Hewlett Packard Enterprise

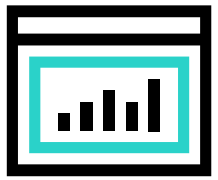
Hewlett Packard Labs

HPE Deep Learning Cookbook:

Recipes to Run Deep Learning Workloads

Natalia Vassilieva, Sergey Serebryakov

Deep learning ecosystem today



Software

Caffe

theano

torch

KALDI

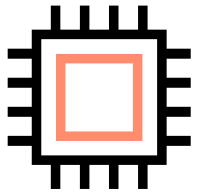
Caffe2

Microsoft CNTK

mxnet

TensorFlow

Chainer



Hardware

NVIDIA

intel

ARM

AMD

Google

HPE's portfolio for deep learning

Government, academia and industries

Financial services

Government and academia

Life Sciences, Health

Autonomous vehicles / Mfg.

HPE POINTNEXT

Advisory, professional and operational services, HPE Flexible Capacity, HPE Datacenter Care for Hyperscale

Compute ideal for training models in data center

HPE SGI 8600

Petaflop scale for deep learning and HPC

HPE Apollo 6500

The enterprise bridge to accelerated computing

HPE Apollo sx40

Maximize GPU capacity and performance with lower TCO

Compute for both training models and inference at edge

HPE Apollo 2000

The bridge to enterprise scale-out architecture

Edge analytics and inference engine

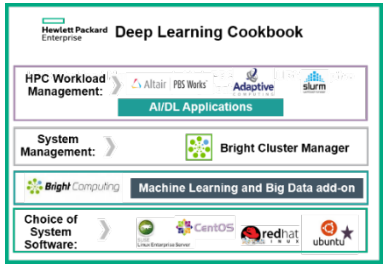
HPE Edgeline EL4000

Unprecedented deep edge compute and high capacity storage; open standards

AI Software Framework

Easy Setup and Flexible OS

Using Bright Computing's distribution of deep learning software development components and workload management tool integration



HPC Storage

HPE Apollo 4520

HPC Data Management Framework Software

Large-scale, storage virtualization & tiered data management platform

Choice of Fabrics

Arista Networking

- Intel® Omni-Path Architecture
- Mellanox InfiniBand
- HPE FlexFabric Network

How to pick the right hardware/software stack?

One size does NOT fit all

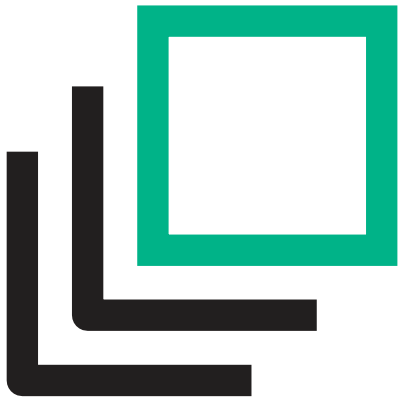
Application

Data type

Data size

Model (topology of artificial neural network):

- How many layers
- How many neurons per layer
- Connections between neurons (types of layers)



