GPU on OpenStack for Science

Deployment and Performance Considerations

Luca Cervigni luca.cervigni@pawsey.org.au
Jeremy Phillips jeremy.phillips@pawsey.org.au

The Pawsey Supercomputing Centre is an unincorporated joint venture between

CSIRO, Curtin University, ECU Edith Cowan University, Murdoch University, The University of Western Australia

and proudly funded by

NCRIS National Research Infrastructure for Australia, An Australian Government Initiative; Government of Western Australia

PAWSEY supercomputing centre

Pawsey Supercomputing Centre

- Based in Perth, Western Australia
- Established in June 2000
- Supercomputing, Data Storage, Cloud Computing, Visualisation

The Pawsey Supercomputing Centre is an unincorporated joint venture between

CSIRO

Curtin University

Australia ECU Edith Cowan University

Murdoch University

The University of Western Australia

and proudly funded by

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

Government of Western Australia

PAWSEY supercomputing centre

Pawsey Supercomputing Centre

- We provide free computing resources and training resources to students, industry personnel, researchers, academics and scientists.
- Two Cray supercomputers and multiple HPC clusters
- 10 PB of live storage and 40 PB of tapes
- **Cloud computing cluster based on OpenStack**

Cloud @ Pawsey

- OpenStack Pike
- 46 compute nodes, 39 storage nodes, 12 service nodes.
- 6 GPU nodes with dual NVIDIA V100 card.
- ~3000 cores and 1PB of raw storage with CEPH

PAWSEY
supercomputing centre

GPUs use Case Examples

Agriculture - processing of multi-spectral imagery from remote sensing

Psychology - using TensorFlow to speed up sampling of large and complex Bayesian models

Biology - using molecular dynamics (MD) simulations to assess the interaction of glycans with their receptor proteins

Astronomy - porting the Australia Telescope Compact Array digital backend from FPGA processing to GPU

Use Case Examples

Classification of Shallow Water Fish

Curtin Institute for Computation
Australian Institute of Marine Science

PAWSEY
supercomputing centre

GPU Nodes

- HPe ProLiant DL380 Gen10
- 2x Intel Xeon 6132 (14 cores, 2.6GHz)
- 384GB RAM
- **2x NVIDIA Tesla V100 16GB PCIE**
- 2x 100Gbps Ethernet

PAWSEY supercomputing centre

CPU Isolation and CPU Pinning

- /etc/default/grub

```
GRUB_CMDLINE_LINUX="quiet intel_iommu=on iommu=pt isolcpus=0-6,8-20,22-27"
```

- /etc/nova/nova.conf

```
vcpu_pin_set=0-6,8-20,22-27
enabled_filters=<...>,NUMATopologyFilter
```

- Hyperthreading disabled

PCI Passthrough

https://docs.openstack.org/nova/latest/admin/pci-passthrough.html

- GPU IDs

```
# lspci -nn | grep -i nvidia
37:00.0 3D controller [0302]: NVIDIA Corporation Device [10de:1db4] (rev a1)
86:00.0 3D controller [0302]: NVIDIA Corporation Device [10de:1db4] (rev a1)
```

- /etc/nova/nova.conf on service node

```
alias={"name":"V100","vendor_id":"10de","product_id":"1db4","device_type":"type-PCI"}
enabled_filters=<...>,PciPassthroughFilter
```

- /etc/nova/nova.conf on the nova compute

```
passthrough_whitelist={"vendor_id":"10de","product_id":"1db4"}
```

Flavour Details

- 7 cores
- 90GB of memory (NUMA node adjacent)
- Direct NUMA access to GPU
- 40GB disk on CEPH

Flavour properties:

```
aggregate_instance_extra_specs:pinned='true',
hw:cpu_policy='dedicated',
pci_passthrough:alias='V100:1'
```

Host aggregate properties:

```
pinned='true'
```

PAWSEY supercomputing centre

2-GPU Flavour

https://docs.openstack.org/nova/pike/admin/cpu-topologies.html

```
openstack flavor create --disk 20 --vcpus 14 --ram 186368 \
  --property aggregate_instance_extra_specs:pinned='true' \
  --property hw:cpu_policy='dedicated' \
  --property pci_passthrough:alias='V100:2' \
  --property hw:numa_nodes=2 \
  <flavour-name>
```

NOTE: each GPU has affinity with a different CPU, therefore it is mandatory to have a NUMA aware flavour.

