New Research Computing Facilities

for Comp Sci.

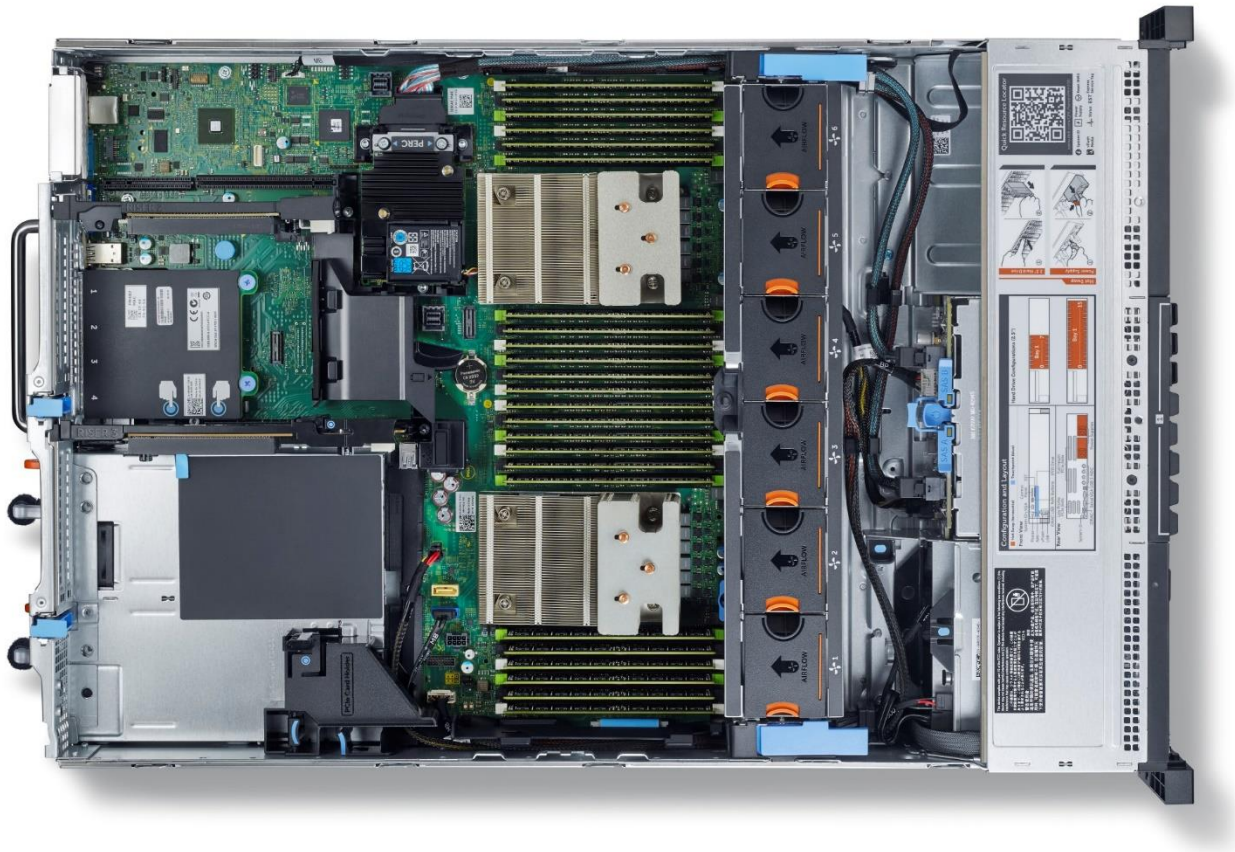
Dr Paul Richmond

Overview

- The department has invested in HPC computing facilities
  - All hosted within ShARC
  - Private DCS access
  - Full ShARC software suite and support
- Big memory nodes
  - Scalable ML with Apache Spark
  - Suited for problems which require large in memory computation
- A GPU Supercomputing system in a box (DGX-1)
  - Deep Learning Machine…
  - Much faster at training DL Networks

