AI for HPC and HPC for AI : Lessons, Gaps and Challenges

@HPC User Forum

Rangan Sukumar, PhD
Office of the CTO, Cray Inc.

AI for HPC and HPC for AI @ Cray

CRAY URIKA-GX → CS-STORM 500NX → CS-STORM 500GT → CRAY XC SERIES → CRAY URIKA-XC

Integrated Analytics and AI platform for Data Preparation and Machine Learning

Dense GPU systems with broad support for NVIDIA® Tesla® Accelerators and FPGAs

Scalable high performance supercomputers with Analytics and AI/DL

Converging hardware, software and systems for performance and productivity

AI for HPC and HPC for AI: Customer Stories

Big Four Consulting Firm

Urika-GX

"Teaming with Cray was a clear choice and allows for versatility combined with speed to tackle big data problems"

-- Deloitte Advisory Cyber Risk Services

Top 5 Global Pharma

CS-Storm Dense GPU Cluster

Supporting core research and development in areas including chem-informatics and large machine and deep learning workloads

Samsung

Cray CS-Storm Dense GPU Cluster

Recent win to support machine and deep learning workloads including autonomous vehicles

The Stanford Research Computing Facility

Cray CS-Storm "XStream" Dense GPU Cluster

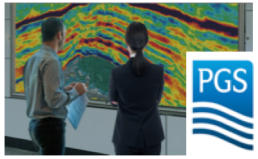
Research in astrophysics, structural biology and bioinformatics, materials modeling, and climate

Recent breakthrough in 3D deep convolutional neural networks for amino acid environment similarity analysis

Argonne National Laboratory

Cray XC40

Predicting how specific patients and tumors respond to different types of drugs using the scalable deep neural network code called CANDLE — or CANcer Distributed Learning Environment



Leading Seismic Processing Services Company

Cray XC40

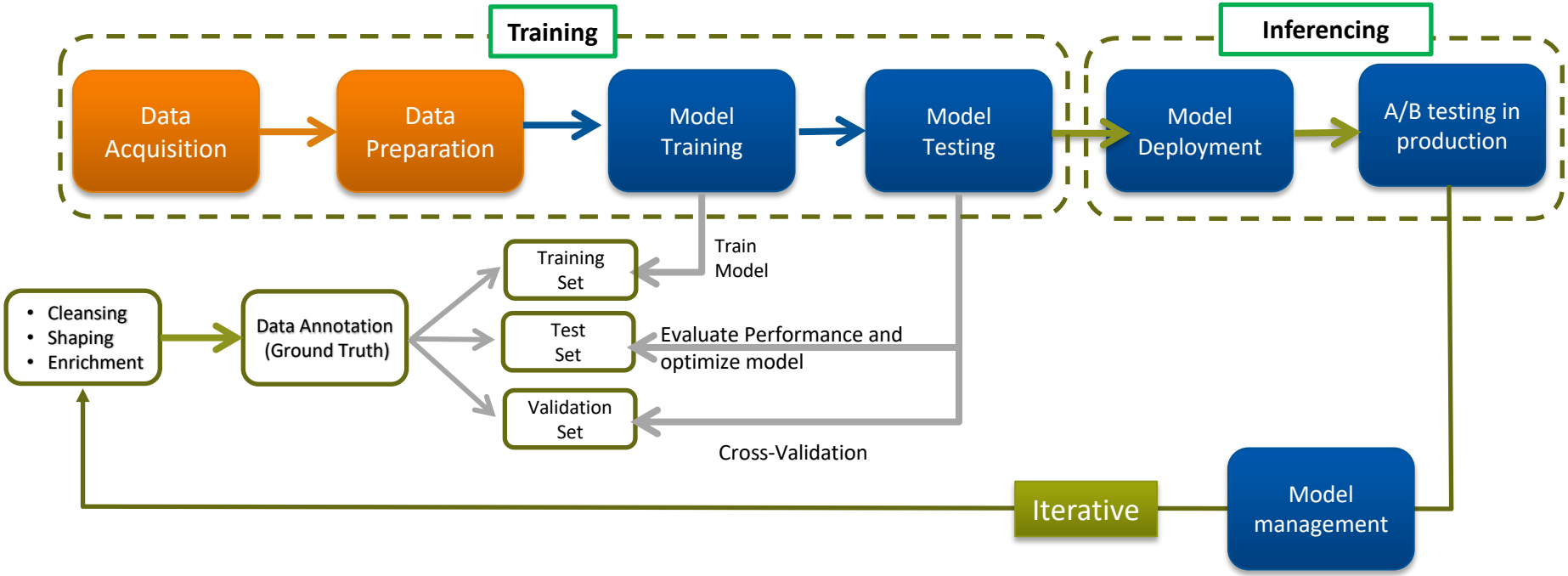
Machine Learning at Scale for Full Waveform Inversion

