# Deploying Deep Learning Systems

Sujay N V
Deep Vision Group
Flipkart

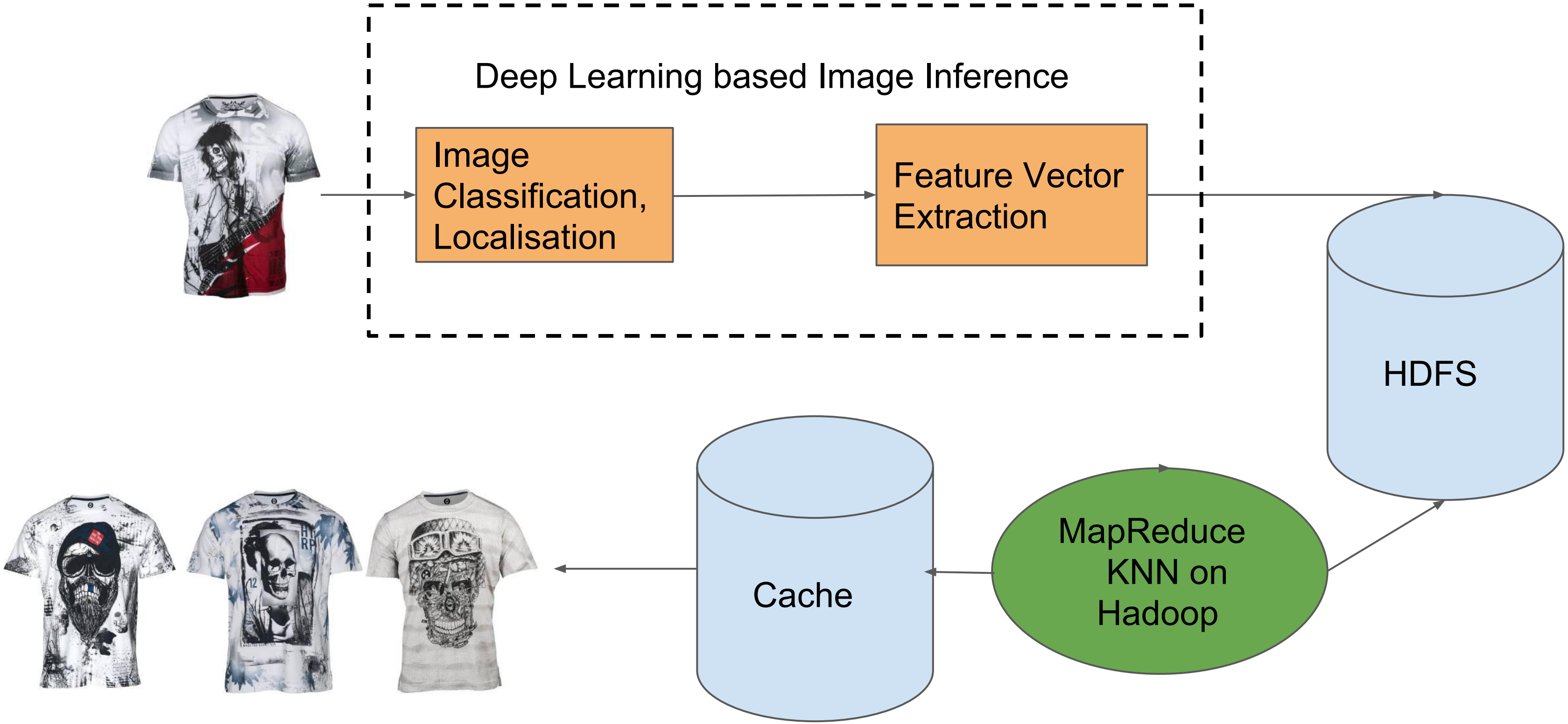
# Contents

- Overview of deployed Visual Similarity Engine
  - Scaling up Nearest Neighbour Search
  - Scaling up Deep Learning Inference across CPUs

- Training models in a distributed setup
  - Data Parallelism
  - Model Parallelism
  - HyperParameter Parallelism

