# Demystifying Hardware Infrastructure Choices for Deep Learning Using MLPerf

DELL EMC

Ramesh Radhakrishnan

THE UNIVERSITY OF TEXAS AT AUSTIN
UT ECE
ELECTRICAL & COMPUTER ENGINEERING

Lizy Kurian John

Snehil Verma

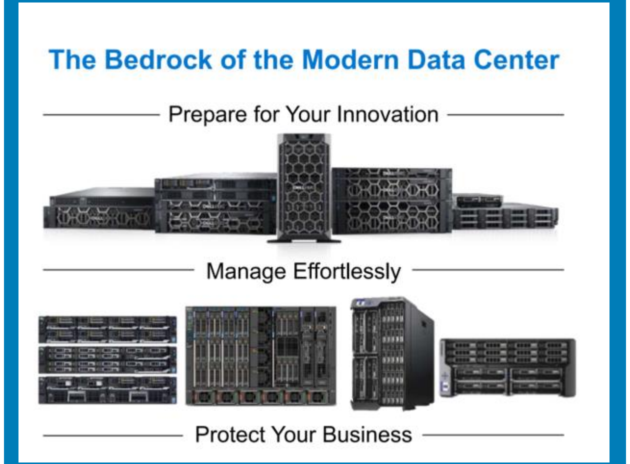
Qinzhe Wu

Bagus Hanindhito

UTSA Engineering

Gunjan Jha

Eugene John




# Server industry trends

The need to simplify and reduce costs driving the
**SOFTWARE DEFINED DATA CENTER**

---

Innovations in memory | Disk | GPU | FPGA
**WORKLOAD ACCELERATION**

## 6X growth in AI

By 2020, 20% of the enterprise infrastructures deployed will be used for AI. Up from 3% in 2017.

---

**ML/DL**

Machine learning/Deep learning emerging to provide better business insight

## 40X growth in edge computing

40% of large enterprises will be integrating edge computing principles into their IT projects by 2021. Up from less than 1% in 2017.

Stats from Gartner

# Infrastructure Options for Deep Learning

| | TITAN | QUADRO | TESLA |
|---|---|---|---|
| Arch/Generation | ← PASCAL / VOLTA / TURING → | | |
| Form Factor | PCIe | | PCIe/SXM2 |
| Capability | CORES \| MEMORY \| FP PRECISION \| MANAGEMENT S/W | | |
| Interconnect | PCIe / NVLink Bridge | | PCIe / NVLink |
| System Design | PCIe Domain \| PCIe Switch \| NVLink Topology | | |
| Multi-GPU Scaling | INTERCONNECT LATENCIES & BANDWITH \| GPU DIRECT P2P | | |

multi-gpu jobs

***We use MLPerf Benchmark suite to quantify performance impact of GPU & System technology choices***

# The Evolution of Deep Learning Benchmarks

**DeepBench**

Benchmark basic operations for DNN

**TF_CNN_Bench**

TensorFlow

**DAWNBench**

Deep Learning Training and Inference Competition

Focused on narrow domain

Uses throughput as metric (ignoring accuracy)

No governing body

synthetic data

**MLPerf**

Accelerate innovation in DL hardware, systems, software & algorithms

MLPerf

Coverage of different DL domains

Improved metrics – Time & Accuracy

Reproducibility of results

Representation from Industry and Academia

MLPerf enables fair comparison of competing systems yet encourages innovation to improve the state-of-the-art of ML

# MLPerf Benchmark v0.5

**IMAGE CLASSIFICATION**
RESNET-50

**LANGUAGE TRANSLATION**
RNN GNMT
Transformer

**RECOMMENDATION**
NCF

**OBJECT DETECTION**
SSD
MASK-RCNN

**REINFORCEMENT LEARNING**
MINI GO

GPU platforms used in initial submission – *8 & 16 GPU Tesla V100-SXM2 NVLink* Platform

Limited conclusions can be drawn about GPU technology choices

Submissions included container build files, data sets and tuning parameters used in the run

## Systems Evaluated

# Dell Precision and Dell EMC GPU Optimized Portfolio

**Precision 5820**
**Quadro GV100 (2)**
1CPU, 2xGV100-PCIe

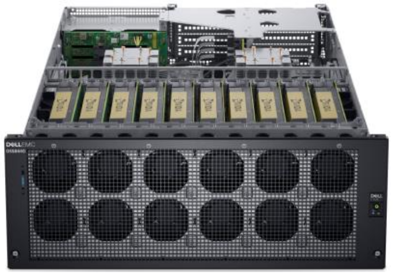
**PowerEdge T640**
**Tesla V100 (4)**
2CPU, 4xV100-PCIe

**PowerEdge C4140**
**Config B: Tesla V100 (4)** 2CPU, 4xV100-PCIe (PCIe Switch)
**Config K: Tesla V100 SXM2 (4)** 2CPU, 4xV100-SXM2 NVL (PCIe Switch)
**Config M: Tesla V100 SXM2 (4)** 2CPU, 4xV100-SXM2 NVL
*NVLink*

**PowerEdge R740**
**Tesla V100 (3)**
2CPU, 3xV100-PCIe

**PowerEdge 940XA**
**Tesla V100 (4)**
4CPU, 4xV100-PCIe

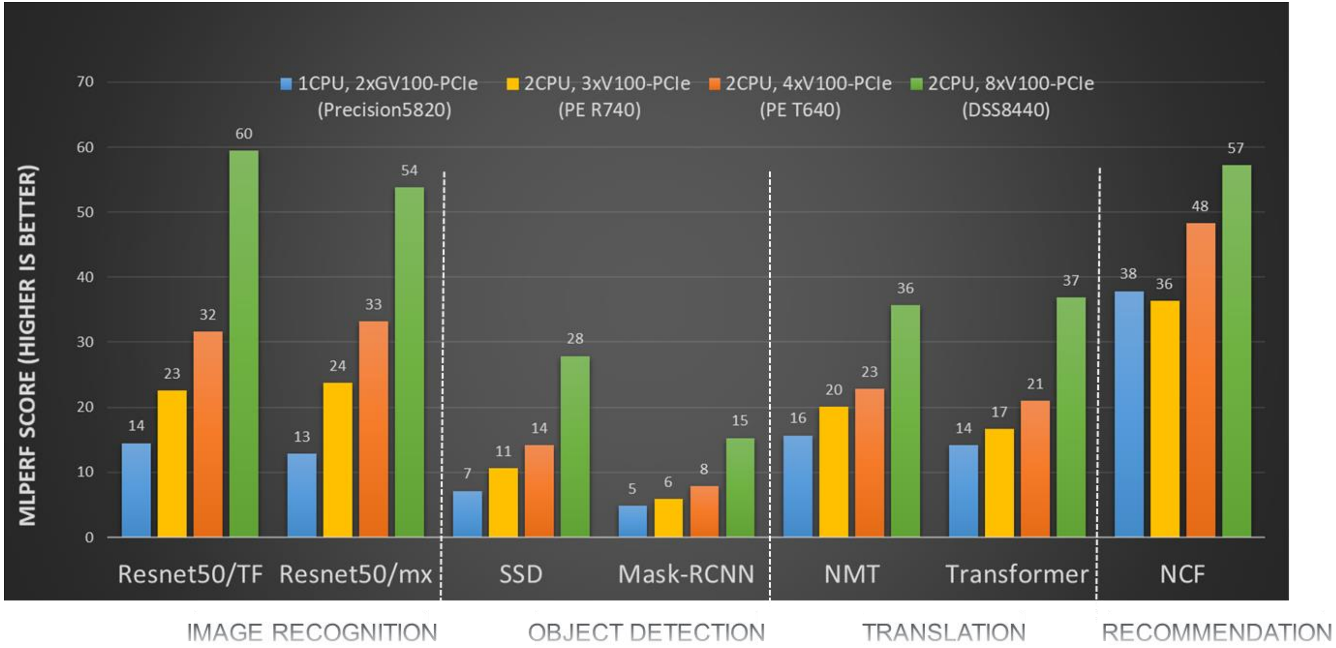
**Dell DSS8440**
**Tesla V100 (4/8/10)**
2CPUs, 8xV100-PCIe
(PCIe Switch)



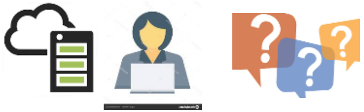
# BENCHMARKING

## *MLPerf Scores – Dell Technologies Portfolio (2GPU/3GPU/4GPU/8GPU)*

*Score = Speedup relative to a Pascal P100*




GPUs to train a DNN model in a single work day? In 4 hours? 2 hours!



I like flexibility of PCIe GPUs. What is the performance difference in training times between a PCIe and NVLink system?



