Singularity for GPU and Deep Learning

Twin Karmakharm
Research Software Engineer
University of Sheffield

30th June 2017

The RSE Sheffield team

- Leads
  - Mike Croucher
  - Paul Richmond
- Members
  - Tania Allard
  - Mozhgan Kabiri Chimeh
  - Will Furnass
  - Twin Karmakharm
  - Anna Krystalli

Research Software Engineering Sheffield. NVIDIA GPU RESEARCH CENTER The University Of Sheffield.

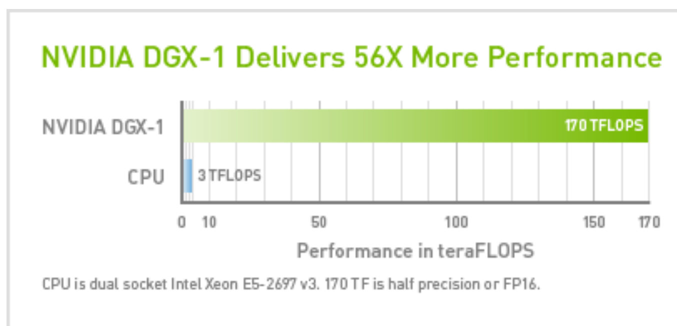
Contents

The ShARC cluster

- Sheffield Advanced Research Computer, new High Performance Computing (HPC) cluster at Sheffield
  - CentOS 7 with Son of Grid Engine (SoGE) scheduler
  - Infiniband interconnect
- 124 Normal memory nodes
  - 2x8 core processor (64GB RAM, 4GB per processor)
- 4 Large memory nodes
  - 2x8 core processor (256GB RAM, 16GB per processor)
- 8 GPU units with Nvidia K80
- + other private K80 and P100 racks

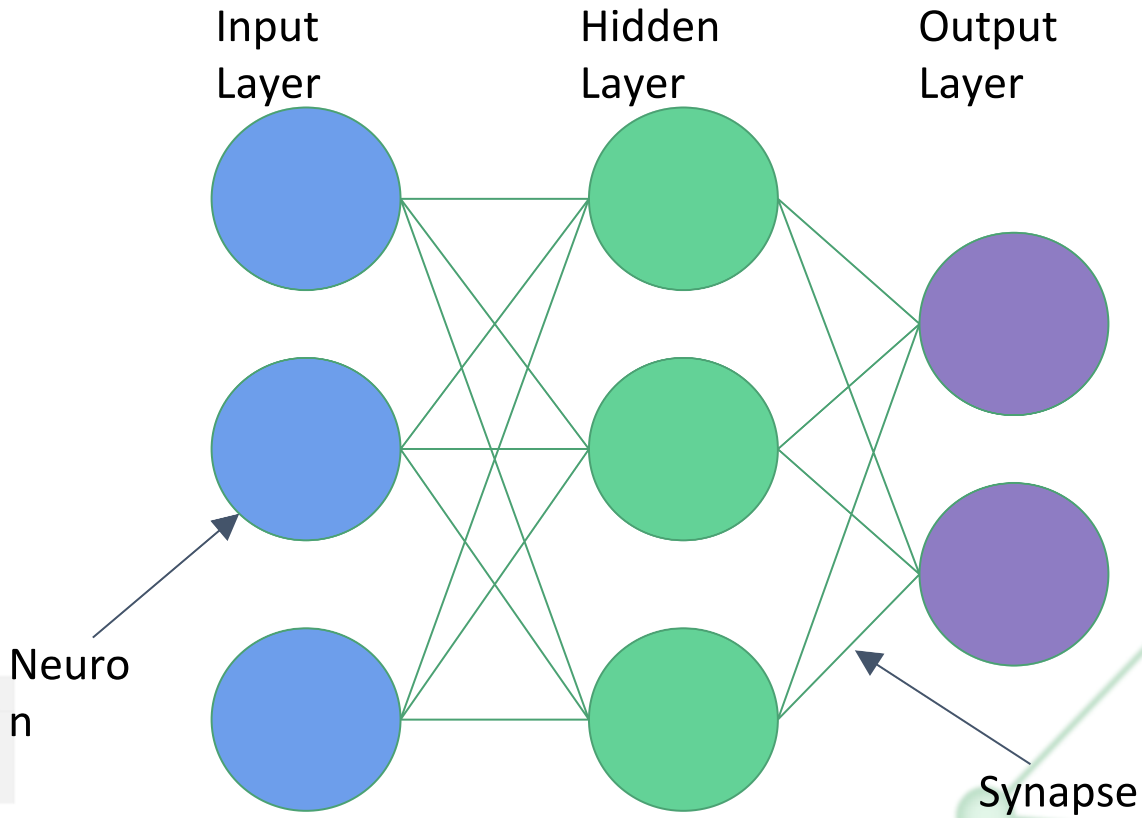
The ShARC cluster: DGX-1

- 8xNvidia P100 GPUs (16GB each) – 170TFLOPS of computation
- Dual 20-core Intel Xeon E5-2698 v4 2.2Ghz
- 512GB RAM



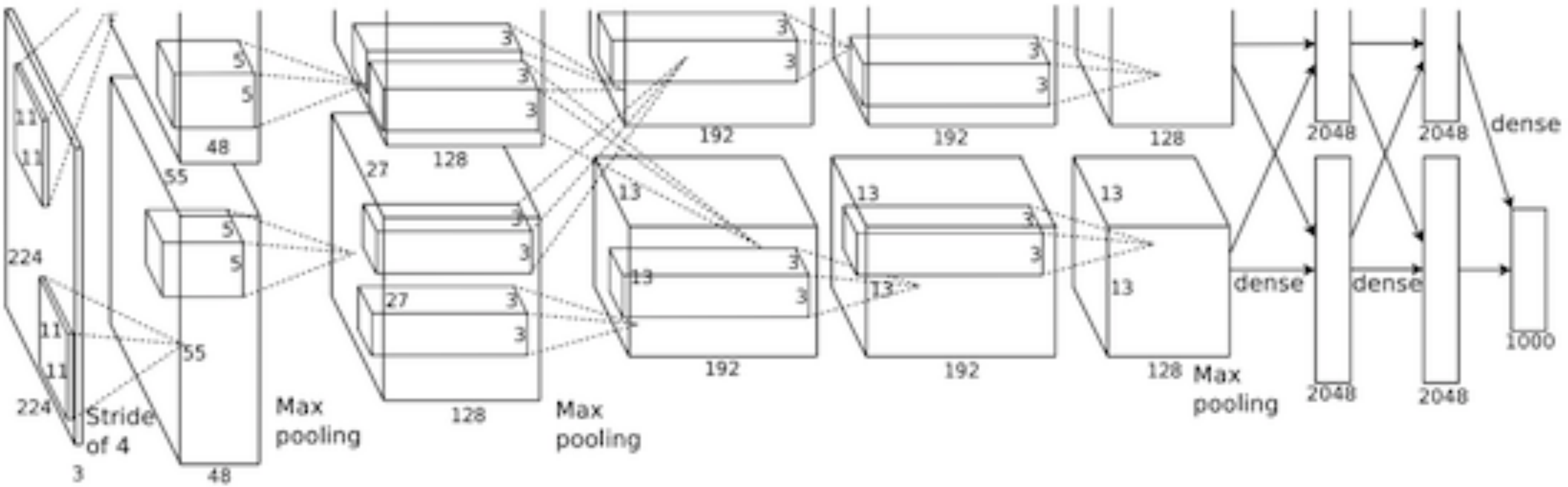
What is Deep Learning (DL)?

