



What to know about NeSI GPUs Should I use GPUs for my research?

6th May 2021



New Zealand eScience Infrastructure

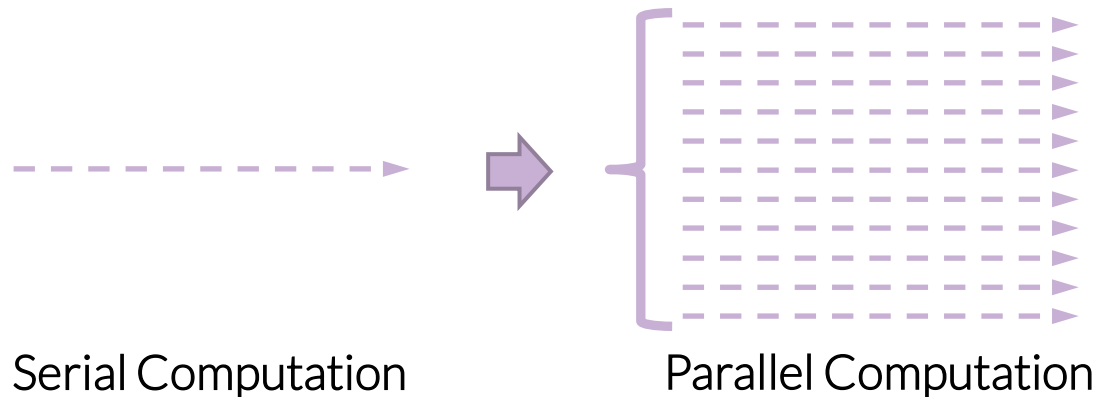
Overview

1. What is a GPU, and what are they good for?
2. How can I use a GPU?
3. Which GPUs can I access on NeSI?
4. How can I get access?
5. Q&A



What is a GPU, and what are they good for?

What is a GPU?



- GPU = Graphics Processing Unit
- Designed for applying the same operation to many pieces of data in parallel, can be much faster than a CPU
- Less flexible, lower frequency, and smaller caches than CPU, but many, many, many cores (thousands)
- Typically run as a co-processor (integrated or external)

What is a GPU?

*"If you were plowing a field, which would you rather use:
two strong oxen or 1024 chickens?"*

Seymour Cray



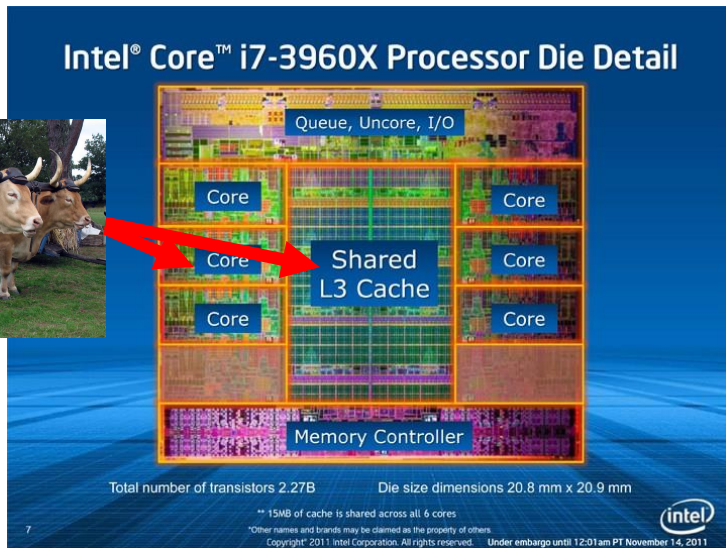
By Monster1000 – Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=1826910>