Benchmark result not verified by MLPerf. MLPerf name and logo are trademarks. See www.mlperf.org for more information

# Impact of GPU Features and System Design

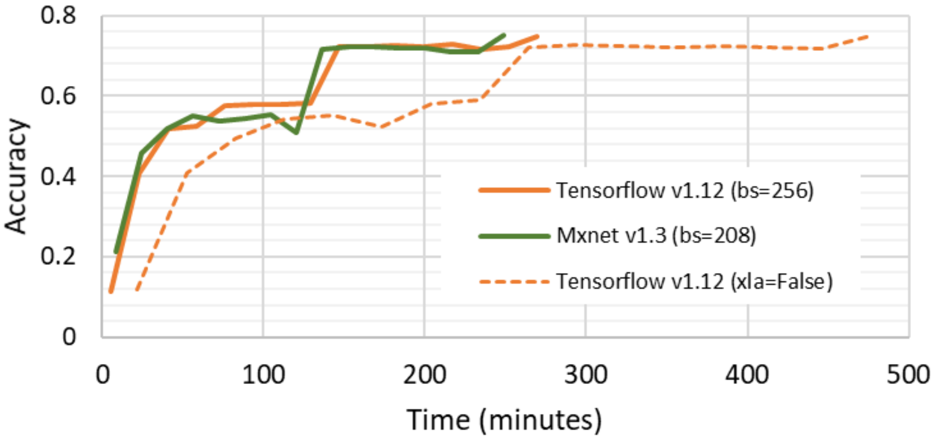
- **Workload Characterization**
  - Time to Accuracy plots
  - Roofline Analysis
  - Batch Size vs. Accuracy
  - Framework Performance
- **GPU Comparison**
  - Titan, Quadro, Tesla
  - GPU Memory: 16GB vs. 32GB
  - Clock Speed: SXM2 vs. PCIe
- **System Profiling**
  - CPU and GPU utilization trends
  - GPU Interconnect Utilization
- **GPU Scaling**
  - 2 GPU Workstation vs. 4 GPU Server vs. 8 GPU Server
  - 1 to 8 GPU Scaling within a single Server
- **GPU Interconnect Topology**
  - CPU PCIe Root Complex vs. PCIe Switch configuration
  - NVLink vs. PCIe topologies

## Workload Characterization

## Framework Comparison

*4xV100-SXM2 16GB (NVLink)*

Image classification - Resnet50




| Image Classification | Number of epochs | Average time per epoch (min) |
|---|---|---|
| TensorFlow v1.12 (Google) | 61 | 4.42 |
| | | |
| Mxnet v1.3.0 (Nvidia) | 62 | 4.01 |

| GPU Kernels (Common) |
|---|
| volta_fp16_s884cudnn_fp16_128x128_ldg8_relu_f2f_exp_interior_nhwc_tn_v1 |
| volta_s884cudnn_fp16_64x64_sliced1x4_ldg8_wgrad_idx_exp_interior_nhwc_nt |
| volta_fp16_s884cudnn_fp16_128x128_ldg8_dgrad_f2f_exp_small_nhwc_tt_v1 |
| volta_fp16_s884cudnn_fp16_128x128_ldg8_relu_f2f_exp_small_nhwc_tn_v1 |
| volta_s884cudnn_fp16_128x128_ldg8_wgrad_idx_exp_interior_nhwc_nt |
| dgrad_1x1_stride_2x2 |
| Volta_fp16_s884cudnn_fp16_256x64_ldg8_relu_f2f_exp_small_nhwc_tn_v1 |

8 GPU kernels shared (CuDNN v7.4)

*~34% execution time in Mxnet;*

*~45% execution time in TensorFlow*

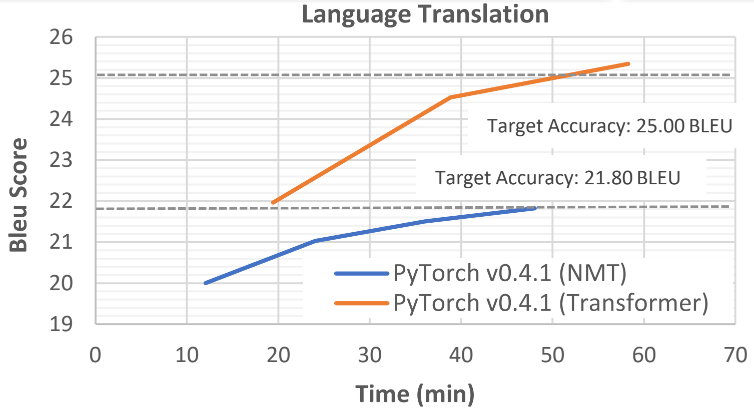
*TensorFlow and Mxnet take advantage of optimized DNN primitives available in CuDNN (profiling shows 8 CuDNN kernels that are common across the 2 runs)*

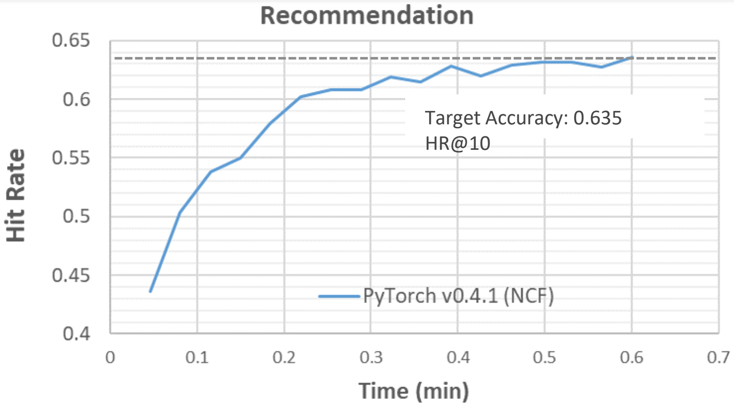
*XLA Just-In-Time Compile is critical to get performance on par with Mxnet and other frameworks*

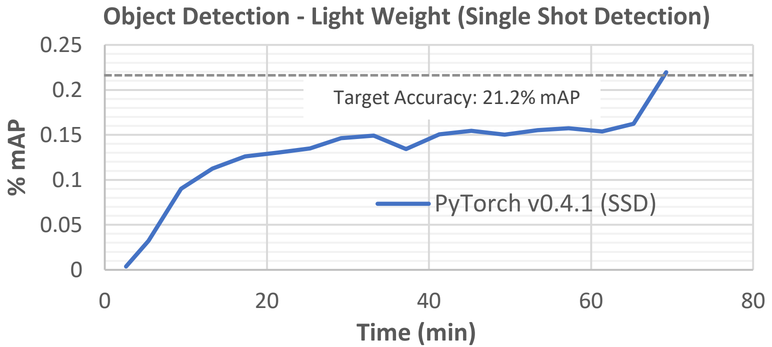
# Workload Characterization

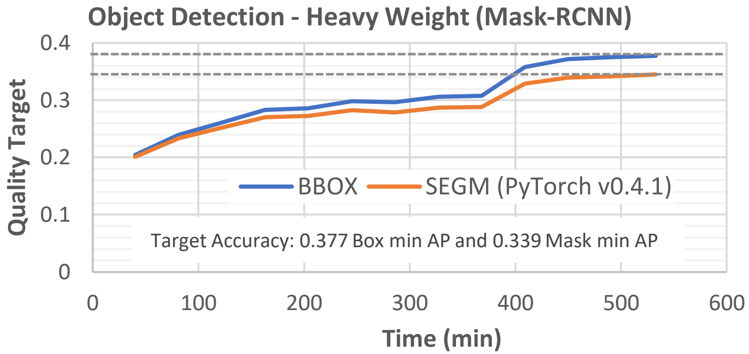
## Time to Accuracy plot

*4xV100-SXM2 16GB (NVLink)*



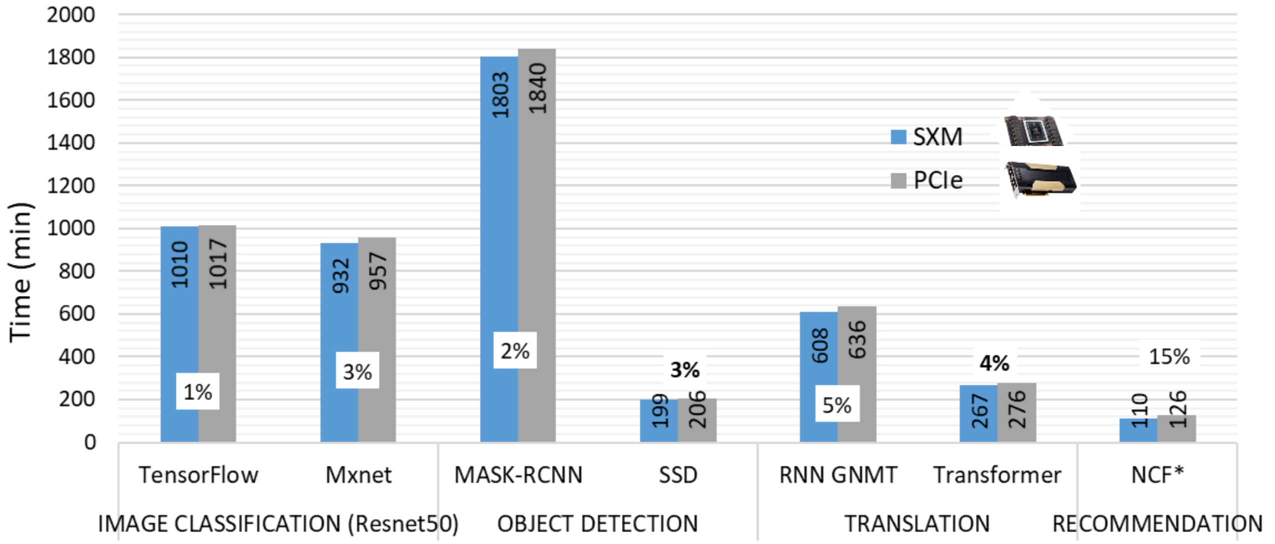

*Training times vary from less than 1 minute (NCF) to 9 hours for Mask-RCNN on a 4 GPU server*
*All models train in under a work day (8 hours) on a 4 GPU NVLink system*