NUMA details

- Each Xeon 6132 has two NUMA nodes
- The VM is configured to use NUMA node adjacent memory, for lower latency and better performance.

```
ubuntu:~$ numactl --hardware
available: 4 nodes (0-3)
node 0 cpus: 0 1 2 3 4 5 6
node 0 size: 96404 MB
node 0 free: 427 MB
node 1 cpus: 7 8 9 10 11 12 13
node 1 size: 96766 MB
node 1 free: 89443 MB
node 2 cpus: 14 15 16 17 18 19 20
node 2 size: 96766 MB
node 2 free: 2366 MB
node 3 cpus: 21 22 23 24 25 26 27
node 3 size: 96766 MB
node 3 free: 92125 MB
node distances:
node   0   1   2   3
  0:  10  21  31  31
  1:  21  10  31  31
  2:  31  31  10  21
  3:  31  31  21  10
```

Pinning and GPU-CPU Affinity

```
Ubuntu: virsh vcpuinfo instance-00003dcc
VCPU:           0
CPU:            0
State:          running
CPU time:       25.1s
CPU Affinity:   y--------------------------

CPU:            1
State:          running
CPU time:       2.0s
CPU Affinity:   -y-------------------------

VCPU:           2
CPU:            2
State:          running
CPU time:       2.5s
CPU Affinity:   --y------------------------
```

```
VCPU:           3
CPU:            3
State:          running
CPU time:       5.7s
CPU Affinity:   ---y-----------------------

VCPU:           4
CPU:            4
State:          running
CPU time:       20.5s
CPU Affinity:   ----y----------------------

VCPU:           5
CPU:            5
State:          running
CPU time:       2.2s
CPU Affinity:   -----y---------------------

VCPU:           6
CPU:            6
State:          running
CPU time:       2.7s
CPU Affinity:   ------y--------------------
```

GPU0	GPU1	CPU Affinity
GPU0	X	0-6
GPU1	SYS	14-20

PAWSEY supercomputing centre

Benchmark Configuration

- Baremetal nodes have multiple CPUs and GPUs therefore performances had to be tuned to be comparable to our default GPU instance flavor.
- VM flavor are configured to use a single NUMA node and to access only that NUMA node adjacent memory.

BareMetal vs VM: 7cores+1GPU

Bare Metal (BM)	Virtual Machine (VM)
Removing GPU from PCI bus via:	7 cores (1 complete NUMA node)
`echo 1 > /sys/bus/pci/devices/0000:86:00.0/remove`	90 GB of RAM (adjacent to the same NUMA node)
Switching off cores:	
`echo 0 > /sys/devices/system/cpu/cpu7/online`	
Local SSD storage	40Gb volume on CEPH

