Deep Learning Inference on Openshift with GPUs

OpenShift Commons, Seattle, Dec 10 2018

Tripti Singhal
Product Manager, NVIDIA Deep Learning Software
Tushar Katarki
Product Manager, AI on OpenShift

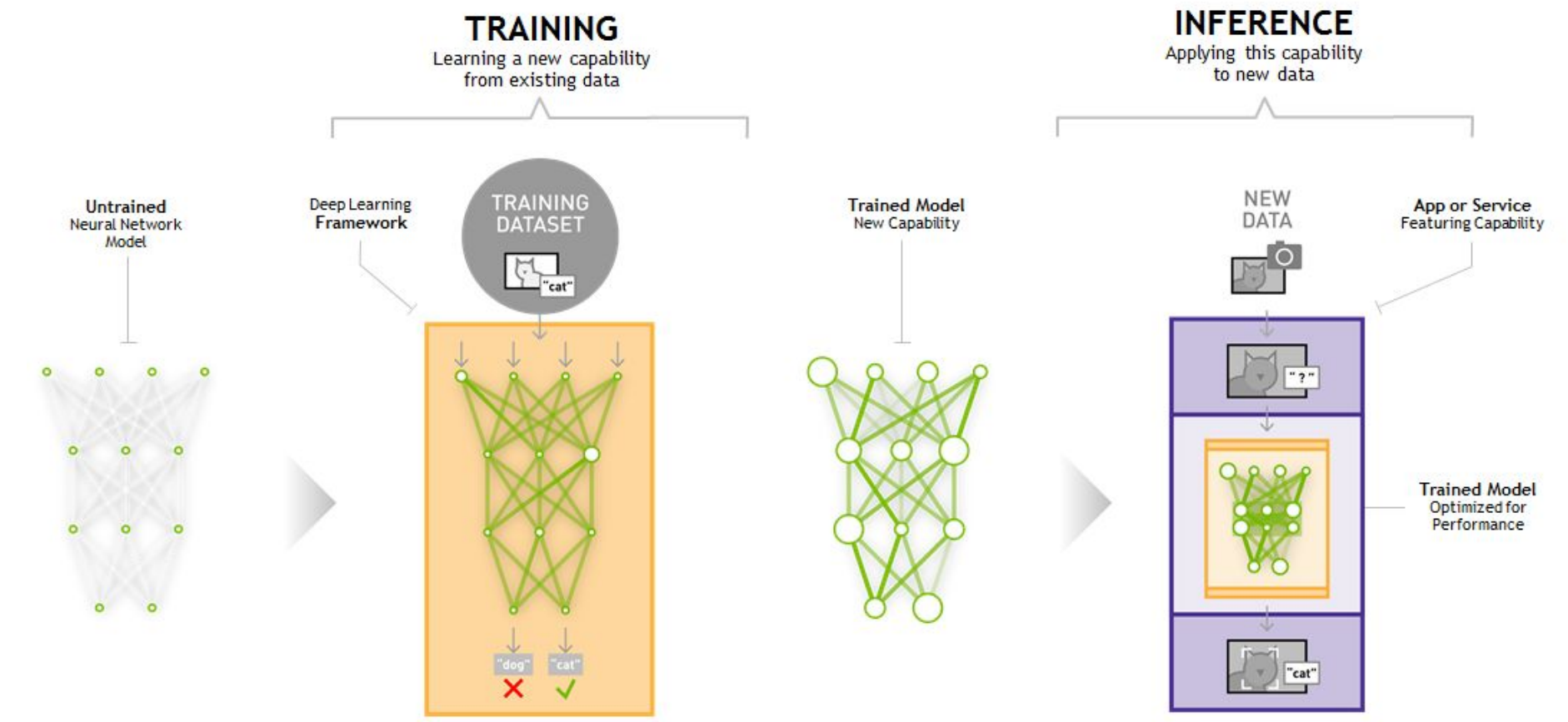
redhat.

