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

Workflow in your AI Cloud Infrastructure

https://jameskle.com/writes/deep-learning-infrastructure-tooling
# Hardware Accelerator Landscape

<table>
<thead>
<tr>
<th>Vendors</th>
<th>GPU</th>
<th>FPGA</th>
<th>ASIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA®, AMD®, INTEL®</td>
<td>NVIDIA®, AMD®, INTEL®</td>
<td>Xilinx®, INTEL® (Altera)</td>
<td>Google TPU, AI Chips</td>
</tr>
<tr>
<td>Development Frameworks</td>
<td>OpenCL, CUDA</td>
<td>OpenCL</td>
<td>OpenCL, TensorFlow</td>
</tr>
<tr>
<td>Machine Learning Lifecycle</td>
<td>Training</td>
<td>Inference</td>
<td>Inference</td>
</tr>
</tbody>
</table>

FPGA: Field-Programmable Gate Array  
ASIC: Application-Specific Integrated Circuit  
TPU: Tensor Processing Unit
NVIDIA® GPU Virtualization
Why Choose NVIDIA

Software Ecosystem

- Speech Understanding
- Image Analysis
- Language Processing
- GPU accelerated DL Frameworks (Caffe, Torch, Theano)
- Performance libraries (cuDNN, cuBLAS) - Highly optimized
- CUDA - Best Parallel Programming Toolkit
- GPU - World’s best DL Hardware

Hardware Performance

https://becominghuman.ai/nvidia-and-the-gpu-contribution-to-the-ai-world-of-self-driving-cars-1f00e3212508
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

<table>
<thead>
<tr>
<th>SPECviewperf 13</th>
<th>creo-02</th>
<th>energy-02</th>
<th>maya-05</th>
<th>medical-02</th>
<th>sw-04</th>
</tr>
</thead>
<tbody>
<tr>
<td>vGPU 16Q</td>
<td>0.263</td>
<td>1.473</td>
<td>0.223</td>
<td>0.903</td>
<td>0.377</td>
</tr>
<tr>
<td>Passthrough</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SPECviewperf 13</th>
<th>creo-02</th>
<th>energy-02</th>
<th>maya-05</th>
<th>medical-02</th>
<th>sw-04</th>
</tr>
</thead>
<tbody>
<tr>
<td>vGPU 16Q</td>
<td>0.943</td>
<td>1.619</td>
<td>1.192</td>
<td>1.915</td>
<td>1.127</td>
</tr>
<tr>
<td>Passthrough</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

- 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 ftt, 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.

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