- A sub-category of Machine Learning
- Uses neural networks with many hidden layers



What is Deep Learning (DL)?

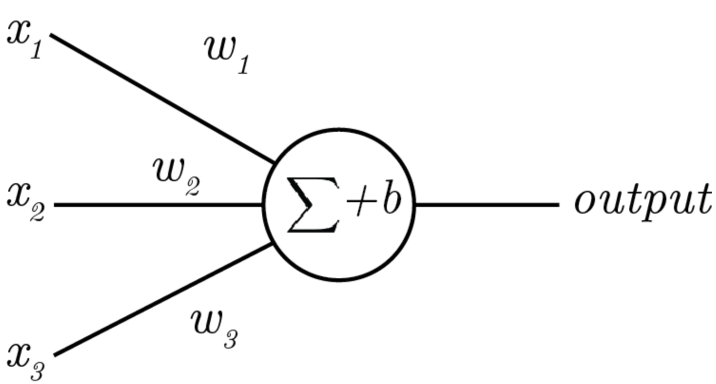
- Hierarchy of representation/feature extractors



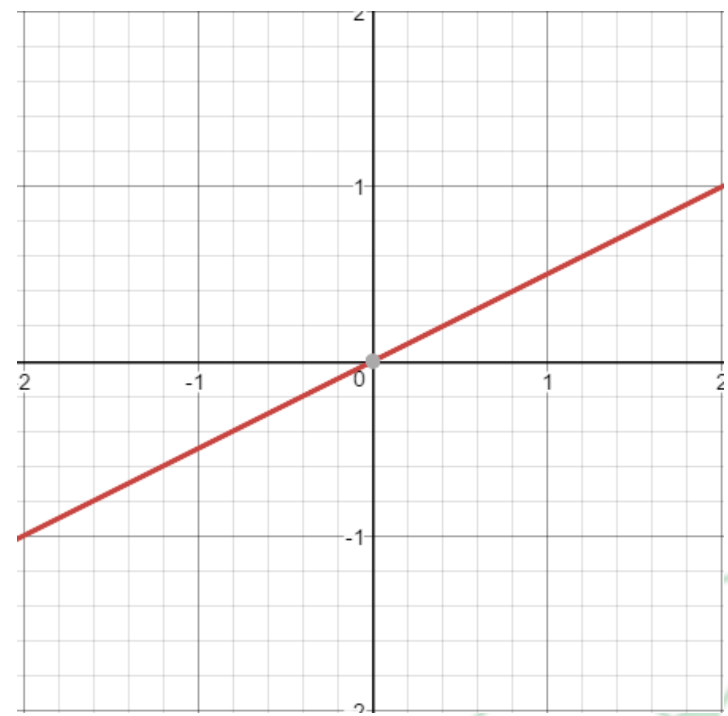
[Alexnet – Krizhevsky et al. 2012]

A Neuron (Perceptron)

- Output is a sum of all input plus bias

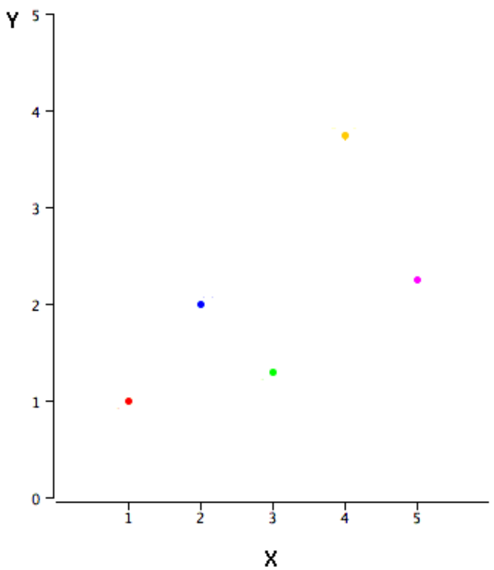


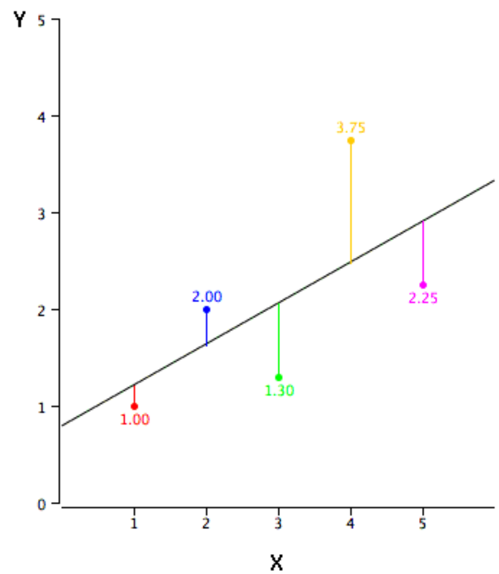
$$output = \sum_j \{x_j w_j\} + b$$

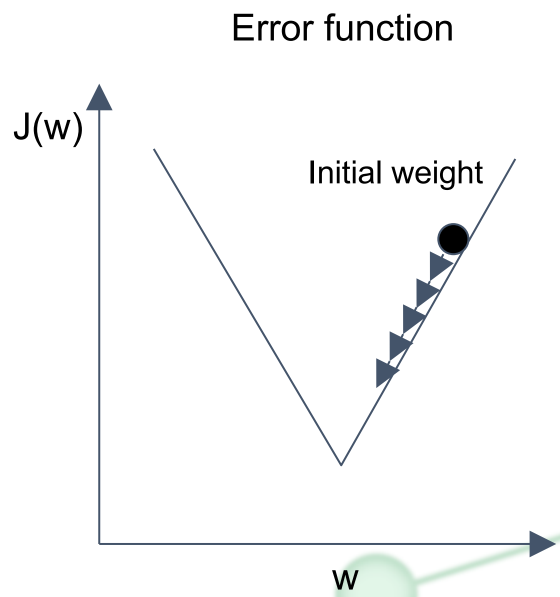


Linear Regression

- Fitting an optimal line through a data set by minimising error







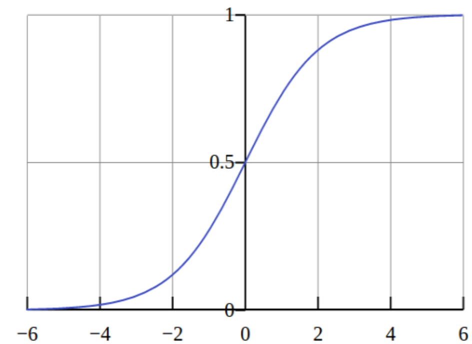
[http://onlinestatbook.com/2/regression/intro.html]

