SUSE

Accelerate Your AI Cloud Infrastructure

A Virtualization Perspective

Liang Yan – SUSE Labs

Liang Yan

Sr. Software Engineer

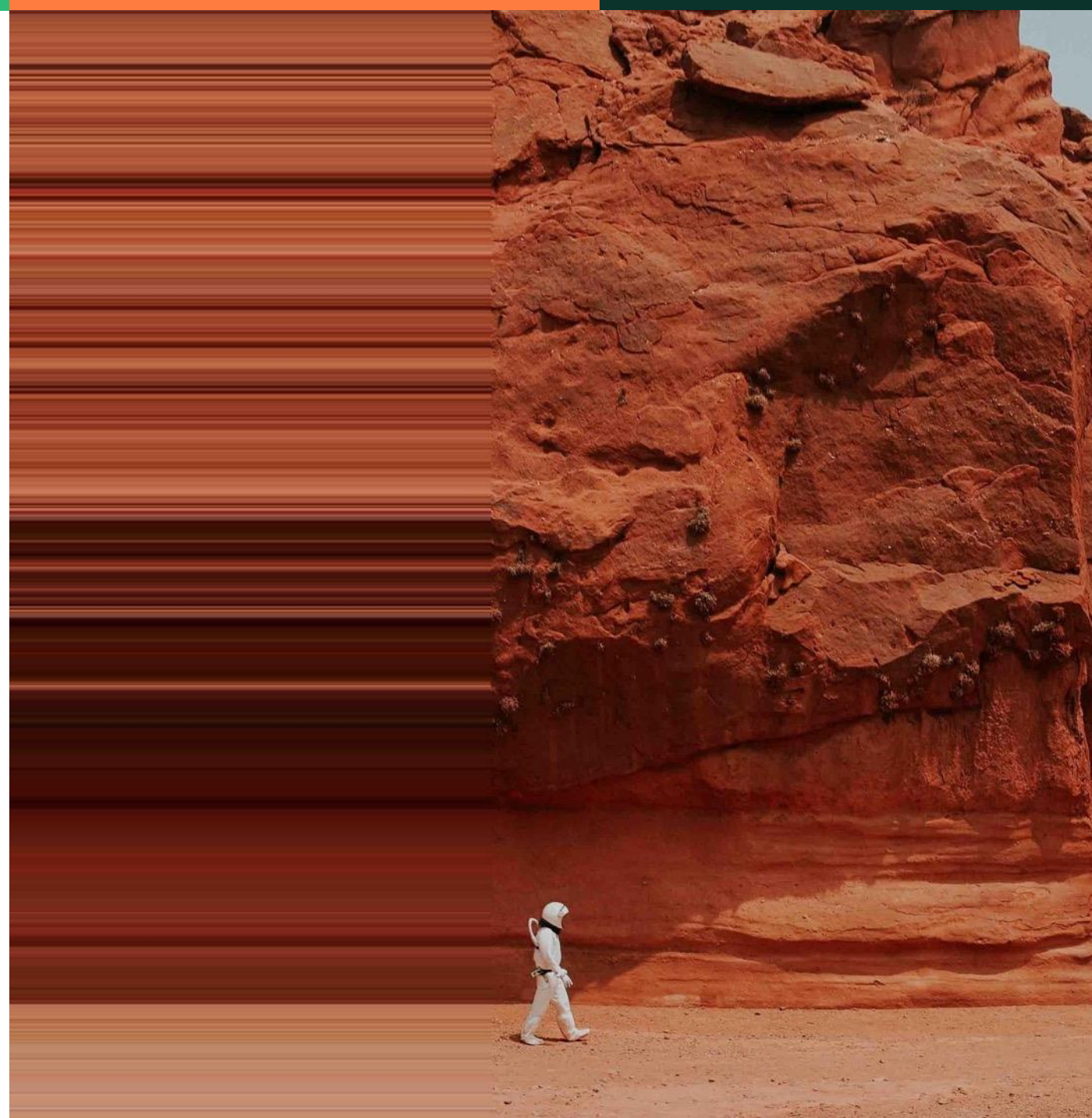
Focus on GPU and ARM64 Virtualization

Work closely with vendors on feature development and performance optimization, deliver customized solutions to customers.