Big Memory Nodes

- 3 x Dell R730 nodes. Each has;
  - 768GB DDR4 RAM,
  - OmniPath connection (100Gb/s),
  - Xeon E5-2630 v3s (AVX2, FMA, 16-core)
- 48GB/core vs 4GB/core for standard ShARC nodes.
- Currently being tested
- Used on Scalable ML course

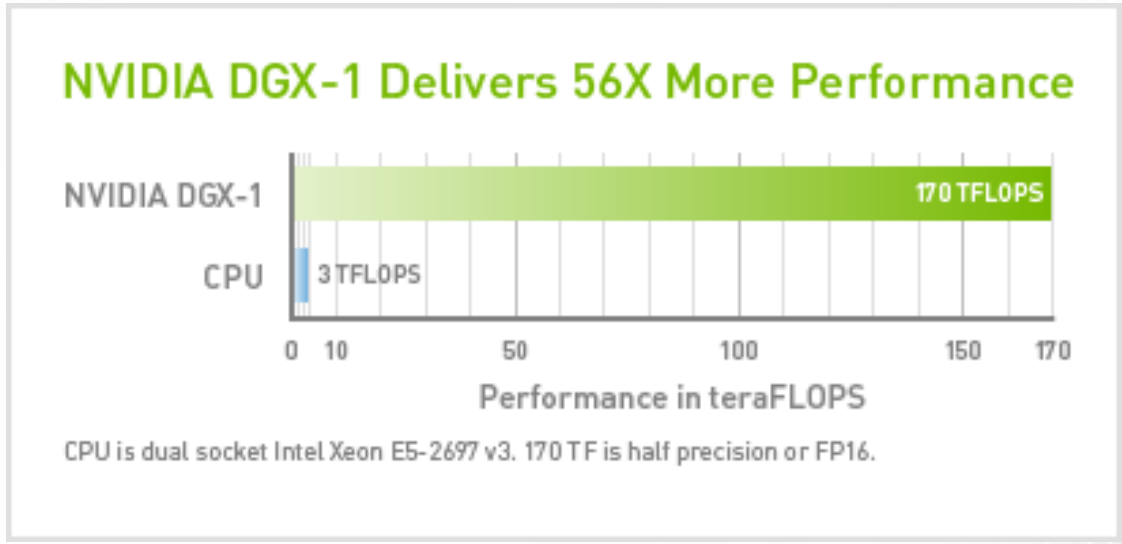


The University Of Sheffield.

