Making use of and accounting for GPUs

Daniel Traynor, GridPP39

Overview

- Given that we don't yet have accounting for GPUs setup I'm planning to talk more about practical GPU issues:

- Purchase, consumer vs enterprise card comparisons pros and cons.

- Deployment, batch system integration, Cuda (Nvidia tax?),

- Accounting

Accelerated computing

A discussion on accelerated computing would include: GPUs (Nvidia, AMD), Intel MICs, Intel AVX, FPGAs. Also Cloud providers provide software services via dedicated APIs that are backed up by dedicated ASICs (e.g. Goggle tensor processing unit (TPU)).

We will limit are discussion to GPUs.

Nvidia has done a lot of “pump priming” of the HPC and enterprise General Purpose GPUs market. Developed their own proprietary software (CUDA) which now dominates.This means we know have an NVIDIA tax.

Will limit discussion to NVIDIA GPUs

Procurement

My GPUs

1* NVIDIA K40c
recycled
Dell workstation

Free

2* NVIDIA K80
HPE DL380

~£10K

1* NVIDIA 1080 Ti
founders edition.
Alienware Aura 2017

~£2.5K

My GPUs

- Got a free K40c from NVIDA via academic program. Used existing workstation with power and space to house the card. Proved we could use in batch and grid infrastructure.

- Brought HPE DL380 (2*K80 = 4 GPU slots).

- Brought “off the shelf” Dell Alienware PC. Able to buy via a framework agreement and get delivery in <10days (good for end of financial year).

- Further development will probably need to allocate dedicated resources.

This is NOT my GPU

System Overview

Introducing the world's first Deep Learning Supercomputer, the DGX-1. Powered by eight NVIDIA Tesla V100 GPU accelerators specifically built for deep learning and machine learning, the DGX-1 will provide you with the fastest possible research so you can explore multiple network architectures and manage and collect datasets to speed up the delivery of data. Available for purchase, free proof of concept trial and rental in the cloud. Please email servers@scan.co.uk for more information.

Ubuntu Server Linux OS

System Powered By

Total PC Price

Sub Total	£230336.48 (inc VAT)
Delivery	£11.99
Total	£230348.47 inc VAT

3U rackmount HPC server, 3200W redundant PSU, dual Intel 10 Gigabit LAN, 4 x Mellanox MCX455A-ECAT Infiniband EDR, 1 x GB management LAN	£0.00 inc VAT

Intel CPUs

2 x Intel Xeon E5 2698 v4 CPUs, each with 20 cores plus HyperThreading, 2.2GHz	£0.00 inc VAT

Memory

512GB ECC Registered DDR4 LRDIMM 2133MHz	£0.00 inc VAT

NVIDIA Computing Processor

8 x NVIDIA Tesla V100, each with 5,120 CUDA cores, 640 Tensor cores and 16GB RAM; system total 40,960 CUDA cores and 5,120 Tensor cores delivering 960 teraFLOPS performance at FP16 (half-precision)	£0.00 inc VAT

Storage

1 x 480GB Intel S3610 for OS, 4 x 1.92TB SSDs in RAID 0 for data	£0.00 inc VAT

Warranty & Support

3 year warranty with next day shipment for replacement parts. Dedicated phone support 24x7.	£230336.48 inc VAT