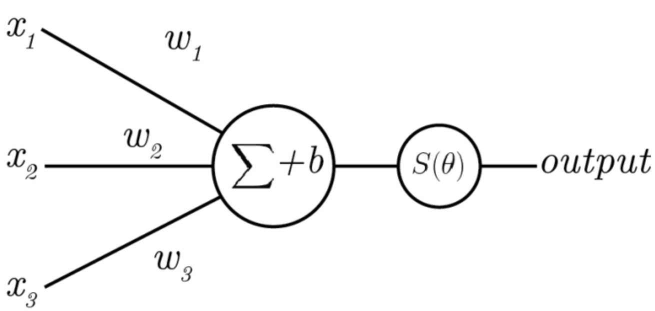
Non-linear activation for real-world problems

- Apply non-linearity over output (Sigmoid in this case)
- Output is between 0 and 1, value in between is 'confidence'

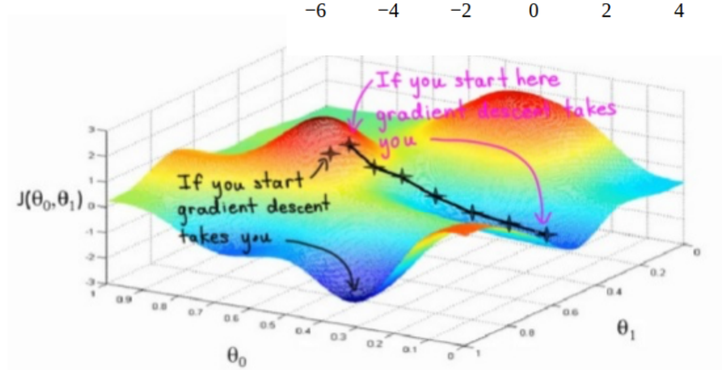
Sigmoid activatior

$$S(t) = \frac{1}{1+e^{-t}}$$





$$output = \frac{1}{1 + e^{\sum_j \{x_j w_j\} + b}}$$



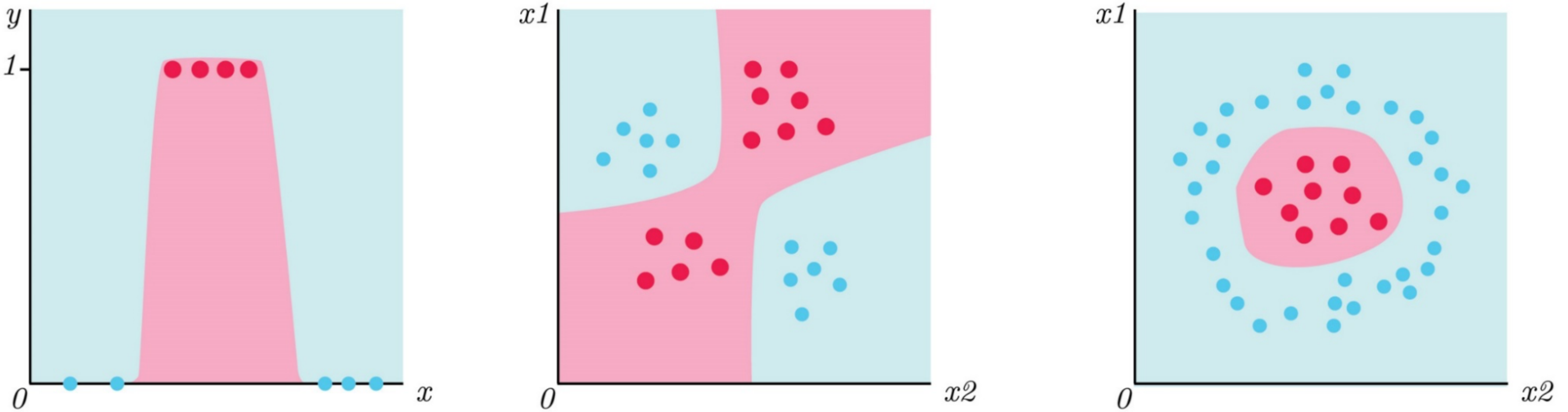
[Andrew Ng – Machine Learning module]

Research Software Engineering Sheffield.

NVIDIA GPU RESEARCH CENTER

Logistic regression

- Classification of data



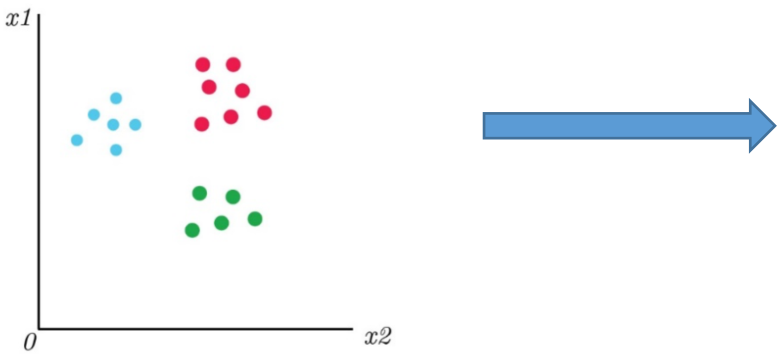
Research Software Engineering Sheffield.