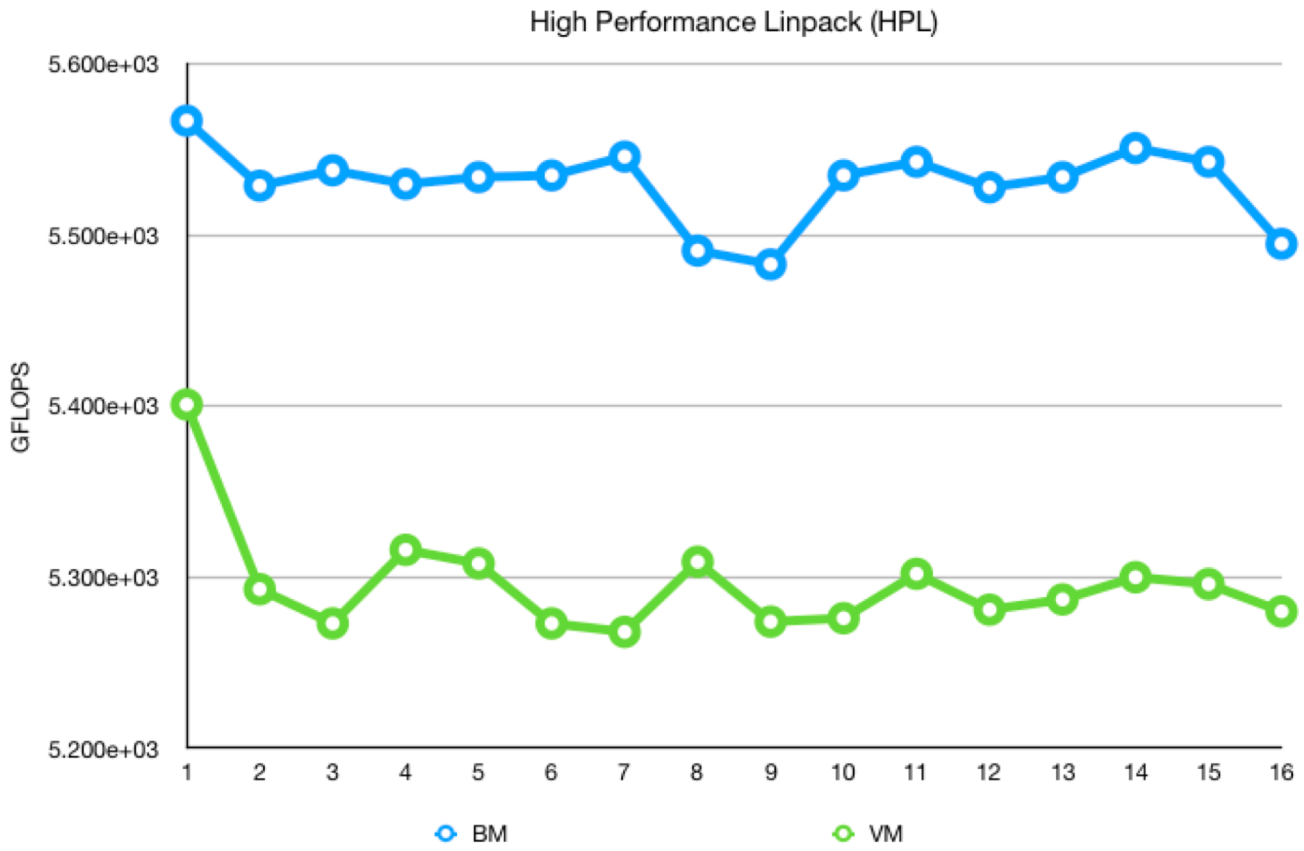
Benchmark 1: High Performance LINPACK

BM Avg: 5530 Gflop/s

VM Avg: 5296 Gflop/s

~4.2% faster in BM

Benchmark Settings:
N = 44000
NB = 256 384 512
Number of runs: 16
Using all CPU cores (threading) and the GPU



