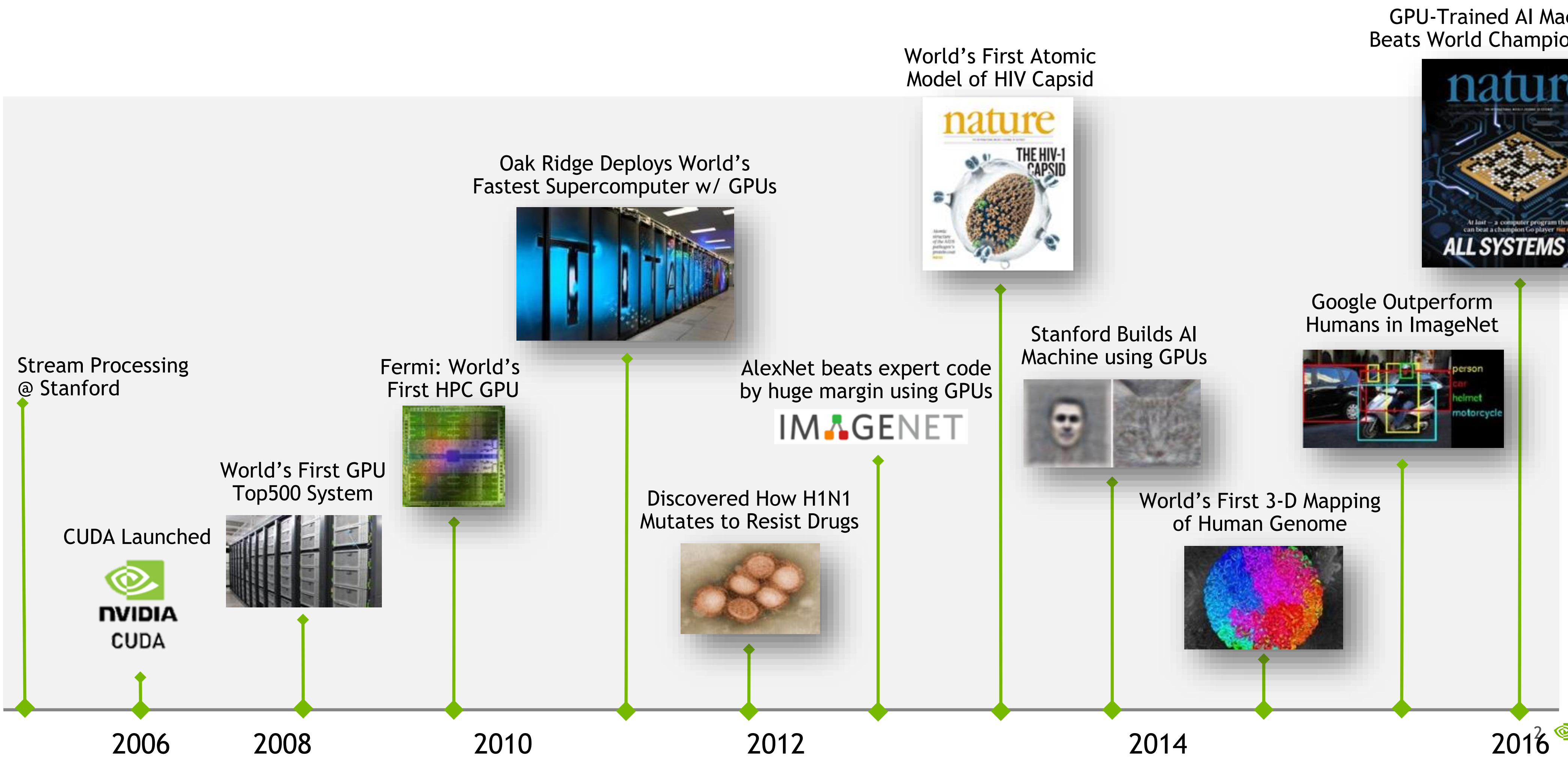
Deep Learning and HPC

Bill Dally, Chief Scientist and SVP of Research

January 17, 2017

NVIDIA.

A Decade of Scientific Computing with GPUs



GPUs Enable Science

TITAN

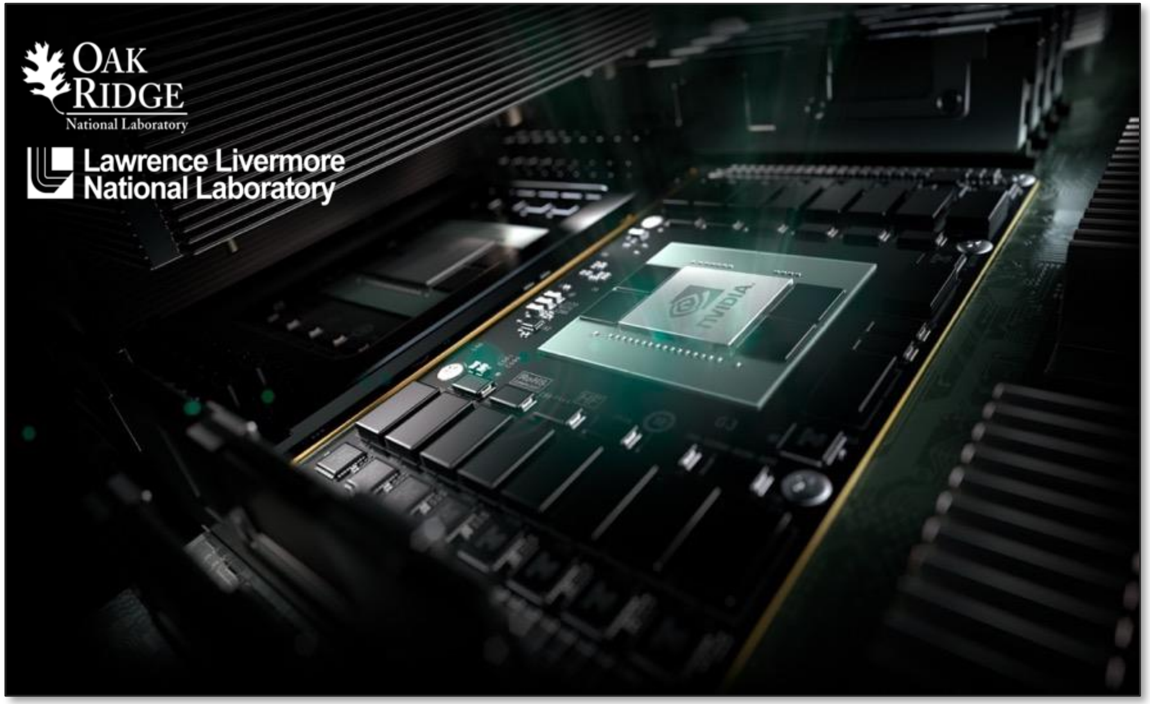
18,688 NVIDIA Tesla K20X GPUs

27 Petaflops Peak: 90% of Performance from GPUs

17.59 Petaflops Sustained Performance on Linpack

U.S. to Build Two Flagship Supercomputers

Pre-Exascale Systems Powered by the Tesla Platform



Summit & Sierra Supercomputers

100-300 PFLOPS Peak

IBM POWER9 CPU + NVIDIA Volta GPU

NVLink High Speed Interconnect

40 TFLOPS per Node, >3,400 Nodes

2017

DGX SATURNV

World's Most Efficient AI Supercomputer

TOP 500 The List.

Fastest AI Supercomputer in TOP500
4.9 Petaflops Peak FP64 Performance
19.6 Petaflops DL FP16 Performance
124 NVIDIA DGX-1 Server Nodes

THE GREEN 500

Most Energy Efficient Supercomputer
#1 on Green500 List
9.5 GFLOPS per Watt
2x More Efficient than Xeon Phi System

FACTOIDS

13 DGX-1 Servers in Top500

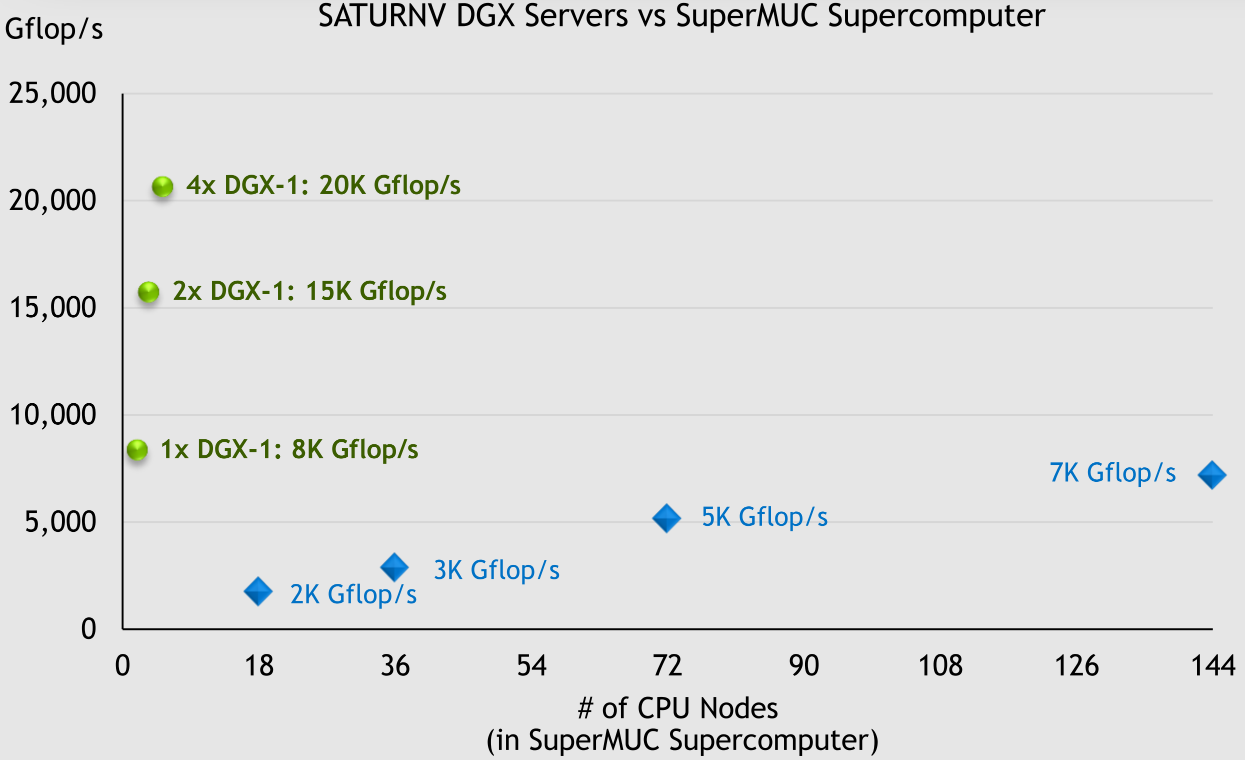
38 DGX-1 Servers for Petascale supercomputer

55x less servers, 12x less power vs CPU-only supercomputer of similar performance

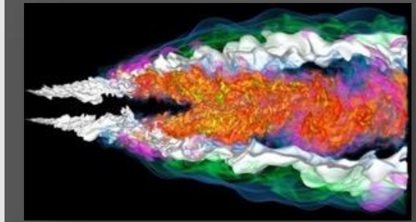
EXASCALE APPLICATIONS ON SATURNV

LQCD- Higher Energy Physics

SATURNV DGX Servers vs SuperMUC Supercomputer

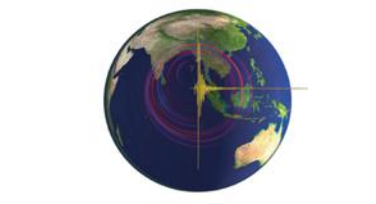


of CPU Servers to Match Performance of SATURNV



2,300 CPU Servers

S3D: Discovering New Fuel for Engines

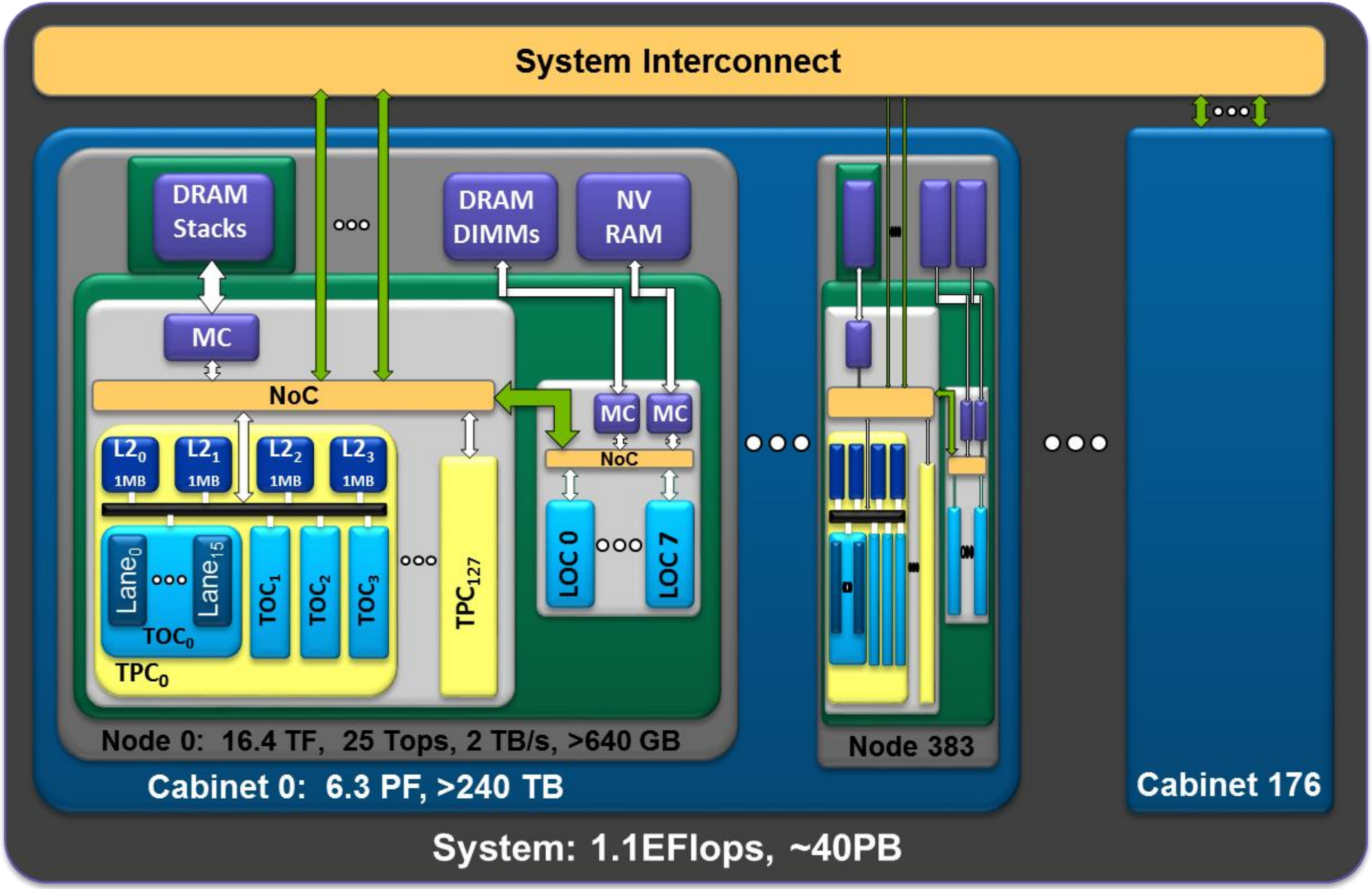


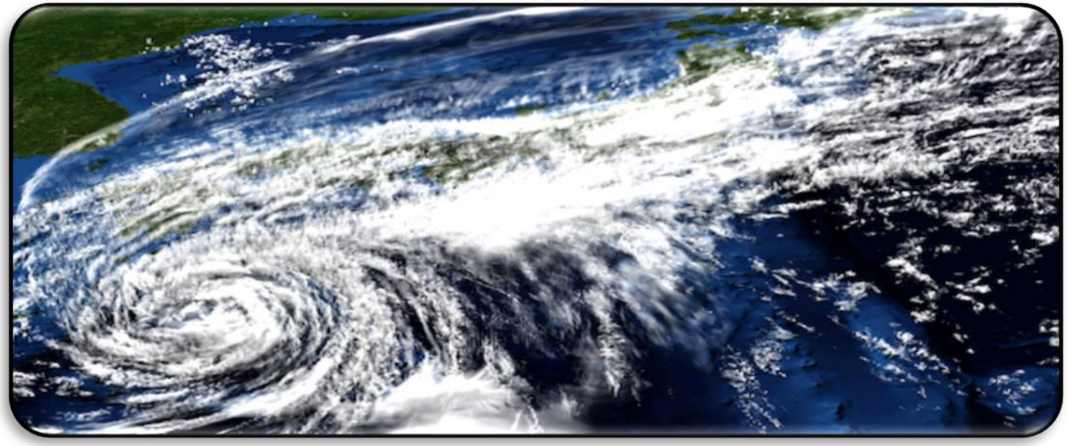
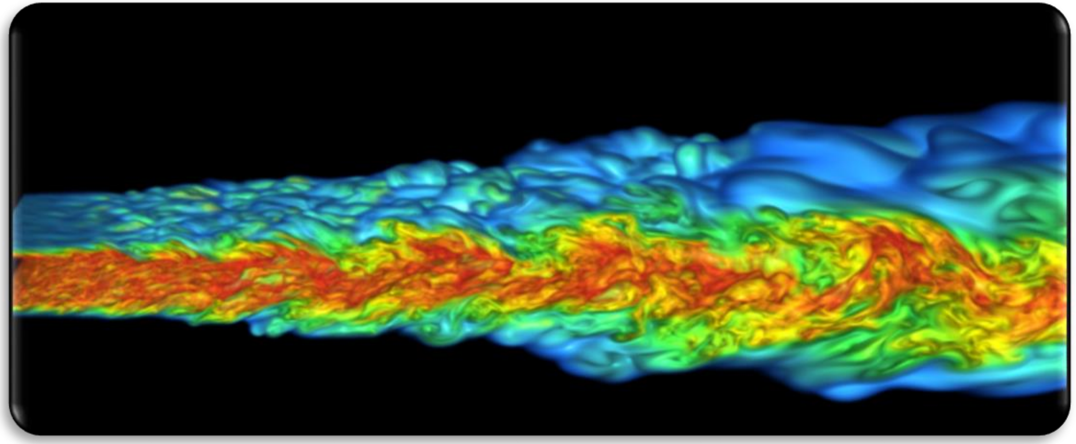
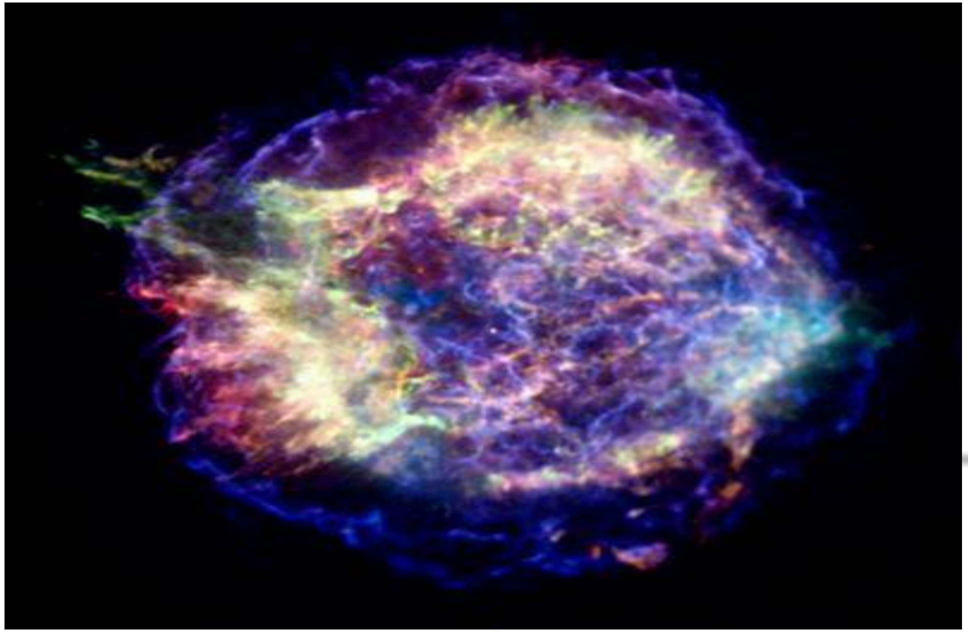
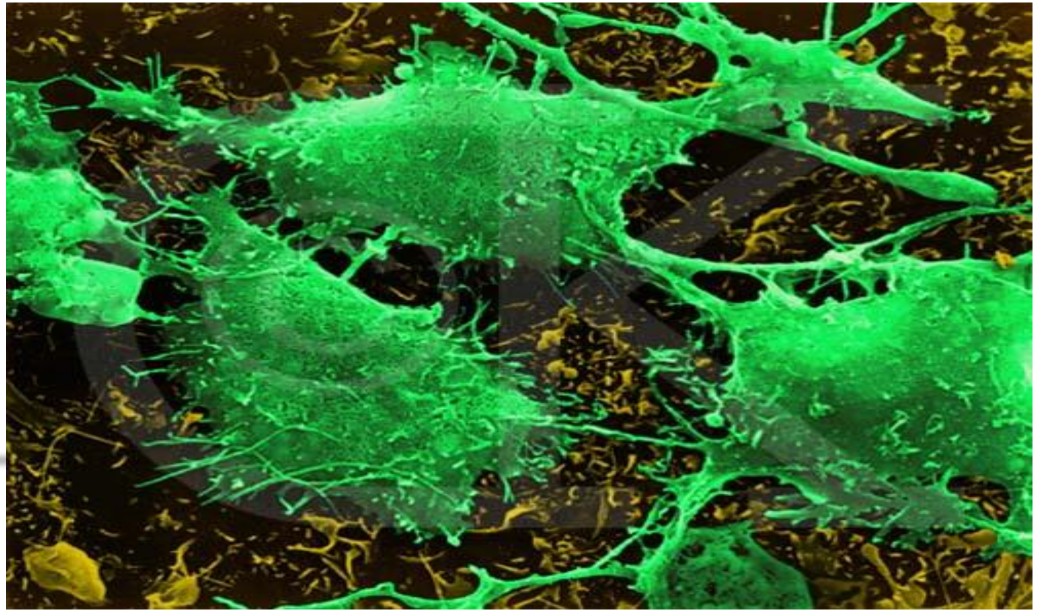
3,800 CPU Servers