POC Research on AI/ML accelerator virtualization and hybrid-LightVMs

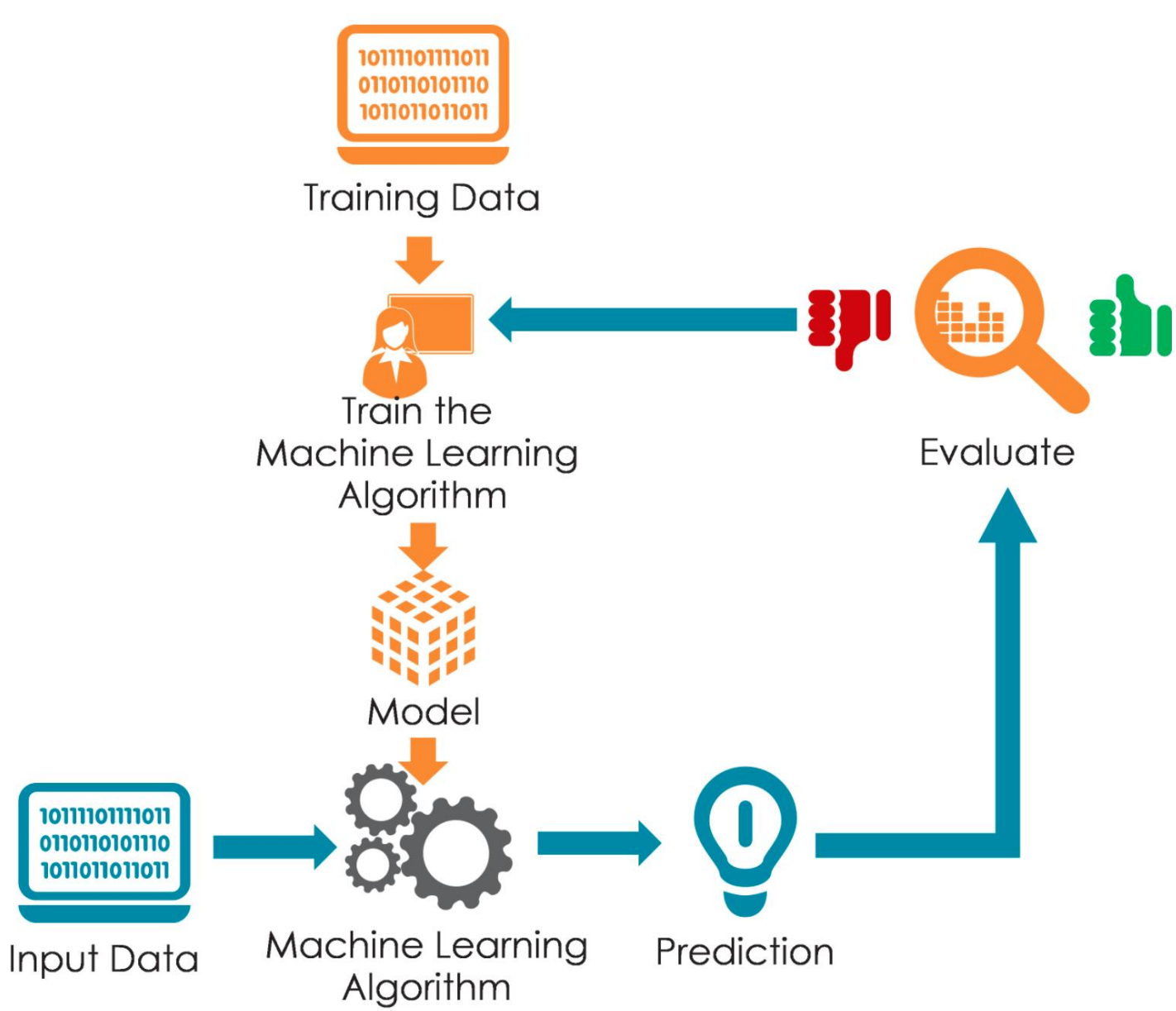
Outline

- Background
  - AI Cloud
  - Hardware Accelerator
- NVIDIA® GPU Virtualization
- SUSE® GPU Virtualization
- Futures