Popular models

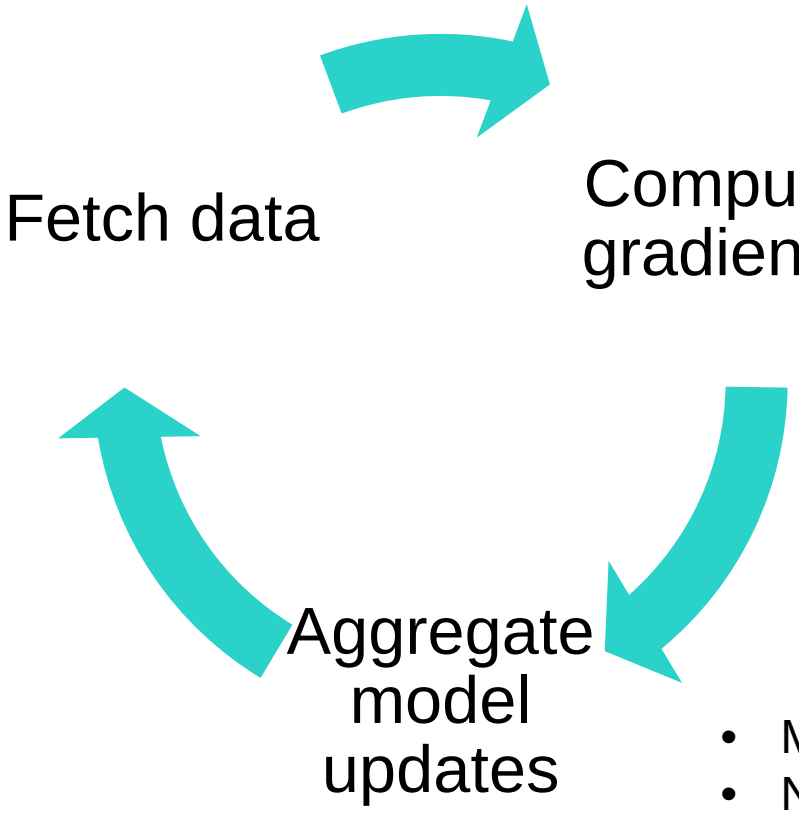
Name	Type	Model size (# params)	Model size (MB)	GFLOPs (forward pass)
AlexNet	CNN	60,965,224	233 MB	0.7
GoogleNet	CNN	6,998,552	27 MB	1.6
VGG-16	CNN	138,357,544	528 MB	15.5
VGG-19	CNN	143,667,240	548 MB	19.6
ResNet50	CNN	25,610,269	98 MB	3.9
ResNet101	CNN	44,654,608	170 MB	7.6
ResNet152	CNN	60,344,387	230 MB	11.3
Eng Acoustic Model	RNN	34,678,784	132 MB	0.035
TextCNN	CNN	151,690	0.6 MB	0.009

Popular models

Name	Type	Model size (# params)	Model size (MB)	GFLOPs (forward pass)
AlexNet	**CNN**	**60,965,224**	**233 MB**	**0.7**
GoogleNet	CNN	6,998,552	27 MB	1.6
VGG-16	CNN	138,357,544	528 MB	15.5
VGG-19	CNN	143,667,240	548 MB	19.6
ResNet50	CNN	25,610,269	98 MB	3.9
ResNet101	CNN	44,654,608	170 MB	7.6
ResNet152	**CNN**	**60,344,387**	**230 MB**	**11.3**
Eng Acoustic Model	RNN	34,678,784	132 MB	0.035
TextCNN	CNN	151,690	0.6 MB	0.009