SPECFEM3D: Simulating Earthquakes

QUDA version 0.9beta, using double-half mixed precision
DDalphaAMG using double-single

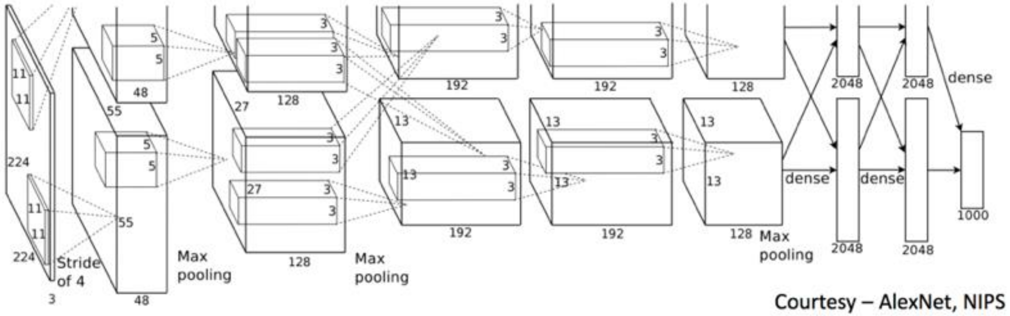
Exascale System Sketch



GPUs Enable Deep Learning

GPUs + Data + DNNs



Courtesy – AlexNet, NIPS 2012

THE STAGE IS SET FOR THE AI REVOLUTION

Deep learning with COTS HPC systems

Adam Coates ACOATES@CS.STANFORD.EDU
Brody Huval BRODYH@STANFORD.EDU
Tao Wang TWANGCAT@STANFORD.EDU
David J. Wu DWU4@CS.STANFORD.EDU
Andrew Y. Ng ANG@CS.STANFORD.EDU
Stanford University Computer Science Dept., 353 Serra Mall, Stanford, CA 94305 USA

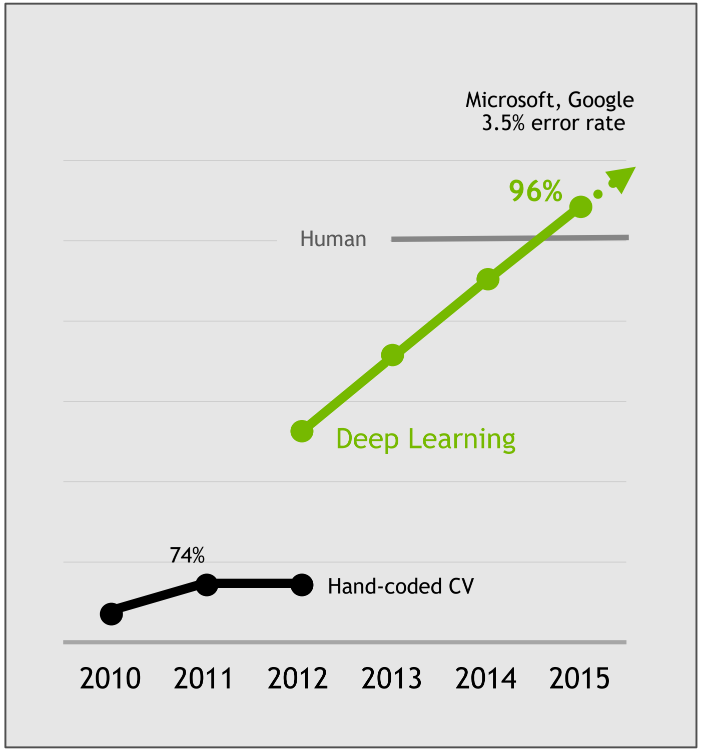
Bryan Catanzaro BCATANZARO@NVIDIA.COM
NVIDIA Corporation, 2701 San Tomas Expressway, Santa Clara, CA 95050

Abstract

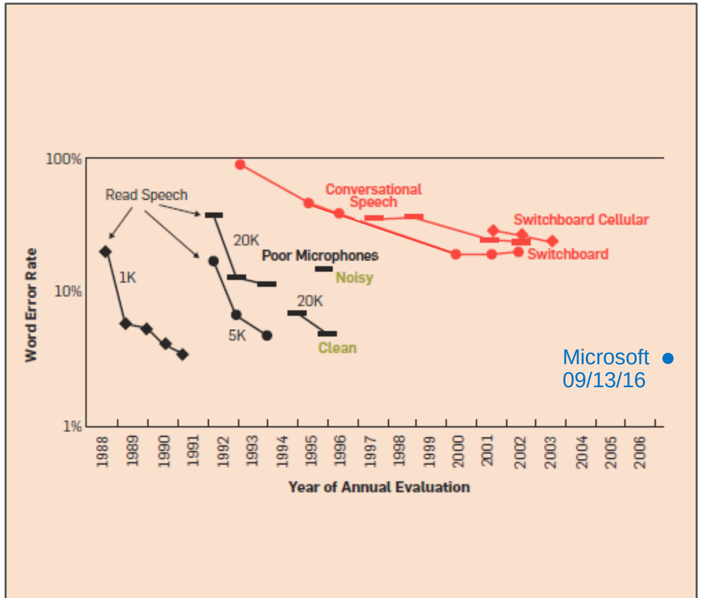
Scaling up deep learning algorithms has been shown to lead to increased performance in benchmark tasks and to enable discovery of complex high-level features. Recent efforts to train extremely large networks (with over 1 billion parameters) have relied on cloud-like computing infrastructure and thousands of CPU cores. In this paper, we present technical details and results from our own system based on Commodity Off-The-Shelf High Performance Computing (COTS HPC) technology: a cluster of GPU servers with Infiniband interconnects and MPI. Our system is able to train 1 billion parameter networks on just 3 machines in a couple of days, and we show that it can scale to networks with over 11 billion parameters using just 16 machines. As this infrastructure is much more easily marshaled by others, the approach enables much wider-spread research with extremely large neural networks.

to detect objects when trained from unlabeled images alone (Coates et al., 2012; Le et al., 2012). The very largest of these systems has been constructed by Le et al. (Le et al., 2012) and Dean et al. (Dean et al., 2012), which is able to train neural networks with over 1 billion trainable parameters. While such extremely large networks are potentially valuable objects of AI research, the expense to train them is overwhelming: the distributed computing infrastructure (known as "DistBelief") used for the experiments in (Le et al., 2012) manages to train a neural network using 16000 CPU cores (in 1000 machines) in just a few days, yet this level of resource is likely beyond those available to most deep learning researchers. Less clear still is how to continue scaling significantly beyond this size of network. In this paper we present an alternative approach to training such networks that leverages inexpensive computing power in the form of GPUs and introduces the use of high-speed communications infrastructure to tightly coordinate distributed gradient computations. Our system trains neural networks at scales comparable to DistBelief with just 3 machines. We demonstrate the ability to train a network with

2012: Deep Learning researchers worldwide discover GPUs



2015: ImageNet — Deep Learning achieves superhuman image recognition



"The Microsoft 2016 Conversational Speech Recognition System." W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig. 2016

2016: Microsoft's Deep Learning system achieves new milestone in speech recognition

A New era of computing

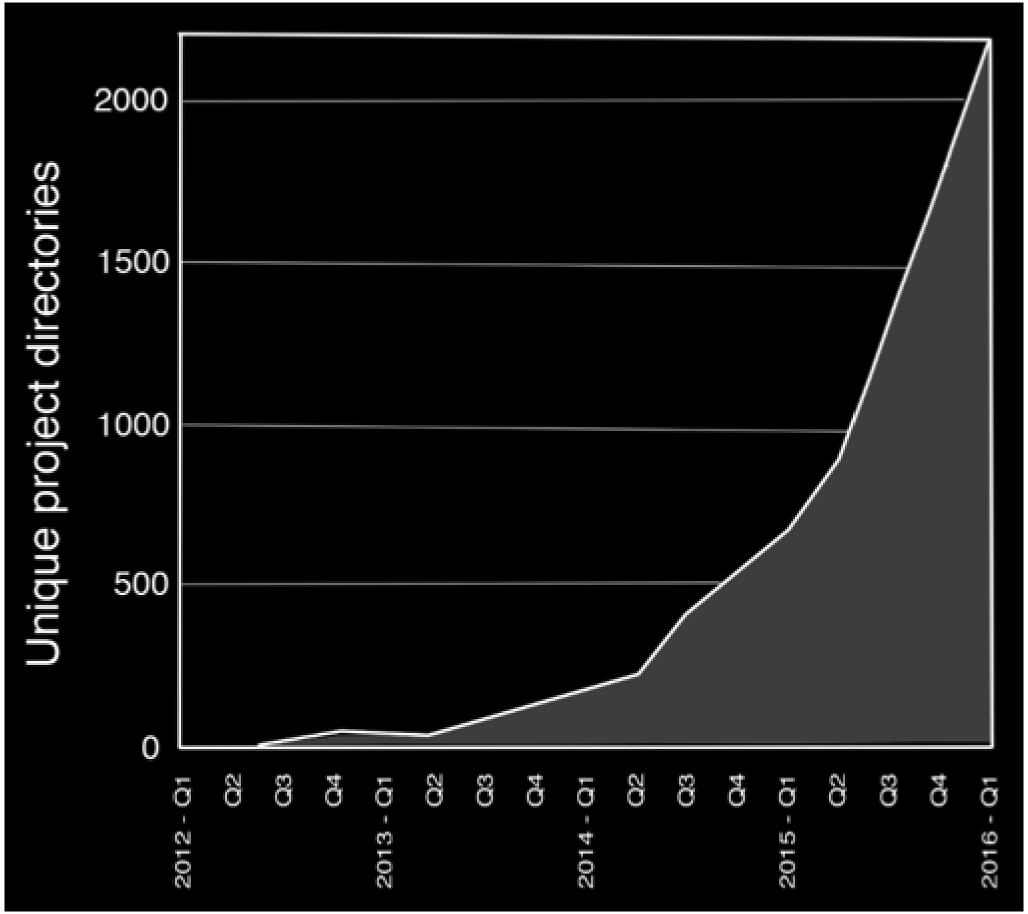
PC INTERNET

MOBILE-CLOUD

AI & INTELLIGENT DEVICES

Deep Learning Explodes at Google

Android apps
Drug discovery
Gmail
Image understanding
Maps
Natural language understanding
Photos
Robotics research
Speech
Translation
YouTube

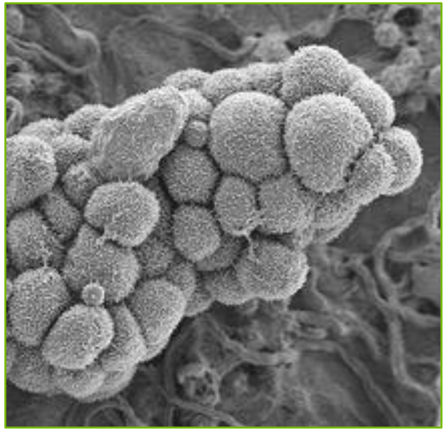


Jeff Dean's talk at TiECon, May 7, 2016

Deep Learning Everywhere

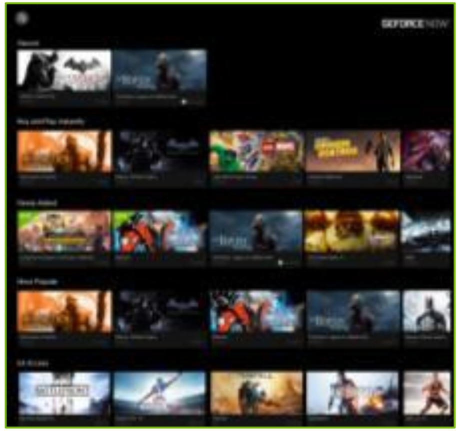
INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery



MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Now "Superhuman" at Many Tasks

Speech recognition

Image classification and detection

Face recognition

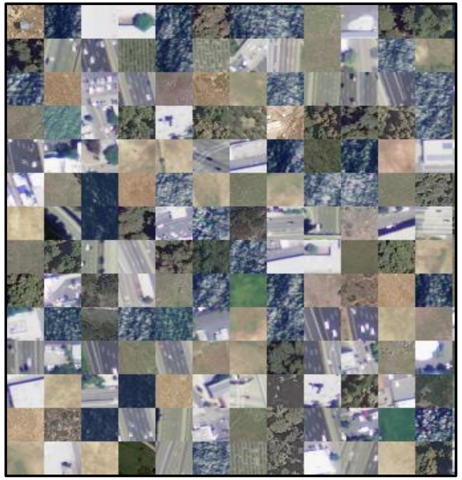
Playing Atari games

Playing Go

Deep Learning Enables Science

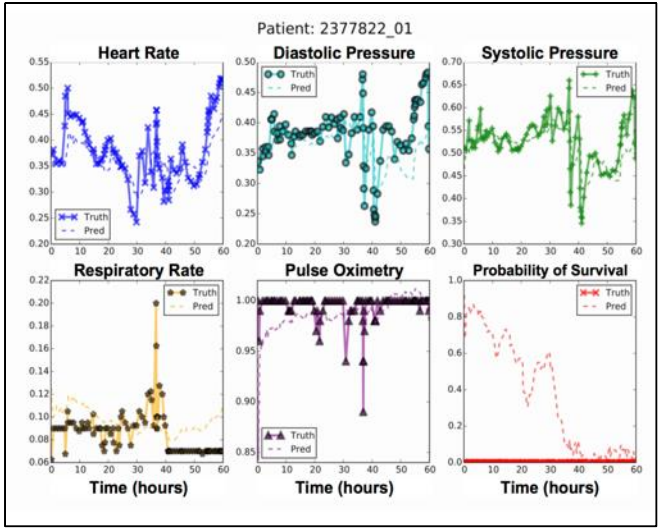
Deep learning enables SCIENCE

NASA AMES



Classify Satellite Images for Carbon Monitoring

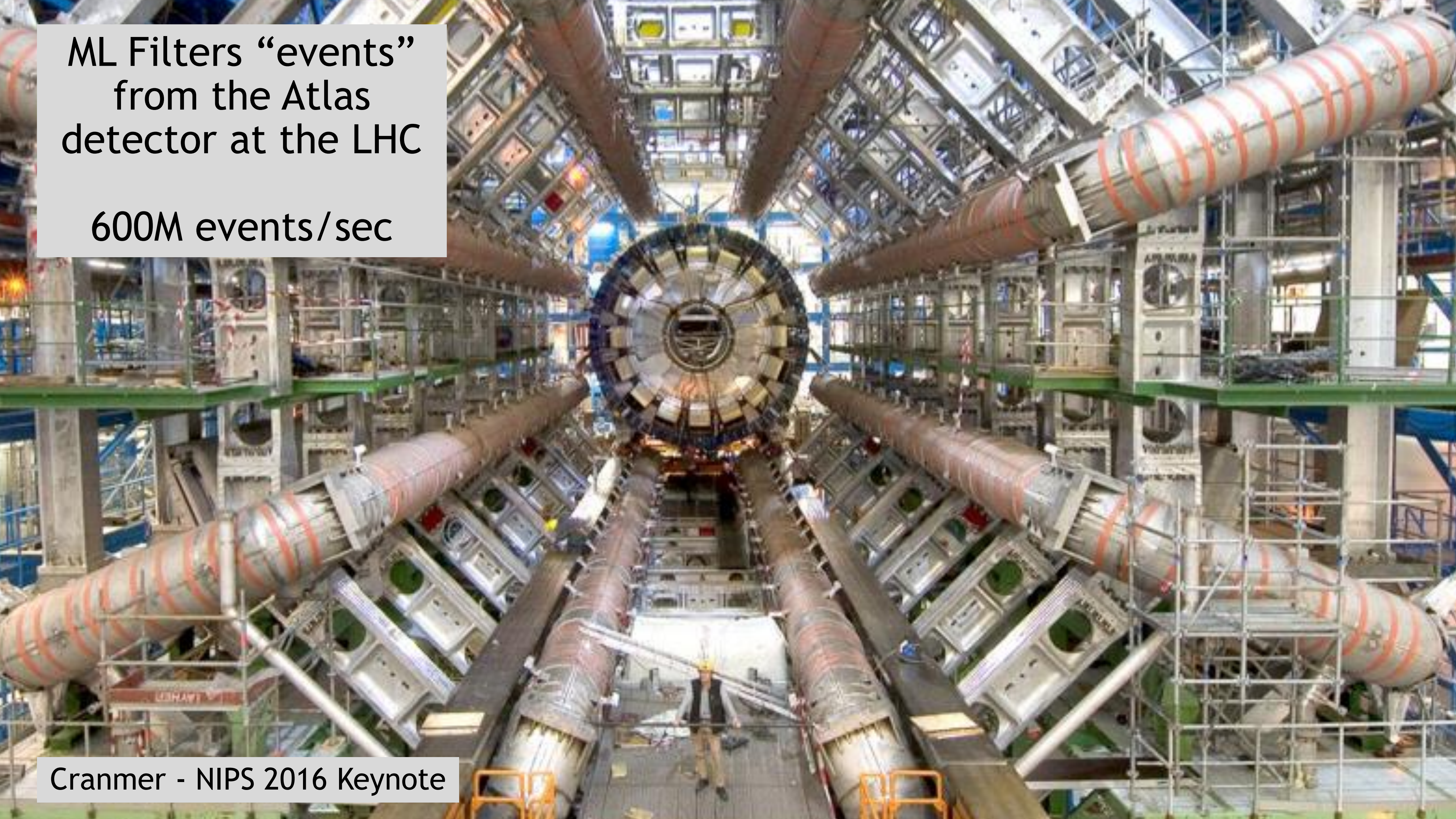
Children's Hospital LOS ANGELES
We Treat Kids Better



