Scaling AI from HPC to Cloud

Pradeep K Dubey

Intel Fellow and Fellow of IEEE

Director Parallel Computing Lab, Intel Labs

IXPUG, Sep 25, 2018, Intel Portland, OR

Machines:
Number Crunching
AND
Decision Making

Division of Labor Between Man and Machine Is Getting Disrupted:
Faster than Anyone Predicted!

MILS: Machine Intelligence Led Services

Mills

Information
Revolution

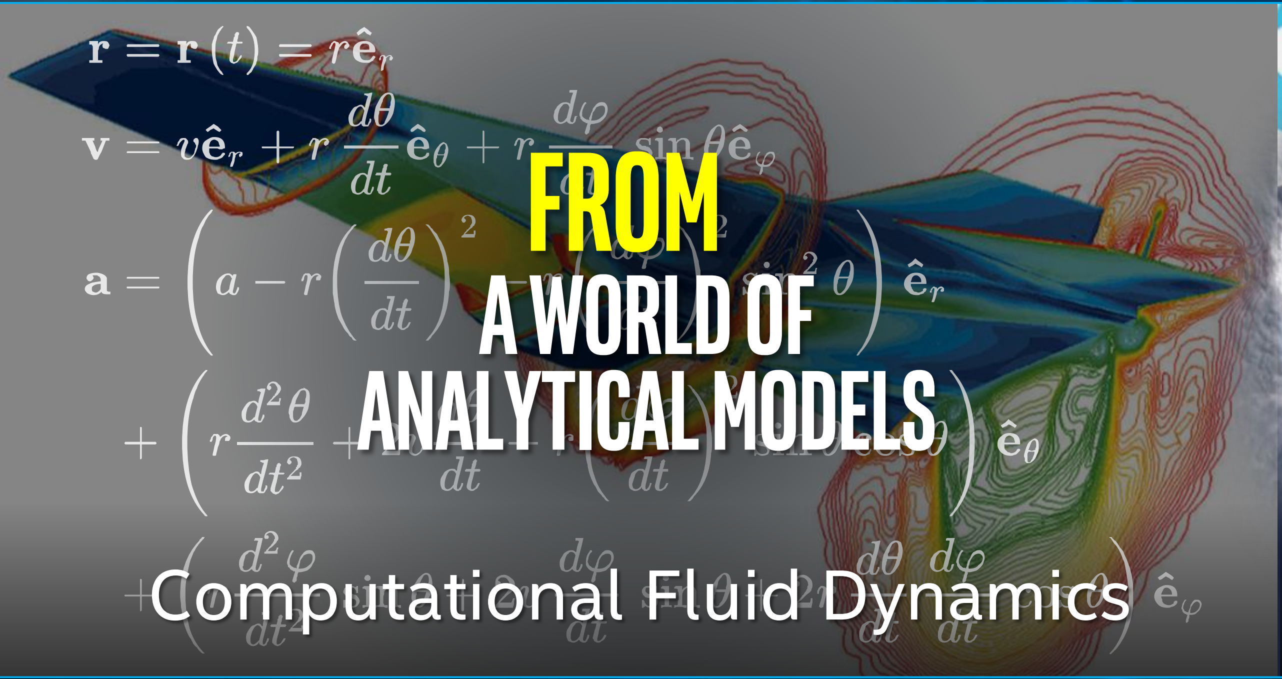
MILS

Intelligence Too Big for a Single Machine

"We're seeing a rebirth of artificial intelligence driven by the cloud, huge amounts of data and the learning algorithms of software,"
Larry Smarr, founding director of the California Institute for Telecommunications and Information Technology
http://bits.blogs.nytimes.com/2014/06/11/intelligence-too-big-for-a-single-machine/