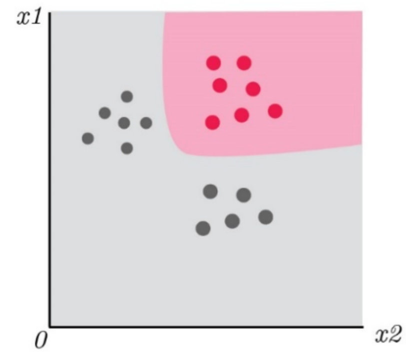
NVIDIA GPU RESEARCH CENTER

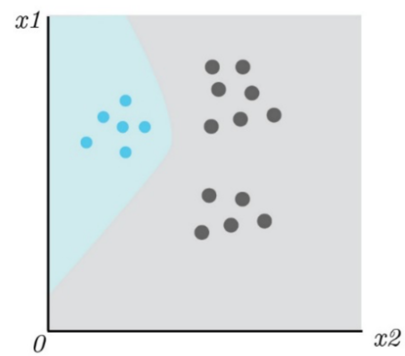
The University Of Sheffield.

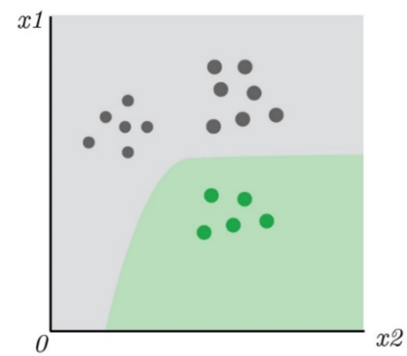
Multi-class logistic regression

- Classify each separately
- One NN output for each classification



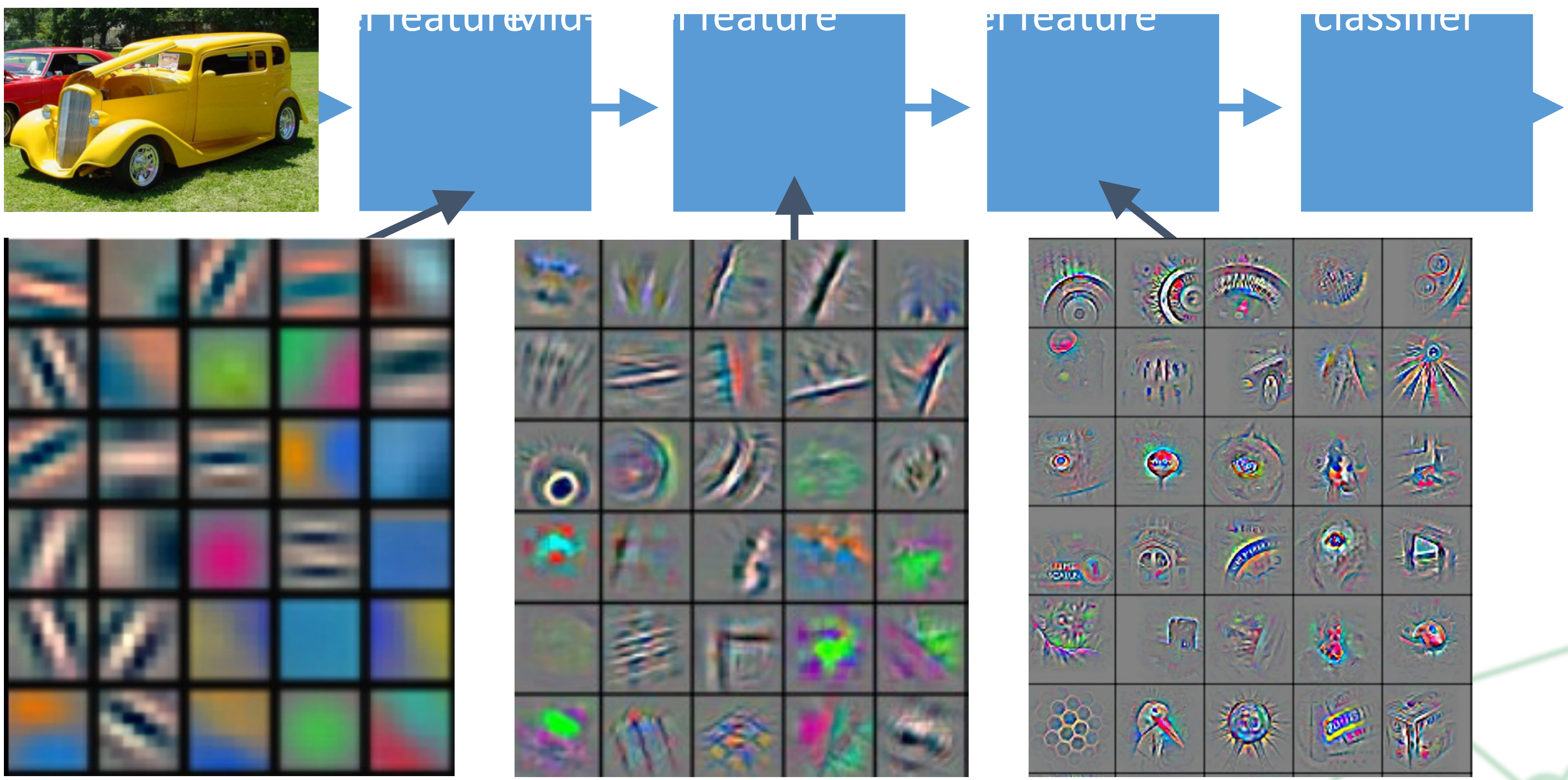






DL – Learning representation/features

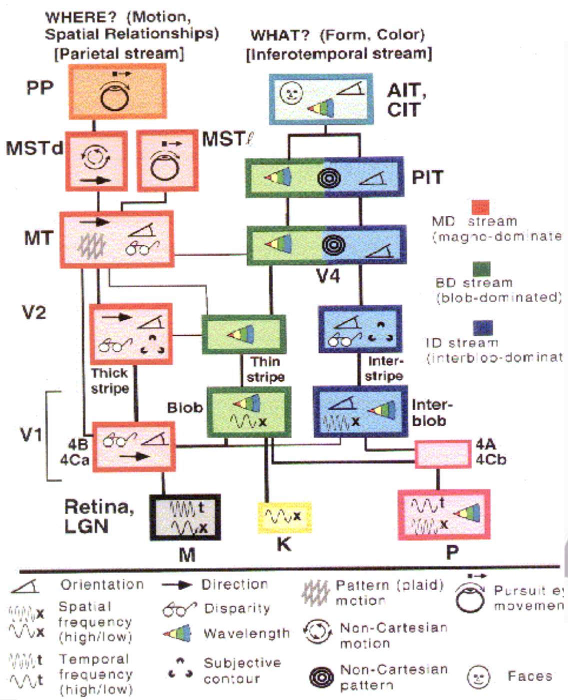
- Hierarchy of features



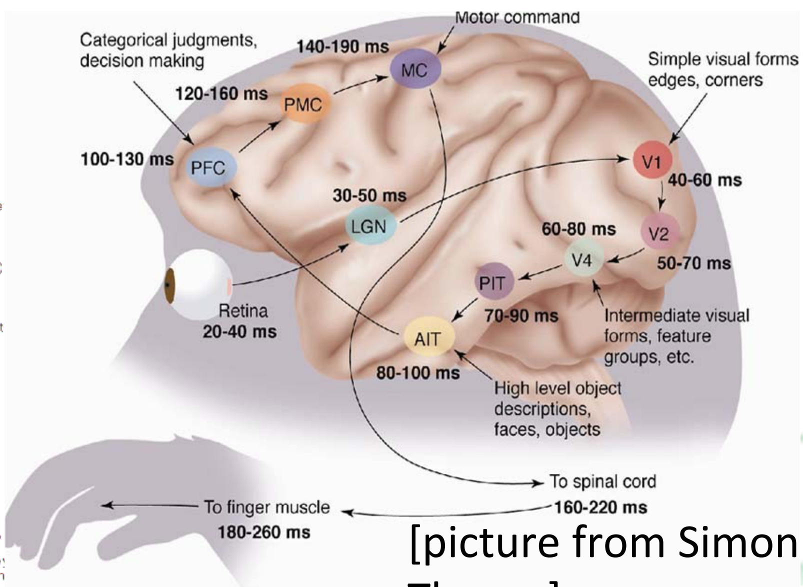
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

The mammalian visual cortex is hierarchical

- The ventral (recognition) pathway in the visual cortex has multiple stages Retina - LGN - V1 - V2 - V4 - PIT - AIT ....
- Lots of intermediate representations



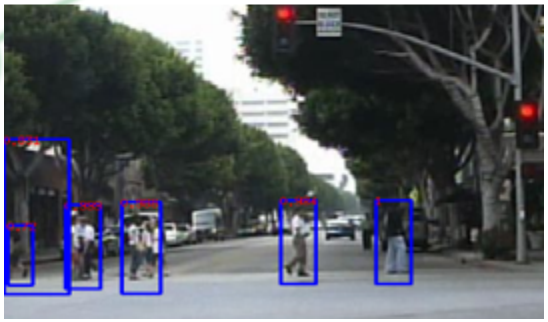
[Gallant & Van Essen]



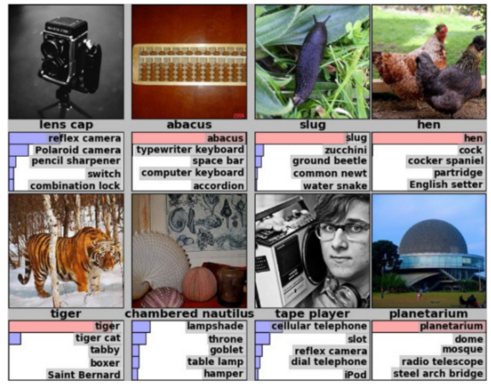
[picture from Simon Thorpe]

What is Deep Learning used for?

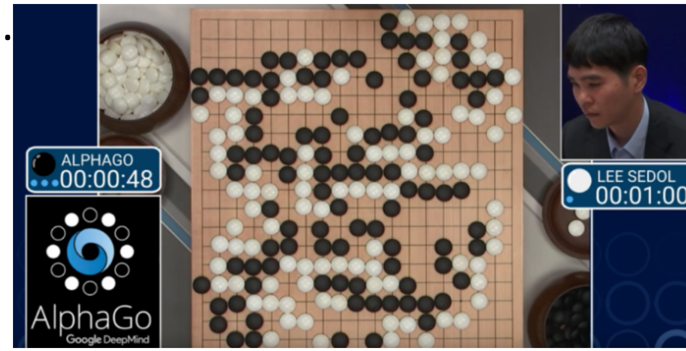
1.



2.

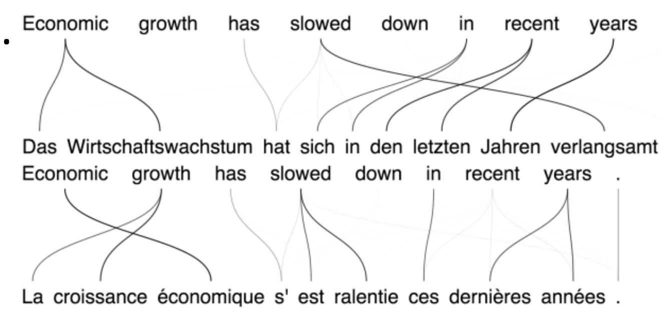


3.



4.

5.



1. https://news.developer.nvidia.com/real-time-pedestrian-detection-using-cascades-of-deep-neural-networks