NVIDIA GPU RESEARCH CENTER

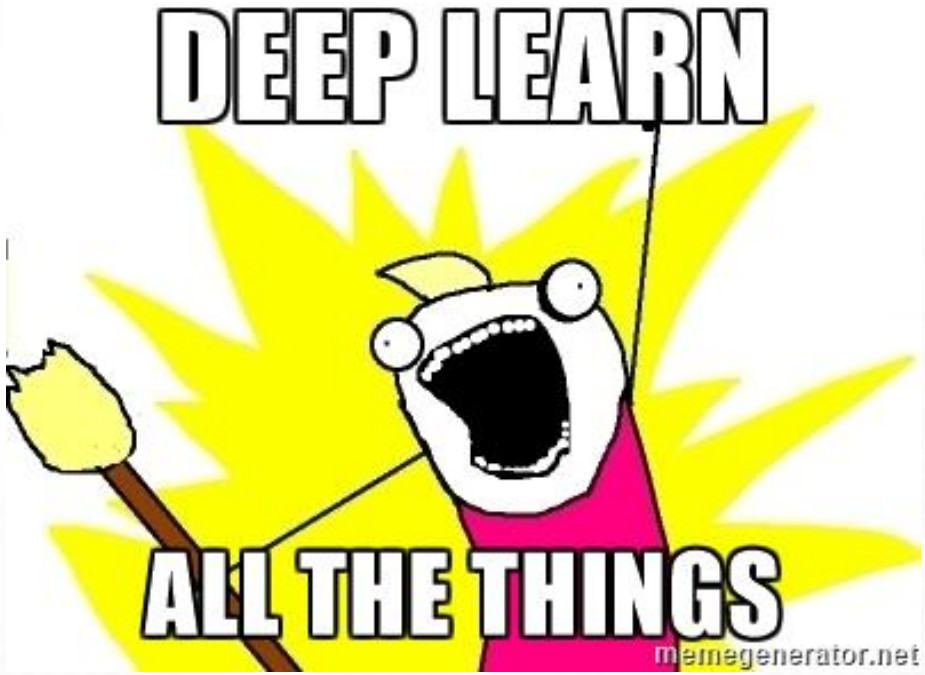
The DGX-1

- Single Node (custom built)
  - 8 x NVIDIA P100 GPUs (16GB each)
  - Dual 20-core Intel Xeon E5-2698 v4 2.2Ghz
  - 512GB RAM
- Huge amount of performance
  - 170 TFLOPS
  - Number 1 spot in top500 in 2004



DGX-1 Use Cases

- ❑Deep Learning
- ❑Deep Learning
- ❑Deep Learning
- ❑Deep Learning
- ❑…
- ❑GPU Computing

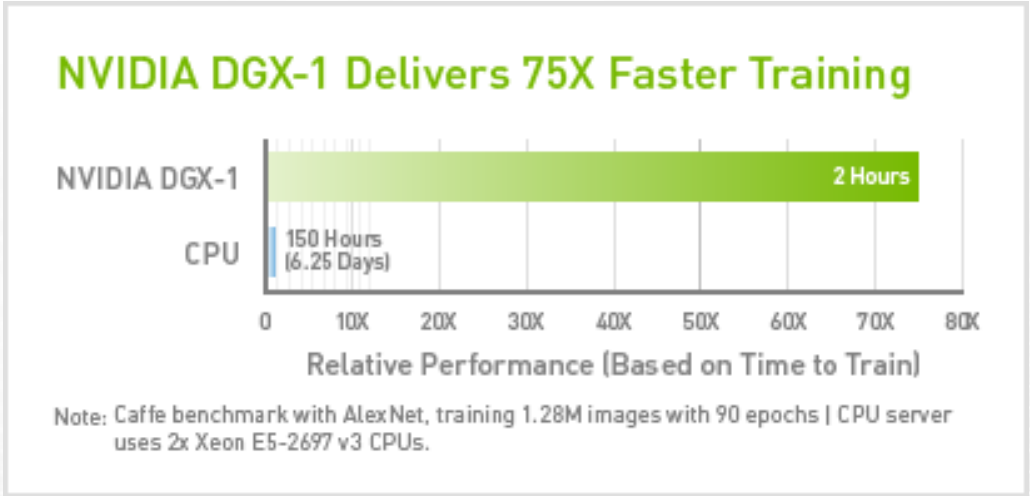


The University Of Sheffield.

NVIDIA GPU RESEARCH CENTER

Why Deep Learning on GPUs

- Training Deep Learning Network = Matrix Multiplications
  - GPUs are fantastic at this
  - Addition of fast memory bandwidth avoids this being memory bound
- DL does not require high precision
  - GPUs have optimised FP16 performance
- Training can be distributed
  - You can get near linear speedups by using more GPUS