Distributed training with data parallelism

- Mini-batch size
- IO bandwidth & latency
- Pre-processing on a fly

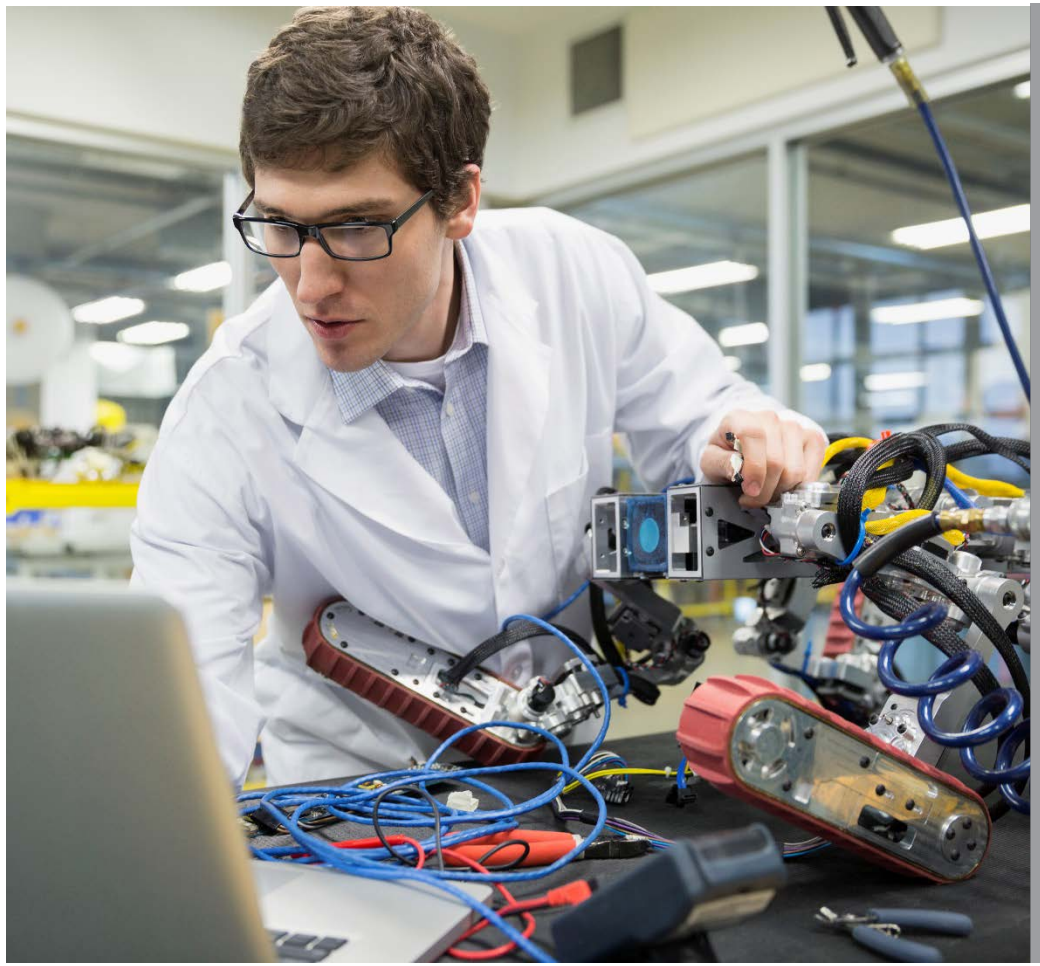


- Replica batch size
- Forward/backward computational complexity
- Computational power of a node

- Model size
- Number of workers/nodes
- Interconnect bandwidth & latency

HPE Deep Learning Cookbook

Overview



Comprehensive set of tools to guide the choice of the best hardware and software environment for different deep learning workloads.

- Eliminate the “guesswork” with Deep Learning Performance Guide: tap into a massive pool of performance data to reason on optimal hardware configuration for your workload
- Validate the configuration of your hardware and software environment with Deep Learning Benchmarking Suite
- Get started fast with one of our Reference Designs as a default technology recipe

https://developer.hpe.com/platform/hpe-deep-learning-cookbook/home

HPE Deep Learning Cookbook

Main components

HPE Deep Learning Benchmarking Suite

Automated benchmarking tool to collect performance of different deep learning workloads on various hardware and software configurations.

Available on GitHub

HPE Deep Learning Performance Guide

NEW

A web-based tool to guide a choice of optimal hardware and software configuration via analysis of collected performance data and applying performance models.

To be released in 2018
Will be hosted by HPE

Reference Designs

Reference hardware/software stacks for particular classes of deep learning workloads.

Image Classification
Reference Designs released

HPE Deep Learning Benchmarking Suite

HPE Deep Learning Benchmarking Suite

A tool to benchmark various DL frameworks and models. To be open sourced in November 2017

- 8 frameworks, 1 inference runtime
  - TensorFlow, BVLC/NVIDIA/Intel Caffe, Caffe2, MXNet, PyTorch, TensorRT
- 18 models (AlexNet, GoogleNet, ResNets, VGGs …)
- Host / docker benchmarks
- Docker files for all supported frameworks / SW combinations

Resource monitoring:

- GPU/CPU/memory utilization

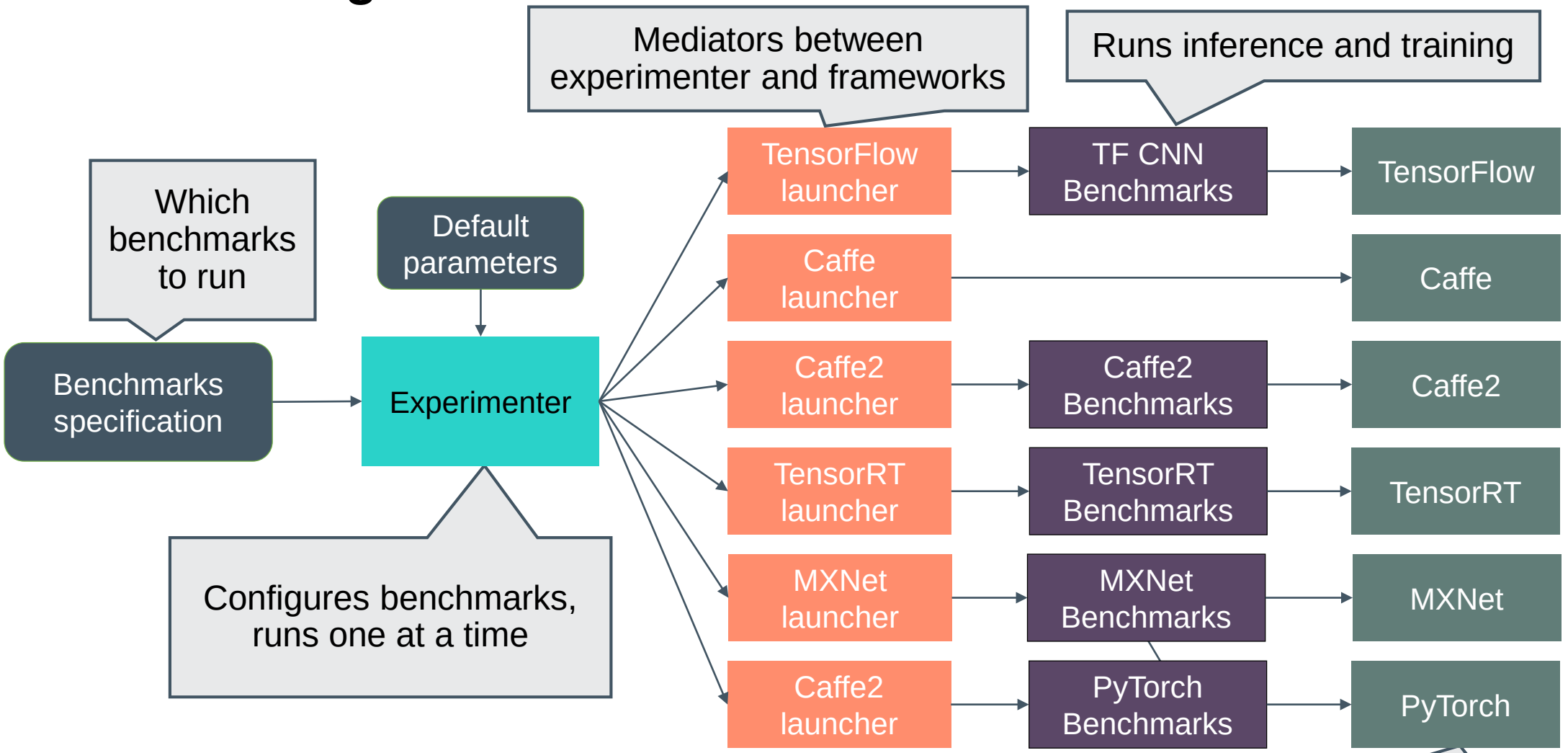
Report builders:

- Weak/strong scaling
- Benchmark statistics and charts

More information:

- GitHub https://github.com/HewlettPackard/dlcookbook-dlbs
- Documentation https://hewlettpackard.github.io/dlcookbook-dlbs