# GPU Comparison

## Tesla V100: PCIe vs. SXM2

*1xV100-16GB*

| | Tesla V100 PCIe | Tesla V100 SXM2 |
|---|---|---|
| GPU Architecture | **NVIDIA Volta** | |
| NVIDIA Tensor Cores | **640** | |
| NVIDIA CUDA® Cores | **5,120** | |
| Double-Precision Performance | **7 TFLOPS** | **7.8 TFLOPS** |
| Single-Precision Performance | **14 TFLOPS** | **15.7 TFLOPS** |
| Tensor Performance | **112 TFLOPS** | **125 TFLOPS** |
| GPU Memory | **32GB /16GB HBM2** | |
| Memory Bandwidth | **900GB/sec** | |
| ECC | **Yes** | |
| Interconnect Bandwidth | **32GB/sec** | **300GB/sec** |
| System Interface | **PCIe Gen3** | **NVIDIA NVLink** |
| Form Factor | **PCIe Full Height/Length** | **SXM2** |
| Max Power Comsumption | **250 W** | **300 W** |
| Thermal Solution | **Passive** | |
| Compute APIs | **CUDA, DirectCompute, OpenCL™, OpenACC** | |



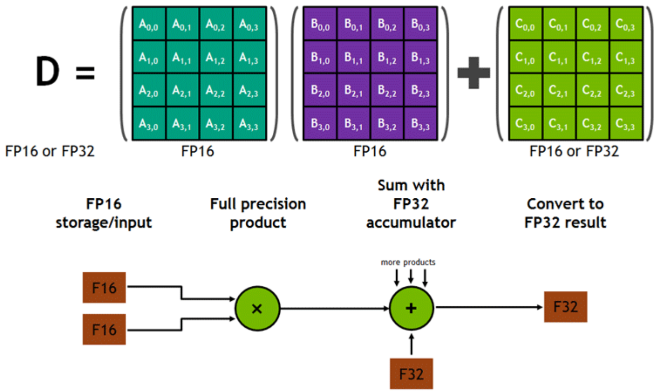

* time is in seconds

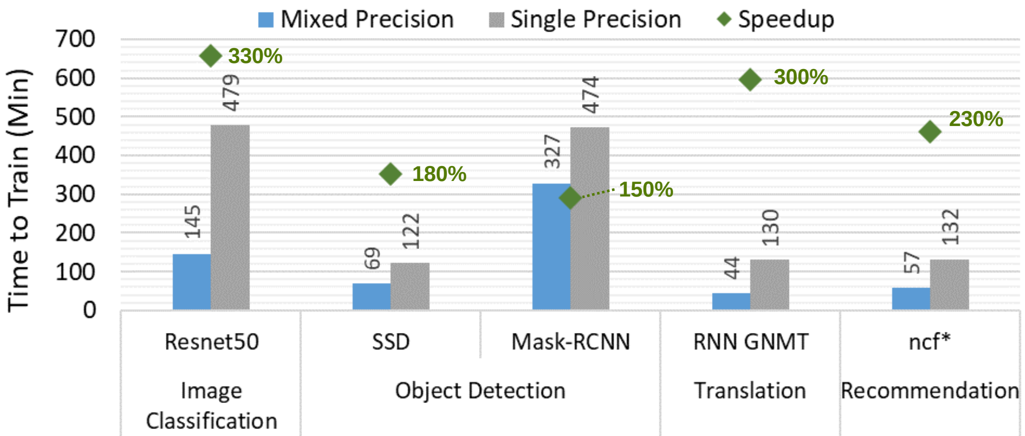
*1-5% speedup for single GPU training jobs*
*30 minutes on a 30 hour training job*

# GPU Comparison

## Tesla V100-PCIe: Single vs. Mixed Precision Training

- Training method that uses different numerical precisions (FP16 & FP32)
- Decrease Memory consumption (2x)
- Reduce training  & inference times by using WMMA (tensor cores)




**NVIDIA Deep Learning SDK**
https://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html

* time is in seconds

**Automatic Mixed Precision (AMP)** NGC 19.03 release
https://developer.nvidia.com/automatic-mixed-precision

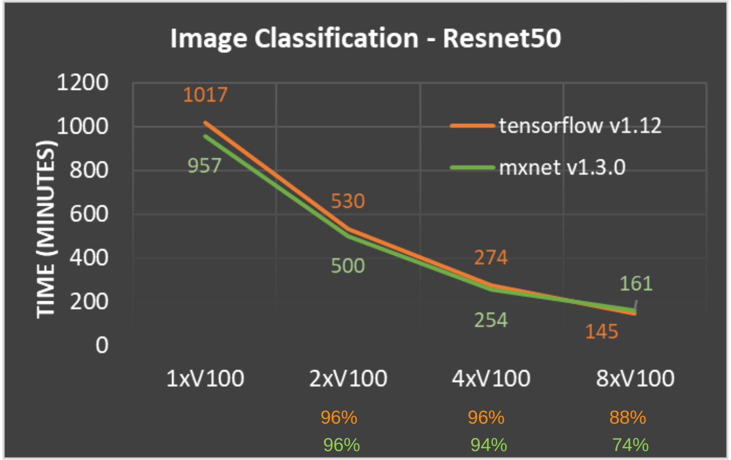
new

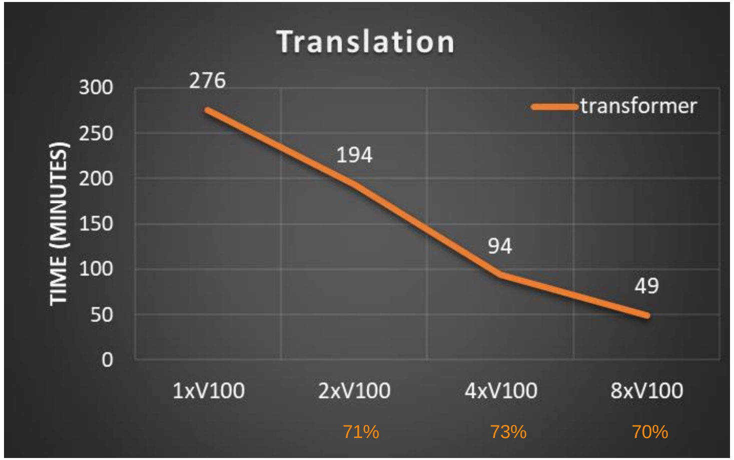
*150-330% speedup across benchmarks tested*
*330 minutes reduction for Resnet50 (70% reduction in training time)*

# GPU Scaling

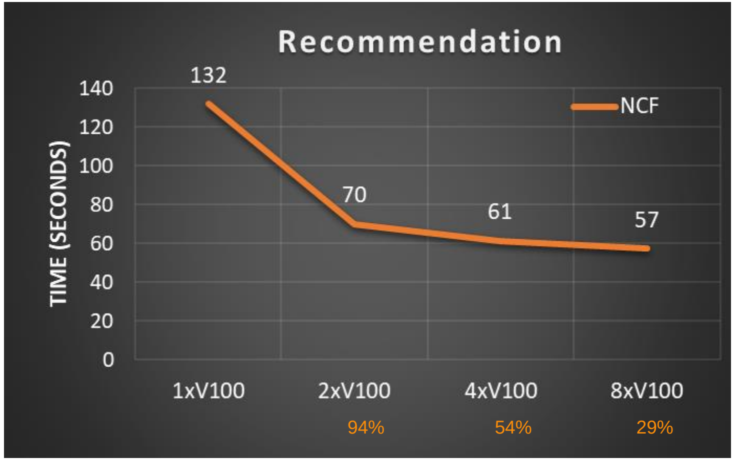
## *1 to 8 GPU Scaling*

*8xV100-PCIe 16GB Server, DSS8440*








*Scaling efficiency is shown with 1 GPU as the baseline*

*At 8 GPUs, scaling efficiency is over 80% for Resnet50 (TF) & SSD*

*At 4 GPUs, Resnet50 (TF & Mx) and SSD exhibit scaling efficiency over 80%*

*At 2 GPUs, Resnet50 (TF & Mx), SSD, Mask-RCNN and NCF exhibit scaling efficiency over 80%*