Determine Drug Treatments to Increase Child's Chance of Survival

OAK RIDGE
National Laboratory

Analyze Obituaries on the Web for Cancer-related Discoveries

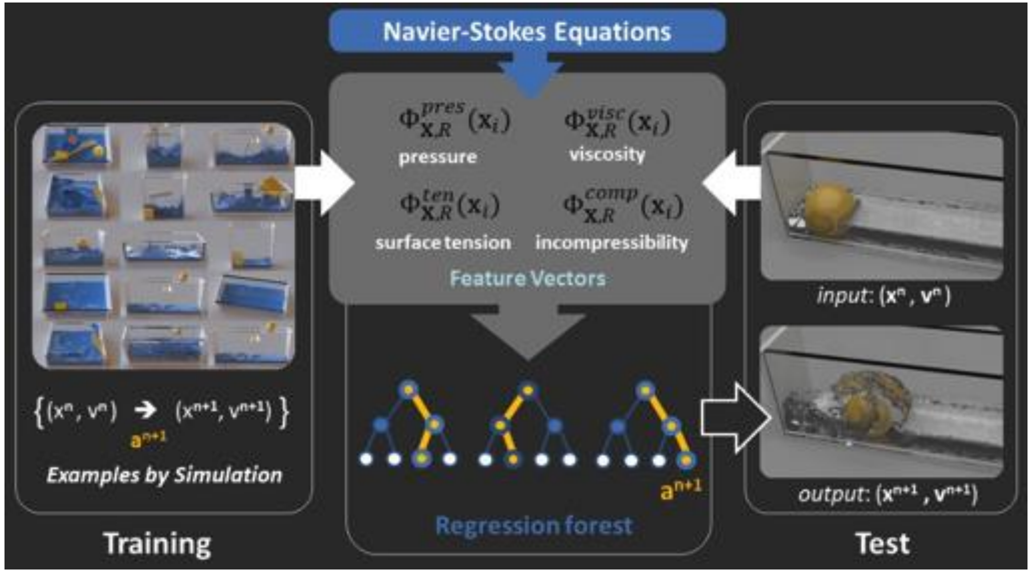


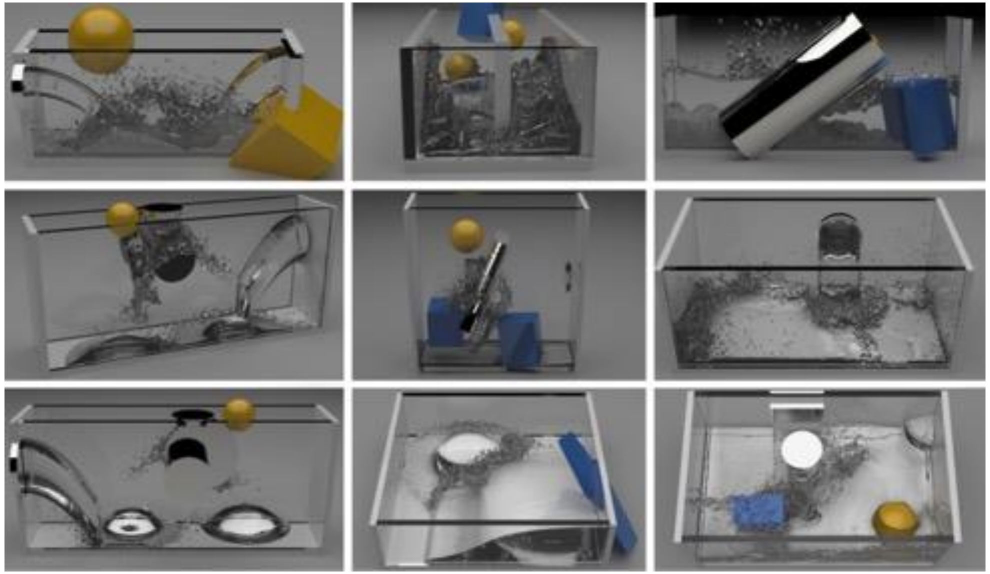
ML Filters “events” from the Atlas detector at the LHC

600M events/sec

Cranmer - NIPS 2016 Keynote

Using ML to Approximate Fluid Dynamics





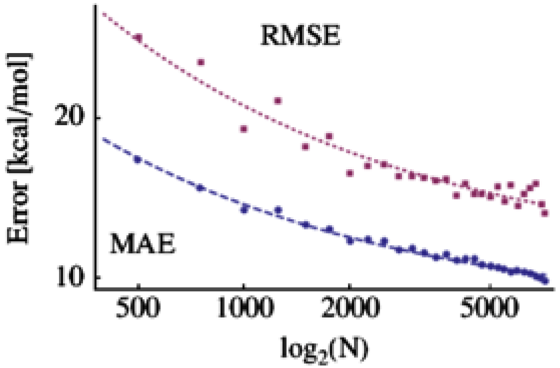
"… Implementation led to a speed-up of one to three orders of magnitude compared to the state-of-the-art position-based fluid solver and runs in real-time for systems with up to 2 million particles"

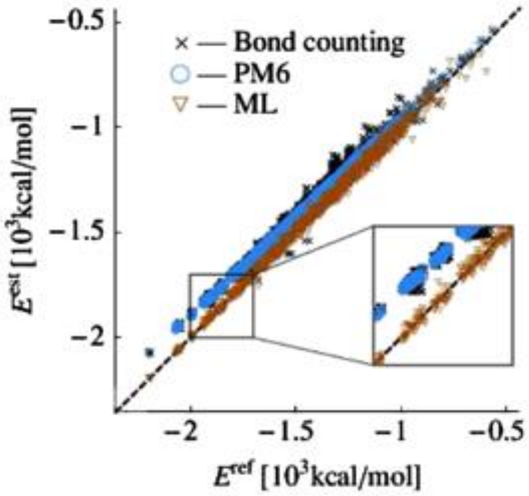
"Data-driven Fluid Simulations using Regression Forests" http://people.inf.ethz.ch/ladickyl/fluid_sigasia15.pdf

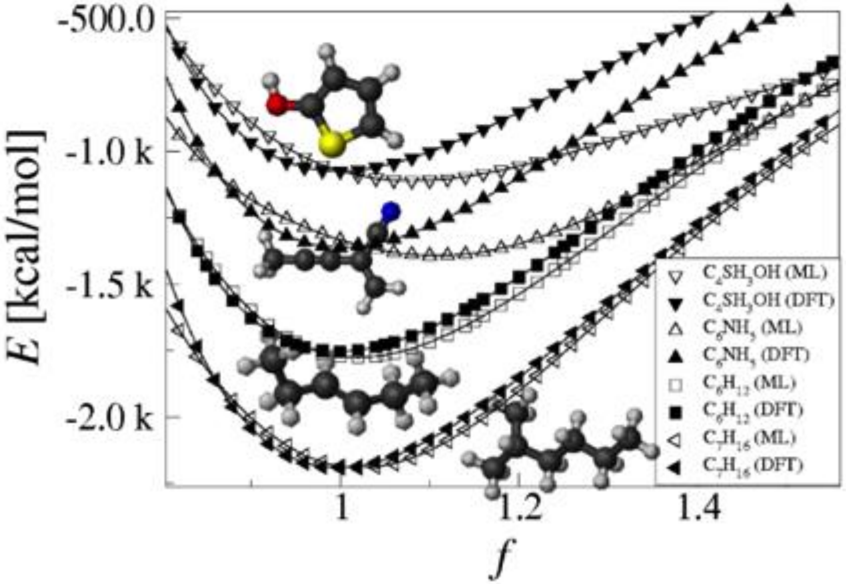
Fluid Simulation with CNNs

Tompson et al. “Accelerating Eulerian Fluid Simulation With Convolutional Networks,” arXiv preprint, 2016

Using ML to Approximate Schrodinger Equation





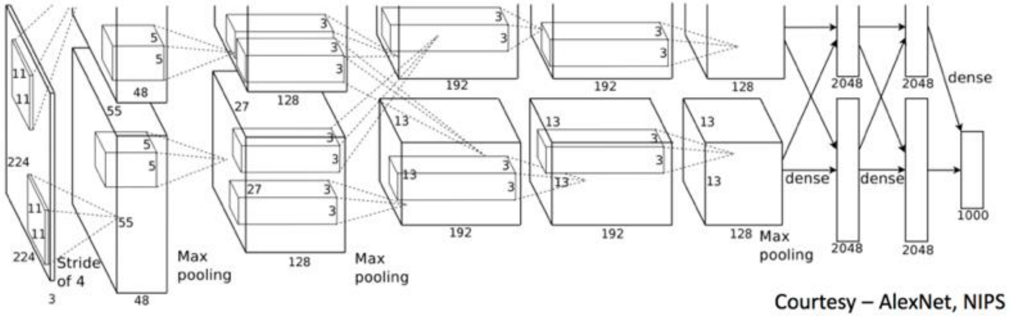


"For larger training sets, N >= 1000, the accuracy of the ML model becomes competitive with mean-field electronic structure theory—at a fraction of the computational cost."

"Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning", Rupp et al., Physical Letters

Deep Learning has an insatiable demand for computing performance

GPUs enabled Deep Learning



Courtesy – AlexNet, NIPS 2012

GPUs now Gate DL Progress

Important Property of Neural Networks

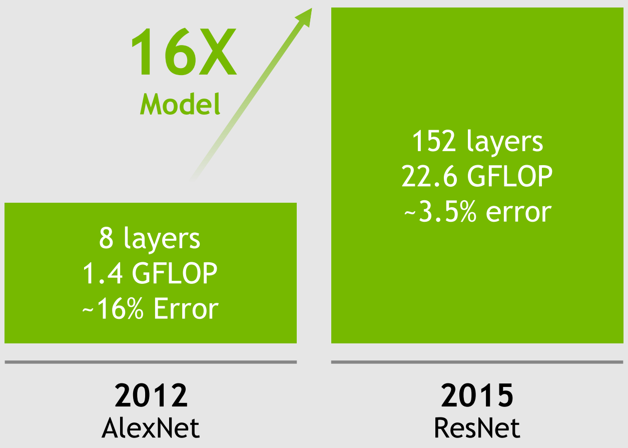
Results get better with

more data +
bigger models +
more computation

(Better algorithms, new insights and improved techniques always help, too!)

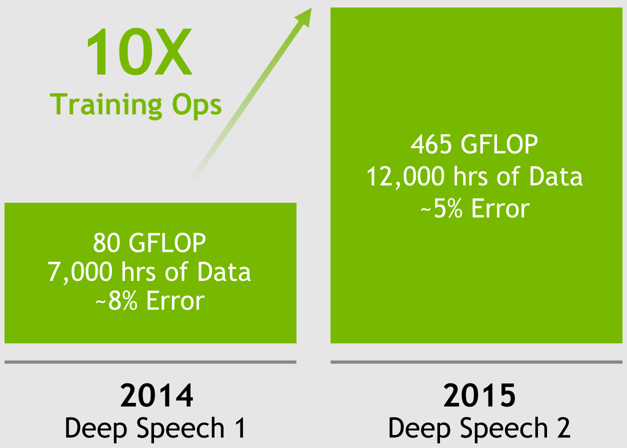
Google

IMAGE RECOGNITION



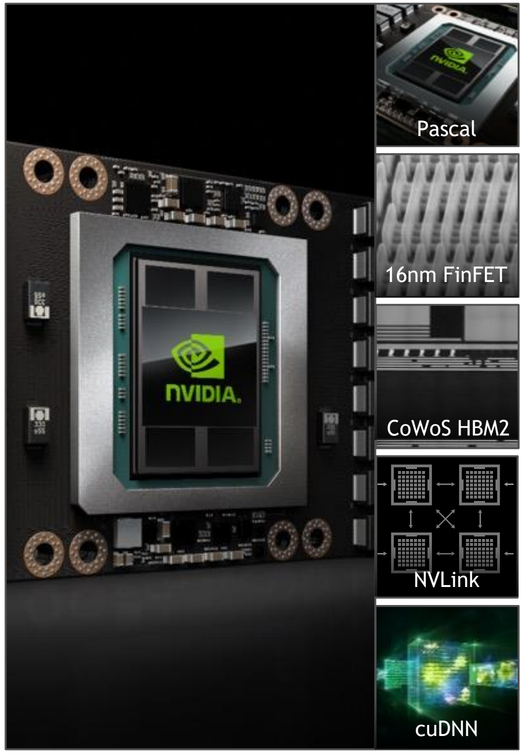
Microsoft

SPEECH RECOGNITION



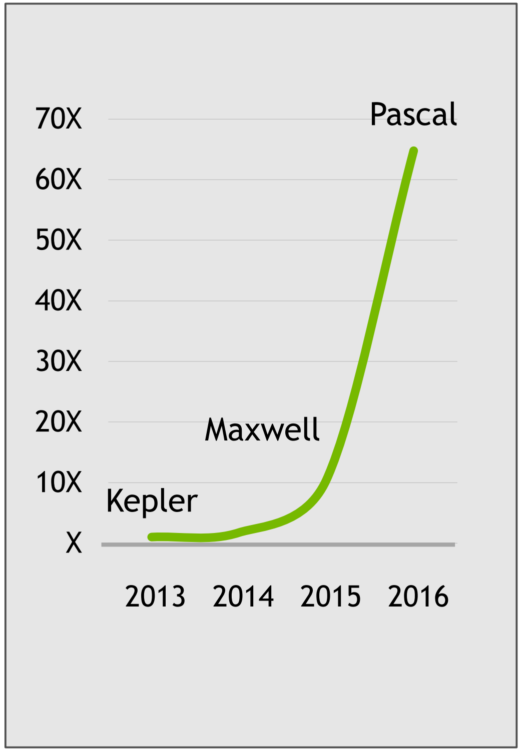
Baidu 百度

Pascal "5 Miracles" Boost Deep Learning 65X



Pascal — 5 Miracles

NVIDIA DGX-1 Supercomputer



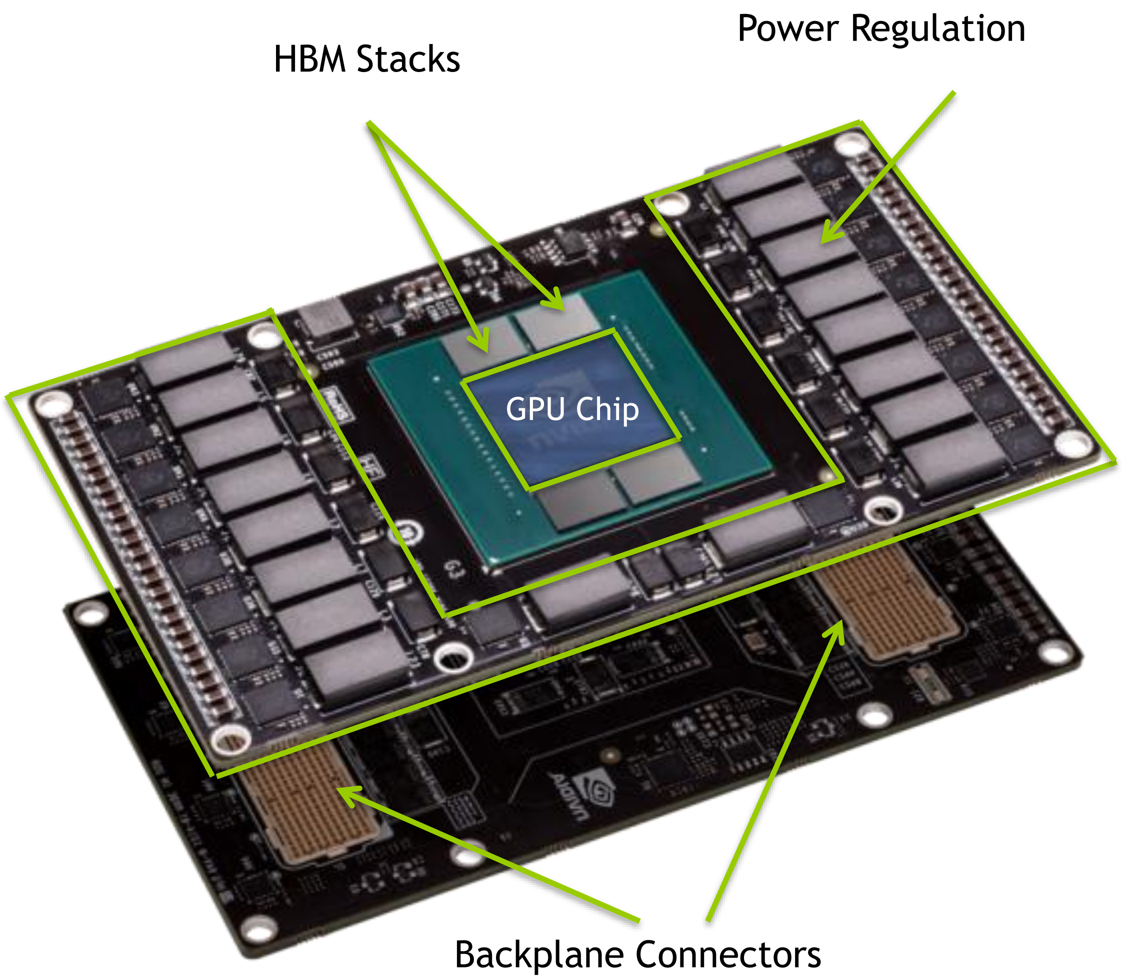
65X in 4 yrs

Caffe
Chainer
Microsoft CNTK
mxnet
PaddlePaddle
Baidu Deep Learning
TensorFlow
theano
torch

Accelerate Every Framework

Chart: Relative speed-up of images/sec vs K40 in 2013. AlexNet training throughput based on 20 iterations. CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory, Ubuntu 14.04. M40 datapoint: 8x M40 GPUs in a node P100: 8x P100 NVLink-enabled.

