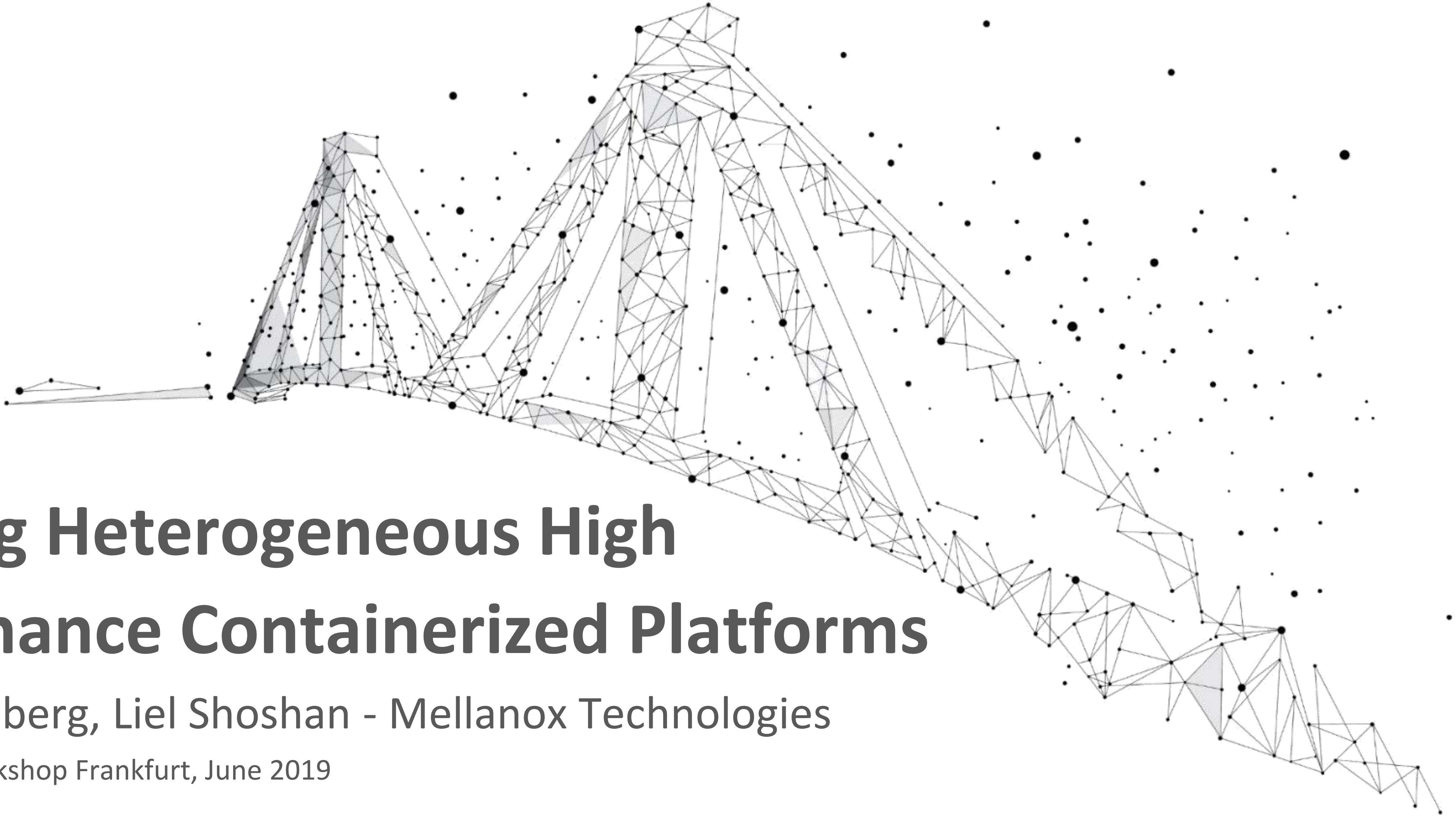


Enabling Heterogeneous High Performance Containerized Platforms

Dror Goldenberg, Liel Shoshan - Mellanox Technologies

ISC Container Workshop Frankfurt, June 2019

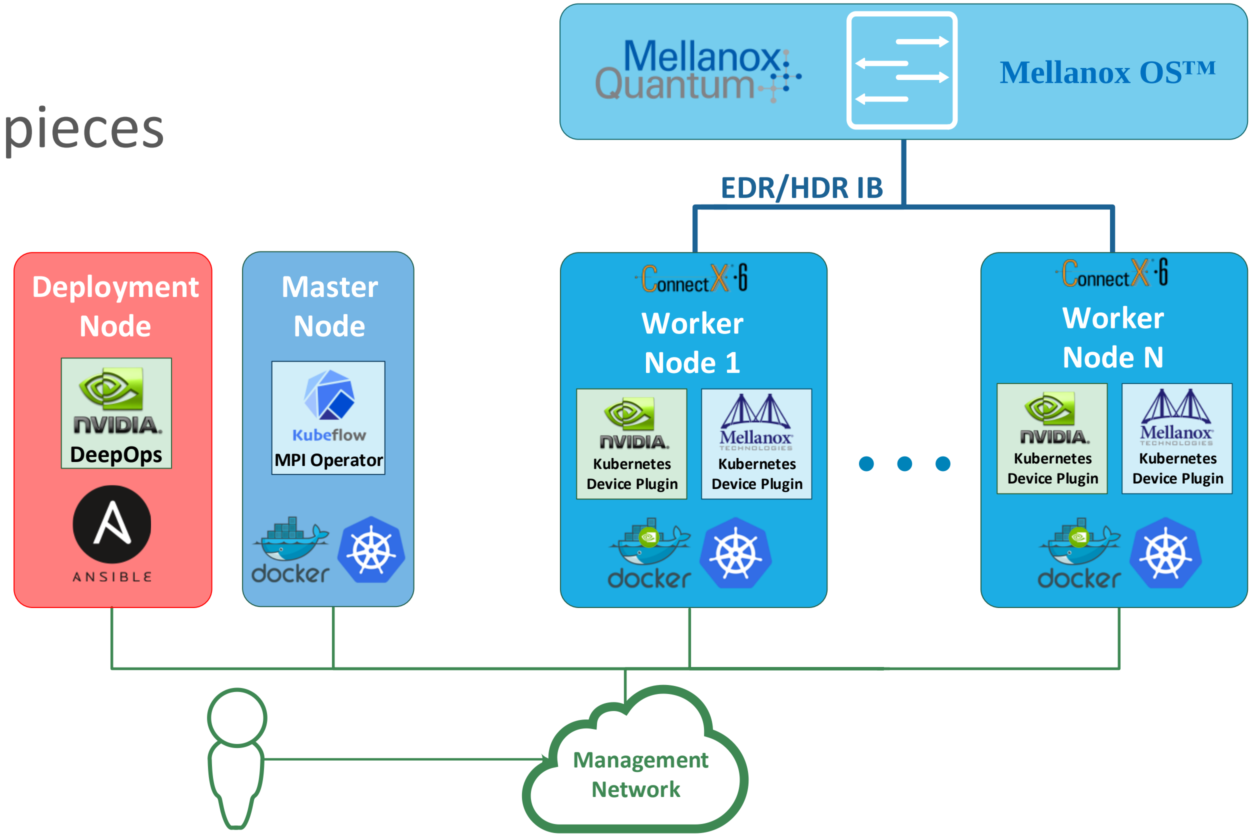
Containerized Applications for Heterogeneous Architectures

- Heterogeneous cluster architectures allows parallel computing that relies on CPU and GPU
- Used for HPC & ML
  - Leveraging high speed, low latency, smart interconnects to speed-up data computation
- GPUDirect RDMA technology improves GPU-GPU communication and eliminates CPU involvement
- Kubernetes serves as a useful way of distributing compute-intensive work across such clusters
- Containerizing compute-intensive applications poses challenges on configuration, deployment and orchestration of the required system devices

