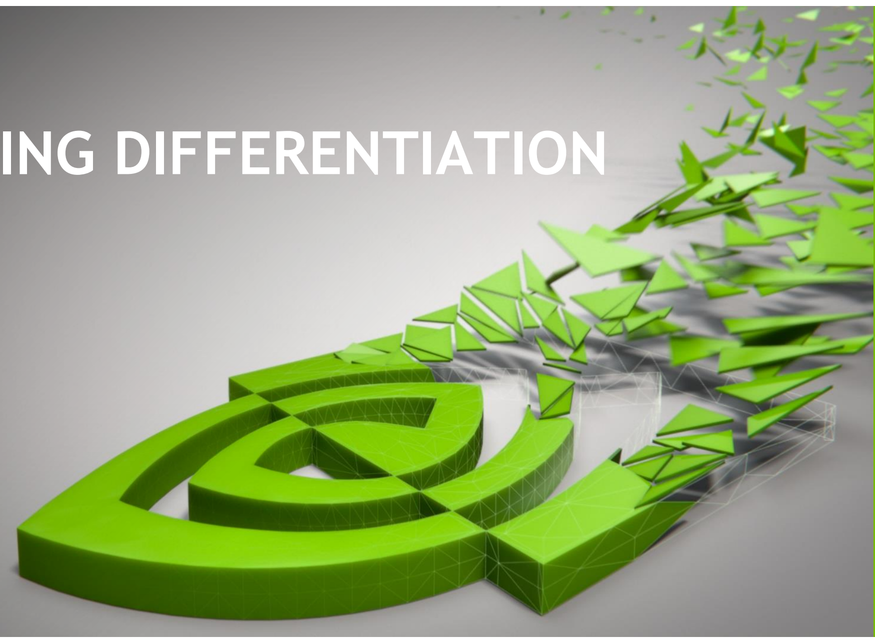
ENDURING DIFFERENTIATION

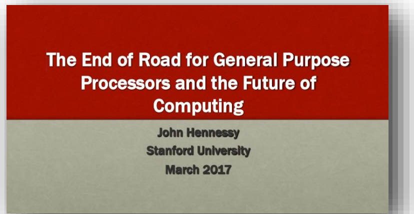
Timothy Lanfear

NVIDIA

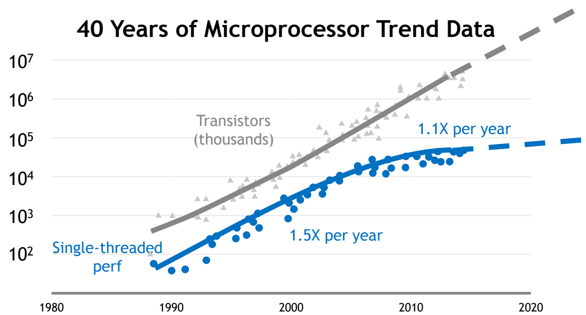


WHERE ARE WE?

LIFE AFTER DENNARD SCALING

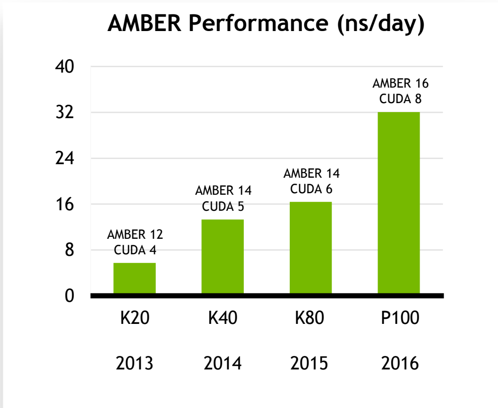

The End of Road for General Purpose Processors and the Future of Computing

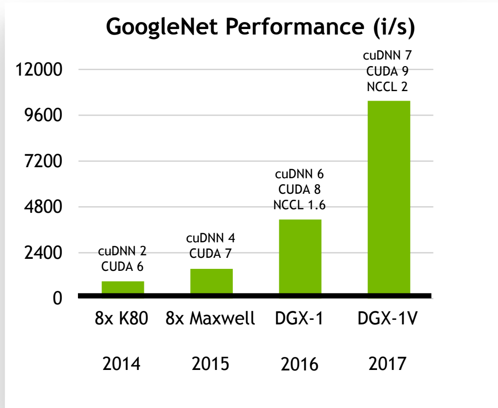
John Hennessy
Stanford University
March 2017



