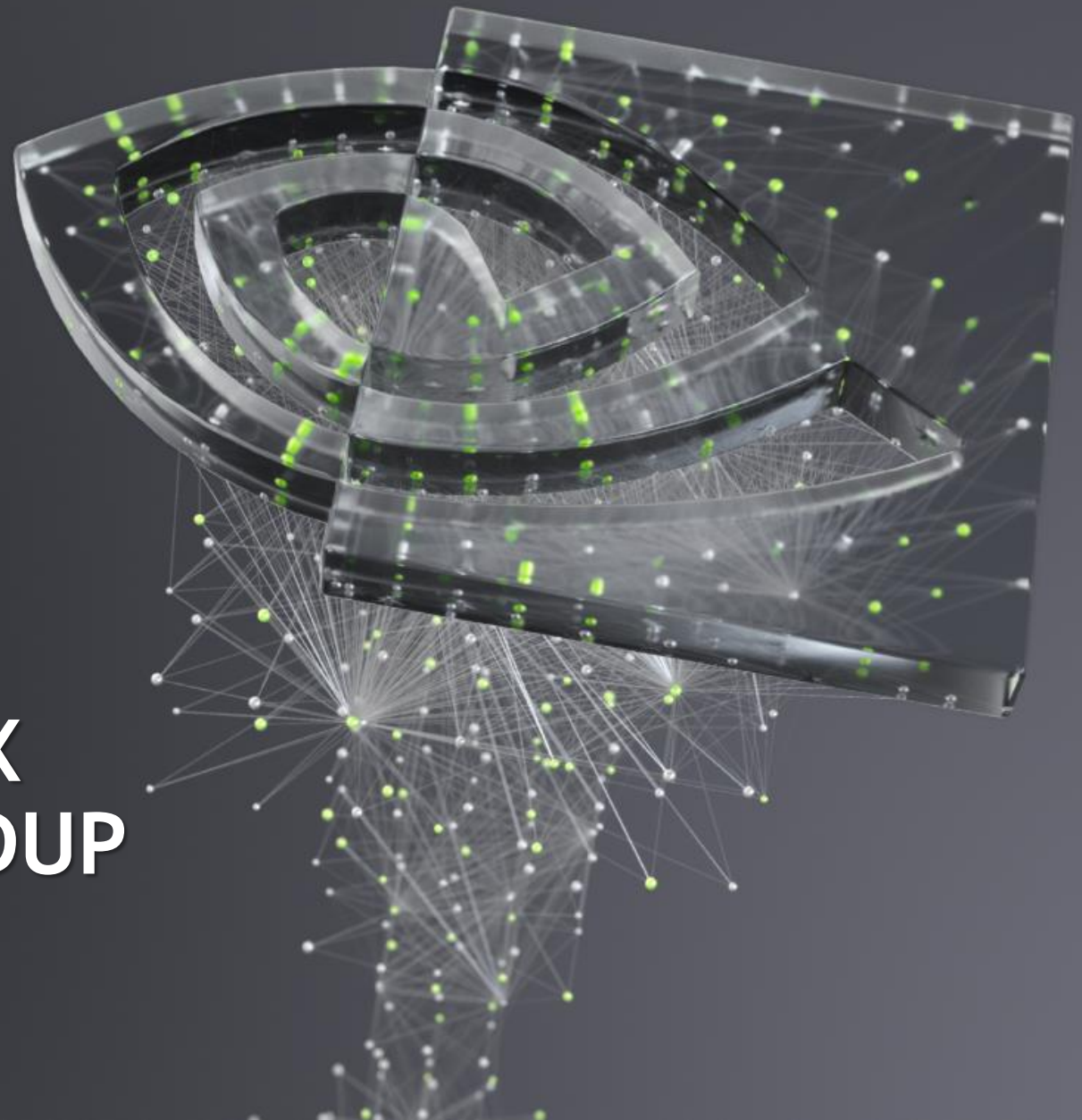
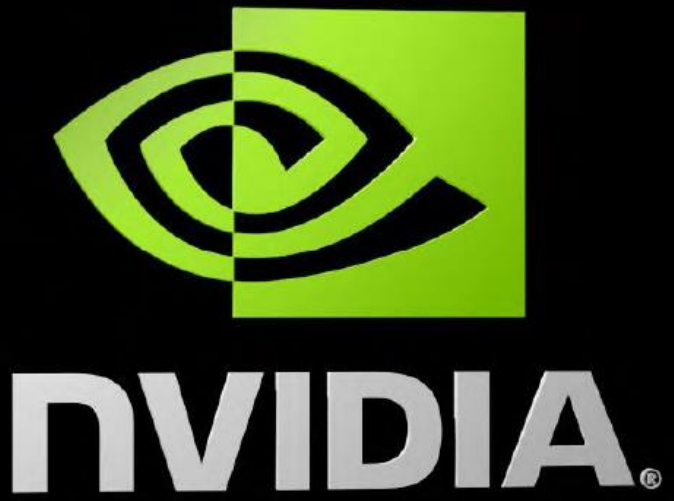


