# High Performance Computing and Data Infrastructure

## Dissecting DGX workstation



**2024-2025 @ Università di Trieste**

# Agenda

- DGX overview: datacenter in a box
- DGX hardware elements and all the rest:
  - nvidia-ampere-architecture-whitepaper.pdf
- Exploring DGX
  - Lab sessions..

# NVIDIA DGX Station A100 Offers Researchers AI Data-Center-in-a-Box

World's Only Petascale Integrated AI Workgroup Server, Second-Gen DGX Station Packs Four NVIDIA A100 GPUs, Debuts with up to 320GB of GPU Memory to Bring AI into Offices and Labs

November 16, 2020

# The machine


Figure 36. NVIDIA DGX 100 System

# The computational kernel: A100 40GB RAM

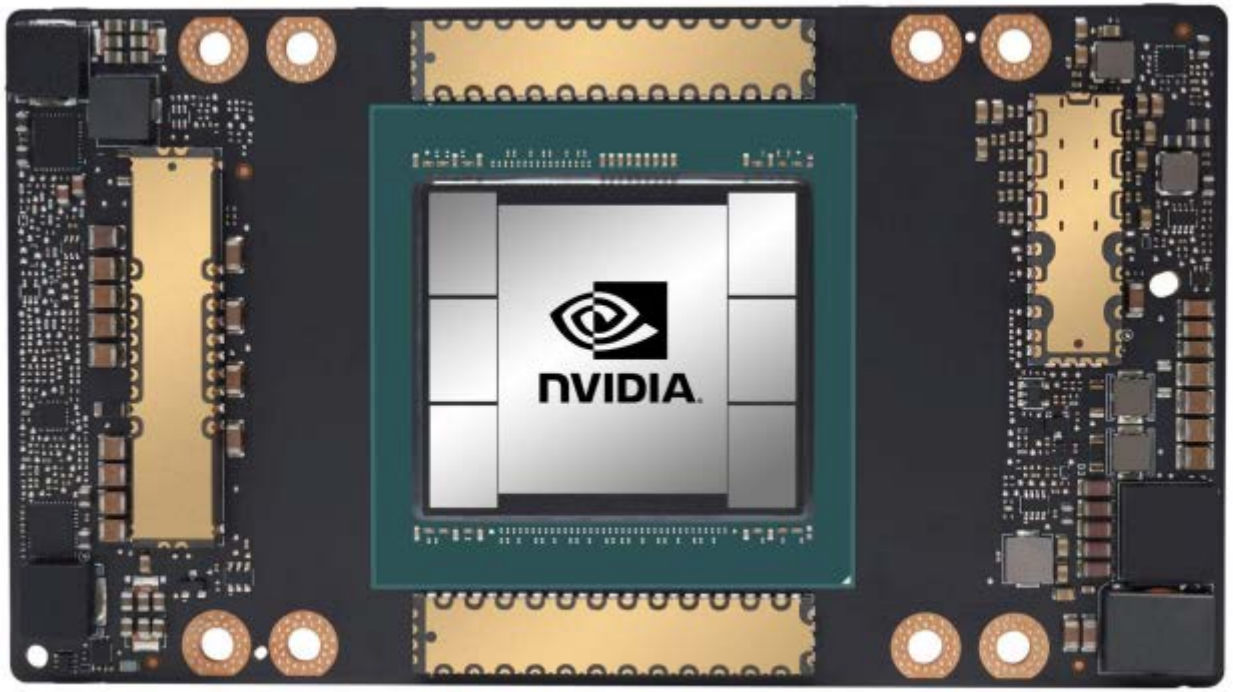
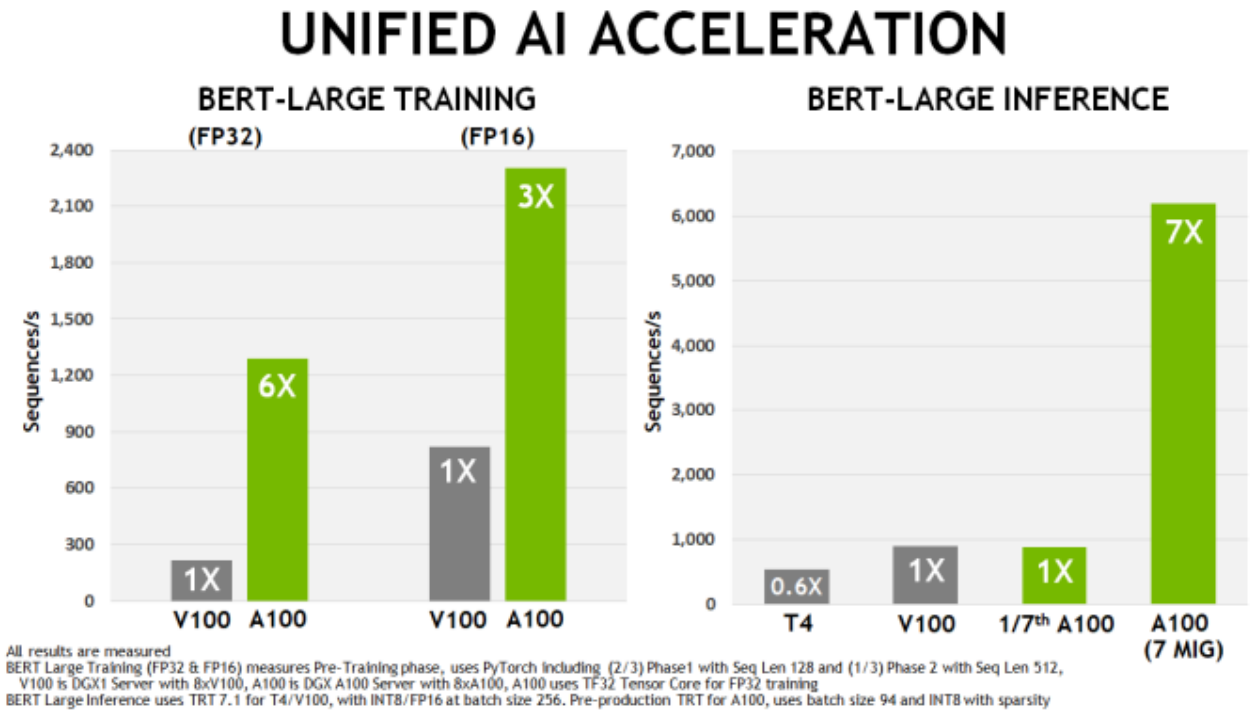



Figure 3. NVIDIA A100 GPU on new SXM4 Module

# Declared performance: AI




A100 GPU performance in BERT deep learning training and inference scenarios compared to NVIDIA Tesla V100 and NVIDIA Tesla T4.

