NVIDIA TO ACCELERATE THE HPC-AI CONVERGENCE

Workshop on the Convergence of ML & HPC
Gunter Roeth
gunter.r@nvidia.com
March 2020
GRAND CHALLENGES REQUIRE MASSIVE COMPUTING

REINVENTING THE LI-ION BATTERY
3M Node Hours | 7 Days on Titan

CLOUD RESOLVING CLIMATE SIMULATIONS
100M Node Hours | 840 Days on Piz Daint

UNDERSTANDING HIV’S STRUCTURE
10M node Hours | 16 Days on BlueWaters
SOMETHING NEW:
AI + HPC = REVOLUTION
BIG DATA IN SCIENCE
Big Science ingests/outputs Big Data

Large Hadron Collider

Johns Hopkins Turbulence Database

Square Kilometer Array
AI WORKFLOW FOR HPC

DATA

- SIMULATION (FP64/FP32)
- TRAINING (FP32/FP16)
- REGRESSION TESTING (FP16/INT8)
- INFERENC (FP16/INT8)

DATA

- REGRESSION SET
- NEW DATA

ERRORS
THE CONVERGENCE OF HPC * AI
Integrating the Third and Fourth Pillars of Scientific Discovery

HPC
40+ years of algorithms based on first principles theory

AI
New algorithms and models with potential to increase model size and accuracy

Dramatically Improves Accuracy and /or Time-to-Solution at Large Scale

- Commercially viable fusion energy
- Improve or validate the Standard Model of Physics
- Clinically viable precision medicine
- Understanding cosmological dark energy and matter
- Climate/weather forecasts with ultra-high fidelity
AI FOR HPC
Transformative Tool To Accelerate The Pace of Scientific Innovation

- **90% accuracy**
  - Fusion Sustainment
  - Clean Energy

- **33% Faster**
  - Track Neutrinos
  - Particle Physics

- **5,000X Faster**
  - Process LIGO Signal
  - Understanding Universe

- **300,000X Faster**
  - Predict Molecular Energetics
  - Drug Discovery

- **70% accuracy**
  - Score Protein Ligand
  - Drug Discovery

- **11% higher accuracy**
  - Monitor Earth’s Vital Climate

- **14X Faster**
  - Generate Bose-Einstein Condensate (Physics)

- **Weeks to 10 milliseconds**
  - Analyze Gravitational Lensing
  - Astrophysics

- **33% Faster**
  - Track Neutrinos
  - Particle Physics

- **70% accuracy**
  - Score Protein Ligand
  - Drug Discovery

- **11% higher accuracy**
  - Monitor Earth’s Vital Climate

- **14X Faster**
  - Generate Bose-Einstein Condensate (Physics)

**Improves Accuracy**
Enabling realization of full scientific potential

**Accelerates Time to Solution**
Unlocking the use of science in exciting new ways
INTELLIGENT HPC
DL Driving Future HPC Breakthroughs

- Trained networks as solvers
- Super-resolution of coarse simulations
- Low- and mixed-precision
- Simulation for training, network in production

Pre-processing
- Select/classify/augment/distribute input data
- Control job parameters

Simulation

Post-processing
- Analyze/reduce/augment output data
- Act on output data

From calendar time to real time?
THE SHAPE OF AI SUPERCOMPUTING
VOLTA TENSOR CORE GPU FUELS WORLD'S FASTEST SUPERCOMPUTER
Fused HPC and AI Computing In a Unified Platform

Genomics (CoMet)
World’s First Exascale Run
Finding Genes-to-disease Connection
Same accuracy as FP64 w/ Tensor Core

Quantum Chemistry (QMCPack)
Simulate New Materials
High-Temperature Semiconductors

150X Over Titan
50X Over Titan

Summit Supercomputer
Oakridge National Labs
AI: 3 Exaflops
HPC: 122 Petaflops

Measured performance: Summit node vs Titan node
AI: A NEW COMPUTING PARADIGM

DL AND HPC - JOINTLY SOLVE NEW PROBLEMS, BETTER
AI SUPERCOMPUTING IS HERE
Extending The Reach of HPC By Combining Computational & Data Science

S8242 – DL for Computational Science, Jeff Adie & Yang Juntao
Presented ~20 Success Stories of DL in Computational Science
(GTC on-demand: http://on-demand-gtc.gputechconf.com)
Computational Chemistry
AI Quantum Breakthrough

Background
Developing a new drug costs $2.5B and takes 10-15 years. Quantum chemistry (QC) simulations are important to accurately screen millions of potential drugs to a few most promising drug candidates.

