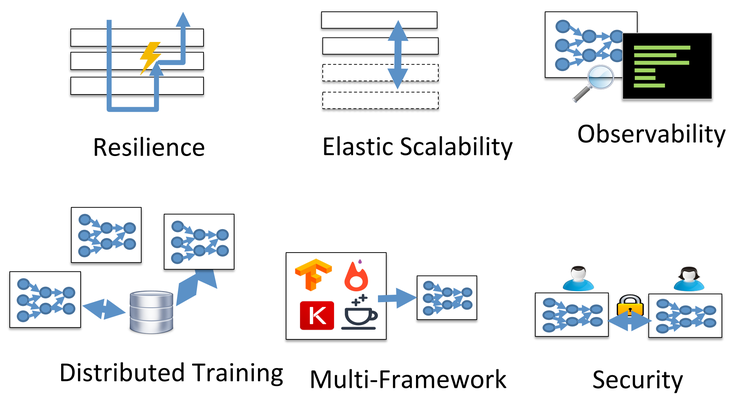
IBM Research **AI**

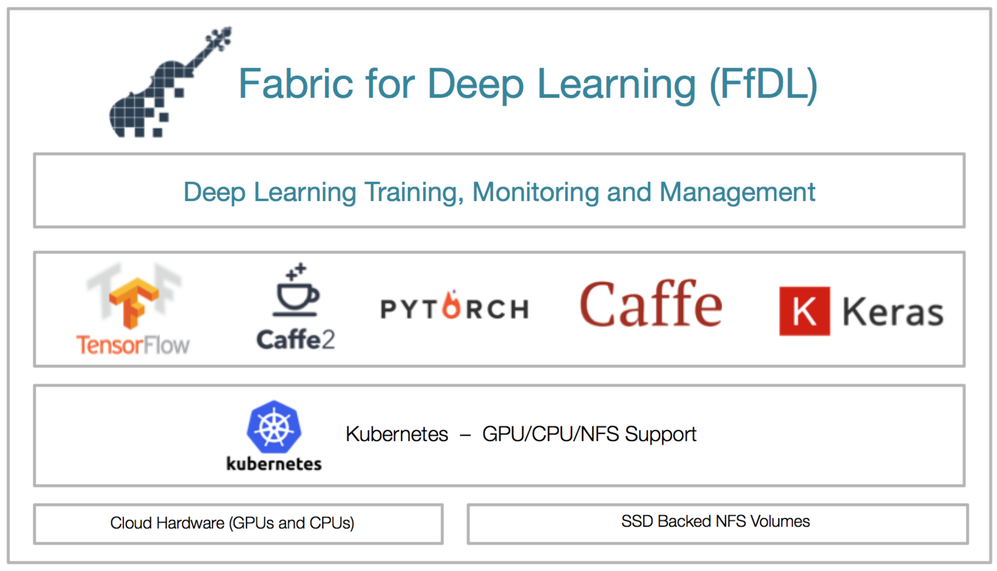
Impact of System Resources on Performance of Deep Neural Network

Parijat Dube and Zehra Sura

Training deep neural networks (DNNs) requires intensive resources for computation and memory/ storage. It is important to enable rapid development, experimentation, and testing of DNNs by improving the performance of these codes. This requires understanding what system resources are exercised by deep learning codes, to what degree the utilization of different resources is impacted by changes in the compute intensity or size of data being processed by the neural network, and the nature of the dependencies between different resource bottlenecks. We are performing an extensive empirical evaluation by varying execution parameters and running experiments with different configurations of DNN training jobs. The goal is to understand how to tailor system resources and training hyperparameters to the needs of a deep learning job, accounting for both the DNN model and dataset.