PGS applied machine learning optimization techniques such as regularization and steering to determine the velocity model in a Full Waveform Inversion seismic imaging workload.

AI for HPC and HPC for AI: Customer Journey

Axis of Maturity	Variables/Options	Example Use-Case	Figure-of-merit
Explainable Intelligence	Meta-learning, Lower-order physics approximation	Fraud Detection (Provenance is important)	**Interpretability**
System Architecture	Commodity Clusters, HPC Clusters, Supercomputers	Full-motion video analytics	**Scalability/ Performance**
Multi-node Architectures	Interconnects: InfiniBand, Ethernet, Proprietary	Autonomous Driving	**Throughput**
Node Architecture / Density	Processing units: CPUs, GPUs, Other Accelerators Density: # of units, # of sockets / unit, CPU:GPU ratio	Voice commands, Speech2Text	**Time-to-accuracy**
Infrastructure Investment	Workstation, Cloud-access, Co-location, On-premise	Insurance claim estimation from pictures, genotype-phenotype map	**Cost (Hardware efficiency)**
Hyper-parameter Optimization	Grid, Random, Bayesian, Evolutionary	Call-center automation, Chatbots	**Generalizability**
Network Topology	Deep, Convolution, Recurrent, Generative-Adversarial, Auto-encoders, Long-short-term memory networks etc.	ImageNet Challenge	**Accuracy (Statistical Efficiency)**
Toolkit Selection	TensorFlow, Caffe2, MXNet, CNTK, BigDL, etc.	Robotics class, computer vision	**Ease-of-use**
DL Problem Formulation	Training and Inference	Handwritten character recognition	**Time to Solution / ROI**