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

Solution
Researchers at the University of Florida and the University of North Carolina leveraged GPU deep learning to develop ANAKIN-ME, to reproduce molecular energy surfaces with super speed (microseconds versus several minutes), extremely high (DFT) accuracy, and at 1-10/millionths of the cost of current computational methods. Essentially the DL model is trained to learn Hamiltonian of the Schrodinger equation.

Impact
Faster, more accurate screening at far lower cost
NEURAL NETWORK MODEL APPROACH

Training set: ~20M DFT data points.
Molecules with 1 to 8 atoms from GDB database


![Diagram showing accuracy and time for different methods like ANI-1 Potential, DFT & HF, Semi-empirical QM, Force fields, and CCSD(T).]
Computational Mechanics
FEA UPDATED WITH NEURAL NETWORK

FEA trained deep neural network for surrogate modelling of estimated stress distribution. Deepvirtuality, a spinoff from Volkswagen Data:Lab under Nvidia Inception Program has demonstrate with their software aimed for a quicker prediction of structural data.

An demonstration of Structure Born Noise of a V12 Engine with Deepvirtuality
Torsional Frequencies of a Car Body by Deepvirtuality
Eulerian Fluid Simulation
Approximating PDE solutions

Algorithm 1 Euler Equation Velocity Update

1: Choose a time-step $\Delta t$
2: Advection and Force Update to calculate $u_t^*$:
3: Advect scalar components through $u_{t-1}$
4: Self-advection velocity field $u_{t-1}$
5: Add external forces $f_{body}$
6: Add vorticity confinement force $f_{vc}$
7: Pressure Projection to calculate $u_t$:
8: Solve Poisson eqn to find $p_t$ that satisfies $\nabla \cdot u_t = 0$
9: Apply velocity update $u_t = \Delta u_t + u_{t-1}$
10: Modify velocity to satisfy boundary conditions.

"Accelerating Eulerian Fluid Simulation With Convolutional Networks", Thompson et al., 2016
SimRes at SC19 in Denver
Physics Informed NN
Vortex induced vibrations problem of flow past a circular cylinder. (eta)

Incompressible Navier-Stokes equations
Train first Data Driven Networks (DDNN)
AUTOMOTIVE AERODYNAMICS
DATA DRIVEN METHODS

Pros & Cons

- Need to generate a lot of Simulations (accuracy dependent on the simulation code)
- No Physics Awareness; Generalizability may be limited
- Not very efficient for Complex 3D Geometries/Curved Surfaces
- Interpolation/Extrapolation Errors

+ Not dependent on Physics
Use Physics Informed NN (PINNs)
A Neural Network Architecture for Computational Mechanics/Physics problems

- Point Cloud for 3D Geometries
- Physics Driven & Physics Aware Networks (respects the governing PDEs, Multi-disciplinary)
- Performance optimized for GPU tensor cores
PHYSICS DRIVEN METHODS

Special Considerations

- **Problem Modeling:**
  - Complex Geometries

- **Sampling Insensitivity**

- **Network Architecture:**
  - Faithfully represent the Physics with initial & boundary conditions
  - Architectural Requirements for $n^{th}$ order derivatives
  - Loss Convergence Acceleration
  - Activation Functions
  - Gradients & Discontinuities
  - Global vs. Local
Results of Physics Informed NN (PINNs)
STEADY STATE: 2D LID DRIVEN CAVITY SIMNET VS. OPENFOAM

U velocity difference = 0.2%
V velocity difference = 0.4%
Heat Sink -
* Temperatures to not exceed the design criteria

Objectives -
* Similar accuracy as the Solver
* Geometry representation with Point Clouds
* Multiple simultaneous parametrized & unparametrized geometries

Physics involved - CFD & Heat Transfer

*Ansys IcePack used for Simulation (**) we kindly acknowledge Ansys’s support **)
NETWORK ARCHITECTURE
Multi-Physics Neural Networks

Multi-Physics PDEs
CFD (with turbulence) - 2\textsuperscript{nd} Order PDE
Heat Transfer in Solids & Fluid

