Exascale Computing:
The Coming Integration of Simulation, Data and Machine Learning

Rick Stevens
Argonne National Laboratory
The University of Chicago

Crescat scientia; vita excolatur
Rick Stevens

Position: Associate Laboratory Director for Computing, Environment and Life Sciences, Argonne National Laboratory and Professor of Computer Science, University of Chicago

Research Background: Parallel Computing Software, Visualization, Automated Theorem Proving, Bioinformatics, Computer Architecture

Research Interests: Exascale Computing, Machine Learning for Medicine and Science, Neuroscience, Automation of Scientific Discovery

Personal Interests: Camping, Cooking, Drones, Politics, Philosophy, Natural History
Aurora 2021 (A21) The first US Exascale System

Architecture supports three ways of computing

- Large-scale Simulation (PDEs, traditional HPC)
- Data Intensive Applications (scalable science pipelines)
- Deep Learning and Emerging Science AI (training and inferencing)
Machine Learning Interest is Exploding
The Cartoon Form

Traditional Programming

Data → Computer → Output
Program → Computer

Machine Learning

Data → Computer
Output → Program “Model”

Training

New Data
New Output

Inferencing
Machine Learning is becoming a major element of scientific computing applications

Across the DOE lab system hundreds of examples are emerging

– From fusion energy to precision medicine
– Materials design
– Fluid dynamics
– Synthetic Biology
– Structural engineering
– Intelligent sensing
– Etc.
# Targets for Exascale Computers

<table>
<thead>
<tr>
<th>Simulation Applications</th>
<th>Big Data Applications</th>
<th>Deep Learning Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Materials Science</td>
<td>• APS Data Analysis</td>
<td>• Drug Response Prediction</td>
</tr>
<tr>
<td>• Cosmology</td>
<td>• HEP Data Analysis</td>
<td>• Scientific Image Classification</td>
</tr>
<tr>
<td>• Molecular Dynamics</td>
<td>• LSST Data Analysis</td>
<td>• Scientific Text Understanding</td>
</tr>
<tr>
<td>• Nuclear Reactor Modeling</td>
<td>• SKA Data Analysis</td>
<td>• Materials Property Design</td>
</tr>
<tr>
<td>• Combustion</td>
<td>• Metagenome Analysis</td>
<td>• Gravitational Lens Detection</td>
</tr>
<tr>
<td>• Quantum Computer Simulation</td>
<td>• Battery Design Search</td>
<td>• Feature Detection in 3D</td>
</tr>
<tr>
<td>• Climate Modeling</td>
<td>• Graph Analysis</td>
<td>• Street Scene Analysis</td>
</tr>
<tr>
<td>• Power Grid</td>
<td>• Virtual Compound Library</td>
<td>• Organism Design</td>
</tr>
<tr>
<td>• Discrete Event Simulation</td>
<td>• Neuroscience Data Analysis</td>
<td>• State Space Prediction</td>
</tr>
<tr>
<td>• Fusion Reactor Simulation</td>
<td>• Genome Pipelines</td>
<td>• Persistent Learning</td>
</tr>
<tr>
<td>• Brain Simulation</td>
<td></td>
<td>• Hyperspectral Patterns</td>
</tr>
<tr>
<td>• Transportation Networks</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Big Data Applications**
- APS Data Analysis
- HEP Data Analysis
- LSST Data Analysis
- SKA Data Analysis
- Metagenome Analysis
- Battery Design Search
- Graph Analysis
- Virtual Compound Library
- Neuroscience Data Analysis
- Genome Pipelines

**Deep Learning Applications**
- Drug Response Prediction
- Scientific Image Classification
- Scientific Text Understanding
- Materials Property Design
- Gravitational Lens Detection
- Feature Detection in 3D
- Street Scene Analysis
- Organism Design
- State Space Prediction
- Persistent Learning
- Hyperspectral Patterns
## Targets for Exascale Computers