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

GPU-ACCELERATED PERFORMANCE

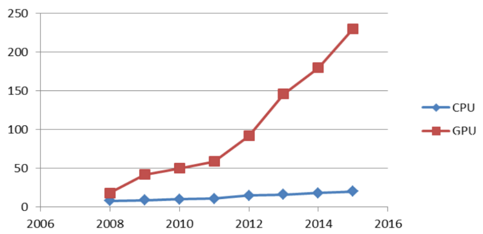




TESLA PLATFORM ADVANTAGE

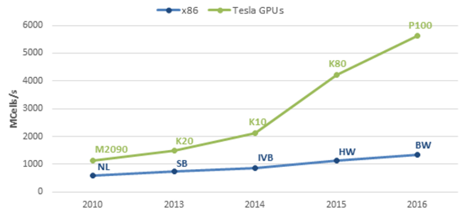
Delivered value grows over time

Life Sciences (AMBER)



10X Perf in 8 Years

Oil & Gas (RTM)



6X Perf in 6 Years

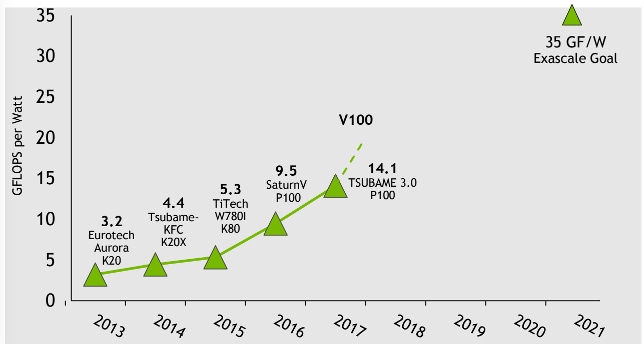
Amber performance: Nano Seconds Per Day delivered on 1xServer with GPUs and CPUS

GPU-ACCELERATED EFFICIENCY

13/13 Greenest Supercomputers
Powered by Tesla P100

TSUBAME 3.0
Kukai
AIST AI Cloud
RAIDEN GPU subsystem
Piz Daint
Wilkes-2
GOSAT-2 (RCF2)
DGX Saturn V
Reedbush-H
JADE
Facebook Cluster
Cedar
DAVIDE

On Track To Meet Exascale Goal



Top GPU Systems in Green500 Lists and NVIDIA Projections for V100

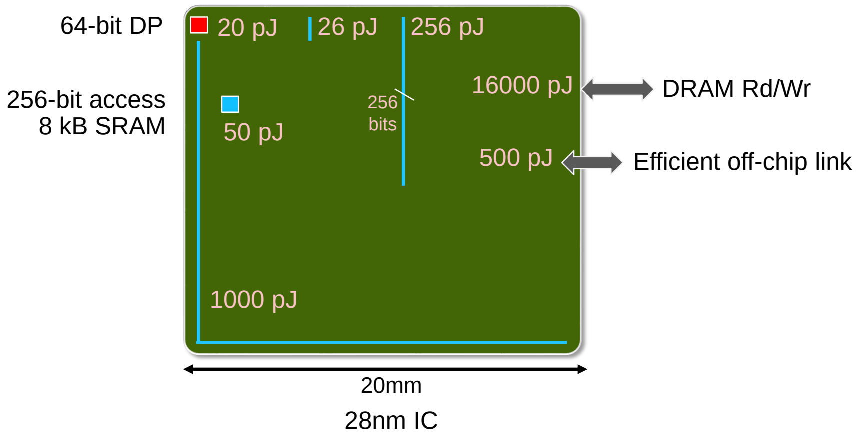
HOW ARE WE DOING THIS?

And, is our differentiation sustainable?

- What are the most important dimensions of our differentiation?
- Why are GPUs so much more efficient than CPUs?
- How can we continue scaling performance/efficiency as Moore's Law fades?
- Why can't competitors replicate GPU efficiency, performance, scaling, etc., with lots of weak CPU cores? (e.g., Intel KNC/KNL/KNM)
- How is optimizing GPUs for AI affecting their suitability for HPC?

