Acceleration-as-a-Service
Exploiting Virtualised GPUs for a Financial Application

Blesson Varghese
University of St Andrews, UK
varghese@st-andrews.ac.uk
www.blessonv.com

Javier Prades, Carlos Reano and Federico Silla
Universitat Politecnica de Valencia
Spain
Overview - 1

- Background
- GPU Virtualisation Framework
- Financial Risk Case Study
- Experiments
- Results
- Conclusions
Hardware accelerators are commonly used in High-Performance Computing (HPC) systems.

A number of Top500 and Green500 supercomputers are supported by hardware accelerators.

Allows for heterogeneity by using regular processors with accelerators (for example, GPUs or Phis).
Best way to set up clusters with accelerators (GPUs):
- GPU for each processor
  - High performance/cost ratio
  - Higher power consumption
  - Under utilisation of accelerators
- Alternate set up:
  - Fewer GPUs on the nodes
  - Access GPUs on-demand
Numerous challenges arise:
- How can GPUs on a server be made available as a service?
  - Virtualisation technologies are required
  - Providing virtual GPUs as a service
  - ‘Acceleration-as-a-Service’ (AaaS)
- How can GPUs be efficiently shared between nodes?
  - Use virtual GPUs concurrently for different application by multiple nodes
  - Virtual GPU cannot be exclusively locked to a single application

Solution
- GPU virtualisation frameworks (for example, rCUDA)
GPU Virtualisation Framework – 1

- rCUDA (Remote CUDA)
  - [http://rcuda.net/](http://rcuda.net/)
  - Developed at the Universitat Politecnica de Valencia
  - Supports NVIDIA CUDA 6.5 runtime API and driver API
  - Partners are Mellanox Technologies and NVIDIA
Scenarios supported by rCUDA
- Single/multiple threads of an application obtaining AaaS from single/multiple remote GPU(s)
- Multiple applications obtaining AaaS from single/multiple remote GPU(s)
Financial Risk Case Study – 1

- Example – Aggregate Risk Analysis
  - Monte Carlo simulation
  - Scenario planning
    - if an earthquake occurs, then how much will be the probable loss?

- Who and Why?
  - Government and Planners
  - Emergency Response
  - Insurers and Reinsurers
Financial Risk Case Study – 2

- Characteristic of Simulation
  - Computationally intensive
  - Done in the back end of a pipeline
  - Slow and cannot produce rapid results
  - Typically run periodically, but not as and when required

- Solution
  - AaaS is a good option to enhance existing capabilities
  - Deliver a most cost effective, good performance, optimally utilised solution
Financial Risk Case Study – 3

- **Algorithm**
  - **Portfolio:** covers risk due to catastrophes like earthquake events
    - Comprises Programs, Layers, Event-Loss Tables (ELTs)
  - **Year Event Table (Simulation Grid)**
    - One million Trials, each Trial covers 1,000 Events
    - Gives different views of how Events can occur in a year
  - **Size of data:**
    - 1 Layer covering 16 ELTs ~ 50 GB
  - **Objective**
    - Aggregate the losses of a Layer for a Portfolio

---

**Algorithm 1: Aggregate Risk Analysis**

```
Input : YET, ELT, PF
Output: YLT
1 for each Program, P, in PF do
2   for each Layer, L, in P do
3     for each Trial, T, in YET do
4       for each Event, E, in T do
5         for each ELT covered by L do
6           Lookup E in the ELT and find corresponding loss, \( l_E \)
7           Apply Financial Terms to \( l_E \)
8           \( l_T \leftarrow l_T + l_E \)
9         end
10     end
11   end
12 end
13 end
14 end
15 end
16 Populate YLT using \( l_T \)
```
Experiments

- **Platform**
  - 1027GR-TRF Supermicro nodes
    - Two Intel Xeon E5-2620 v2 processors (each six cores)
    - 2.1 GHz and 32 GB of DDR3 SDRAM memory
    - Supports InfiniBand FDR and QDR
    - One NVIDIA Tesla K20 GPUs available on each node, or
    - Four K20 GPUs on a single node
  - **Set up**
    - 1 Portfolio with 1 Program covering 1 Layer
    - YET – 1 million trials
    - Each trial has ~1,000 events
    - 16 ELTs
Results – 1

▪ On a Single GPU
Results – 2

- On multiple (4) GPUs
Results – 3

- On 64 threads per block
Conclusions

- GPU virtualisation facilitates ‘Acceleration-as-a-Service’
- rCUDA is a GPU virtualisation framework
- Financial risk applications can benefit from virtual GPUs
- Little (to no) performance overhead of using rCUDA over CUDA
  - Data transfer performance is better in rCUDA than CUDA

Thank you for your attention