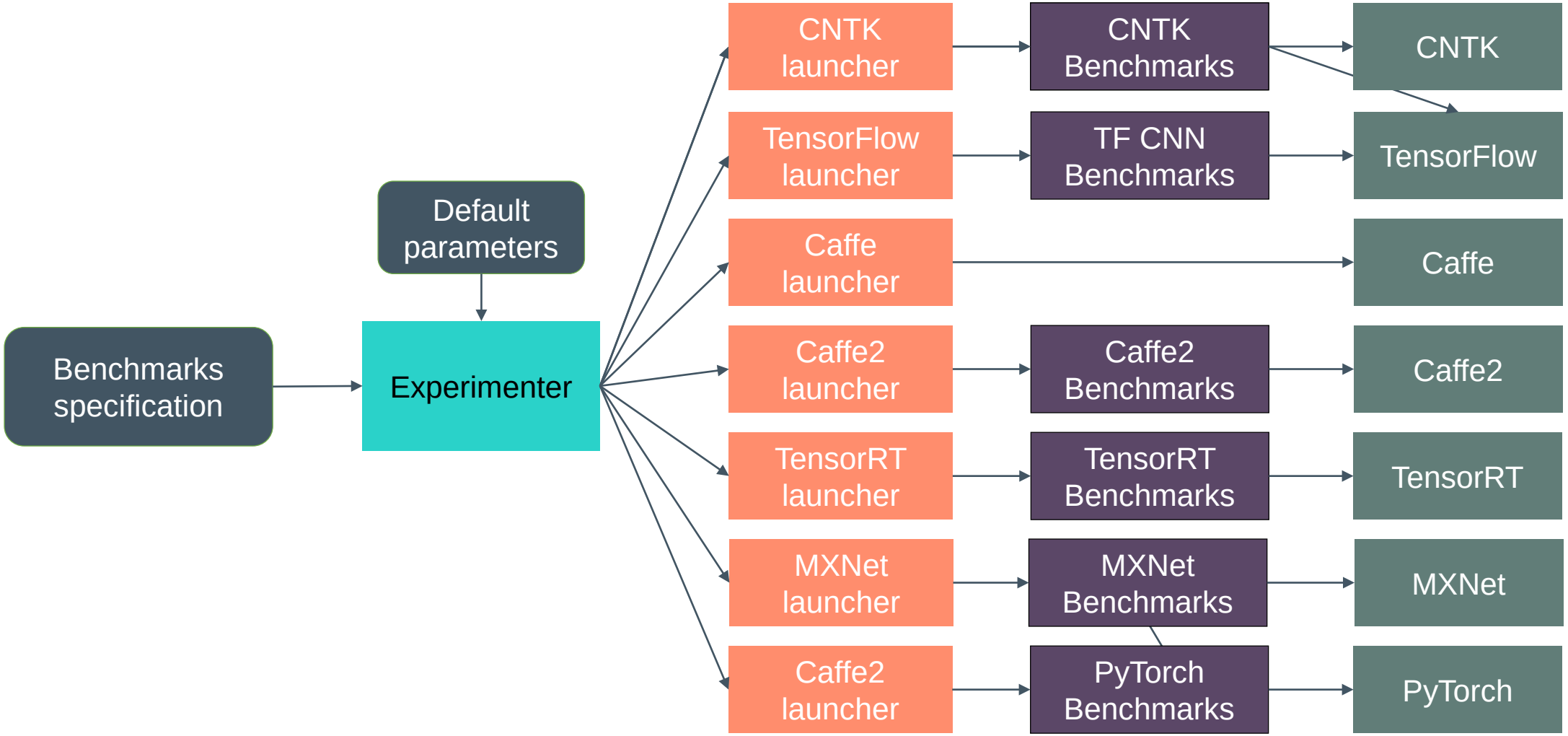
Benchmarking Suite Architecture



GitHub: https://github.com/HewlettPackard/dlcookbook-dlbs

Documentation: https://hewlettpackard.github.io/dlcookbook-dlbs

How to expand



GitHub: https://github.com/HewlettPackard/dlcookbook-dlbs

Documentation: https://hewlettpackard.github.io/dlcookbook-dlbs

Quick Start

```
git clone https://github.hpe.com/labs/dlcookbook.git        # Install benchmarking suite
cd ./dlcookbook/docker && ./build tensorflow/cuda8-cudnn6   # Build TensorFlow docker image
cd .. && export PYTHONPATH=$(pwd)/python:$PYTHONPATH        # Setup python paths
python ./python/dlbs/experimenter.py run \                  # Benchmark ...