Pascal GP100



- 10 TeraFLOPS FP32
- 20 TeraFLOPS FP16
- 16GB HBM – 750GB/s
- 300W TDP
- 67GFLOPS/W (FP16)
- 16nm process
- 160GB/s NV Link

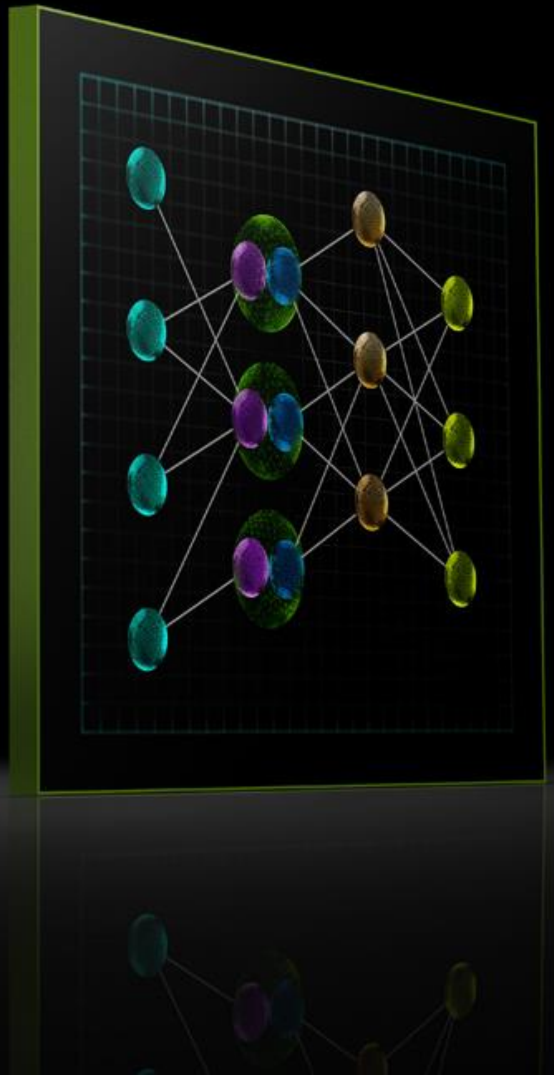
TESLA P4 & P40

INFERENCING ACCELERATORS

Pascal Architecture | INT8

P4 : 50W | 40X Energy Efficient versus CPU

P40: 250W | 40X Performance versus CPU



TensorRT

PERFORMANCE OPTIMIZING INFERENCING ENGINE

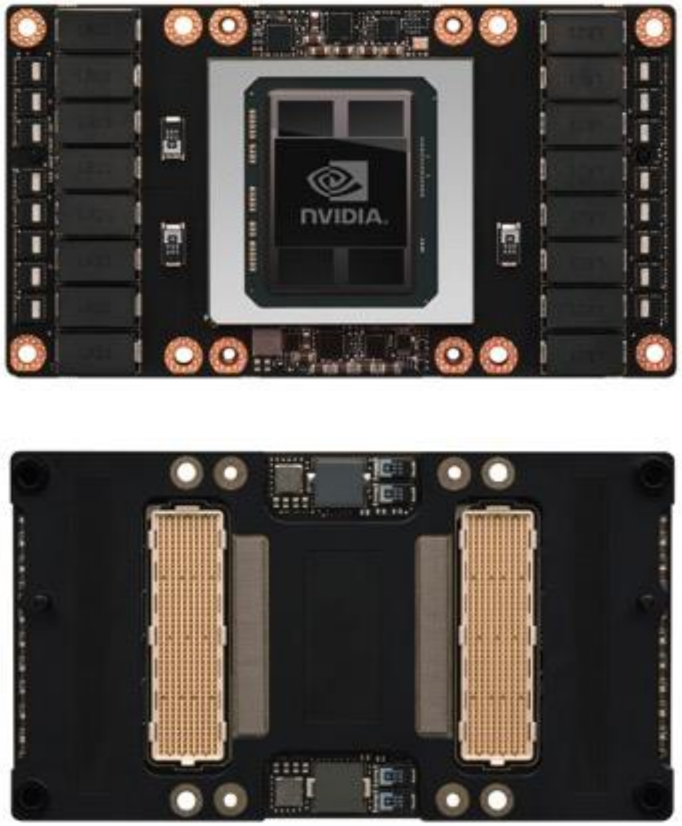
FP32, FP16, INT8 | Vertical & Horizontal Fusion | Auto-Tuning

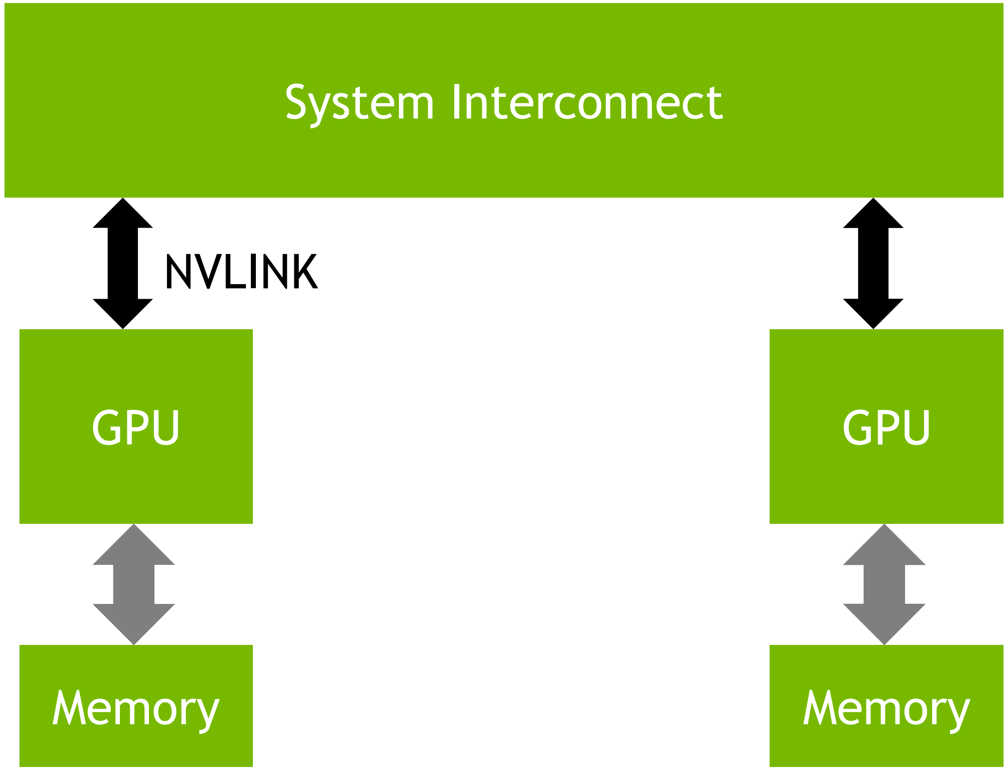
VGG, GoogLeNet, ResNet, AlexNet & Custom Layers

Available Today: developer.nvidia.com/tensorrt

NVLINK enables scalability

NVLINK - Enables Fast Interconnect, PGAS Memory





NVIDIA DGX-1

WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER

170 TFLOPS

8x Tesla P100 16GB

NVLink Hybrid Cube Mesh

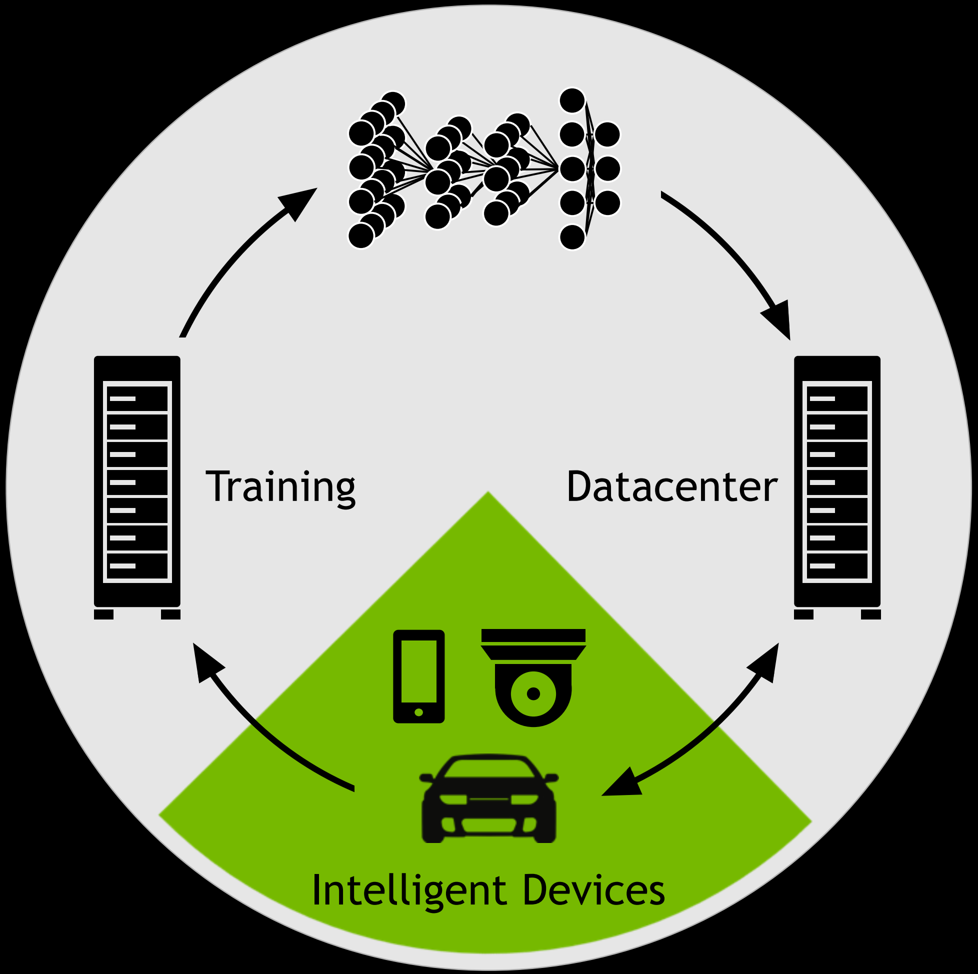
Optimized Deep Learning Software

Dual Xeon

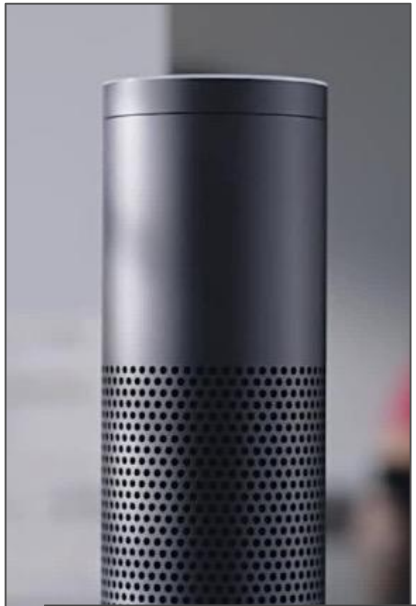
7 TB SSD Deep Learning Cache

Dual 10GbE, Quad IB 100Gb

3RU – 3200W

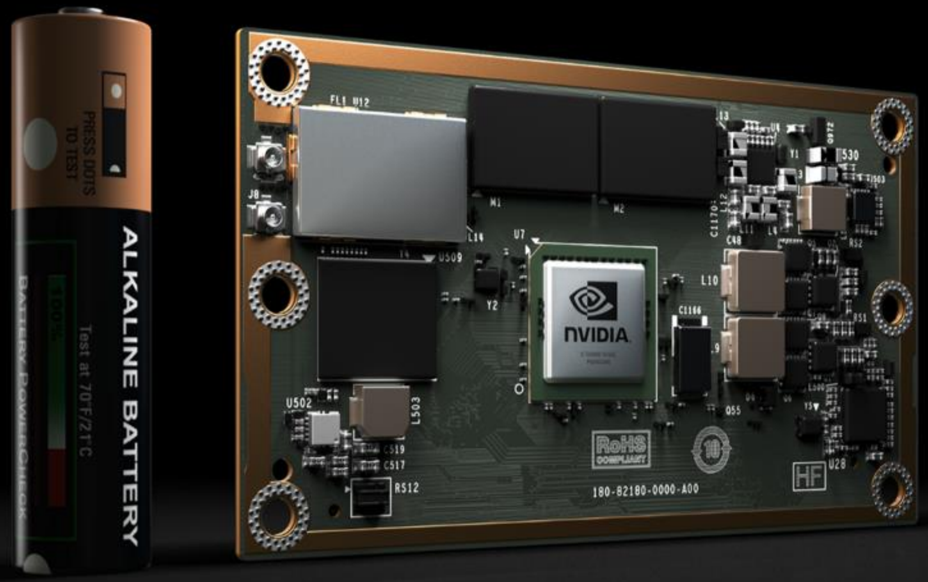

Training
Datacenter
Intelligent Devices

“Billions of INTELLIGENT devices”



“Billions of intelligent devices will take advantage of DNNs to provide personalization and localization as GPUs become faster and faster over the next several years.”

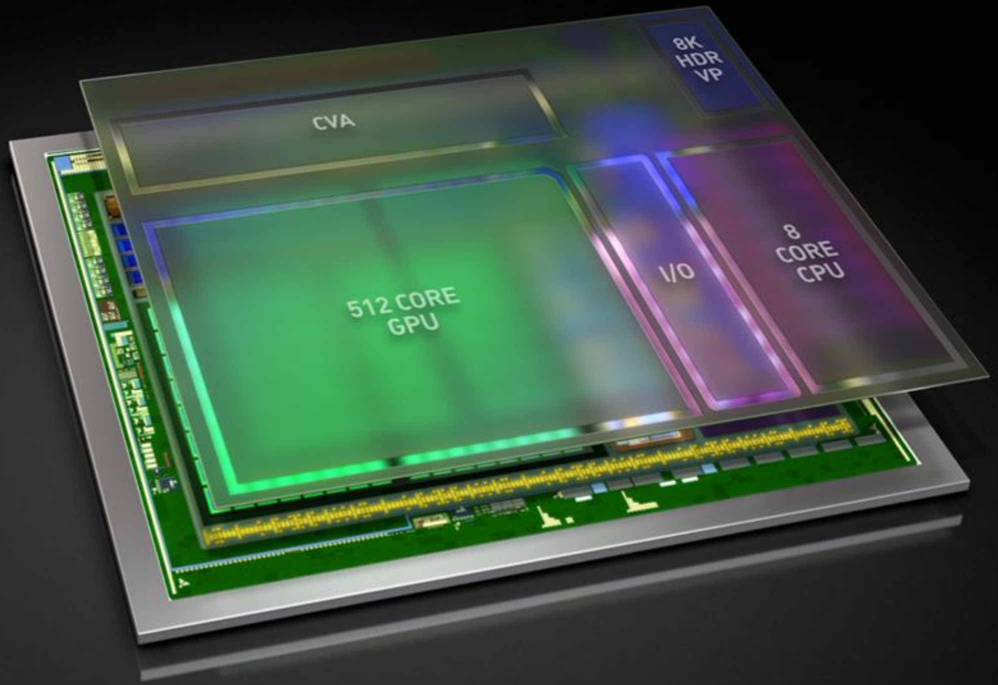
— Tractica



JETSON TX1

EMBEDDED AI SUPERCOMPUTER

10W | 1 TF FP16 | >20 images/sec/W



INTRODUCING XAVIER

AI SUPERCOMPUTER SOC

7 Billion Transistors 16nm FF

8 Core Custom ARM64 CPU

512 Core Volta GPU

New Computer Vision Accelerator

Dual 8K HDR Video Processors

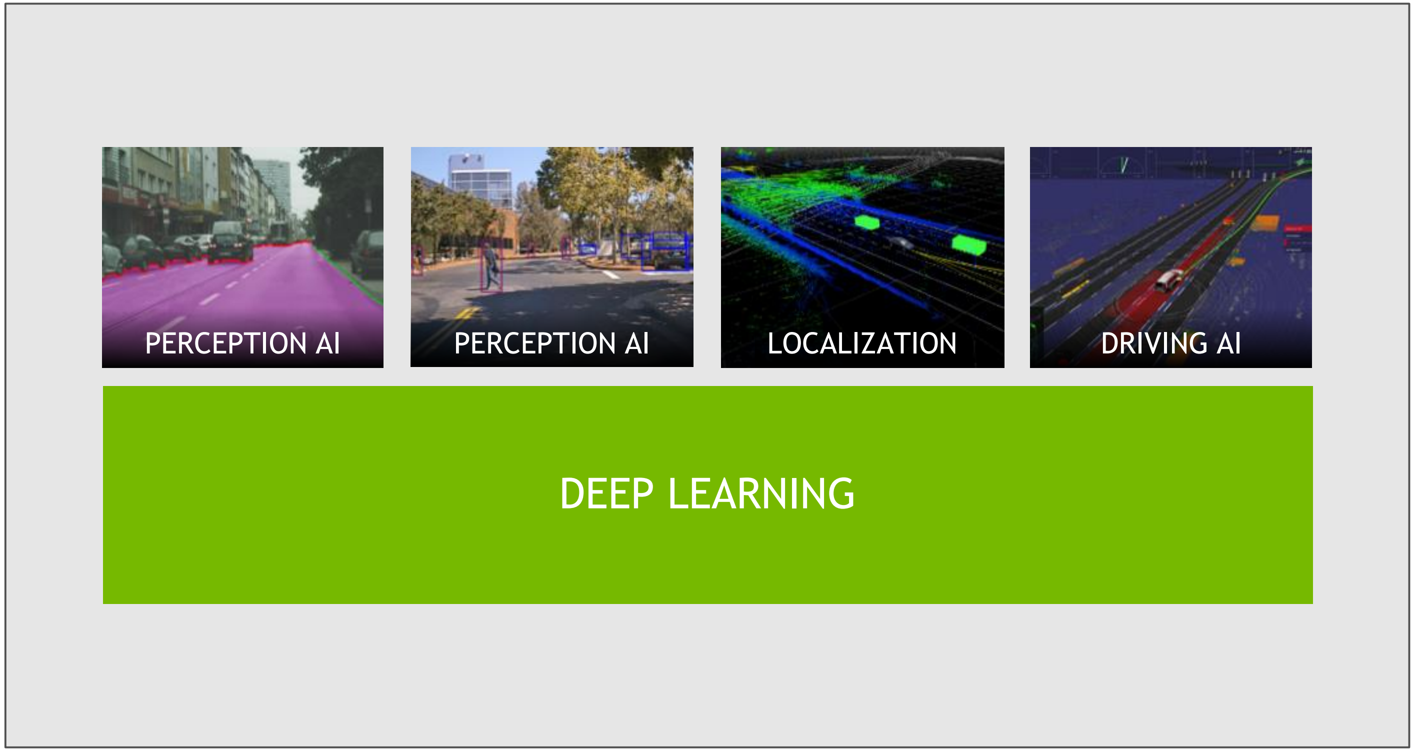
Designed for ASIL C Functional Safety

20 TOPS DL

160 SPECINT

20W

AI TRANSPORTATION — $10T INDUSTRY

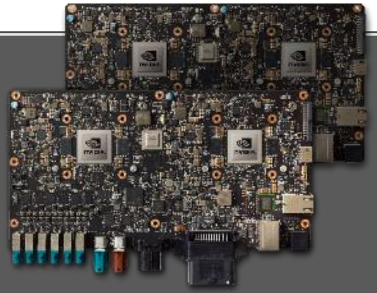


NVIDIA DRIVE PX 2

AutoCruise to Full Autonomy — One Architecture

AutoCruise

AutoChauffeur



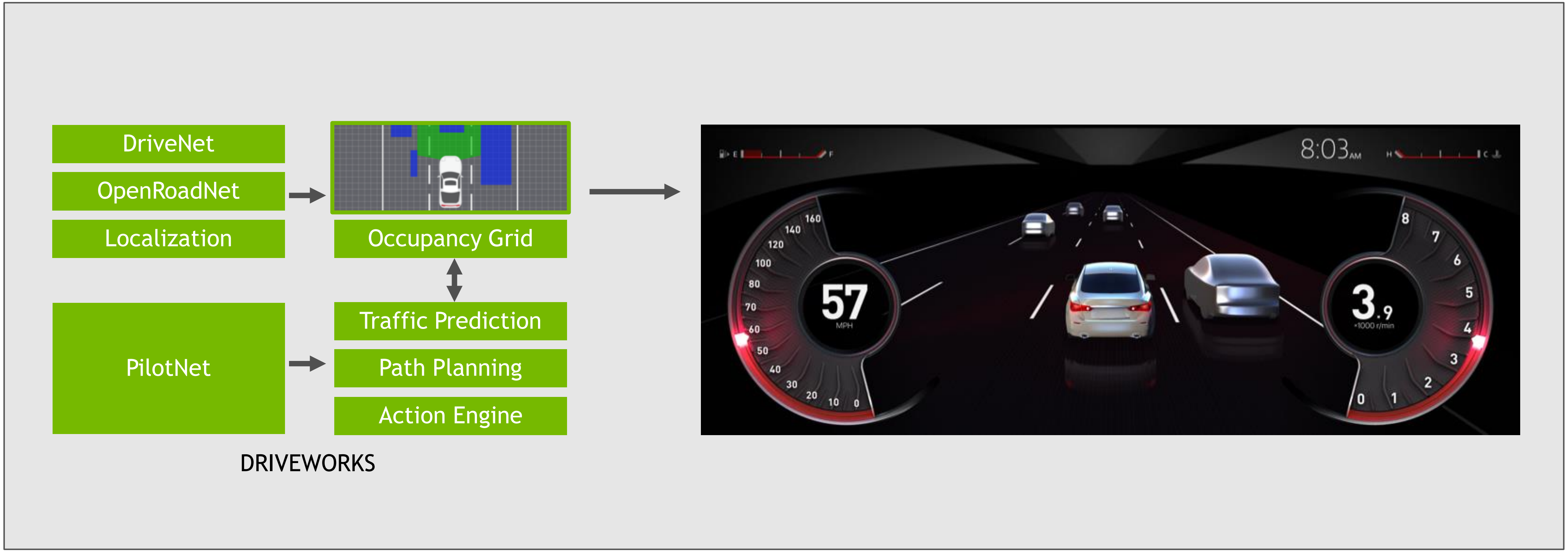
Full Autonomy

AUTONOMOUS DRIVING

Perception, Reasoning, Driving
AI Supercomputing, AI Algorithms, Software
Scalable Architecture

