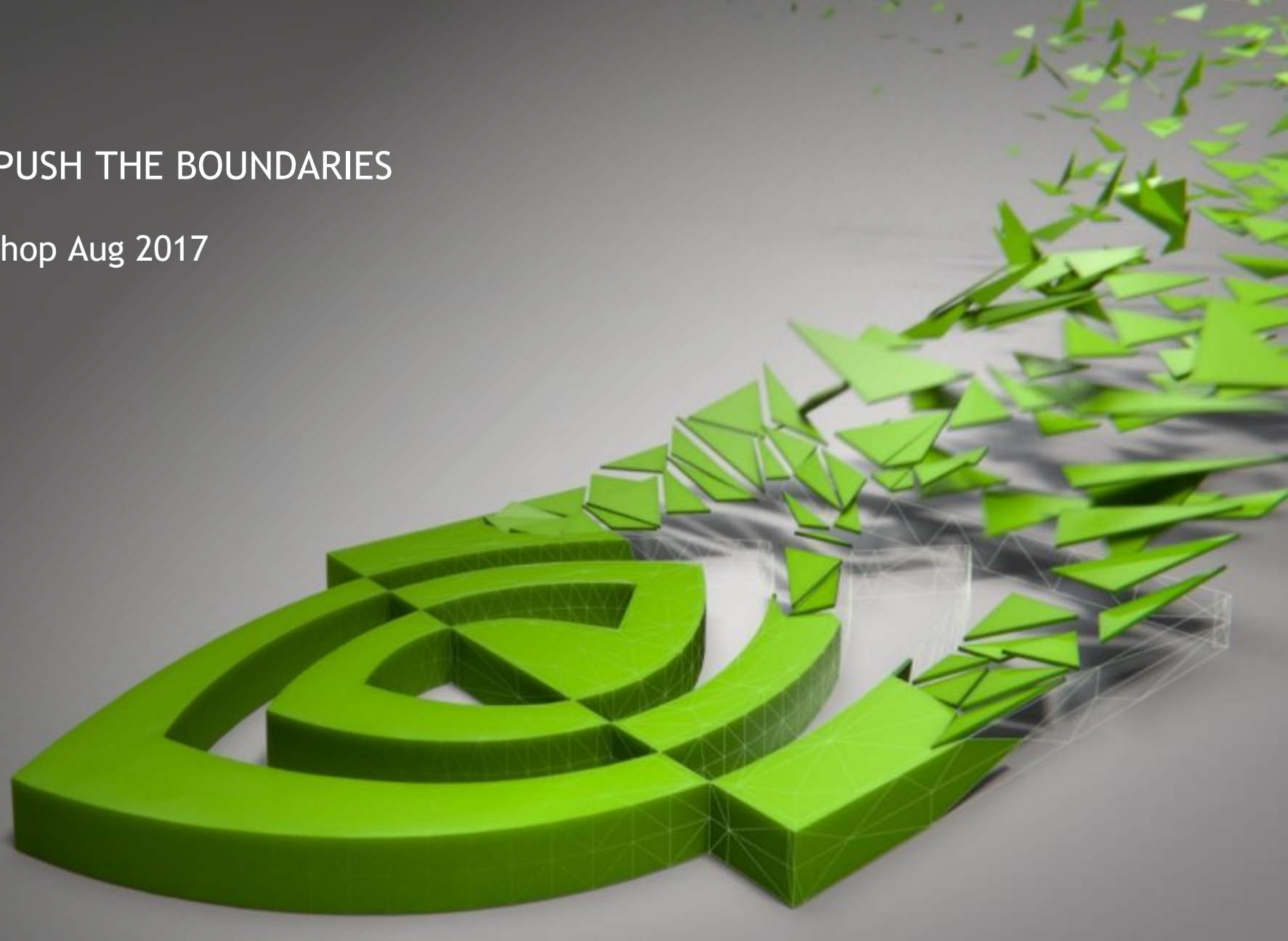
VOLTA

CONTINUING TO PUSH THE BOUNDARIES

OpenSHMEM Workshop Aug 2017

Brad Rees Ph.D.

NVIDIA



FOREWORD

This talk is about NVIDIA's latest GPU and not OpenSHMEM

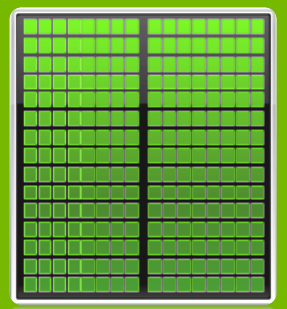
Sreeram Potluri will be presenting on NVIDIA's NVSHMEM work Tuesday at 2pm

Efficient Breadth First Search on Multi-GPU Systems using GPU-centric OpenSHMEM

"It's time to start planning for the end of Moore's Law, and it's worth pondering how it will end, not just when."

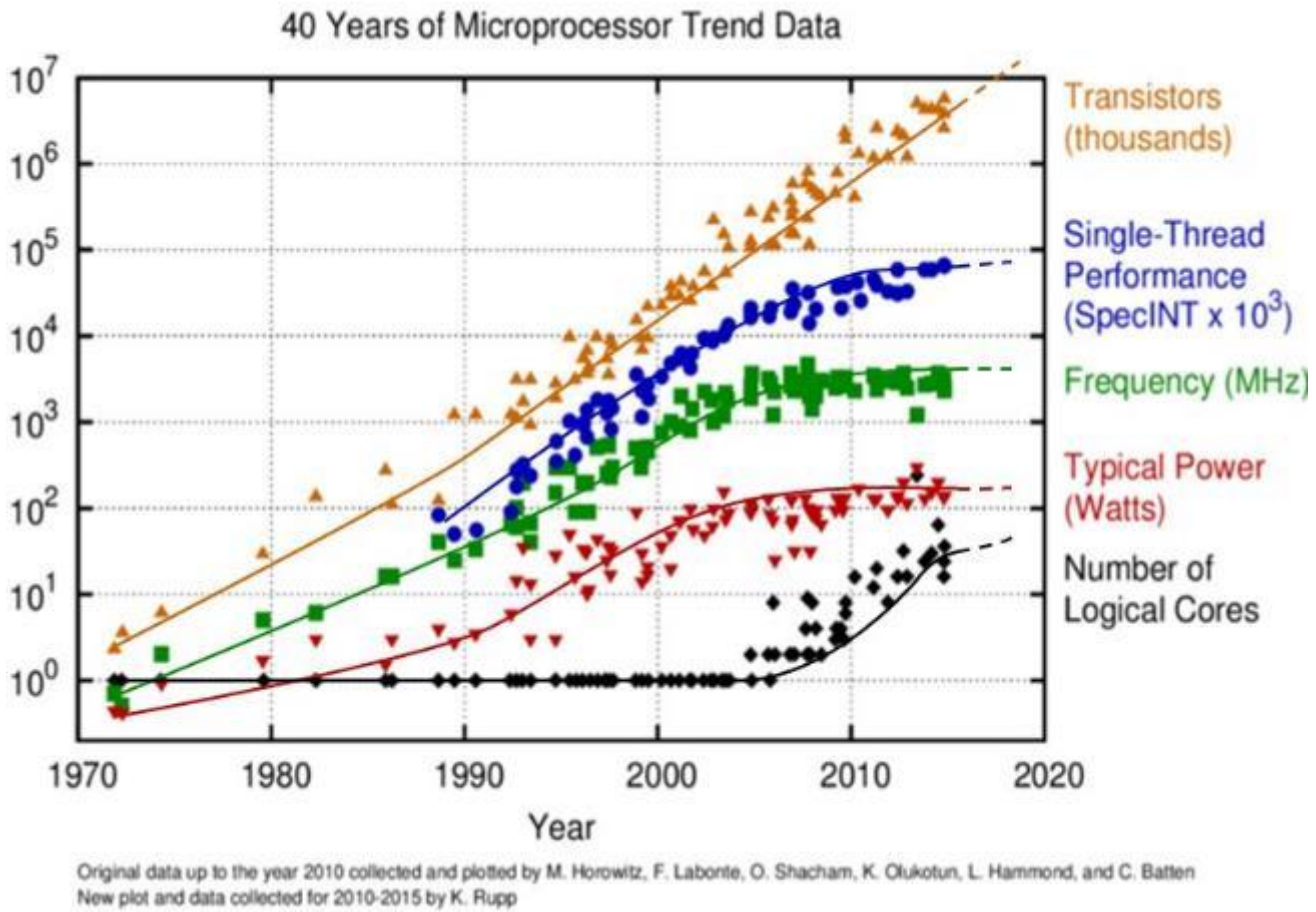
Robert Colwell
Director, Microsystems Technology Office, DARPA