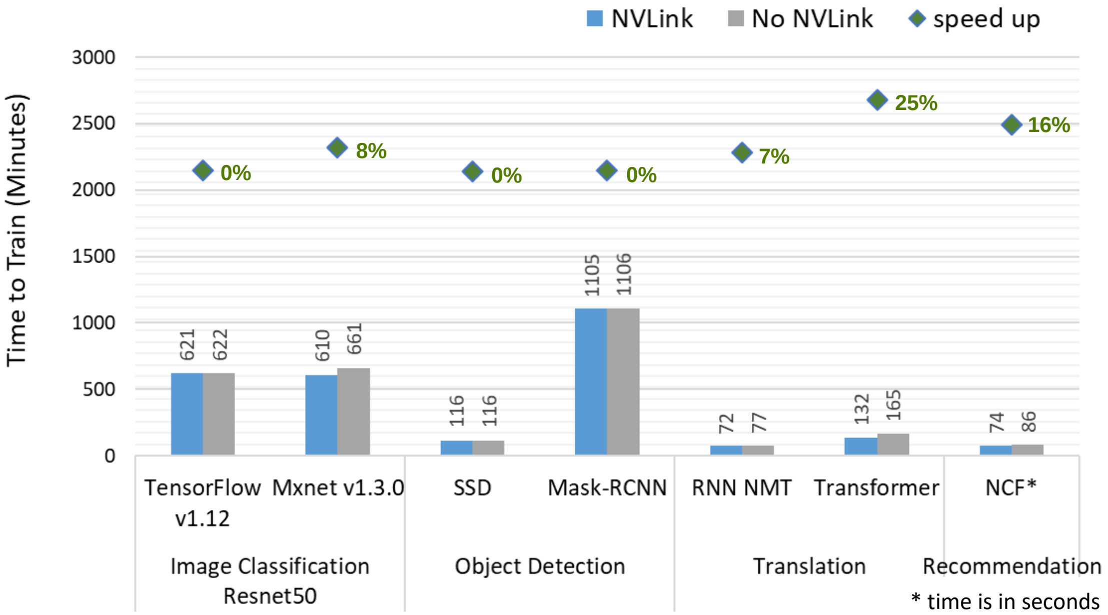
# GPU Interconnect Topology

## *NVLink Bridge*

*2xGV100-PCIe 32GB, Workstation 5820*

*nvidia-smi topo -m*

| | GPU0 | GPU1 | CPU Affinity |
|---|---|---|---|
| GPU0 | X | NV4 | 0-35 |
| GPU1 | NV4 | X | 0-35 |




*For a 2 GPU training job, the performance gains from NVLink ranges from 0%-25%*

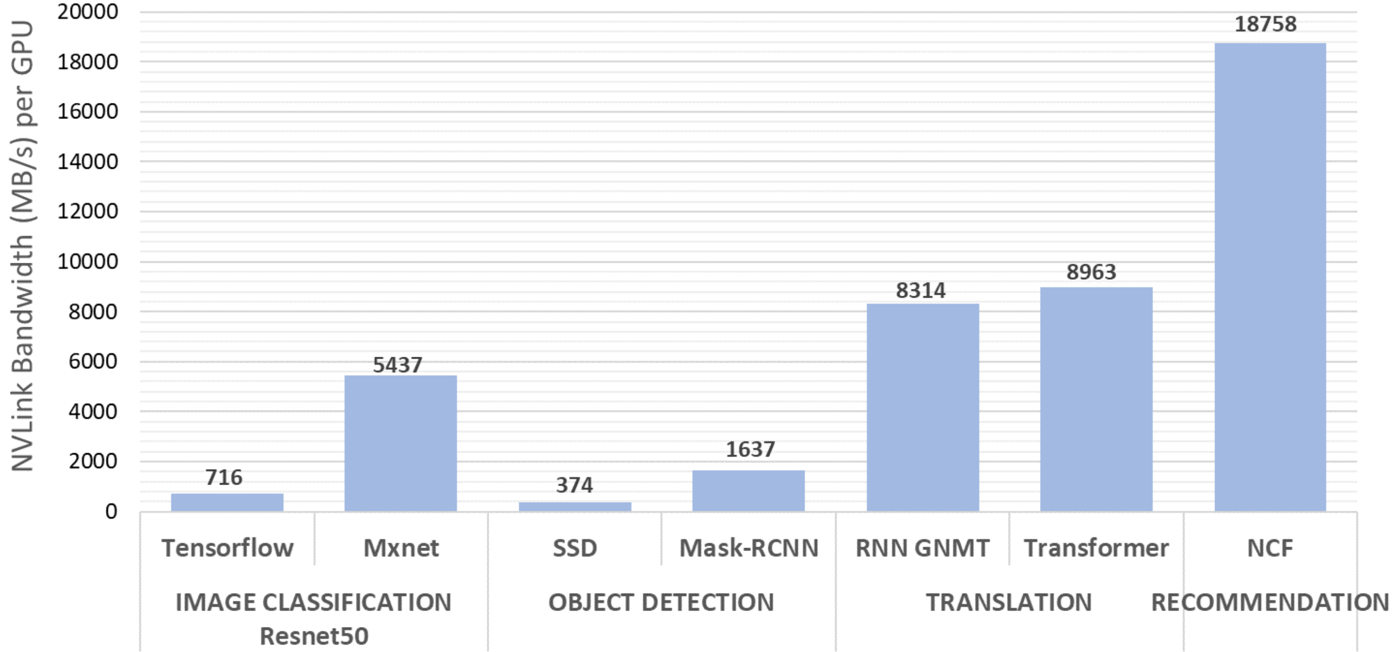
*This translates as a 50 minute savings in training time on Resnet50 (Mxnet) @ 8% speedup*

*33 minutes on Transformer (25% speedup)*

# System Profiling

## NVLink Bandwidth Utilization

*4xV100-SXM2 16GB (NVLink)*

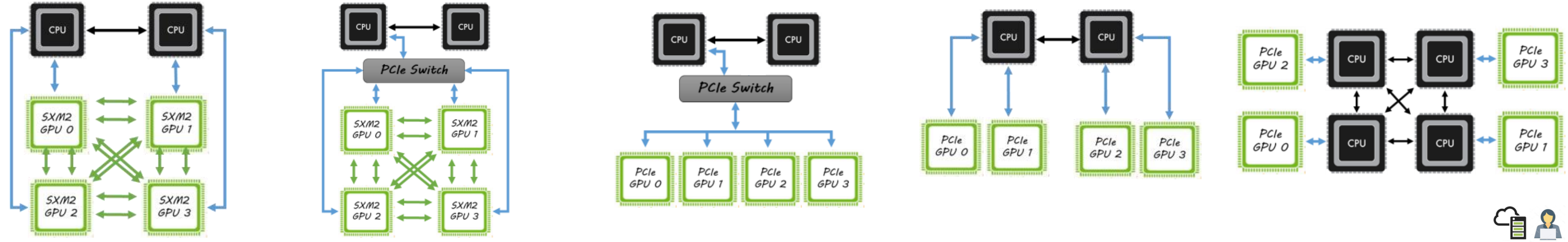



*GPUDirect P2P bandwidth is highest for Resnet50/Mxnet, Translation and Recommendation benchmarks*