<table>
<thead>
<tr>
<th>Simulation Applications</th>
<th>Big Data Applications</th>
<th>Deep Learning Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Materials Science</td>
<td>• APS Data Analysis</td>
<td>• Drug Response Prediction</td>
</tr>
<tr>
<td>• Cosmology</td>
<td>• HEP Data Analysis</td>
<td>• Scientific Image</td>
</tr>
<tr>
<td>• Molecular Dynamics</td>
<td>• LSST Data Analysis</td>
<td>Classification</td>
</tr>
<tr>
<td>• Nuclear Reactor Modeling</td>
<td>• SKA Data Analysis</td>
<td>• Scientific Text</td>
</tr>
<tr>
<td>• Combustion</td>
<td>• Metagenome Analysis</td>
<td>Understanding</td>
</tr>
<tr>
<td>• Quantum Computer Simulation</td>
<td>• Battery Design Search</td>
<td>• Materials Property Design</td>
</tr>
<tr>
<td>• Climate Modeling</td>
<td>• Graph Analysis</td>
<td>• Gravitational Lens</td>
</tr>
<tr>
<td>• Power Grid</td>
<td>• Virtual Compound</td>
<td>Detection</td>
</tr>
<tr>
<td>• Discrete Event Simulation</td>
<td>Library Generation</td>
<td>• Feature Detection in 3D</td>
</tr>
<tr>
<td>• Fusion Reactor Simulation</td>
<td>• Neuroscience Data</td>
<td>• Street Scene Analysis</td>
</tr>
<tr>
<td>• Brain Simulation</td>
<td>• Analysis</td>
<td>• Organism Design</td>
</tr>
<tr>
<td>• Transportation Networks</td>
<td>• Genome Pipelines</td>
<td>• State Space Prediction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Persistent Learning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Hyperspectral Patterns</td>
</tr>
<tr>
<td>Application</td>
<td>Summary</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>HACC</td>
<td>Particle/N-Body, FFT</td>
<td></td>
</tr>
<tr>
<td>LAMMPS</td>
<td>Classical MD</td>
<td></td>
</tr>
<tr>
<td>QMCPACK</td>
<td>Many-Body Theory</td>
<td></td>
</tr>
<tr>
<td>Nekbone</td>
<td>Unstructured Grids, Spectral Element</td>
<td></td>
</tr>
<tr>
<td>LSST-SVM</td>
<td>Multi-class classification or regression analysis using support vector machines. Datasets consistent with those expected from future cosmological surveys.</td>
<td></td>
</tr>
<tr>
<td>Tomography Reconstruction</td>
<td>Fourier Time method. Used 1D and 2D FFTs, interpolation, and approximation functions. Dataset includes images.</td>
<td></td>
</tr>
<tr>
<td>FCMA (SGEMM &amp; SSYRK)</td>
<td>Interactions among brain regions in functional magnetic resonance imaging Data is a stream of 3D human brain data (volumes of voxel) over time, 4D data</td>
<td></td>
</tr>
<tr>
<td>Candle Pilot 1 (P1B2, P1B3)</td>
<td>Convolution Neural Nets (CNN), Multilayer Perceptrons (MLP) Datasets include gene sequences, drug responses, drug descriptors.</td>
<td></td>
</tr>
<tr>
<td>Candle Pilot 3 (P3)</td>
<td>Hierarchical Attention Networks, Multi-task Learning Datasets include ontology reports</td>
<td></td>
</tr>
<tr>
<td>Imaging (Inference)</td>
<td>Convolution Neural Nets, GANs, MLP. Datasets include images and potentially experimental settings</td>
<td></td>
</tr>
</tbody>
</table>
Simulation Targets and Needs

Performance Goal: >50x over 20PF systems

Typical Targets
• Materials and Physics
• Cosmology and Astronomy
• Climate and Energy
• Chemistry and Fluids
• Particles and Fields
• Flow in Networks

