The Impact of AI on HPC ... “Everywhere”

Rob Farber
CEO TechEnablement.com

Contact info AT techenablement.com for consulting, teaching, writing, and other inquiries
GPUs - From games to supercomputers to Deep Learning. CPUs – General purpose to DL devices

GPUs evolved from pushing pixels

CPU evolved from running applications
- Failure of Dennard’s scaling laws caused switch to multicore
- Parallelism via GPUs and Intel Xeon Phi is now the norm
- CPUs are now competitive!

Intel Xeon Scalable

NVIDIA

AMD

...
CPUs are also good for AI

• Don’t get locked into thinking “have to have a GPU”
Jump on the AI/Machine learning bandwagon

- **Electronic Brain** (1943)
- **Perceptron** (1957)
- **XOR** (1969)
- **Multilayer Perceptron** Backprop (1986)

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1940</td>
<td>Electronic Brain</td>
<td>(1943)</td>
</tr>
<tr>
<td>1950</td>
<td>Perceptron</td>
<td>(1957)</td>
</tr>
<tr>
<td>1960</td>
<td>Electronic Brain</td>
<td>(1943)</td>
</tr>
<tr>
<td>1980</td>
<td>Electronic Brain</td>
<td>(1943)</td>
</tr>
<tr>
<td>1990</td>
<td>Electronic Brain</td>
<td>(1943)</td>
</tr>
<tr>
<td>2000</td>
<td>Electronic Brain</td>
<td>(1943)</td>
</tr>
<tr>
<td>2010</td>
<td>Electronic Brain</td>
<td>(1943)</td>
</tr>
<tr>
<td>2020</td>
<td>Electronic Brain</td>
<td>(1943)</td>
</tr>
</tbody>
</table>
“Computational Universality” An XOR ANN

- The example of XOR nicely emphasizes the importance of hidden neurons:
  - They re-represent the input such that the problem becomes linearly separable.

- Networks with hidden units can implement any Boolean function -> Computational Universal devices!

- Networks without hidden units cannot learn XOR
  - Cannot represent large classes of problems

---

**NetTalk**
Sejnowski, T. J. and Rosenberg, C. R. (1986) NETtalk: a parallel network that learns to read aloud, Cognitive Science, 14, 179-211
Big hardware idea 1: SIMD

High-performance from the past
• Space and power efficient
• Long life via a simple model

The Connection Machine $30M

Farber: general SIMD mapping:
“Most efficient implementation to date”
(Singer 1990), (Thearling 1995)

2013 Results courtesy Intel and NVIDIA

16,384 GPU results 2014
courtesy ORNL
A general SIMD mapping:

$$\text{Optimize}(\text{LMS\_Error} = \text{objFunc}(p_1, p_2, \ldots p_n))$$

**Step 1:** Broadcast parameters

**Step 2:** Calculate partials

**Step 3:** Sum partials to get energy

---

**GPU 1:**

Examples: 0, N-1

$p_1, p_2, \ldots p_n$

---

**GPU 2:**

Examples: N, 2N-1

$p_1, p_2, \ldots p_n$

---

**GPU 3:**

Examples: 2N, 3N-1

$p_1, p_2, \ldots p_n$

---

**GPU 4:**

Examples: 3N, 4N-1

$p_1, p_2, \ldots p_n$

---

**Host**
See a path to exascale
(MPI can map to thousands of GPU or Processor nodes)

Always report “Honest Flops”

\[
\text{Effective Rate} = \frac{\text{Total Op Count}}{T_{\text{broadcast}} + T_{\text{objectfunc}} + T_{\text{reduce}}}
\]
Application to Bioinformatics

**NetTalk**
Sejnowski, T. J. and Rosenberg, C. R. (1986) NETtalk: a parallel network that learns to read aloud, Cognitive Science, 14, 179-211


The phoneme to be pronounced

**T/F Exon region**

Internal connections

```
t e X t
```
Predicting binding affinity
(The closer you look the greater the complexity)
The question for computational biology

How do we know you are not playing expensive computer games with our money?
Utilize a blind test (Example reference when starting my drug discovery company)

Internal connections

Possible hexamers

\[ 20^6 = 64M \]