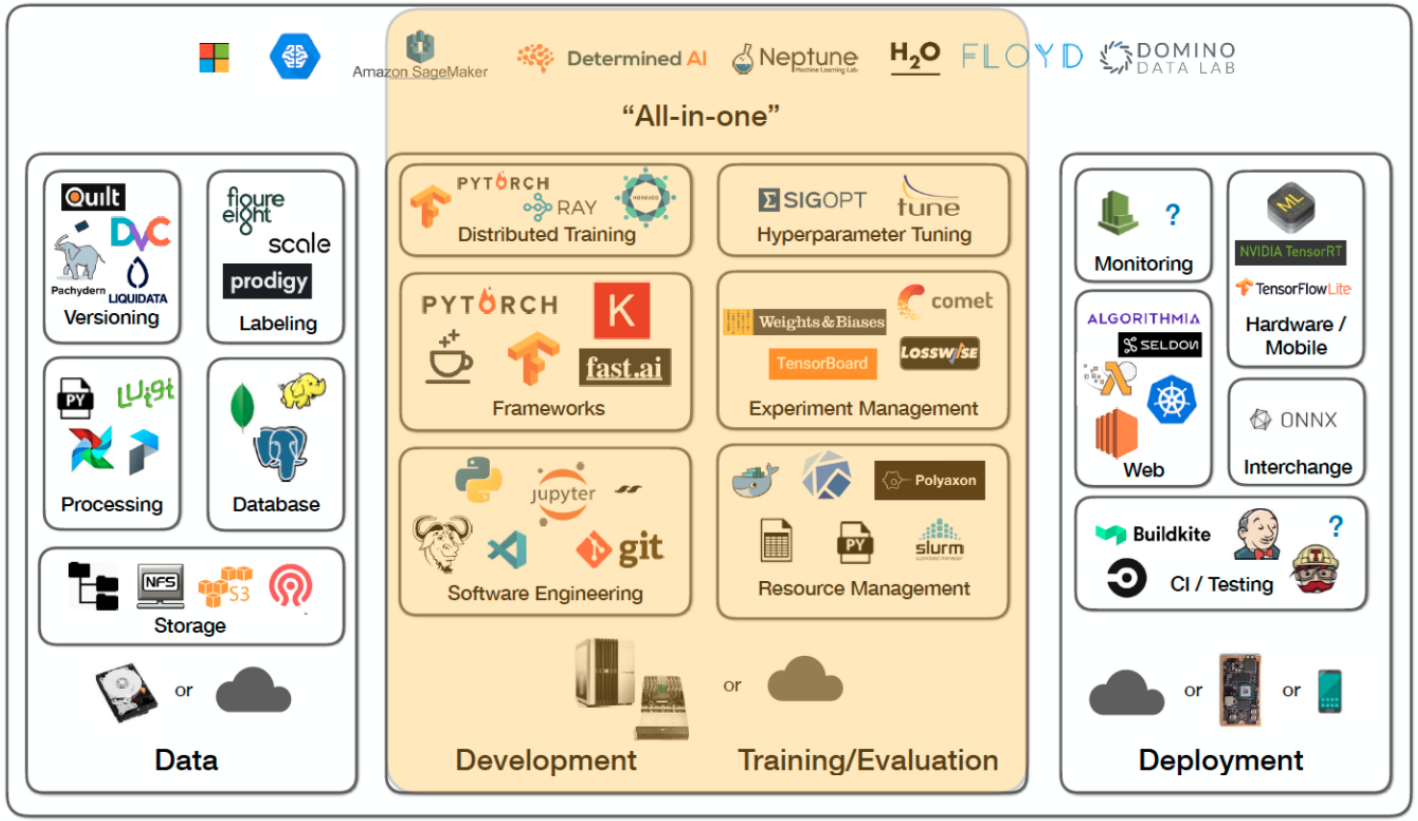
Background

Workflow to a Data Scientist



https://www.7wdata.be/big-data/building-the-machine-learning-infrastructure/

Workflow in your AI Cloud Infrastructure



https://jameskle.com/writes/deep-learning-infrastructure-tooling

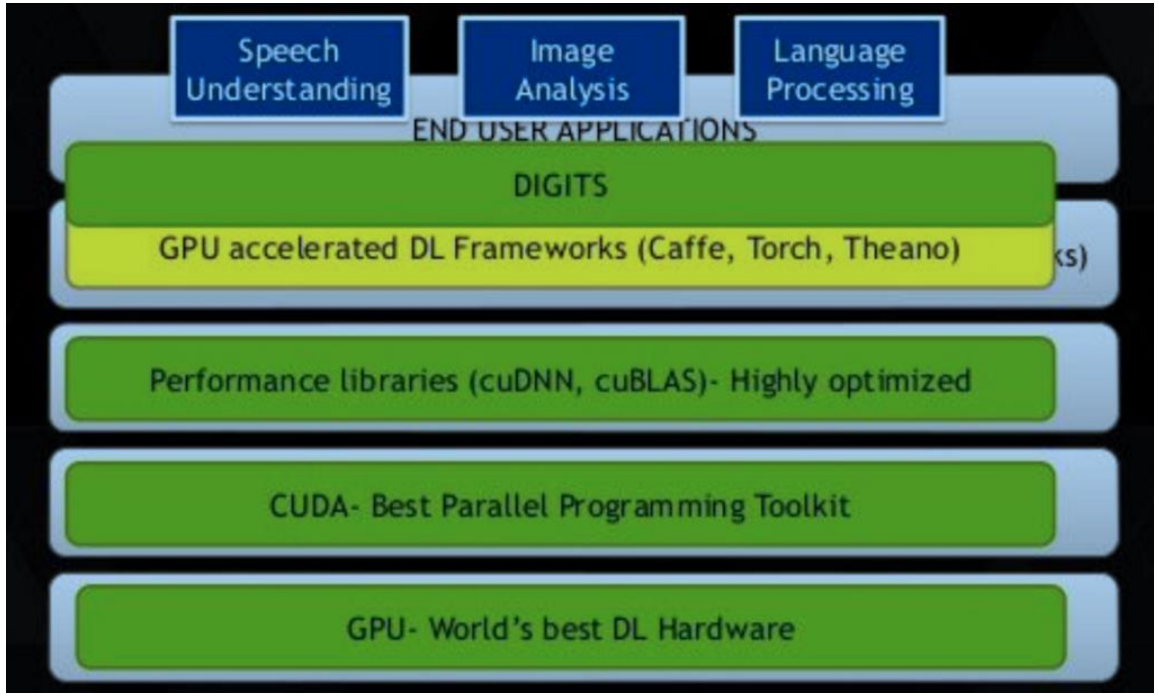
Hardware Accelerator Landscape

	GPU	FPGA	ASIC
