Boosting Performance and Earnings of Cloud Computing Deployments with rCUDA

Federico Silla
Universitat Politècnica de València
Spain
1. Using CUDA GPUs from virtual machines
2. rCUDA: GPU virtualization
3. Performance of rCUDA with one virtual machine
4. Performance of rCUDA with several virtual machines
5. Conclusions
1. Using CUDA GPUs from virtual machines
2. rCUDA: GPU virtualization
3. Performance of rCUDA with one virtual machine
4. Performance of rCUDA with several virtual machines
5. Conclusions
How to access the GPU in the native domain from the inside of a virtual machine?
Using CUDA GPUs from virtual machines

- The **PCI passthrough technique** can be used to assign the GPU to a virtual machine.
- However … the GPU is assigned in an **exclusive** way.
- Concurrent usage of the GPU is **not possible**.

**Computer hosting several KVM virtual machines**

- KVM Guest Linux 1
- KVM Guest Linux 2
- KVM Guest Linux 3
- KVM Guest Linux n

**Diagram elements:**
- vETH
- SW BRIDGE
- Gb ETH
- STOP
- GO
- PCI PT
- KVM Host Linux
- Host HW
Using CUDA GPUs from virtual machines

• ... the amount of virtual machines using CUDA acceleration cannot be larger than the amount of GPUs present in the host

virtual machines ≤ GPUs
GPU virtualization allows as many virtual machines as required to share the GPU in the host.
1. Using CUDA GPUs from virtual machines
2. rCUDA: GPU virtualization for CUDA
3. Performance of rCUDA with one virtual machine
4. Performance of rCUDA with several virtual machines
5. Conclusions
rCUDA … CUDA … they sound similar
Basic behavior of CUDA
Basics of GPU computing

Application
CUDA libraries
A software technology that enables a more flexible use of GPUs in computing facilities

No GPU

rCUDA is a development by Universitat Politècnica de València, Spain
Basics of rCUDA

rCUDA is a development by Universitat Politècnica de València, Spain
Basics of rCUDA

rCUDA is a development by Universitat Politècnica de València, Spain
rCUDA GPU virtualization envision

- rCUDA allows a new vision of a GPU deployment, moving from the usual cluster configuration:

  ![Physical configuration diagram](image)

  to the following one:

  ![Logical configuration diagram](image)
Performance of applications using rCUDA

- Several applications executed with CUDA and rCUDA
  - K20 GPU and FDR InfiniBand
  - K40 GPU and EDR InfiniBand

Lower is better
Performance of applications using rCUDA

EDR InfiniBand and P100 GPU

BarraCUDA

CUDA-MEME

Lower is better

Lower is better
Why the good performance of rCUDA?

The low overhead of applications using rCUDA is due to:

- Data copies with rCUDA attaining higher bandwidth to the remote GPU than CUDA does to the local GPU
- Some internal synchronization mechanisms faster in rCUDA than in CUDA
- … a very careful implementation of the rCUDA framework …

“Ideas Are Easy, Implementation Is Hard”

Guy Kawasaki, marketing specialist and Silicon Valley venture capitalist
Example of performance with P2P copies

CUDA model

rCUDA model

rCUDA scenario 1

rCUDA scenario 2

rCUDA provides the same semantics as CUDA
Example of performance with P2P copies

rCUDA scenario 2

Bandwidth (MB/s)

Copy Size (MB)

Higher is better

- K20 CUDA
- K20 FDR rCUDA
- K40 CUDA
- K40 EDR rCUDA

GPU Technology Conference 2017
1. Using CUDA GPUs from virtual machines
2. rCUDA: GPU virtualization for CUDA
3. **Performance of rCUDA with one virtual machine**
4. Performance of rCUDA with several virtual machines
5. Conclusions
Using rCUDA to access the GPU

• In clusters where InfiniBand is not available, the rCUDA server may be placed in the native domain and the rCUDA client would be placed inside the VMs.

• The virtual network provided by the hypervisor would be used to exchange data between the rCUDA clients and the rCUDA server.

• This configuration allows the use of more than one GPU at the host.
Using rCUDA to access the GPU

- If InfiniBand is available, the rCUDA server can be placed in another node.
- Several GPUs can be provided to the VMs, either in a single remote node or in several remote nodes.
Application performance with KVM

FDR InfiniBand + K20 !!

- LAMMPS
- CUDA-MEME

- CUDA
- CUDA VM-PT
- rCUDA non-VM
- rCUDA VM IB
- rCUDA VM Local

- CUDASW++
- GPU-BLAST
Outline

1. Using CUDA GPUs from virtual machines
2. rCUDA: GPU virtualization for CUDA
3. Performance of rCUDA with one virtual machine
4. **Performance of rCUDA with several virtual machines**
5. Conclusions
Let’s use a computer with two GPUs and four virtual machines:

- Two virtual machines use one GPU each (PCI passthrough)
- Two virtual machines must run applications on CPU
With rCUDA, the four virtual machines can share both GPUs. The two GPUs can be either in the same host or in other computer.
Performance comparison

- Each of the 4 virtual machines execute as many instances as possible of one of the 4 following applications:
  - LAMMPS (red color in the plot below)
  - NAMD (green)
  - GPU-Blast (blue)
  - Fluidsim (yellow)

- For each experiment, applications are shifted across virtual machines

Sharing GPUs among applications increases the overall amount of executed jobs
1. Using CUDA GPUs from virtual machines
2. rCUDA: GPU virtualization for CUDA
3. Performance of rCUDA with one virtual machine
4. Performance of rCUDA with several virtual machines
5. Conclusions
Conclusions

- rCUDA allows GPUs to be shared among several virtual machines
- Applications do not need to be modified in order to use rCUDA
- Performance with rCUDA when GPUs are not shared is not significantly reduced
- Overall performance is increased when GPUs are shared among virtual machines
Get a free copy of rCUDA at
http://www.rcuda.net
More than 800 requests world wide

@rcuda_

rCUDA is a development by Universitat Politècnica de València, Spain
Get a free copy of rCUDA at
http://www.rcuda.net
More than 800 requests world wide

rCUDA is a development by Universitat Politècnica de València, Spain
Thanks!

Questions?

rCUDA is a development by Universitat Politècnica de València, Spain