1k – 2k pseudo-random (hexamer, binding) affinity pairs

Approx. 0.001% sampling

Hill climbing to find high affinity

Learn:

\[ \text{Affinity}^{\text{Antibody}} = f(A_0, \ldots, A_5) \]

Predict P,C,T,N,S,L has the highest binding affinity

Confirm experimentally

\[ f(P, C, T, N, S, L) \]

\[ f(F, F, F, F, F, V) \]

\[ f(F, F, F, F, F, L) \]

\[ f(F, F, F, F, F, F) \]
Time series

Learn:

\[ X_{t+1} = f(X_t, X_{t-1}, X_{t-2}, ...) \]

\[ X_{t+2} = f(X_{t+1}, X_t, X_{t-1}, ...) \]

\[ X_{t+3} = f(X_{t+2}, X_{t+1}, X_t, ...) \]

\[ X_{t+4} = f(X_{t+3}, X_{t+2}, X_{t+1}, ...) \]

Iterate

Works great! (better than other methods at that time)

Designing ANNs for Integration and Bifurcation analysis – “training a netlet”


ANN schematic for continuous-time identification. (a) A four-layered ANN based on a fourth order Runge-Kutta integrator. (b) ANN embedded in a simple implicit integrator.

(a) Periodic attractor of the Van der Pol oscillator for $\gamma = 1.0$, $\delta = 4.0$ and $\omega = 1.0$. The unstable steady state in the interior of the curve is marked +. (b) ANN-based predictions for the attractors of the Van der Pol oscillator shown in (a).
Dimension reduction

- The curse of dimensionality
  - People cannot visualize data beyond 3D + color
  - Search volume rapidly increases with dimension
    - Queries return too much data or no data
Jump on the AI/Machine learning bandwagon AGAIN!

Electronic Brain (1943)  Perceptron (1957)  XOR (1969)


Wildly Overpromised


- Adjustable Weights
- Weights are not Learned
- Learnable Weights and Threshold
- XOR Problem
- Solution to nonlinearly separable problems
- Big computation, local optima and overfitting
- Hierarchical feature Learning

STOP XOR
People are “rediscovering” ANNs in an Exascale framework

All the previous work is now orders of magnitude faster!

• Time series
• Bifurcation analysis
• Locally Weighted Linear Regression (LWLR)
• Neural Networks
• Naive Bayes (NB)
• Gaussian Discriminative Analysis (GDA)
• k-means
• Logistic Regression (LR)
• Independent Component Analysis (ICA)
• Expectation Maximization (EM)
• Support Vector Machine (SVM)
• Others: (MDS, Ordinal MDS, etcetera)
This is not an AI history talk – Deep learning is commercially viable encompassing much of what we do

Speech recognition in noisy environments (Siri, Cortana, Google, Baidu, ...)

Better than human face recognition

Self-driving cars

Internet Search  •  Robotics  •  Self guiding drones  •  Much, much, more

Plus everything discussed previously: time series, system modeling, drug discover, and more
A change in the scientific mindset: ANNs in physical models

- Previously: ANNs are black boxes – how can you guarantee they won’t add non-physical artifacts to my model?
- Today: CERN has shown that GANs (Generative Adversarial Networks) can learn physics-based distributions that cannot be distinguished from real distributions.

It does not matter which generator one uses so long as the generated distributions are indistinguishable from the real data distributions – Sophia Vallecorsa, CERN physicist, Openlab.

Time to generate an electron shower:

- Full Monte Carlo: **17000.00 msec**
- GAN running on GTX 1080: **00000.04 msec**
ANNs evaluate simulation results

• Deep learning ANNs recognize storm types in faster-than-realtime climate models.
  • Too much data for a human
  • Validation via historical data
  • Evaluate storm distributions in the future predictions
Expect significant algorithm and HW retooling

Today: General-purpose devices like CPUs and GPUs
Tomorrow: FPGA and ASIC Data Flow devices
Near term: Neuromorphic will change everything
Far term: Quantum: Seems to be ideal for training (once we have the qubits)

Roughly four different camps

- CPU
- GPU
- FPGA
- Chips
Ubiquitous Low precision AI hardware is changing HPC

- NLAFET: Reduced precision BLAS
  - Synchronization-reducing algorithms
  - Communication-reducing algorithms
  - Mixed precision methods (2x speed of ops and 2x speed for data movement)
  - Autotuning
  - Fault resilient algorithms
  - Reproducibility of results

- LIBXSMM: batched reduced precision small matrix operations => very high percentage of peak!