THE NEW FRONTIER

Inside - Out ⇄ Outside - In



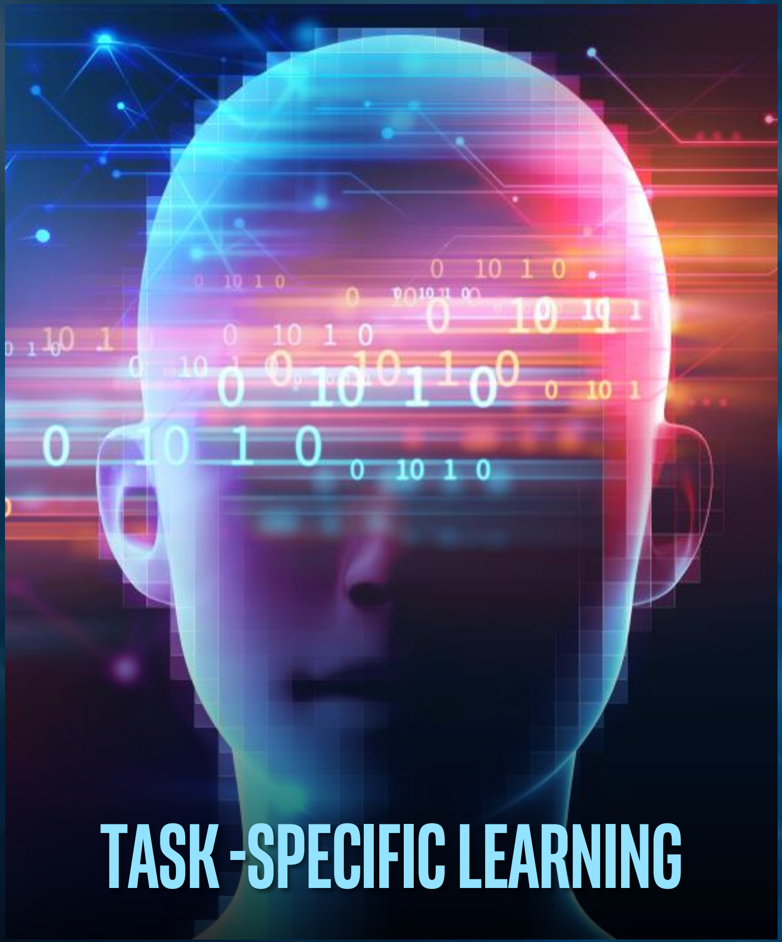
Start with Mathematical Model

Model → Simulate → Predict

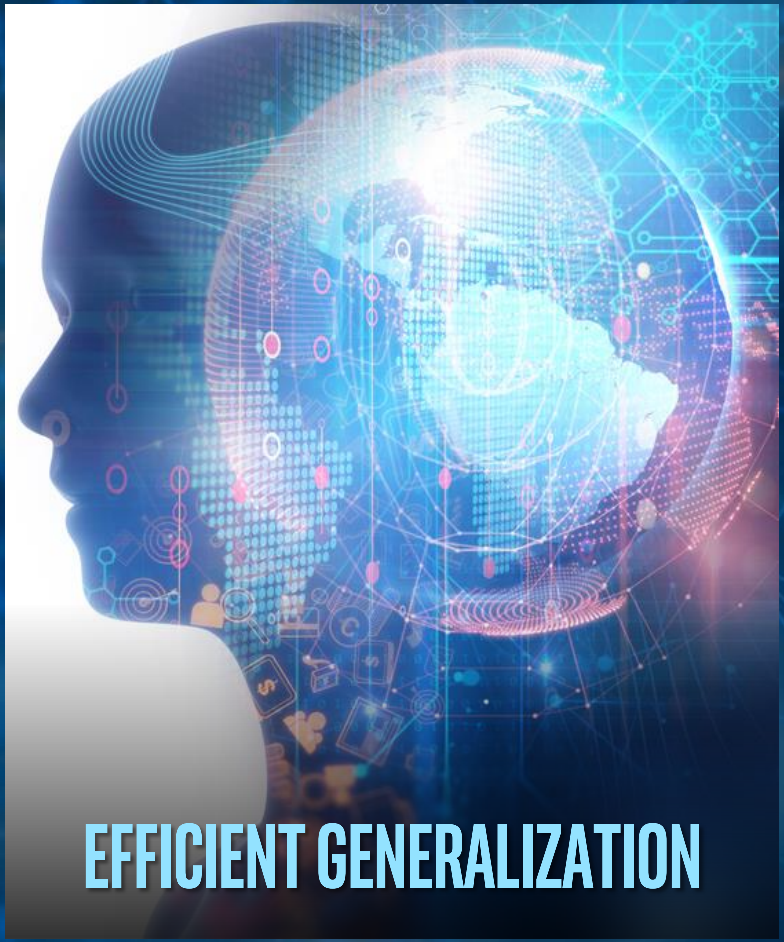
Start with Data

Initial State → Increment → Steer

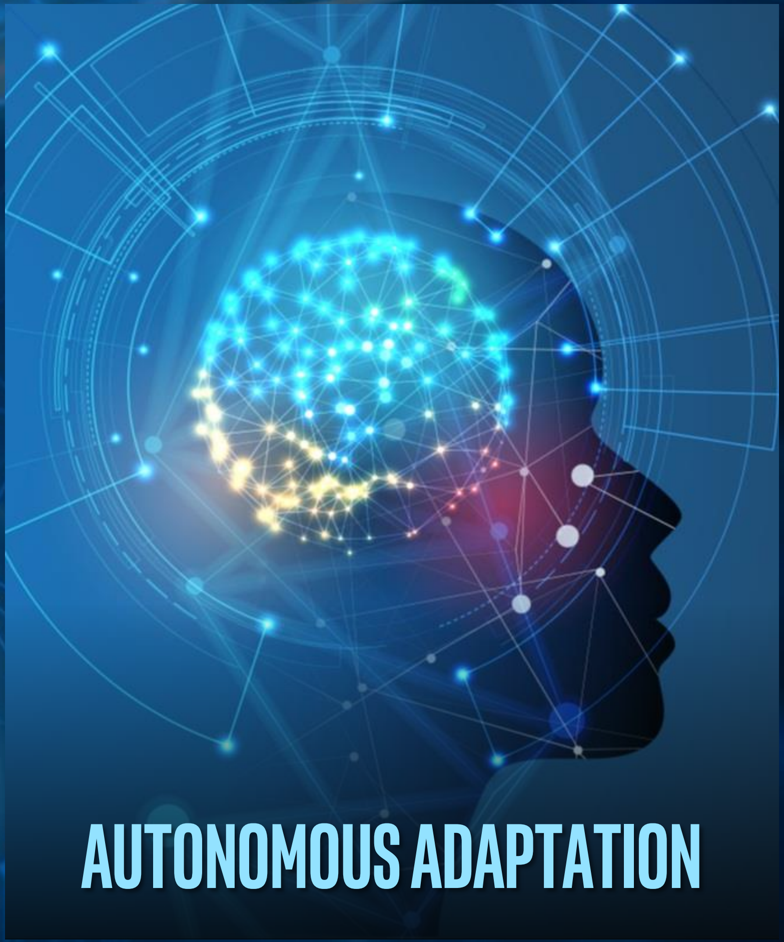
PROGRESSING TOWARDS HIGHER FORMS OF *INTELLIGENCE*



TASK -SPECIFIC LEARNING



EFFICIENT GENERALIZATION



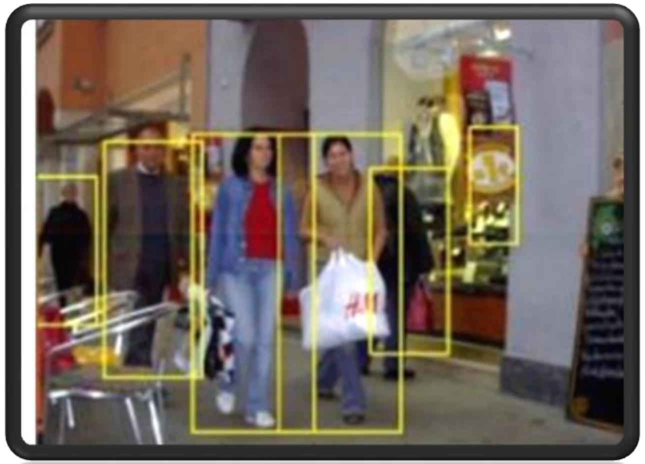
AUTONOMOUS ADAPTATION

INTEL'S AI RESEARCH STACK

SCIENCE:

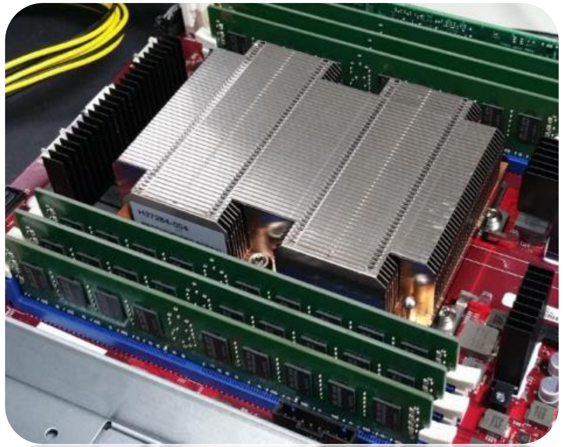
COGNITION AND NEUROSCIENCE

MIND'S EYE: A 3+ YEAR ONGOING COLLAB BETWEEN INTEL & PRINCETON



COMPUTER SCIENCE:

SCALABLE ALGORITHMS FOR LEARNING AND DECISION MAKING



COMPUTER SYSTEMS:

END-TO-END PIPELINE

DATA MANAGEMENT, MODEL TRAINING AND DEPLOYMENT

AI: What makes it hard and fun!

Better model building

LEARNING WITH LESS DATA AND SUPERVISION

DEEP NEURAL NETWORKS GETTING AUGMENTED: DATA-DRIVEN + ANALYTICAL + MEMORY

LEARNING MODELS THAT ARE EASIER TO REASON

CONTINUOUS LEARNING FOR MISSION-CRITICAL AI

More efficient and pervasive model deployment

THROUGHPUT, ACCURACY, AND MODEL SIZE TRADEOFFS: SPARSIFICATION AND PRUNING

SELF-LEARNING AND PERSONALIZATION AT THE EDGE

Compute architecture needs of AI

REDUCING ARITHMETIC PRECISION WHILE PRESERVING ACCURACY: ALL 32 → 16, 8, 4, 2 …
FEEDING THE COMPUTE ← MEMORY AND NETWORK; COMPUTE NEAR NETWORK AND MEMORY
DOMAIN-SPECIFIC ARCHITECTURES → TRADITIONAL, SPATIAL, NEUROMORPHIC, QUANTUM