Figure 4. Unified AI Acceleration for BERT-LARGE Training and Inference

# Declared performance: HPC

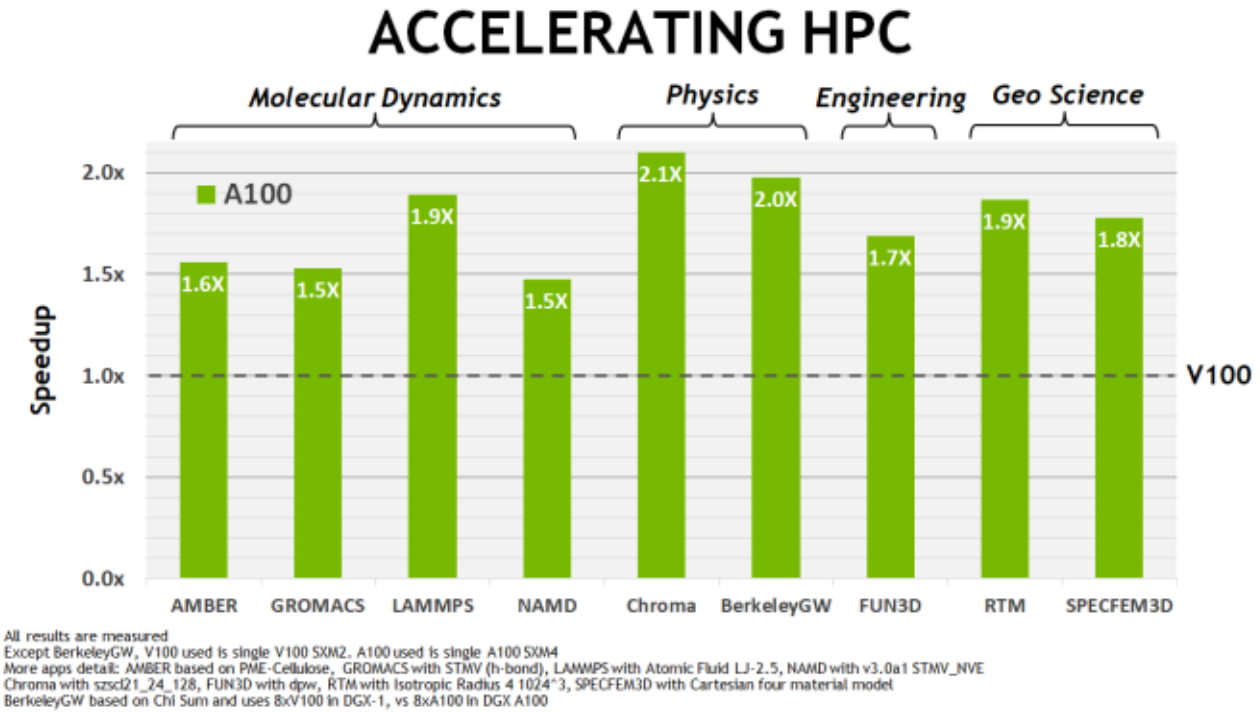



Figure 5. A100 GPU HPC application speedups compared to NVIDIA Tesla V100

# TENSOR CORES: A REFRESHER

## Introduced on NVIDIA Volta V100 GPU

Tensor Cores are …

… special hardware execution units

… built to accelerate deep learning

… executing matrix multiply operations

Volta Tensor Cores

FP16/FP16 and FP16/FP32 modes

Turing Tensor Cores

+ INT8/INT32, INT4/INT32, INT1/INT32

# Third generation Tensor cores...

- Third-Generation Tensor Core together with a new Sparsity feature to deliver a further doubling of throughput with respect to V100
- New TensorFloat-32 (TF32) to accelerate FP32 input/output data running 10x faster than V100 FP32 FMA operations or 20x faster with sparsity.
- For FP16/FP32 mixed-precision DL, the A100 Tensor Core delivers 2.5x the performance of V100, increasing to 5x with sparsity.

## A100 GPU Streaming Multiprocessor (SM)

The new SM in the NVIDIA Ampere architecture-based A100 Tensor Core GPU significantly increases performance, builds upon features introduced in both the Volta and Turing SM architectures, and adds many new capabilities.