# GPU Interconnect Topology

## *NVLink and PCIe Topology Comparisons*

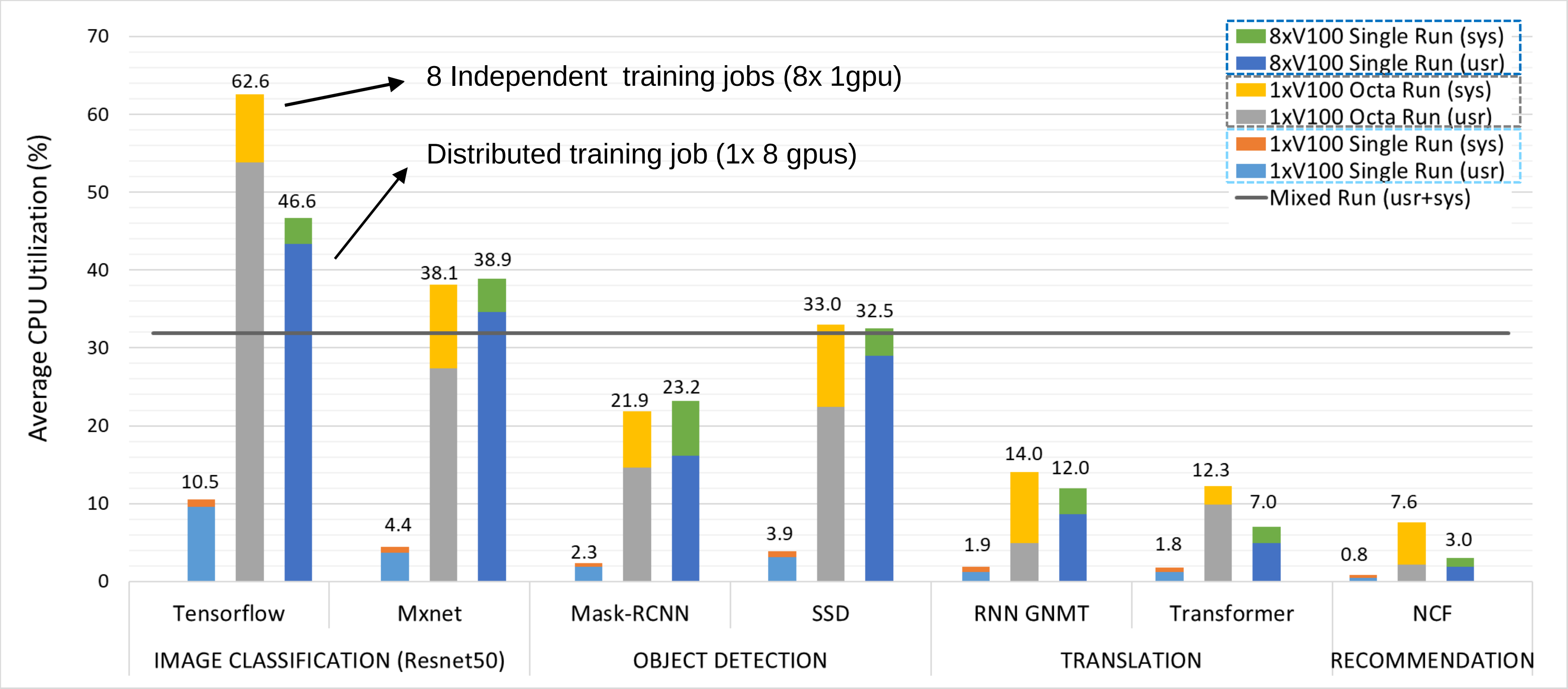
| 2S-4xV100-SXM2 NVL (1CPU:2GPU) | 2S-4xV100-SXM2 NVL+SW (1CPU:4GPU) | 2S-4xV100-PCIe PCIe SW (1CPU:4GPU) | 2S-4xV100-PCIe (1CPU:2GPU) | 4S-4xV100-PCIe (1CPU:1GPU) |
|---|---|---|---|---|




## System Profiling

# *Distributed Training vs. Single GPU Compare 8xV100-PCIe*

### Average CPU Utilization



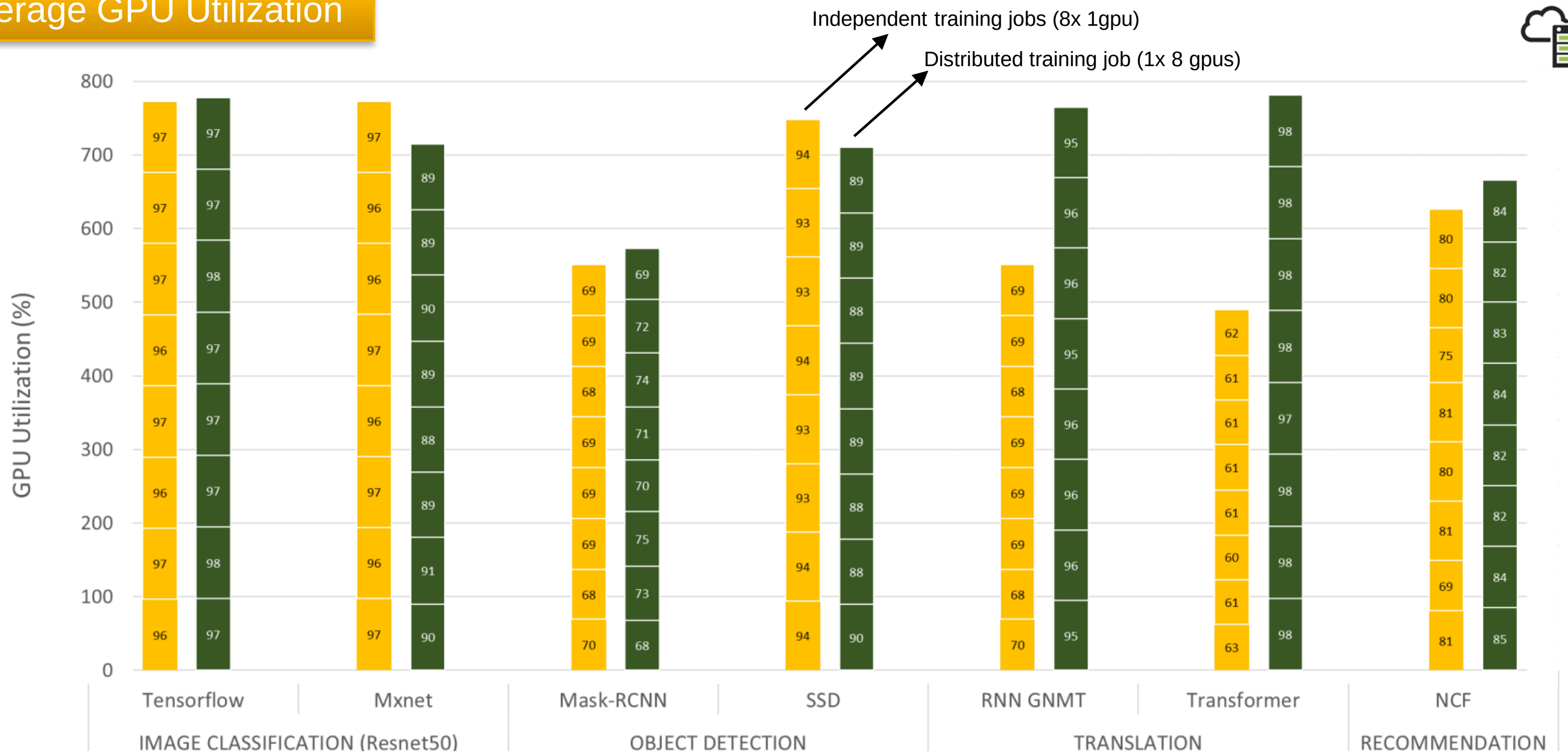



Benchmark result not verified by MLPerf. MLPerf name and logo are trademarks. See www.mlperf.org for more information

# System Profiling

## *Distributed Training vs. Single GPU Compare 8xV100-PCIe*

### Average GPU Utilization






Benchmark result not verified by MLPerf. MLPerf name and logo are trademarks. See www.mlperf.org for more information

# Key Messages

MLPerf is a valuable tool to evaluate impact of GPU technologies and its impact on Deep Learning Training workloads

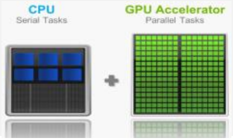
- Performance improvements in Frameworks/Libraries already being accelerated due to MLPerf

Dave the Data Scientist

- Use Nvidia tools to monitor GPU utilization and Scaling Efficiency
- Single node performance sufficient to train complex models in a single workday (Tesla V100)
- Mixed-Precision has significant impact on training performance (150%-330%)

CPU utilization varies considerably between the different benchmarks

- increases with #GPUs and type of DNN
- Offload to GPUs is an option for some DL pipelines




- Choose GPU platforms that meet power, cost, density & flexibility requirements for your training workloads
  - For tests that involve substantial inter-GPU communication, NVLink improves performance (up to 40%) for distributed training scenarios
  - Advances in PCIe topology closing the gap for some use cases

# Acknowledgements


- *Guy Laporte, Liz Raymond, April Berman, Rengan Xu, Frank Han*, Shreya Shah **Dell EMC**
- *Marc Hammons, David Patschke* **Dell Inc**
- *Paulius Micikevicius, Aman Arora, Michael Andersch* **Nvidia**
- *Sreepathi Pai* **Univ of Rochester**