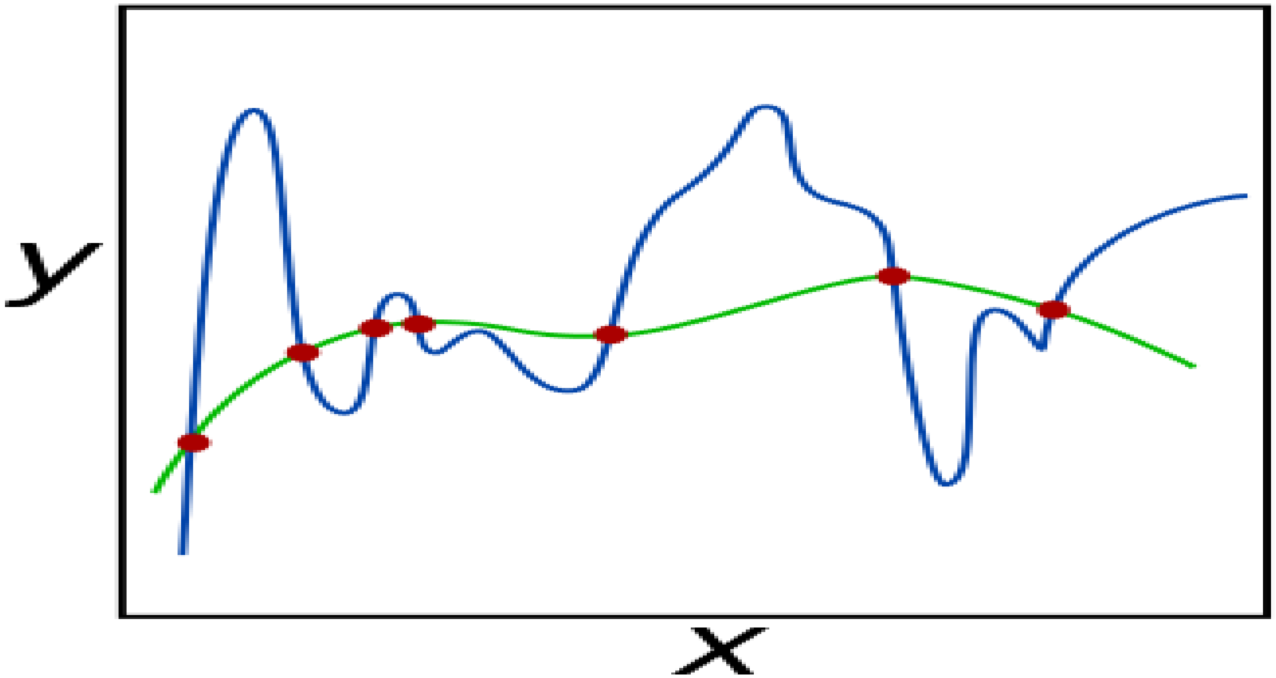
Productivity and Scaling needs of AI

STRONG-SCALING AI TO HPC SCALE ON CLOUD INFRASTRUCTURE: LARG BATCH SIZE AND 2nd ORDER METHODS

DELIVERING PERFORMANCE-PRODUCTIVITY: FaaS AND HIGHER ORDER LANGUAGE CONSTRUCTS

Low precision numeric motivation for AI is similar, Yet different …

- Similar: Energy saving and denser flops, higher throughput
- Different: Intrinsic Regularization → Generalization Of Learned Information



Capturing the dynamic range

Int16 based training (KNM/LakeCrest):

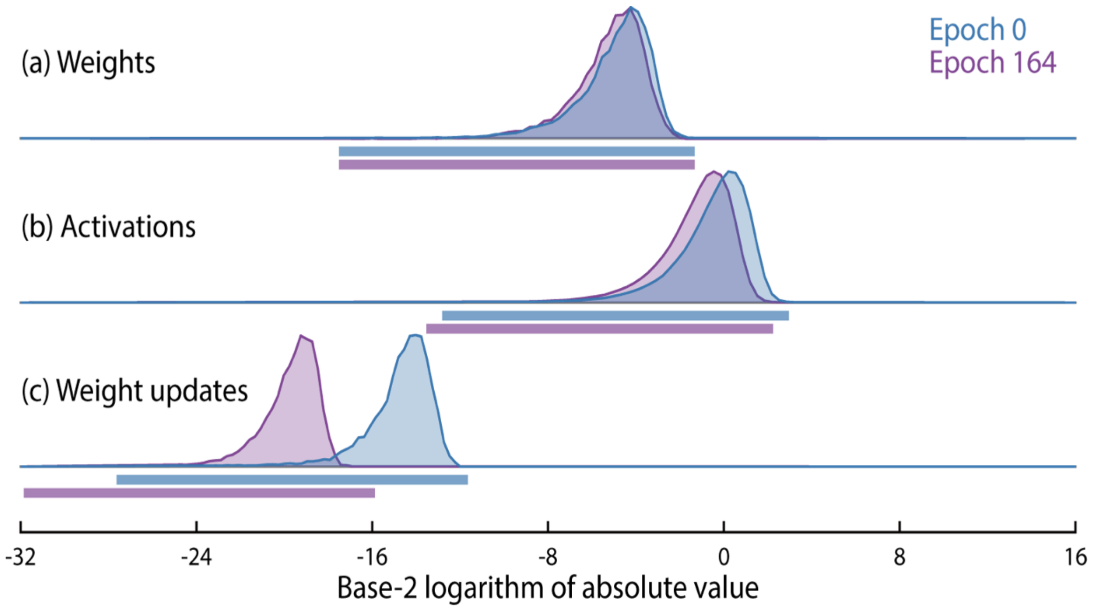
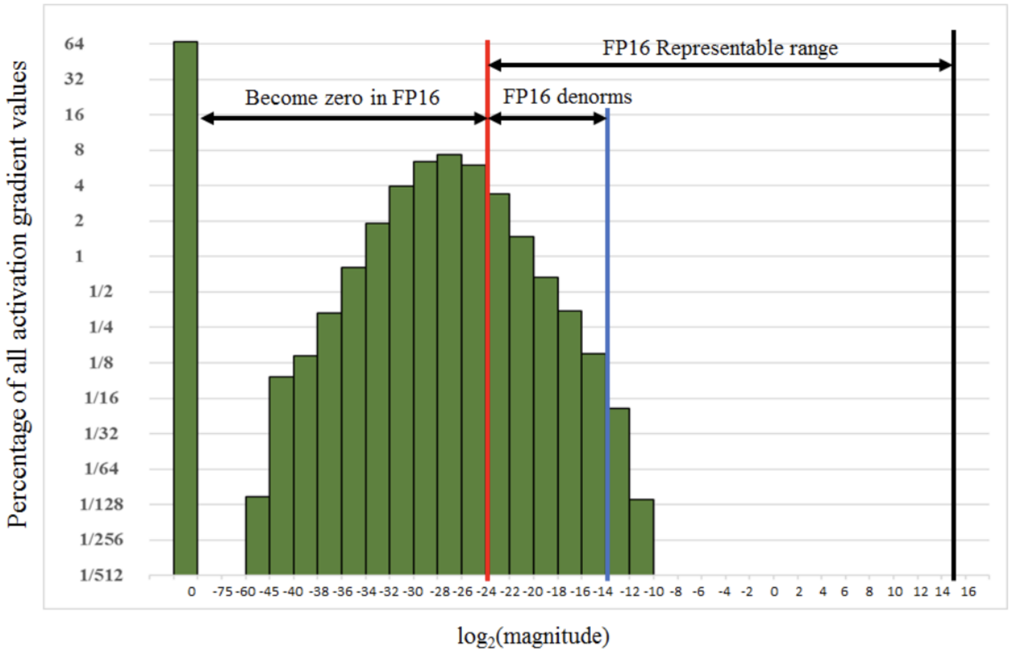
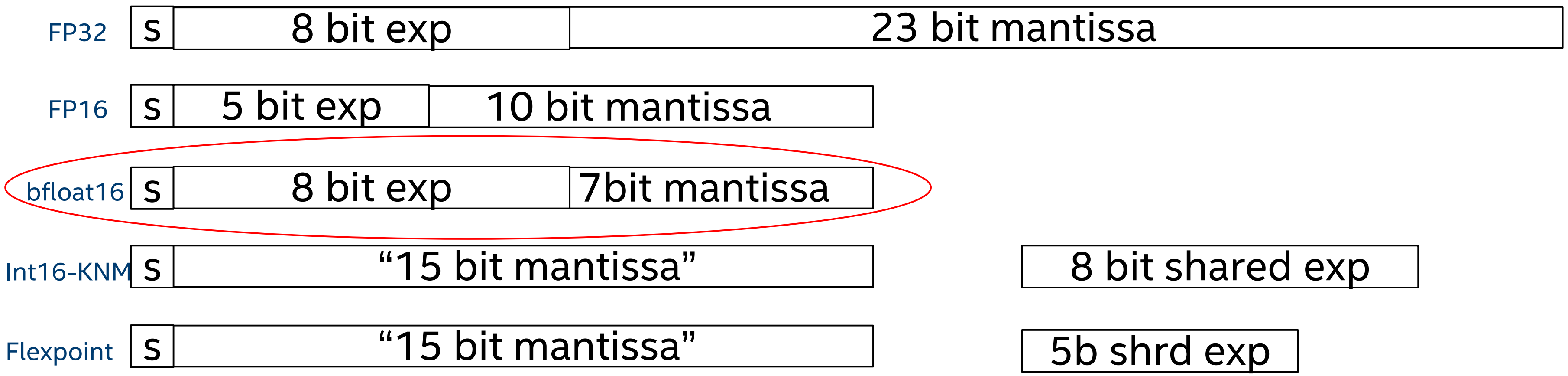