NVIDIA & MELLANOX INNOVATION ROUNDUP

John Ashley
General Manager, Financial Services and Technology
NVIDIA


NVIDIA

Mellanox
TECHNOLOGIES

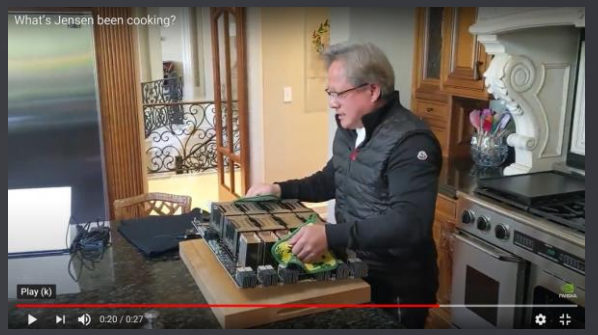
THREE HIGHLIGHTS FROM GTC 2020

Sorry, Jensen was busy, you get me…

Bluefield 2 & ConnectX-6
Smarter, Faster NICs

Spectrum 3
Single Chip Ethernet Switches

NVIDIA A100
The Next Generation GPU



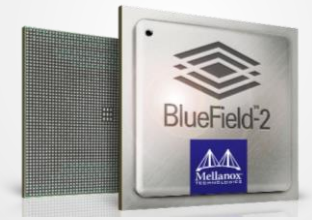
BLUEFIELD-2 - ULTIMATE DATA PROCESSING UNIT

High Performance IPU

- 8 Arm® A72 CPUs @ 2GHz-2.5GHz
- Embedded ConnectX-6 Dx controller
- Dual port 10/25/100/200G ETH
- Dual EDR/Single HDR InfiniBand
- PCIe Gen 3.0/4.0 x16

Advanced Hardware Accelerations

- Networking : RDMA, CC, QoS
- Storage : NVMeoF, Dedup (SHA), CRC64
- Storage : (De)Compression, EC, RAID
- Virtualization : OVS, VirtIO, SR-IOV



BlueField-2

Scalability and Programmability

- SmartNIC and as a controller in system
- User's applications to run in ARM
- Integrated control and data planes

Security

- SmartNICs isolated from main host
- Secured Hardware (Root-of-Trust)
- Crypto offloads : IPsec, TLS, AES-XTS
- Regular Expression (RegEx) and DPI

NVIDIA

ConnectX NICs
Fast and Secure

	Extreme Security, Scale Performance	Agile Future Proof Secured 25G
Feeds &Speeds	2x100/1x200Gbs PAM4 PCIe 4.0 x16	2x25/1x50Gbs NRZ PCIe x8 Gen 4.0
Security	IPsec TLS	IPsec
Virtualization	ASAP² VirtIO ASAP² Enhancements	
RoCE	RoCE /IPsec ZTR at Scale	
Time Sensitive Networking	Enhanced PTP 1588v2 Full 5T for 5G PPS In/Out	Enhanced PTP 1588v2