Newish Targets
• Quantum Computer Simulation
• Brain Simulation

System Requirements
• Compute intensity (fp64)
• Memory Bandwidth (HBM2e)
• Memory Capacity (6-8PB)
• Large Coherency Domain (~TB)
• Unified Address Space CPU/Acc
• High Bisection Bandwidth (~PB/s)
• High Injection Bandwidth (~100’s GB/s)
• Lowish Comm Latency (~us)
• Sustained IO Performance (~10’s TB/s)
### Exascale Applications Will Address National Challenges

Summary of current DOE Science & Energy application development projects

<table>
<thead>
<tr>
<th>Nuclear Energy (NE)</th>
<th>Climate (BER)</th>
<th>Chemical Science (BES, BER)</th>
<th>Wind Energy (EERE)</th>
<th>Combustion (BES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerate design and commercialization of next-generation small modular reactors*</td>
<td>Accurate regional impact assessment of climate change*</td>
<td>Biofuel catalysts design; stress-resistant crops</td>
<td>Increase efficiency and reduce cost of turbine wind plants sited in complex terrains*</td>
<td>Design high-efficiency, low-emission combustion engines and gas turbines*</td>
</tr>
<tr>
<td>Climate Action Plan; SMR licensing support; GAIN</td>
<td>Climate Action Plan</td>
<td>Climate Action Plan; MGI</td>
<td>Climate Action Plan</td>
<td>2020 greenhouse gas and 2030 carbon emission goals</td>
</tr>
</tbody>
</table>

* Scope includes a discernible data science component
Exascale Applications Will Address National Challenges
Summary of current DOE Science & Energy application development projects

**Nuclear Energy (NE)**
Accelerate design and commercialization of next-generation small modular reactors*
Climate Action Plan; SMR licensing support; GAIN

**Climate (BER)**
Accurate regional impact assessment of climate change*
Climate Action Plan

**Chemical Science (BES, BER)**
Biofuel catalysts design; stress-resistant crops
Climate Action Plan; MGI

**Wind Energy (EERE)**
Increase efficiency and reduce cost of turbine wind plants sited in complex terrains*
Climate Action Plan

**Combustion (BES)**
Design high-efficiency, low-emission combustion engines and gas turbines*
2020 greenhouse gas and 2030 carbon emission goals

* Scope includes a discernible data science component
### Summary of current DOE Science & Energy application development projects

**Materials Science (BES)**
- Find, predict, and control materials and properties: property change due to hetero-interfaces and complex structures
  - MGI

**Nuclear Physics (NP)**
- QCD-based elucidation of fundamental laws of nature: SM validation and beyond SM discoveries
- 2015 Long Range Plan for Nuclear Science; RHIC, CEBAF, FRIB

**Nuclear Materials (BES, NE, FES)**
- Extend nuclear reactor fuel burnup and develop fusion reactor plasma-facing materials*
- Climate Action Plan; MGI; Light Water Reactor Sustainability; ITER; Stockpile Stewardship Program

**Accelerator Physics (HEP)**
- Practical economic design of 1 TeV electron-positron high-energy collider with plasma wakefield acceleration*
- >30k accelerators today in industry, security, energy, environment, medicine

**Materials Science (BES)**
- Protein structure and dynamics; 3D molecular structure design of engineering functional properties*
  - MGI; LCLS-II 2025 Path Forward

* Scope includes a discernible data science component
Exascale Applications Will Address National Challenges
Summary of current DOE Science & Energy application development projects

Materials Science (BES)
Find, predict, and control materials and properties: property change due to hetero-interfaces and complex structures
MGI

Nuclear Physics (NP)
QCD-based elucidation of fundamental laws of nature: SM validation and beyond SM discoveries
2015 Long Range Plan for Nuclear Science; RHIC, CEBAF, FRIB

Nuclear Materials (BES, NE, FES)
Extend nuclear reactor fuel burnup and develop fusion reactor plasma-facing materials*
Climate Action Plan; MGI; Light Water Reactor Sustainability; ITER; Stockpile Stewardship Program

Accelerator Physics (HEP)
Practical economic design of 1 TeV electron-positron high-energy collider with plasma wakefield acceleration*
>30k accelerators today in industry, security, energy, environment, medicine