Figure 4. Distributions of tensor scales in a deep neural network and their evolution during training.

FP16 based training (GPUs):



The Training Datatype Choices



BFloat16: To be supported across *all* Intel Deep Learning Training Platforms
First Xeon Platform with BFloat16: Cooper Lake

MLSL : Key features & ideas

Abstraction:

- MLSL abstracts communication patterns and backend and supports data/model/hybrid parallelism

Flexibility:

- C, C++, Python languages are supported out of box

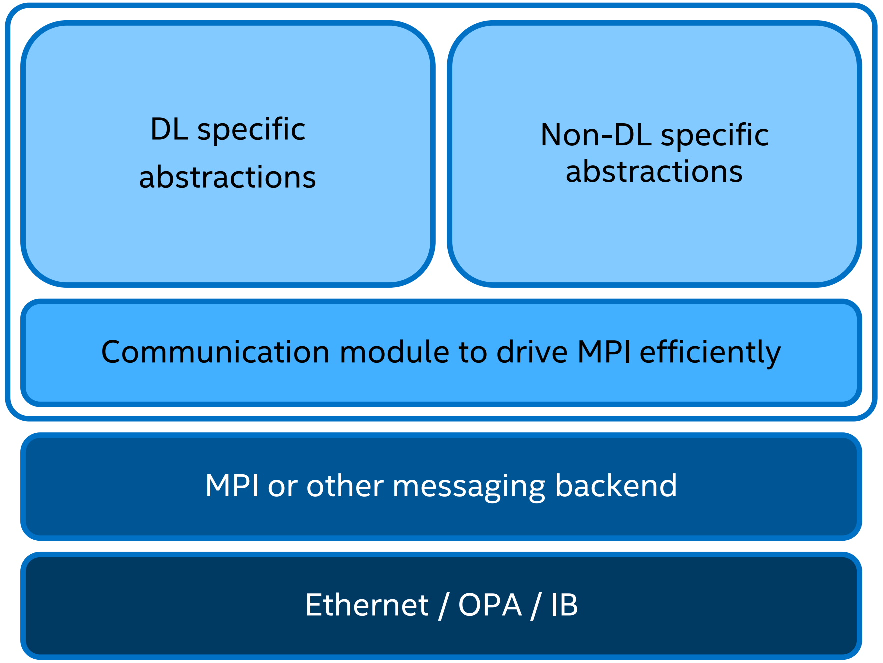
Usability

- MLSL API is being designed to be applicable to variety of popular FWs

Optimizations:

- MLSL uses not only the existing MPI functionality, but also extensions
- Domain awareness to drive MPI in a performant way
- Best performance across interconnects– transparent to frameworks

MLSL Architecture



MLSL : Collective API

Goal:

- Ease of enabling graph-based frameworks (allreduce op)

Collective Ops supported (non-blocking):

- Reduce/Allreduce
- Alltoall(v)
- Gather/Allgather(v)
- Scatter, Reduce_Scatter
- Bcast

Features:

- High performance (EP-based)
- Efficient asynchronous progress
- Prioritization (WIP)

```
/*Create MLSL environment*/
Environment env = Environment::GetEnv();
env.Init(&argc, &argv);

/* Create distribution
 * Arguments define how compute resources are split
 * between GROUP_DATA and GROUP_MODEL
 * Example below: all nodes belong to GROUP_DATA*/
Distribution* distribution = env.CreateDistribution(nodeCount, 1);

/*Handle for non-blocking comm operation*/
CommReq cr;

/*Start non-blocking op*/
distribution->AllReduce(sendbuffer, recvbuffer, size, DT_FLOAT, RT_SUM, GROUP_ALL, &cr);

/*Blocking wait call*/
env.Wait(&cr);
```

MLSL: Features

Current features:

- ✓ Non-blocking DL Layer and Collective interface
- ✓ Python/C++/C bindings
- ✓ Asynchronous communication progression
- ✓ Optimized algorithms
- ✓ Support for data, model, hybrid parallelism
- ✓ Initial support for quantization – available in IntelCaffe/MLSL
- ✓ Built-in inversed prioritization (through env. variable) – available in IntelCaffe/MLSL

- **Upcoming features (in development or research):**
  - ✓ Explicit prioritization API
  - ✓ Sparse data allreduce
  - ✓ Gradient quantization and compression
  - ✓ Cloud native features

https://github.com/intel/MLSL

Scale-out in Cloud environment

DAWNbench:

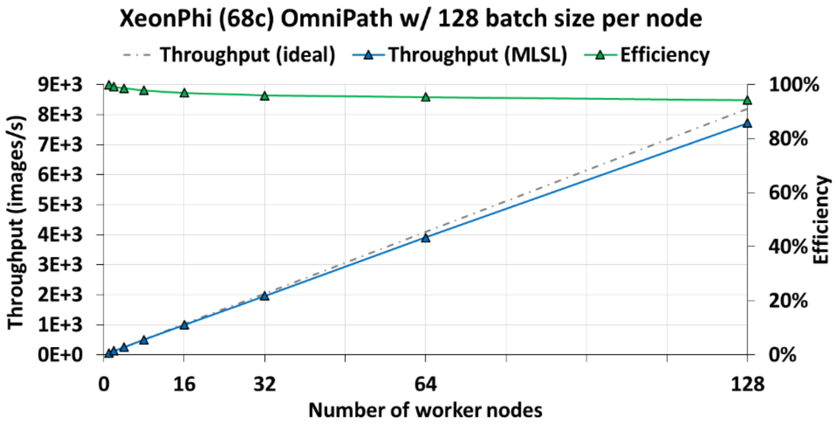
Apr 2018	ResNet50 *Intel(R) Corporation* source	3:25:55	N/A	93.02%	128 nodes with Xeon Platinum 8124M / 144 GB / 36 Cores (Amazon EC2 [c5.18xlarge])	Intel(R) Optimized Caffe
Apr 2018	ResNet56 *Intel(R) Corporation* source	3:31:47	N/A	93.11%	128 nodes with Xeon Platinum 8124M / 144 GB / 36 Cores (Amazon EC2 [c5.18xlarge])	Intel(R) Optimized Caffe
Apr 2018	ResNet50 *Intel(R) Corporation* source	6:09:50	N/A	93.05%	64 nodes with Xeon Platinum 8124M / 144 GB / 36 Cores (Amazon EC2 [c5.18xlarge])	Intel(R) Optimized Caffe