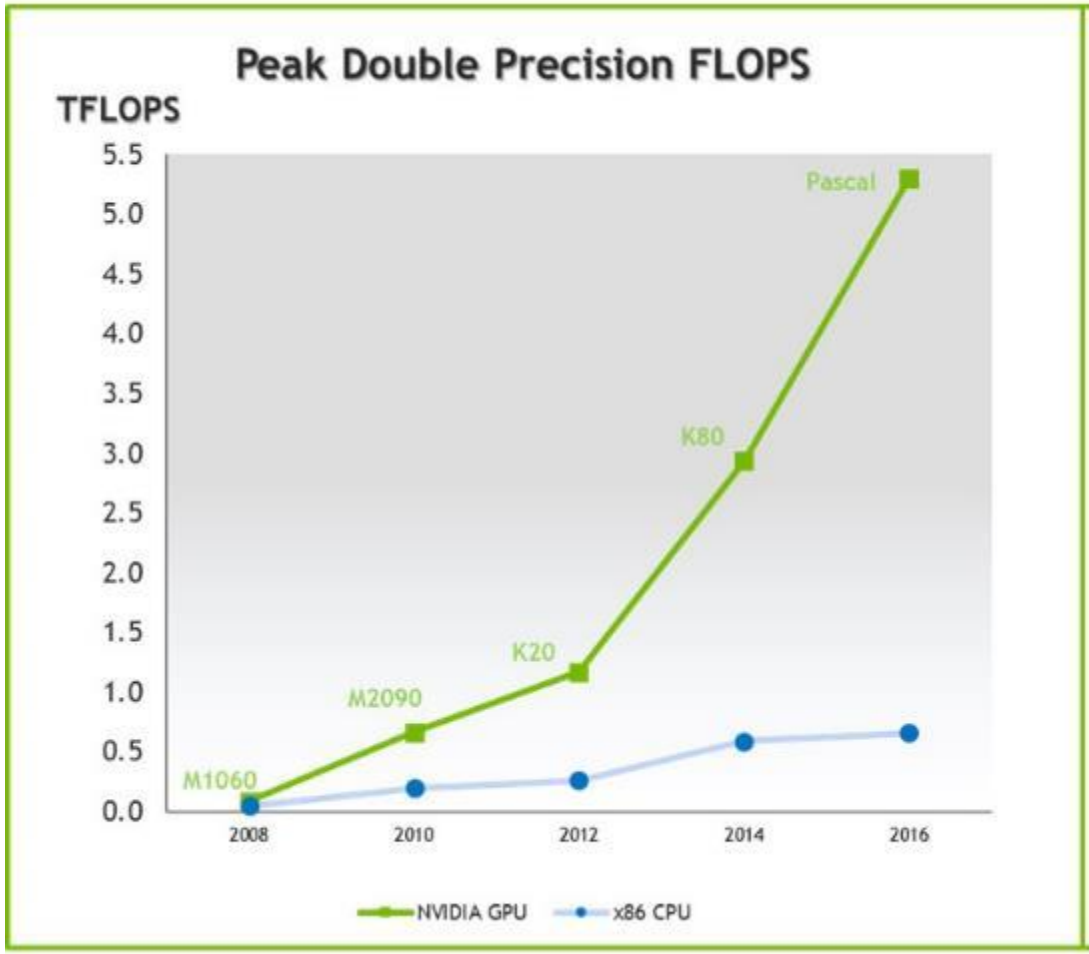
GPU Accelerator

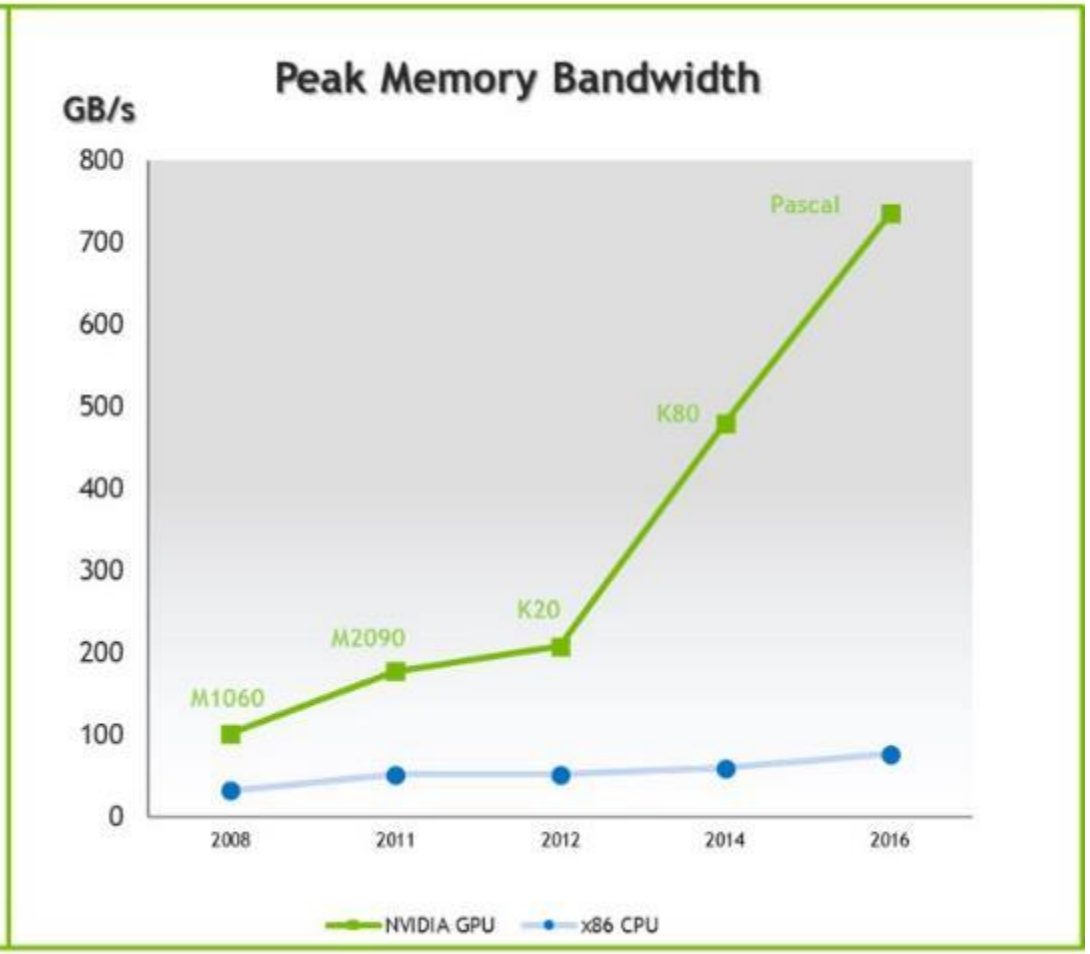


+

CPU





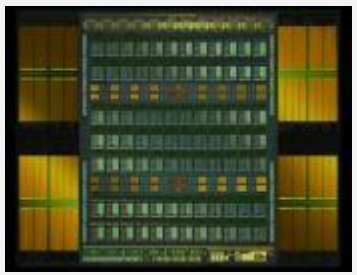


PERFORMANCE GAP
(PRE-VOLTA)

INTRODUCING VOLTA

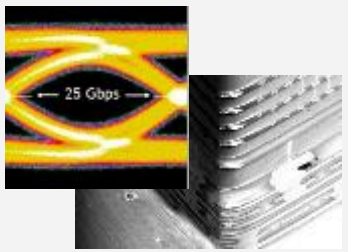
5,120 CUDA cores | **640 NEW** Tensor cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS | **120 Tensor TFLOPS**
16GB HBM2 @ 900 GB/s | 300 GB/s NVLink

Volta Architecture



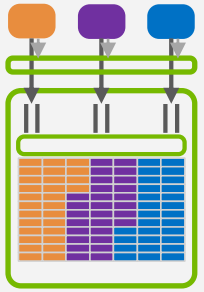
Most Productive GPU

Improved NVLink & HBM2



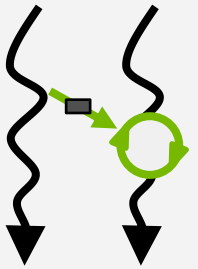
Efficient Bandwidth

Volta MPS



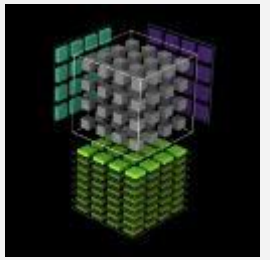
Inference Utilization

Improved SIMT Model



New Algorithms

Tensor Core

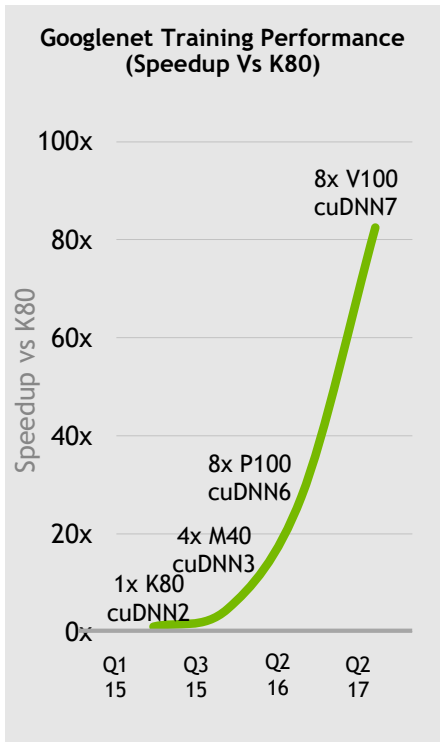


120 Programmable TFLOPS Deep Learning

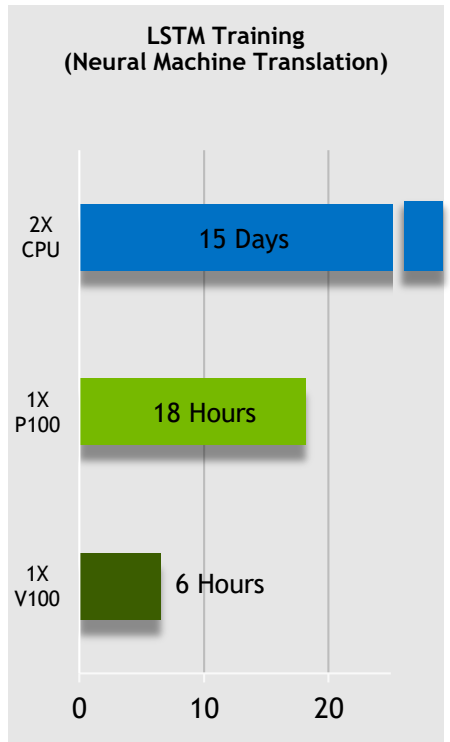
The Fastest and Most Productive GPU for Deep Learning and HPC