       -Pexp.framework='"tensorflow"' \                     #    TensorFlow framework
       -Vexp.model='["resnet50", "alexnet"]' \              #    with ResNet50 and AlexNet models
       -Vexp.gpus='["0", "0,1", "0,1,2,3"]' \               #    run on 1, 2 and 4 GPUs
       -Pexp.log_file='"${HOME}/${exp.id}.log"'             #    and write results to these files
python ./python/dlbs/logparser.py ${HOME}/*.log             # Parse log files and print summary
```

HPE Deep Learning Performance Guide

HPE Deep Learning Performance Guide

Hewlett Packard Labs

Choose configurations | New | Add new +

Close

Axis

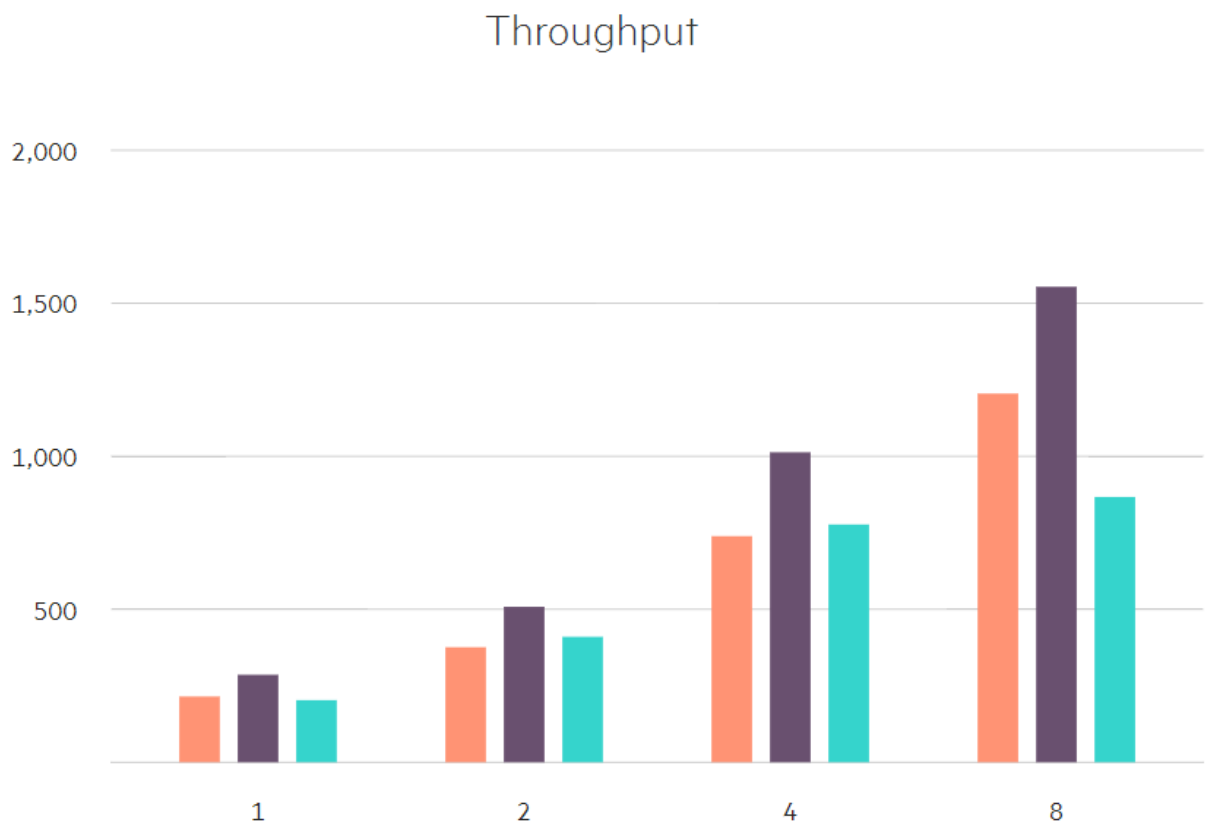
Vertical axis variable	Horizontal axis variable	Axis range
Throughput	Num processors	1, 2, 4, 8