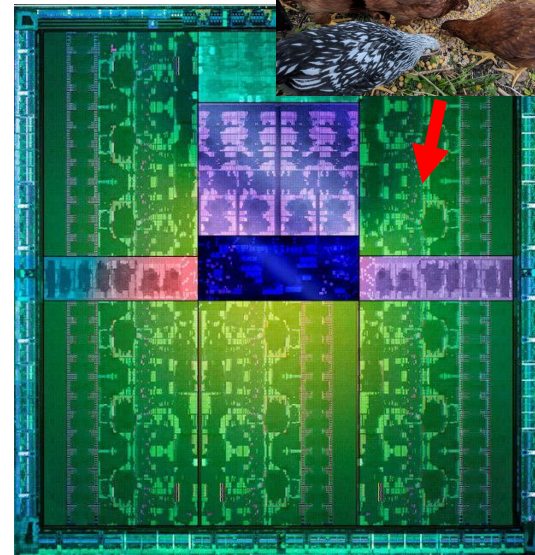
By Rbreidbrown – Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=79919470>

What is a GPU?

... today we will be using chickens - in some fields... 😊



Intel Core i7 CPU



Nvidia GK110 GPU

What are they used for?

Applications

- Graphics processing ☺ - OpenGL, Vulkan, ...
- Gaming – Physics models, ...
- Scientific Computation - CFD, materials science, ...
- Signal processing – Software defined radio, ...
- Image processing – Photo manipulation, ...
- Video processing – Video editing, ...
- Machine Learning - Neural networks, ...
- Bioinformatics – Genomics, ...
- Computer Vision – Robotics, self-driving cars, ...
- ...

So... Should I be using a GPU?

Short answer: yes, if you can... but not always!

- Yes, if you do a lot of numerical work in large batches that can be computed in parallel
- No, if you only have modest amounts of numerical work, or many small batches

Also keep in mind that:

- Supercomputing relies more and more on GPUs
- The future is parallel - GPU-friendly workloads will (often) also run fast on modern CPUs
- GPU memory is still limited – but getting larger
- Some alternatives exist (e.g., TPUs for machine learning)



How can I use a GPU?

How can I use a GPU?

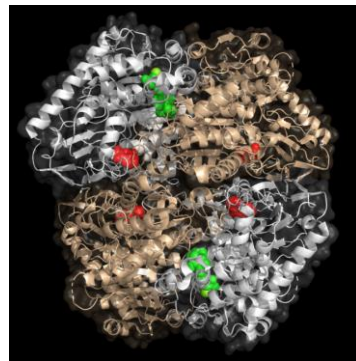
1. Your software package comes with GPU support
 - More and more packages support GPU (TensorFlow, PyTorch, GROMACS, NAMD, ...)
 - Many packages are provided as containers
 - You may need to check your license in some cases
2. Your code can use GPU libraries
 - Mainly standard compute libraries: LAPACK, BLAS, sparse linear algebra, FFT, ...
 - Just requires relinking your program or small source code modifications in many cases

How can I use a GPU?

3. Your programming language supports GPUs
 - gpuArrays (Matlab), cupy (Python), C++17 parallel STL, Fortran “*do concurrent*”, ...
 - Recompile with an offloading compiler, limited support
4. Directive-based APIs
 - OpenMP 4.5, OpenACC for C/C++/Fortran
 - Avoids the need to fork your code, limited support
 - May require algorithmic changes and refactoring
5. Low-level programming models/APIs
 - CUDA (Nvidia), OpenCL, [HIP (AMD), Vulkan]
 - Most powerful but may require algorithmic changes and at least partial code rewriting

How can I use a GPU?

Case Study – Software supports GPU



- Protein Modelling with NAMD– Dr Wanting Jiao (VUW)
- Very computationally expensive
- Use NAMD container provided by Nvidia
- P100 GPU \approx 3 Māui Skylake nodes (120 cores)
- A100 GPU x1.7 faster than 2 Māui nodes

How can I use a GPU?

Case Study – Software supports GPU



- Natural language processing project with Qiming Bao (UoA)
- Project uses Fairseq Deep Learning toolkit
- A100 2x faster than P100 with small batch size
- A100 10x faster with larger batch size (requires more memory than available on P100)

How can I use a GPU?