REVOLUTIONARY AI PERFORMANCE

Faster DL Training Performance

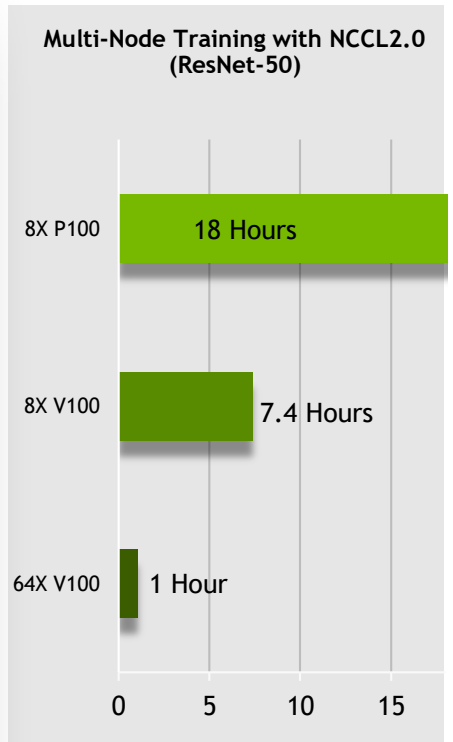


Over 80x DL Training Performance in 3 Years



3X Reduction in Time to Train Over P100

Neural Machine Translation Training for 13 Epochs |German ->English, WMT15 subset | CPU = 2x Xeon E5 2699 V4 | V100 performance measured on pre-production hardware.

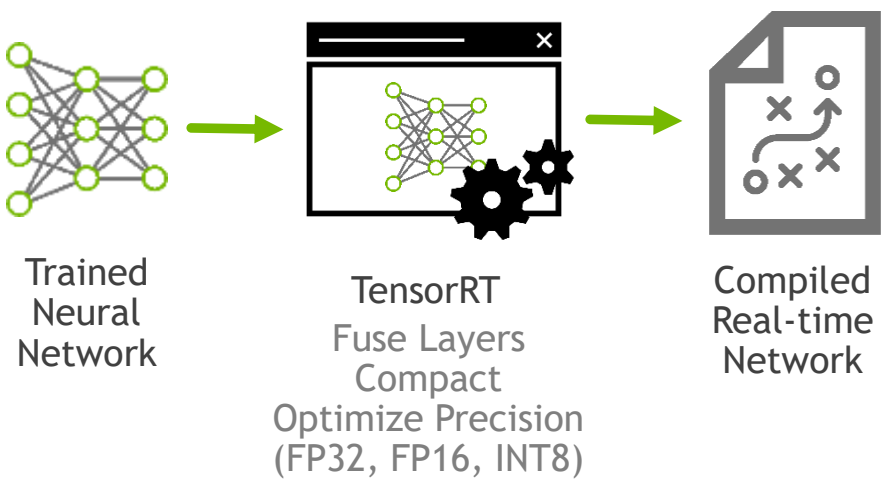


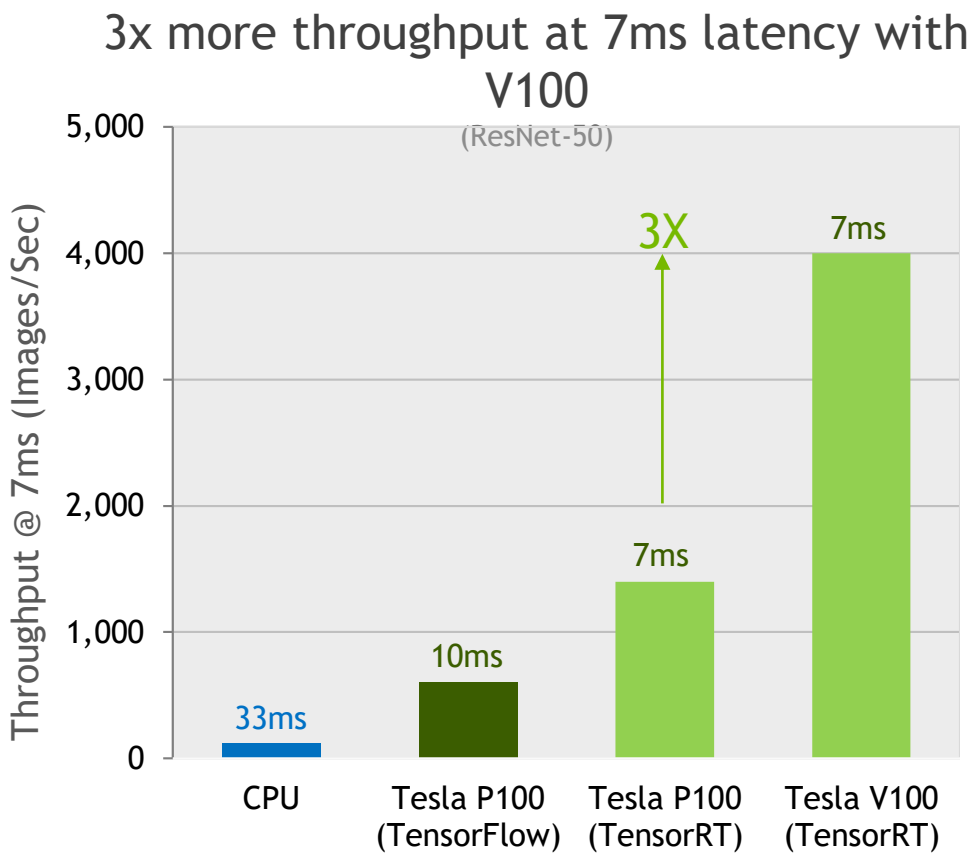
85% Scale-Out Efficiency Scales to 64 GPUs
Microsoft Cognitive Toolkit

ResNet50 Training for 90 Epochs with 1.28M images dataset | Cognitive Toolkit with NCCL 2.0 | V100 performance measured on pre-production hardware.

VOLTA DELIVERS 3X MORE INFERENCE THROUGHPUT

Low Latency performance with V100 and TensorRT

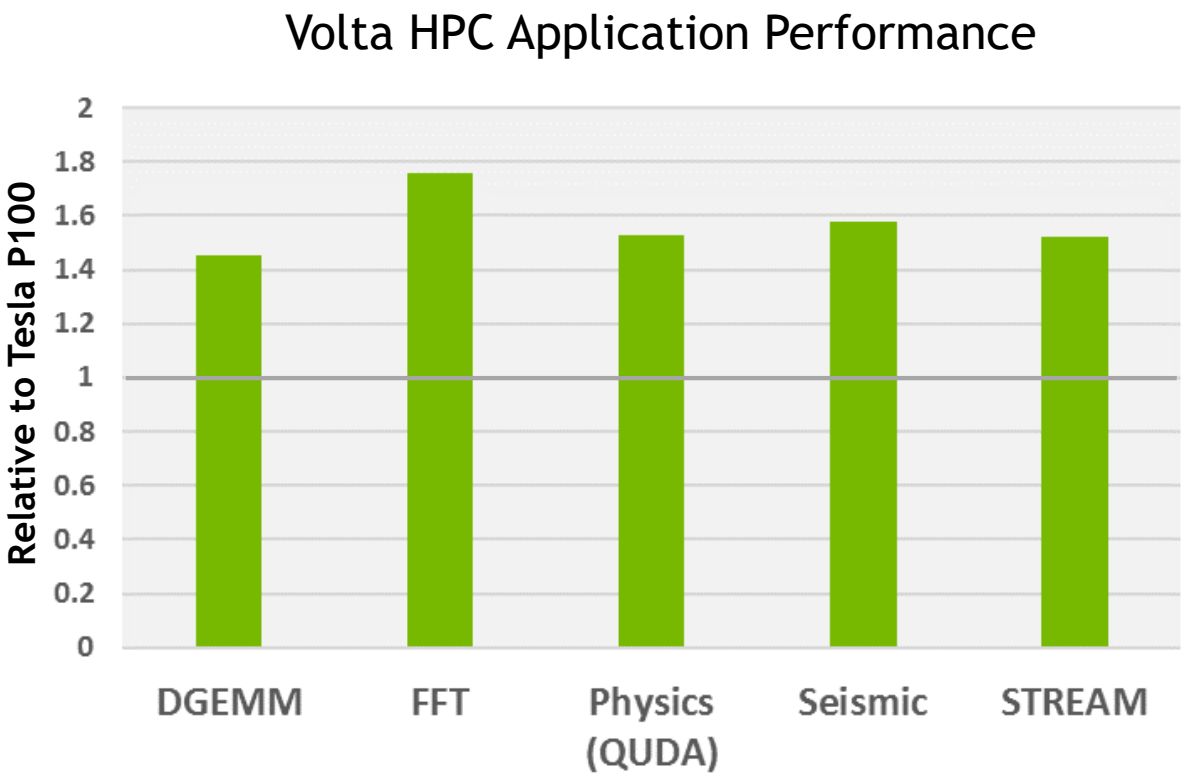




CPU Server: 2X Xeon E5-2660 V4; GPU: w/P100, w/V100 (@150W) | V100 performance measured on pre-production hardware.