2. http://danielnouri.org/notes/2014/01/10/using-deep-learning-to-listen-for-whales/
3. https://deepmind.com/research/alphago/
4. http://www.33rdsquare.com/2015/01/what-do-you-need-to-know-about-deep.html
5. https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

Research Software Engineering Sheffield.

NVIDIA GPU RESEARCH CENTER

The University Of Sheffield.

What is Deep Learning used for?

- Handwriting Recognition
  - convert written letters in to digital letters
- Language Translation
  - translate spoken and or written languages (e.g. Google Translate)
- Speech Recognition
  - convert voice snippets to text (e.g. Siri, Cortana, and Alexa)

What is Deep Learning used for?

- Image Classification
  - label images with appropriate categories (e.g. Google Photos)
- Autonomous Driving
  - enable cars to drive

What is Deep Learning used for?

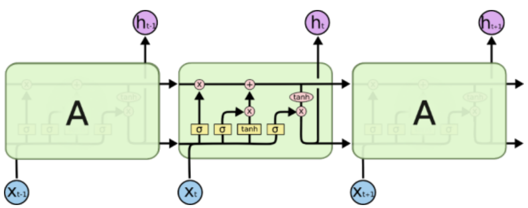
- Examples of DL use at the University of Sheffield:
  - MultiMT - Multi-modal machine translation
  - Audio source location with microphone arrays
  - Identification of sleep apnoea
  - AVCOGHEAR - multi-modal hearing aid (vision + audio)

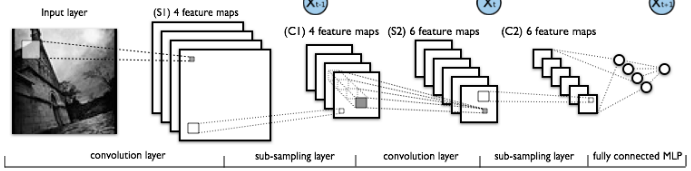
Why is it possible now?