Case Study – GPU Libraries



- Tropical Circulation Model– Dr Gilles Bellon (UoA)
- Runtime dominated by matrix multiplication
- Offload BLAS routine “dgemm” to GPU with cuBLAS
- P100: “dgemm” x1.4 faster than 16 CPU cores with Intel MKL, comparable to 32 cores

How can I use a GPU?

Case Study – Directive-based APIs



- Marketing Insights– Dr Damien Mather (UoO)
- Runtime dominated by log determinant calculation
- Use OpenACC directives to offload computation
- P100: 13x faster than 1 CPU core, similar to 40 cores
- A100: similar to 160 cores, 3x faster than P100



Which GPUs can I access on NeSI?

Which GPUs can I access on NeSI?



Nvidia Tesla P100
Image: Nvidia

- Data-centre grade GPU – fast double precision math!
- Good for...
 - Smaller scientific problem sizes (unable to scale to A100s)
 - Smaller machine learning workloads (e.g., smaller amounts of training data, less complex models)
 - Software that lacks A100 support (e.g., older versions of TensorFlow)

Which GPUs can I access on NeSI?



Nvidia Tesla A100
Image: Nvidia

- Latest hardware – more compute cores, tensor cores, and much more GPU memory
- Supports multi-user GPU partitioning (“MIG mode”)
- Good for...
 - Large scientific compute – can be 2-3x faster than P100
 - Large machine learning workloads
 - Memory-demanding codes

Which GPUs can I access on NeSI?

A100 Launch

- Only Machine Learning projects will get access initially
- Please email support@nesi.org.nz if you are interested
- ... and follow our news channels for further updates...

Join our mailing list at <https://www.nesi.org.nz/>
(training alerts, newsletters, event announcements, etc.)

Follow us on social channels



@NeSI_NZ



New Zealand eScience Infrastructure



How can I get access?

How can I get access?

- Ask for an allocation via www.nesi.org.nz/apply
- Allocation pricing (P100)
 - Mahuika CPU: 1 “compute unit” per hour***
 - Mahuika GPU: 40 “compute units” per device-hour (plus CPUs and host memory)
 - Māui Ancil: GPU included in Māui allocation
- A100 pricing – *coming soon*

*** See support.nesi.org.nz for details

How can I get access?

If you already have an allocation...

Request a GPU using a Slurm directive (batch mode):

```
#SBATCH --gpus-per-node=1
```

Or request a GPU for interactive access:

```
salloc --gpus-per-node=1
```

Or use a GPU on JupyterHub:

```
jupyter.nesi.org.nz
```


Computational Science consultancy

- A service offered to NeSI platform users, generally at no cost to the researcher
- NeSI Research Software Engineers work directly with research group members
- Goal is to **raise the capability** of the research group

Our Research Software Engineers can assist with:

- **Workflow parallelisation** – allowing more inputs to be processed simultaneously
- **Software parallelisation** – use of technologies such as OpenMP or MPI to process one single input more quickly
- **Code optimisation** – redesign of algorithms to improve overall speed or efficiency of resource use
- **Improving I/O performance** – speed up reading from or writing to the disk, or to reduce the amount of data that must be read or written
- **Porting to GPU** – accelerate code by offloading computations to a coprocessor
- **Improving software sustainability** – introducing best practices such as version control and unit testing



Contact support@nesi.org.nz



Q&A

Q&A

... over to you!

Additional slides

Which GPUs can I access on NeSI?

	Nvidia P100	Nvidia A100
Max Clock Rate	1303 GHz	1410 GHz
Single Precision	9340 GFLOPS	19500 GFLOPS
Double Precision	4670 GFLOPS	9700 GFLOPS
Memory	16 GB	40 GB
Extra Features		Tensor Cores
Total Number	13 GPUs	8 GPUs
Availability	Now	Very soon 🕶️