SUMMARY

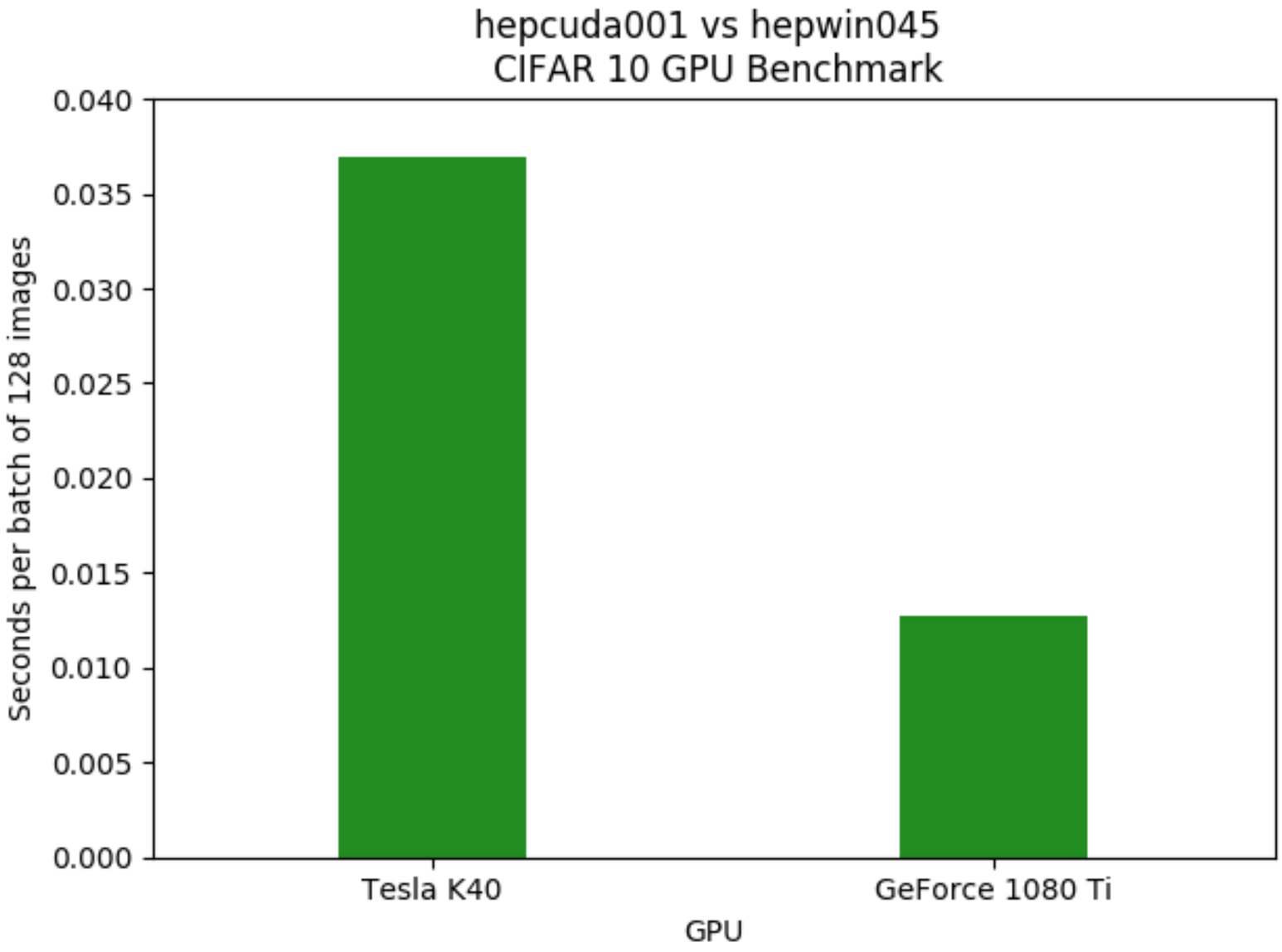
CUSTOMER REVIEWS

Q&A

Performance comparisons

	K40	1080 Ti	V100
RAM	12GB EEC	11GB	16GB EEC
32bit(GFLOPS)	~5,000	~11,000	~14,000
64bit (GFLOPS)	1:3	1:32	1:2
16bit(GFLOPS)	NA	1:64	2:1
8bit (GFLOPS)	NA	4:1	8:1

Performance comparisons



Training for a convolutional neural network was run for 100,000 batches, with a batch size of 128 images (they are 32x32 images). Shown is the mean time taken per batch (Tom Charmen, QMUL).

Tensor flow machine learning, Floating point

Pros and cons of consumer GPUs

	1080 Ti	V100
64 bit	horrendous	awesome
32 bit	awesome	awesome
16 bit	horrendous	awesome
8 bit	awesome	awesome
Warranty	not in server[1]	maybe no the same as server?
ECC	NO	YES
Price	relatively awesome	horrendous

1) Running GeForce GPUs in a server system will void the warranty

note also power consumption, compute capabilities, memory bandwidth.