*RN50:

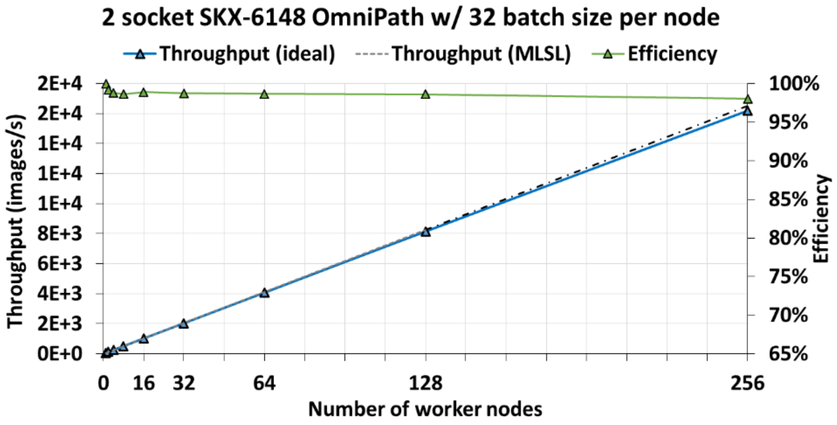
- 81 epochs for 64 nodes
- 85 epochs for 128 nodes

Scale-out in HPC environment

- **IntelCaffe**: MLSL–based multinode solution; **Horovod, nGraph:** WIP
- MLSL is enabled in Baidu's DeepBench
- SURFSara: used IntelCaffe/MLSL to achieve ResNet50 time-to-train record (~40 minutes, 768 SKX) *
- UC-Berkeley, TACC, and UC-Davis: 14 minutes TTT for ResNet50 with IntelCaffe/MLSL (2048 KNL) **



TensorFlow scaling on IA

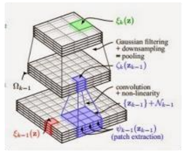


IntelCaffe/MLSL

Deep Learning at 15PF!

Deep Learning Applied to Science Problems in High Energy Physics and Climate Simulation

Novel Hybrid Parameter Scheme



Highest Performance and Scaling Reported for Deep Learning To Date:

15 PF peak, sustained 13.27 PF on 9K Cori nodes *

NERSC-STANFORD-INTEL COLLABORATION *

Common Tool Chain of MKL-DNN, MLSL, IntelCaffe Scales DL from 100s to 1000s of Xeon and Xeon Phi nodes: benchmarks and science apps

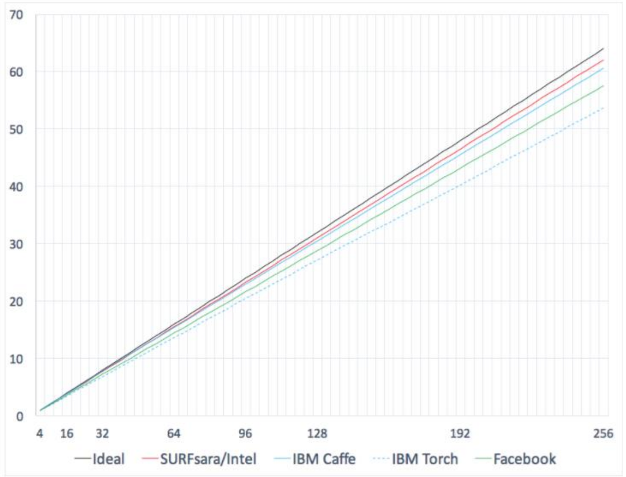
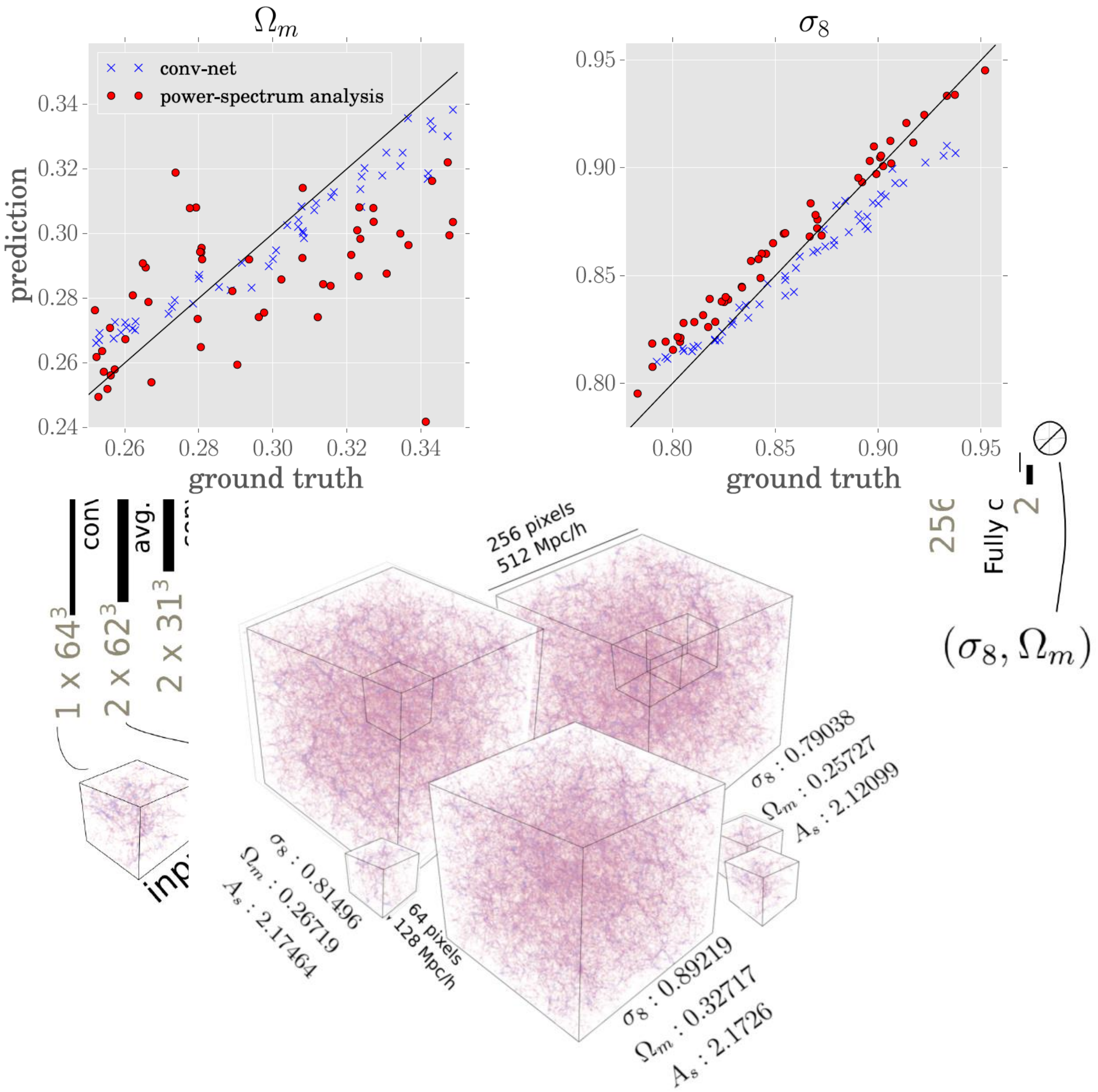


Fig 1. Scaling efficiency on Stampede2 (speedup vs number of workers). This plot starts from scaling on 4 workers, which has a scaling factor of 1.