Single Chip Ethernet Switch Solutions

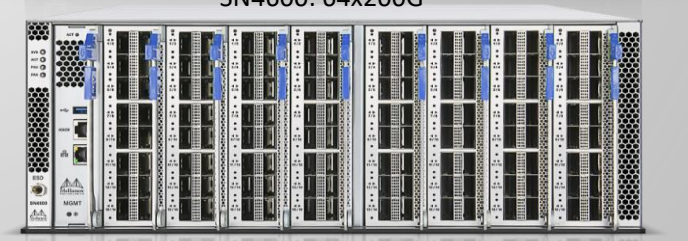
Spectrum-3

GA Q3 2020

SN4700: 32x400G

SN4600C: 64x100G

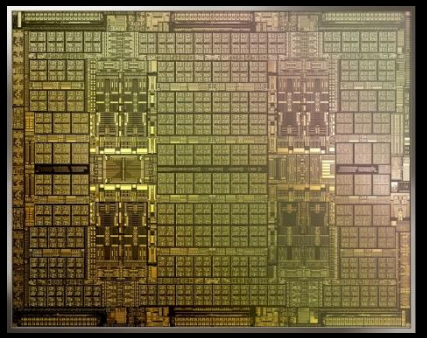
SN4600: 64x200G



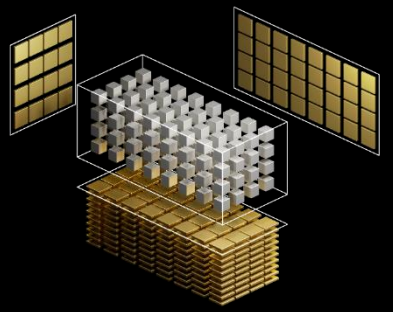
SN4800: 128x100G on a SINGLE CHIP

NVIDIA

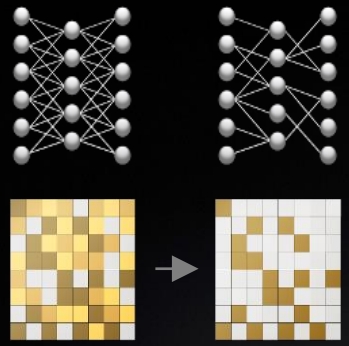
5 MIRACLES OF A100



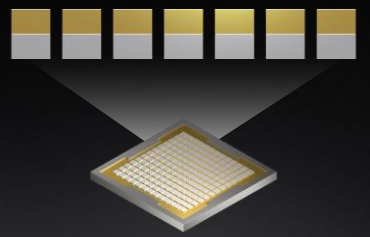
Ampere
World's Largest 7nm chip
54B XTORS, HBM2



3rd Gen Tensor Cores
Faster, Flexible, Easier to use
20x AI Perf with TF32



New Sparsity Acceleration
Harness Sparsity in AI Models
2x AI Performance



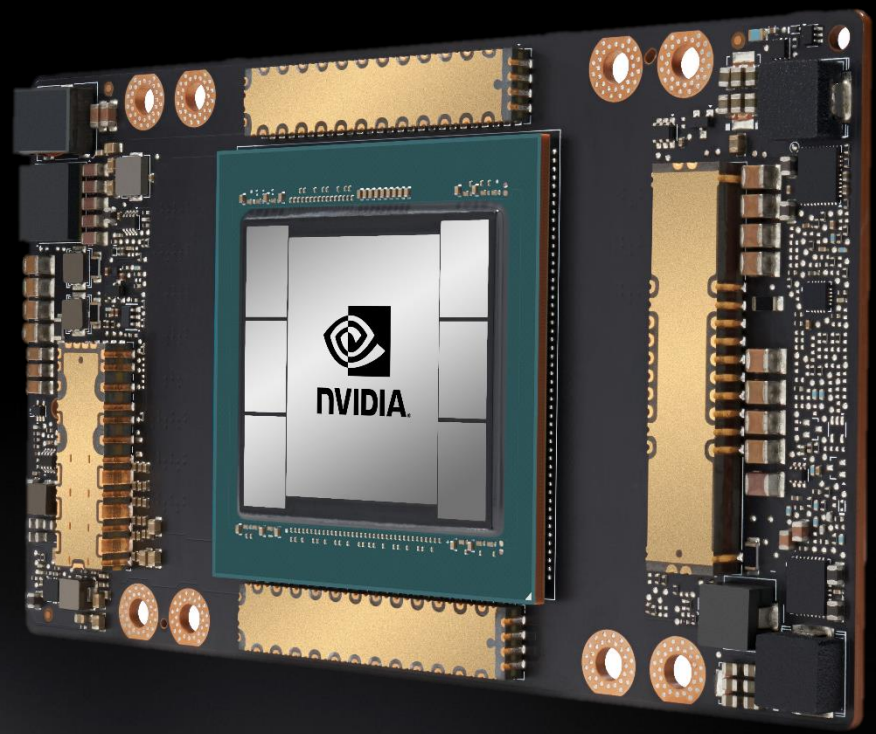
New Multi-Instance GPU
Optimal utilization with right sized GPU
7x Simultaneous Instances per GPU

3rd Gen NVLINK and NVSWITCH
Efficient Scaling to Enable Super GPU
2X More Bandwidth

NVIDIA.