# BACKUP

DELL EMC

# MLPerf Benchmark v0.5

https://www.mlperf.org


**Categories**

- Open vs. Closed
- Cloud vs. On-Premise

**Metrics**

- Metrics to capture both performance and quality
- Cloud Scale
- Variance

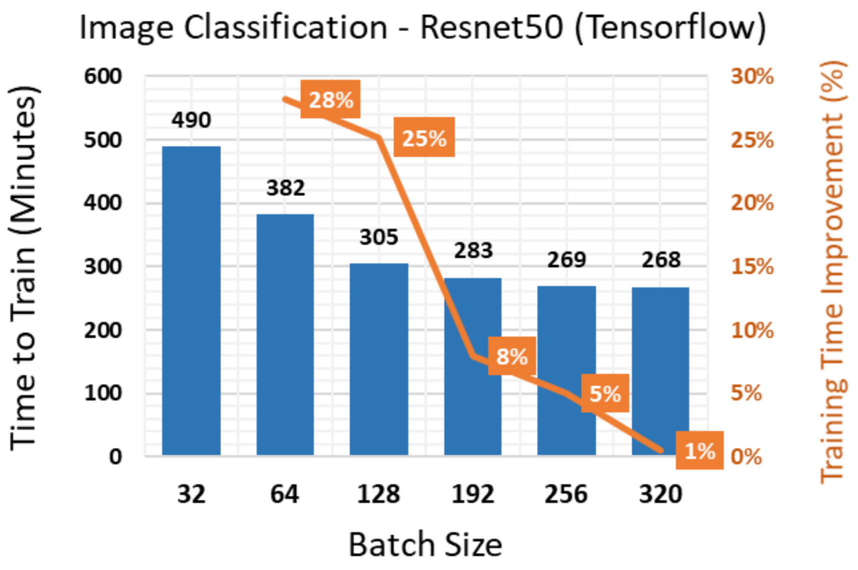
**Code**

- Docker containers for reproducibility
- Agile development model for rapid iteration
- peer review process

| Domain | Model | Dataset | Performance Metric | Use Cases |
|---|---|---|---|---|
| Image Classification | Resnet-50 | ImageNet | Top-1 Classification Accuracy | Google Shopper, Facebook, Google Goggles, Xbox 360 |
| Object detection | SSD, Mask RCNN | Microsoft COCO | mAP | Video surveillance, Pedestrian detection, Anomaly detection |
| Translation | RNN GNMT, Transformer | WMT17 | BLEU scores | Google Translate, Skype |
| Recommendation | Neural Collaborative Filtering | MovieLens 20 Million (ml-20m) | Hit Rate | Product recommendation by Amazon, Netfix recommendations, Spotify |
| Reinforcement Learning | Minigo | Data from games played during benchmarking | # of correct predictions / # of predictions attempted | Traffic Light Control, Robotics, Bidding and Advertising, AlphaGo Zero |

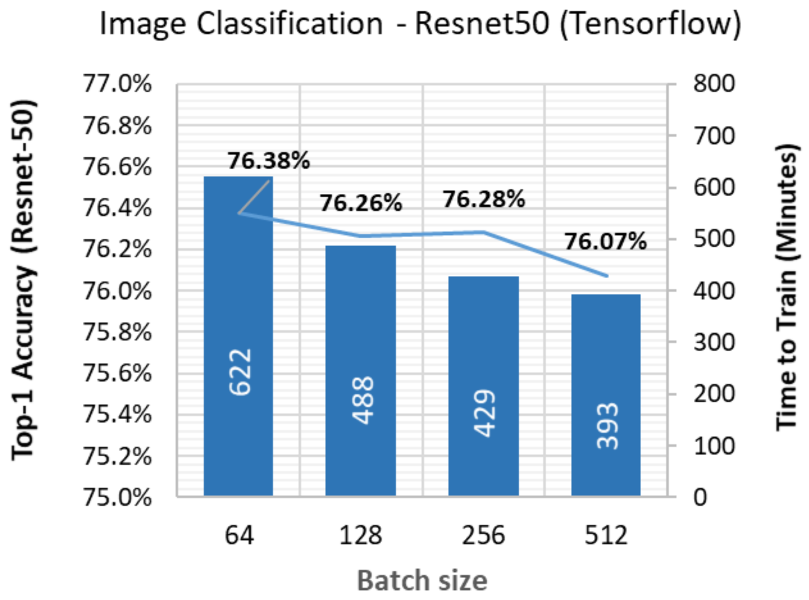
# Workload Characterization

## Batch Size




Target Accuracy = 74.9%

Test Platform: 4xV100-SXM2 16GB NVLink




Max Epochs=90

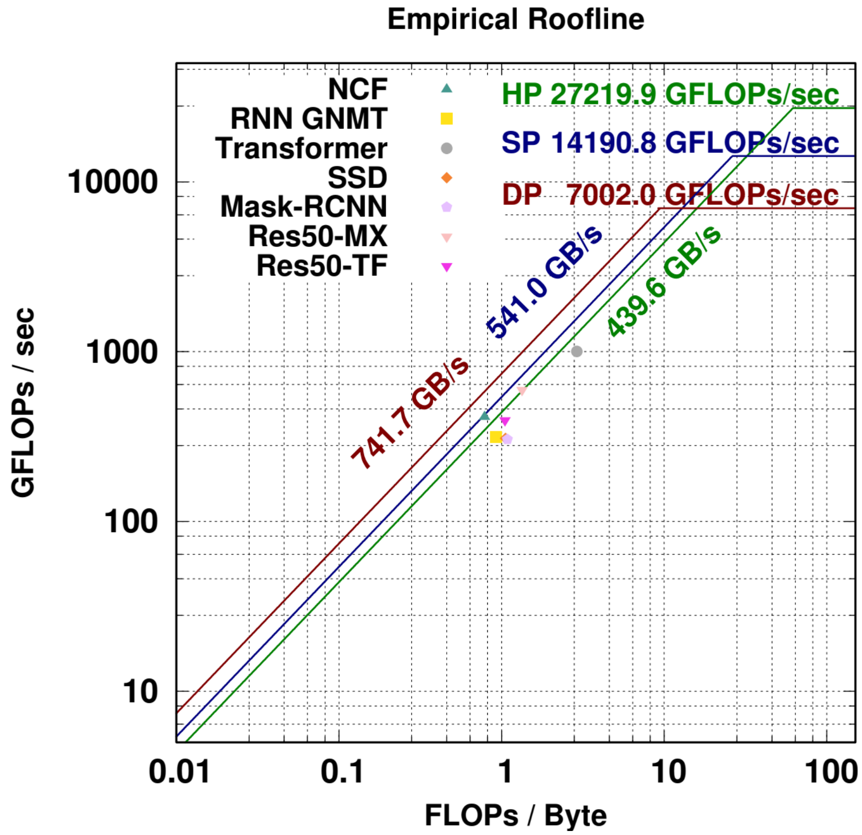
Test Platform: 4xV100-PCIe 32GB

# Roofline Analysis

*1xV100-PCIe 32GB*

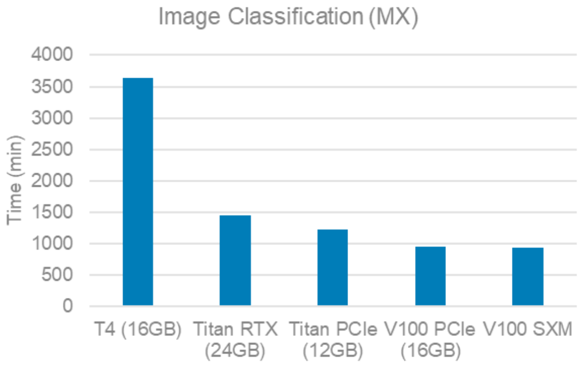
Roofline can be used to assess the quality of attained performance

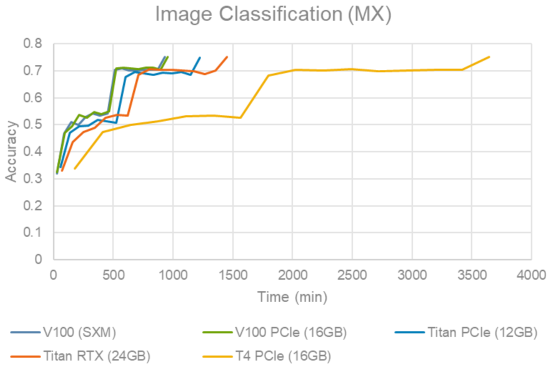
- Arithmetic Intensity is the ratio of total floating-point operations to total data movement
- Kernels near the roofline are making good use of computational resources
- Translation (Transformer) has highest data reuse
- RNN, SSD, Mask-RCNN have similar characteristics

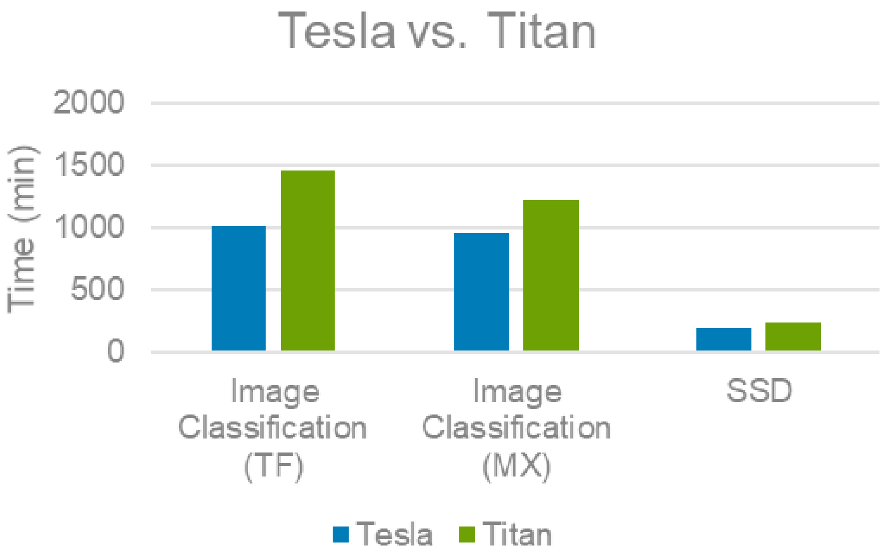


# GPU Comparison

## Titan, Quadro and Tesla Compare





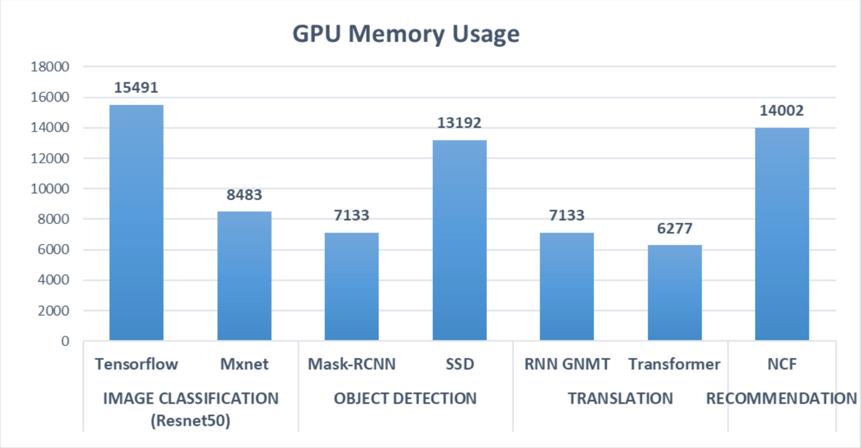



Benchmark results not verified by MLPerf. MLPerf name and logo are trademarks. See www.mlperf.org for more information

# GPU Comparison

# Tesla V100-PCIe: 16GB vs. 32GB

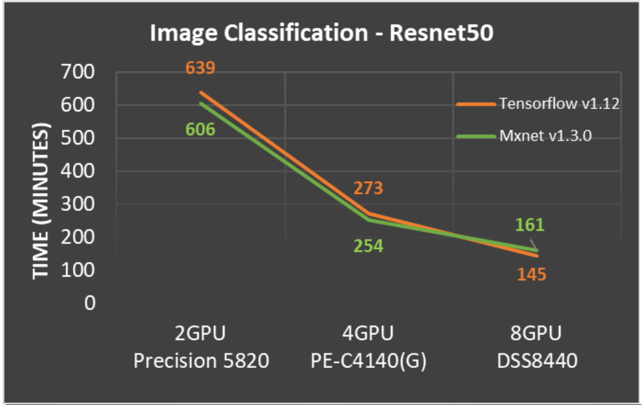
- We compare V100 16GB and 32GB
- T640 Resnet50/TensorFlow Results (256 vs. 512 Batch size)
  - 18260 vs16781
- RNN 512 vs. 256 vs 128
  - 2993 vs. 2656 vs. 8551
- 940xa
  - Mxnet Result (1664 vs 832 Batch Size)
    - 16663 vs 15994
  - RNN_Translation(512 vs 256)
    - 3000 vs 2656)
  - Translation 10240 vs. 5120 (batch size)
    - 4188 vs 5321
  - RCNN
    - 33202 vs. 39460
- Object Detection
  - Images=4 vs. images=8
  - 4xV100 33698 vs 28164

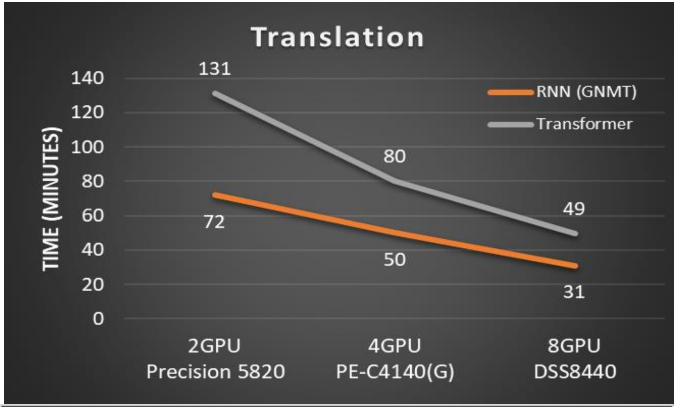


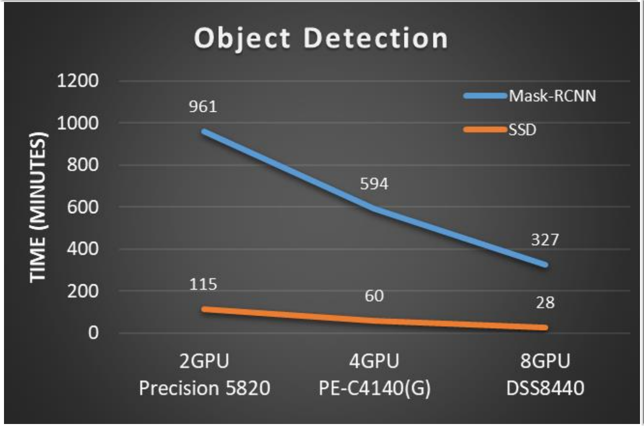
# GPU Scaling

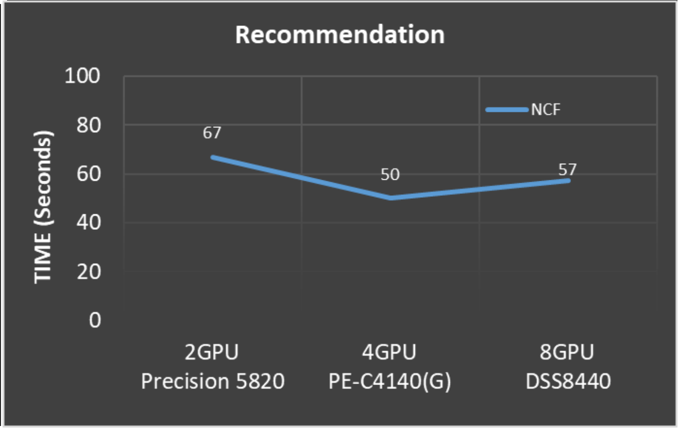
## 2GPU Workstation → 4GPU PCIe Server → 8 GPU PCIe Server

| Workstation | 1U Server | 4U Server |
|---|---|---|
| System | | |
| 2xGV100 | 4xV100 | 8xV100 |
| *DL TFLOPs (mixed-precision)* | | |
| 237 | 480 | 960 |
| *Total HBM2 Memory* | | |
| 64GB | 64GB | 128GB |
| GPU-GPU Bandwidth | | |
| 200 GB/s | 32 GB/s | 32 GB/s |
| *GPU TDP* | | |
| 500W | 1000W | 2000W |