Exploring the universe with deep learning *

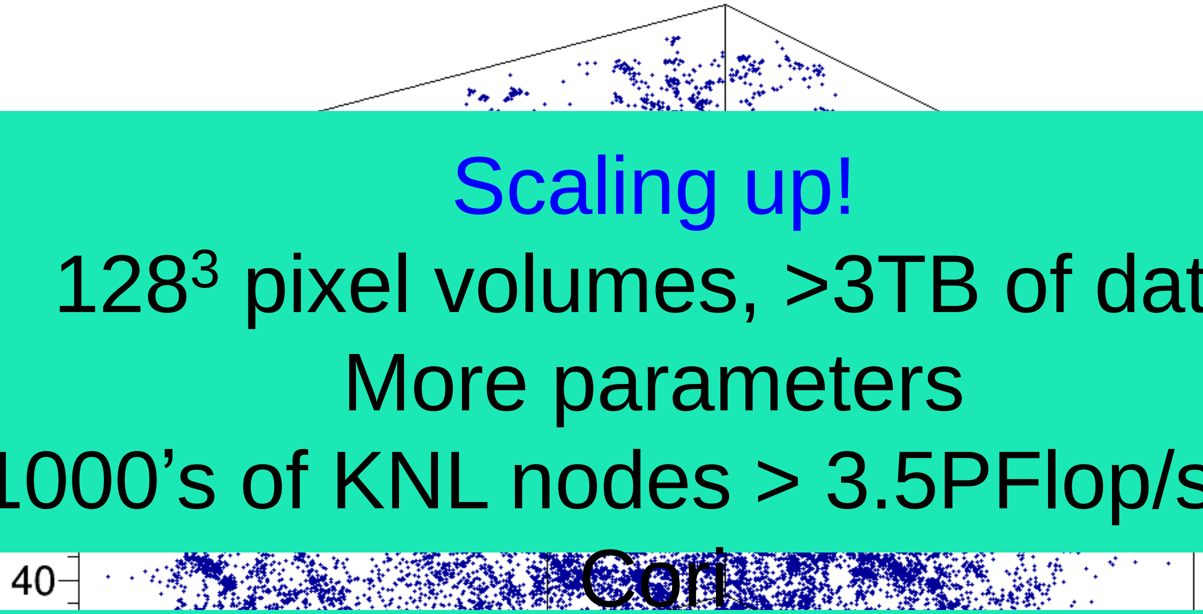


- Using deep 3D convolutional networks for volumetric representation of dark-matter simulations
- *First work* to predict cosmological constants (Ω_M , σ_8) from simulations!
- Outperforms traditional methods at parameter estimation using "cosmological models"

Slide courtesy: Prabhat at NERSC
Contributors: NERSC, Cray, Intel, UC Berkeley

* Ravanbaksh, Oliva, Fromenteau, Price, Ho, Schneider, Poczos; arxiv.org/abs/1711.02033

Exploring the universe with deep learning *



Scaling up!

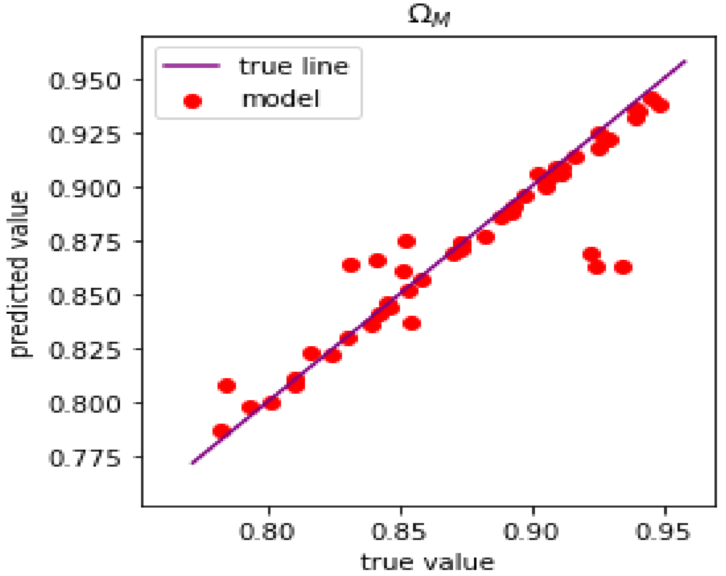
128^3 pixel volumes, >3TB of data

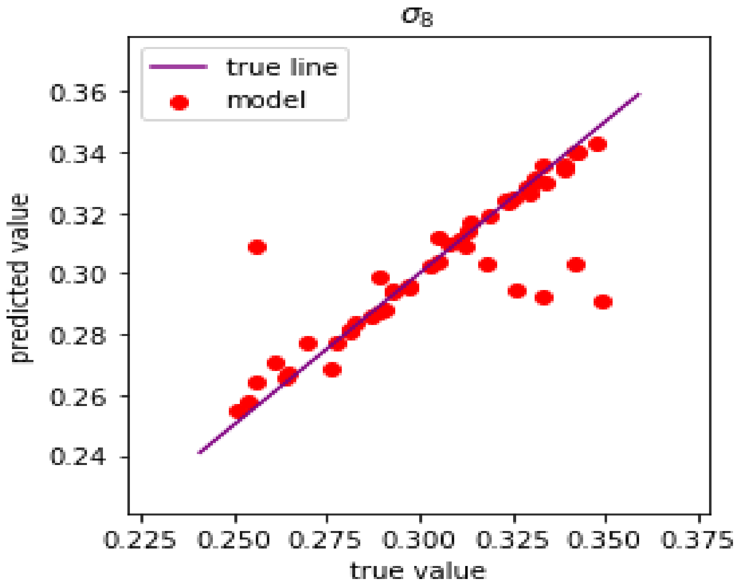
More parameters

1000's of KNL nodes > 3.5PFlop/s on Cori

2048-node run achieved comparable to the experimental uncertainty for Ω_M and σ_8, and almost 5x smaller for N_s.