NVIDIA

AGENDA

- Deep Learning - Inference
- NVIDIA TensorRT Inference Server
  - Features
  - Architecture
  - Ecosystem
  - Performance
  - Demo
- Road Ahead for Deep Learning on Openshift
- References and Kubecon Highlights

DEEP LEARNING

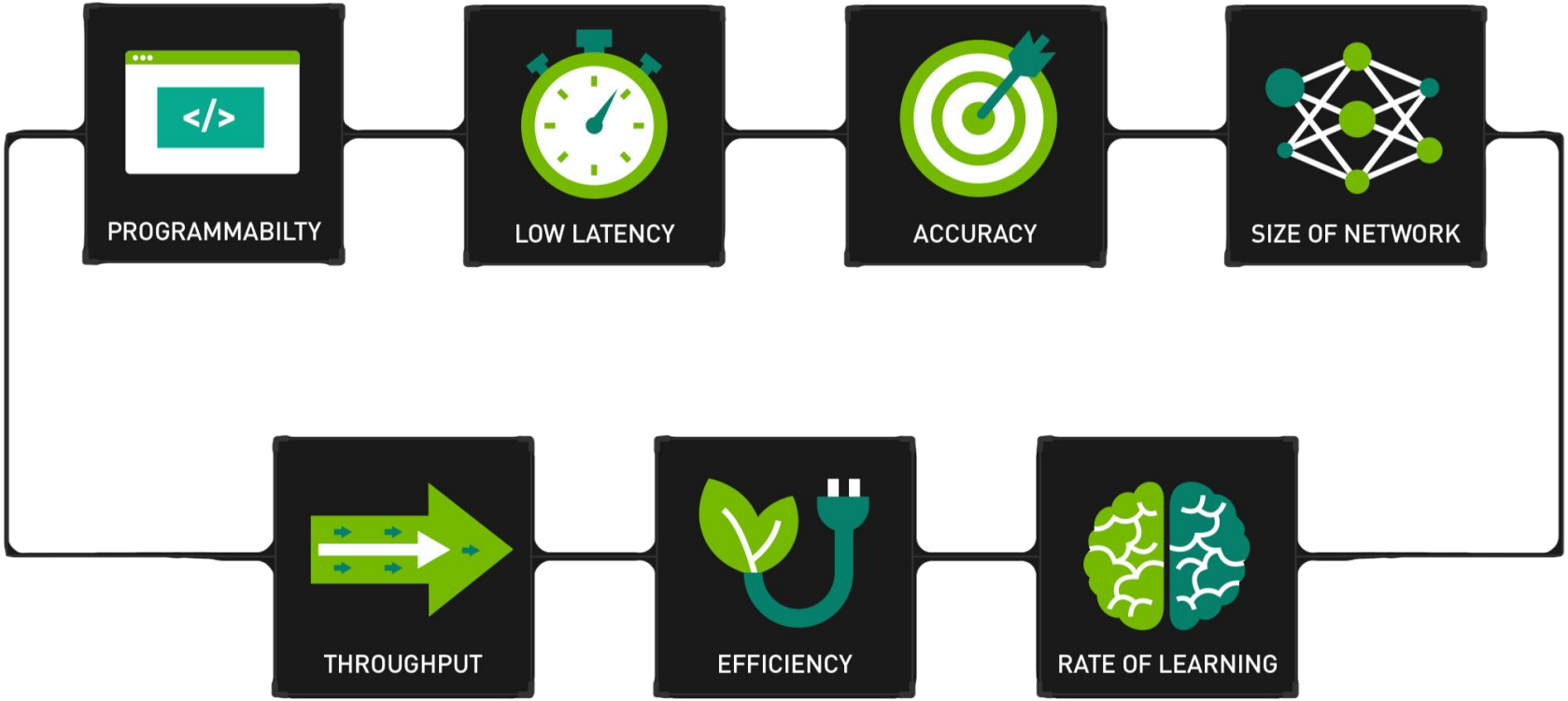


redhat.

NVIDIA

PLASTER

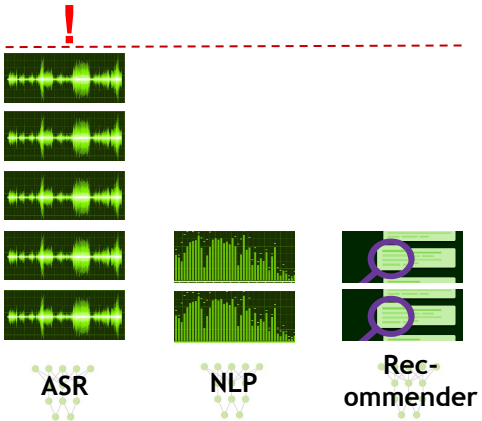
Why use GPUs for inference?