Challenges in Building a K8s HPC/ML Cluster

- Deploying a K8s HPC cluster requires installation of various device drivers, libraries and toolkits
  - On the node level
    - Nvidia Driver, CUDA toolkit, cuDNN, MLNX_OFED, GPUDirect, Docker, K8s, etc.
  - On the orchestration level
    - Nvidia Device Plugin, RDMA Device Plugin, K8s CNI, Kubeflow, etc.
  - On the container level
    - Tensorflow, Horovod, MLNX_OFED, OpenMPI, etc.

- One of the biggest challenges is making all these code pieces up and running in an easy and consistent manner

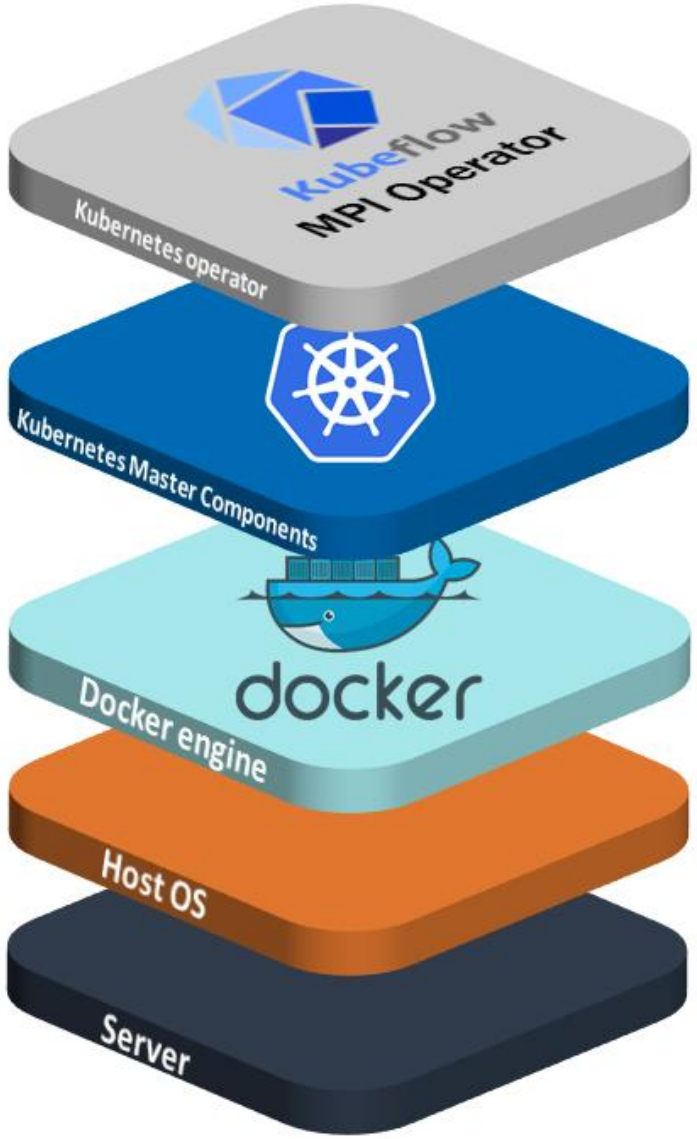


The Solution

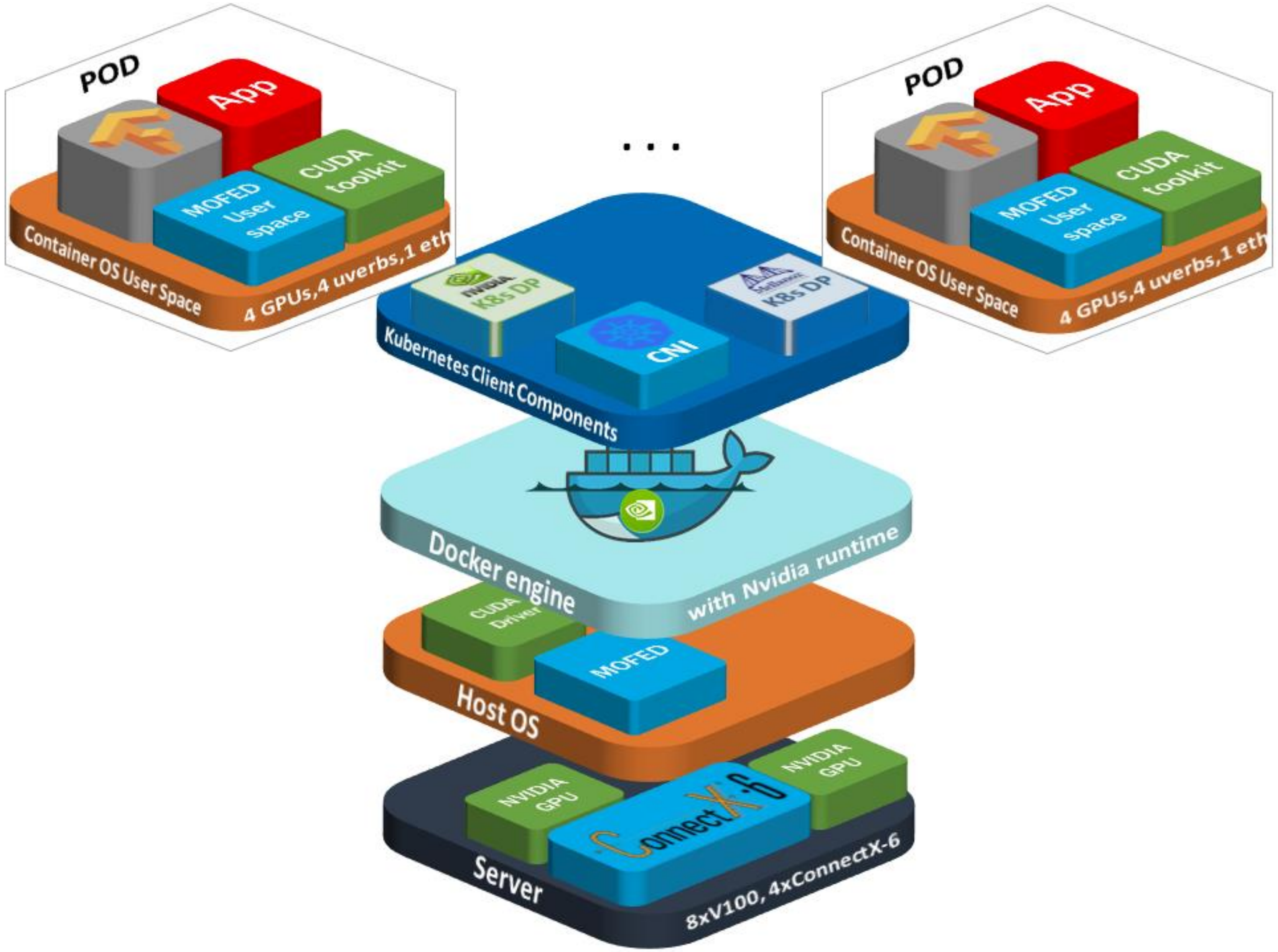
- The following projects speed up deployment time, while making cluster installation vastly simpler
  - DeepOps
    - Facilitates deployment of multi-node GPU and RDMA K8s clusters for ML and HPC environments
    - Employs best practices when setting storage and configuring authentication and user access
  - Kubeflow
    - Kubernetes-native platform for developing, orchestrating, deploying and running scalable and portable ML workloads
    - Provides a straightforward way to deploy best-of-breed open-source systems for ML to diverse infrastructures
    - Helps support reproducibility and collaboration in ML workflow lifecycles
    - MPI Operator
      - Makes it easy to run allreduce-style
  - Mellanox addons for DeepOps
    - Ansible playbook for MLNX_OFED, GPU Direct and K8s device plugin
- Reference deployment guides can be found on community.Mellanox.com and docs.Mellanox.com