ANNOUNCING Driveworks alpha 1

OS FOR SELF-DRIVING CARS



NVIDIA BB8 AI CAR

Nvidia AI self-driving cars in development

Baidu

nuTonomy

Volvo

TomTom

WEpods

NVAIL

AI Pioneers Pushing state-of-the-art

Carnegie Mellon University

USI/SUPSI IDSIA

Massachusetts Institute of Technology

University of Oxford

NTU

NYU

University of Toronto

Stanford University

Peking University

DFKI German Research Center for Artificial Intelligence

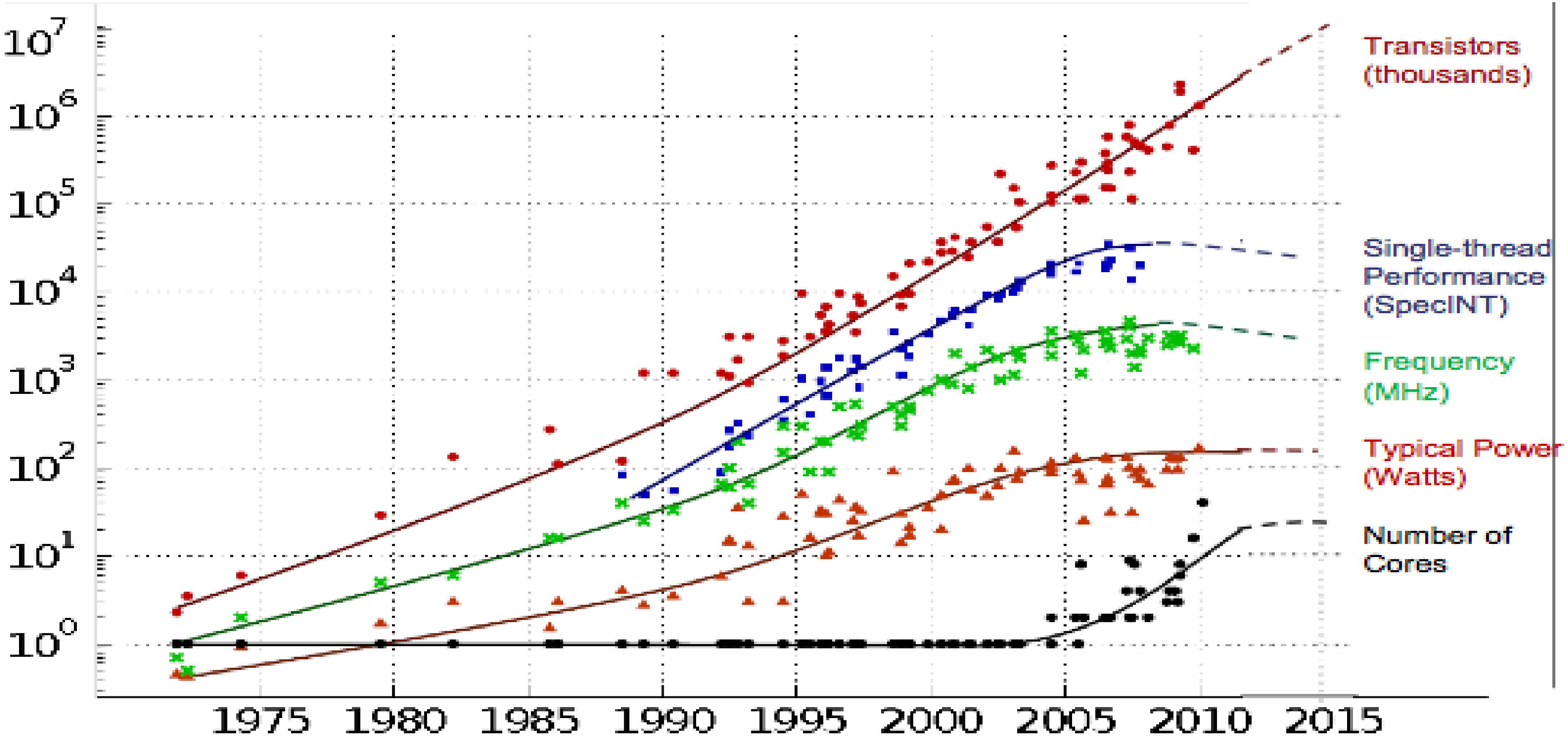
- Reasoning, Attention, Memory — Long-term memory for NN
- End-to-end training for autonomous flight and driving
- Generic agents — Understand and predict behavior
- RNN for long-term dependencies & multiple time scales
- Unsupervised Learning — Generative Models
- Deep reinforcement learning for autonomous AI agents
- Reinforcement learning — Hierarchical and multi-agent
- Semantic 3D reconstruction

東京大学
THE UNIVERSITY OF TOKYO

Yasuo Kuniyoshi
Professor, School of Info Sci & Tech
Director, AI Center (Next Generation Intelligence Science Research Center)
The University of Tokyo

Challenge:
Provide Continued Performance Improvement

But Moore's Law is Over



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

C Moore, *Data Processing in ExaScale-ClassComputer Systems,* Salishan, April 2011

Its not about the FLOPs

- DFMA 0.01mm^2 10pJ/OP – 2GFLOPs

16nm chip, 10mm on a side, 200W

A chip with 10^4 FPUs:

100mm^2

200W

20TFLOPS

Pack 50,000 of these in racks

1EFLOPS

10MW

Overhead

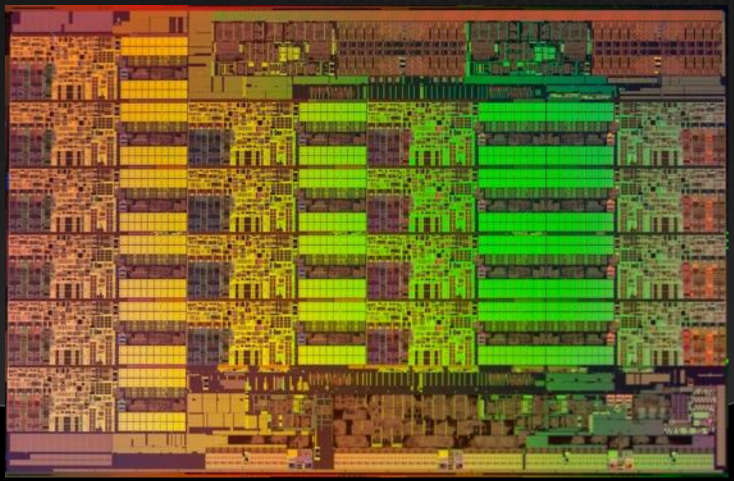
Locality

CPU

126 pJ/flop (SP)

Optimized for Latency

Deep Cache Hierarchy



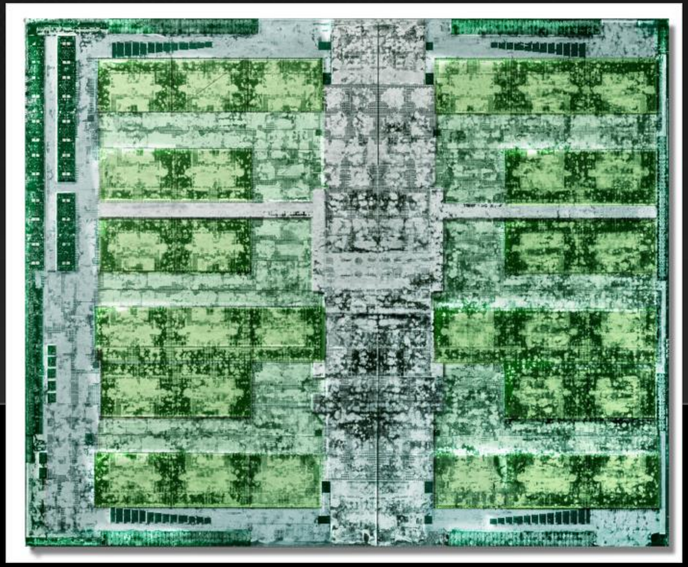
Broadwell E5 v4
14 nm

GPU

28 pJ/flop (SP)

Optimized for Throughput

Explicit Management
of On-chip Memory



Pascal
16 nm

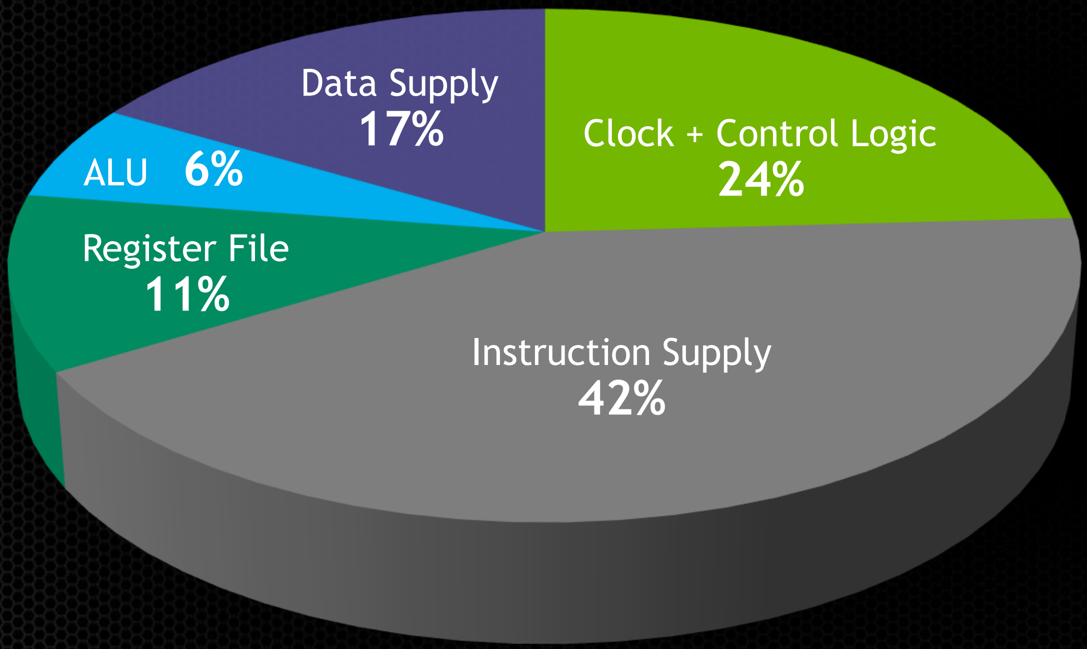
NVIDIA

Fixed-Function Logic is Even More Efficient

	Energy/Op
CPU (scalar)	1.7nJ
GPU	30pJ
Fixed-Function	3pJ

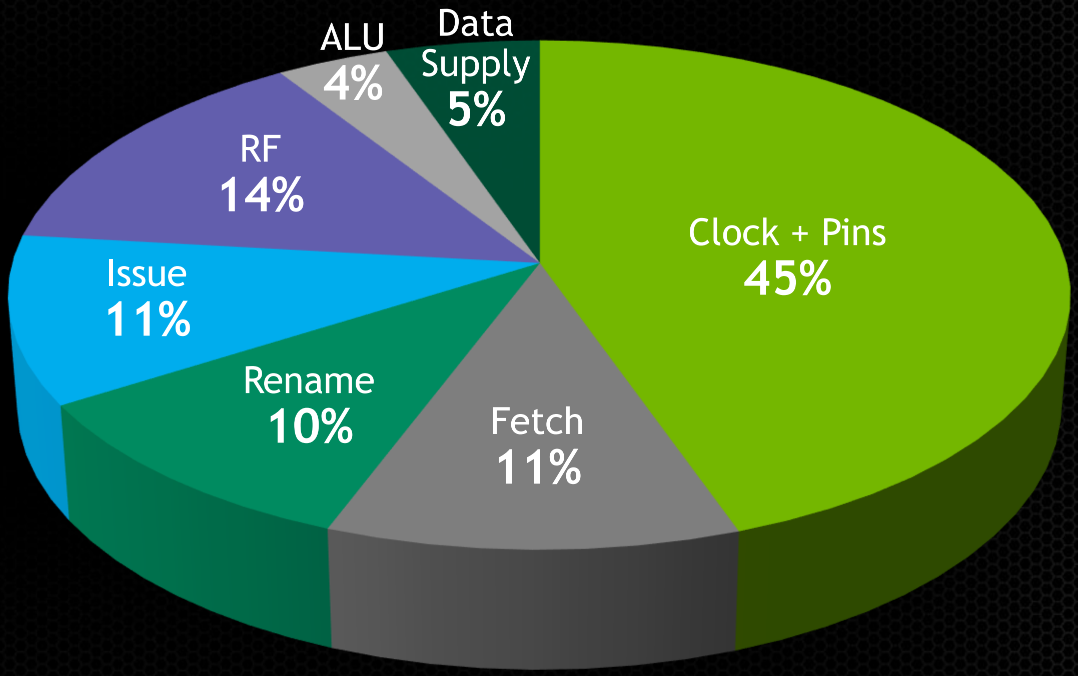
How is Power Spent in a CPU?

In-order Embedded

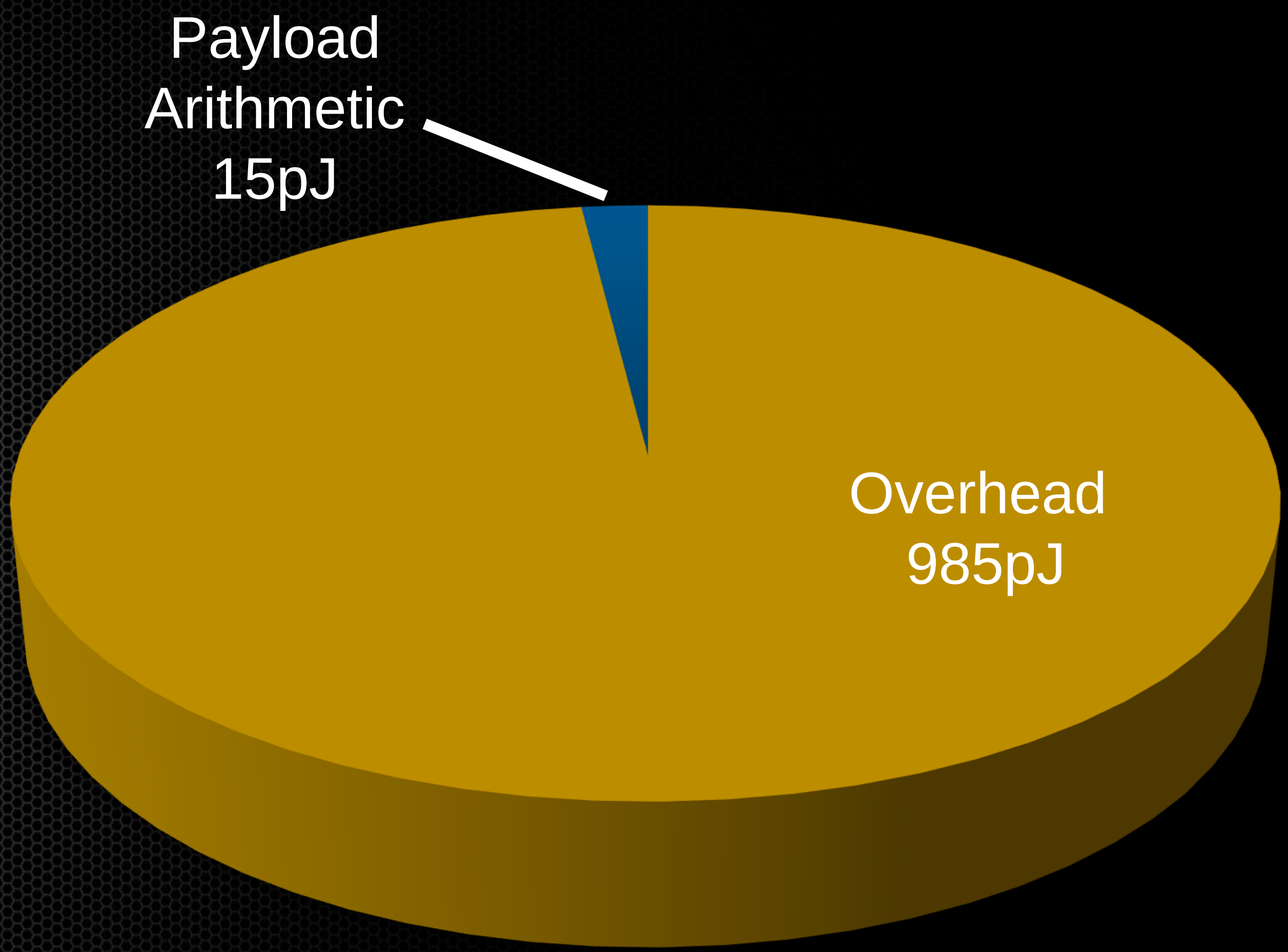


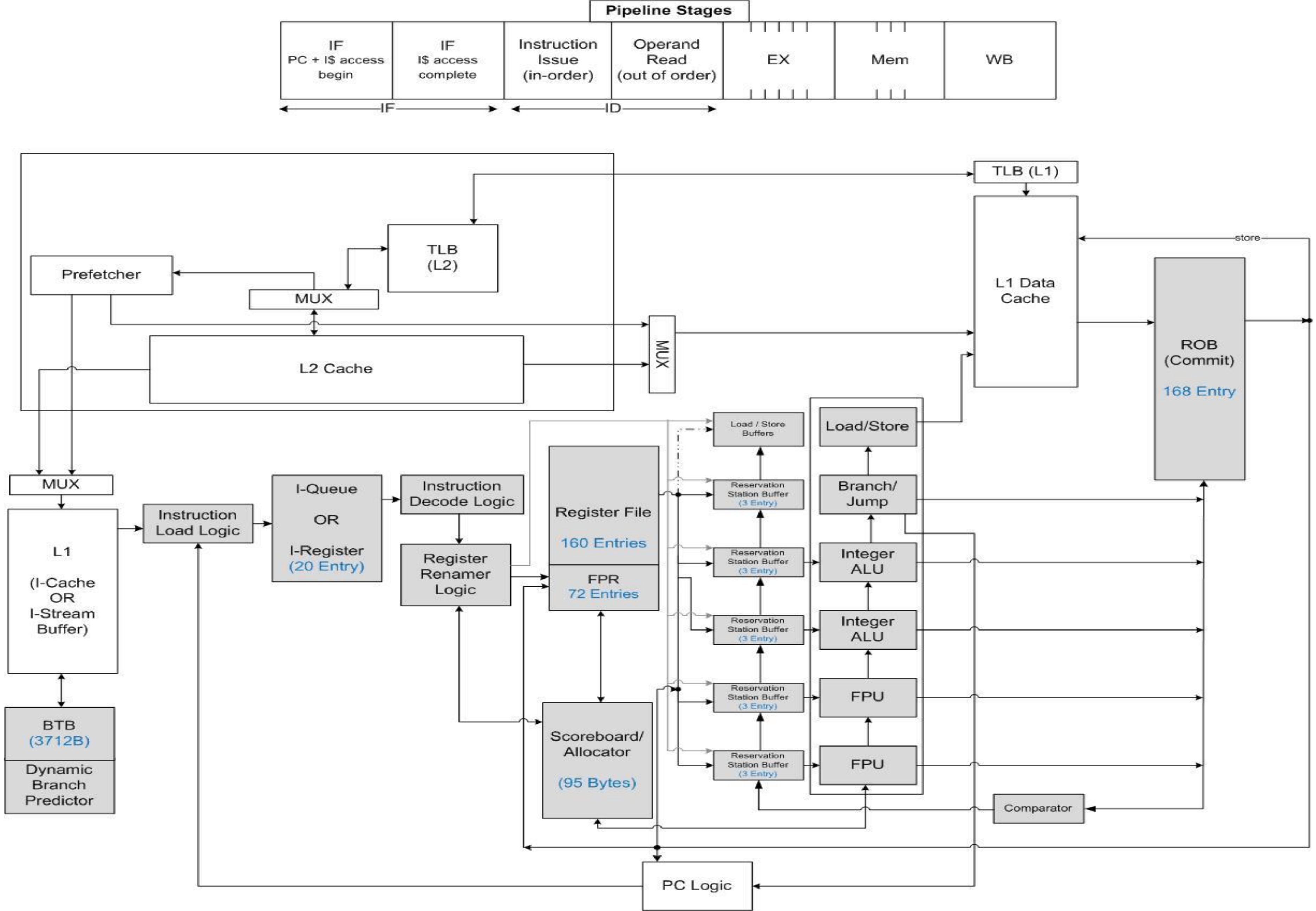
Dally [2008] (Embedded in-order CPU)