INEFFICIENCY LIMITS INNOVATION

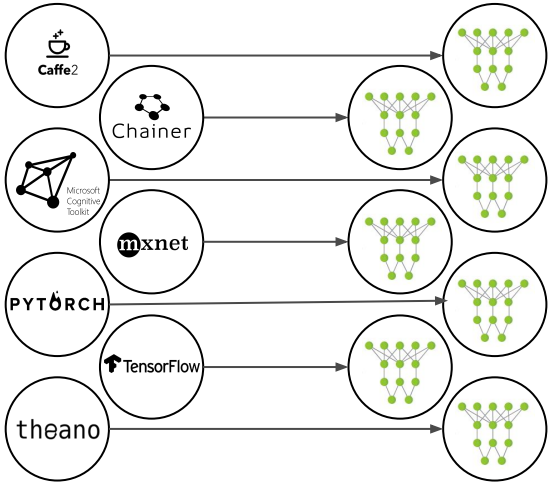
Difficulties with Deploying Data Center Inference

Single Model Only



Some systems are overused while others are underutilized

Single Framework Only



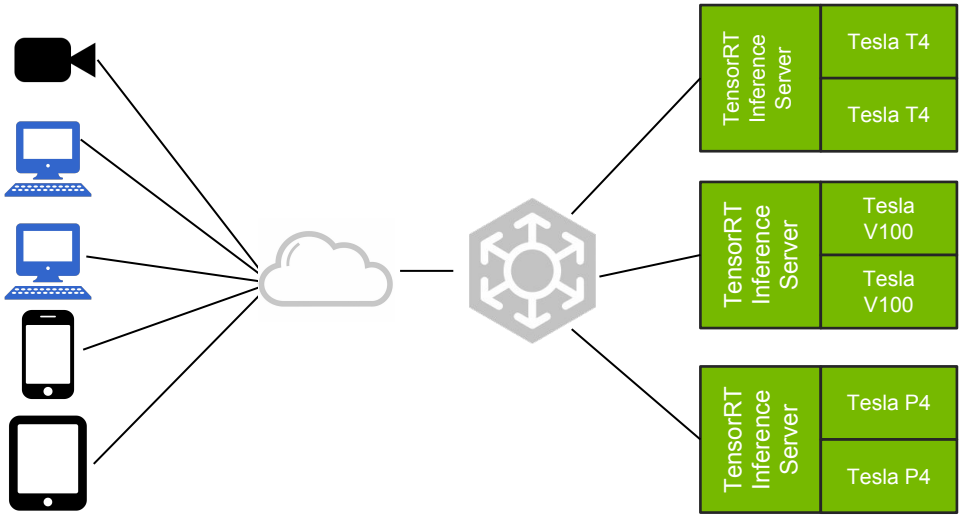
Solutions can only support models from one framework

Custom Development

Developers need to reinvent the plumbing for every application

NVIDIA TENSORRT INFERENCE SERVER

Production Data Center Inference Server



Maximize real-time inference performance of GPUs

Quickly deploy and manage multiple models per GPU per node

Easily scale to heterogeneous GPUs and multi GPU nodes

Integrates with orchestration systems and auto scalers via latency and health metrics

Now open source for thorough customization and integration

CURRENT FEATURES

Concurrent Model Execution

Multiple models (or multiple instances of same model) may execute on GPU simultaneously

Eager Model Loading

Any mix of models specified at server start. All models loaded into memory.

CPU Model Inference Execution

Framework native models can execute inference requests on the CPU

Metrics

Utilization, count, and latency

Dynamic Batching

Inference requests can be batched up by the inference server to 1) the model-allowed maximum or 2) the user-defined latency SLA

Multiple Model Format Support

TensorFlow GraphDef/SavedModel
TensorFlow and TensorRT GraphDef
TensorRT Plans
Caffe2 NetDef (ONNX import path)