- Distributed Training and **TensorFlow**

# Visual Similarity Engine - Batch Pipeline



Deep Learning based Image Inference

Image Classification, Localisation

Feature Vector Extraction

HDFS

MapReduce KNN on Hadoop

Cache

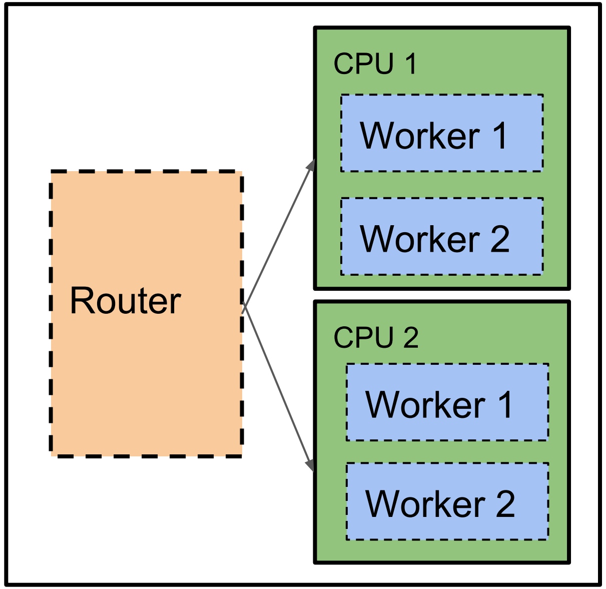
# Visual Similarity - Scaling up for real time applications

## KNN

- **Exact KNN Search** (Brute Force)
    - May not meet latency requirements
- **Approximate Nearest Neighbour Search**
    - Clustering
    - KD-Tree - Not suitable for very high dimensional vectors (4096)
    - Locality Sensitive Hashing - Drops accuracy by 10-12 %
    - Deep Hash - Learning hash functions

## Image Inference

- Ideal to have GPUs, can make do with CPUs

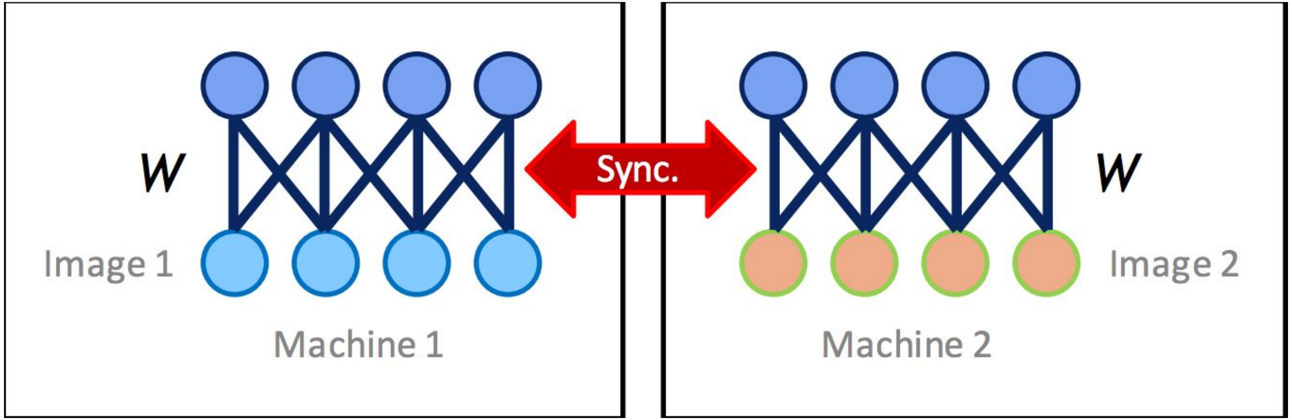


Router - Worker configuration

# Training Deep Learning models

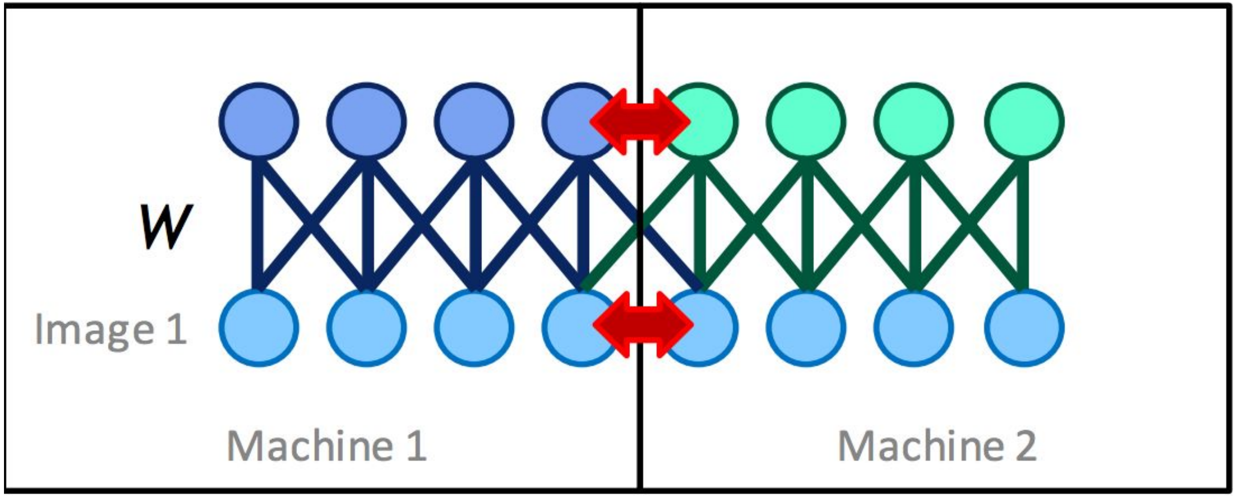
- **Use GPUs** !! (days worth of CPU effort can be achieved in hours on GPU)
- Most models fit into GPU memory for common applications - Single GPU machine suffices
- What if the model does not fit into memory?
- How do you leverage multiple GPUs?