NVIDIA A100 DETAILED SPECS

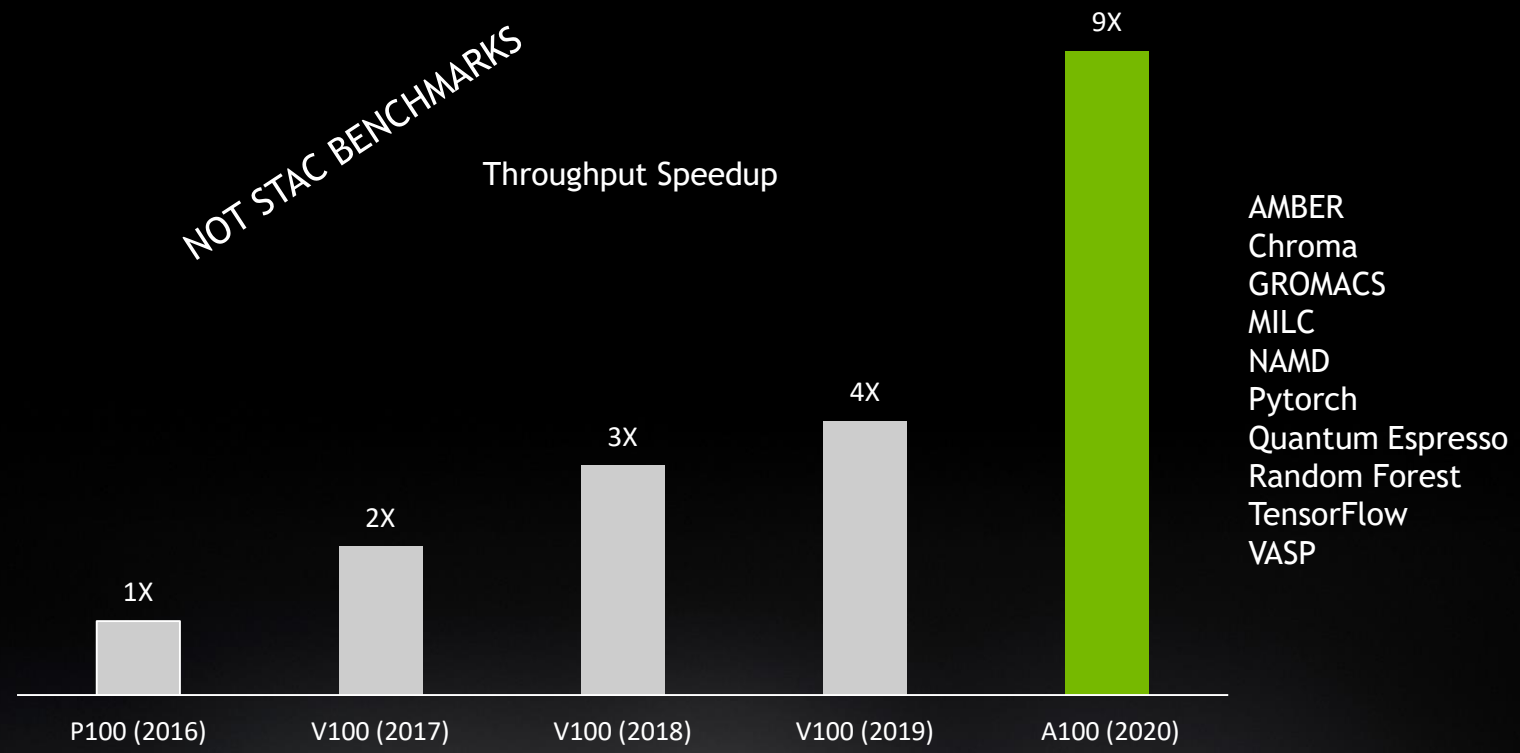
	Peak Performance
Transistor Count	54 billion
Die Size	826 mm^2
FP64 CUDA Cores	3,456
FP32 CUDA Cores	6,912
Tensor Cores	432
Streaming Multiprocessors	108
FP64	9.7 teraFLOPS
FP64 Tensor Core	19.5 teraFLOPS
FP32	19.5 teraFLOPS
TF32 Tensor Core	156 teraFLOPS | 312 teraFLOPS*
BFLOAT16 Tensor Core	312 teraFLOPS | 624 teraFLOPS*
FP16 Tensor Core	312 teraFLOPS | 624 teraFLOPS*
INT8 Tensor Core	624 TOPS | 1,248 TOPS*
INT4 Tensor Core	1,248 TOPS | 2,496 TOPS*
GPU Memory	40 GB @ 1.5 TB/s
Interconnect	NVLink 600 GB/s PCIe Gen4 64 GB/s
Multi-Instance GPUs	Various Instance sizes with up to 7MIGs @5GB
Form Factor	4/8/16 SXM GPUs in HGX A100
Max Power	400W (SXM)



* Includes Sparsity

9X MORE PERFORMANCE IN 4 YEARS

Beyond Moore's Law With Full Stack Innovation



Geometric Mean of application speedups vs. P100 : Benchmark Application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT Large Fine Tuner], Quantum Espresso [AUSURF112-jR]; Random Forest FP32 [make_blobs (160000 x 64 : 10)], TensorFlow [ResNet-50], VASP 6 [Si Huge], |GPU node: with dual-socket CPUs with 4x P100, V100, or A100 GPUs.