ENERGY EFFICIENCY

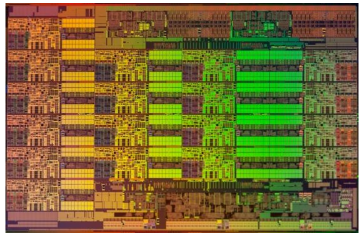
COMPUTATION VERSUS COMMUNICATIONS



CPU
126 pJ/flop (SP)

Optimized for Latency

Deep Cache Hierarchy

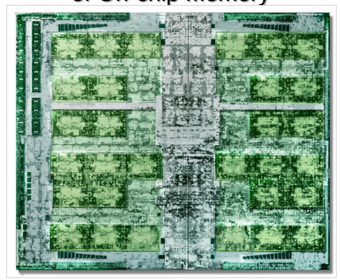


Broadwell E5 v4
14 nm

GPU
28 pJ/flop (SP)

Optimized for Throughput

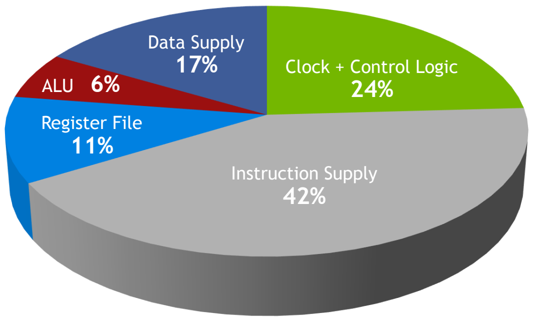
Explicit Management of On-chip Memory



Pascal
16 nm

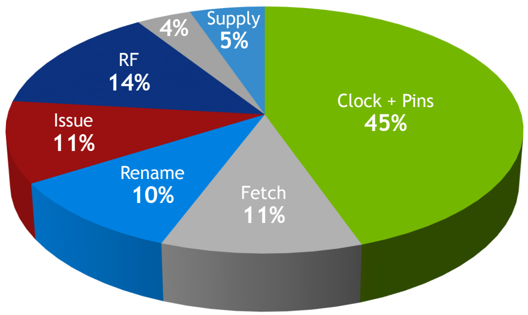
HOW IS POWER SPENT IN A CPU?

In Order, Embedded

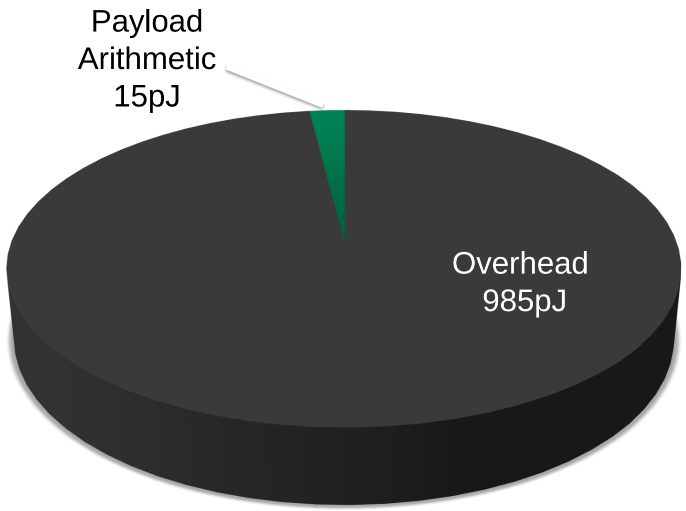


Dally [2008] (Embedded in-order CPU)