Scientists reexamining precision

- The HiCMA library: million by million matrix operations in gigabytes of workstation memory
DAGs are also changing HPC

- NLAFET uses DAG scheduling to support CPUs, GPUs, and Data Flow architectures

**Runtime DAG scheduling**

- Every node has the **symbolic DAG representation**
  - Only the (node local) frontier of the DAG is considered
  - Distributed Scheduling based on remote completion notifications
- Background remote **data transfer automatic with overlap**
- NUMA / Cache aware Scheduling
  - Work Stealing and sharing based on memory hierarchies
From HPC centers to IoT edge devices

HPC centers will soon be part of a data network, not the center of the HPC effort.

• Read the BDEC report (Big Data and Extreme-scale Computing)

• Look at Frontera at TACC – an “agile” HPC supercomputer!
  • Largest academic supercomputer at a University
  • Leverages the older Wrangler effort data analytics system
  • The Argonne lab-wide data service based on the Petrel service.

• More ...
Not bad for surface fitting – but what’s next?

- Lapedes and Farber, “How Neural Networks Work,” showed that during training the neural network is actually fitting a “bumpy” multi-dimensional surface.

- Inference:
  - Interpolates between points on the surface, or
  - Extrapolates out from points on the surface.

Birthday card: “Herpes Bathtub!”

Inside the card: “Happy Birthday! (Damn voice recognition)”
Neuromorphic – the next step!

• “[We] demonstrate that neuromorphic computing ... can implement deep convolution networks that approach state-of-the-art classification accuracy across eight standard datasets encompassing vision and speech, perform inference while preserving the hardware’s underlying energy-efficiency and high throughput.”

Accuracy of different sized networks running on one or more TrueNorth chips to perform inference on eight datasets. For comparison, accuracy of state-of-the-art unconstrained approaches are shown as bold horizontal lines (hardware resources used for these networks are not indicated).
Build big and very, very power efficient networks (e.g. “brains”)

PNAS (8/9/16) Convolutional networks for fast, energy-efficient neuromorphic computing, Steven K. Essera et. al.

- Use integrate-and-fire spiking neurons
- Data sets run between 1,200 and 2,600 frames/s and using between 25 and 275 mW
- (effectively >6,000 frames/s per watt)

A system roughly the neuronal size of a rat brain
It’s easy “From “Hello World” to exascale”

```c
int main()
{
    cout << "Hello World" << endl;
    // load data and initialize parameters
    init();
    #pragma acc data
    copyin(param[0:N_PARAM-1])
    pcopyin(example[0:nExamples*EXAMPLE_SIZE-1])
    {
        optimize( objFunc ); // the optimizer calls the objective function
    }
    return 0;
}

double objFunc(...)
{
    double err=0.;
    #pragma acc parallel loop reduction(+:err)
    #pragma omp parallel for reduction(+ : err)
    {
        err = 0.;
        for(int i=0; i < nExamples; i++) {
            // transform
            float d=myFunc(i, param, example, nExamples, NULL);
            //reduce
            err += d*d;
        }
    }
    return sqrt(err);
}
```

Intel Core™ i7

NVIDIA TESLA
## Organizing HPC + AI Convergence

### Operation
HPC + AI couple simulation with live data in real time detection/control system

Experimental/simulated data is used to train a NN, where resulting inference engineering is used to for real-time detection/control of an experiment or clinical delivery system. The NN is improved as new simulated/live data is acquired.

### Augmentation
HPC + AI combined to improve simulation time to science > orders of magnitude

Experimental/simulated data is used to train a NN that is used to replace all or significant runtime portions of a conventional simulation. The NN is improved continuously as new simulated/live data is acquired.