Significantly lower than prior CNN based results





* CosmoFlow: "Using Deep Learning to Learn the Universe at Scale" Amrita Mathuriya, Deborah Bard, Peter Mendygral, Lawrence Meadows, James Arnemann Lei Shao, Siyu He, Tuomas Kärnä, Daina Moise, Mike Ringenburg, Prabhat, Victor Lee; Accepted at SC'18

Slide courtesy: Prabhat at NERSC
Contributors: NERSC, Cray, Intel, UC Berkeley

Extreme scale *de novo* metagenome assembly *

Problem & Challenges

- Metagenomics is the leading technology in studying the uncultured microbial diversity, microbiome structure and function.
- Given overlapping, short, erroneous genome fragments we want to assemble metagenomes that are up to 1,000,000 times longer.
- Challenges: Erroneous sequences, polymorphisms, variable coverage and repeats in the genomes.
- Memory and computational requirements make the processing of massive datasets prohibitive.

Analogy: "*Shred all the books in a library into small pieces and reconstruct the books given only the shreds*"

** Extreme scale *de novo* metagenome assembly, E. Georganas, R. Egan, S. Hofmeyr, E. Goltsman, A. Buluc, B. Arndt, A, Tritt, L. Oliker, K. Yelick, to appear at Supercomputing '18 → Best Paper Finalist

Extreme scale *de novo* metagenome assembly … continued

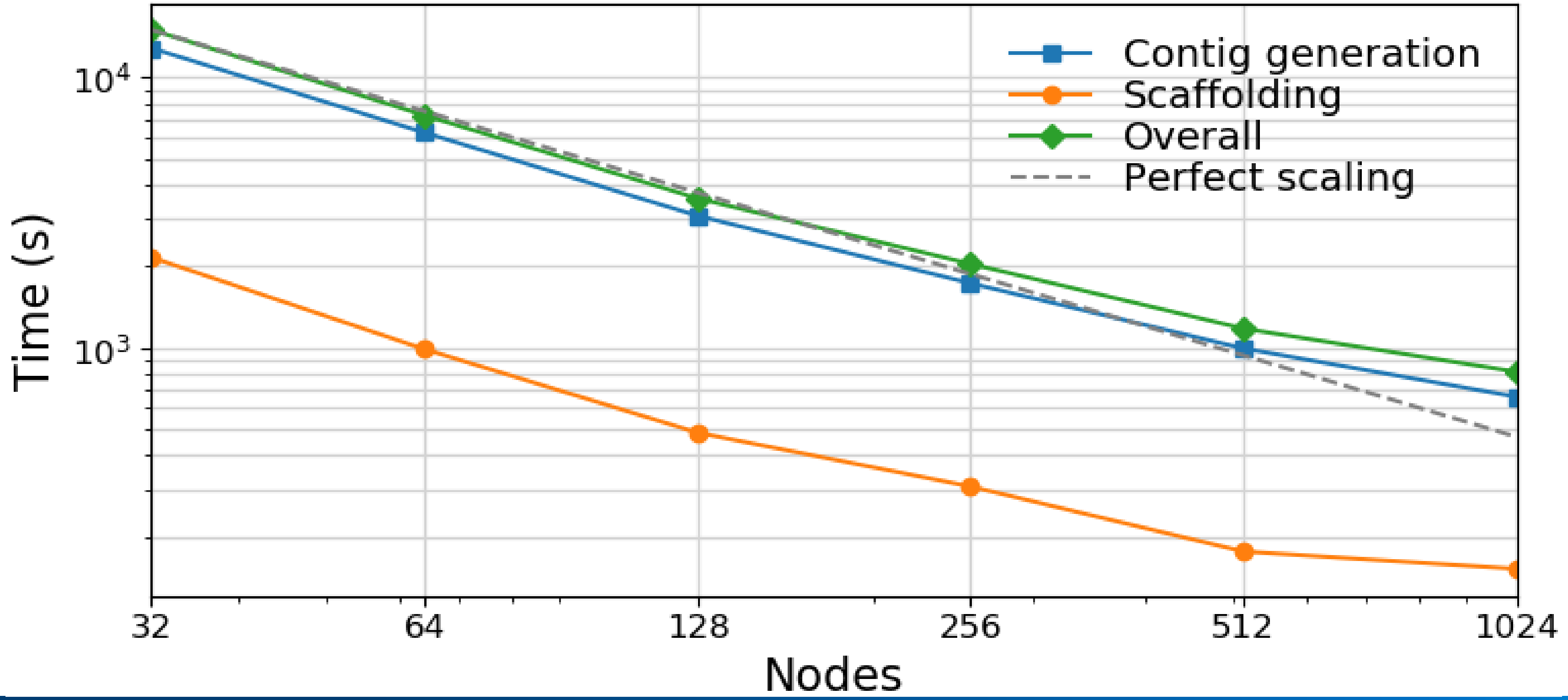
Solution

- **MetaHipmer**: Novel metagenome assembly algorithms and distributed memory parallelization for both memory size and speed.
- Global address space programming model via Unified Parallel C to facilitate irregular accesses across the aggregate machine memory.
- Parallel graph algorithms and distributed hash tables, optimized for the statistical characteristics of the assembly process.
- Communication-avoiding algorithms, one-sided communication, software caches, dynamic message aggregation, hardware atomics

** Extreme scale *de novo* metagenome assembly, E. Georganas, R. Egan, S. Hofmeyr, E. Goltsman, A. Buluc, B. Arndt, A, Tritt, L. Oliker, K. Yelick, to appear at Supercomputing '18 → Best Paper Finalist

Extreme scale *de novo* metagenome assembly * … continued

MetaHipmer is transformative: Full assembly of the 2.6 TByte Twitchell Wetlands environmental sample **-- the largest, high-quality de novo metagenome assembly completed to date** → Quality of MetaHipmer matches or exceeds state-of-the-art assemblers. This grand challenge problem took 3 hours and 25 minutes on 512 nodes of Cori supercomputer.

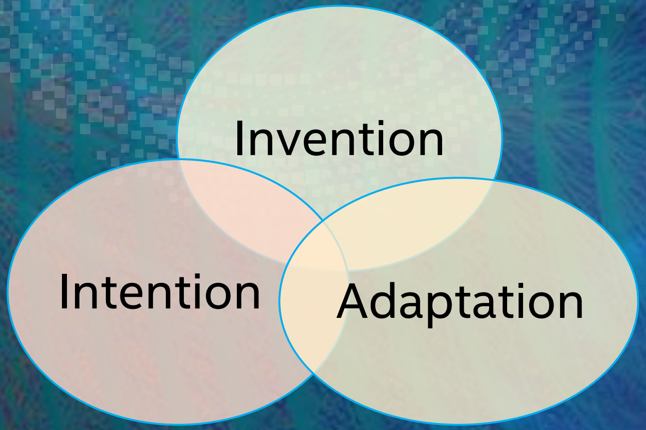


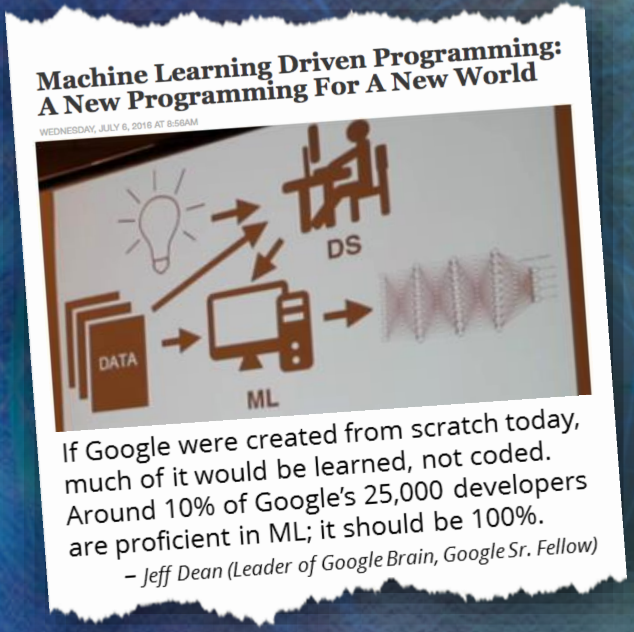
** Extreme scale *de novo* metagenome assembly, E. Georganas, R. Egan, S. Hofmeyr, E. Goltsman, A. Buluc, B. Arndt, A, Tritt, L. Oliker, K. Yelick, to appear at Supercomputing ‘18 → Best Paper Finalist

The AI Grand Challenge?