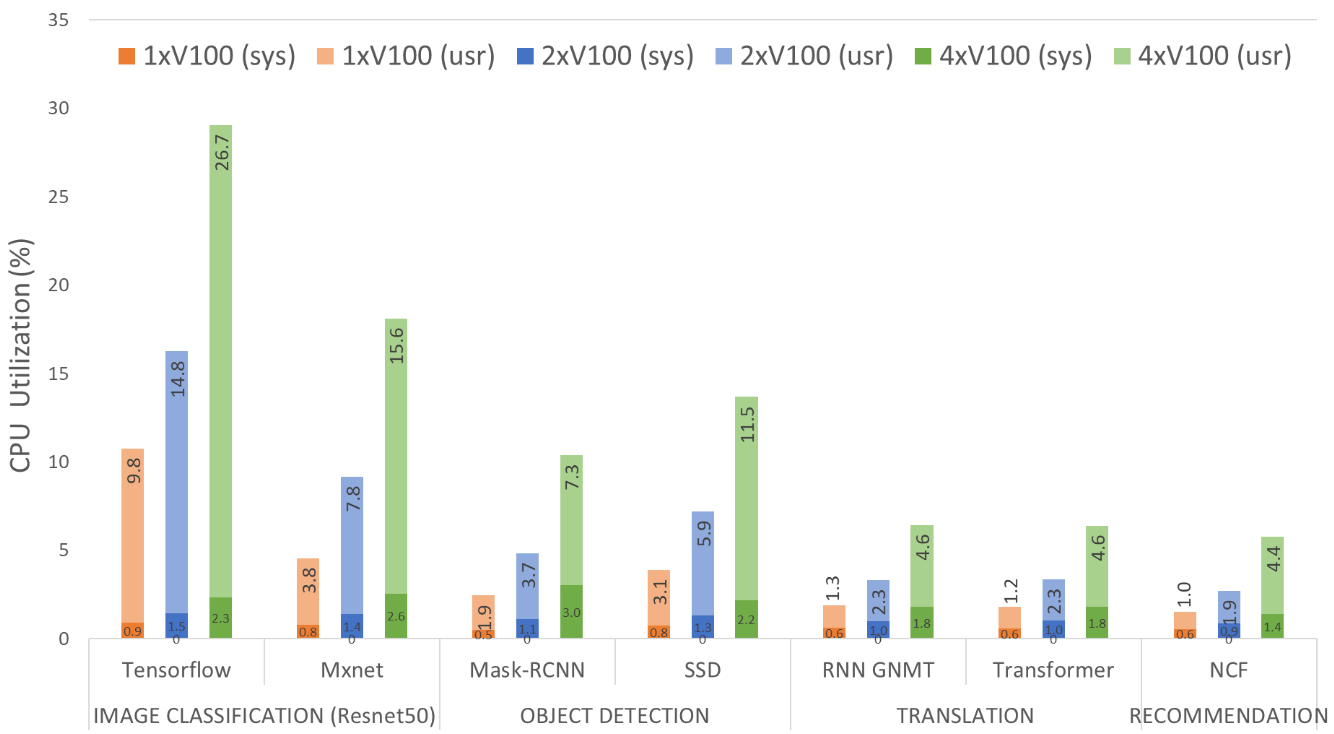



# System Profiling

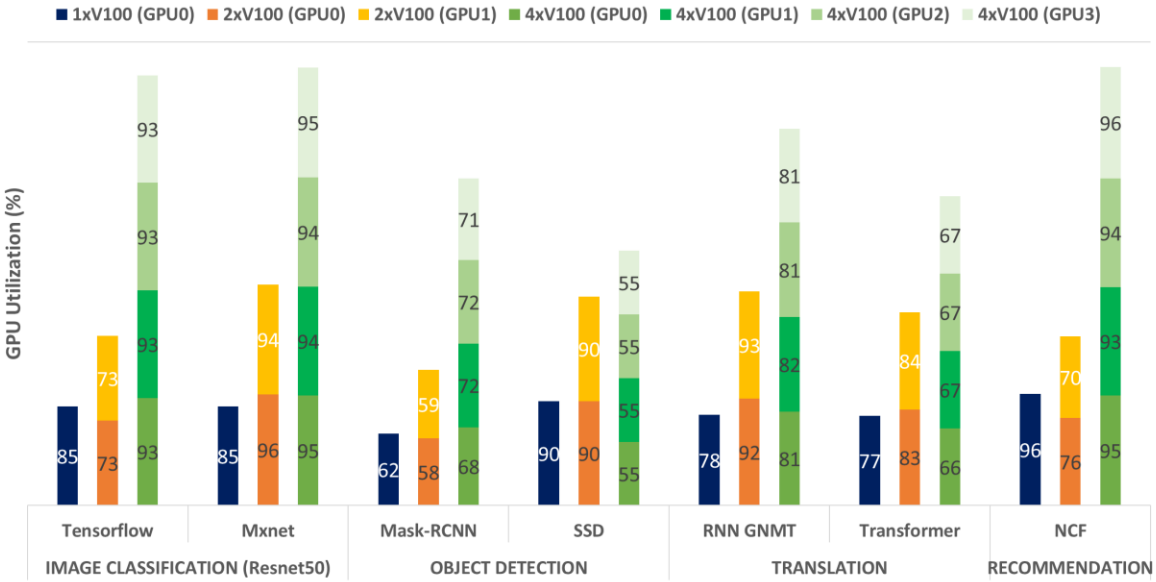
## *CPU Utilization Trends*

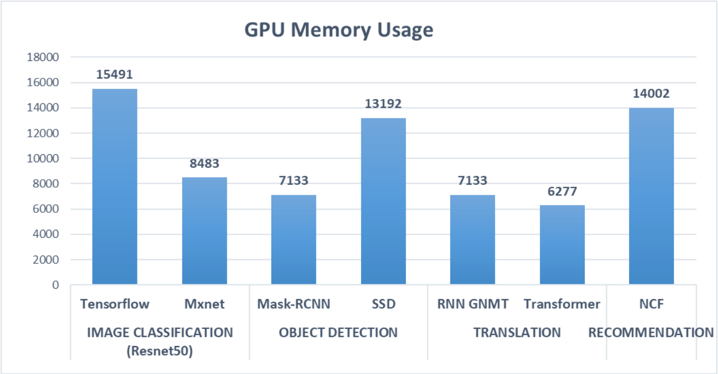
*4xV100-SXM2 16GB (NVLink)*



# System Profiling

## *GPU Utilization Trends* *4xV100-SXM2 16GB (NVLink)*





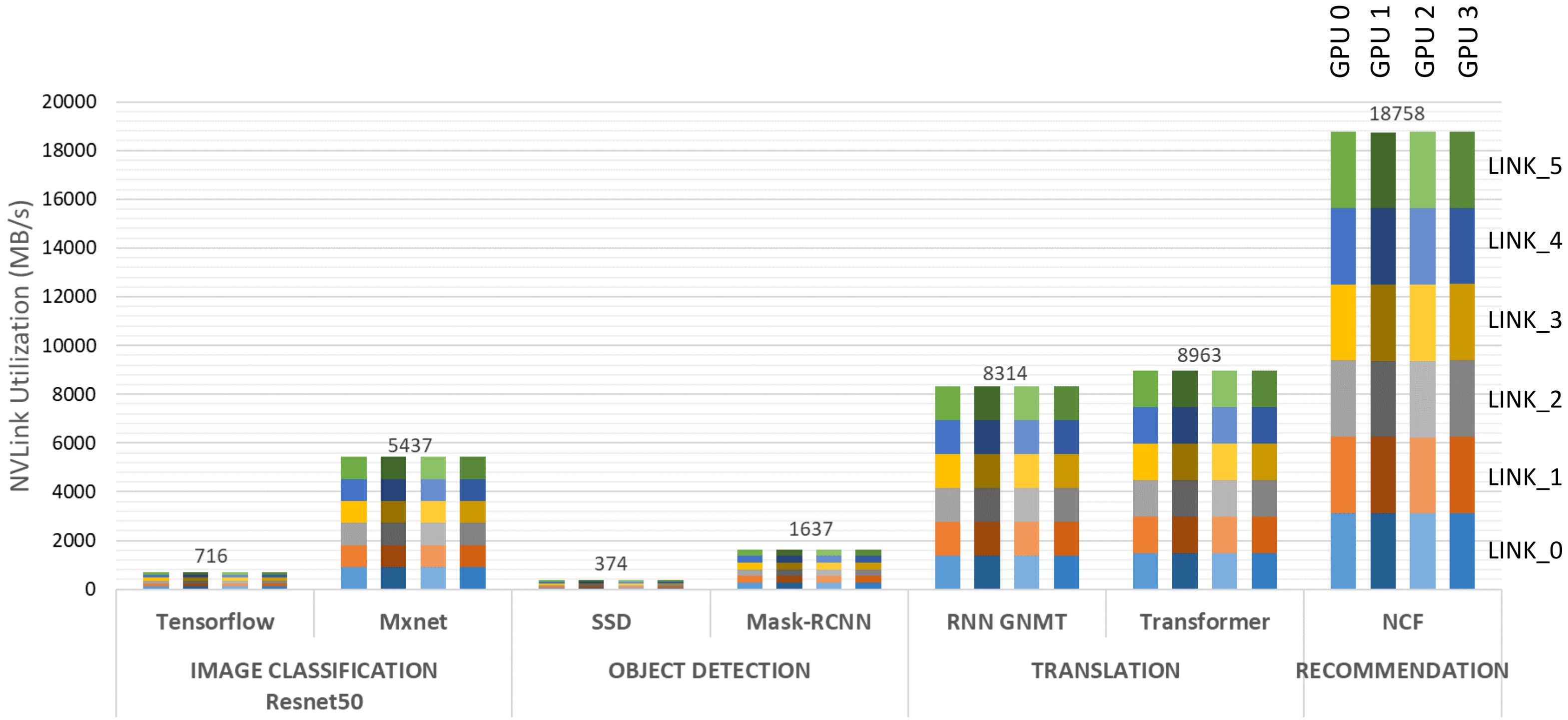



DELL EMC

# System Profiling

## NVLink Utilization

### 4xV100-SXM2 16GB (NVLink)



DELL EMC