K8s HPC Cluster

Master



Worker node

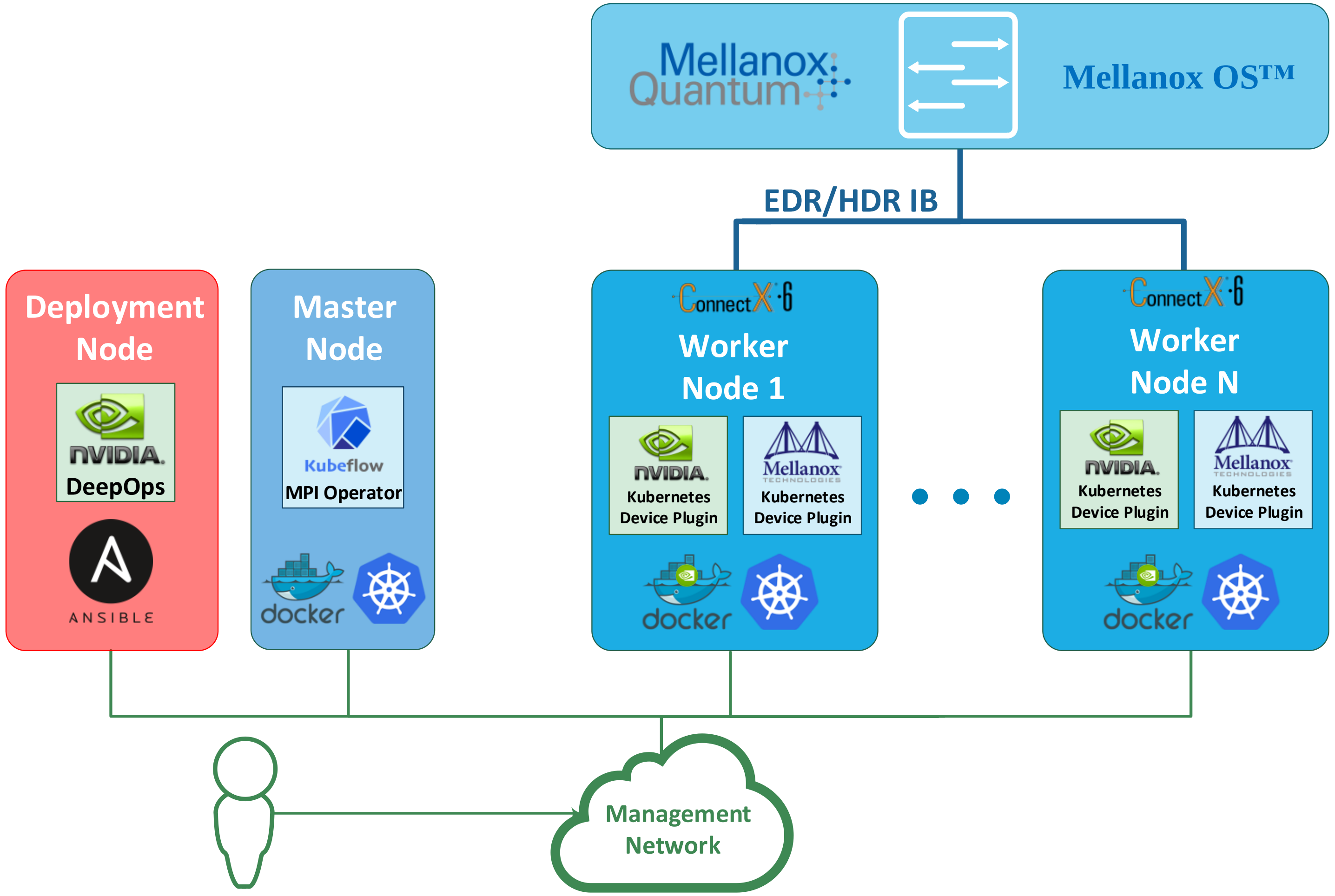


Performance Tests

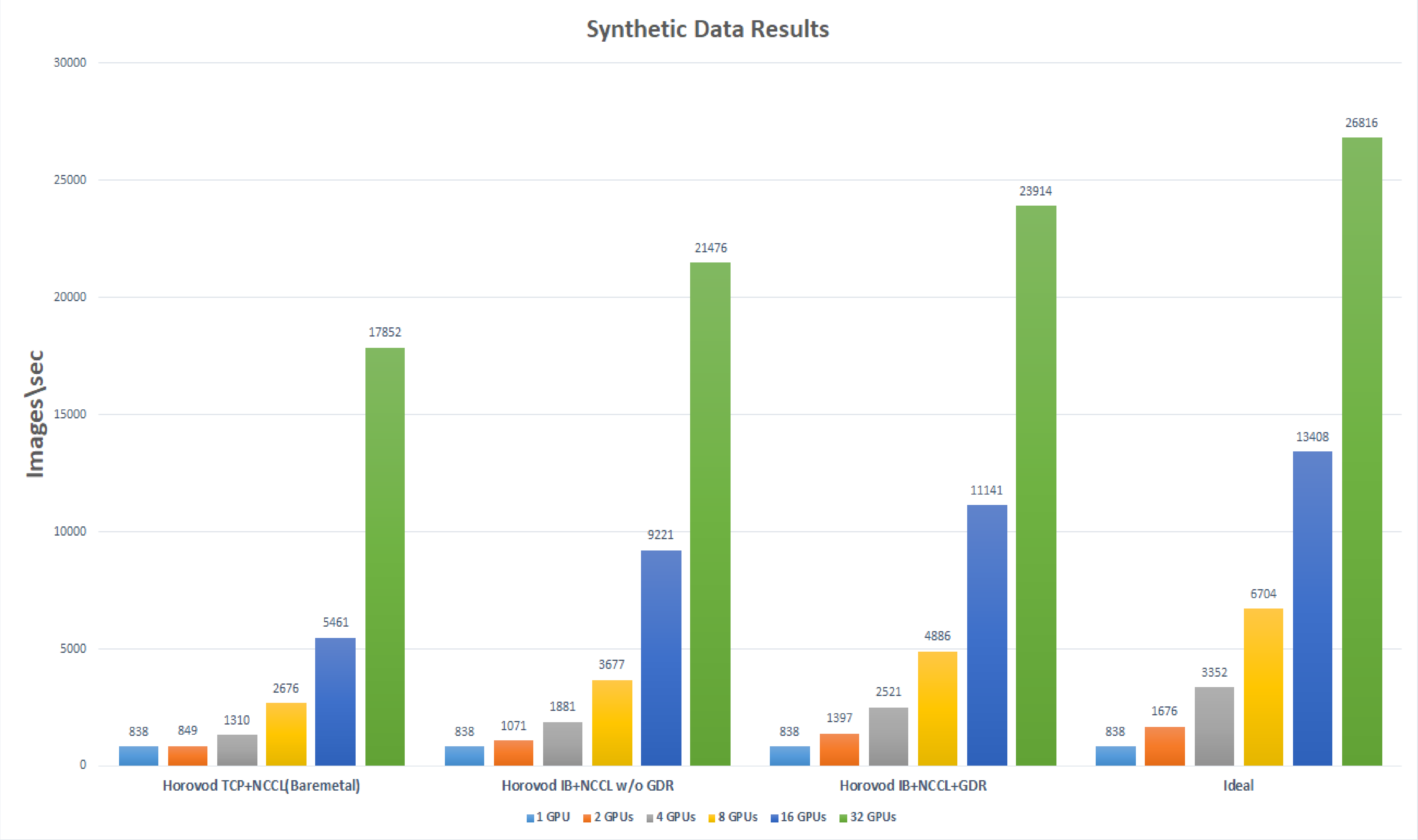
Testing Environment

- Topology
  - Nodes
    - Deployment node
    - Master node
    - 4 x Worker nodes
      - Each node has 8 NVidia Tesla GPU cards and 4 Mellanox ConnectX-6 adapters
  - Containers
    - Each worker node runs 1 Pod