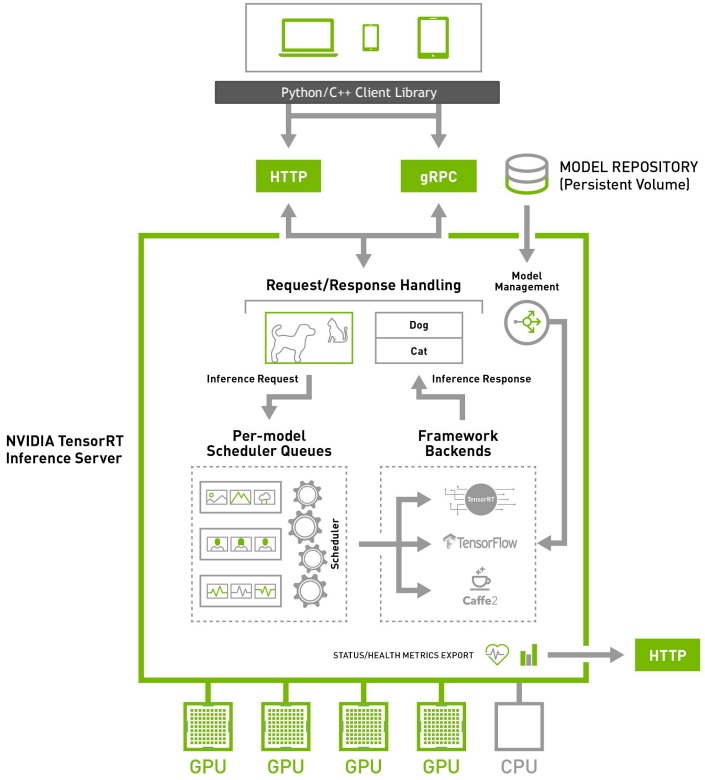
Mounted Model Repository

Models must be stored on a locally accessible mount point

TensorRT

INFERENCE SERVER ARCHITECTURE

Available with Monthly Updates



Python/C++ client libraries

Server HTTP REST API/gRPC

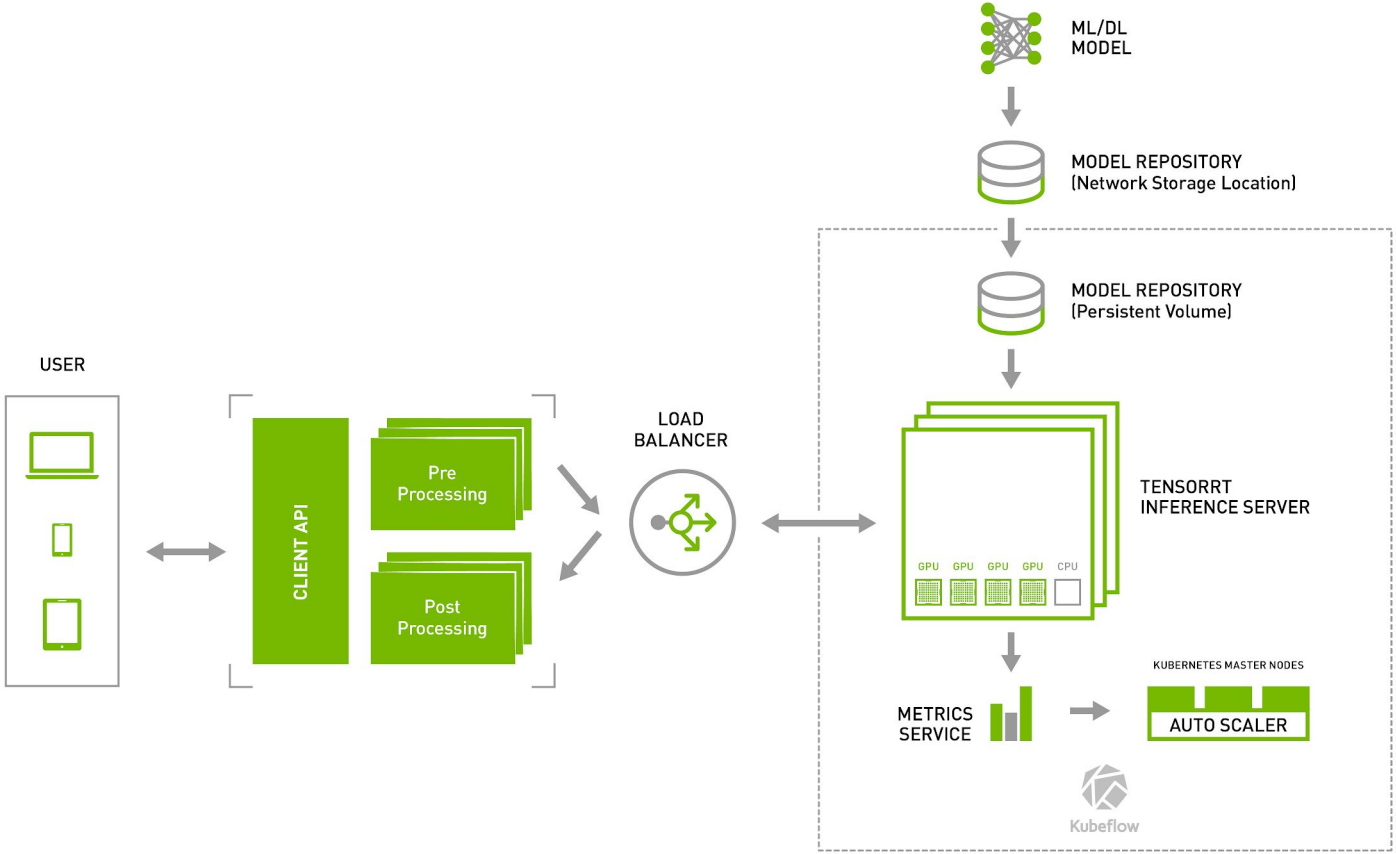
Request/Response Handling

Per-model Scheduler Queues

Framework Backends

