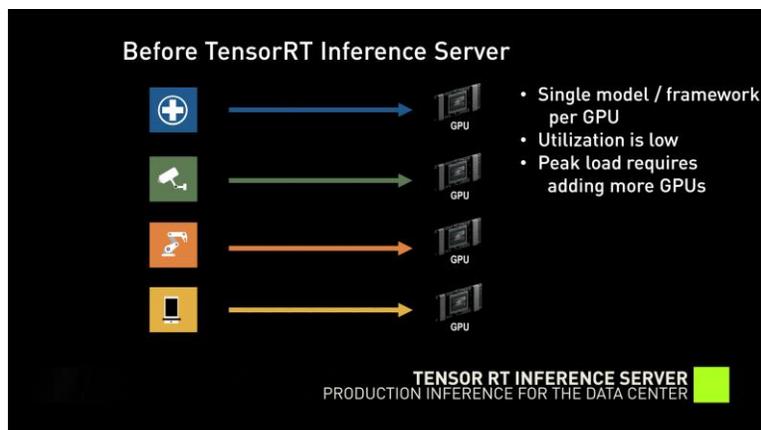
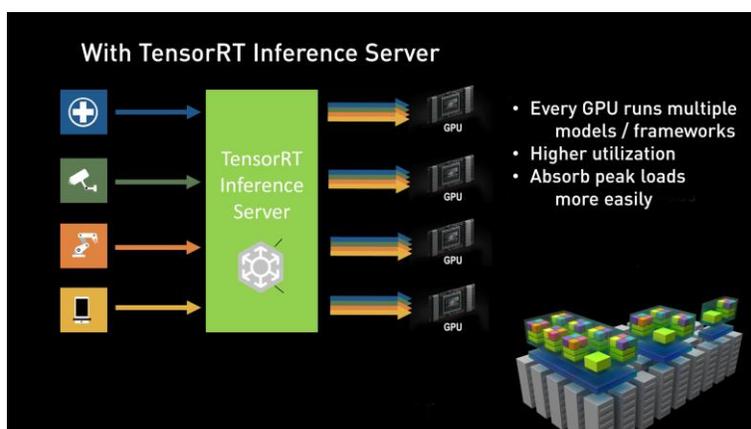


TensorRT Inference Server: Transcript



Introduction

- Whether it's performing object detection in images or video, recommending restaurants, or translating the spoken word, inference is the mechanism that allows applications to derive valuable information from trained AI models.
- However many inference solutions are one-off designs that lack the performance and flexibility to be seamlessly deployed in modern production data center environments.



- NVIDIA TensorRT Inference Server lets you simplify the deployment of inference applications in data centers.
- Delivered as a ready-to-deploy container from NGC and as an open source project, TensorRT Inference Server is a microservice that enables applications to use AI models in data center production.
- It supports the top AI frameworks and custom backends, and it maximizes utilization by running multiple models concurrently per GPU and across multiple GPUs with dynamic request batching.
- TensorRT Inference Server also seamlessly supports Kubernetes with health and latency metrics, and integrates with KubeFlow for simplified deployment.

Demo

- Here is a typical deployment where each one of the eight GPUs is dedicated to running only one neural network model such as voice recognition, product recommendation, or even medical imaging
- In this case, we're classifying pictures of flowers on only the first two GPUs shown here in blue.
- Note the current flower identification inference demand is 1200 images per second, and the GPUs are keeping up.
- *Action: Highlight demand/delivered, and blue load at the same time*



- In a data center production environment, inference agility is critical.
- Because it's impossible to fully predict demand, failure to efficiently manage unique inference workloads can result in underutilized GPUs while STILL failing to meet the demand for that specific inference need.



- If the demand for flower identification increases significantly, the GPUs providing inference for that model are overwhelmed even though the overall data center utilization is still quite low.
- *Action:*
increase the load to 5k
highlight demand/delivered + overwhelmed servers + overall data center utilization
- Currently, each GPU is ONLY able to focus its efforts on one neural network model at a time.
- This is where TensorRT Inference Server can help.
- *Action turn on TRTIS*



- *Action: highlight Inf Server ON + 5k demand is met*
- TensorRT Inference Server runs any model on any GPU - maximizing resource utilization.
- Now, we're meeting each unique inference demand, and servicing multiple inference request types on every GPU, shown here as different colors on each server's graph.
- Another great benefit of TensorRT Inference Server is improved inference capacity.
- *Action: Increase demand to 15,000, highlight demand, delivered, and loads at the right time*



- Before TensorRT Inference Server, we were hitting an inference bottleneck with a demand of only 5000 images per second.
- Now we can handle even 15,000 images per second, and the load is balanced amongst all available resources.

Wrap Up

- We now have a common, open source solution for AI inference that supports all the top AI frameworks, and maximizes GPU utilization.
- NVIDIA TensorRT Inference Server enables everyone to focus on what matters most - researchers to focus on creating high-quality trained models, DevOps engineers to focus on deployment, and developers to focus on their applications.