Bar Chart | Line Chart | Table

Data

(InceptionV3)	Model Hadware -	Data Real	Framework TensorFlow	Batch size -	Choose as base
(ResNet50)	Model Hadware -	Data Real	Framework TensorFlow	Batch size -	Choose as base
(VGG16)	Model Hadware -	Data Real	Framework TensorFlow	Batch size -	Choose as base

Add series | Clear series | Share

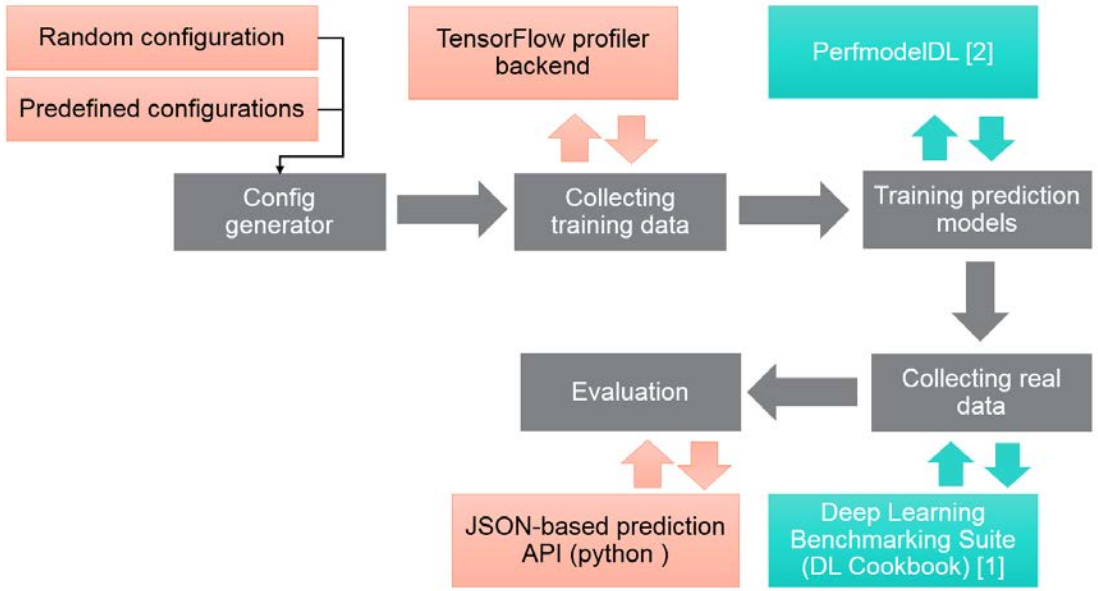


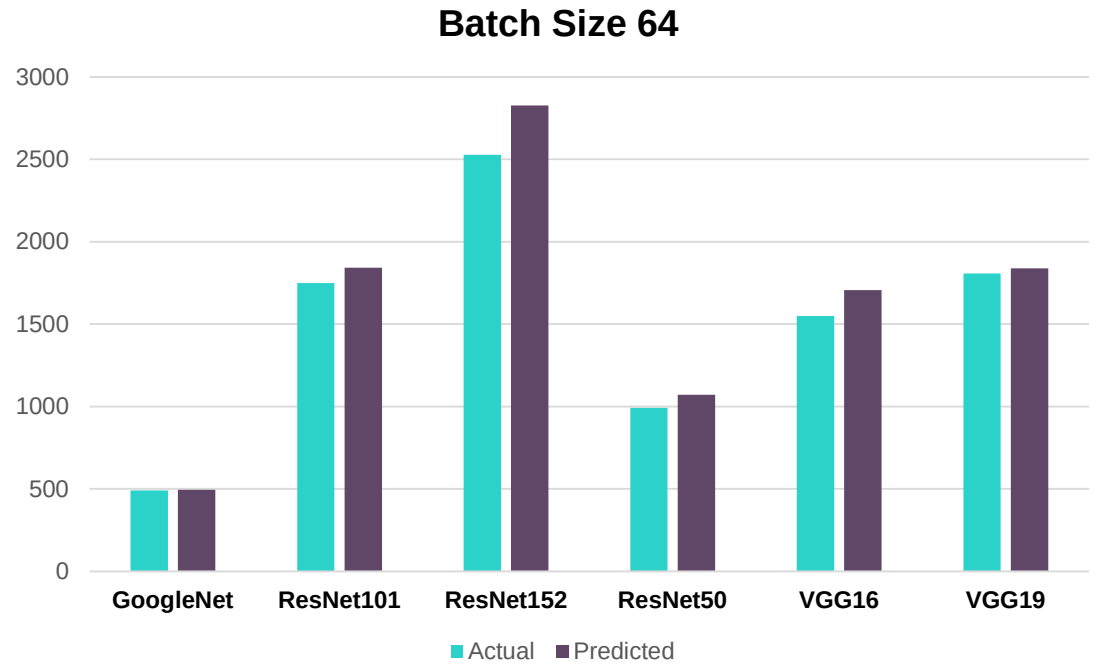
Demo

Performance Models (TensorFlow)

Performance and scalability models for deep learning workloads in different hardware and software environments.

- Computation models are based on machine learning (Support Vector Regression)
- Communication models are based on analytical models





Hewlett Packard Enterprise

Reference Designs

Reference Designs

A reference hardware/software stack for a particular class of workloads

- Image classification
- Natural Language Processing
- Speech Recognition
- Video Analytics

Image Classification Reference Design:

"Get Started" Configuration

Platform	HPE ProLiant DL380 Gen10 Server
CPU	2 x Intel® Xeon® Gold 6128 Processor 3.40 GHz 6 19.25 MB 115W
Memory	>= 64 GB RAM