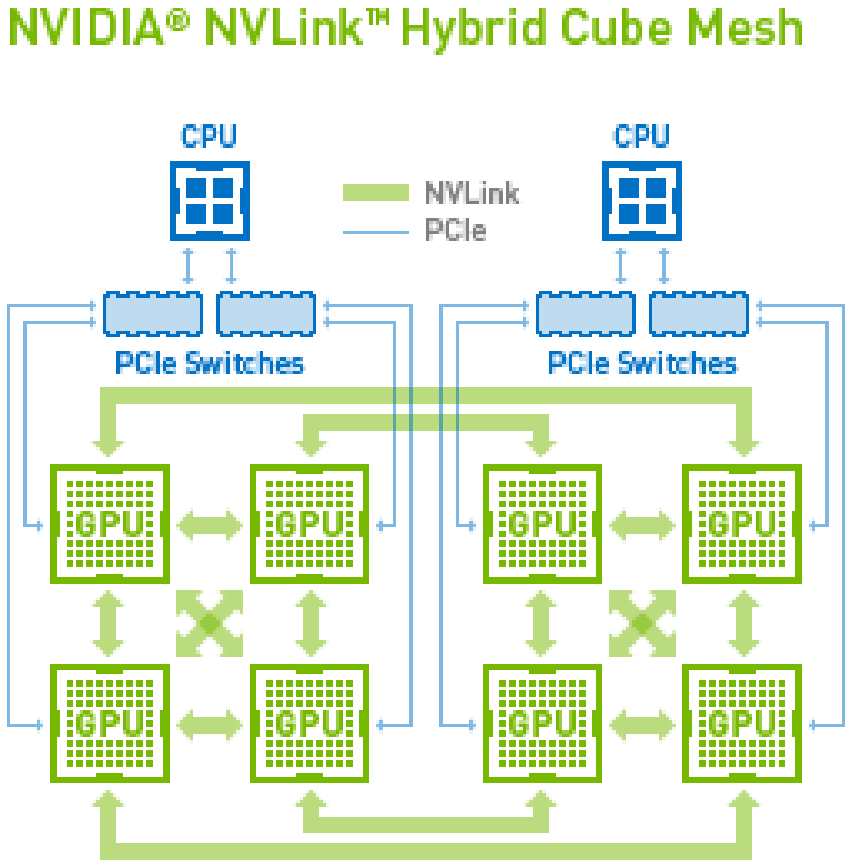


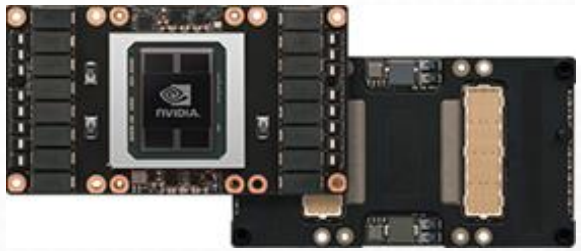
The University Of Sheffield.

NVIDIA GPU RESEARCH CENTER

DGX-1 Special DL Use Cases

- Bigger data sets
  - Get results faster
  - Distribute between the 8 GPUs on the node
- Bigger Networks
  - NVLink between the GPU devices
    - up to 12x faster than PCIe
- Up to 732GB of addressable space
  - *virtually* unified





The University Of Sheffield.

NVIDIA GPU RESEARCH CENTER

Deep Learning Platforms & Frameworks

- **Supported on ShARC**
  - Theano – Python, low-level
  - Tensorflow – Python, low-level with some built-in ML/DL features and visualiser
  - Caffe – High-level, CLI, C++ with Python and Matlab interface
  - Torch – High-level, LUA interface
- High-level Wrappers
  - Keras – Theano & Tensorflow
  - Lasanga (for Theano)
  - sklearn-theano
  - DIGITS – GUI for Caffe and Torch Training

The University Of Sheffield.

NVIDIA GPU RESEARCH CENTER

Let me at it…

- Available to all Comp Sci staff and students
  - **ShARC access**
    - Need to be on the list
    - See ShARC docs for software guidance (http://docs.hpc.shef.ac.uk/en/latest)
  - **Big Memory Nodes**
    - Details soon…
  - **DGX-1 (https://github.com/RSE-Sheffield/GPUComputing)**
    - Request software and updates (via GitHub issue tracker)
    - Ask for help (via GitHub issue tracker)
    - 1 to 1 support
- RSE Support
  - EPSRC want to see specialist SE and support costed in this way

- All queries: rse@shef.ac.uk

The University Of Sheffield.

NVIDIA GPU RESEARCH CENTER