Materials Science (BES)
Protein structure and dynamics; 3D molecular structure design of engineering functional properties*
MGI; LCLS-II 2025 Path Forward

* Scope includes a discernible data science component
### Exascale Applications Will Address National Challenges

Summary of current DOE Science & Energy and Other Agency application development projects

<table>
<thead>
<tr>
<th>Magnetic Fusion Energy (FES)</th>
<th>Advanced Manufacturing (EERE)</th>
<th>Cosmology (HEP)</th>
<th>Geoscience (BES, BER, EERE, FE, NE)</th>
<th>Precision Medicine for Cancer (NIH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict and guide stable ITER operational performance with an integrated whole device model*</td>
<td>Additive manufacturing process design for qualifiable metal components*</td>
<td>Cosmological probe of standard model (SM) of particle physics: Inflation, dark matter, dark energy*</td>
<td>Safe and efficient use of subsurface for carbon capture and storage, petroleum extraction, geothermal energy, nuclear waste*</td>
<td>Accelerate and translate cancer research in RAS pathways, drug responses, treatment strategies*</td>
</tr>
<tr>
<td>ITER; fusion experiments: NSTX, DIII-D, Alcator C-Mod</td>
<td>NNMIs; Clean Energy Manufacturing Initiative</td>
<td>Particle Physics Project Prioritization Panel (P5)</td>
<td>EERE Forge; FE NRAP; Energy-Water Nexus; SubTER Crosscut</td>
<td>Precision Medicine in Oncology; Cancer Moonshot</td>
</tr>
</tbody>
</table>

* Scope includes a discernible data science component
### Exascale Applications Will Address National Challenges

Summary of current DOE Science & Energy and Other Agency application development projects

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
</table>
| **Magnetic Fusion Energy (FES)** | Predict and guide stable ITER operational performance with an integrated whole device model*  
                                 | ITER; fusion experiments: NSTX, DIII-D, Alcator C-Mod                                                                                      |
| **Advanced Manufacturing (EERE)** | Additive manufacturing process design for qualifiable metal components*  
                                  | NNMIs; Clean Energy Manufacturing Initiative                                                                                             |
| **Cosmology (HEP)**             | Cosmological probe of standard model (SM) of particle physics: Inflation, dark matter, dark energy*  
                                  | Particle Physics Project Prioritization Panel (P5)                                                                                       |
| **Geoscience (BES, BER, EERE, FE, NE)** | Safe and efficient use of subsurface for carbon capture and storage, petroleum extraction, geothermal energy, nuclear waste*  
                                      | EERE Forge; FE NRAP; Energy-Water Nexus; SubTER Crosscut                                                                                   |
| **Precision Medicine for Cancer (NIH)** | Accelerate and translate cancer research in RAS pathways, drug responses, treatment strategies*  
                                              | Precision Medicine in Oncology; Cancer Moonshot                                                                                           |

* Scope includes a discernible data science component
Exascale Applications Will Address National Challenges
Summary of current DOE Science & Energy application development seed projects

<table>
<thead>
<tr>
<th>Seismic (EERE, NE, NNSA)</th>
<th>Carbon Capture and Storage (FE)</th>
<th>Chemical Science (BES)</th>
<th>Urban Systems Science (EERE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable earthquake hazard and risk assessment in relevant frequency ranges*</td>
<td>Scaling carbon capture/storage laboratory designs of multiphase reactors to industrial size</td>
<td>Design catalysts for conversion of cellulosic-based chemicals into fuels, bioproducts</td>
<td>Retrofit and improve urban districts with new technologies, knowledge, and tools*</td>
</tr>
<tr>
<td>DOE Critical Facilities Risk Assessment; urban area risk assessment; treaty verification</td>
<td>Climate Action Plan; SunShot; 2020 greenhouse gas/2030 carbon emission goals</td>
<td>Climate Action Plan; SunShot Initiative; MGI</td>
<td>Energy-Water Nexus; Smart Cities Initiative</td>
</tr>
</tbody>
</table>

* Scope includes a discernible data science component
# Exascale Applications Will Address National Challenges