Configuration

Install gpu driver and cuda

- https://developer.nvidia.com/cuda-toolkit
- https://developer.nvidia.com/cuda-downloads
- http://docs.nvidia.com/cuda/cuda-installation-guide-linux/index.html

Install Software

- Install the cuda repo, has cuda software and drivers

```
rpm --install cuda-repo-centos6-<version>.x86_64.rpm
yum clean all; yum install cuda
```

- Disable the opensource driver: /etc/modprobe.d/blacklist-nouveau.conf

```
blacklist nouveau blacklist nouveau
options nouveau modeset=0
```

- Reboot and check install

```
cat /proc/driver/nvidia/version
nvidia-smi
```

- Compile and run samples, start with deviceQuery

```
cuda-install-samples-8.0.sh <dir>
```

Fine Tuning

- Driver will unload itself from time to time. Enable driver persistence mode to stop this.

```
SL6 in rc.local: nvidia-smi -pm ENABLED -i 0
CENTOS7: systemctl enable nvidia-persistenced
```

- Compute exclusive mode makes the driver refuse a context establishment request if another process already holds a context on that GPU.

```
in rc.local: nvidia-smi -c EXCLUSIVE_PROCESS
```

Integrate with SLURM

- /etc/slum/slum.conf

```
...
# Configure support for our GPUs
GresTypes=gpu
...
NodeName=cn456 CPUs=8 Gres=gpu:teslaK40c:1 RealMemory=11845 Sockets=1
CoresPerSocket=4 ThreadsPerCore=2 State=UNKNOWN CoreSpecCount=1
MemSpecLimit=768
```

- /etc/slurm/gres.conf

```
...
NodeName=cn456 Name=gpu Type=teslaK40c File=/dev/nvidia0
```

- /etc/slurm/cgroup.conf

```
...
ConstrainDevices=yes
```

- /etc/slurm/cgroup_allowed_devices_file.conf

```
...
/dev/nvidia*
```

- Submit jobs: `sbatch --gres=gpu:1 -n1 test_gpu.sh`

Integrate with SGE

- Create host complex : qconf -mc

```
#name               shortcut   type         relop requestable consumable default  urgency
#-------------------------------------------------------------------------------------------
…
gpu                 gpu        INT          <=    YES         YES        0        0
…
```

- Add complex attribute to host: qconf -me cn456

```
hostname              cn456.htc.esc.qmul
load_scaling          NONE
complex_values        gpu=1
user_lists            NONE
xuser_lists           NONE
projects              NONE
xprojects             NONE
usage_scaling         NONE
report_variables      NONE
```

- Submit job: `qsub -l gpu=1 testgpu.job`

Integrate with Cream CE

- Development of new CREAM CE version, specifically for CentOS 7.
- https://wiki.egi.eu/wiki/GPGPU-CREAM
- QMUL is using a patch for our SL6 Cream CEs.
- Introduced new JDL parameters that will be passed to the batch system: GPUNumber, MICNumber and GPUModel
- This works with SLURM and should work with SGE.

Other

- I don’t believe ARC CEs do not have dedicated GPU support.

- I understand that Manchester have a GPU queue available on the Grid. They have done this via a dedicated queue. I also believe that they do not operate their GPUS in an exclusive mode.

- I’m unaware of any other UK site with GPUs available on the grid.

- I am aware that a number HEP sites do have GPUs and that various experiments are developing GPU workloads.

Accounting

APEL

- John Gordon -

- APEL grid works on batch logs and we haven't found a batch system that associates the use of a GPU by a batch job and so with a VO or user. This is generally because gpus can be shared.

- In cloud use GPUs are associated to one and only one VM so APEL cloud accounting can associate the use with that VM and hence its user/VO. Since cloud accounting is only done on wallclock time (no consistent way to report cpu usage of VMs) then the time a GPU is attached to the VM is also walltime then we are reporting something comparable for cpu and gpu.

- We have proposed(tested?) an extension to the cloud record to include GPUs. We have a guineapig site. Greg can tell you the latest.

APEL

- Greg Corbett -
- For gird, we need to look into the currently previewed CREAM/C7 release to see if it offers an opportunity for grid based GPU accounting.
- For the cloud development, I don't believe the extension to the cloud record to include GPUs was ever tested, I'll get in touch with the site helping us and try to arrange a test.

Comments

- We have significant performance difference not just between generations of GPUS (K40,P100,V100) but also for different types of calculations (8/16/32/64bit). The option of consumer GPUs at a fraction of the price adds more differences. Difficult to see how this can be accounted for in the accounting.

- I think there is a lot of development work ongoing with GPUs in HEP. Machine learning is a major “disrupting” technology that is here now. conservative HEP code will change but we have time to adapt.

Summary

- Adding GPUs to a cluster is expensive, there are cheeper options but they have limitations.

- The drivers, software and batch system integration is well understood.

- Accounting using wall clock time possible but incomplete.