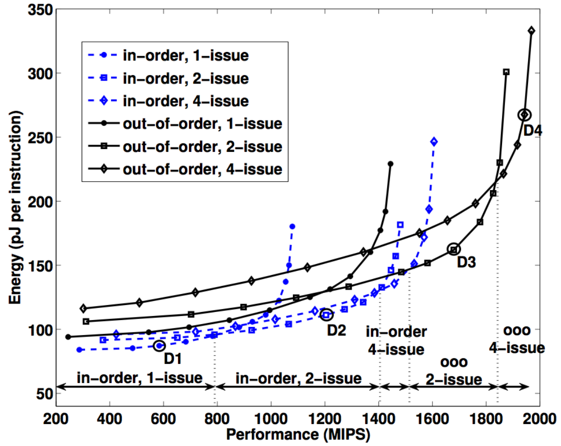
Out of Order, High Performance



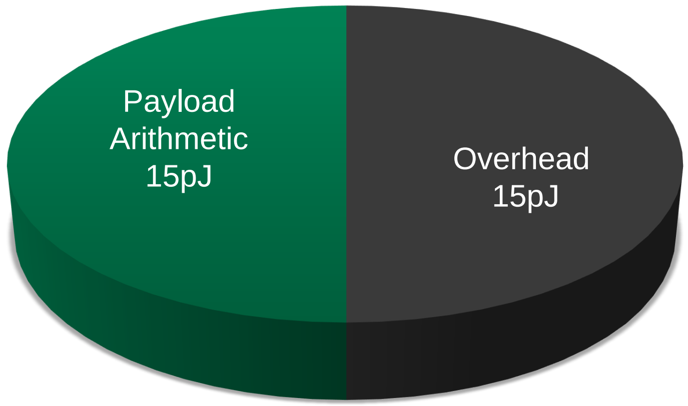
Natarajan [2003] (Alpha 21264)



SIMPLER CORES = ENERGY EFFICIENCY

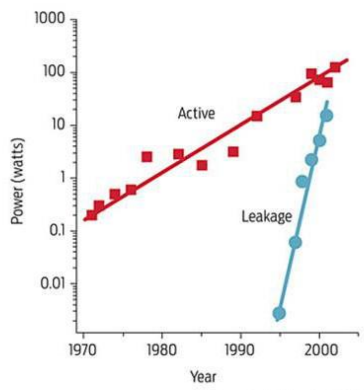


Source: Azizi [PhD 2010]