Fluid-Solid Interface Conditions
\[ \theta^f = \theta^s \]
\[ \kappa^f (\theta^f_x n_x + \theta^f_y n_y) = \kappa^s (\theta^s_x n_x + \theta^s_y n_y) \]

PINN Network Architecture
10 layers for non-Physics Informed Network
10 \times 2n layers for n\textsuperscript{th} order PDEs
50-500 neurons per layer
Swish Activation Function
HEATSINK DESIGN OPTIMIZATION
Physics Informed Neural Net for Coupled CFD-Heat Transfer Problems
Earth Science
ANOMALY DETECTION IN CLIMATE DATA

Identifying “extreme” weather events in multi-decadal datasets with 5-layered Convolutional Neural Network. Reaching 99.98% of detection accuracy. (Kim et al, 2017)

Dataset: Visualization of historic cyclones from JWTC hurricane report from 1979 to 2016

Systemic framework for detection and localization of extreme climate event
EMULATING RRTMG WITH DEEP NEURAL NETWORKS FOR THE ENERGY EXASCALE EARTH SYSTEM MODEL

- Rapid Radiation Transfer Model for GCMs (RRTMG) is the most time-consuming component of General Circulation Models (GCMs).
- Oak Ridge National Laboratory made use of Deep Neural Network to learn from RRTMG model.
Computational Physics
ACCELERATING MODELS FOR ACCELERATORS

Challenge
The Full Order GEANT Simulation is used to model the CERN LHC, NoVA, DUNE and other High Energy Particle Physics Experiments

The GEANT simulation is over 10Mn lines of C++ code with a flat execution profile that takes hours/days to simulate an experiment run, so each experiment uses a Reduced Order Fast Sim

Solution
A Generative Adversarial Network (CaloGAN) was trained on Fast Simulation data and compared against “ground truth” using GEANT output

Impact
The CaloGAN was shown to be 5 Orders of Magnitude faster than the FAST Simulation and nearly as accurate as the GEANT ground truth
Deep learning method named deep filtering was used in the first detection of gravitational wave. Numerical simulated data was used for training deep filtering, a convolutional neural network to replace matched filtering. It provided 20X speed up on single core and potential to be accelerated further with GPU.
HUNTING “GHOST PARTICLES” WITH A BETTER TOOL

Challenge
Neutrino detection experiments are massive and extremely expensive to build where detection sensitivity is directly proportional to size

Solution
Convolutional Neural Net (ImageNet with layers removed)
Trained with data from Full Order Models like GEANT and GENIE
Validated with Full Order Model Data

Impact
Accuracy increased by 33% as a POC, and then improved with further tuning to 49%
An equivalent increase of 15Mn pounds of detector mass
MAKING COMMERCIALLY Viable Fusion Energy Possible

Challenge
Accurate prediction of plasma disruption with enough lead time to shut down or modify the reactor

Solution
Recurrent Neural Net (Custom for Fusion)
Trained and validated with data from Joint European Tokamak (JET)

Impact
Accuracy is >95% with <5% False Alarm with 60ms lead time with higher accuracy at 10ms lead time

Next step: add higher order classifiers to training set, and model improved actuator response time
### NVIDIA ECOSYSTEM FOR CONVERGED METHODS

- **Optimized Applications:** 600+ Apps and 50 Containerized Apps and DL Frameworks....
- **Compilers and 3rd Party Libraries:** PGI, GCC, KOKKOS, RAJA, MAGMA, PETSC, All DL Frameworks....

#### Workflow Support:
- RAPIDS, MERLIN....

<table>
<thead>
<tr>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Algebra</td>
</tr>
<tr>
<td>Parallel Algorithms</td>
</tr>
<tr>
<td>Signal Processing</td>
</tr>
<tr>
<td>Deep Learning</td>
</tr>
<tr>
<td>Machine Learning</td>
</tr>
<tr>
<td>Visualization</td>
</tr>
</tbody>
</table>

**CUDA-X HPC & AI**

<table>
<thead>
<tr>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop Development</td>
</tr>
<tr>
<td>Data Center</td>
</tr>
<tr>
<td>Supercomputers</td>
</tr>
<tr>
<td>GPU-Accelerated Cloud</td>
</tr>
</tbody>
</table>
Merci
gunterr@nvidia.com