Vendors	NVIDIA®, AMD®, INTEL®	Xilinx®, INTEL® (Altera)	Google TPU, AI Chips
Development Frameworks	OpenCL, CUDA	OpenCL	OpenCL, TensorFlow
Machine Learning Lifecycle	Training	Inference	Inference

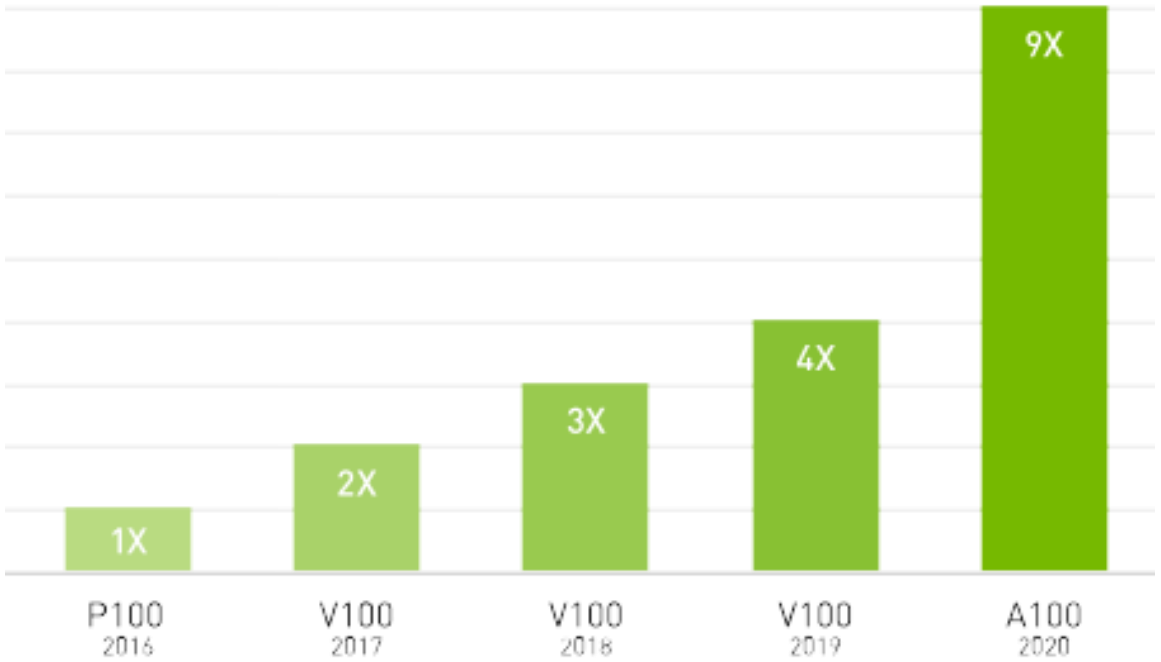
FPGA: Field-Programmable Gate Array
ASIC: Application-Specific Integrated Circuit
TPU: Tensor Processing Unit