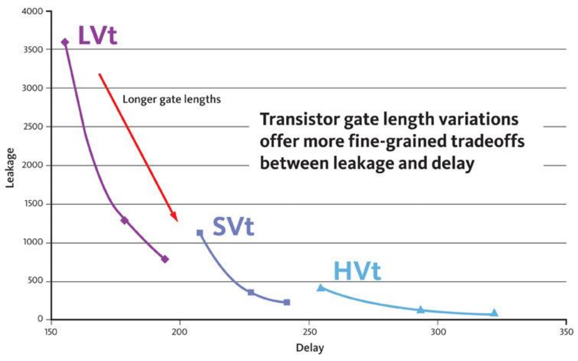
THROUGHPUT PROCESSORS

RISE OF LEAKAGE



Source: Gordon Moore, Intel; IEEE

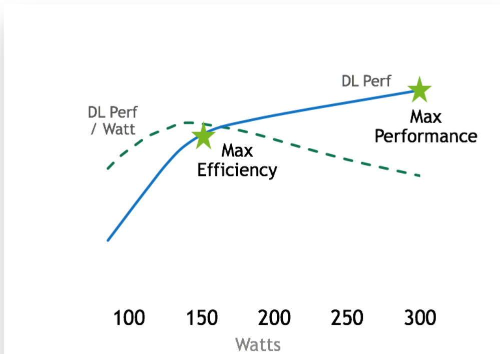
FREQUENCY VS. LEAKAGE



Source: Gordon Moore, Intel; IEEE

OPTIMIZED FOR DATACENTER EFFICIENCY

40% More Performance in a Rack



80% Perf at Half the Power

V100
Max Performance

13 KW Rack
4 Nodes of 8xV100

13
ResNet-50 Networks Trained Per Day

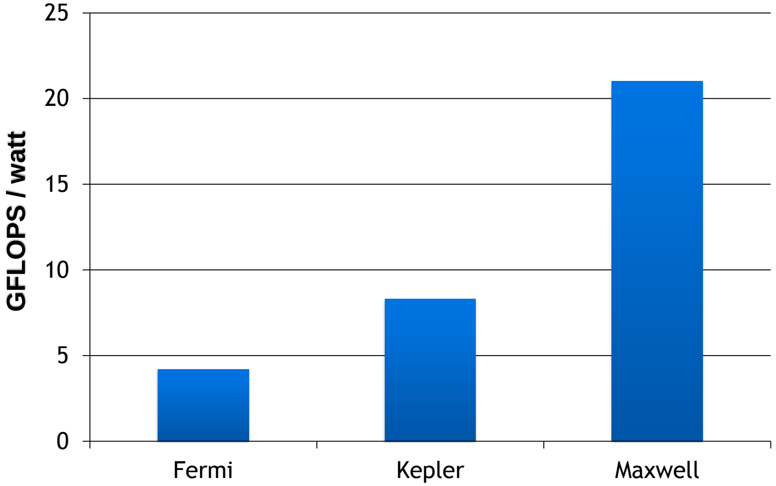
V100
Max Efficiency

13 KW Rack
7 Nodes of 8xV100

18
ResNet-50 Networks Trained Per Day

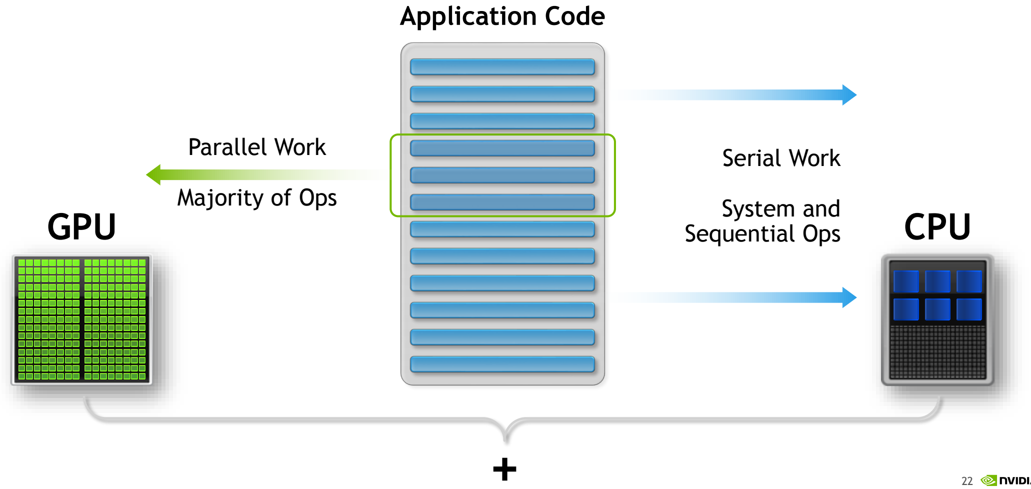
ResNet-50 Training, Max Efficiency run with V100@160W | V100 performance measured on pre-production hardware.

SP ENERGY EFFICIENCY @ 28 NM



HETEROGENEOUS COMPUTING

OPTIMIZING SERIAL/PARALLEL EXECUTION

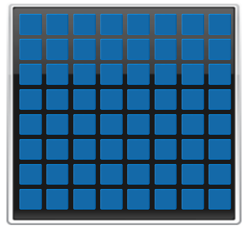


TWO TYPES OF ACCELERATORS

Many-Weak-Cores (MWC) Model

Single CPU Core for Both Serial & Parallel Work

Xeon Phi (*And Others*)
Many Weak Serial Cores

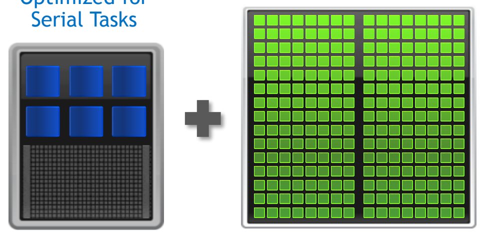


Heterogeneous Computing Model

Complementary Processors Work Together

CPU
Optimized for
Serial Tasks

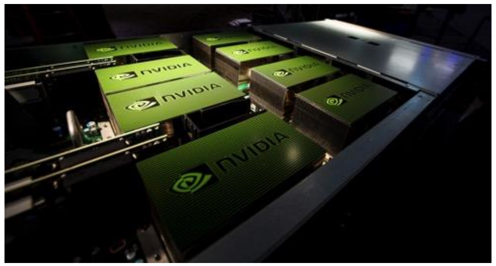
GPU Accelerator
Optimized for
Parallel Tasks