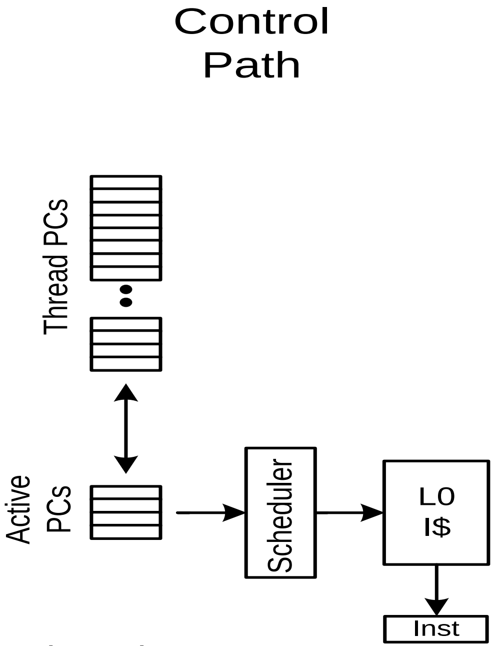
OOO Hi-perf

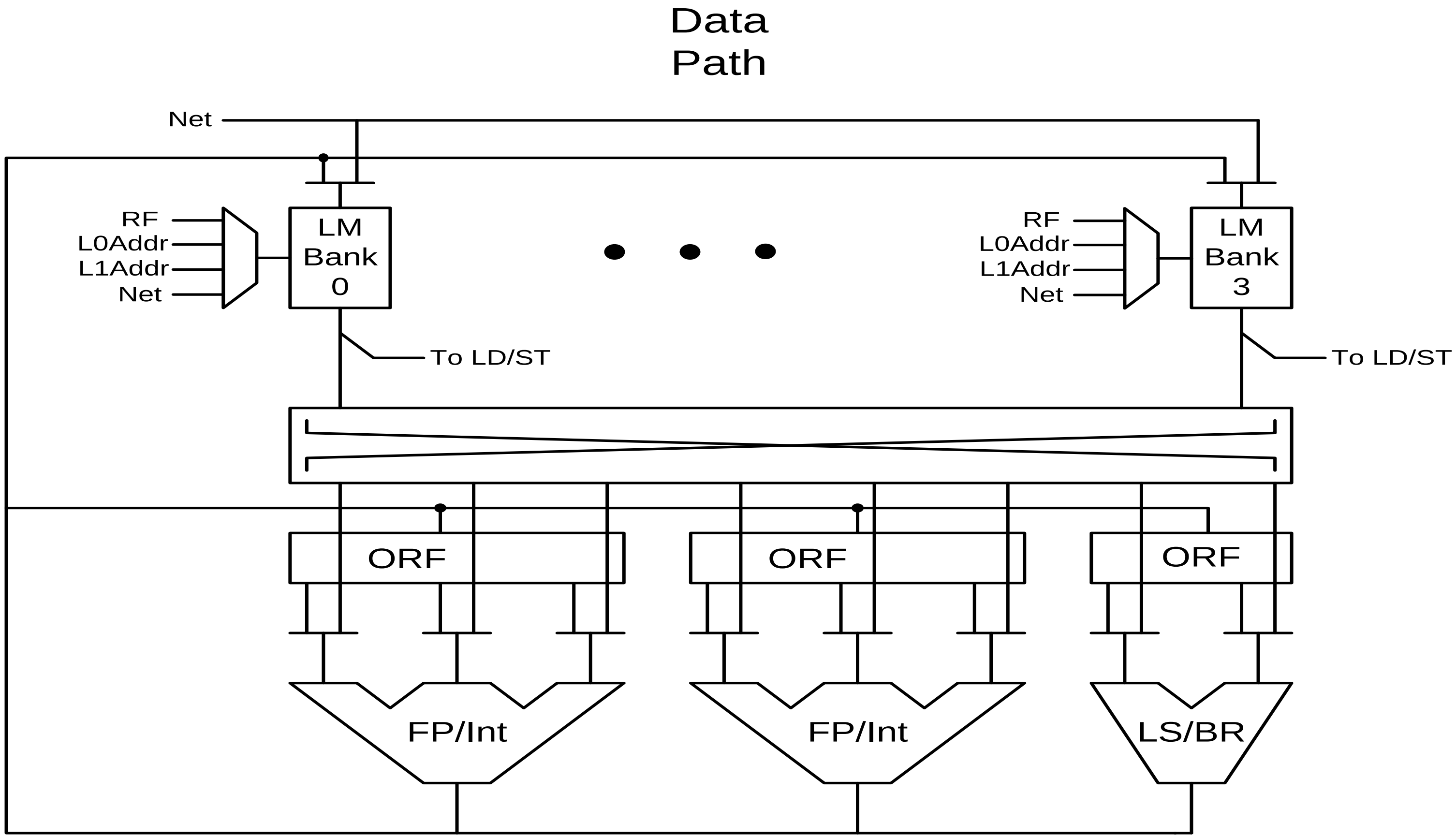


Natarajan [2003] (Alpha 21264)


Payload
Arithmetic
15pJ
Overhead
985pJ

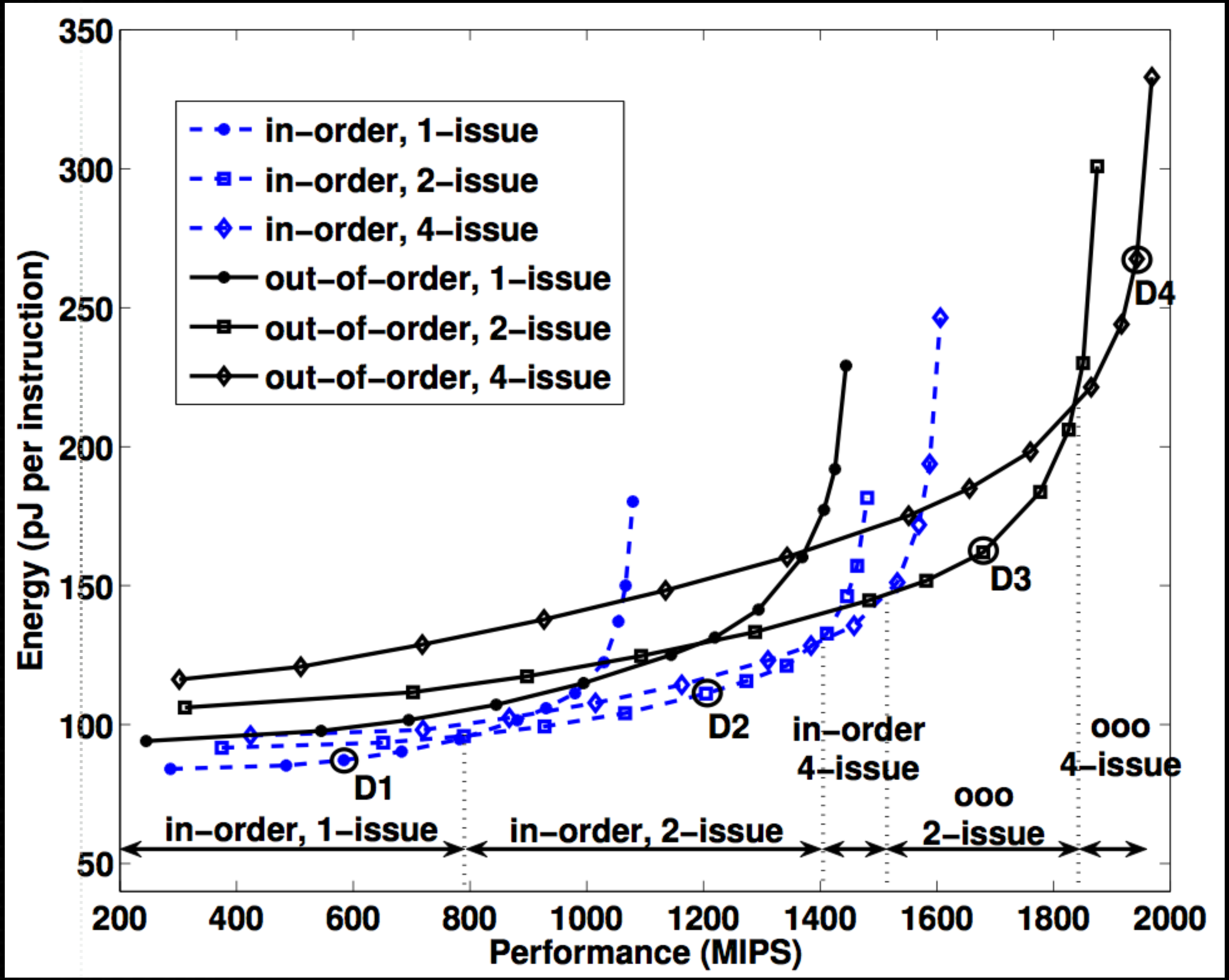

Pipeline Stages
IF PC + I$ access begin
IF I$ access complete
Instruction Issue (in-order)
Operand Read (out of order)
EX
Mem
WB
IF
ID
TLB (L1)
TLB (L2)
Prefetcher
MUX
L2 Cache
MUX
L1 Data Cache
store
ROB (Commit)
168 Entry
Load / Store Buffers
Load/Store
Reservation Station Buffer (3 Entry)
Branch/ Jump
Integer ALU
Integer ALU
FPU
FPU
MUX
L1 (I-Cache OR I-Stream Buffer)
Instruction Load Logic
I-Queue OR I-Register (20 Entry)
Instruction Decode Logic
Register Renamer Logic
Register File 160 Entries
FPR 72 Entries
Scoreboard/ Allocator (95 Bytes)
BTB (3712B)
Dynamic Branch Predictor
Comparator
PC Logic





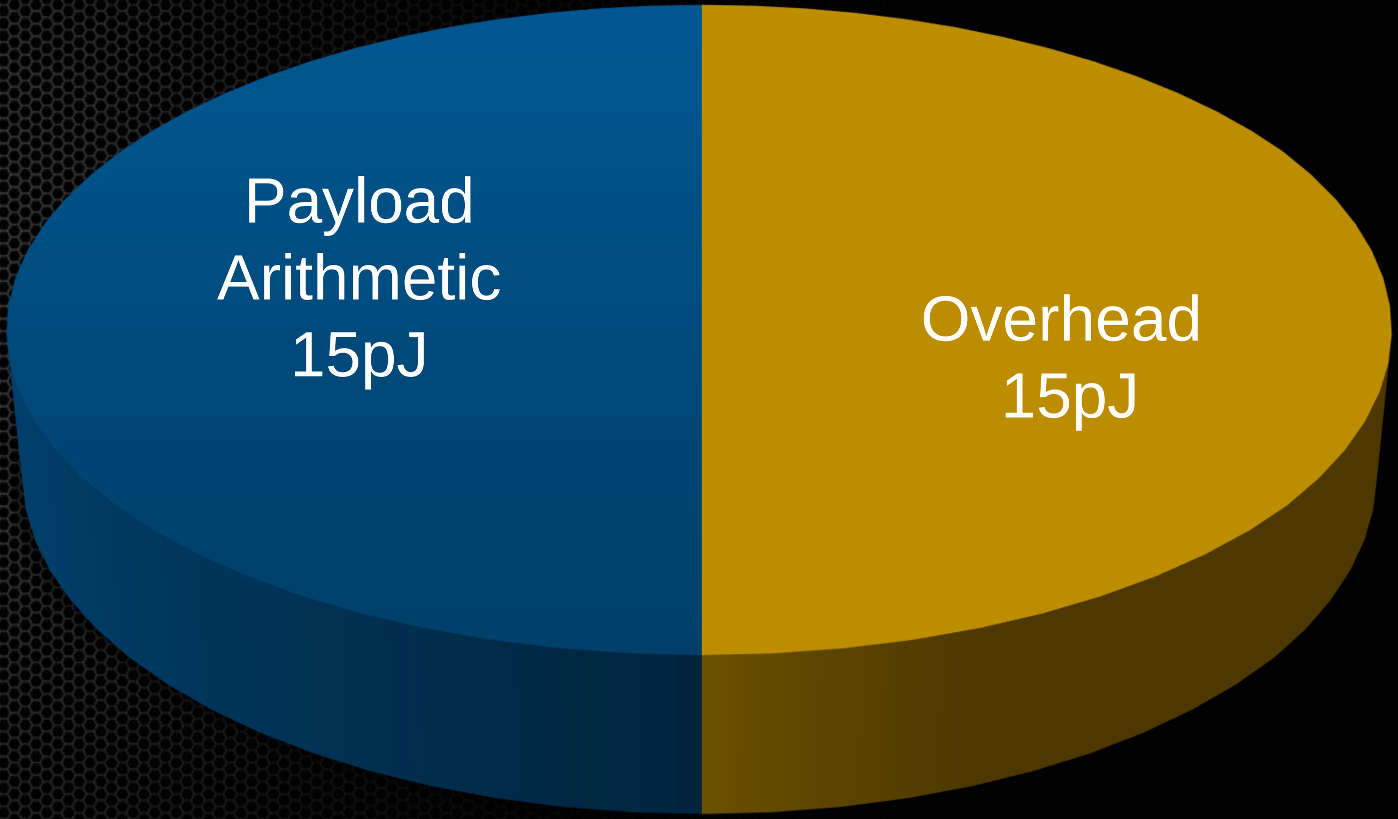
64 threads
4 active threads
2 DFMAs (4 FLOPS/clock)
ORF bank: 16 entries (128 Bytes)
L0 I$: 64 instructions (1KByte)
LM Bank: 8KB (32KB total)

Simpler Cores = Energy Efficiency

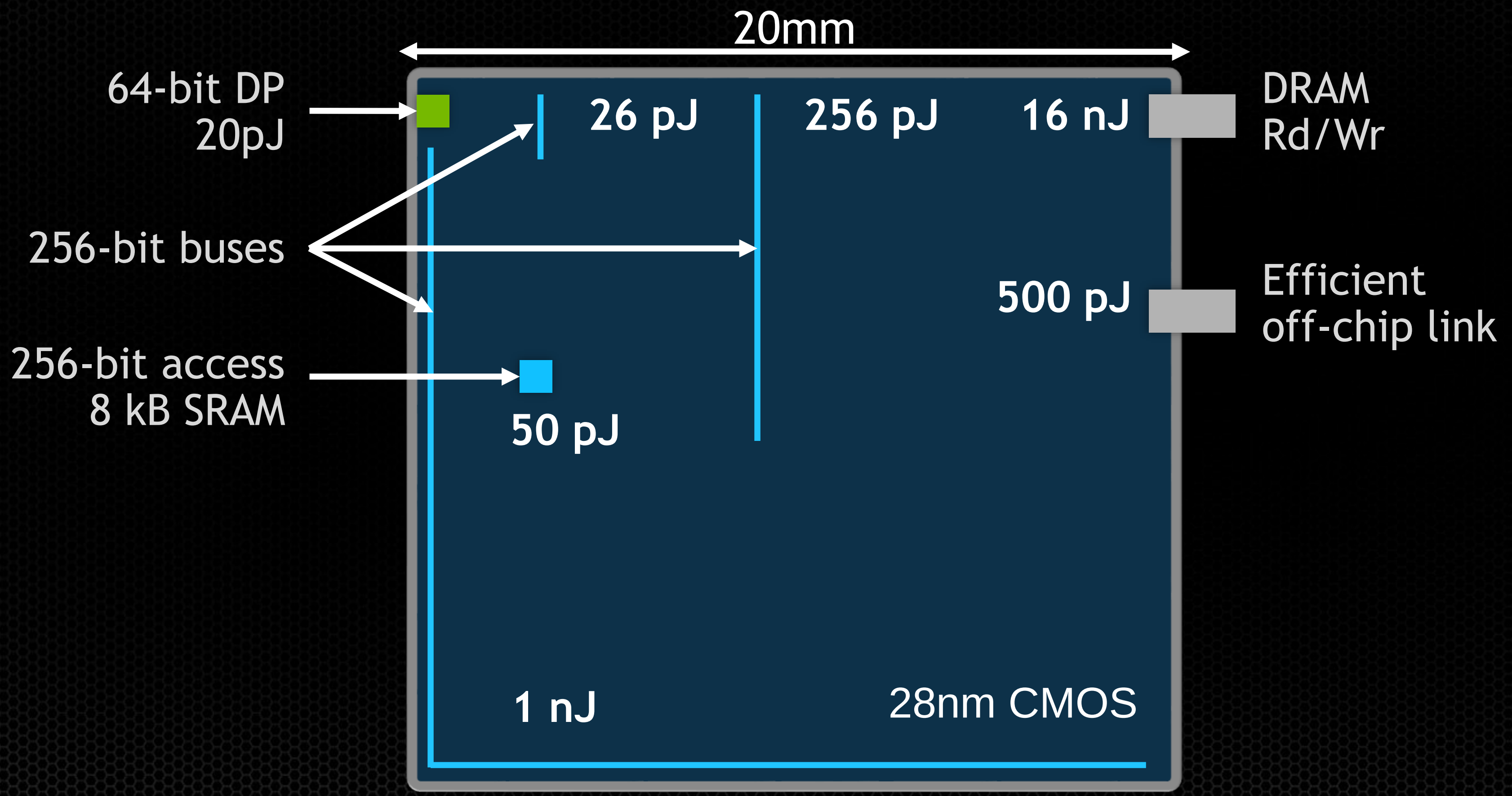


Source: Azizi [PhD 2010]

NVIDIA.


Payload
Arithmetic
15pJ
Overhead
15pJ

Communication Dominates Arithmetic



Energy Shopping List

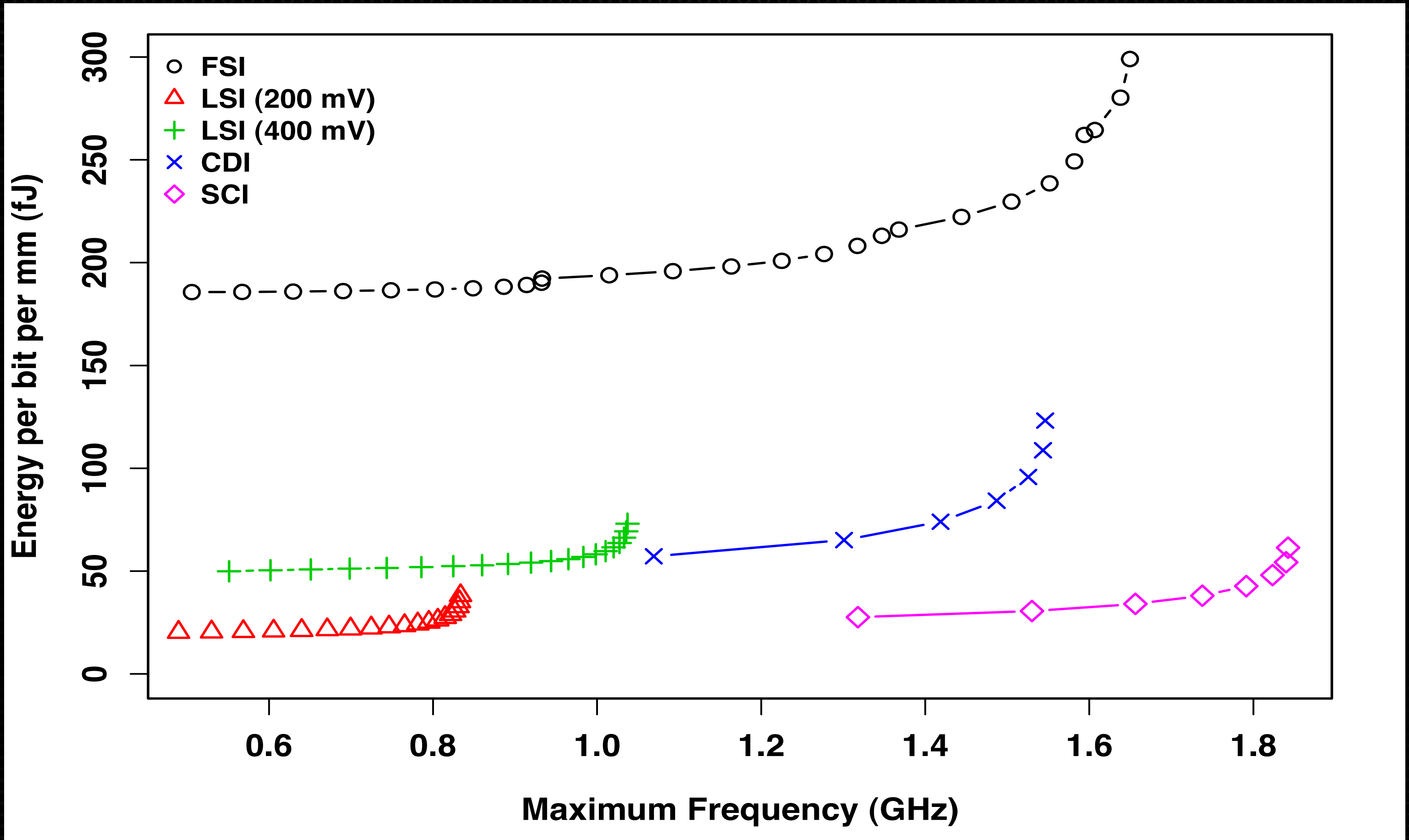
Processor Technology	40 nm	10nm
Vdd (nominal)	0.9 V	0.7 V
DFMA energy	50 pJ	7.6 pJ
64b 8 KB SRAM Rd	14 pJ	2.1 pJ
Wire energy (256 bits, 10mm)	310 pJ	174 pJ