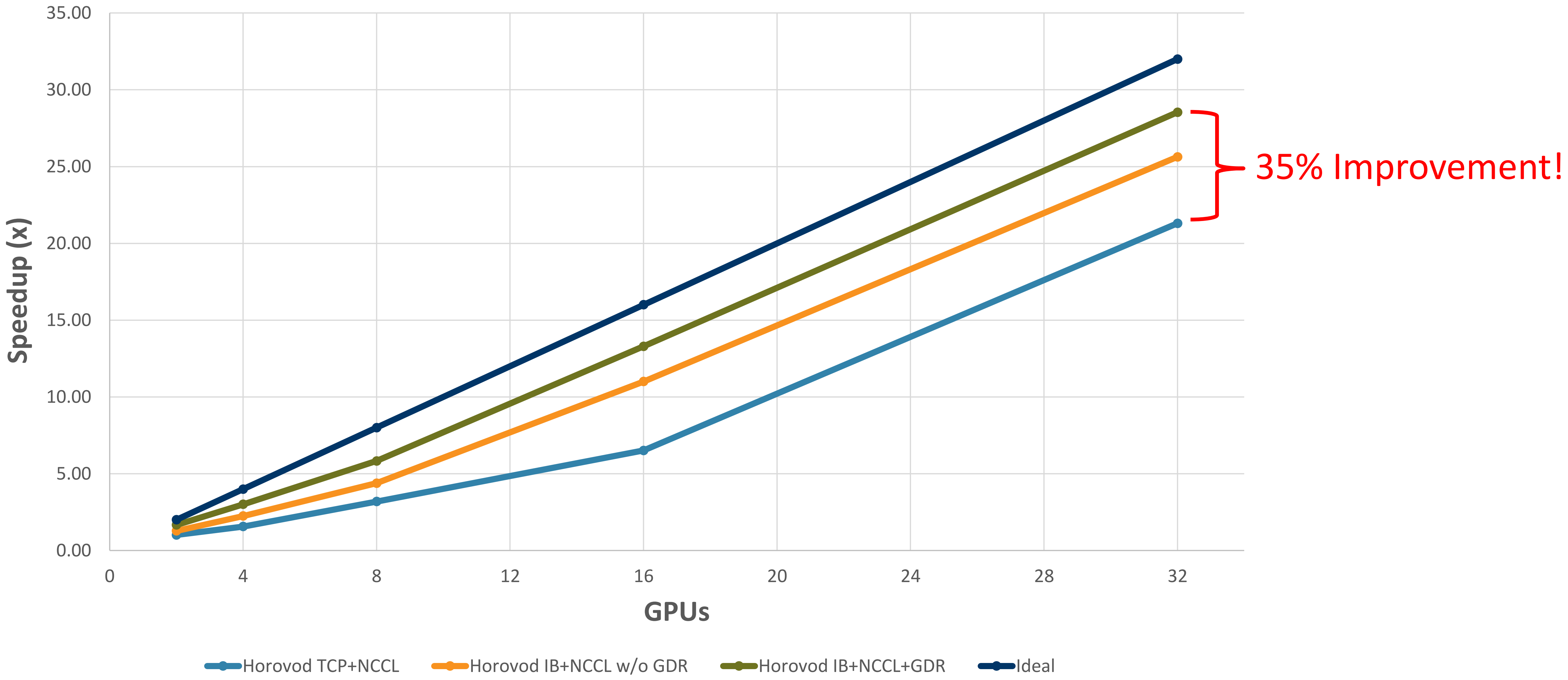
- Benchmark
  - TensorFlow v1.12.0
  - Type: Synthetic
  - Batch size: 32
  - Resnet50 model