ROAD TO EXASCALE

Volta to Fuel Most Powerful US Supercomputers

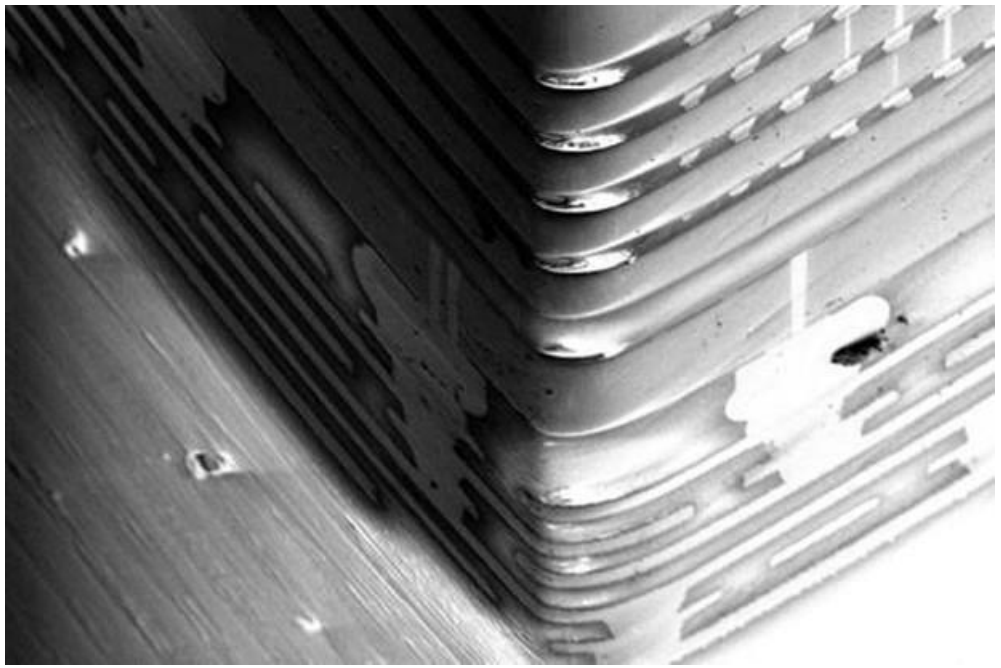


System Config Info: 2X Xeon E5-2690 v4, 2.6GHz, w/ 1X Tesla P100 or V100. V100 measured on pre-production hardware.

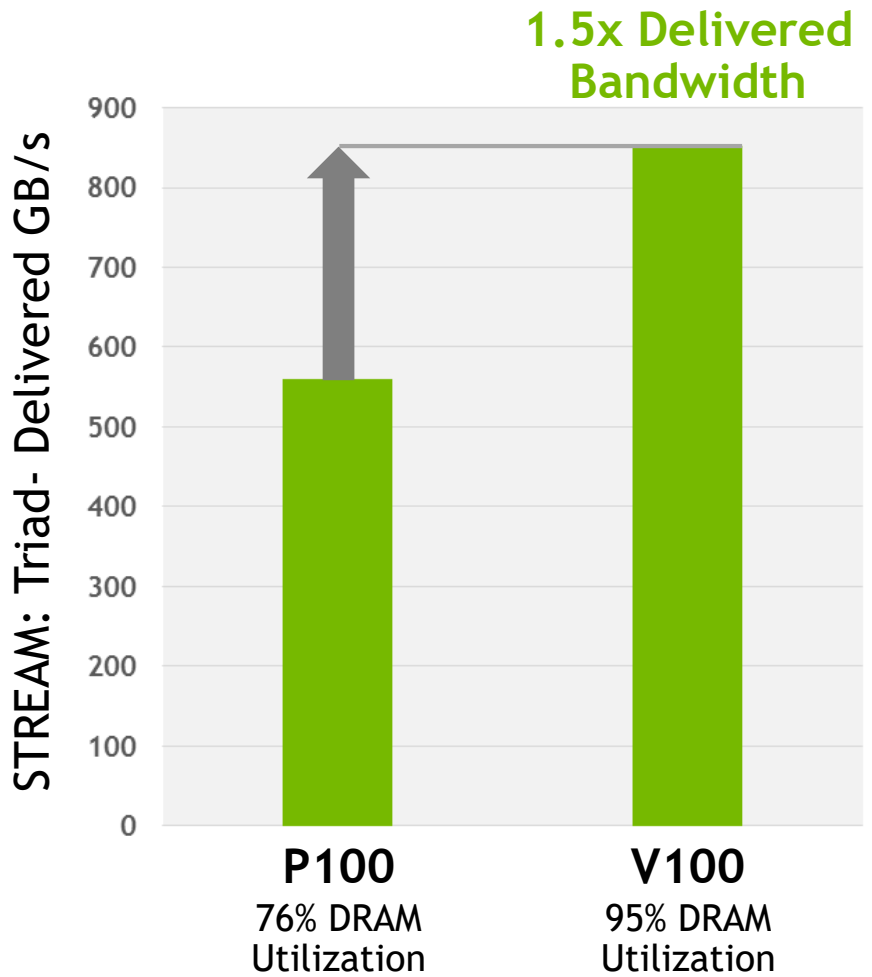
GPU PERFORMANCE COMPARISON

	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 peak bandwidth	720 GB/s	900 GB/s	1.2x
NVLink peak 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

NEW HBM2 MEMORY ARCHITECTURE



HBM2 stack



V100 measured on pre-production hardware.

STATE OF UNIFIED MEMORY

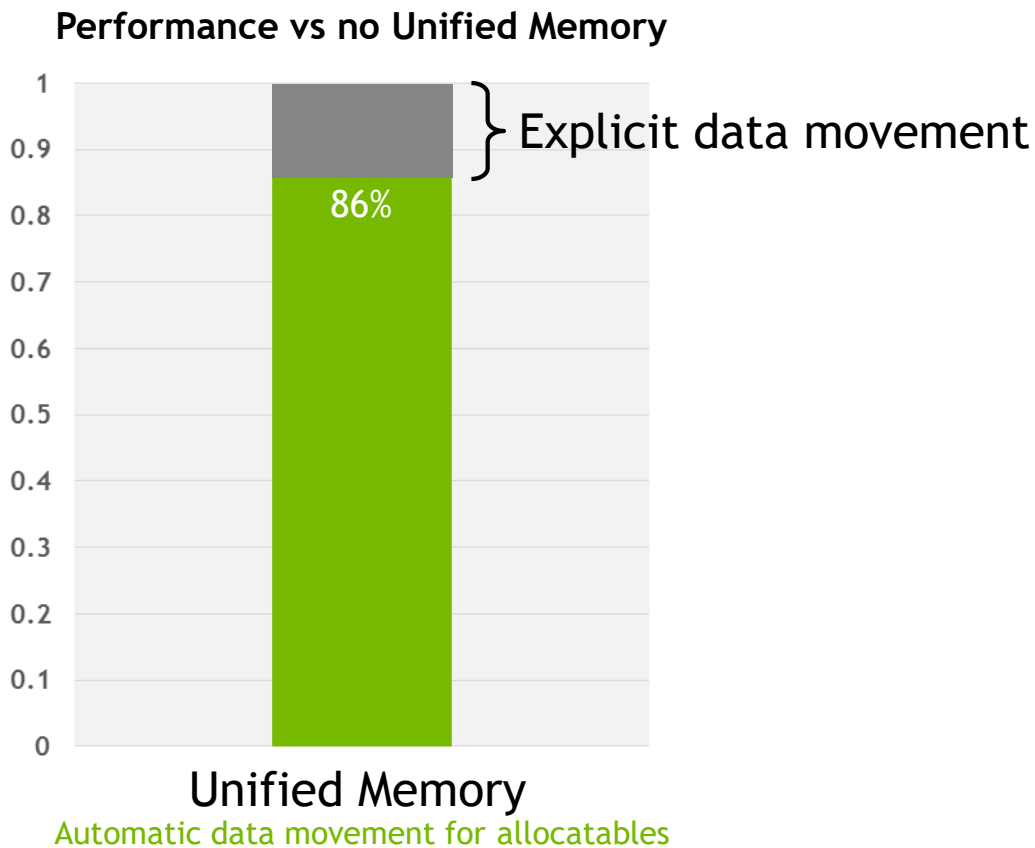
High performance, low effort

PGI OpenACC on Pascal P100

Geometric mean across all 15 SPEC ACCEL™ benchmarks

UM Perf vs. Explicit movement:

86% PCI-E, 91% NVLink



PGI 17.1 Compilers OpenACC SPEC ACCEL™ 1.1 performance measured March, 2017. SPEC® and the benchmark name SPEC ACCEL™ are registered trademarks of the Standard Performance Evaluation Corporation.