Enterprise Computing: AI Workflows Leveraging HPC



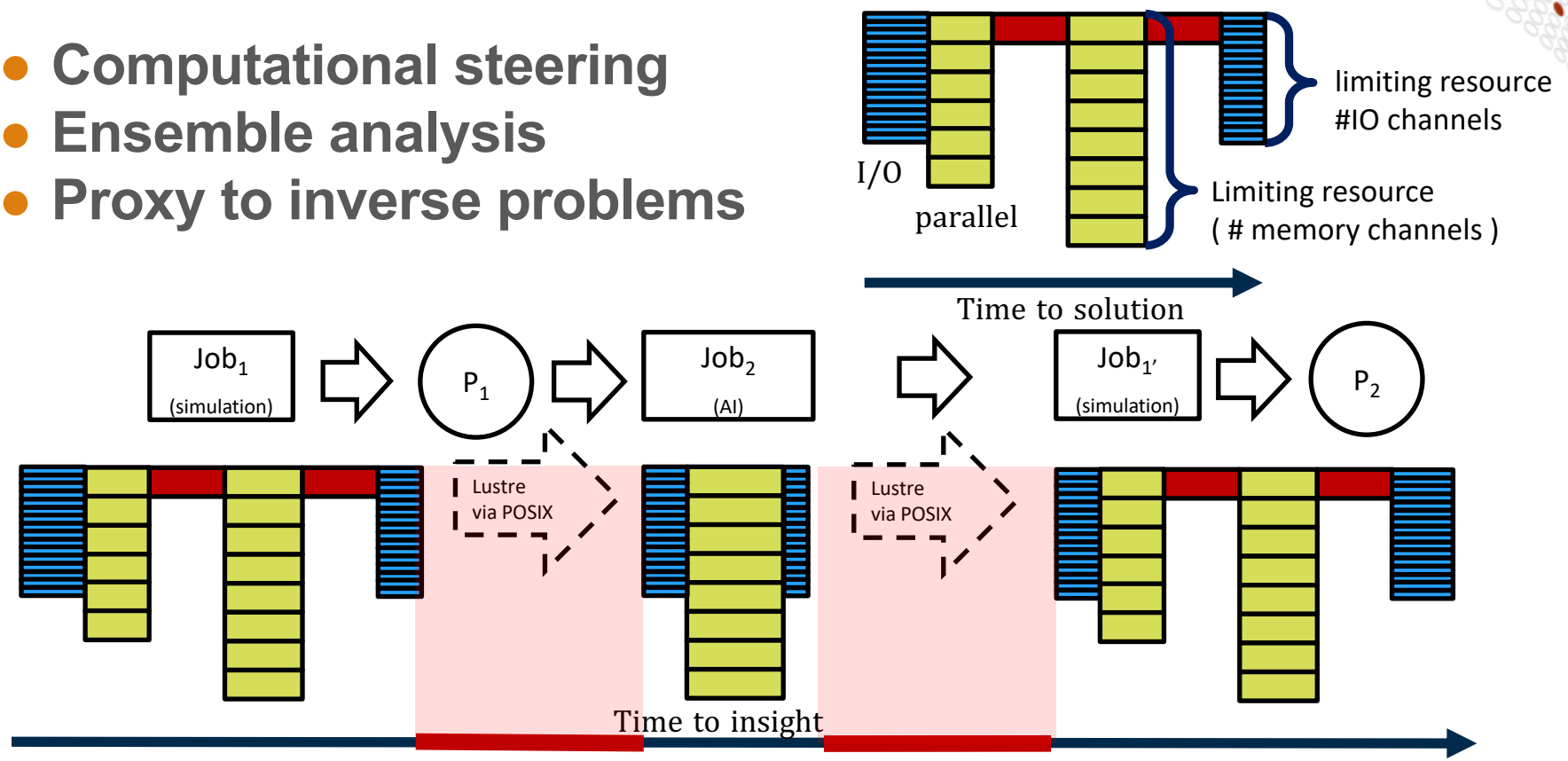
Enterprise Computing: Emerging Trends

Training patterns determine choice of supporting infrastructure for storage and i/o

Training	Use-case	Data size growth in unit time	Time to quality metric today	# of xPUs
Continuous	Internet-of-things	1:1	O(minutes)	O(10)
Cadence	Uber Eats prediction	n:1 (n>>1)	O(days)	O(10)
Delta	Speech (rare words)	n:1 (n~1)	O(days)	O(1)
One-time	Lower-order physics approximations	10-100n:1	O(weeks)	O(100+)
Throughput	Speech and speaker detection	1:# of users	O(days)	O(100+)

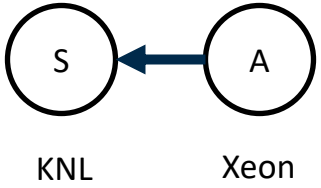
Scientific Computing: HPC Workflows Leveraging AI