Big Data: Sensor data, structured and unstructured text, images, audio, video, databases

Large training data set

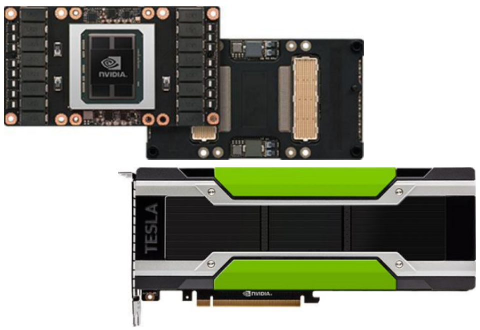
Convolution, LSTM





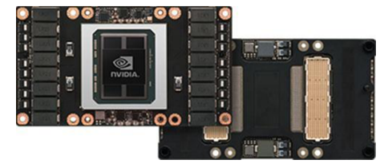
New Algorithms

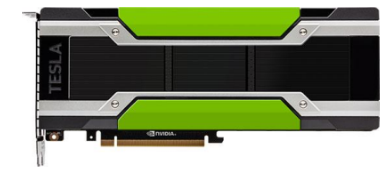
GPUs, TPUs, etc.



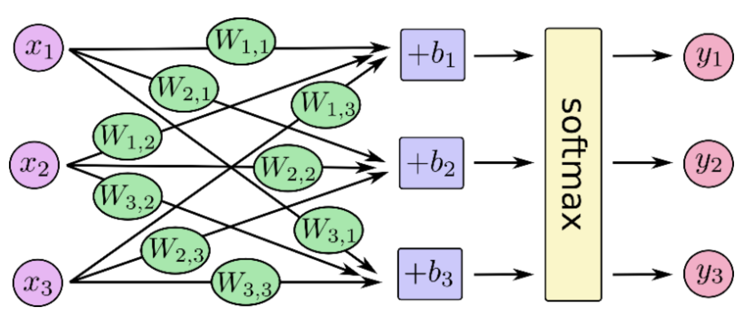
Hardware

Why use GPUs?





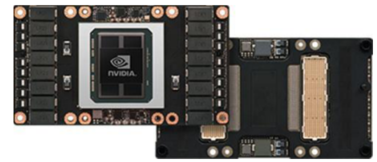
- GPUs have massively parallel architecture
- Designed for fast parallel floating point and matrix operations
- NNs are essentially large floating point and matrix multiplication problems

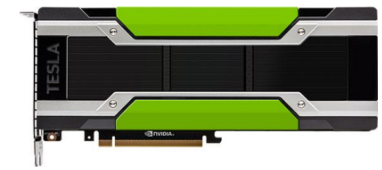


$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{softmax}\begin{pmatrix} W_{1,1}x_1 + W_{1,2}x_2 + W_{1,3}x_3 + b_1 \\ W_{2,1}x_1 + W_{2,2}x_2 + W_{2,3}x_3 + b_2 \\ W_{3,1}x_1 + W_{3,2}x_2 + W_{3,3}x_3 + b_3 \end{pmatrix}$$

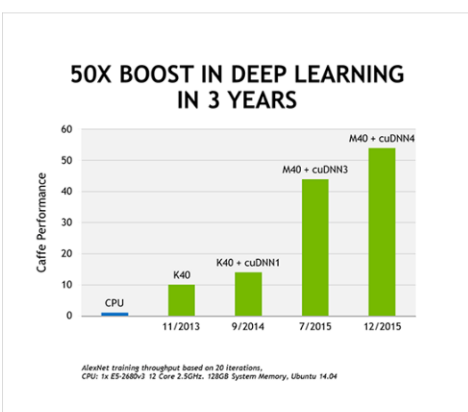
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{softmax}\left(\begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}\right)$$

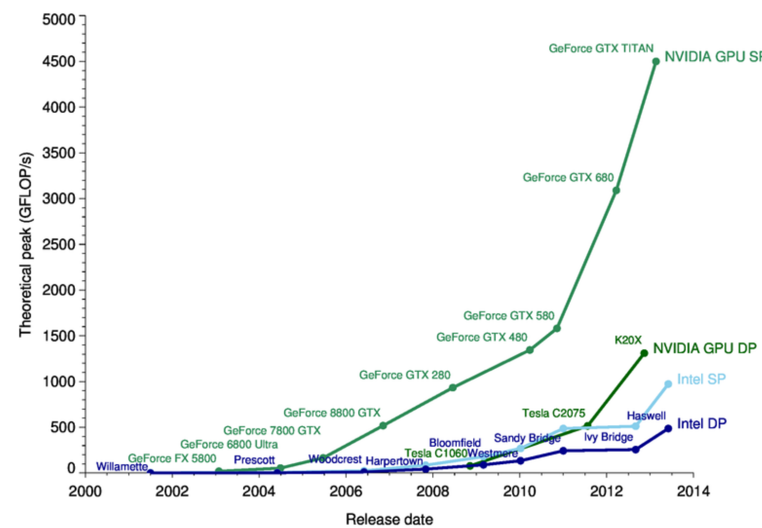
Why use GPUs?





- Even inexpensive consumer hardware can be used to massively speed up calculation
- Architectures now being optimised for NN
- GPU makers are creating low-level frameworks optimised for NN computation e.g. cuDNN
- Support in most DL packages





Why use Singularity?

- Rapid deployment of complex software stack
- Easy to share & test
- Reproducibility
- Avoid dependencies
- Wealth of pre-built images, especially from Docker
- We'd like to make a single image work for
    - Workstation
    - ShARC
    - JADE (Tier 2 HPC centre)
    - Cloud

Research Software Engineering Sheffield.

NVIDIA GPU RESEARCH CENTER

The University Of Sheffield.

Deployment: Global

- Modules
  - Software installed on public network drive
  - Module files sets the correct environment path
  - No root access, no package managers, installs must be isolated
  - Great for monolithic packages or licensed software e.g. Matlab
  - Very long turnaround - limited admin resources
    - Difficult to test with users

Deployment: Local

- Build to home directory from source
  - Provide build instructions/build scripts
  - Complex for new users
  - Install scripts can be brittle
  - Redundant build effort
- Anaconda (python virtual environment)
  - Many DL packages are Python-based
  - Enables installation of Python packages to user's home environment
  - Not good for mixing and matching compiled source and pre-built packages
  - Conda package can be created for C++ installs instead

Deep Learning Platforms & Frameworks

- Theano
- Tensorflow
- Caffe/Caffe2
- Torch/PyTorch
- MatConvNet
- Mxnet
- Deeplearning4j
- Chainer
- CNTK

Users require more than just the frameworks

- Combination of software
  - Tensorflow + emergence + Qt
  - Torch + OpenNMT
- Custom software/stack
  - Neurokernel - fruit fly brain
- As a web service
  - DIGITS
- Audio, video, image and text pre & post processing
  - E.g. OpenCV, SOX