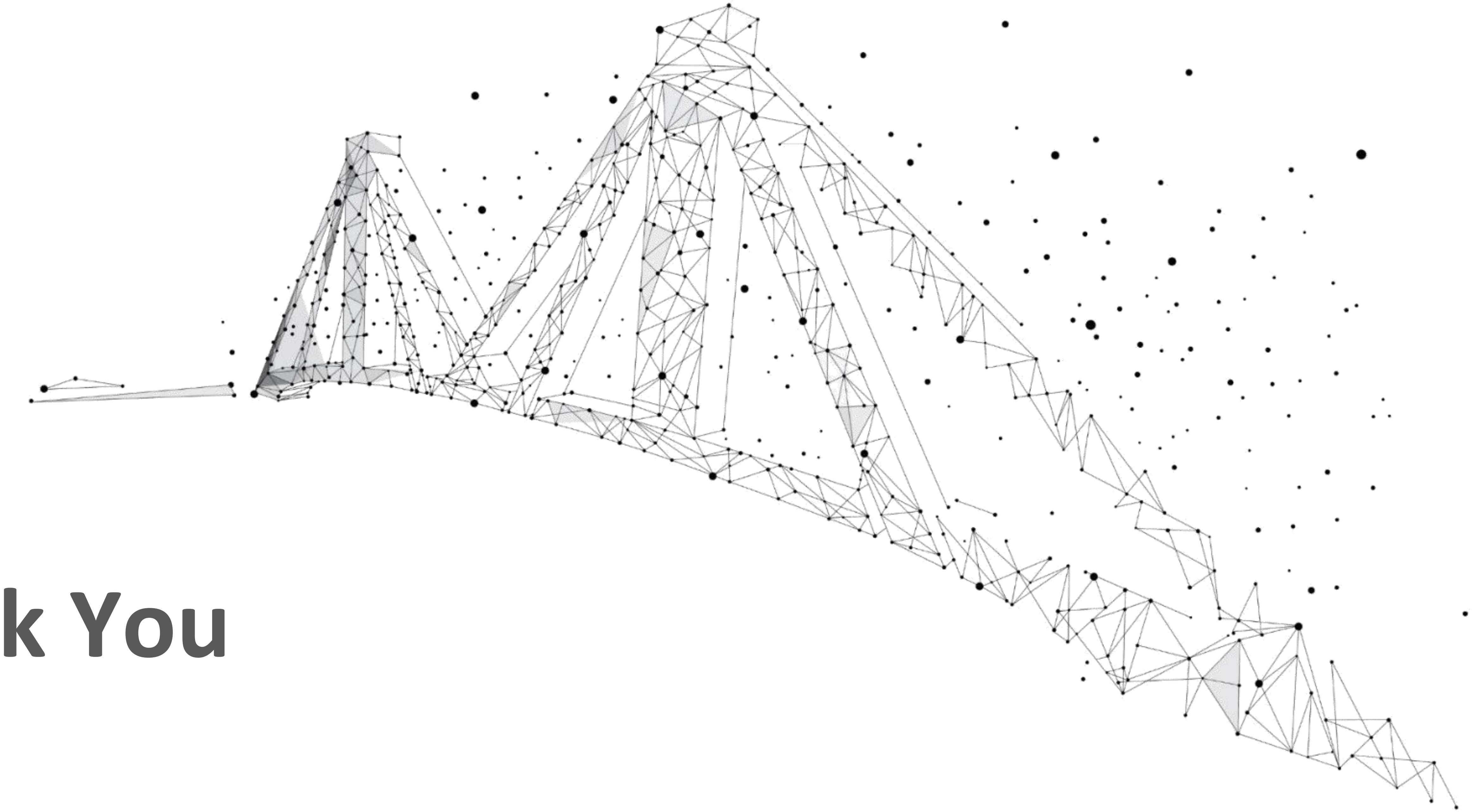
Resnet50 Performance Results



Resnet50 Container Performance Results



35% Improvement!



Thank You