- Computational steering
- Ensemble analysis
- Proxy to inverse problems

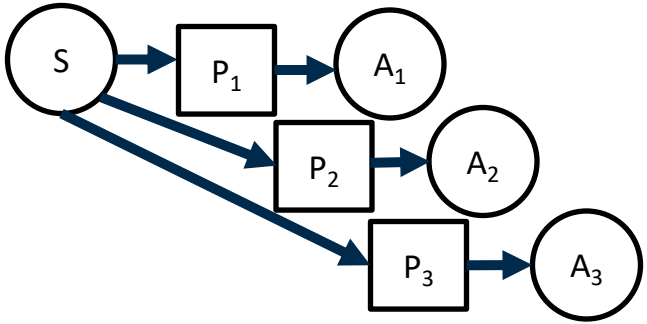


Scientific Computing: Creative Workflows

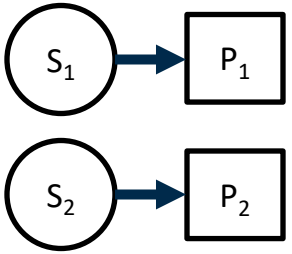
Async DAG-execution



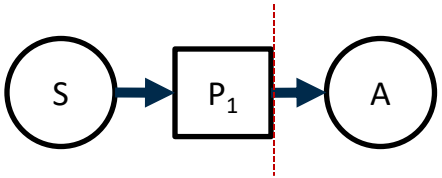
Dataflow semantics



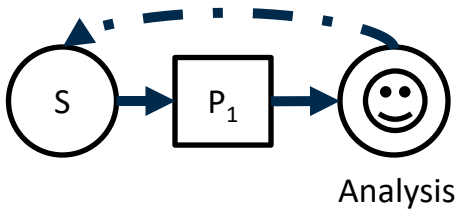
Data Management & Curation



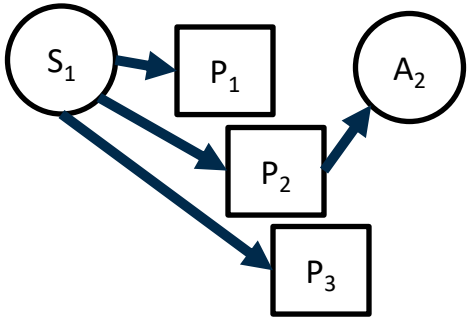
Data-driven notification



Human in-the-loop



Data-dependent workflow



Lessons/Differences: Architecture

	Scientific Computing	Enterprise Computing
Primarily used for	Solving equations	Search/Query, Machine learning
Philosophy	Send data to compute	Send compute to data
Efficiency via	Parallelism	Distribution
Scaling expectation	Strong (scale-up)	Weak (scale-out)
Programming model	MPI, OpenMP, etc.	Map-reduce, SPMD, etc.
Popular languages	FORTRAN, C++, Python	Java, Scala, Python, R
Design strength	Multi-node communication using an interconnect	Built-in job fault tolerance over Ethernet
Access model	On-premise	Cloud
Preferred algebra	Dense Linear	Set-theoretic / Relational
Memory access	Predictable	Random
Storage	Centralized, POSIX/RAID	Decentralized, Duplication

Lessons/Differences: Workflows+Workload

	Scientific Computing	Enterprise Computing
Data (Structured)	Vector, Matrix, Tensor	Table, Key-Values, Objects
Data (Unstructured)	Mesh, Images (Physics-based)	Documents, Images (Camera)
Visualization	Voxel, Surface, Point Clouds	Word Cloud, Parallel Coordinates, BI Tools
Validation	Cross-validation (ROC curves, statistical significance)	Manual / Subject matter expert, A/B testing
Extract, Transform, Load	Fourier, Wavelet, Laplace, etc. Cartesian, Radial, Toroidal, etc.	File-format transformations e.g. CSV to VRML
Search (Query)	Properties such as periodicity, self-similarity, anomaly, etc.	SQL, SPARQL, etc. (Sum, Average, Group by)
Funding Model	Non-profit grand challenge (Answer matters)	Value-driven (Cost matters)