NVIDIA® GPU Virtualization

Why Choose NVIDIA



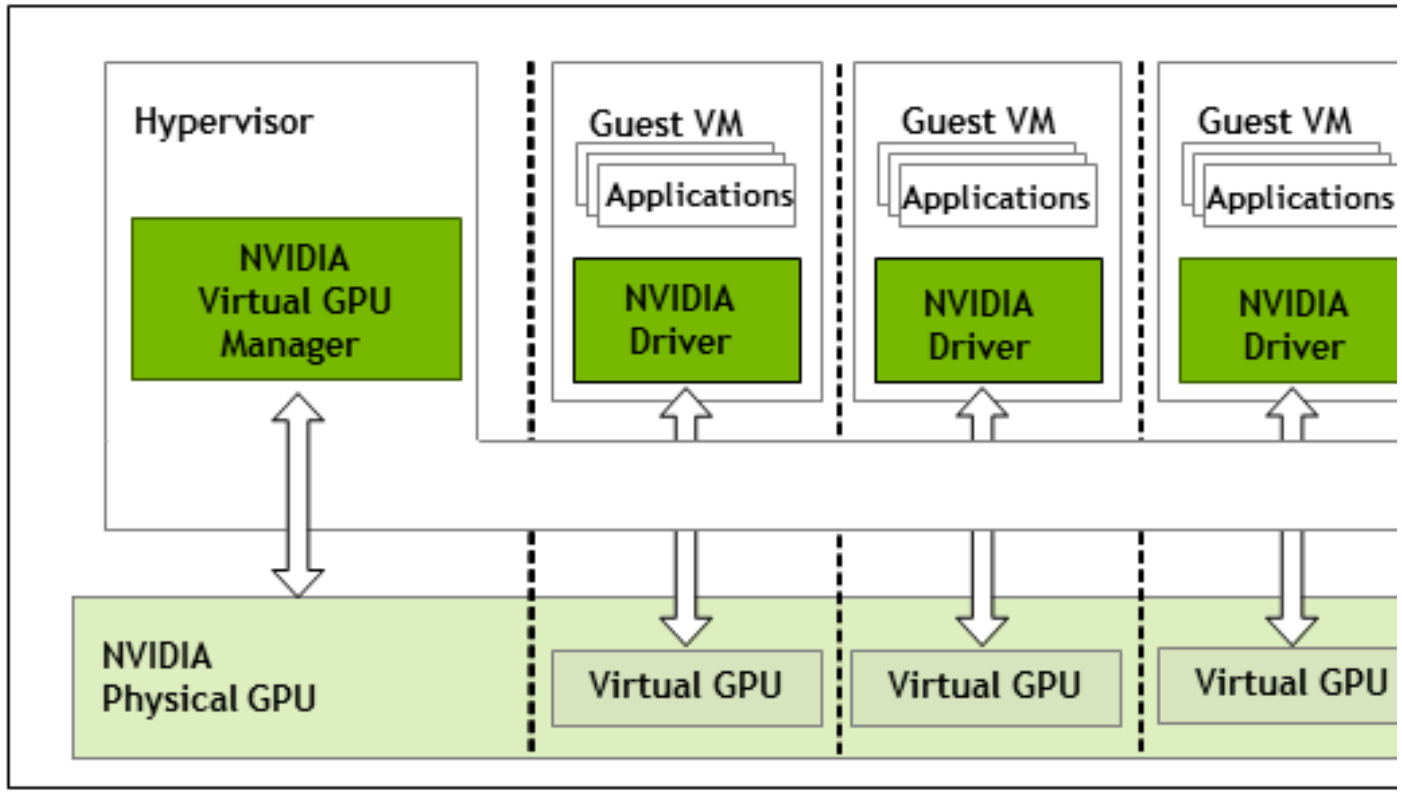
Software Ecosystem



Hardware Performance

https://becominghuman.ai/nvidia-and-the-gpu-contribution-to-the-ai-world-of-self-driving-cars-1f00e3212508
http://www.nvidia.com/object/grid-certified-servers.html