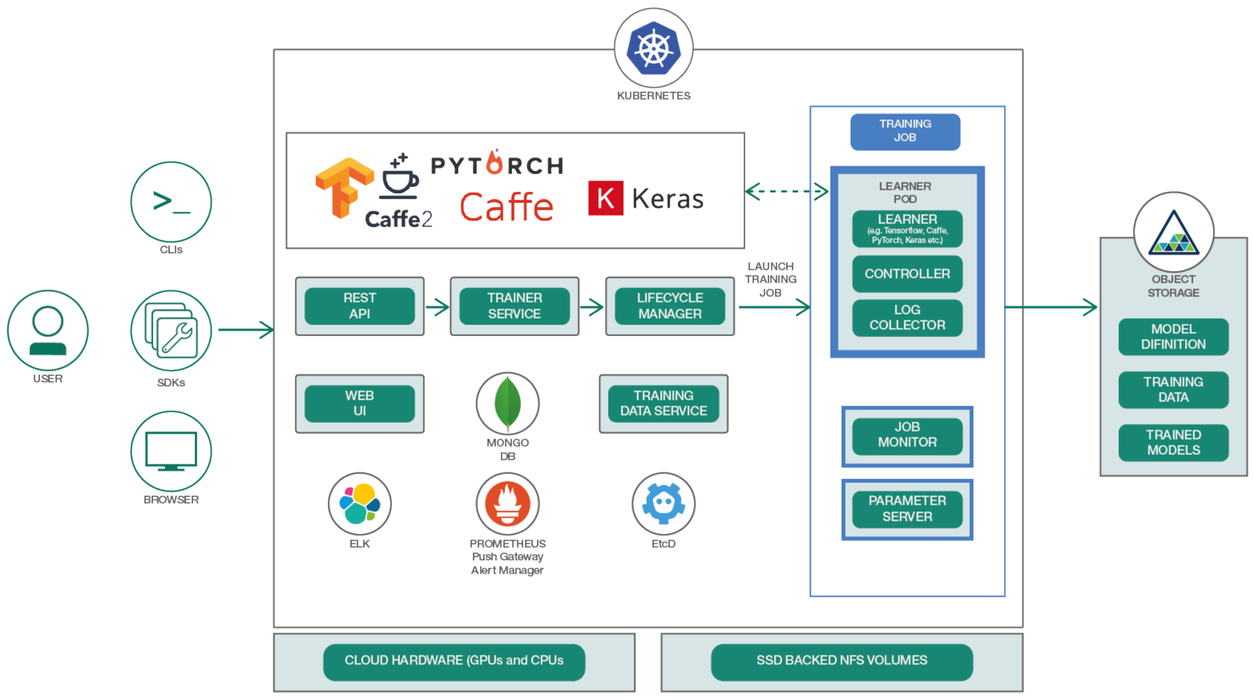
IBM Fabric for Deep Learning



https://developer.ibm.com/code/2018/03/20/fabric-for-deep-learning/

https://www.ibm.com/cloud/deep-learning

FfDL Architecture



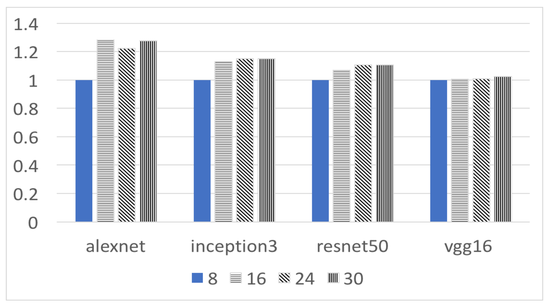
https://developer.ibm.com/code/2018/03/20/democratize-ai-with-fabric-for-deep-learning/

Factors affecting DNN Performance

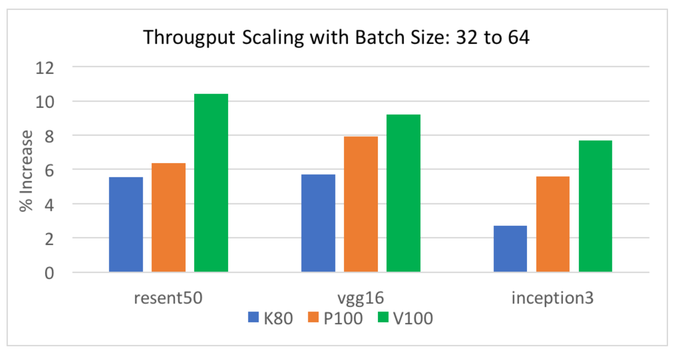
Parameter	Dimension	Example Choices
DNN Model	Model size Number of layers Neurons per layer Interconnection topology Compute intensity of functions	AlexNet Inception3 ResNet50 VGG16
Framework	NN libraries Inter-gpu communication Native distribution support	Tensorflow Pytorch Caffe/Caffe2
Dataset	Modality Size Encoding Training and testing datasets	Imagenet Cifar Places TREC
Batch Size	Number of data samples used in one iteration of the training job	32, 64, 218, 256, 512
Job Resources	Number of CPU threads Type of accelerators Number of accelerators	2, 4, 8, 16, 32, 64 NVIDIA K80, P100, V100 1, 2, 4, 8, 16, 32

Evaluation – Throughput Scaling

- **Scaling with number of CPU threads** on P100 cards:
  - 8 to 30 CPU threads: 1.28x (alexnet), 1.03x (vgg16)



- **Scaling with batch size**
  - Depends on GPU speed, memory, and DNN model size



Future Directions

- Expand configuration parameters for a thorough empirical evaluation
- Customize system resources and training hyperparameters for a DNN training job
- Account for the elastic environment of a shared cloud system:
  - Delays due to sharing
  - Delays due to resource failures
  - Opportunities due to additional resource availability
- Continuous feedback and control mechanism to interact with users and service modules