable universe hypothesis (CUH) – Max Tegmark**
More relevant to cognitive computing ...
More relevant to cognitive computing ...

Answer: We don’t know!
Viewpoints of consciousness (Roger Penrose)

A. All thinking (including consciousness) is computation – *strong* or *hard* Artificial Intelligence

B. Consciousness is a physical process that can be simulated computationally (known physics), but simulation cannot by itself evoke consciousness (simulation of a process $\neq$ process itself) – *weak* or *soft* Artificial Intelligence
Ex Machina (2015 film – A vs B)
Viewpoints of consciousness (Roger Penrose)

A. All thinking (including consciousness) is computation – *strong* or *hard* Artificial Intelligence

B. Consciousness is a physical process that can be simulated computationally (known physics), but simulation cannot by itself evoke consciousness (simulation of a process ≠ process itself) – *weak* or *soft* Artificial Intelligence

C. Consciousness is a physical process that cannot be simulated computationally – new physics

D. Consciousness cannot be scientifically explained – mysticism
Proof (?) that there is more than computation (RP)

• Let $A$ be an algorithmic procedure taking two inputs, $i$ and $j$, such that if $A(i, j)$ terminates then $C_i(j)$ does not terminate

• $A$ is known and believed to be sound – encapsulates all the procedures available for demonstrating that computations do not stop

• It does not need to be perfect – $A(i, j)$ may not terminate when $C_i(j)$ does not terminate

• Make $i = j$, then if $A(j, j)$ terminates then $C_j(j)$ does not terminate

• But $A(j, j) = C_k(j)$, for some $k$

• Make $j = k$, then if $A(k, k)$ terminates then $C_k(k)$ does not terminate

• But $A(k, k) = C_k(k)$

• Then if $A(k, k)$ terminates then $A(k, k)$ does not terminate

• Therefore, $A(k, k)$ does not terminate – does not encompass all our knowledge!
Conclusions

• The new Summit/Sierra supercomputers are ushering a new era of large-scale scientific/artificial intelligence computing

• Even more significant, scientific/artificial intelligence computing are being integrated into new large-scale applications

• Will integration at the application level continue to drive integration at the hardware level, or are we going to see more divergence?

• Current machine learning/cognitive computing research will not give us thinking/conscious machines

• And it doesn’t have to – that is not the purpose

• “Is the human mind computable?” – One of the most important scientific questions we face today!
• Marvin Minsky. **Computation: Finite and Infinite Machines.** *Prentice-Hall.*

• Roger Penrose. **The Large, the Small and the Human Mind.** *Cambridge University Press.*

• Roger Penrose. **Shadows of the Mind.** *Oxford University Press.*

• Hava T. Siegelmann. **Computation Beyond the Turing Limit.** *Science,* vol. 268, no. 5210.

• Robert I. Soare. **Turing Oracle Machines, Online Computing, and Three Displacements in Computability Theory.**


• Max Tegmark. **Our Mathematical Universe: My Quest for the Ultimate Nature of Reality.** *Penguin Random House.*

• Masafumi Oizumi, Larissa Albantakis, Giulio Tononi. **From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0.**
  [https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003588](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003588)
On a personal note ...