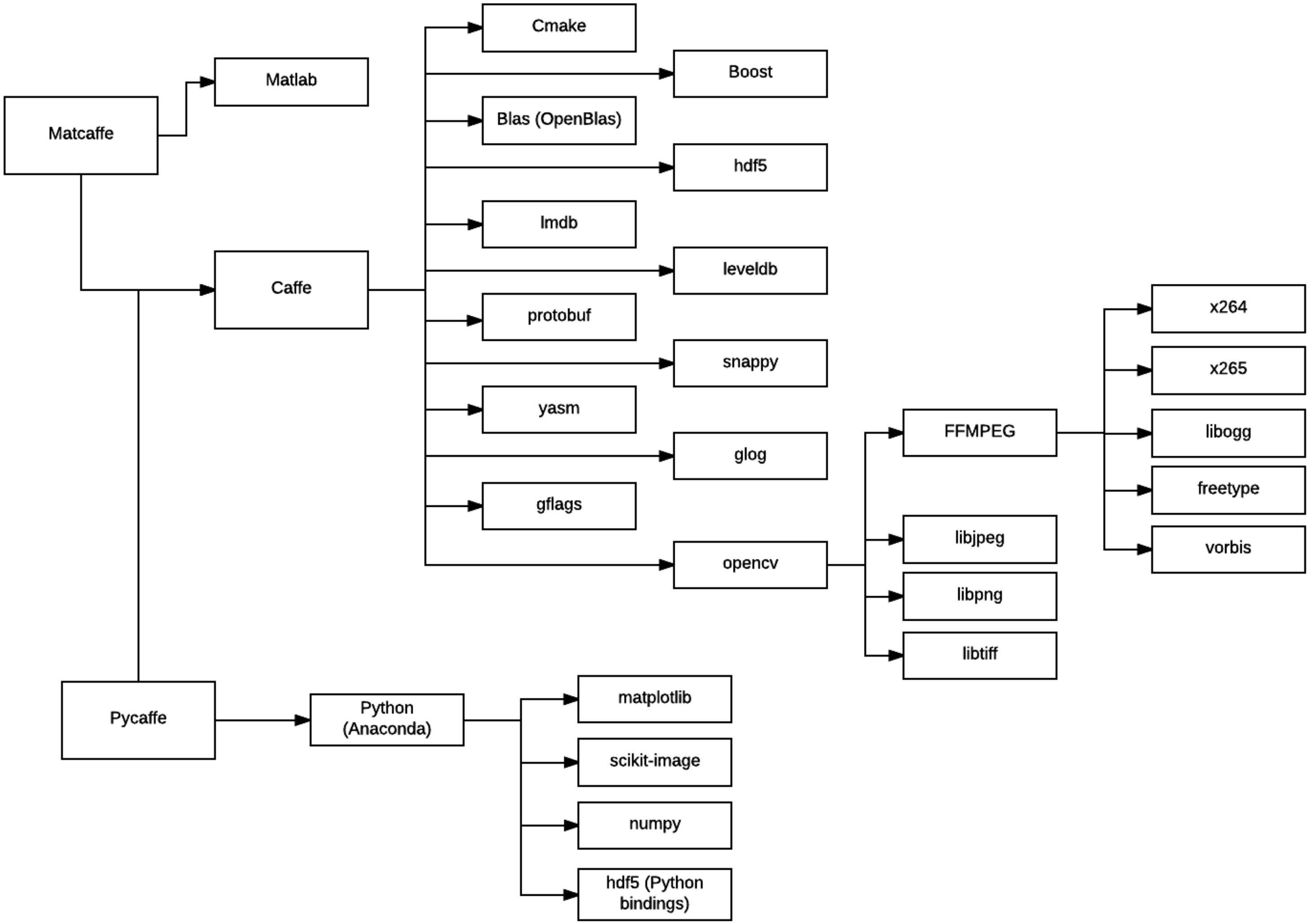
A Case Study: Deploying Caffe

- High-level deep learning package
- Great performance for training and inferencing, written in C++
- Used in production environments

A Case Study: Deploying Caffe

- Install brittle even with package managers
- Builds must be isolated, multiple versions are offered for repeatability of experiments and compatibility of code
- Caffe has > 15 Dependencies
- Module file creation/update
  - Update slow to update and refresh, dependent on sysadmins
  - Difficult to share modules for testing