PAWSEY supercomputing centre

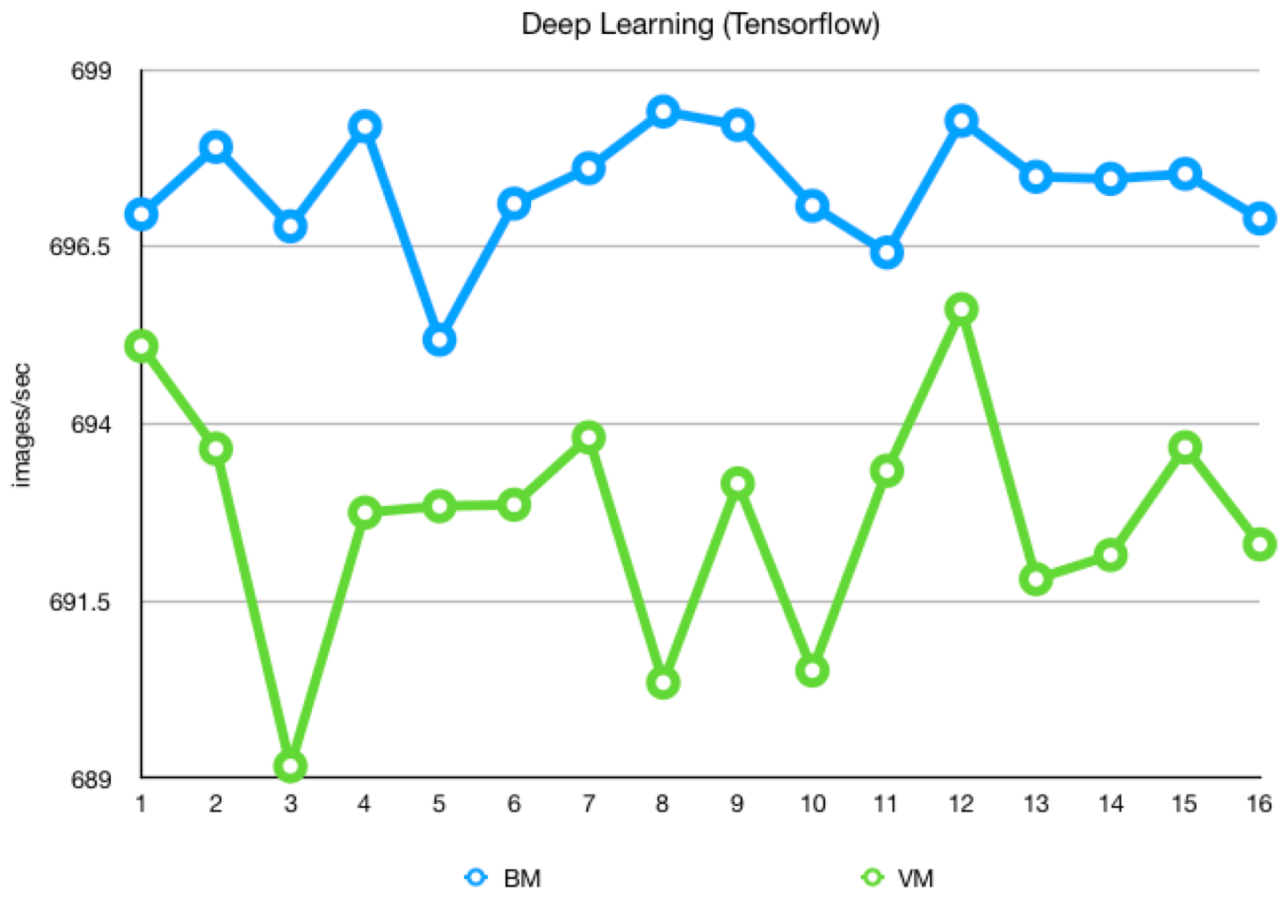
Benchmark 2: Tensorflow

BM Avg: 697.34 images/sec

VM Avg: 692.69 images/sec

~0.6% faster in BM

Benchmark settings:
Resnet50 benchmark
Tensorflow 1.11.0
Precision: fp16
Batch size: 128
Num batches: 100
Number of runs: 16
Using all CPU cores (load ~110%) and GPU