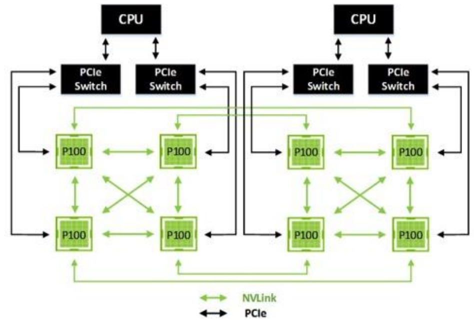


EXTENSIBILITY

NVLINK: A MEMORY FABRIC, NOT A NETWORK

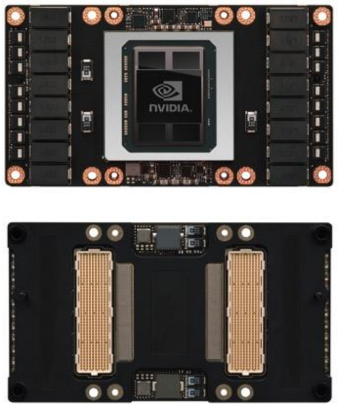
DGX-1: 8 NVLink-Connected GPUs

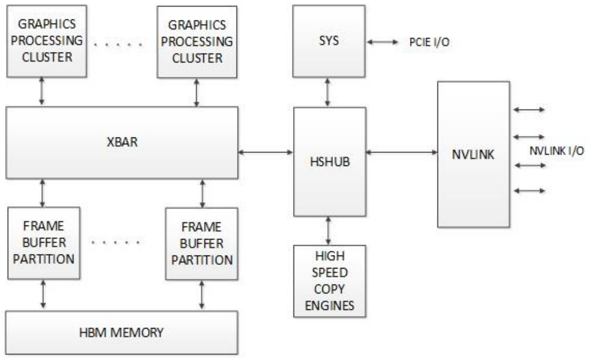




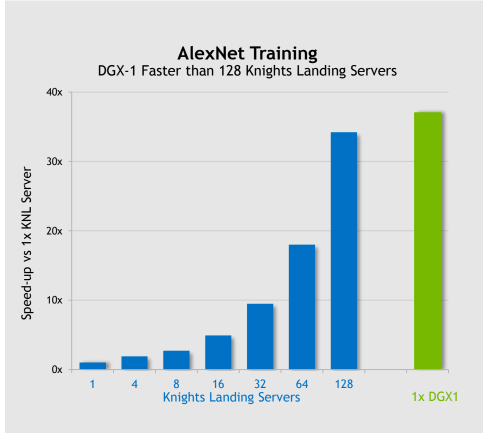
LATENCY HIDING FOR LOAD/STORE/ATOMICS

Where are the NICs? There are no NICs.