Summary of current DOE Science & Energy application development seed projects

<table>
<thead>
<tr>
<th>Seismic (EERE, NE, NNSA)</th>
<th>Carbon Capture and Storage (FE)</th>
<th>Chemical Science (BES)</th>
<th>Urban Systems Science (EERE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable earthquake hazard and risk assessment in relevant frequency ranges*</td>
<td>Scaling carbon capture/storage laboratory designs of multiphase reactors to industrial size</td>
<td>Design catalysts for conversion of cellulosic-based chemicals into fuels, bioproducts</td>
<td>Retrofit and improve urban districts with new technologies, knowledge, and tools*</td>
</tr>
<tr>
<td>DOE Critical Facilities Risk Assessment; urban area risk assessment; treaty verification</td>
<td>Climate Action Plan; SunShot; 2020 greenhouse gas/2030 carbon emission goals</td>
<td>Climate Action Plan; SunShot Initiative; MGI</td>
<td>Energy-Water Nexus; Smart Cities Initiative</td>
</tr>
</tbody>
</table>

* Scope includes a discernible data science component
Exascale Applications Will Address National Challenges
Summary of current DOE Science & Energy application development seed projects

**Metagenomics (BER)**
- Leveraging microbial diversity in metagenomic datasets for new products and life forms*
- Climate Action Plan; Human Microbiome Project; Marine Microbiome Initiative

**Astrophysics (NP)**
- Demystify origin of chemical elements (> Fe); confirm LIGO gravitational wave and DUNE neutrino signatures*
- 2015 Long Range Plan for Nuclear Science; origin of universe and nuclear matter in universe

**Power Grid (EERE, OE)**
- Reliably and efficiently planning our nation’s grid for societal drivers: rapidly increasing renewable energy penetration, more active consumers*
  - Grid Modernization Initiative; Climate Action Plan

* Scope includes a discernible data science component
Exascale Applications Will Address National Challenges
Summary of current DOE Science & Energy application development seed projects

Metagenomics (BER)
Leveraging microbial diversity in metagenomic datasets for new products and life forms*
Climate Action Plan; Human Microbiome Project; Marine Microbiome Initiative

Astrophysics (NP)
Demystify origin of chemical elements (> Fe); confirm LIGO gravitational wave and DUNE neutrino signatures*
2015 Long Range Plan for Nuclear Science; origin of universe and nuclear matter in universe

Power Grid (EERE, OE)
Reliably and efficiently planning our nation’s grid for societal drivers: rapidly increasing renewable energy penetration, more active consumers*
Grid Modernization Initiative; Climate Action Plan

* Scope includes a discernible data science component
Data Analysis Targets and Needs

Performance Goal: >50x over 20PF systems

Instrument data
- APS data streams
- Connectome Imaging
- LSST data streams
- ARM radar data streams
- HEP detector data streams

Simulation data
- Climate output analysis
- Cosmology output analysis
- Detector simulation output analysis

System Requirements
- Streaming IO bandwidth (~10’s TB/s)
- Large RAM for in memory (~PBs)
- Large coherency domain (~TB)
- Flexible control flow (CPU+X)
- Flexible SW Stack (x86+ Containers)
- Workflow support (Spark, Swift)
- Scripting Language Support (Python)
- Accelerated xform libraries
- Embedded machine learning
- Embedded simulation
APS data volumes already equal those from LHC

Source: Francesco de Carlo
Local flows already exceed those of LHC

Argonne data flows in TB/day (estimates)

Advanced Photon Source

Argonne Leadership Computing Facility

Other sources that remain to be quantified
Reconstructing Brain Connectivity from stacked EM images
Machine Learning Targets and Needs
Performance Goal: >50x over 20PF systems

Targets
• Materials (Energy Storage)
• Dark Matter (Lensing)
• Organism Design
• Advanced Manufacturing
• Cancer Therapeutics (Drugs)
• Traumatic Brain Injury
• Genomics (AMR, CVD)
• Software Design/Improvement
• Theory Integration with ML

Systems Requirements
• Compute Intensity (fp32, bfp16)
• Large Coherency domain (~TB)
• Large Memory (~PB)
• Unified Address Space CPU/Acc
• Data Sparsity Support (S/G)
• Sustained IO Performance (~10’s TB/s)
• High Injection Bandwidth (~100’s GB/s)
• Flexible SW Stack (x86+ Containers)
• Workflow support (Spark, Swift)
• Scripting Language Support (Python)
Personalized Cancer Therapy

1. Molecular Profiling

2. Prognostic Markers
   - Markers predictive of drug sensitivity/resistance
   - Markers predictive of adverse events
Modeling Cancer Drug Response

\[ R = f(T, D_1, D_2) \]

- **Drug(s)**
  - descriptors
  - fingerprints
  - structures
  - SMILES
  - dose

- **Response**
  - IC50
  - GI50
  - % growth
  - Z-score

- **Tumor**
  - gene expression levels
  - SNPs
  - protein abundance
  - microRNA
  - methylation

- **Drug Concentration in Log scale**
Machine Learning Models with UQ

[High-]Throughput Experiments

Interesting Biology

Model Uncertainty
CANDLE: Deep Learning Meets HPC

Exascale Needs for Deep Learning
• Automated Model Discovery
• Hyper Parameter Optimization
• Uncertainty Quantification
• Flexible Ensembles
• Cross-Study Model Transfer
• Data Augmentation
• Synthetic Data Generation
• Reinforcement Learning
ECP-CANDLE: CANcer Distributed Learning Environment

CANDLE Goals

- Develop an exscale deep learning environment for cancer and DOE mission applications
- Build on open source deep learning frameworks
- Optimize for CORAL and exascale platforms
- Support all three Cancer pilot project needs for deep learning
- Collaborate with DOE computing centers, HPC vendors and ECP co-design and software technology projects
CANDLE Software Stack

Hyperparameter Sweeps, Data Management (e.g. DIGITS, Swift, etc.)

Network description, Execution scripting API (e.g. Keras, Mocha)

Tensor/Graph Execution Engine (e.g. Theano, TensorFlow, LBANN-LL, etc.)

Architecture Specific Optimization Layer (e.g. cuDNN, MKL-DNN, etc.)

Workflow
Scripting
Engine
Optimization
CANDLE System Architecture

CANDLE Supervisor

Workflow Manager (Swift-T EMEWS)

Hyperparameter Optimization Frameworks
Hyperopt, mlrMBO, Spearmint

CANDLE Specifications
Benchmark Spec
Hyperparameter Spec
Hardware Spec

ML/DL Benchmarks
Pilot 1
Pilot 2
Pilot 3

CANDLE Database
Metadata Store
Model Store
Data API

Benchmarks
Datasets
Models
Experiments
Runs
Model Descriptions
Model Weights

Hardware Resources
ALCF
NERSC
OLCF
Theta, Cooley
Cori
Titan, SimmitDev

Integrator Website

CANDLE Specifications

Hyperparameter Optimization Frameworks
Hyperopt, mlrMBO, Spearmint

Hardware Resources
ALCF
NERSC
OLCF
Theta, Cooley
Cori
Titan, SimmitDev

Integrator Website
GitHub and FTP

• ECP-CANDLE GitHub Organization:
  • https://github.com/ECP-CANDLE

• ECP-CANDLE FTP Site:
  • The FTP site hosts all the public datasets for the benchmarks from three pilots
Integration of Simulation and ML

• Steering of simulations and Planning N moves ahead
  – ML/RL making decisions what to do next
• Embedding ML into Simulation
  – Replacing explicit functions/kernels with learned models
  – Trading accuracy for speed/power improvements ±7% for 2x ?
• Tuning or Customization of Kernels and Parameters
  – Customization of force fields in MD simulations (most accurate H₂O sim)
• Function/Property association
  – VAE to map latent representation to properties and generating candidates
• Student Teacher Model for Learning
  – Augment training data with simulation generated ground truth
(a) Standard: student network learns from teacher guidance (soft loss) and ground truth (hard loss).
Teacher-Student Sim/Network Model

Revised Model: Student Learns from data, from hints from simulation and from enhanced Ground Truth.
Integrating ML and Simulation

Figure 3: Overview of how data at all steps will be integrated using machine learning. The orange square boxes represent the three types of data in this project: kinases, drugs, and their interactions at various levels. The green rounded boxes denote the variety of MD simulations for free energy calculation. Each blue arrow represents an ML model; they combine in a joint predictive model that integrates all datasets.
Not just Deep Learning

• Last five years many problems have fallen to deep neural networks and “convolutional neural networks” (representational learning)
• AI is also advancing in “Re-enforcement Learning” where the reward is delayed
  – Game Playing, Stock Portfolios, Planning, Preference Systems, etc.
• Re-enforcement learning uses an intermediate entity called an "agent" driven by a “policy” to evaluate “moves” in the “game”. Policy gets updated based on delayed feedback. Policy needs to give a good estimate of the value of alternative moves. Agent/Policy plays the game, machine learning improves the policy. Policies themselves might be machine learning based.
Figure 1: An illustration of DeepCube. The training and solving process is split up into ADI and MCTS. First, we iteratively train a DNN by estimating the true value of the input states using breadth-first search. Then, using the DNN to guide exploration, we solve cubes using Monte Carlo Tree Search. See methods section for more details.
The New HPC + AI “Paradigm”

- Simulation
- Data Analysis
- Machine Learning
- Visualization
Exascale Programming

• MPI + X { where X could be OpenMP, CUDA, other (UPC) + libraries }
• C/C++ is likely ahead of Fortran for software support
• Think ~10K separate address spaces via message passing
• Think ~100 CPU threads per address space
• Think 4-16 accelerator contexts per node address space
• Some support for data objects + traditional filesystem I/O
• Manual checkpointing (system MTBF < ~4 days)
• In most cases > 90% of the flops are in the accelerator
• In most cases > 4x-8x of the memory bandwidth is in the accelerator
Exascale Software Stack

• Single Unified stack with resource allocation and scheduling across all pillars and ability for frameworks and libraries to seamlessly compose
• Minimize data movement: keep permanent data in the machine via distributed persistent memory while maintaining availability requirements
• Support standard file I/O and path to memory coupled model for Sim, Data and Learning
• Isolation and reliability for multi-tenancy and combining workflows
Towards an Integrated Sim, Data, Learn Stack

- HPC, Analytics and Big Data, AI and Machine Learning
- **Domain Platform Abstractions**
  - HPC
  - Big Data Analytics
  - AI ML DL
- **Domain Runtime Environments**
  - (Domain-aware RM plug-ins)
- **Global Resource Management**
  - Multi-domain Resource Manager
- **Resource Provisioning**
  - (Compute, Network, Storage)
    - Bare-metal Provisioning (e.g., xCAT, Warewulf, Ironic)
    - SDI Virtualized Provisioning (e.g., OpenStack, AWS, Azure, Google, Containers)
- **Infrastructure Abstractions**
  - Compute (Xeon, Xeon Phi, FPGA)
  - Storage Abstractions (e.g., POSIX, Object, Block, HDFS, DAOS)
  - Networking (OmniPath, Ethernet, IB)
- **Resource Pools**
  - (Public & Private)
    - Object Stores (e.g., RADOS (Ceph), AWS S3, Swift, Lustre OST)
Goal 2025 Automate and Accelerate

High Throughput Laboratory Synthesis, Characterization, Imaging

- Data Generation
- In Vitro and In Vivo Experiments
- Hypothesis Testing

High Performance Computing and Accelerated Machine Learning

- Data Analysis
- Simulation + Machine Learning
- Experimental Design + Uncertainty Quantification
Acknowledgements

Many thanks to DOE, NSF, NIH, DOD, ANL, UC, Moore Foundation, Sloan Foundation, Apple, Microsoft, Cray, Intel and IBM for supporting my research group over the years.