A Case Study: Deploying Caffe

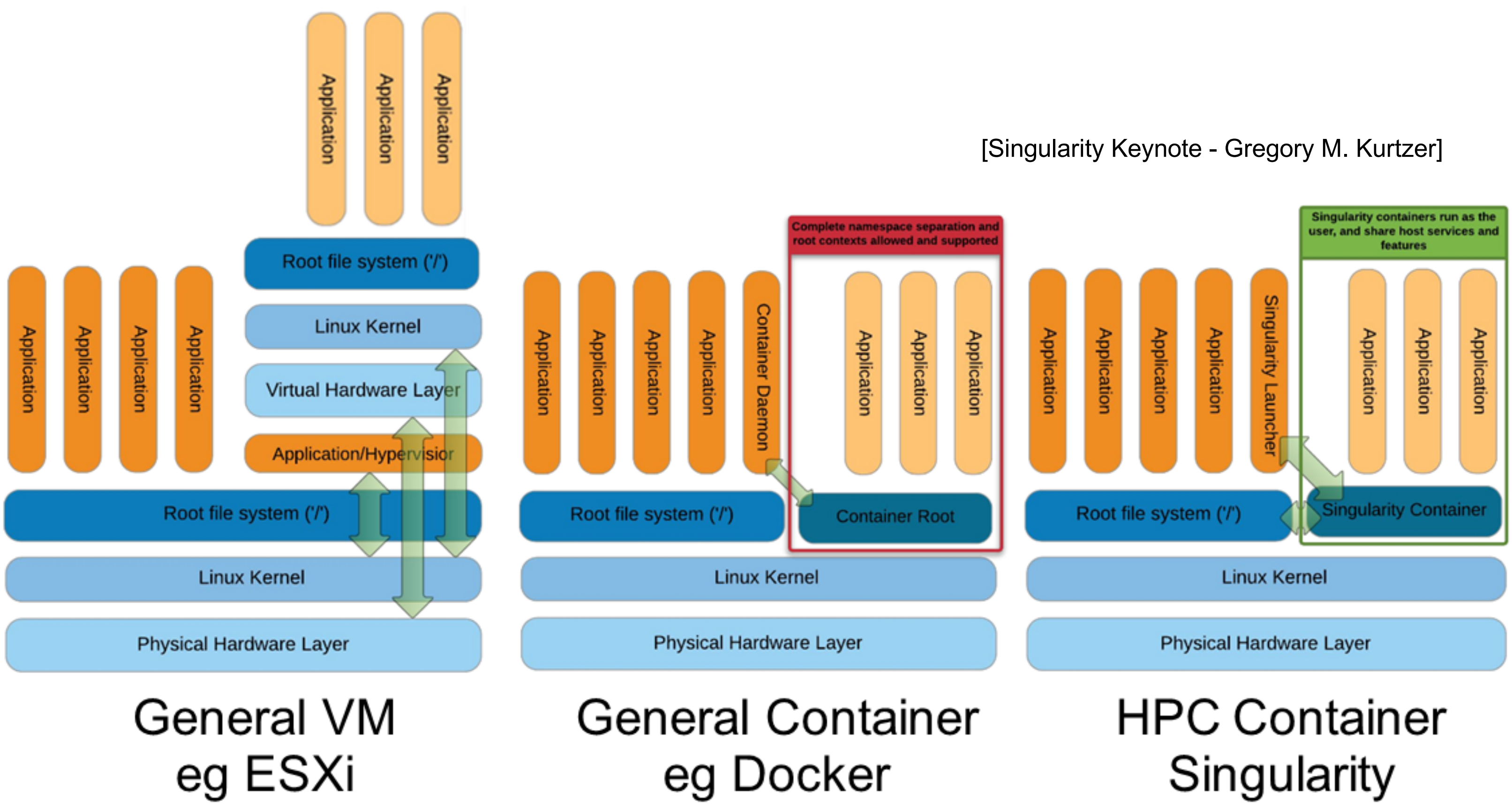


A Case Study: Deploying Caffe

- With Singularity:
  - Can pull pre-built image directly from Docker Hub
  - No difference in performance
  - Easier to share test images
  - Users can create own images or download images pre-built images/use provided definition files
  - Keep an index of images with associated metadata
    - Image ID, OS, available software, versions, etc.
  - But GPUs does not work out of the box (feature still experimental)

Singularity: Enabling GPUs in images

- Unlike VMs, Singularity uses kernel sharing



[Singularity Keynote - Gregory M. Kurtzer]

Singularity: Enabling GPUs in images

- This means:
  - GPU driver has to be installed on the host (kernel module + libraries and executables)
  - GPU driver files (libraries and executable) must also be accessible in the image

Singularity: Enabling GPUs in images

- Embedding files directly in the image makes it not portable
  - Requires all nodes to have same driver
  - When updating the driver on host, driver files must be updated in every image

Singularity: Enabling GPUs in images at Sheffield

- Host sets of supported driver files on a public network location, isolated from other lib files
- Every image created has additional folders (/nvlib and /nvbin) which is mounted to the correct driver files

In %post

```
echo 'export PATH="/nvbin:$PATH"' >> /environment
echo 'export LD LIBRARY_PATH="/nvlib:$LD_LIBRARY_PATH"'
   >> /environment
```

In the config file:

```
bind path = /mynvdriver/v367.56:/nvbin
bind path = /mynvdriver/v367.56:/nvlib
```

Research Software Engineering Sheffield.

NVIDIA GPU RESEARCH CENTER

The University Of Sheffield.

Singularity: Enabling GPUs in images at Sheffield

- Configuration file per node-GPU configuration
- The same approach can be used for MPI cluster that has Infiniband (OFED driver)
- And probably for other similar driver installs

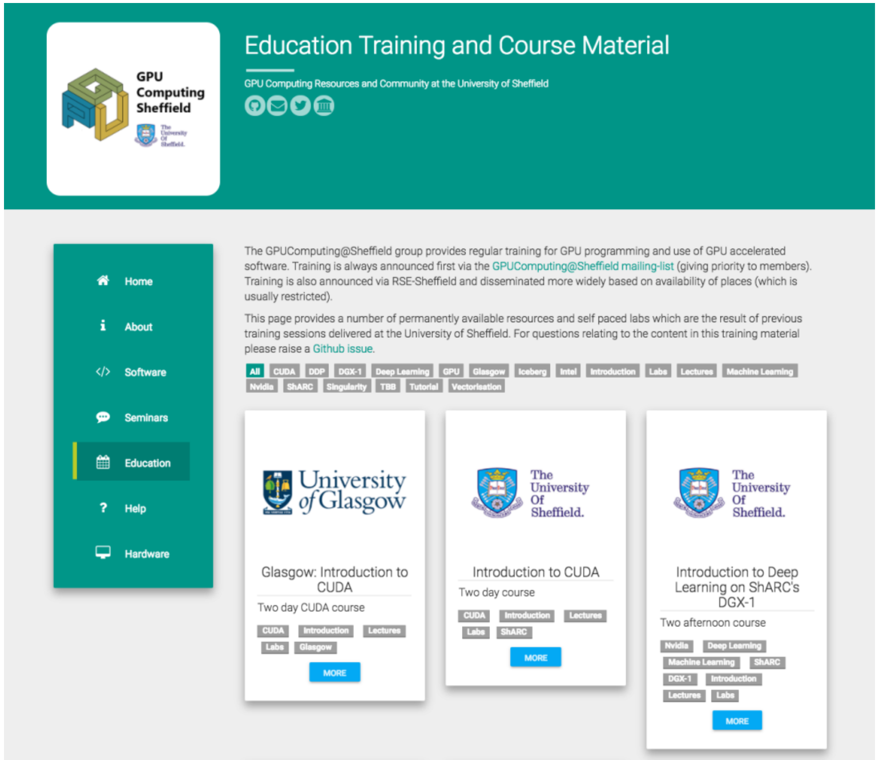
GPU Management: Ensuring Utilisation

- ShARC uses Son of Grid Engine (SoGE) scheduler
  - No GPU locking, everybody uses 0th GPU as default
  - No GPU utilisation monitoring

GPU Management: Ensuring Utilisation

- CUDA_VISIBLE_DEVICES env flag used to lock GPUs
  - Flag set outside Singularity image works inside it
- Prolog script
  - uses the proc interface to check GPU exist
  - creates a lock directory for each GPU requested
    - POSIX directory operation is atomic
- Epilog script unlocks the GPU(s) after a job is finished
- We're still working on utilisation monitoring, potentially using Nvidia DCGM

Tutorial for enabling GPUs on singularity images:
http://gpucomputing.shef.ac.uk/education/creating_gpu_singularity



Thank you!

Any questions?

Research Software Engineering Sheffield.

NVIDIA GPU RESEARCH CENTER

The University Of Sheffield.