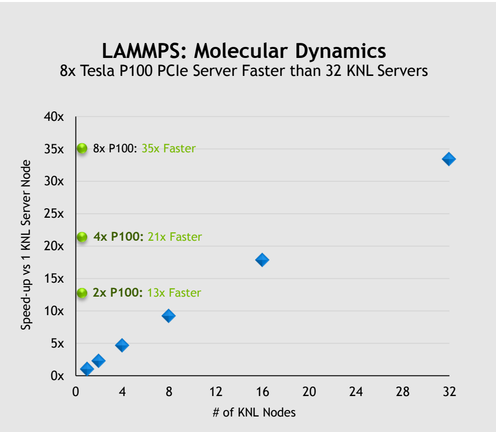


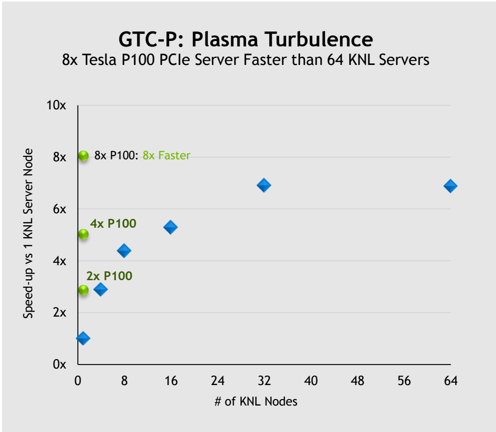


STRONG SCALING

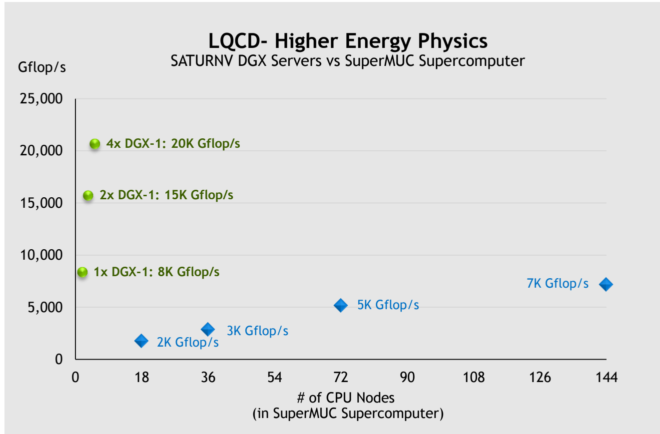


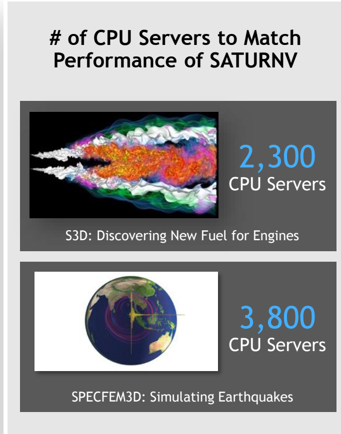
STRONG SCALING





STRONG SCALING





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

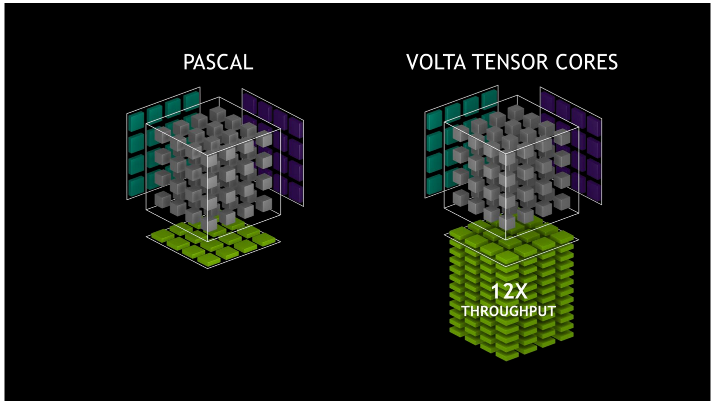
NEW TENSOR CORE

New CUDA TensorOp instructions and data formats

4×4 matrix processing array

$D_{FP32} = A_{FP16} \times B_{FP16} + C_{FP32}$

Optimized for deep learning

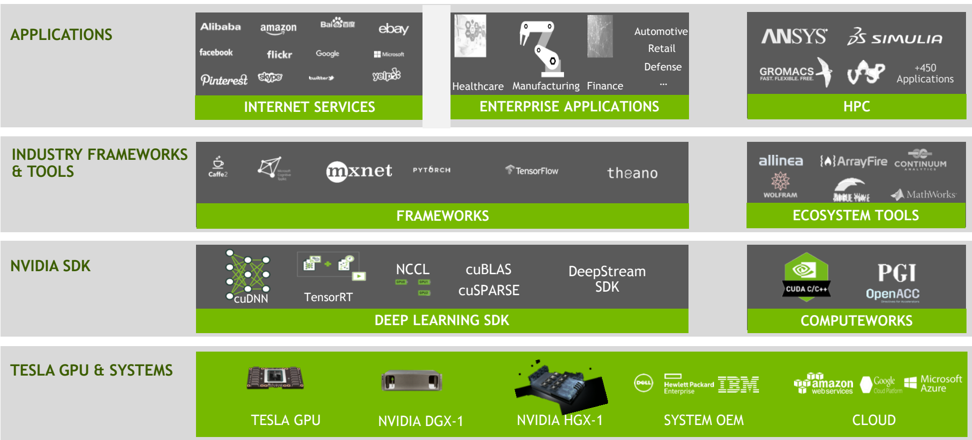


Activation Inputs   Weights Inputs   Output Results

TESLA PLATFORM

TESLA IS A PLATFORM

World's Leading Data Center Platform for Accelerating HPC and AI



MULTIPLE GROWTH MARKETS

CONCLUSION

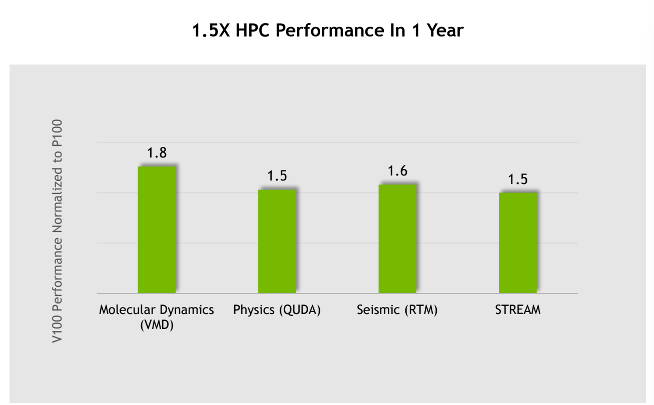
PASCAL TO VOLTA

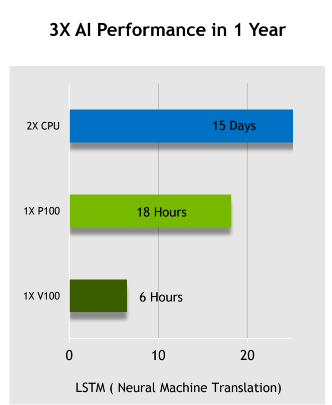
Architecture with Technology

- Area: ~600 mm² → ~800 mm² (~33% more area)
- Process: ~ small Pascal → Volta improvement (a few percent)
- Clocks: similar dynamic range, power limited
- Memory BW (sustained): 50% improvement