### Modulation
HPC + AI combined to reduce the number of runs needed for a parameter sweep

Experimental/simulated data used to train a NN which steers simulation/experiment between runs. The steering NN can be trained continuously as new simulated/live data is acquired.

---

**Potential for Breakthroughs in Scientific Insight**
So much more, you have been great, Thank You!

Rob Farber
CEO TechEnablement.com

Contact info AT techenablement.com for consulting, teaching, writing, and other inquiries
Numerous use cases I don’t have time to present
Background
The aLIGO (Advanced Laser Interferometer Gravitational Wave Observatory) experiment successfully discovered signals proving Einstein’s theory of General Relativity and the existence of cosmic Gravitational Waves. While this discovery was by itself extraordinary it is a step to the ultimate goal to combine multiple observational data sources that not only hear but also see to the complete spectrum of data in real time.

Challenge
The initial aLIGO discoveries were successfully completed using classic data analytics. The processing pipeline used hundreds of CPU’s where the bulk of the detection processing was done offline. The latency is far outside the range needed to activate resources, such as optical, infrared or radio telescopes which observe phenomena in the electromagnetic spectrum in time to “see” what aLIGO can “hear”.

Solution
A DNN was developed and trained with simulated data and verified using from the CACTUS/Einstein Toolkit. The DNN was shown to produce better accuracy with latencies 4500x faster than the original CPU based pattern matching waveform detection.

Impact
Faster and more accurate detection of gravitational waves with the potential to steer other observational data sources.
Background
The “Grand Challenge” of fusion energy would offer the humankind changing opportunity to provide clean, safe energy for millions of years.
ITER is a $25B international experiment to develop the prototype to demonstrate commercially viable fusion reactor.

Challenge
The plasma in a fusion reactor is highly turbulent at the edges of the flow, and disruptions can occur that break the magnetic confinement, which can cause damage to the physical reactor.
It is critical to predict when a disruption will occur to prevent damage and maintain safe operation.
Traditional simulation and ML approaches were 65% to 85% accurate with 5% false alarm rate.

Solution
DL network called FRNN using Theano exceeds today’s best accuracy results. It scales to 200 Tesla K20s, and with more GPUs, can deliver higher accuracy. Current level of accuracy is 95% prediction with 5% false alarm rate.

Impact
Vision is to operate ITER with FRNN, operating and steering experiments in real-time to minimize damage and down-time.
The simulation for the LHC particle collider is known as GEANT and it is numerically intensive, requires significant compute cycles as part of the scientific workflow and is extremely hard to optimize for modern CPU or GPU type architectures. The High Luminosity LHC is expected to generate 10X the volume of data, and the compute load on GEANT is in turn expected to challenge future computing systems within a flat budget profile.

A GAN (caloGAN) was constructed by a team at Lawrence Berkeley Lab, CERN and Yale that was customized for Calorimeter Experiments similar to ATLAS in the LHC. A training data set was developed using 100,000 GEANT events as input. 50 epochs were then used on 18xK80's to train the GAN with KERAS and TensorFlow.

Impact

The CaloGAN is more accurate and showed up to 5 Orders of Magnitude performance speed up relative to the original simulation on a single K80 GPU.
Background
It takes 14 years and $2.5 Billion to develop 1 drug
Higher than 99.5% failure rate after the drug discovery phase

Challenge
QC simulation is computationally expensive - it takes 5 years to compute on CPUs
So researchers use approximations, compromising on accuracy. To screen 10M drug candidates.

Solution
Researchers at the University of Florida and the University of North Carolina leveraged GPU deep learning to develop a custom framework ANAKIN-ME, to reproduce molecular energy surfaces with super speed (microseconds versus several minutes), extremely high (DFT) accuracy, and at up to 6 orders of magnitude improvement in speed.

Impact
Speed and accuracy could start a revolution in computational chemistry — and forever change the way we discover the medicines of the future.
Converged HPC Accelerates Drug Discovery

The “drug discovery” phase of the development process involves exploring all the different possible combinations of protein molecules (targets) and drug chemical compounds to ensure the drug will do what it’s designed to do. Classic Molecular Dynamics simulations are very time-consuming and expensive. Machine Learning models have been designed to help predict probability of the target molecules interacting with the drug chemical compounds, but still require significantly more performance to deliver improved accuracy.