A process in which some or all of the steps of turning a user's intent into an executable program are automated.

Machine Programming*




Machine Learning Driven Programming: A New Programming For A New World

WEDNESDAY, JULY 6, 2016 AT 8:56AM

If Google were created from scratch today, much of it would be learned, not coded. Around 10% of Google's 25,000 developers are proficient in ML; it should be 100%.

– *Jeff Dean (Leader of Google Brain, Google Sr. Fellow)*

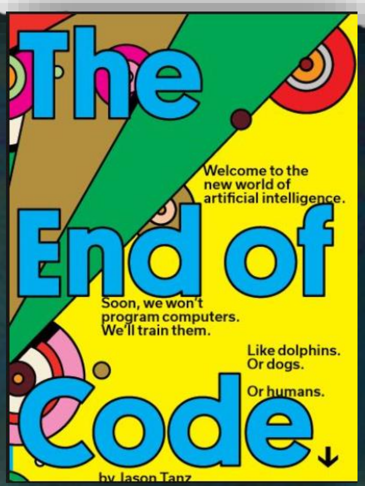

WIRED

Soon We Won't Program Computers. We'll Train Them Like Dogs

JASON TANZ BUSINESS 05.17.16 06:50 AM

SUBSCRIBE

SOON WE WON'T PROGRAM COMPUTERS. WE'LL TRAIN THEM LIKE DOGS


The End of Code

Welcome to the new world of artificial intelligence.

Soon, we won't program computers. We'll train them.

Like dolphins. Or dogs.

Or humans.

by Jason Tanz

* Three Pillars paper (MAPL '18): Justin Gottschlich, Armando Solar-Lezama, Nesime Tatbul, Michael Carbin, Martin Rinard, Regina Barzilay, Saman Amarasinghe, Joshua B Tenenbaum, Tim Mattson; https://arxiv.org/abs/1803.07244

We are at an unprecedented convergence of massive compute with massive data …
This confluence will have a lasting impact on both how we do computing and what computing can do for us!

Thank You

Notice and Disclaimers

Notice: This document contains information on products in the design phase of development. The information here is subject to change without notice. Do not finalize a design with this information. Contact your local Intel sales office or your distributor to obtain the latest specification before placing your product order.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY RELATING TO SALE AND/OR USE OF INTEL PRODUCTS, INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT, OR OTHER INTELLECTUAL PROPERTY RIGHT. Intel products are not intended for use in medical, life saving, or life sustaining applications. Intel may make changes to specifications, product descriptions, and plans at any time, without notice.

All products, dates, and figures are preliminary for planning purposes and are subject to change without notice.

Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined.“ Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The Intel products discussed herein may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

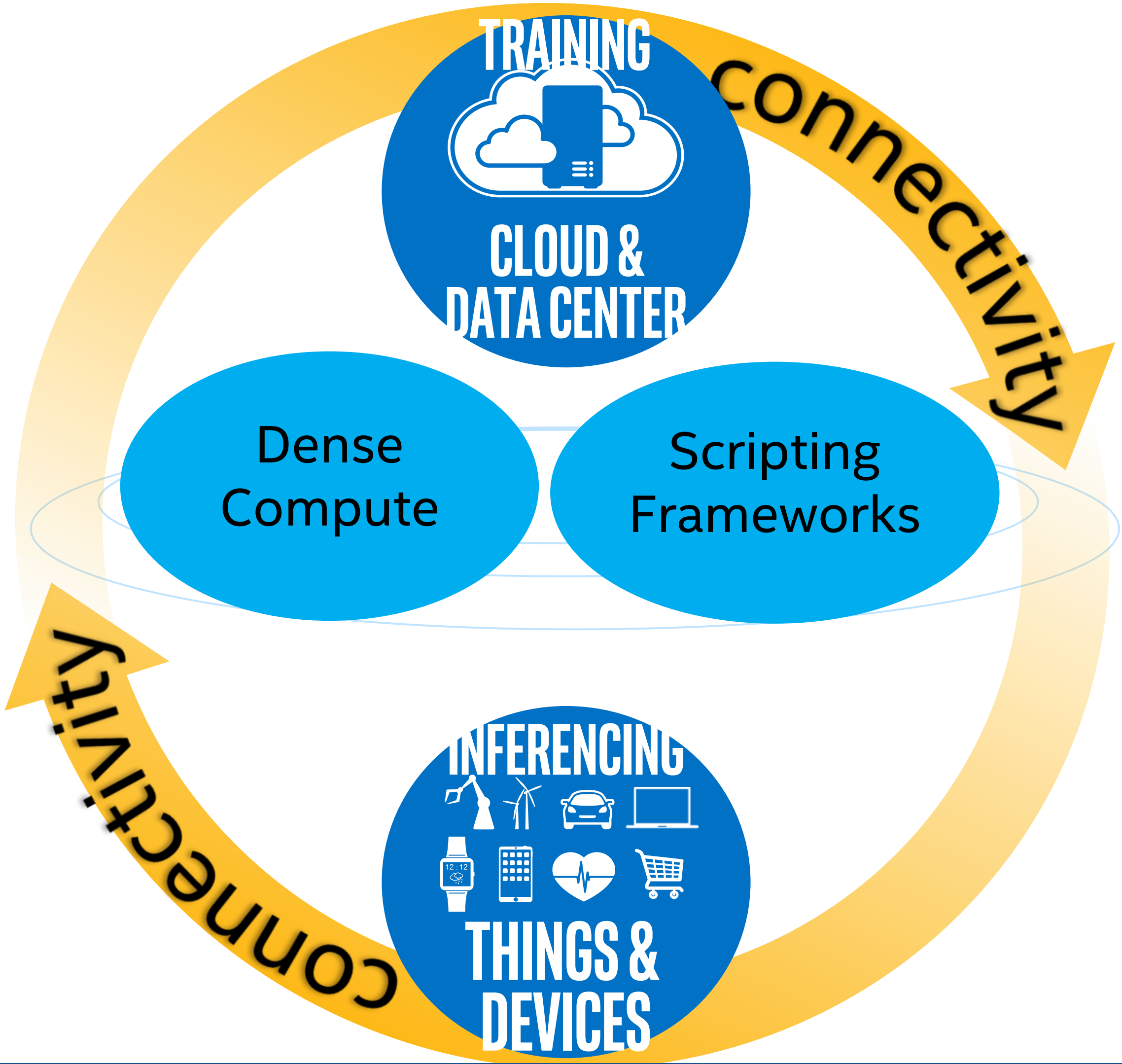
Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's website at http://www.intel.com.

Notice and Disclaimers Continued …

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804

Virtuous Cycle of Compute



AI Needs More Compute Faster: 55% revenue CAGR … >$47 billion in 2020 *

* Source: IDC Worldwide Semiannual Cognitive/Artificial Intelligence Systems Spending Guide, Oct 2016