Memory Technology	45 nm	16nm
DRAM interface pin bandwidth	4 Gbps	50 Gbps
DRAM interface energy	20-30 pJ/bit	2 pJ/bit
DRAM access energy	8-15 pJ/bit	2.5 pJ/bit

FP Op lower bound
=
4 pJ

Keckler [Micro 2011], Vogelsang [Micro 2010]

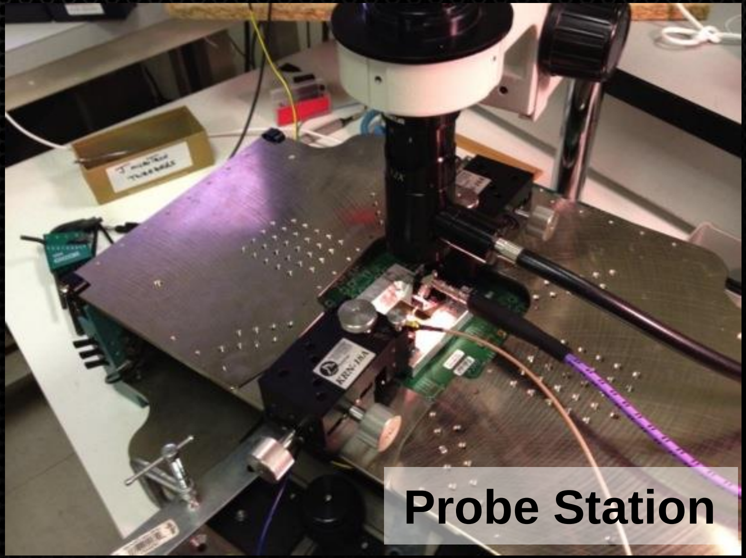
NVIDIA



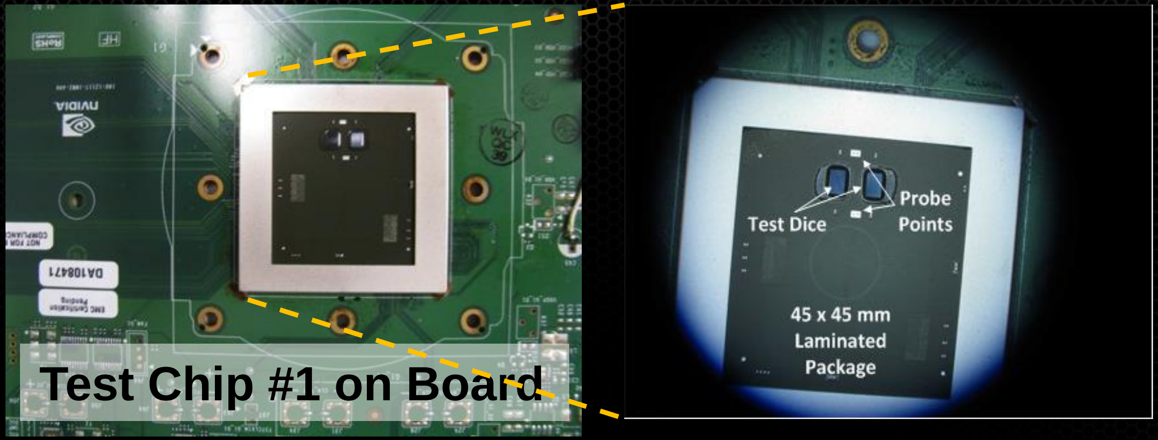
FSI
LSI (200 mV)
LSI (400 mV)
CDI
SCI
Energy per bit per mm (fJ)
0
50
100
150
200
250
300
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Maximum Frequency (GHz)

NVIDIA.

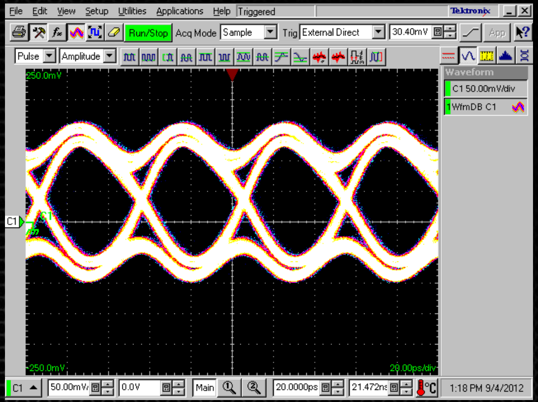
GRS Test Chips



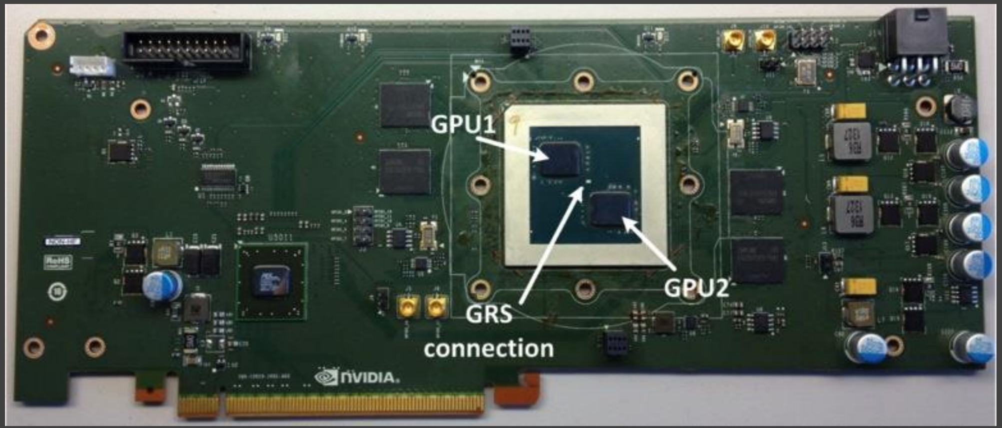
Probe Station



Test Chip #1 on Board



Eye Diagram from Probe



Test Chip #2 fabricated on production GPU

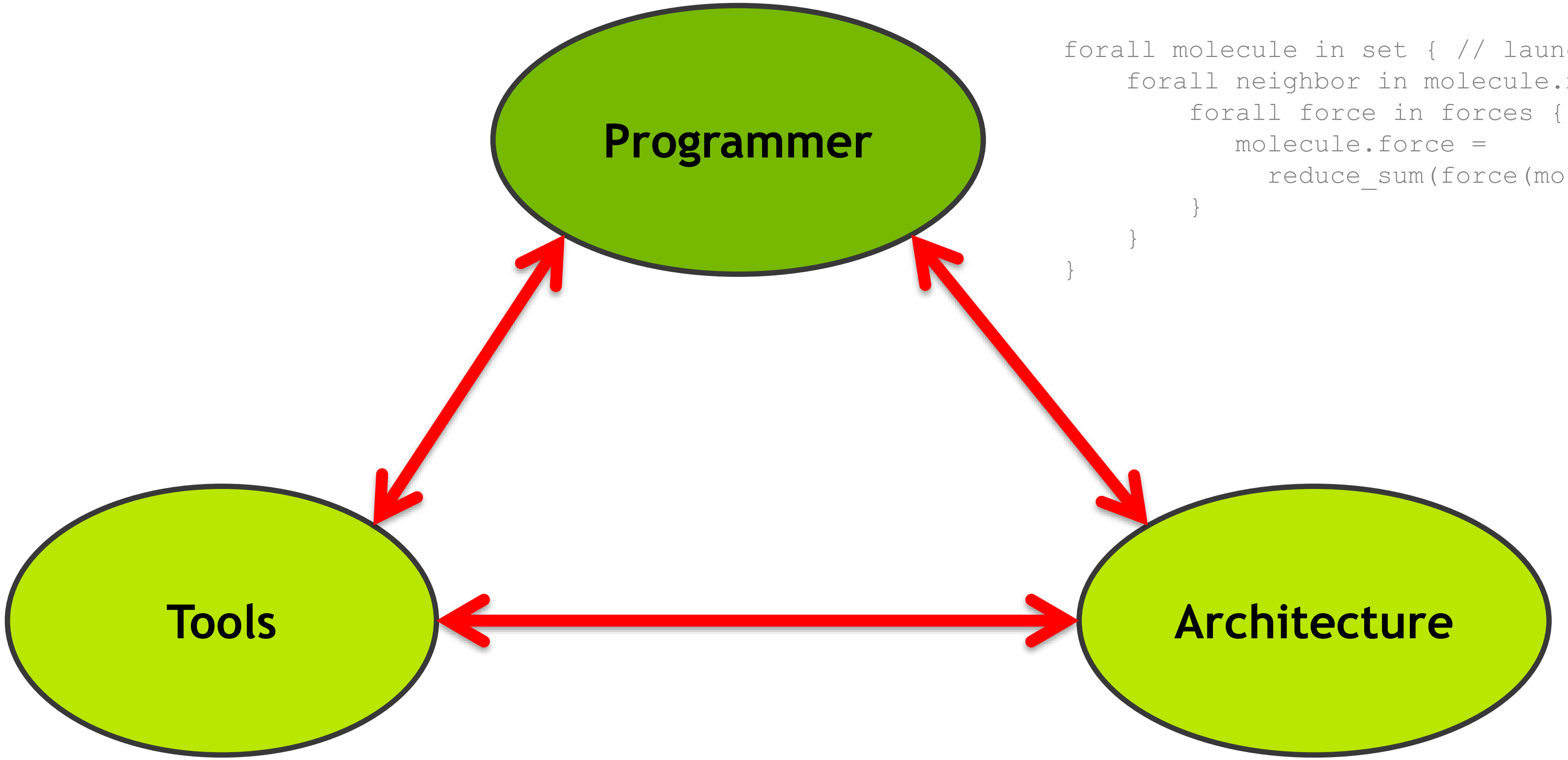
Poulton et al. ISSCC 2013, JSSCC Dec 2013

Efficient Machines

Are Highly Parallel
Have Deep Storage Hierarchies
Have Heterogeneous Processors

Target Independent Programming

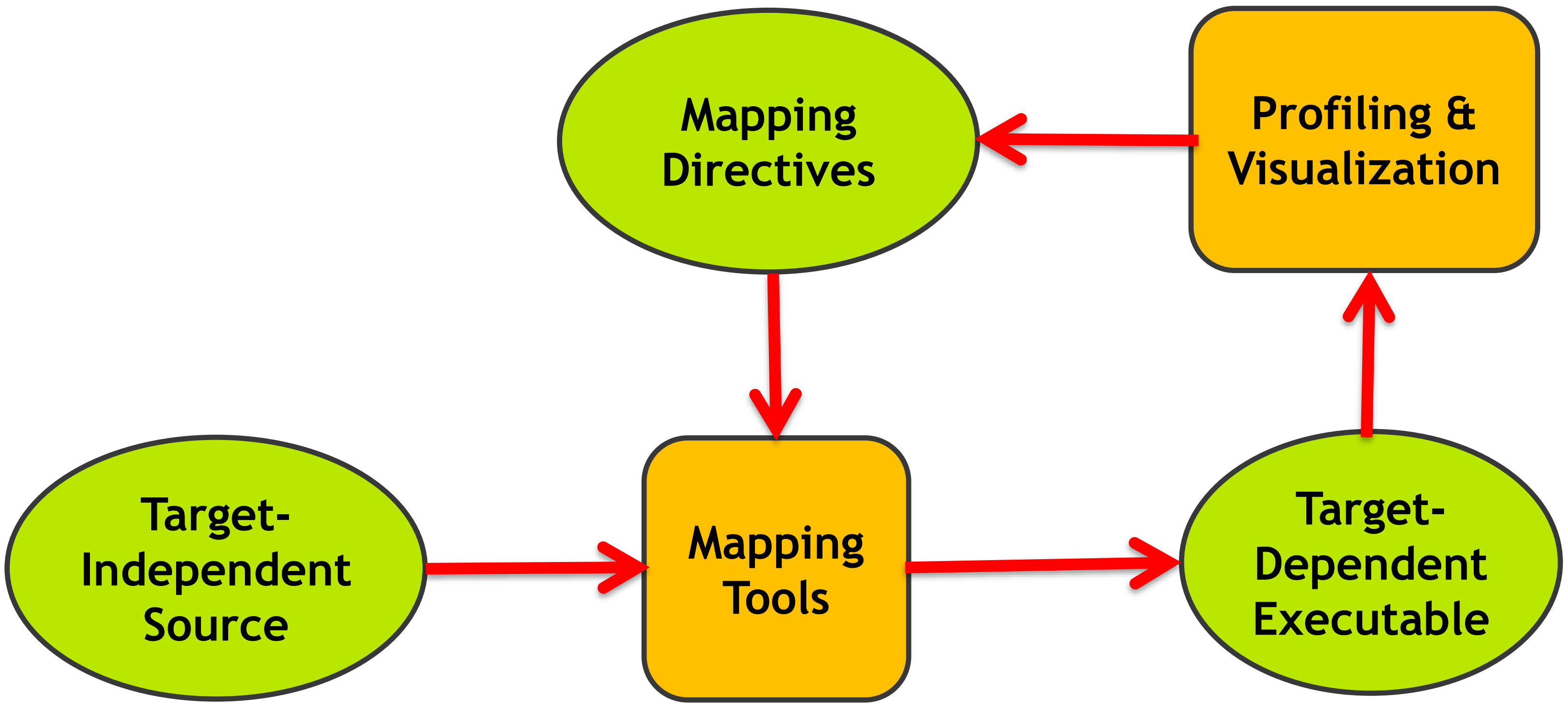
Programmers, tools, and architecture
Need to play their positions




```
forall molecule in set { // launch a thread array
    forall neighbor in molecule.neighbors { //
        forall force in forces { // doubly nested
           molecule.force =
             reduce_sum(force(molecule, neighbor))
        }
    }
}
```

Map foralls in time and space
Map molecules across memories
Stage data up/down hierarchy
Select mechanisms

Exposed storage hierarchy
Fast comm/sync/thread mechanisms


Mapping Directives
Profiling & Visualization
Target-Independent Source
Mapping Tools
Target-Dependent Executable

Legion Programming Model

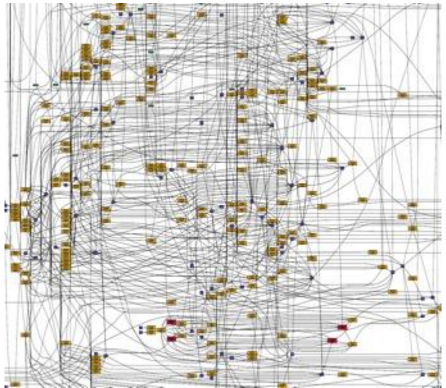
Separating program logic from machine mapping

Legion Program → **Legion Runtime** ← **Legion Mapper**

Target-independent specification

Compute target-specific mapping

Schedule

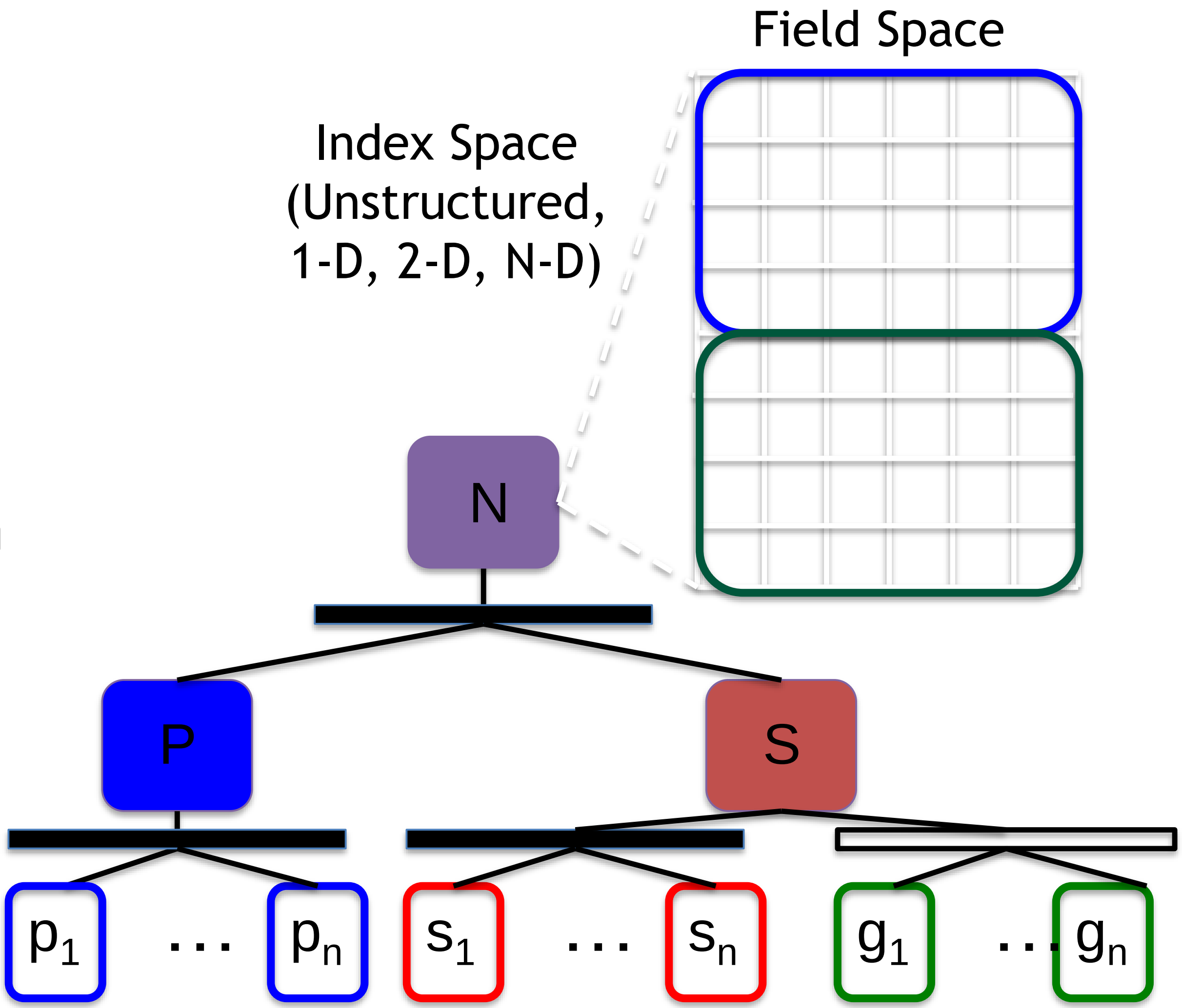


The Legion Data Model: Logical Regions

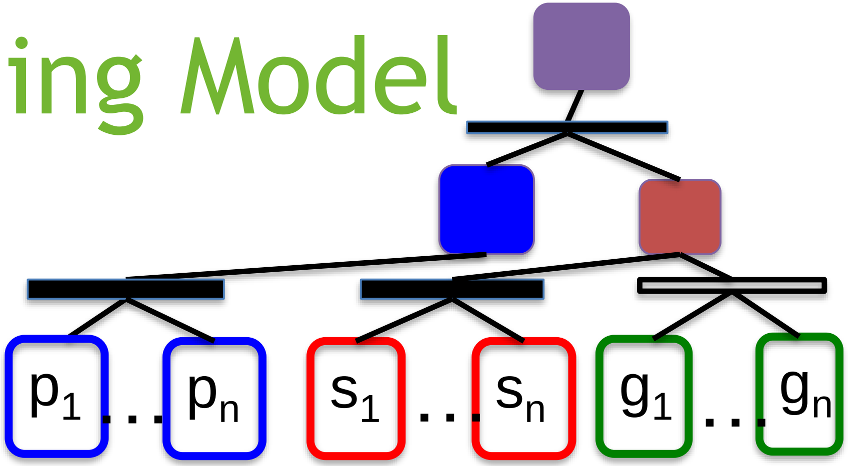
Main idea: logical regions
- Describe data abstractly
- Relational data model
- No implied layout
- No implied placement