UNIFIED MEMORY BENEFITS

Deep Copy – Array of Strings example (NxN)

```
char **data;
data = (char**)malloc(N*sizeof(char*));
for (int i = 0; i < N; i++)
        data[i] = (char*)malloc(N);

char **d_data;
char **h_data = (char**)malloc(N*sizeof(char*));
for (int i = 0; i < N; i++) {
        cudaMalloc(&h_data2[i], N);
        cudaMemcpy(h_data2[i], h_data[i], N, ...);
}
cudaMalloc(&d_data, N*sizeof(char*));
cudaMemcpy(d_data, h_data2, N*sizeof(char*), ...);

gpu_func<<<...>>>(data, N);
```

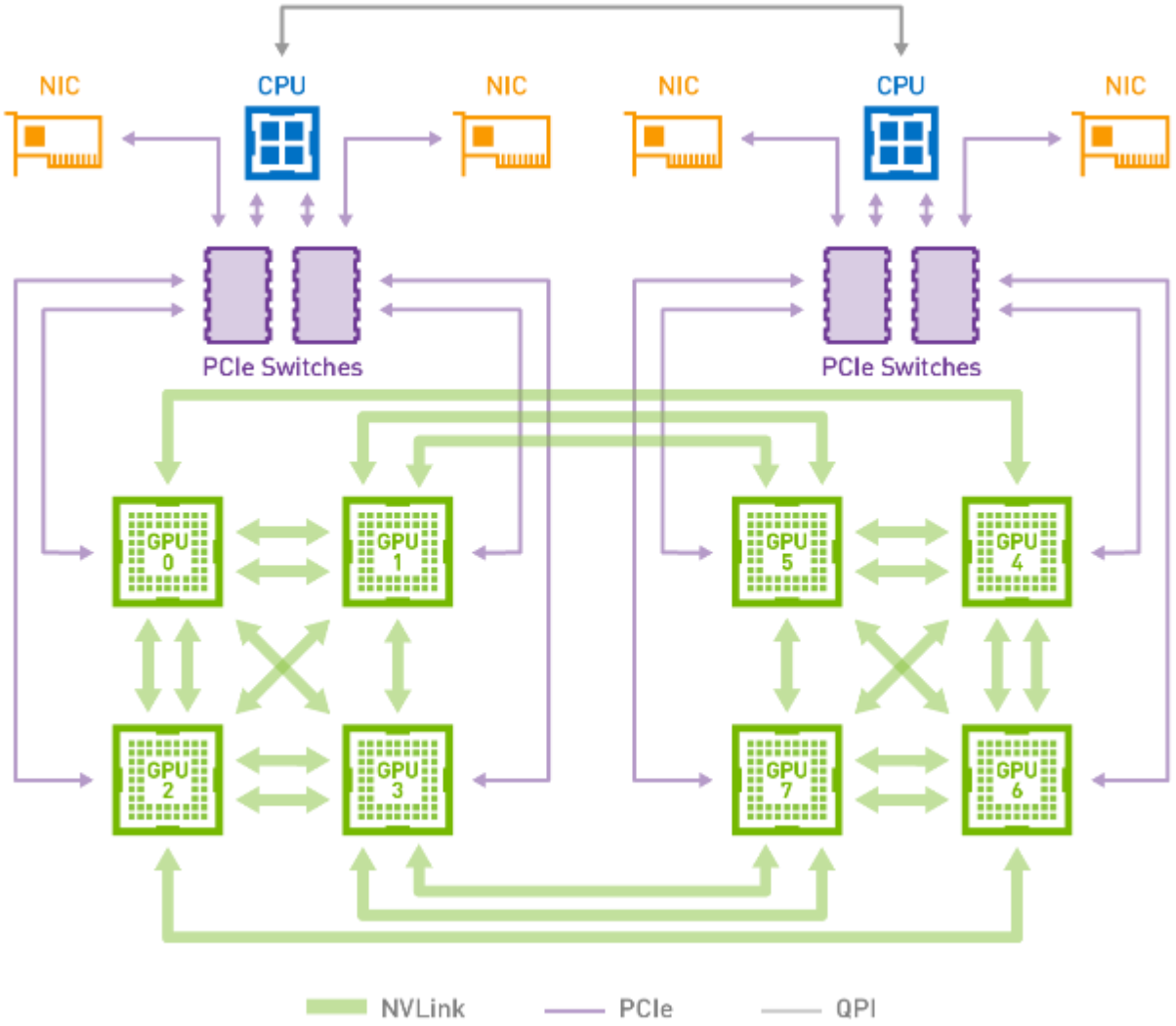
```
char **data;
data = (char**) cudaMallocManaged(N*sizeof(char*));
for (int i = 0; i < N; i++)
        data[i] = (char*) cudaMallocManaged(N);

gpu_func<<<...>>>(data, N);
```

VOLTA NVLINK

300GB/sec

50% more links

25% faster signaling



NEW TENSOR CORE BUILT FOR AI

120 Tensor TFLOPS of DL Performance

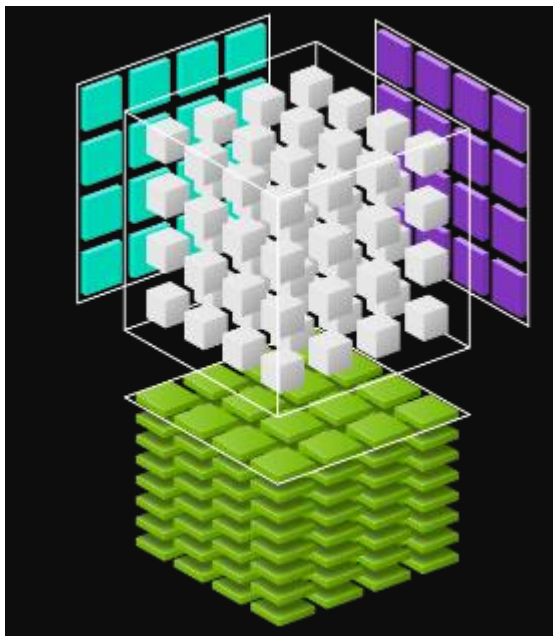
MATRIX DATA OPTIMIZATION:

Dense Matrix of Tensor Compute

TENSOR-OP CONVERSION:

FP32 to Tensor Op Data for Frameworks

VOLTA-OPTIMIZED cuDNN



VOLTA TENSOR CORE

4x4 matrix processing array
D[FP32] = A[FP16] + B[FP16] + C[FP32]
Optimized For Deep Learning

Caffe2

Microsoft Cognitive Toolkit

mxnet

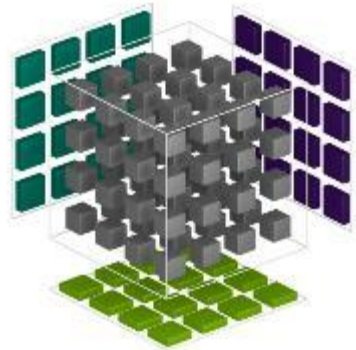
PYTORCH

TensorFlow

ALL MAJOR FRAMEWORKS

TENSOR CORE

4x4x4 Warp Matrix Multiply and Accumulate (WMMA)



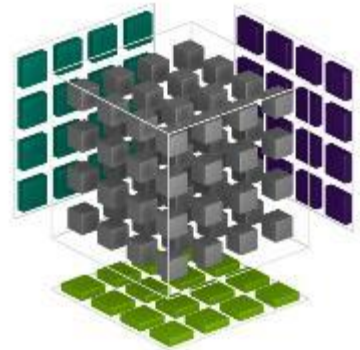
$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

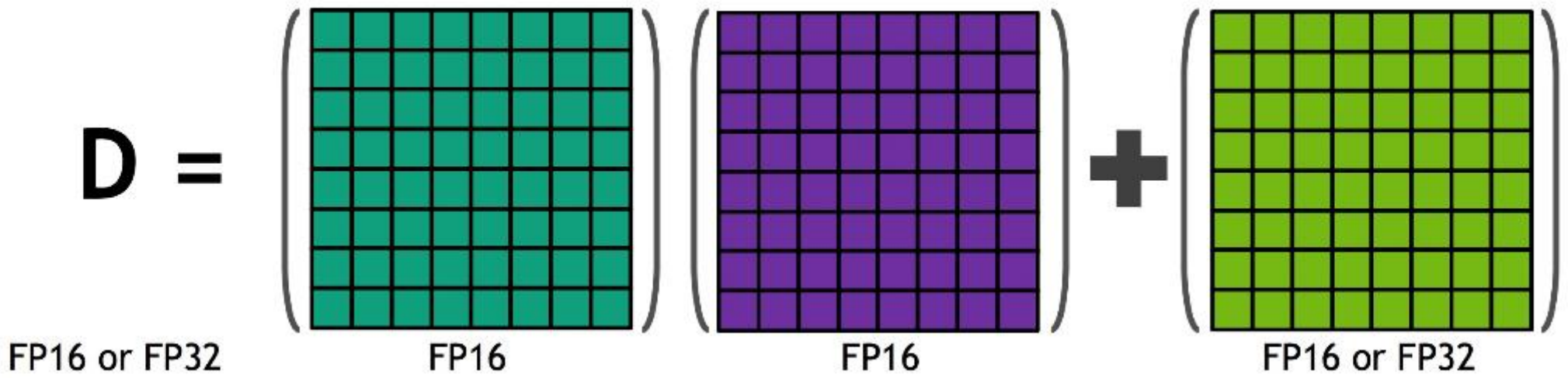
FP16 or FP32 — FP16 — FP16 — FP16 or FP32

$$D = AB + C$$

TENSOR CORE

16x16x16 Warp Matrix Multiply and Accumulate (WMMA)