Status/Health Metrics

INFERENCE SERVER ECOSYSTEM



redhat.
NVIDIA.

GREAT PERFORMANCE FOR MULTIPLE MODEL DEPLOYMENTS OF RN50

RN50 with 50ms latency SLA across various deployments

- CPU: TensorFlow FP32
- GPU - V100 16GB: TensorFlow FP32
- GPU - V100 16GB: TensorRT FP16

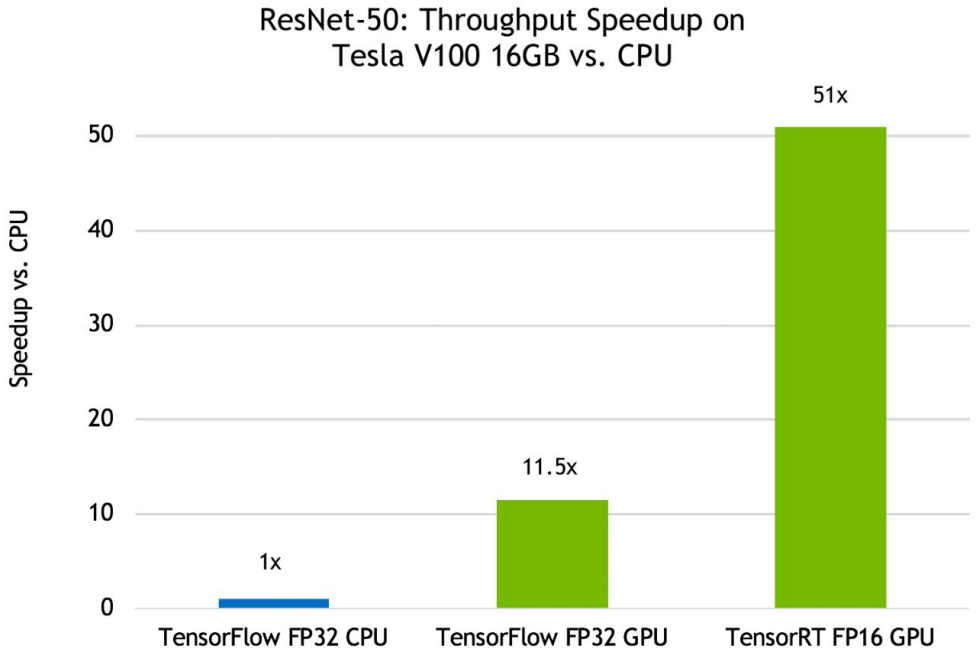
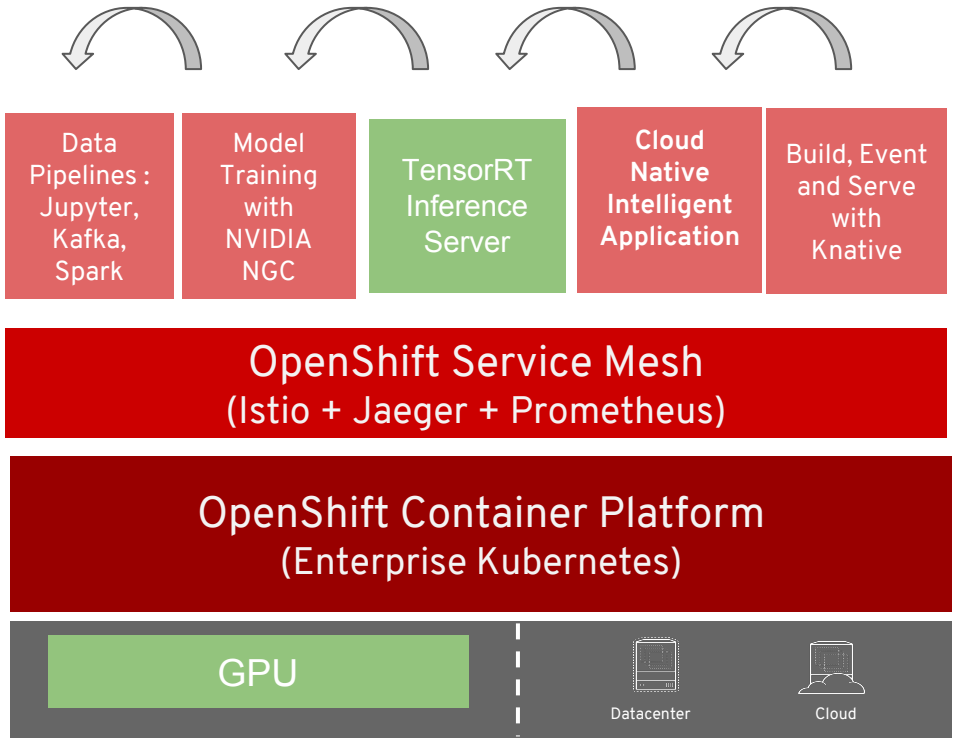