- Communications (NVLink): 160 GB/s → 300 GB/s (almost double!)
- AI (Tensor Cores): ~20 TFLOPS → 120 TFLOPS (~6x!)

REVOLUTIONARY PERFORMANCE FOR HPC AND AI

Single Platform For Data Science and Computation Science



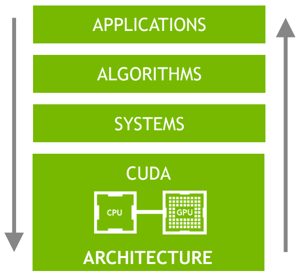


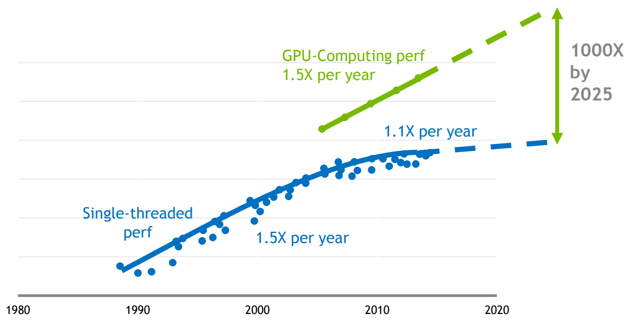
NMT Training for 13 Epochs |German ->English, WMT15 subset |System Config Info: CPU = 2x Xeon E5 2699 V4 w/ P100s or V100s |
QUDA, RTM, STREAM | System Config Info: 2X Xeon E5-2690 v4, 2.6GHz, w/ 1X Tesla P100 or V100.
VMD | System Config Info: Xeon E5-2698v3 w/ 1x Tesla P100 and E5-2697Av4 w/ 1x Tesla V100 | V100 measured on pre-production hardware.

GPU PERFORMANCE COMPARISON

	P100	V100	Ratio
Training acceleration	10 TOPS	120 TOPS	12x
Inference acceleration	21 TFLOPS	120 TOPS	6x
FP64/FP32	5/10 TFLOPS	7.5/15 TFLOPS	1.5x
HBM2 Bandwidth	720 GB/s	900 GB/s	1.2x
NVLink Bandwidth	160 GB/s	300 GB/s	1.9x
L2 Cache	4 MB	6 MB	1.5x
L1 Caches	1.3 MB	10 MB	7.7x

GPU TRAJECTORY





Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp