Big Data and Machine Learning on Virtual Graphics Processing Units

Adrian Sabou
Computer Science Department
Technical University of Cluj-Napoca
adrian.sabou@cs.utcluj.ro

CLOUDUT Project

- **Titlu:** Cloud Cercetare UTCN – CLOUDUT (http://cloudut.utcluj.ro)
- **MySMIS ID:** 124493
- **Contract nr:** 235/ 21.04.2020

UNIVERSITATEA TEHNICĂ DIN CLUJ-NAPOCA

- **Tip Proiect:** Program Operațional Competitivitate 2014-2020 (POC)
- **Axa prioritara 1**: Cercetare, dezvoltare tehnologică și inovare (CDI) în sprijinul competitivității economice și dezvoltării afacerilor
- Acțiunea 1.1.2: Dezvoltarea unor rețele de centre CD, coordonate la nivel național și racordate la rețele europene și internaționale de profil și asigurarea accesului cercetătorilor la publicații științifice și baze de date europene și internaționale
- **Finanțare:** Fonduri Europene pentru Dezvoltare Regională, Valoarea totală: 4.955.000 RON, din care 4.950.000 RON din fonduri Europene.

Welcome to the Era of AI

Deep Learning is Everywhere [7]

Cloud Infrastructure

Minimum requirements:

- 20 x 16 core CPU processors, 2GHz, support for VMWare virtualization and hyperthreading
- 2 AI processing servers. Each server is equipped with 2 x 20 core processors, 512GB RAM, 1TB storage, 2 x GPUs with 640 tensor cores, 32 GB, support for virtualization
- 16GB RAM per CPU core, storing capacity 70TB, RAID 5
- 25Gbps internal and external connectivity

Cloud Infrastructure

- 2 x Dell Poweredge R740 servers
  - (each with) 2 x NVIDIA V100 GPUs, 32 GB

Dell Poweredge R740 [6]

Optimized for workload acceleration

Optimized for AI and GPGPU Computing

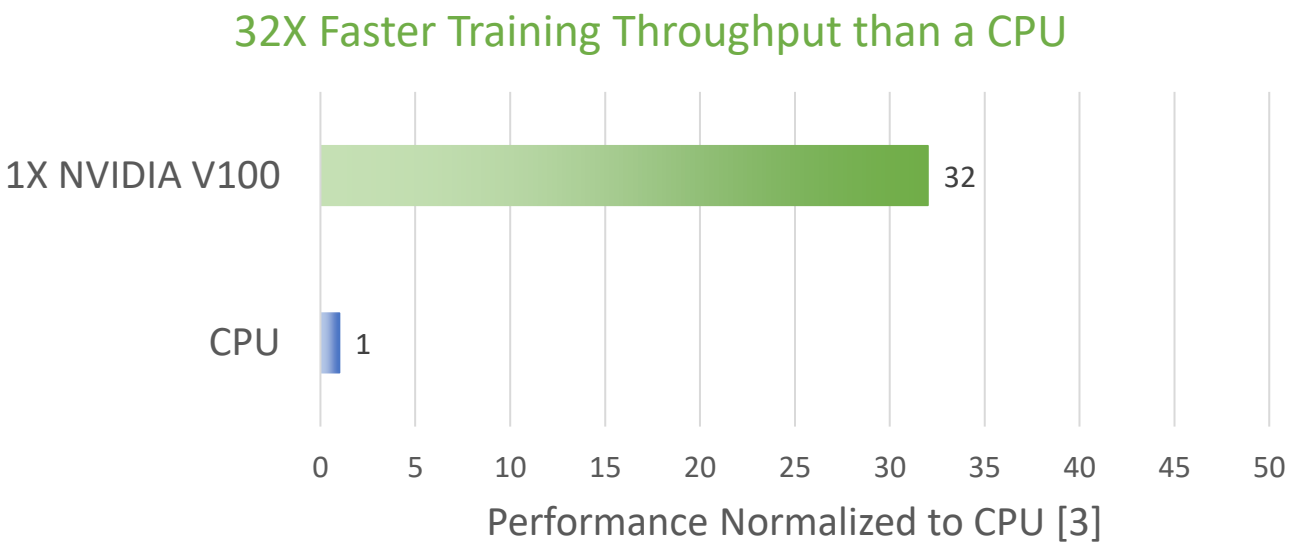
NVIDIA V100 (PCIe) [1]

NVIDIA V100 – Specifications

Specification	V100 PCIe	V100 SXM2	V100S PCIe
GPU Architecture	NVIDIA Volta		
NVIDIA Tensor Cores	640		
NVIDIA CUDA Cores	5120		
Double-precision Performance	7 TFLOPS	7.8 TFLOPS	8.2 TFLOPS
Single-precision Performance	14 TFLOPS	17.7 TFLOPS	16.4 TFLOPS
Tensor Performance	112 TFLOPS	125 TFLOPS	130 TFLOPS
GPU Memory	32 GB / 16 GB HBM2		32 GB HBM2
Memory Bandwidth	900 GB/sec		1134 GB/sec
Compute APIs	CUDA, DirectCompute, OpenCL, OpenACC		

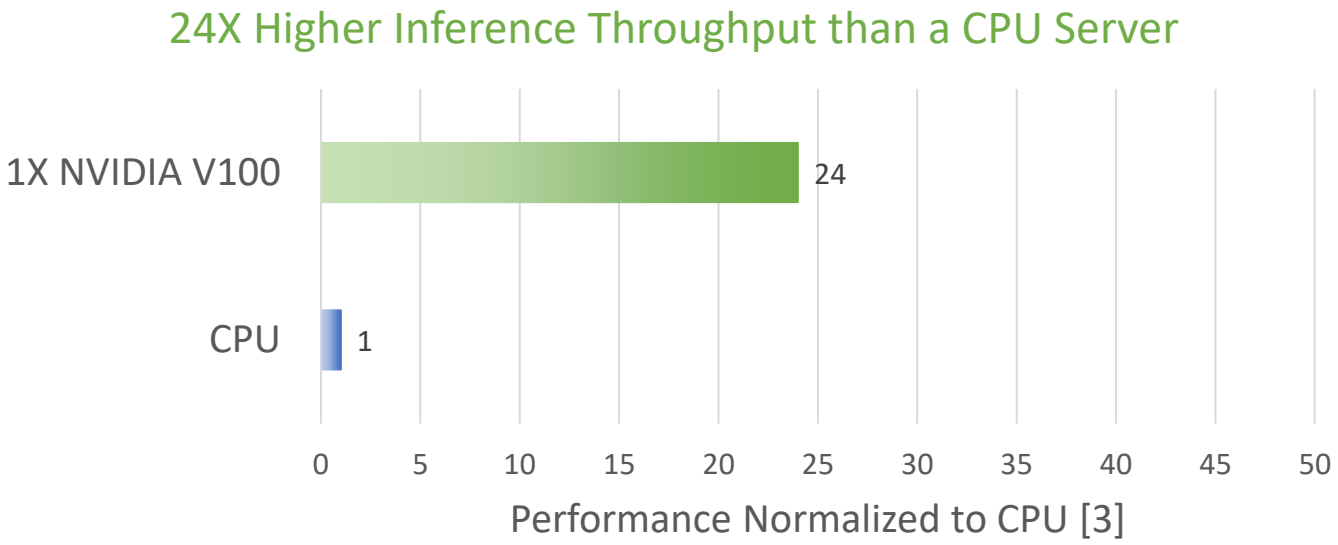
NVIDIA V100 Specifications [3]

NVIDIA V100 – Acceleration



ResNet-50 training, dataset: ImageNet2012, BS=256 | NVIDIA V100 comparison: NVIDIA DGX-2™ server, 1x V100 SXM3-32GB, MXNet 1.5.1, container=19.11-py3, mixed precision, throughput: 1,525 images/sec | Intel comparison: Supermicro SYS-1029GQ-TRT, 1 socket Intel Gold 6240@2GHz/3.9Hz Turbo, Tensorflow 0.18, FP32 (only precision available), throughput: 48 images/sec [3]

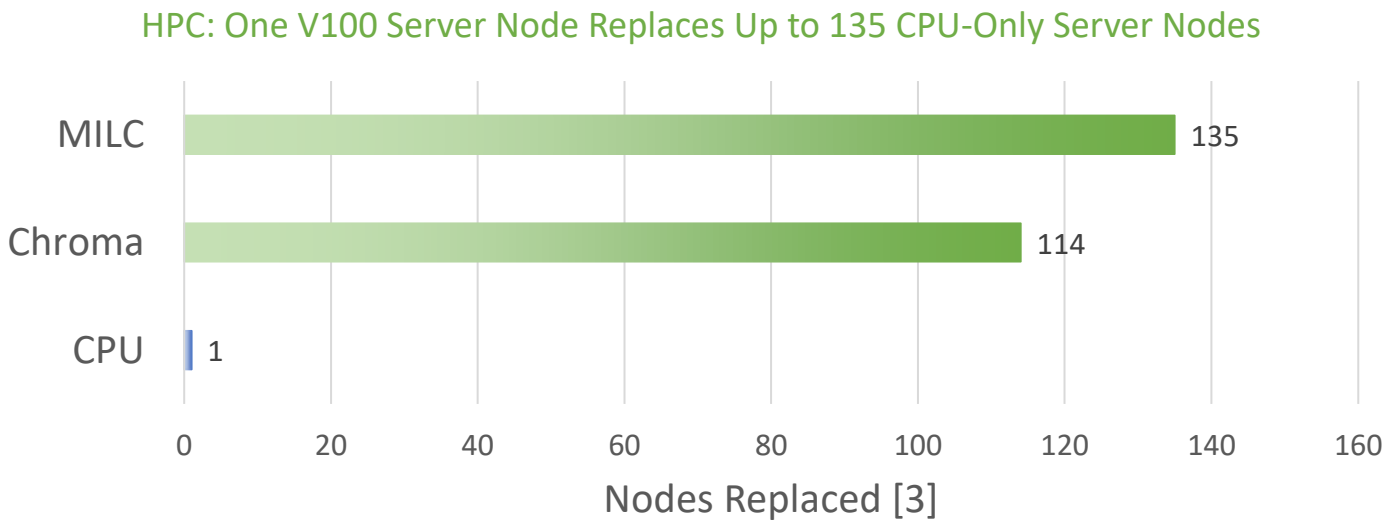
NVIDIA V100 – Acceleration



BERT Base fine-tuning inference, dataset: SQuADv1.1, BS=1, sequence length=128 | NVIDIA V100 comparison: Supermicro SYS4029GP-TRT, 1x V100-PCIE-16GB, pre-release container, mixed precision, NVIDIA TensorRT™ 6.0, throughput: 557 sentences/sec | Intel comparison: 1 socket Intel Gold 6240@2.6GHz/3.9Hz Turbo, FP32 (only precision available), OpenVINO MKL-DNN v0.18, throughput: 23.5 sentences/sec [3]

More details regarding Deep Learning Performance -
https://developer.nvidia.com/deep-learning-performance-training-inference

NVIDIA V100 – Acceleration

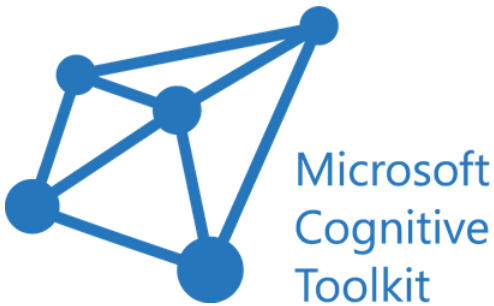


16x V100-SXM2-32GB in NVIDIA HGX-2™ | Application (dataset): MILC (APEX Medium) and Chroma (szscl21_24_128) | CPU server: dual-socket Intel Xeon Platinum 8280 (Cascade Lake) [3]

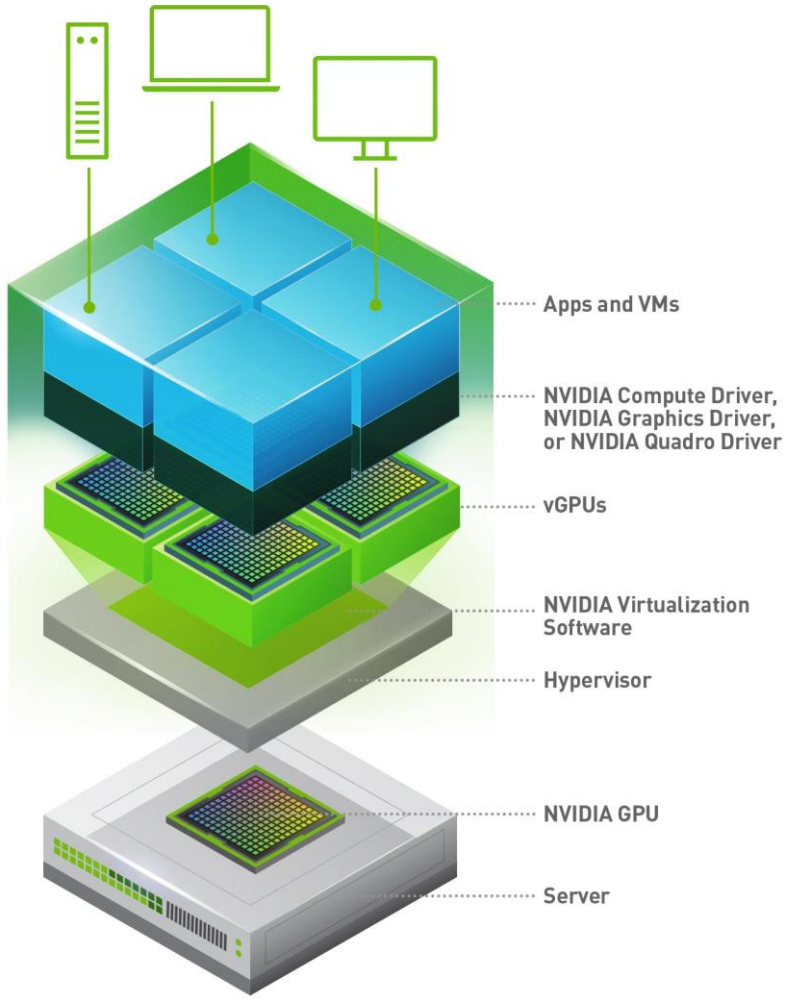
More details regarding HPC Performance -
https://developer.nvidia.com/hpc-application-performance

NVIDIA V100 – DL and HPC

- Compatible with the most popular Deep Learning frameworks and with 600+ HPC applications [3]



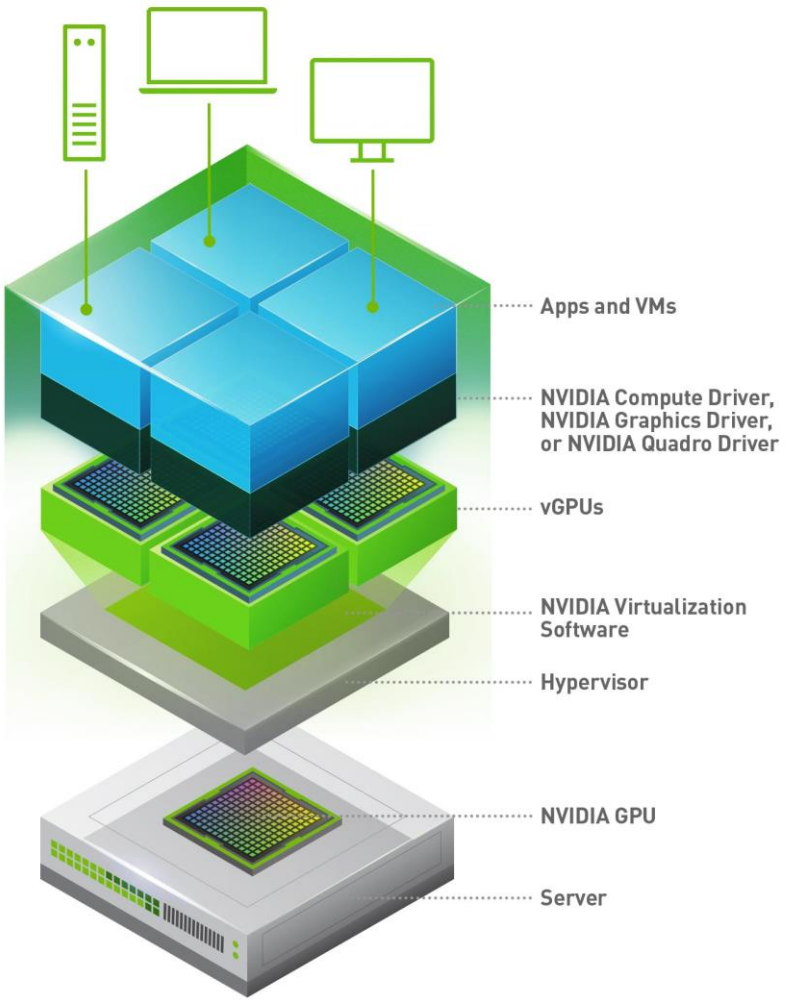
NVIDIA V100 – Virtualization



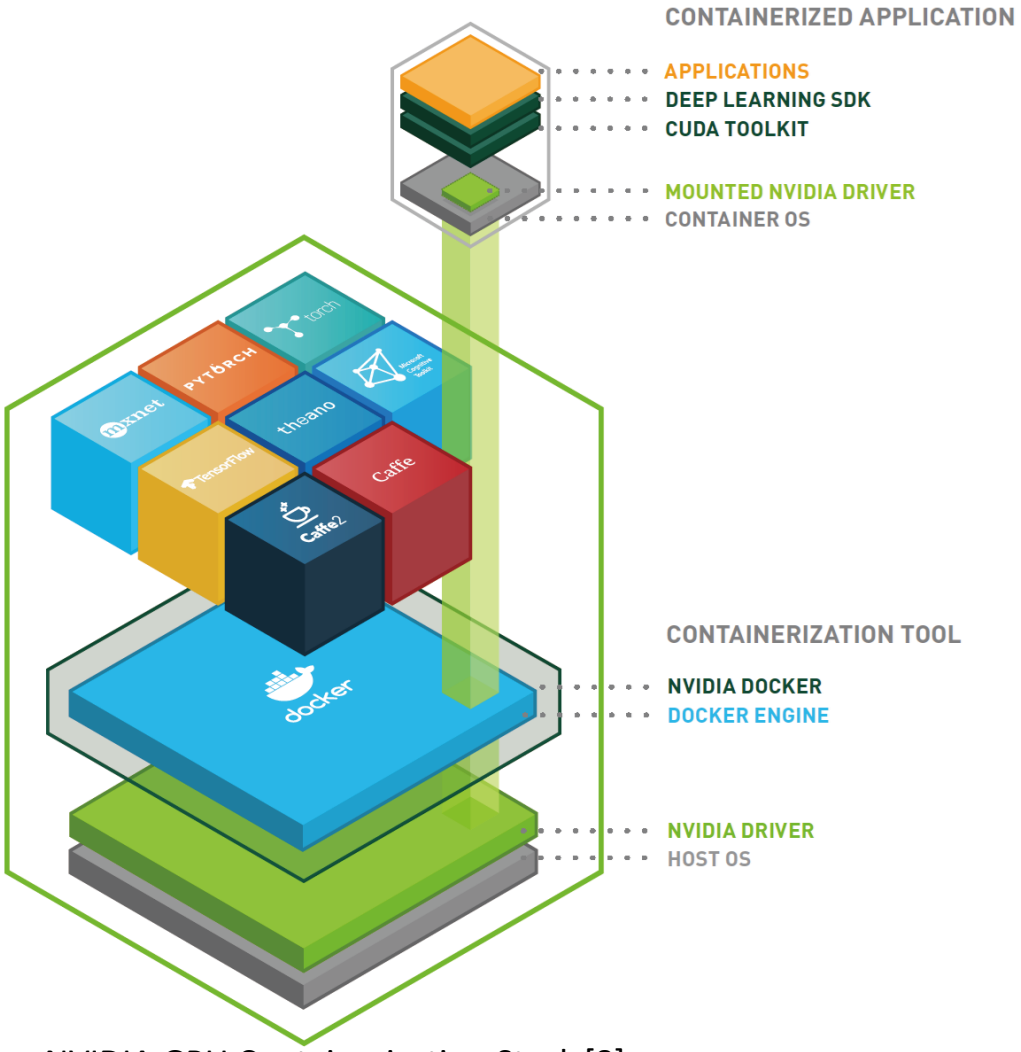
NVIDIA GPU Virtualization Stack [4]

- GPU Virtualization
- NVIDIA Virtual GPU Software Solutions
  - NVIDIA GRID (Virtual PC (GRID vPC) and Virtual Apps (GRID vApps)) (VDI)
  - NVIDIA Quadro Virtual Workstation (Graphics Processing)
  - NVIDIA Virtual Compute Server (Virtualization for complex AI and Compute tasks)

Virtualization vs. Containerization



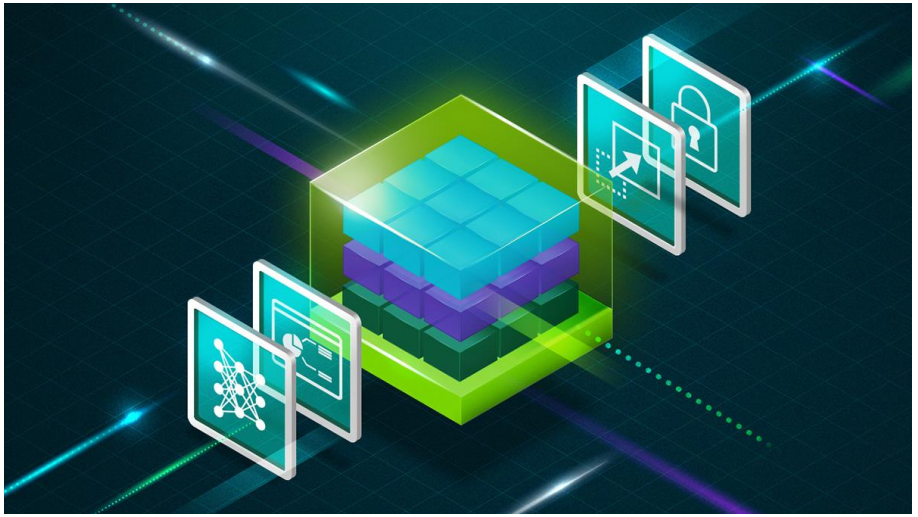
NVIDIA GPU Virtualization Stack [4]



NVIDIA GPU Containerization Stack [8]

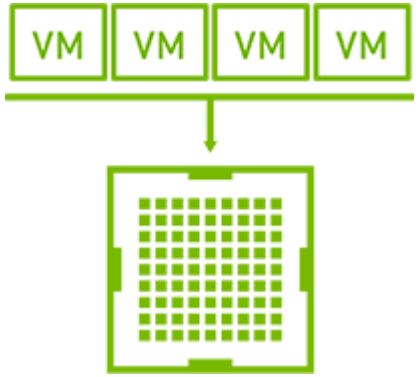
NVIDIA Virtual Compute Server

- Software solution to virtualize computation for AI, Deep Learning, and Data Science
  - Artificial Intelligence
  - Deep Learning
  - Data science
  - High Performance Computing



NVIDIA Virtual Compute Server [5]

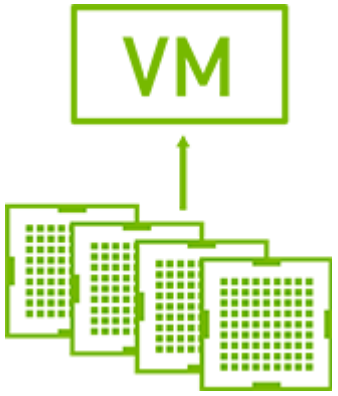
NVIDIA Virtual Compute Server



GPU Sharing [5]

- GPU Sharing
  - Fractional GPU sharing with NVIDIA vGPU technology
  - Multiple VMs sharing the same GPU
  - Maximizing utilization

- GPU Aggregation
  - A VM can access more than one GPU
  - Multi GPU Computing – GPUs aren't directly interconnected
  - Peer-to-peer Computing – GPUs are interconnected through NVLink for higher bandwidth



GPU Aggregation [5]

NVIDIA Virtual Compute Server

Management and Monitoring [5]

- Management and Monitoring
  - Support for app-, guest-, and host-level monitoring
  - Live VM migration
  - Suspend, resume, thresholds

- Multi-Instance GPU
  - Specific for NVIDIA A100 Tensor Core GPU
  - Partitioning the cores into up to seven instances
  - Provisioning a VM on an instance

Multi-Instance GPU [5]

Conclusions

- GPUs – enormous potential for accelerating AI and GPGPU Computation
- GPU Virtualization – Maximum flexibility
- NVIDIA vGPU Software Solutions
  - NVIDIA Virtual Compute Server – AI and GPGPU Compute Tasks
  - GPU Sharing – Maximize utilization
  - GPU Aggregation – Flexible Multi GPU Computing

References

[1] NVIDIA, "NVIDIA V100 Tensor Core GPU," 2020. [Online]. Available: https://www.nvidia.com/en-us/data-center/v100/.

[2] NVIDIA, "NVIDIA A100 Tensor Core GPU," 2020. [Online]. Available: https://www.nvidia.com/en-us/data-center/a100/.

[3] NVIDIA, "NVIDIA V100 Tensor Core GPU Datasheet," 2020. [Online]. Available: https://images.nvidia.com/content/ technologies/volta/pdf/volta-v100-datasheet-update-us-1165301-r5.pdf.

[4] NVIDIA, "NVIDIA Virtual GPU Technology," 2020. [Online]. Available: https://www.nvidia.com/en-us/data-center/virtual-gpu-technology/.

[5] NVIDIA, "NVIDIA Virtual Compute Server," 2020. [Online]. Available: https://www.nvidia.com/en-us/data-center/virtual-compute-server/.

[6] Dell, "PowerEdge R740 Rack Server," 2020. [Online]. Available: https://www.dell.com/en-us/work/shop/povw/poweredge-r740.

[7] NVIDIA, "NVIDIA Deep Learning Institute," 2020. [Online]. Available: https://www.nvidia.com/en-us/deep-learning-ai/education/.

[8] NVIDIA, "NGC CONTAINER - User Guide," 2020. [Online]. Available: https://docs.nvidia.com/ngc/pdf/NGC-User-Guide.pdf.

Thank you for your attention!

Adrian Sabou
Computer Science Department
Technical University of Cluj-Napoca
adrian.sabou@cs.utcluj.ro