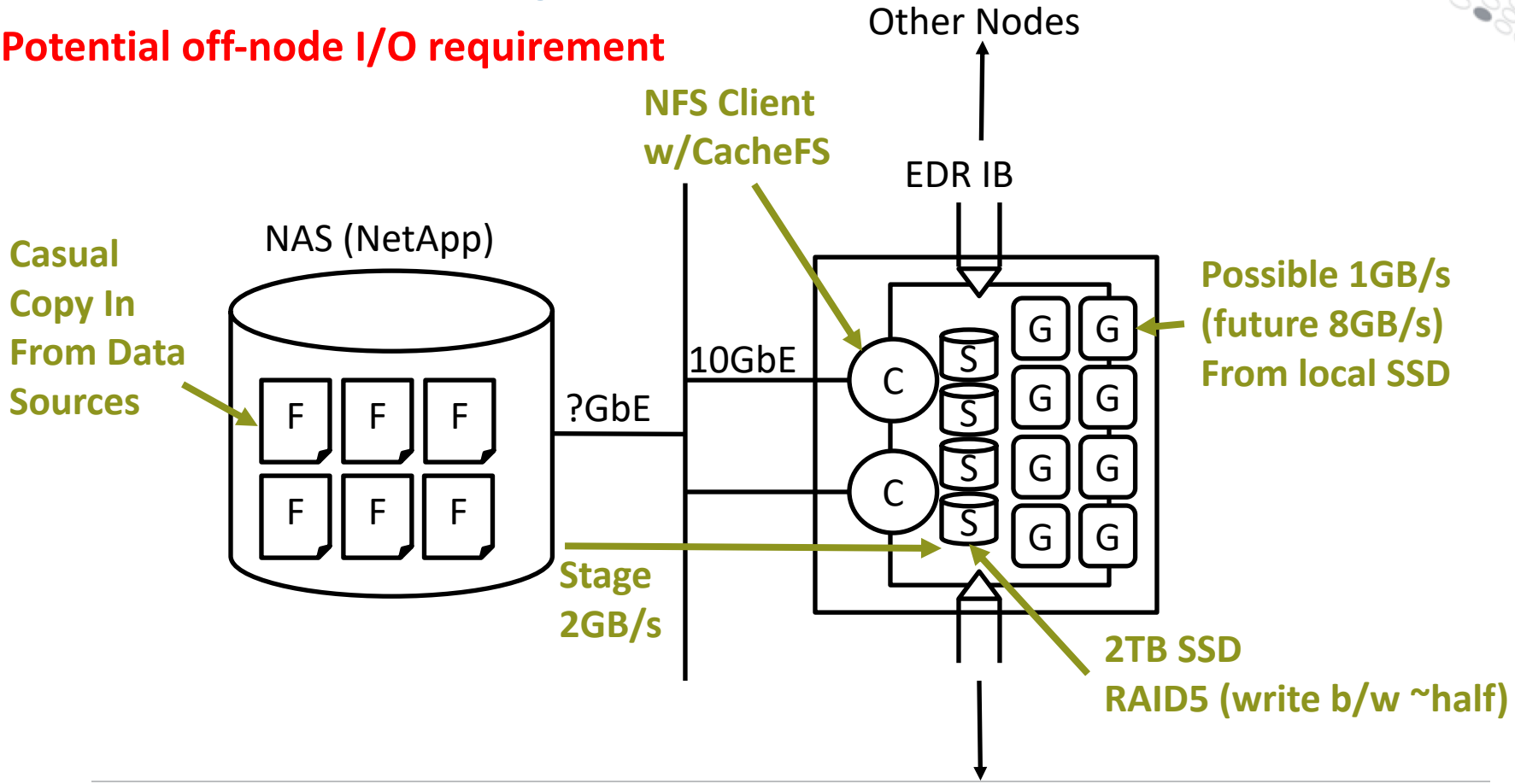
Sukumar, S. R., et al., (2016, December). Kernels for scalable data analysis in science: Towards an architecture-portable future. *In the Proc. Of the 2016 IEEE International Conference on Big Data,* pp. 1026-1031.