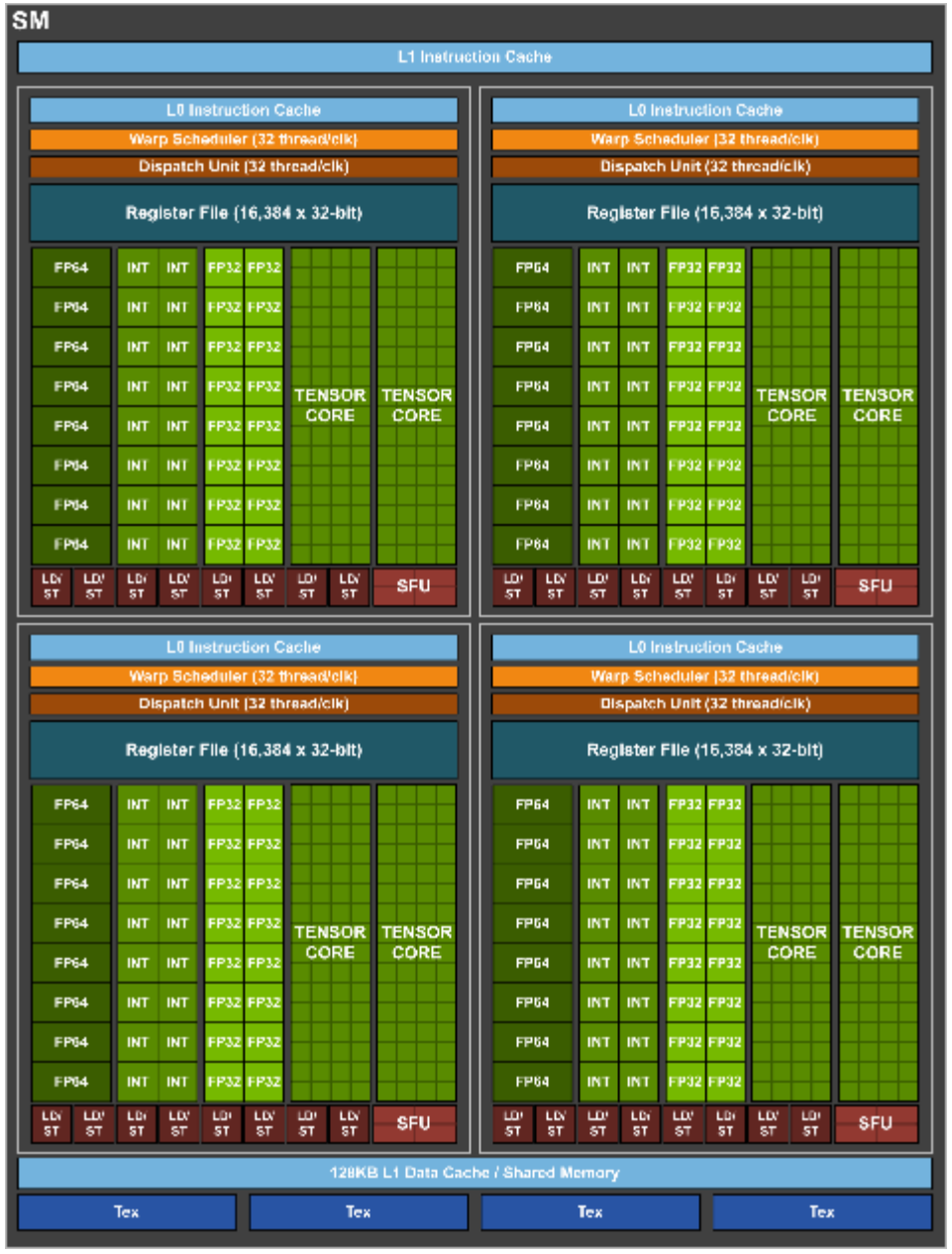




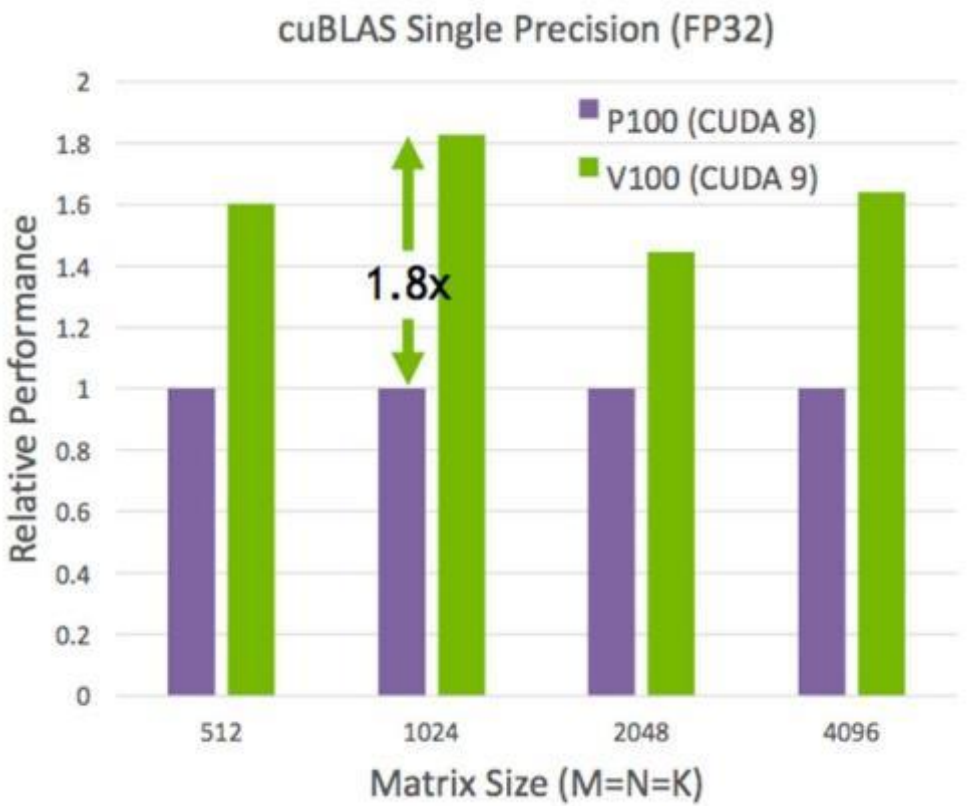
$$D = AB + C$$

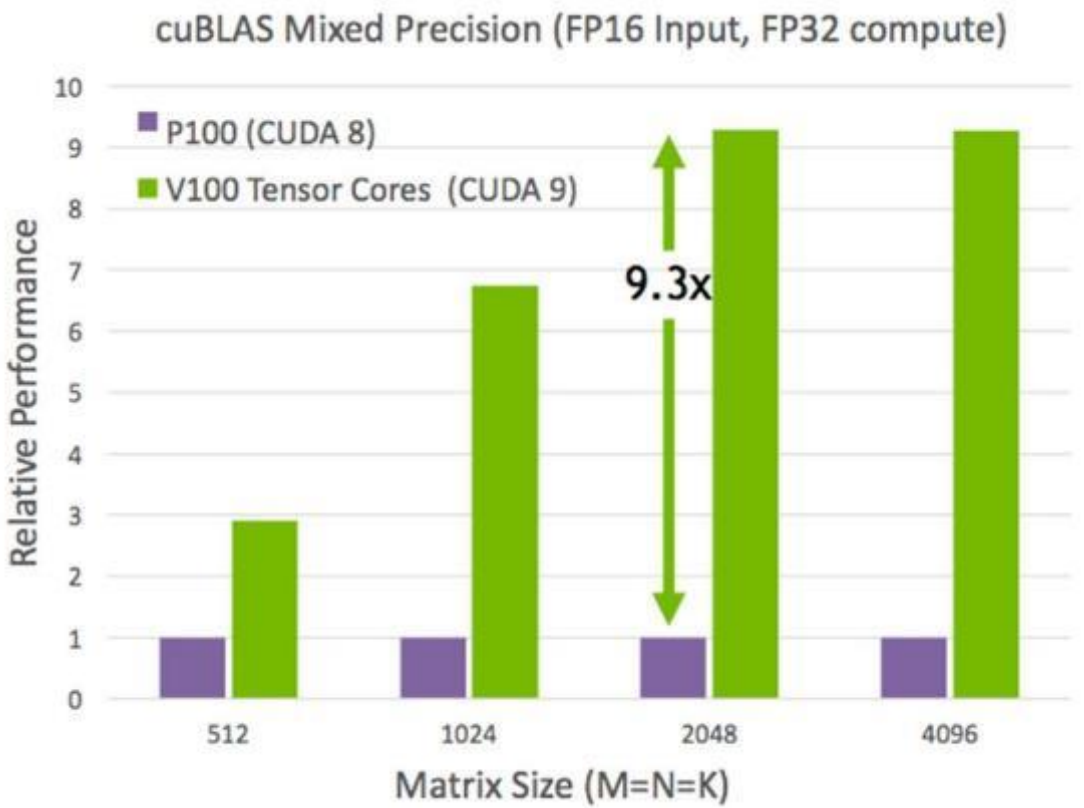
VOLTA GV100 SM

	GV100
FP32 units	64
FP64 units	32
INT32 units	64
Tensor Cores	8
Register File	256 KB
Unified L1/Shared memory	128 KB
Active Threads	2048



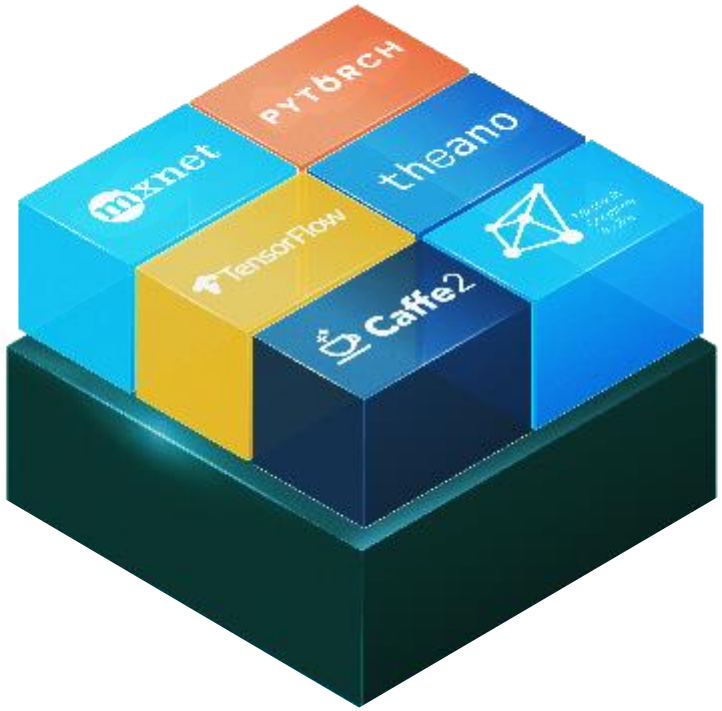
A GIANT LEAP FOR DEEP LEARNING





V100 measured on pre-production hardware.