NVIDIA® GPU Virtualization



- Scalability
  - Multi-vGPU
- Security
  - isolate
  - Avoid system failure
- Flexibility
  - Migration/Live Migration
- Monitoring

SUSE® GPU Virtualization

SUSE Reference Platform: Tests and Results

- Test Setup
  - Host: SUSE Linux Enterprise Server 15 SP2
  - Guests: SUSE Linux Enterprise Server 15 SP2, 15SP1, Windows Server 2019(4 vCPU, 24G)
  - Hardware: HPE ProLiant DL380 Gen9(E5-2650 v3 x 2, 128G), NVIDIA® V100 (PCIe 16G)
  - vGPU: 450.74
  - Benchmarks: LAMMPS, TensorRT, Specperfview
- Functional Tests:
  - Driver
  - CUDA
  - 3D Graphics
  - Remote Display
  - Max vGPUs support
- Performance Tests:
  - vGPU vs Passthrough
  - vGPU across different guest VMs
  - vGPU with different memory configurations
  - vGPU scalability

Graphic Performance Results

SPECviewperf 13	creo-02	energy-02	maya-05	medical-02	sw-04
vGPU 16Q	0.263	1.473	0.223	0.903	0.377
Passthrough	1.000	1.000	1.000	1.000	1.000

Unoptimized

SPECviewperf 13	creo-02	energy-02	maya-05	medical-02	sw-04
vGPU 16Q	0.943	1.619	1.192	1.915	1.127
Passthrough	1.000	1.000	1.000	1.000	1.000

Optimized

- For the experiment, we take first run as warm up, then run three times and take the mean value as result, reboot during each run. For consistency purposes, we run twice for each experiment, the difference is minimal.
- For the optimization: We disabled ftl, ecc from vGPU driver level, we enabled display and manage from libvirt level.
- Data are normalized by passthrough result
- Results are only used as reference.

Compute Performance Results

- For the experiment, we take first run as warm up, then run three times and take the mean value as result, reboot during each run. For consistency purposes, we run twice for each experiment, the difference is minimal.
- Data are normalized by passthrough result
- Results are only used as reference.

	fp32			fp16			int8		
TensorRT 6.0	times	host walltime	99% percentile time	times	host walltime	99% percentile time	times	host walltime	99% percentile time
16C	1.005	1.060	1.070	1.010	1.031	1.022	1.041	1.069	1.039
16Q	1.005	1.035	1.008	1.015	1.038	1.021	1.044	1.085	1.038
4C	1.017	1.051	1.024	1.001	1.024	1.005	1.001	1.036	1.006
4Q	1.005	1.039	1.005	1.017	1.044	1.019	1.001	1.033	0.997
Passthrough	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4C-1	2.557	2.607	2.975	1.868	1.886	2.038	1.002	1.029	1.002
4C-2	1.729	1.762	1.907	1.797	1.821	1.931	1.830	1.879	1.981
4C-3	1.034	1.067	1.269	1.075	1.091	1.138	1.782	1.832	1.943
4C-4	1.747	1.461	2.013	1.178	1.192	1.277	1.198	1.222	1.367

Conclusions

- No major discernible difference on compute performance between vGPU and pass-through, vGPU even has better graphic performance.
- Similar results were achieved across different SUSE Linux Enterprise guest environments (15 SP2 vs 15 SP1)
- During lower workload, vGPU memory size showed no effect on performance (V100-16C vs V100-4C)
- For Compute workload, vGPU model types showed no major differences (V100-16C vs V100-16Q)
- Scalability impacts performance, but still better than expectations(V100-16C vs 4xV100-4C)

Feature Checklist - Review

Remote Display	Graphic Performance	CUDA installation	AI Framework installation
Compute Performance	VM Snapshots	Live Migration	A100 support
Secure boot for vGPU			

Futures

Further Exploration

SUSE Exploration:

- GPU passthrough for ARM64
- vGPU plugin in KubeVirt (Kubernetes scenario)
- vGPU plugin in RUST-VMM

AMD GPU:

- Radeon Instinct GPU + MxGPU GIM(GPU-IOV Module)

ARM GPU

- Mali GPU virtualization for in-vehicle
- ARM platform for GPU

Intel

- Dedicated GPU
- FPGA

Thank you

For more information, contact SUSE at:

+1 800 796 3700 (U.S./Canada)

+49 (0)911-740 53-0 (Worldwide)

Maxfeldstrasse 5

90409 Nuremberg

www.suse.com