Figure 7: GPU performance improvements at FP32 and FP16 precision. CPU: Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz x 36. At ~50ms latency target. Batch size: 1 (CPU), 8 (GPU), 8 (GPU). Model concurrency: 8 (CPU), 8 (GPU), 12 (GPU).


19%
57%

ROAD AHEAD FOR DEEP LEARNING ON OPENSHIFT



Device Manager Support, Priority and Preemption

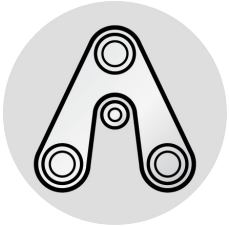
OpenShift/Red Hat Enterprise Linux on Nvidia DGX-1/Tesla - NVIDIA NGC on OpenShift

Preview of TensorRT Inferencing on OpenShift

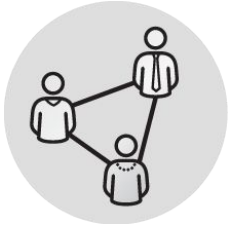
GPU Sharing, Heterogeneous Clusters, GPU Topology, Install experience with Operators

HOW RED HAT SEES AI

Represents a workload requirement for our **platforms** across the hybrid cloud.



Applicable to Red Hat's existing core business in order to increase **Open Source** development and production **efficiency**.



Valuable to our customers as specific services and product capabilities, providing an **Intelligent Platform** experience.

Enable customers to build **Intelligent Apps** using Red Hat products as well as our broader partner ecosystem.

DATA AS THE FOUNDATION

REFERENCES AND KUBECON HIGHLIGHTS

Sessions

- Tue, Dec 11: Scaling AI Inference Workloads with GPUs and Kubernetes - Renaud Gaubert & Ryan Olson, NVIDIA
- Thu, Dec 13: Building DL Through Knative Serverless Framework - Huamin Chen et al, Red Hat

Blogs

- NVIDIA TensorRT Inference Server Boosts Deep Learning Inference
- Kubeflow: GPU Accelerated Inference for Kubernetes with the NVIDIA TensorRT Inference Server and Kubeflow
- Running Nvidia GPUs on OpenShift

Webinar

- Jan 24, Maximizing GPU Utilization for Data Center Inference with NVIDIA TensorRT Inference Server

Booths

- NVIDIA (Booth S86, Hall 4AB)
- Red Hat (Booth D1)

NVIDIA TensorRT Inference Server Open Source on GitHub
OpenShift Commons ML SIG and OpenDataHub

THANK YOU !