Lessons/Differences: AI Deployment

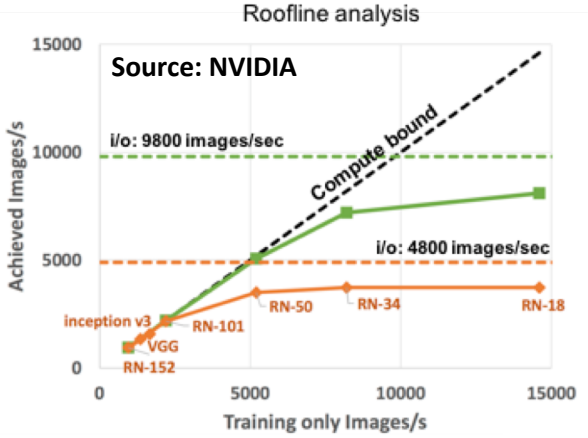
	Scientific Computing	Enterprise Computing
Model	Domain-specific	CNN, RNN, LSTM, GAN etc.
Baseline	Theoretic e.g. Navier Stokes	Humans, Other ML algorithms
Parallelism	Model, Ensemble	Data
Use Case	Computational Steering Proxy models	Speech, Test Image interpretation Hyper-personalization
Source File System	Lustre and GPFS	HDFS, S3, NFS etc.
Figure of Merit	Interpretability, Feasibility	Time-to-accuracy, Model-size
Training Data	O(GBs) per sample, $O(10^3)$ samples, O(10) categories	O(KBs) per sample, $O(10^6)$ samples, $O(10^4)$ categories
Data Model	HDF5, NETCDF	Relational, Document, Key-Value

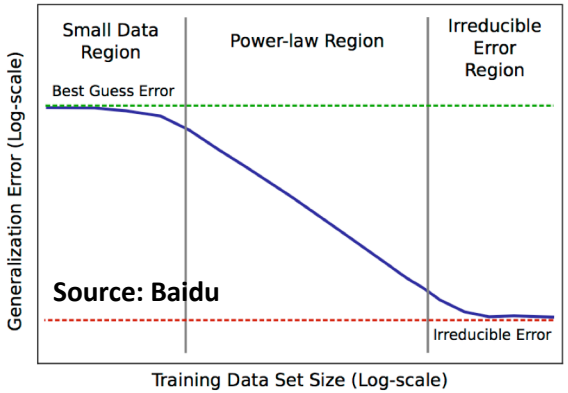
Observations: Today's Hardware

Potential off-node I/O requirement



Observations: Algorithm vs. Data vs. Compute





Opportunity for productivity with strong scaling

ResNet-50 Success	Time-to-accuracy	How many GPUs?	Scalability Efficiency
Facebook (Caffe2)	2 days 1 hour	352 GPUs 256	90% (large-batch)
IBM PowerAI (Caffe)	50 minutes	256 GPUs	95% (large-batch)
Google (TensorFlow)	~24 hours	64 TPUs	>90%
Preferred Networks (Chainer)	15 minutes	1000 GPUs	>90%
Cray @ CSCS (Tensorflow)	<14 minutes	1000 GPUs	~>95%

Productivity is performance and performance translates to productivity...

Observations: Looking Ahead

Figures-of-merit	State-of-practice	Projected 1-2 years ahead
Training-time to best accuracy	5+ days	2+ hours
Model Cost / TB (AWS GPUs)	~$25K (ResNet training on 80 GPUs for 5 days)	~10K
Hardware Efficiency	Network Depth: Flops::20x: 16x (based on AlexNet-2012 and ResNet-2015)	O(Teraflops) / problem
Statistical Efficiency	Depth: Accuracy:: 20x:13+ (based on AlexNet-2012 and ResNet-2015)	O(Teraflops) / problem
Need for compute as data grows	Data: Flops: Error:: 2x: 5x: 3+ (based on DeepSpeech1 and DeepSpeech2)	O(Petaflops) / problem
Training Cadence	~ Monthly	~ Daily
# of models per organization	1x	10-100x

Future of AI-HPC Integrated Systems

- **General purpose flexibility**
  - Commodity-like configurations with custom processors, chips
- **Seamless heterogeneity**
  - CPUs, GPUs, FPGAs, ASICs
- **High-performance interconnects for data centers**
  - MPI and TCP/IP collectives, compute on the network
- **Unified software stack**
  - Programming environment for performance and productivity
- **Workflow optimization**
  - Match growth in compute and data with I/O