USING TENSOR CORES



NVIDIA cuDNN, cuBLAS, TensorRT

Volta Optimized Frameworks and Libraries

```
__device__ void tensor_op_16_16_16(
    float *d, half *a, half *b, float *c)
{
  wmma::fragment<matrix_a, …> Amat;
  wmma::fragment<matrix_b, …> Bmat;
  wmma::fragment<matrix_c, …> Cmat;

  wmma::load_matrix_sync(Amat, a, 16);
  wmma::load_matrix_sync(Bmat, b, 16);
  wmma::fill_fragment(Cmat, 0.0f);

  wmma::mma_sync(Cmat, Amat, Bmat, Cmat);

  wmma::store_matrix_sync(d, Cmat, 16,
                          wmma::row_major);
}
```

CUDA C++

Warp-Level Matrix Operations

VOLTA MULTI-PROCESS SERVICE

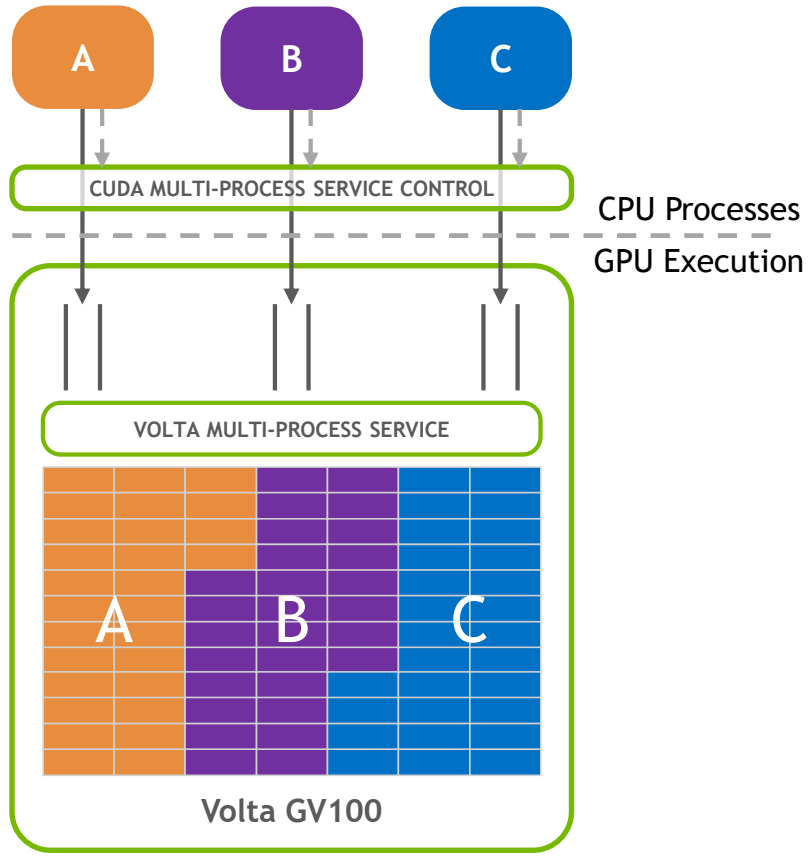
(MPS)

Volta MPS Enhancements:

- Reduced launch latency
- Improved launch throughput
- Improved quality of service with scheduler partitioning
  - More reliable performance
- 3x more clients than Pascal

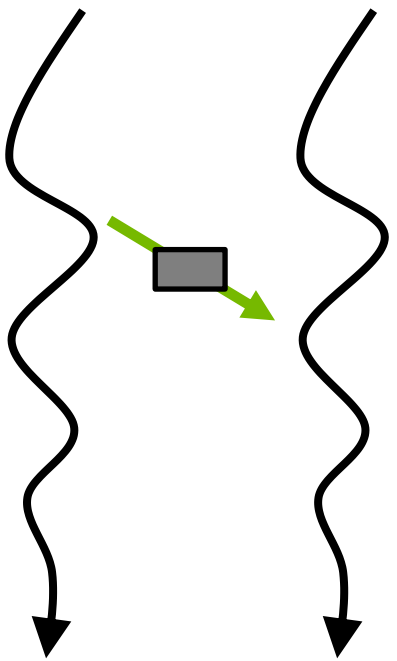
Hardware Accelerated Work Submission

Hardware Isolation



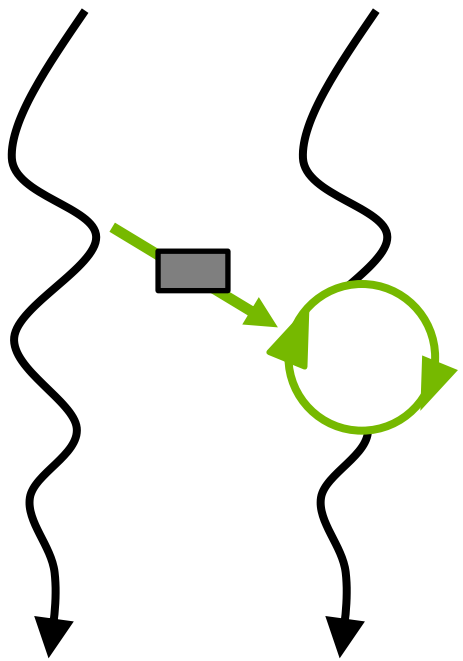
VOLTA: INDEPENDENT THREAD SCHEDULING

Communicating Algorithms



Pascal: Lock-Free Algorithms

Threads cannot wait for messages



Volta: Starvation Free Algorithms

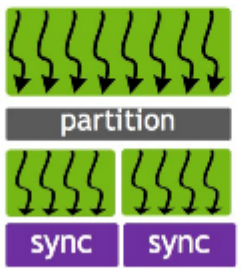
Threads may wait for messages

SYNCHRONIZE AT ANY SCALE

Three Key Capabilities

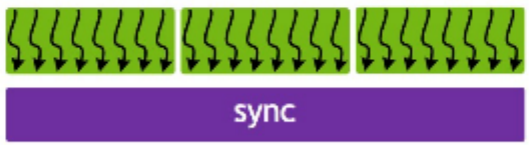
FLEXIBLE GROUPS

Define and Synchronize Arbitrary Groups of Threads

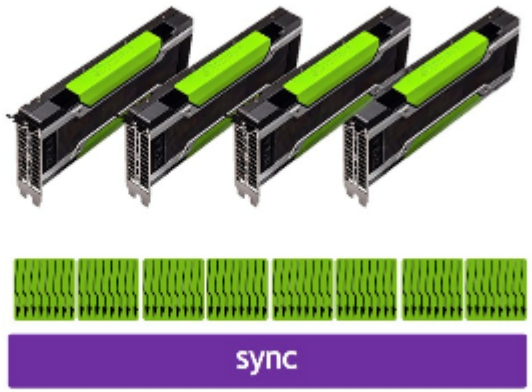


WHOLE-GRID SYNCHRONIZATION

Synchronize Multiple Thread Blocks



MULTI-GPU SYNCHRONIZATION

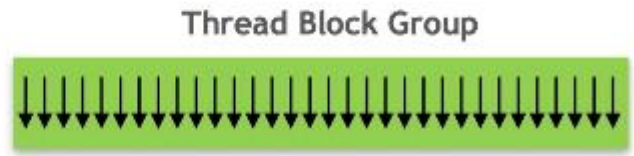


COOPERATIVE GROUPS BASICS

Flexible, Explicit Synchronization

Thread groups are explicit objects in your program

```
thread_group block = this_thread_block();
```



You can synchronize threads in a group

```
block.sync();
```

Create new groups by partitioning existing groups

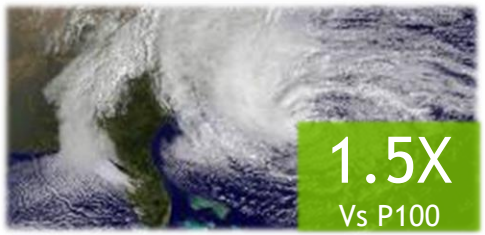
```
thread_group tile32 = tiled_partition(block, 32);
thread_group tile4 = tiled_partition(tile32, 4);
```