Sophisticated partitioning mechanism
- Multiple views onto data

Capture important data properties
- Locality
- Independence/aliasing



The Legion Programming Model



Computations expressed as tasks

- Declare logical region usage
- Declare field usage
- Describe privileges: read-only, read-write, reduce

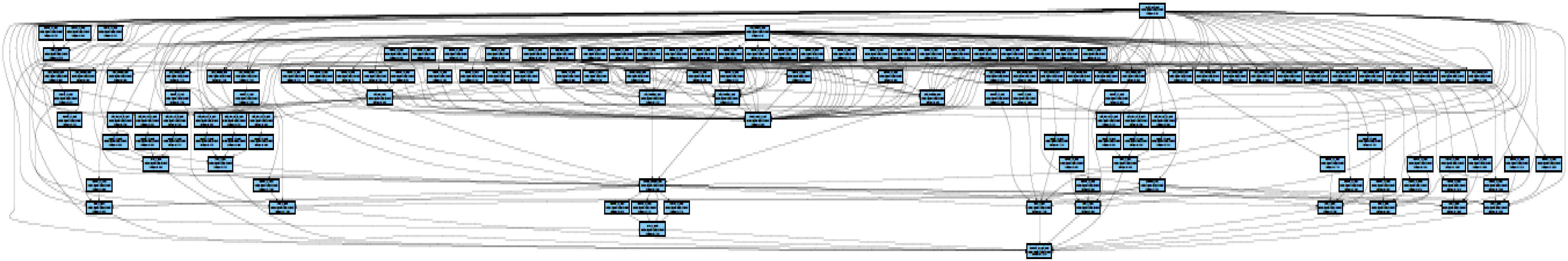
Tasks specified in sequential order

Legion infers implicit parallelism

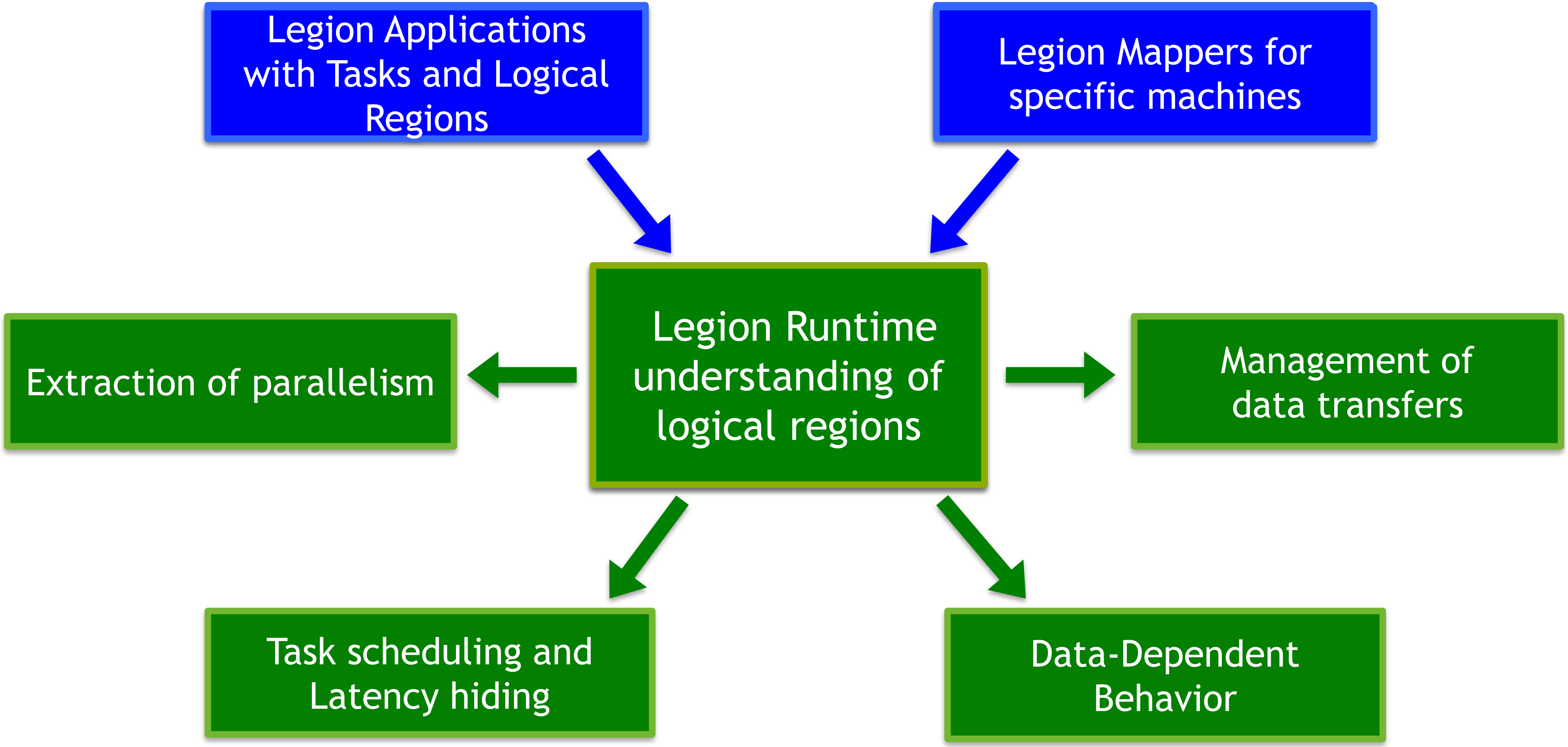
Programs are machine-independent

- Tasks decouple computation
- Logical regions decouple data

calc_currents(piece[0], p_0, s_0, g_0);
calc_currents(piece[1], p_1, s_1, g_1);
distribute_charge(piece[0], p_0, s_0, g_0);
distribute_charge(piece[1], p_1, s_1, g_1);



Legion Runtime System



Evaluation with a Real App: S3D

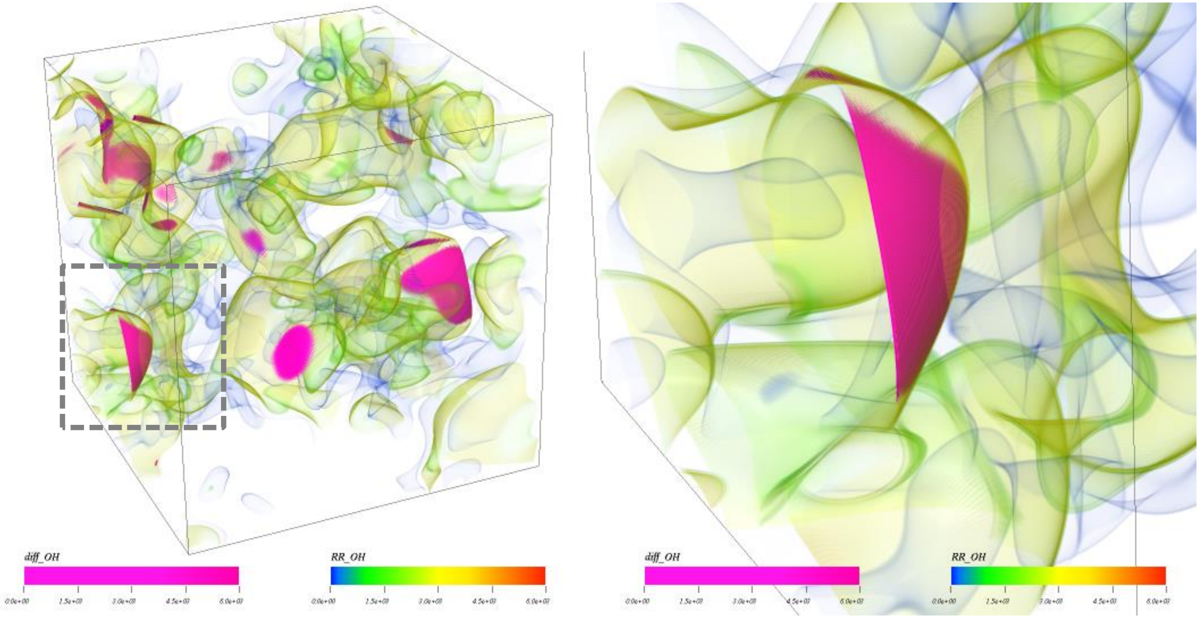
Evaluation with a production-grade combustion simulation

Ported more than 100K lines of MPI Fortran to Legion C++

Legion enabled new chemistry: Primary Reference Fuel (PRF) mechanism

Ran on two of the world's top 10 supercomputers for 1 month

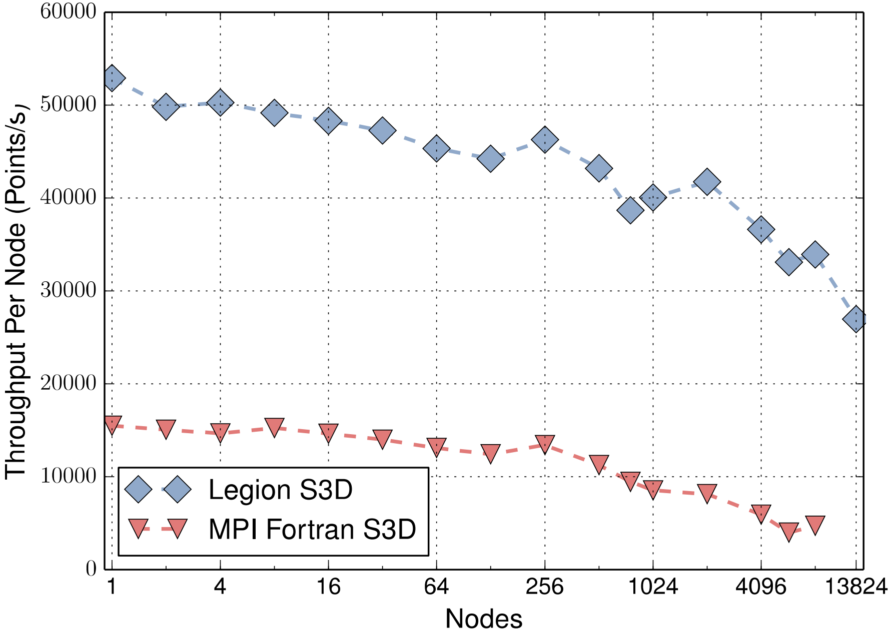
- Titan (#2) and Piz-Daint (#10)



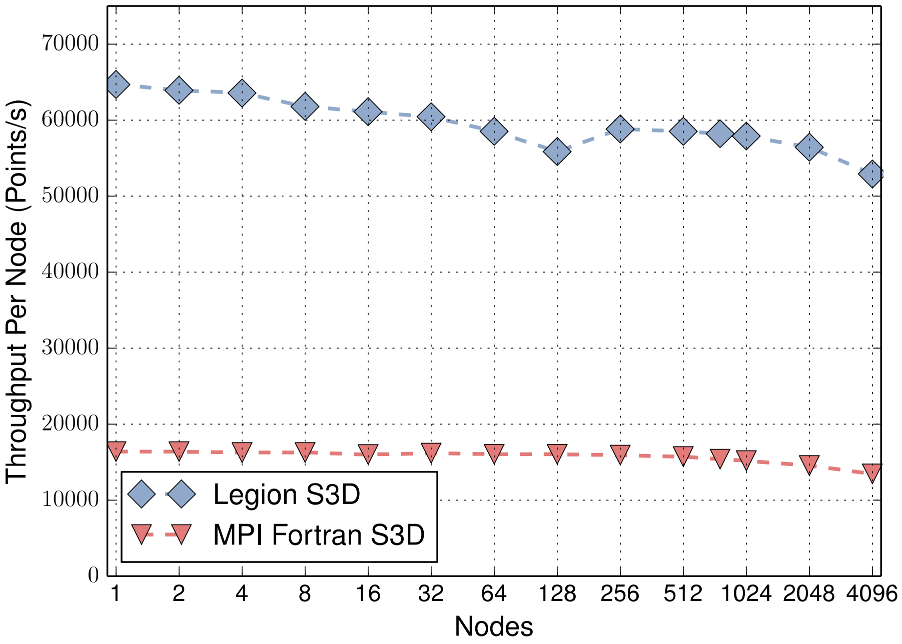
Performance Results: Original S3D

Weak scaling compared to vectorized MPI Fortran version of S3D

Achieved up to 6X speedup



Titan



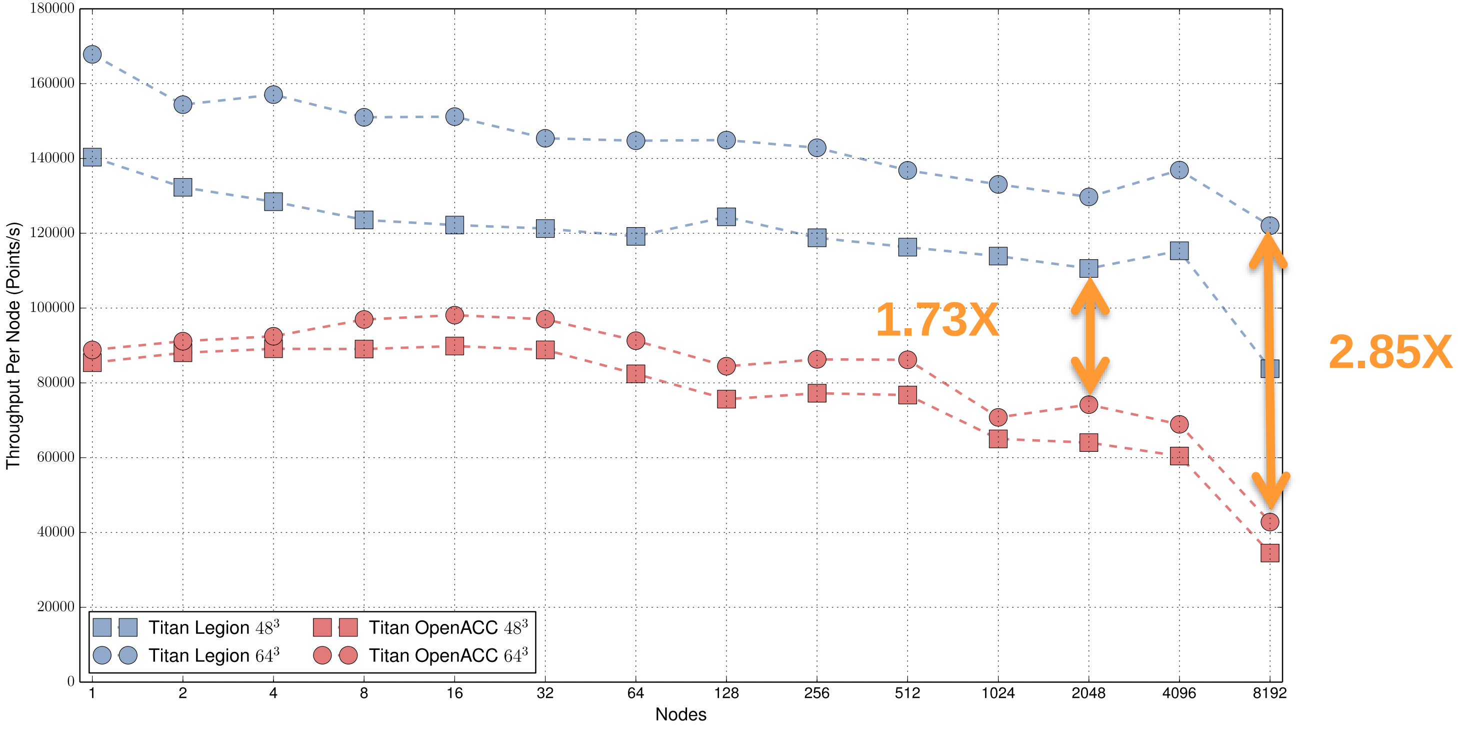
Piz-Daint

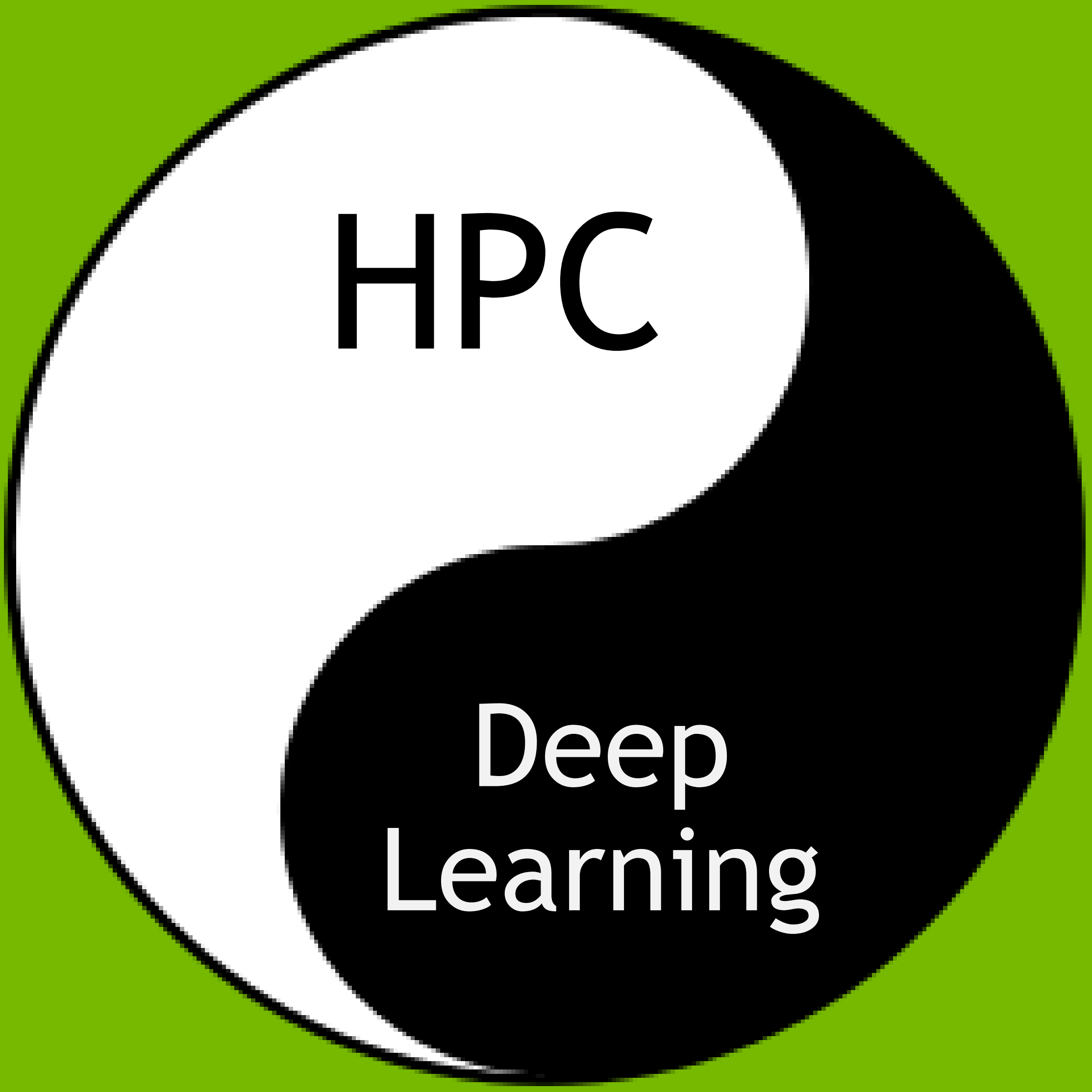
Performance Results: OpenACC S3D

Also compared against experimental MPI+OpenACC version

Achieved 1.73 - 2.85X speedup on Titan

Why? Humans are really bad at scheduling complicated applications




HPC
Deep
Learning

HPC <-> Deep Learning

- HPC has enabled Deep Learning
  - Concepts developed in the 1980s - GPUs provided needed performance
  - Superhuman performance on many tasks – classification, go, …
  - Enabling intelligent devices – including cars
- Deep Learning enables HPC
  - Extracting meaning from data
  - Replacing models with recognition
- HPC and Deep Learning both need more performance – but Moore’s Law is over
  - Reduced overhead
  - Efficient communication
- Resulting machines are parallel with deep memory hierarchies
  - Target-Independent Programming

NVIDIA