GPUs	2 x NVIDIA® Tesla® P100 (1:1 configuration)
Storage	1.6TB Hot Plug SFF SAS SSD, read intensive
Framework	TensorFlow
Deployment	HPE AI Solution for Rapid Software Installation with Bright Computing
Performance Results (Images/sec)1	GoogleNet: 1036 ; InceptionV3: 273 ResNet152: 173 ; ResNet50: 441 VGG16: 287 ; VGG19: 245

1 Results with Real data (ImageNet)

"Scale-up" Configuration

Platform	HPE Apollo 6500 Chassis and Power Shelf HPE ProLiant XL270d Gen9 system tray
CPU	2 x Intel® Xeon® Processor E5-2680V4, 35MB Cache, 2.40 GHz, 14 cores
Memory	256 GB RAM: 16x16GB DDR4
GPUs	8 x NVIDIA® Tesla® P100 (8:1 configuration)
Storage	1.6TB Hot Plug SFF SAS SSD, read intensive
Framework	TensorFlow
Deployment	HPE AI Solution for Rapid Software Installation with Bright Computing
Performance Results (Images/sec)1	GoogleNet: 1370 ; InceptionV3: 870 ResNet152: 588 ; ResNet50: 1291 VGG16: 904 ; VGG19: 792

1 Results with Real data (ImageNet)

Thank you

Natalia Vassilieva
nvassilieva@hpe.com

Sergey Serebryakov
sergey.serebryakov@hpe.com

HPE Deep Learning Cookbook
https://developer.hpe.com/platform/hpe-deep-learning-cookbook/home