The A100 **Third-Generation Tensor Cores** enhance operand sharing and improve efficiency, and add powerful new data types including:

- TF32 Tensor Core instructions which accelerate processing of FP32 data
- IEEE-compliant FP64 Tensor Core instructions for HPC
- BF16 Tensor Core instructions at the same throughput as FP16

Table 1. NVIDIA A100 Tensor Core GPU Performance Specs

| | |
|---|---|
| Peak FP64[1] | 9.7 TFLOPS |
| Peak FP64 Tensor Core[1] | 19.5 TFLOPS |
| Peak FP32[1] | 19.5 TFLOPS |
| Peak FP16[1] | 78 TFLOPS |
| Peak BF16[1] | 39 TFLOPS |
| Peak TF32 Tensor Core[1] | 156 TFLOPS \| 312 TFLOPS[2] |
| Peak FP16 Tensor Core[1] | 312 TFLOPS \| 624 TFLOPS[2] |
| Peak BF16 Tensor Core[1] | 312 TFLOPS \| 624 TFLOPS[2] |
| Peak INT8 Tensor Core[1] | 624 TOPS \| 1,248 TOPS[2] |
| Peak INT4 Tensor Core[1] | 1,248 TOPS \| 2,496 TOPS[2] |

**1 - Peak rates are based on GPU Boost Clock.**
**2 - Effective TFLOPS / TOPS using the new Sparsity feature**

# NVIDIA A100 Tensor Core GPU implementation

- IT includes the following units:
  - 7 GPCs, 7 or 8 TPCs/GPC, 2 SMs/TPC, up to 16 SMs/GPC, 108 SMs
  - 64 FP32 CUDA Cores/SM, 6912 FP32 CUDA Cores per GPU
  - 4 Third-generation Tensor Cores/SM, 432 Third-generation Tensor Cores per GPU
  - 5 HBM2 stacks, 10 512-bit Memory Controllers

Figure 6 shows a full GA100 GPU with 128 SMs. The A100 is based on GA100 and has 108 SMs.

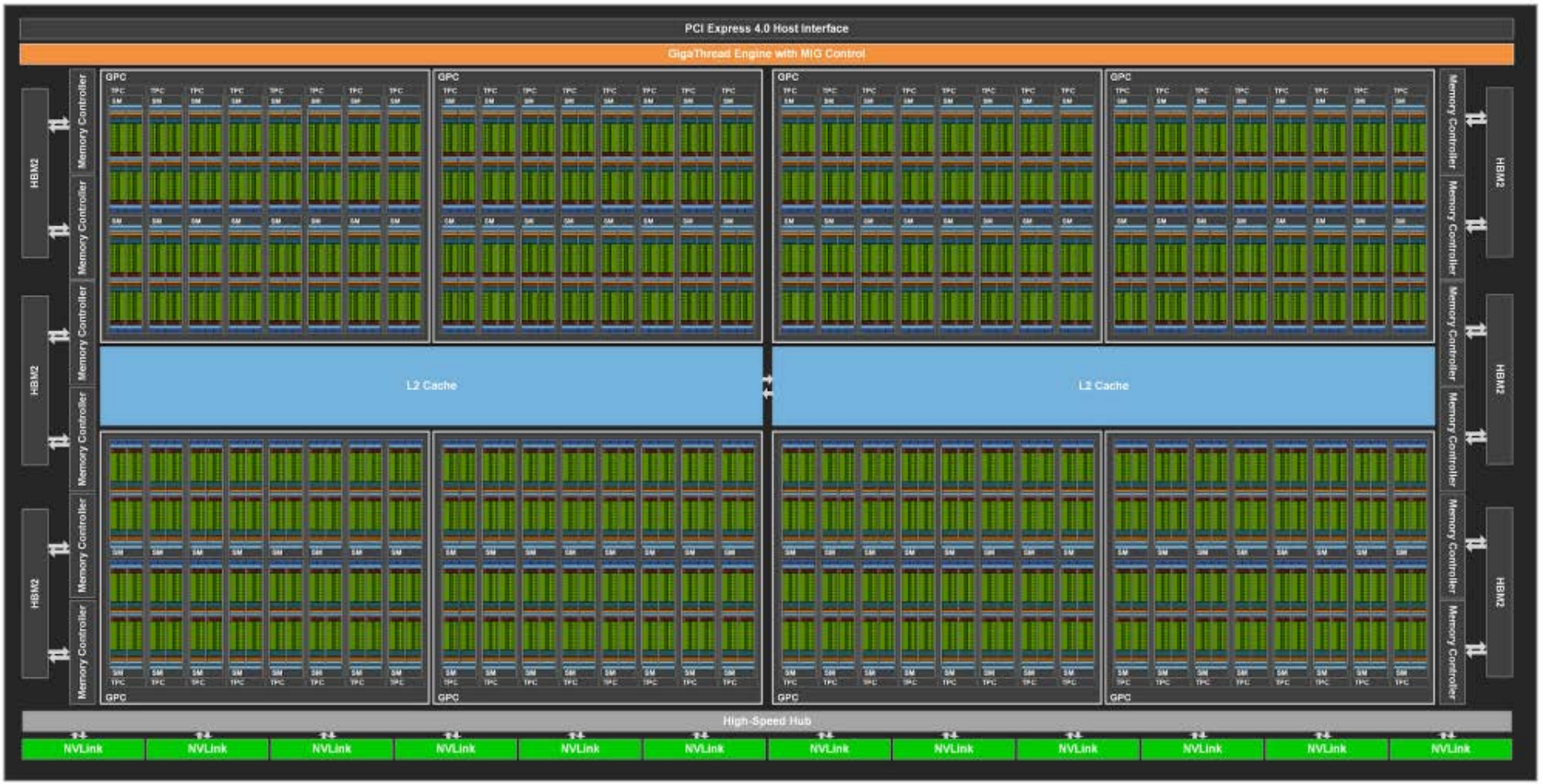
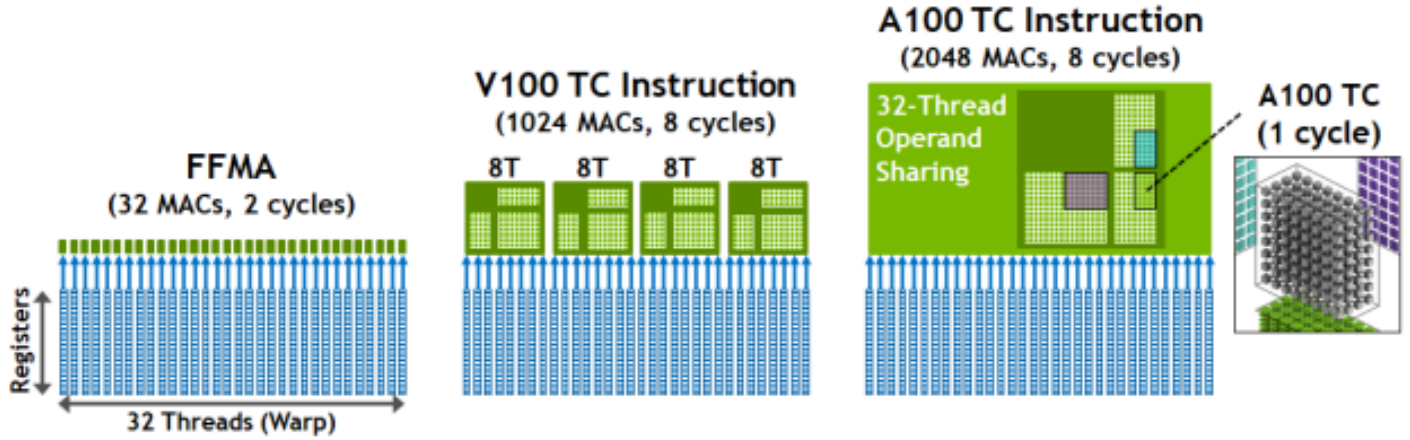



Figure 6. GA100 Full GPU with 128 SMs (A100 Tensor Core GPU has 108 SMs)

# A100 SM


Figure 7. GA100 Streaming Multiprocessor (SM)

## A100 Tensor core: 2x throughput vs. V100, >2x efficiency




| 16x16x16 matrix multiply | FFMA | V100 TC | A100 TC | A100 vs. V100 (improvement) | A100 vs. FFMA (improvement) |
|---|---|---|---|---|---|
| Thread sharing | 1 | 8 | 32 | 4x | 32x |
| Hardware instructions | 128 | 16 | 2 | 8x | 64x |
| Register reads+writes (warp) | 512 | 80 | 28 | 2.9x | 18x |
| Cycles | 256 | 32 | 16 | 2x | 16x |

A100's Tensor Core increases thread sharing by 4x over V100. For a 16x16x16 matrix multiply, A100's enhanced 16x8x16 Tensor Core (TC) instructions improve on V100 by reducing register accesses from 80 to 28, and hardware instructions issued from 16 to 2. Cycle counts are per SM partition. Note: Each V100 8x8x4 TC instruction (CUDA warp-level instruction) is translated into four lower-level MMA hardware instructions.

Figure 14. A100 Tensor Core Throughput and Efficiency

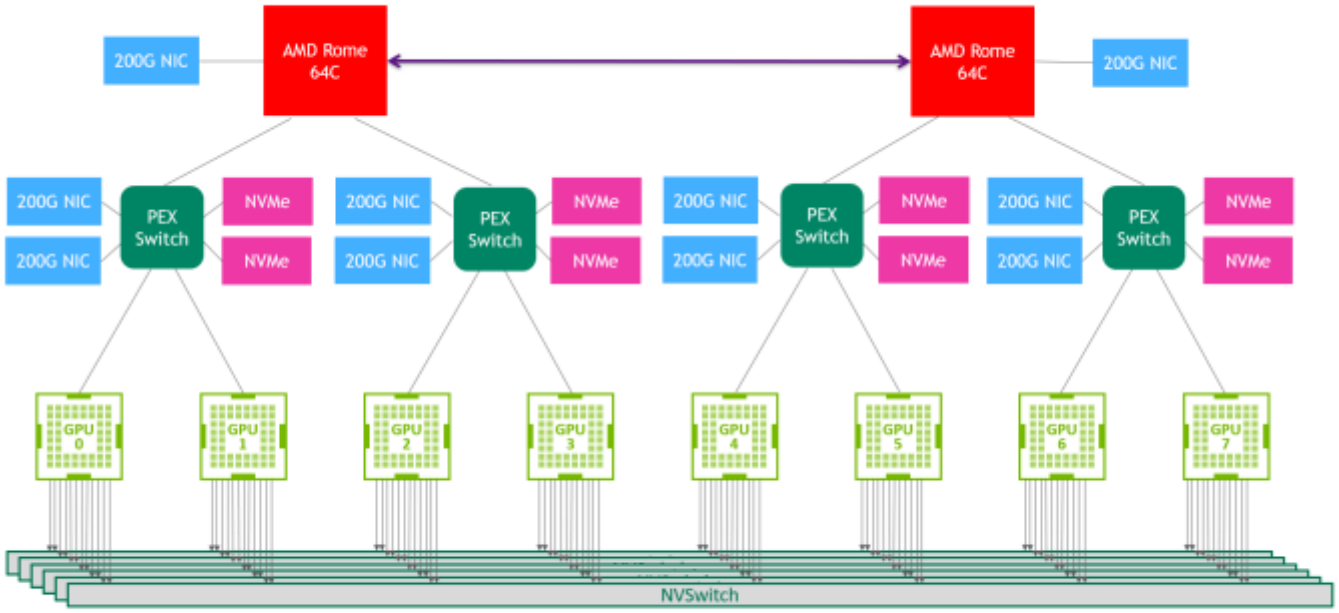
Table 5. Compute Capability: GP100 vs GV100 vs GA100

| GPU Features | NVIDIA Tesla P100 | NVIDIA Tesla V100 | NVIDIA A100 |
|---|---|---|---|
| GPU Codename | GP100 | GV100 | GA100 |
| GPU Architecture | NVIDIA Pascal | NVIDIA Volta | NVIDIA Ampere |
| Compute Capability | 6.0 | 7.0 | 8.0 |
| Threads / Warp | 32 | 32 | 32 |
| Max Warps / SM | 64 | 64 | 64 |
| Max Threads / SM | 2048 | 2048 | 2048 |
| Max Thread Blocks / SM | 32 | 32 | 32 |
| Max 32-bit Registers / SM | 65536 | 65536 | 65536 |
| Max Registers / Block | 65536 | 65536 | 65536 |
| Max Registers / Thread | 255 | 255 | 255 |
| Max Thread Block Size | 1024 | 1024 | 1024 |
| FP32 Cores/ SM | 64 | 64 | 64 |
| Ratio of SM Registers to FP32 Cores | 1024 | 1024 | 1024 |
| Shared Memory Size / SM | 64 KB | Configurable up to 96 KB | Configurable up to 164 KB |

# NVIDIA NVLink® high-speed interconnect

- Third-generation NVLink has a data rate of 50 Gbit/sec per signal pair. A single A100 NVLink provides 25 GB/second bandwidth in each direction
- The total number of links is increased to twelve in A100, versus 6 in V100, yielding 600 GB/sec total bandwidth versus 300 GB/sec for V100.

# Nvlink




Note Third-Generation NVLink connectivity through NVSwitches.

Figure 26. NVIDIA DGX A100 with Eight A100 GPUs

# PCIe Gen 4 with SR-IOV

- The A100 GPU supports PCI Express Gen 4 (PCIe Gen 4) which doubles the bandwidth of PCIe 3.0/3.1 by providing 31.5 GB/sec versus 15.75 GB/sec for x16 connections.
- The faster speed is especially beneficial for A100 GPUs connecting to PCIe 4.0-capable CPUs, and to support fast network interfaces, such as 200 Gbit/sec InfiniBand.
- A100 also supports Single Root Input/Output Virtualization (SR-IOV), which allows sharing and virtualizing a single PCIe connection for multiple processes or Virtual Machines (VMs).

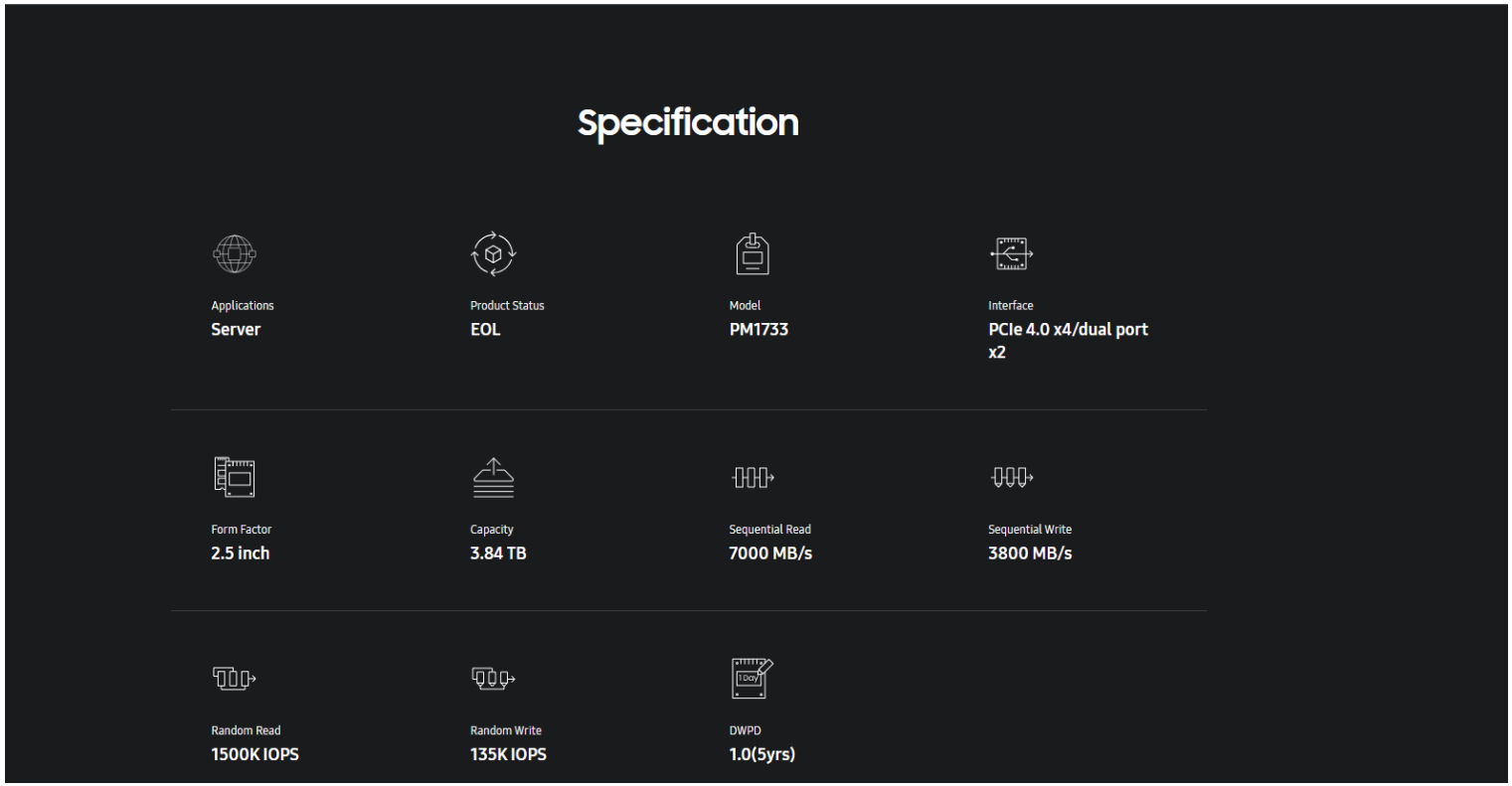
# Memory

- The A100 GPU includes 40 GB of fast HBM2 DRAM memory on its SXM4-style circuit board. The memory is organized as five active HBM2 stacks with eight memory dies per stack. With a 1215 MHz (DDR) data rate the A100 HBM2 delivers 1555 GB/sec memory.

# Local storage: 6 Samsung NVME

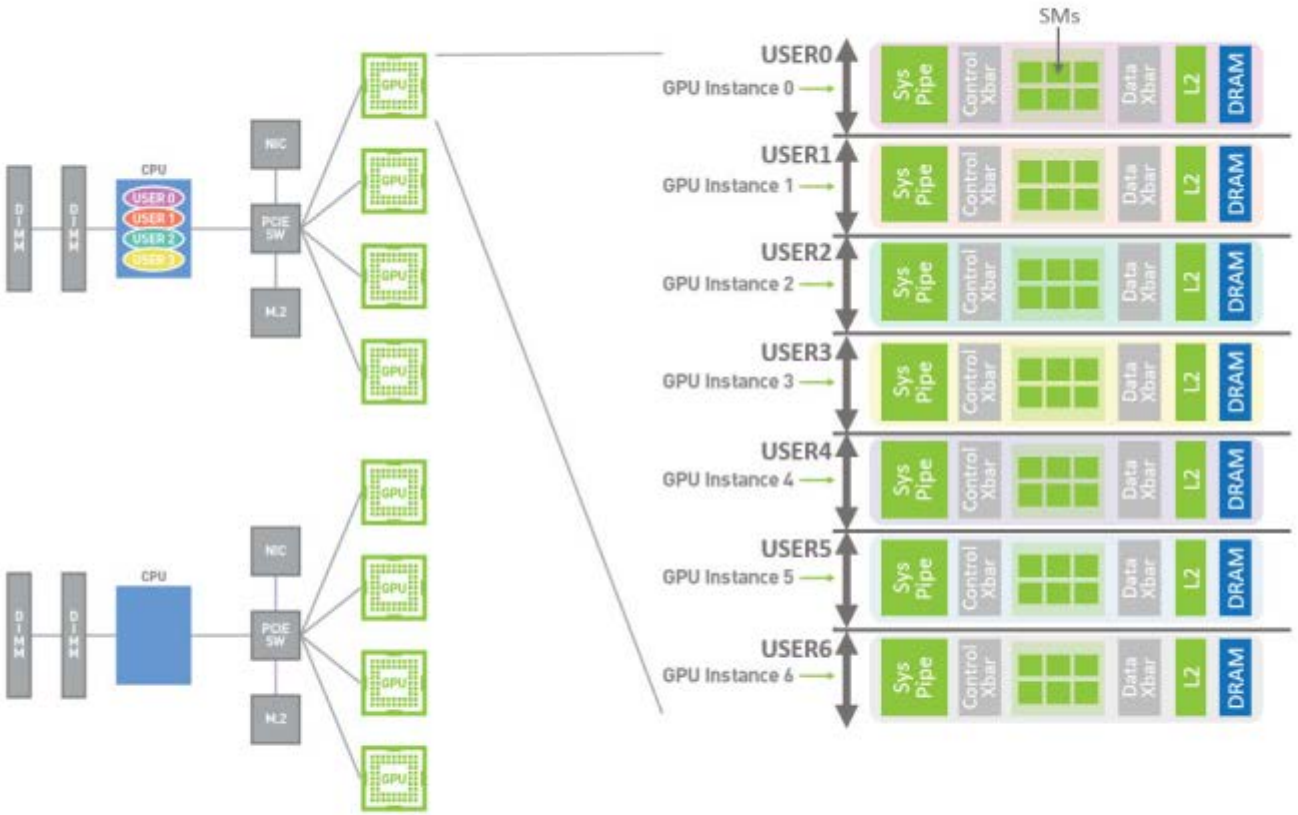
## Today's Highest-Performing NVMe™ SSDs

Samsung's PCIe® Gen 4-enabled PM1733 SSD will have double the throughput capabilities of current Gen 3 SSDs, giving it the highest performance of any SSD on the market today. The two NVMe™ SSD series come in two form factors, 2.5-inch and HHHL, with capacities ranging from 0.8TB to 30.72TB to suit the diverse needs of OEMs worldwide. The drives also ensure endurance of one or three drive writes per day (DWPD) over a five-year period.



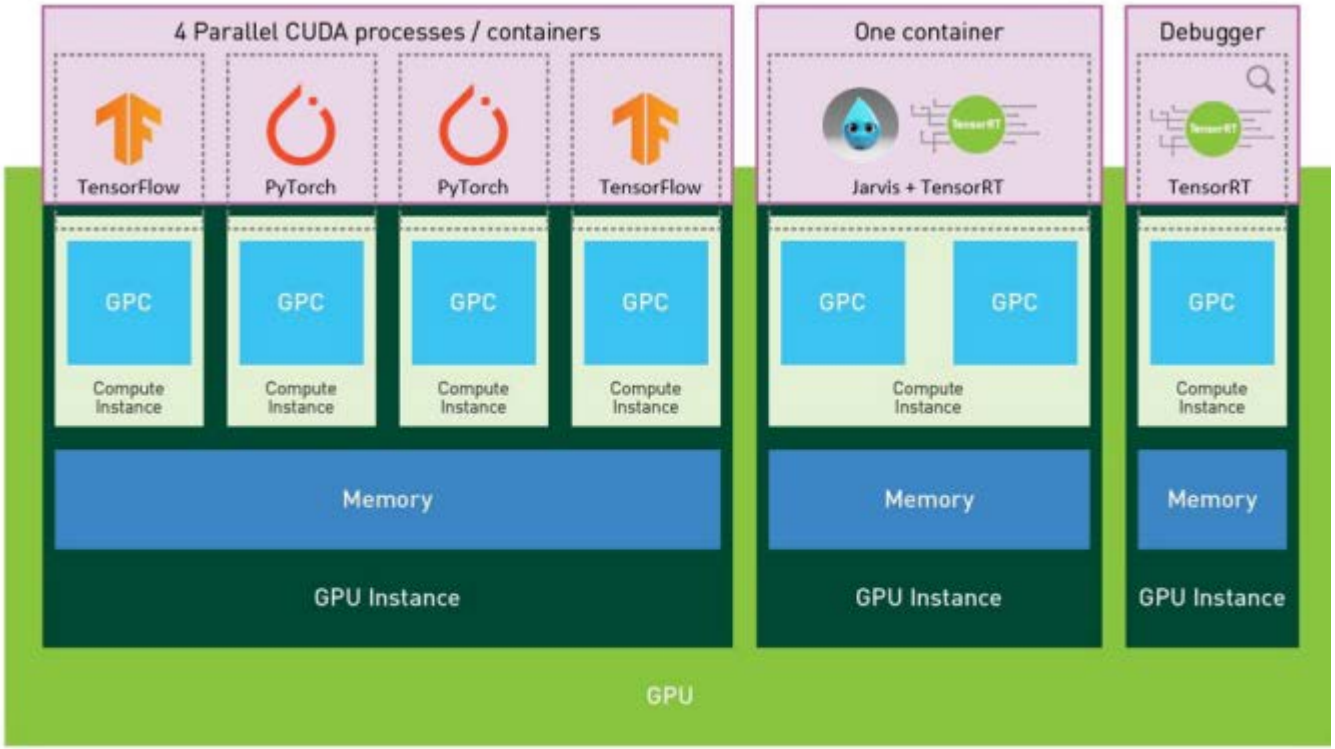
# MIG (Multi-Instance GPU) Architecture

- MIG allows compute resources to be partitioned across different Virtual Machines (VMs), and allows multiple VMs to execute simultaneously while maintaining fault isolation.
- Consistent performance can be maintained even if a VM is migrated to another GPU.
- In addition, better utilization of GPUs can be obtained by packing multiple VMs on the same GPU




This CSP MIG diagram shows how multiple independent users from the same or different organizations can be assigned their own dedicated, protected, and isolated GPU Instance within a single physical GPU. (See MIG configuration and GPU partitioning details below).

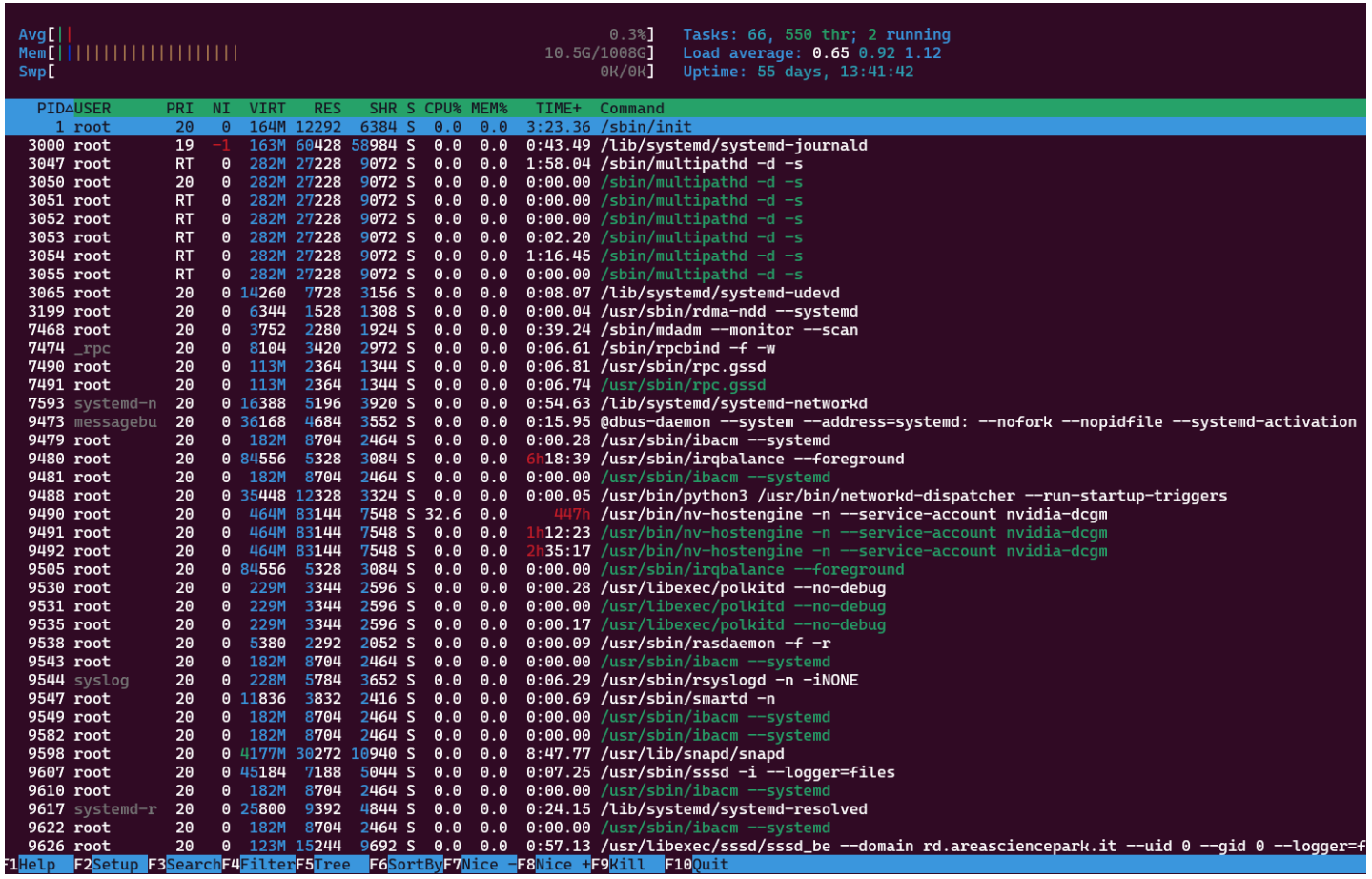
Figure 21. Example CSP MIG Configuration




Example of multiple independent GPU Compute workloads running in parallel using a MIG configuration on an A100 GPU with three GPU Instances and variable numbers of Compute Instances within each GPU Instance.

Figure 23. MIG Configuration with multiple independent GPU Compute workloads

# Htop on DGX


Avg[|| 0.3%] Tasks: 66, 550 thr; 2 running
Mem[|||||||||||||||||| 10.5G/1008G] Load average: 0.65 0.92 1.12
Swp[ 0K/0K] Uptime: 55 days, 13:41:42
PID△USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
1 root 20 0 164M 12292 6384 S 0.0 0.0 3:23.36 /sbin/init
3000 root 19 -1 163M 60428 58984 S 0.0 0.0 0:43.49 /lib/systemd/systemd-journald
3047 root RT 0 282M 27228 9072 S 0.0 0.0 1:58.04 /sbin/multipathd -d -s
3050 root 20 0 282M 27228 9072 S 0.0 0.0 0:00.00 /sbin/multipathd -d -s
3051 root RT 0 282M 27228 9072 S 0.0 0.0 0:00.00 /sbin/multipathd -d -s
3052 root RT 0 282M 27228 9072 S 0.0 0.0 0:00.00 /sbin/multipathd -d -s
3053 root RT 0 282M 27228 9072 S 0.0 0.0 0:02.20 /sbin/multipathd -d -s
3054 root RT 0 282M 27228 9072 S 0.0 0.0 1:16.45 /sbin/multipathd -d -s
3055 root RT 0 282M 27228 9072 S 0.0 0.0 0:00.00 /sbin/multipathd -d -s
3065 root 20 0 14260 7728 3156 S 0.0 0.0 0:08.07 /lib/systemd/systemd-udevd
3199 root 20 0 6344 1528 1308 S 0.0 0.0 0:00.04 /usr/sbin/rdma-ndd --systemd
7468 root 20 0 3752 2280 1924 S 0.0 0.0 0:39.24 /sbin/mdadm --monitor --scan
7474 _rpc 20 0 8104 3420 2972 S 0.0 0.0 0:06.61 /sbin/rpcbind -f -w
7490 root 20 0 113M 2364 1344 S 0.0 0.0 0:06.81 /usr/sbin/rpc.gssd
7491 root 20 0 113M 2364 1344 S 0.0 0.0 0:06.74 /usr/sbin/rpc.gssd
7593 systemd-n 20 0 16388 5196 3920 S 0.0 0.0 0:54.63 /lib/systemd/systemd-networkd
9473 messagebu 20 0 36168 4684 3552 S 0.0 0.0 0:15.95 @dbus-daemon --system --address=systemd: --nofork --nopidfile --systemd-activation
9479 root 20 0 182M 8704 2464 S 0.0 0.0 0:00.28 /usr/sbin/ibacm --systemd
9480 root 20 0 84556 5328 3084 S 0.0 0.0 6h18:39 /usr/sbin/irqbalance --foreground
9481 root 20 0 182M 8704 2464 S 0.0 0.0 0:00.00 /usr/sbin/ibacm --systemd
9488 root 20 0 35448 12328 3324 S 0.0 0.0 0:00.05 /usr/bin/python3 /usr/bin/networkd-dispatcher --run-startup-triggers
9490 root 20 0 464M 83144 7548 S 32.6 0.0 447h /usr/bin/nv-hostengine -n --service-account nvidia-dcgm
9491 root 20 0 464M 83144 7548 S 0.0 0.0 1h12:23 /usr/bin/nv-hostengine -n --service-account nvidia-dcgm
9492 root 20 0 464M 83144 7548 S 0.0 0.0 2h35:17 /usr/bin/nv-hostengine -n --service-account nvidia-dcgm
9505 root 20 0 84556 5328 3084 S 0.0 0.0 0:00.00 /usr/sbin/irqbalance --foreground
9530 root 20 0 229M 3344 2596 S 0.0 0.0 0:00.28 /usr/libexec/polkitd --no-debug
9531 root 20 0 229M 3344 2596 S 0.0 0.0 0:00.00 /usr/libexec/polkitd --no-debug
9535 root 20 0 229M 3344 2596 S 0.0 0.0 0:00.17 /usr/libexec/polkitd --no-debug
9538 root 20 0 5380 2292 2052 S 0.0 0.0 0:00.09 /usr/sbin/rasdaemon -f -r
9543 root 20 0 182M 8704 2464 S 0.0 0.0 0:00.00 /usr/sbin/ibacm --systemd
9544 syslog 20 0 228M 5784 3652 S 0.0 0.0 0:06.29 /usr/sbin/rsyslogd -n -iNONE
9547 root 20 0 11836 3832 2416 S 0.0 0.0 0:00.69 /usr/sbin/smartd -n
9549 root 20 0 182M 8704 2464 S 0.0 0.0 0:00.00 /usr/sbin/ibacm --systemd
9582 root 20 0 182M 8704 2464 S 0.0 0.0 0:00.00 /usr/sbin/ibacm --systemd
9598 root 20 0 4177M 30272 10940 S 0.0 0.0 8:47.77 /usr/lib/snapd/snapd
9607 root 20 0 45184 7188 5044 S 0.0 0.0 0:07.25 /usr/sbin/sssd -i --logger=files
9610 root 20 0 182M 8704 2464 S 0.0 0.0 0:00.00 /usr/sbin/ibacm --systemd
9617 systemd-r 20 0 25800 9392 4844 S 0.0 0.0 0:24.15 /lib/systemd/systemd-resolved
9622 root 20 0 182M 8704 2464 S 0.0 0.0 0:00.00 /usr/sbin/ibacm --systemd
9626 root 20 0 123M 15244 9692 S 0.0 0.0 0:57.13 /usr/libexec/sssd/sssd_be --domain rd.areasciencepark.it --uid 0 --gid 0 --logger=f
F1Help F2Setup F3SearchF4FilterF5Tree F6SortByF7Nice -F8Nice +F9Kill F10Quit