Researchers developed and trained a convolutional neural network accelerated with NVIDIA GPU’s to improve the model performance and prediction accuracy. Ultimately, they improved prediction accuracy from approximately 52% to 70% compared to other machine learning-based models (Vina Docking). (35% relative improvement)
HUNTING “GHOST PARTICLES” WITH DEEP LEARNING

Background
The NoVA experiment managed by Fermi lab comprises 200 scientists at 40 institutions in 7 countries. The goal is to track neutrino's, which are often referred to as the “Ghost Particle”, and detect oscillation which is used to better understand how this super abundant, and elusive particle interacts with matter.

Challenge
The experiment is built underground and is comprised of a main injector beam and two large detector apparatus located 50 miles apart. The near detector is 215 Tons and the Far detector is 15,000 Tons. The experiment can be thought of as a 30 Mn pound detector that takes 2 Mn pictures per second.

The detectability of the current experiment is proportional to the size of the detectors, so increasing the “visibility” is complex and costly.

Solution
A DNN was developed and trained using a data set derived from multiple HPC simulations including GENIE and GEANT using 2 k40 GPU's. the CVN was based on convolutional neural networks used for image processing.

Impact
The result was an overall improvement of 33% in detector mass, where the optimized CVN signal detection optimized efficiency of 49% is a significant gain over the efficiency of 35% quoted in prior art. This would yield a net effect of a 10Mn pound increase the physical detector mass.
AN AI MONITOR OF EARTH’S VITALS

The Earth’s climate has changed throughout history, but in recent years there have been record increases in temperature, glacial retreat and rising sea levels. NASA Ames is using satellite imagery to measure the effects of carbon and greenhouse gas emissions on the planet. To do so, they developed DeepSat – a deep learning framework for satellite image classification trained on a GPU-powered supercomputer. The enhanced satellite imagery will help scientists plan to protect ecosystems and farmers improve crop production.
AI Sheds Light on Key Mysteries of the Universe

Gravitational Lensing generates an image of a faraway galaxy that is distorted into rings and arcs by the gravity of a massive object, such as galaxy cluster. The distortions provide important clues about how mass is distributed in space and how that distribution changes over time, to measure properties of dark matter, size of galaxies, and the expansion of the universe. Today there are roughly 200 known gravitational lenses. However, scientists expect to uncover 200,000 gravitational lenses in the next decade. Using traditional approaches, it can take anywhere from 2 days to 3 months to analyze a single lens image.

SLAC scientists have for the first time used artificial neural networks to analyze gravitational lenses in just 10ms. This will allow researchers to unlock many of the unsolved mysteries of the universe.
AI Accelerates the Production of Ultra Cold Gases

Bose-Einstein Condensate (BEC) is a state of matter formed by cooling a gas to near-zero absolute temperature. BEC is achieved by controlling the intensity of the lasers to trap only the ultra-cold atoms and allowing other atoms to escape. BECs are super sensitive to external disturbances. This makes them suitable for very precise measurements of things like tiny changes in Earth’s magnetic field or gravity.

Researchers at the University of North South Whales used AI to create a BEC gas 14 times faster than conventional methods.
Background
Unexpected fog can cause an airport to cancel or delay flights, sometimes having global effects in flight planning.

Challenge
While the weather forecasting model at MeteoSwiss work at a 2km x 2km resolution, runways at Zurich airport is less than 2km. So human forecasters sift through huge simulated data with 40 parameters, like wind, pressure, temperature, to predict visibility at the airport.

Solution
MeteoSwiss is investigating the use of deep learning to forecast type of fog and visibility at sub-km scale at Zurich airport.

Impact
Forecasting Fog at Zurich Airport
Multiple Examples of AI for earthquake prediction are underway. Earthquake Prediction

Can Artificial Intelligence Predict Earthquakes?

The ability to forecast tremors would be a seismic shift in seismology. But is it a pipe dream? A seismologist is conducting machine-learning experiments to find out.

MIT

PRINCETON UNIVERSITY

SCEC

an NSF + USGS center

Stanford University