Partitioned groups can also synchronize

```
tile4.sync();
```

SINGLE UNIVERSAL GPU FOR ALL ACCELERATED WORKLOADS

BOOSTS ALL ACCELERATED WORKLOADS



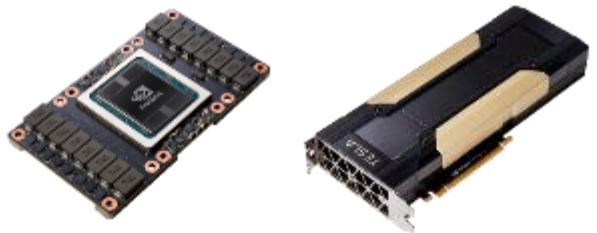
HPC

AI Training

AI Inference

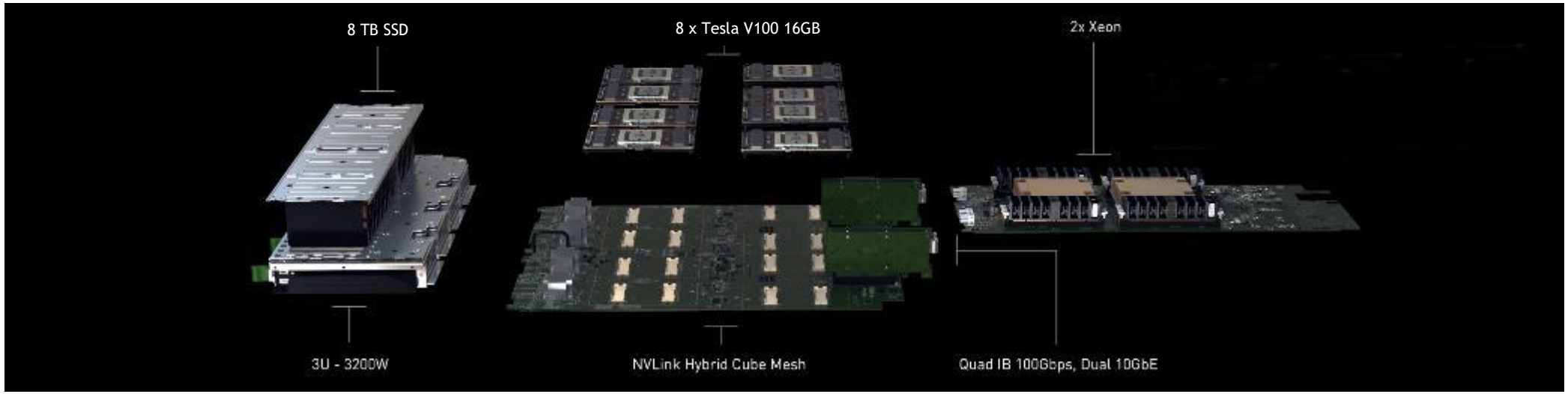
Virtual Desktop

V100 UNIVERSAL GPU



NVIDIA DGX-1

Highest Performance, Fully Integrated V100 HW System

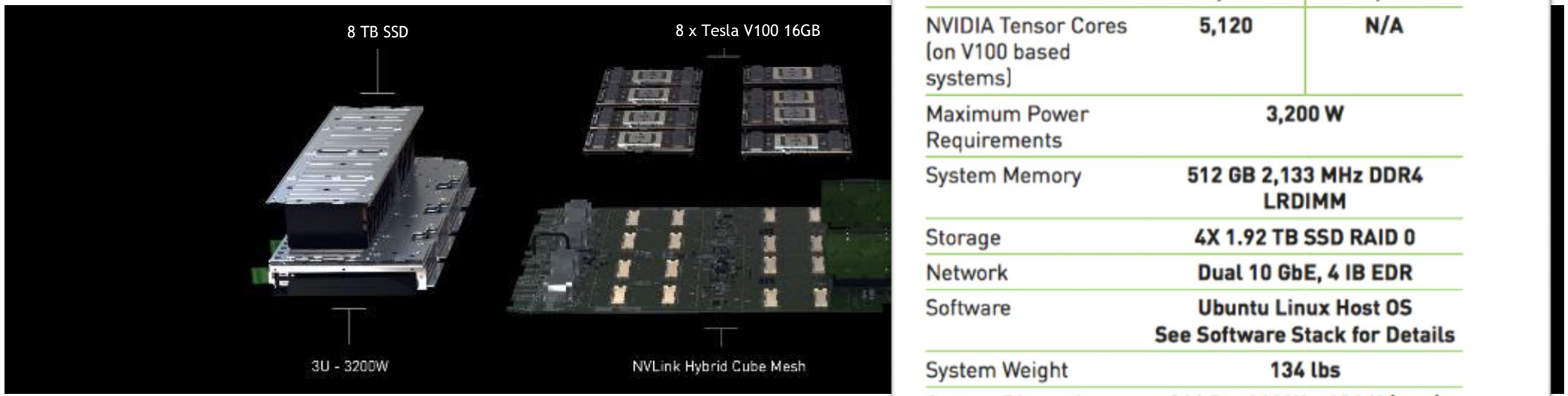


960 TFLOPS | 8x Tesla V100 16GB | 300 GB/s NVLink Hybrid Cube Mesh

2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W

NVIDIA DG

Highest Performance, Fully Integ



960 TFLOPS | 8x Tesla V100 16GB | 300 GI

2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps

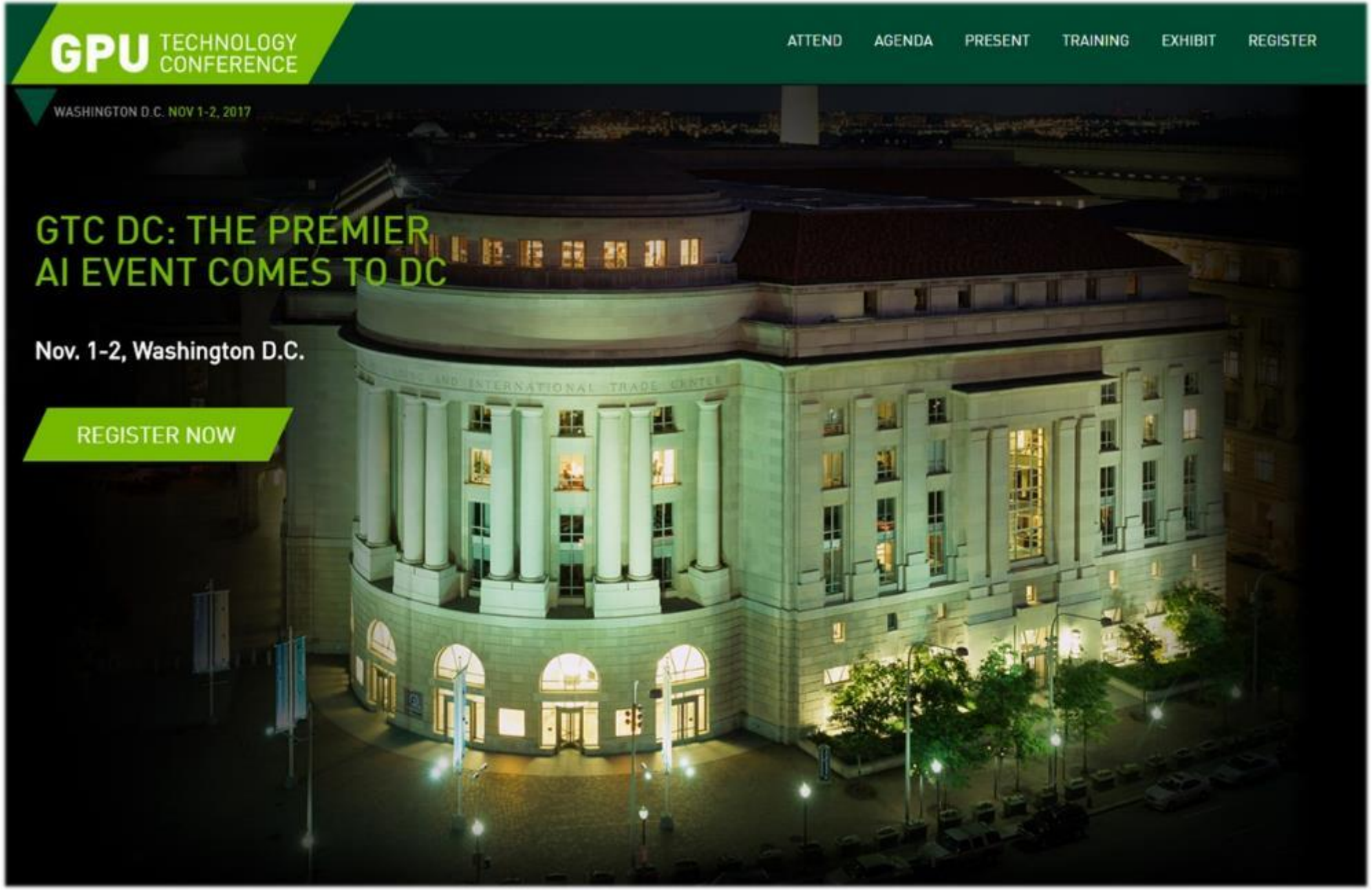
SYSTEM SPECIFICATIONS

GPUs	8X Tesla V100	8X Tesla P100
TFLOPS (GPU FP16)	960	170
GPU Memory	128 GB total system	
CPU	Dual 20-Core Intel Xeon E5-2698 v4 2.2 GHz	
NVIDIA CUDA® Cores	40,960	28,672
NVIDIA Tensor Cores (on V100 based systems)	5,120	N/A
Maximum Power Requirements	3,200 W	
System Memory	512 GB 2,133 MHz DDR4 LRDIMM	
Storage	4X 1.92 TB SSD RAID 0	
Network	Dual 10 GbE, 4 IB EDR	
Software	Ubuntu Linux Host OS See Software Stack for Details	
System Weight	134 lbs	
System Dimensions	866 D x 444 W x 131 H (mm)	
Packing Dimensions	1,180 D x 730 W x 284 H (mm)	
Operating Temperature Range	10–35 °C	

NVIDIA DGX STATION

At a Glance

GPUs	4x NVIDIA® Tesla® V100
TFLOPS (GPU FP16)	480
GPU Memory	16 GB per GPU
NVIDIA Tensor Cores	2,560 (total)
NVIDIA CUDA Cores	20,480 (total)
CPU	Intel Xeon E5-2698 v4 2.2 GHz (20-core)
System Memory	256 GB LRDIMM DDR4
Storage	Data: 3 x 1.92 TB SSD RAID 0 OS: 1 x 1.92 TB SSD
Network	Dual 10 Gb LAN
Display	3x DisplayPort, 4K Resolution
Acoustics	< 35 dB
Maximum Power Requirements	1500 W
Operating Temperature Range	10 - 30 °C
Software	Ubuntu Desktop Linux OS DGX Recommended GPU Driver CUDA Toolkit


GPU TECHNOLOGY CONFERENCE
ATTEND
AGENDA
PRESENT
TRAINING
EXHIBIT
REGISTER
WASHINGTON D.C. NOV 1-2, 2017
GTC DC: THE PREMIER AI EVENT COMES TO DC
Nov. 1-2, Washington D.C.
REGISTER NOW

Thank You !