PAWSEY supercomputing centre

Benchmark 3: NAMD

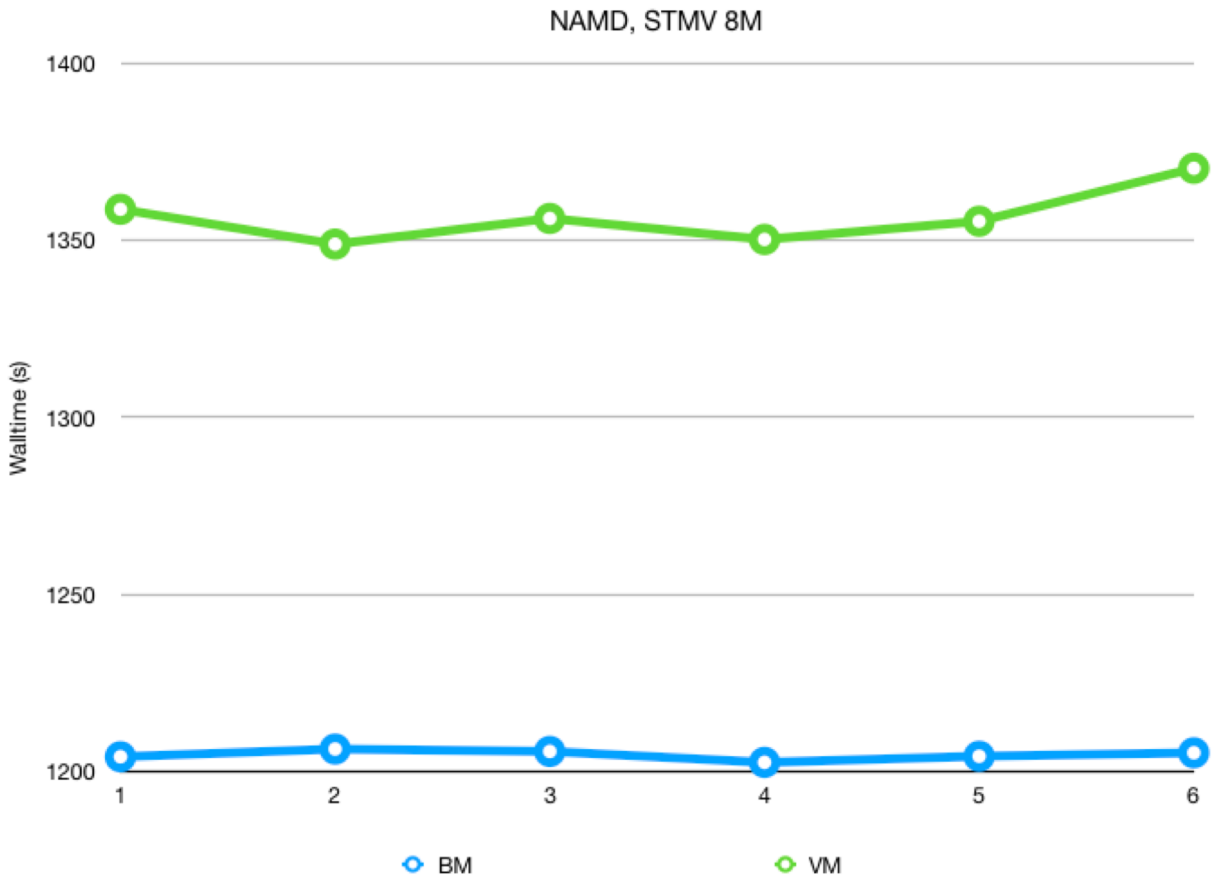
BM Avg: 1204.76 s (walltime)

VM Avg: 1356.61 s (walltime)

~11% faster in BM

NAMD Test Case A, STMV 8M, Unified European Applications Benchmark Suite, PRACE
NAMD version: 2.13b2

Benchmark settings:
Number of runs: 6
Using all CPU cores (charm++) and GPU
CPU load ~700%



PAWSEY supercomputing centre

Open Discussion

Acknowledgements: The Pawsey Supercomputing Centre is supported by $90 million funding as part of the Australian Government's measures to support national research infrastructure under the National Collaborative Research Infrastructure Strategy and related programs through the Department of Education. The Centre would also like to acknowledge the support provided by the Western Australian Government and its Partner organisations.

www.pawsey.org.au

PAWSEY
supercomputing centre

Known Issues

- Each of the network NIC has an affinity with a different CPU.
  - eth1 -> CPU1 and eth2 -> CPU2.
  - But coupled with LACP.
- This configuration can creates issues with vCPU pinning.
  - vCPUs of VM cannot access directly eth2/eth1 without passing through another NUMA node on a different CPU.

Solutions?

- Pinning half vCPUs on CPU1/NUMA0 and half to CPU2/NUMA3 to have both GPU and network access? Even making the flavour NUMA aware increase latencies.
- Changing the networking configuration to eliminate LACP?

Going Ahead and Wish List

We would like to test:

- NVIDIA GRID and vGPUs for Ubuntu/KVM if the NVIDIA binaries will ever be available for Debian.
  - Already available for RHEL for libvirt.

- vGPU support on Queens/Rocky (dependent from previous point)
- RDMA, GPU -> GPU through the network.

Thanks everyone.
Further questions?

Acknowledgements: The Pawsey Supercomputing Centre is supported by $90 million funding as part of the Australian Government's measures to support national research infrastructure under the National Collaborative Research Infrastructure Strategy and related programs through the Department of Education. The Centre would also like to acknowledge the support provided by the Western Australian Government and its Partner organisations.

www.pawsey.org.au

PAWSEY
supercomputing centre