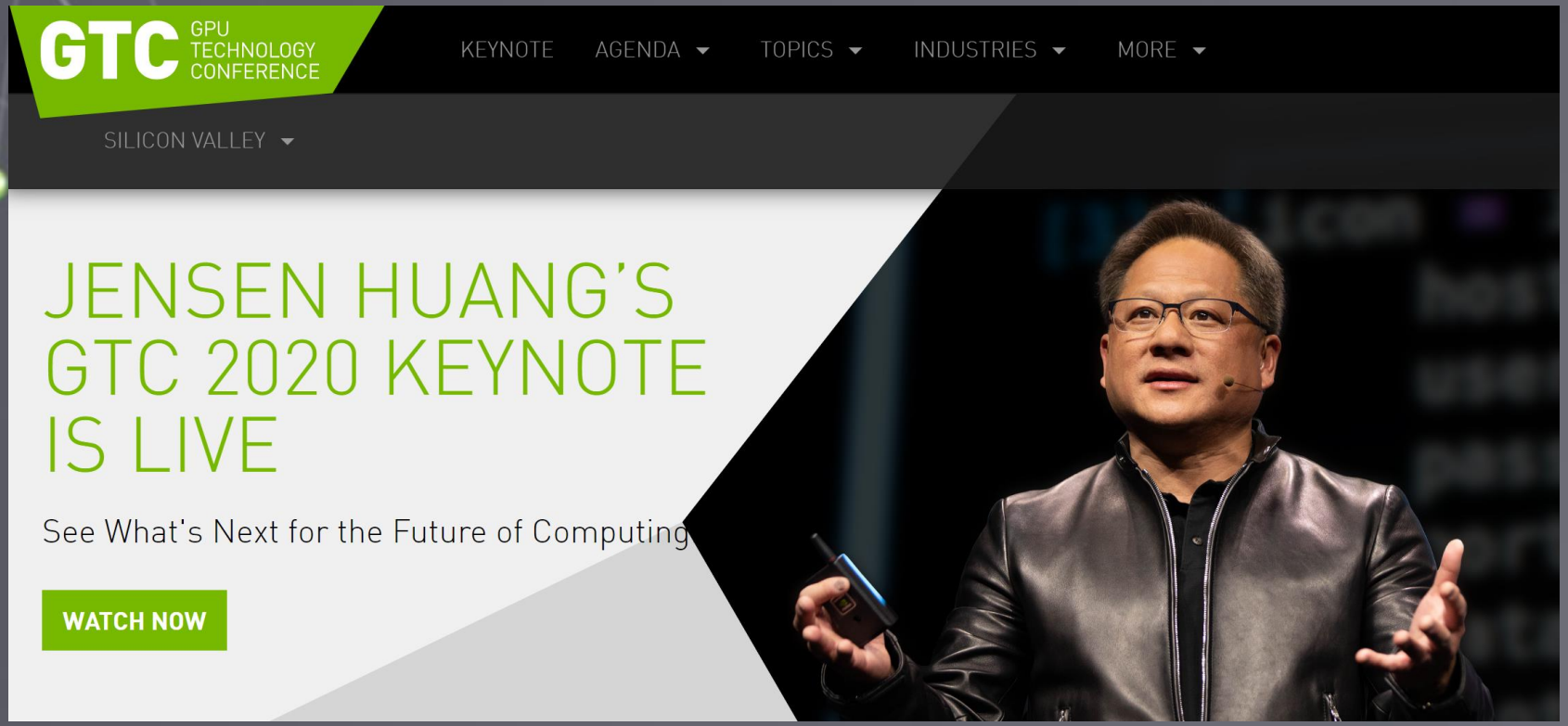
NVIDIA GTC 2020

https://www.nvidia.com/en-us/gtc/

https://www.youtube.com/nvidia

Introduction & NVIDIA Mellanox

NVIDIA A100



jashley@nvidia.com liorp@mellanox.com