# Multi GPU Training - Data Parallelism



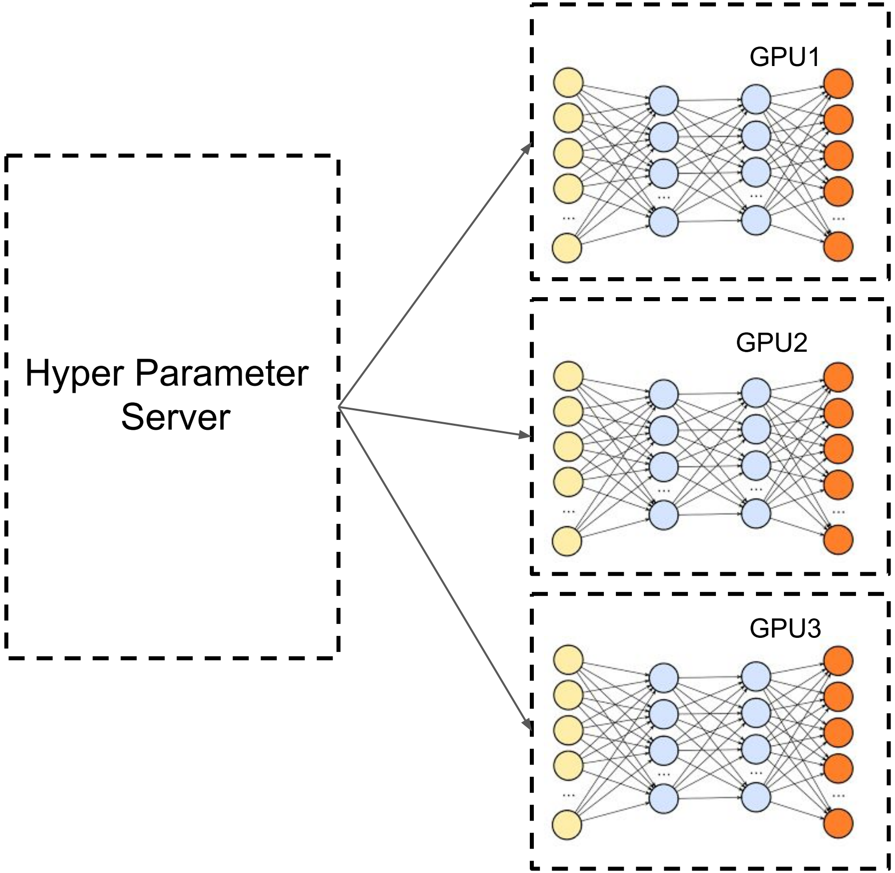
- Different machines receive different batches of data
- Parameters to be synced every iteration (Parameter Server approach)
- Network is the main blocker

# Multi GPU Training - Model Parallelism



- Model distributed across machines
- Useful for models that cannot fit into a single machine
- More frequent communication between GPUs, but a lot less data transfer
- Network is still a major constraint, however better than data parallelism

# HyperParameter Parallelism



- Easiest way to parallelise
- Run different instances of the same model with different hyperparameters

# Distributed Training

- DistBelief (Google’s internal framework)
- Project Adam (Microsoft)
- TensorFlow !
    - Learnt from shortcomings of DistBelief
    - Open Source !
    - Platform independent  - Model can be seamlessly be deployed on GPU, CPU and Mobile
- The Future
    - Tensor Processing Units (TPU - geared for deep learning, support TensorFlow)
    - NVLink, Infiniband - Improve network communication between GPU machines

# THANK YOU