LIQID

Unleashing AI/ML Superpowers

With Liqid CDI, Orange Silicon Valley:

» Transformed a traditional server into an AI powerhouse.

» Scaled performance linearly from one to sixteen GPUs.

» Achieved exceptional ImageNet training and NLP task results.

» Flexibly designed without proprietary components or vendor lock-in.

orange™

Unleashing AI/ML Superpowers: Orange Silicon Valley Transforms Standard Servers with Liqid Matrix CDI

The demand for GPU-intensive workloads is exploding, driven in large part by the rise of artificial intelligence (AI) and machine learning (ML) applications. These applications require significant amounts of processing power to train complex neural networks, analyze large datasets, and perform other compute-intensive tasks. ImageNet training using RESNET50 over TensorFlow and NLP via transformer for PyTorch are two of the most popular AI frameworks used for computer vision and natural language processing (NLP) tasks, respectively. Both frameworks require significant amounts of GPU processing power to achieve optimal performance.

The standard approach to deploying GPUs, which involves physically installing them within server chassis, presents significant performance and cost challenges when addressing the GPU demands of modern workloads. First, the server chassis limit GPU density and performance, at maximum of roughly four for standard servers and eight for GPU-optimized servers. Furthermore, any underutilized GPU resources are trapped within individual servers, and must be manually moved to be shared with other workloads, which constrains performance, squanders resources, and increases costs.

Liqid Matrix transforms GPU scale and efficiency for AI/ML workloads. The underlying technology is Composable Disaggregated Infrastructure (CDI). And as the name implies, PCIe devices including GPUs and NVMe storage are composed into servers from pools of disaggregated resources via software. By eliminating the physical server chassis barrier, resources like GPUs can be provisioned, scaled and migrated in seconds. Furthermore, customers can easily add up to 20 GPUs to their standard 2U servers, instead of purchasing expensive, purpose-built GPU servers that only hold eight.

Orange Silicon Valley (OrangeSV) is a great example of how Liqid helps our customers achieve greater results. A team consisting of members from Orange Silicon Valley and the Orange Innovation Data & AI organizations, conducted natural language processing (NPL) training on a composable server with a standardized English-to-German translation task.

For their testing they used Liqid to expose sixteen NVIDIA A100s (40GB) GPUs and one Liqid LQD-4500 NVMe device to a single Dell PowerEdge R6525 with two AMD EPYC 7502P 32 core processors and 1 TB of memory. All GPUs had Liqid ioDirect P2P technology turned-on, which enables direct GPU-to-GPU and GPU-to-Storage communication.

OrangeSV's Composable GPU Pod

1x Liqid Matrix Software

1x Dell PowerEdge R6525 1U

1x Liqid Director - Management Node

1x Liqid Gen4 PCIe Switch

2x Liqid Gen4 Expansion Chassis (8-Slot)

Data (PCIe Gen4x16)

Management

16x NVDIA A100 40GB GPU Connected to a 1U Server

8x NVIDIA A100 40GB GPU

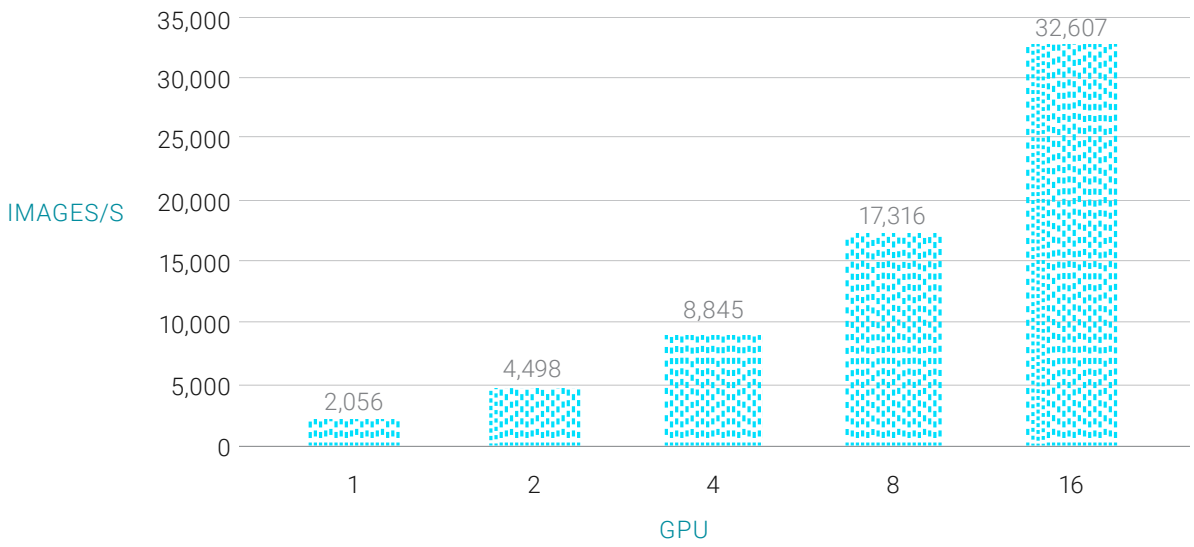
8x NVIDIA A100 40GB GPU

Testing Performed

Throughput where a Resnet50 model was trained with TensorFlow with batch sizes of 512 and 768.

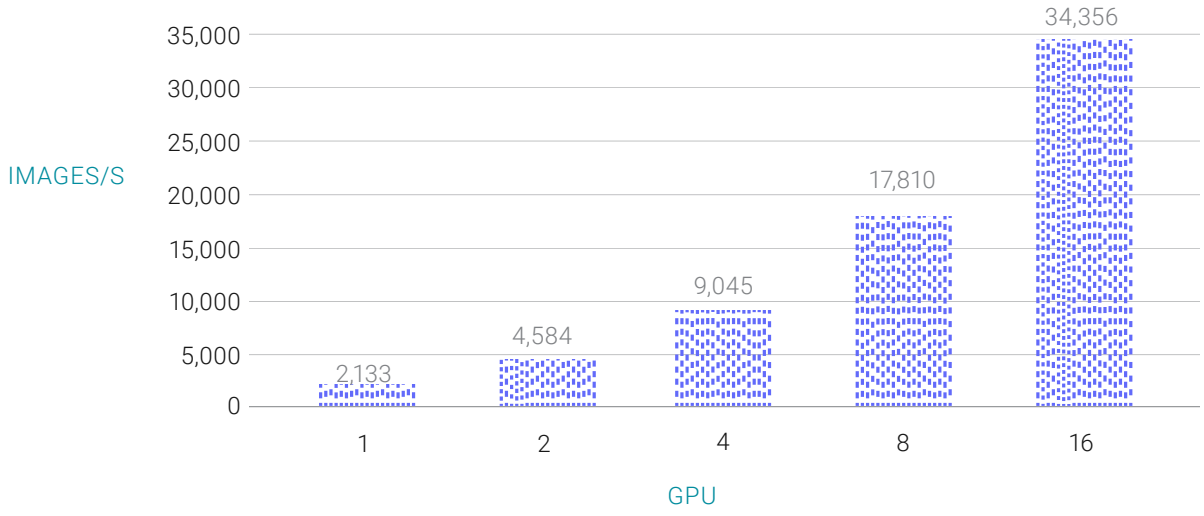
ImageNet Results

NVIDIA A100 running ImageNet training using ResNet 50 v1.5 over TensorFlow, Batch size of 512.



ImageNet Results

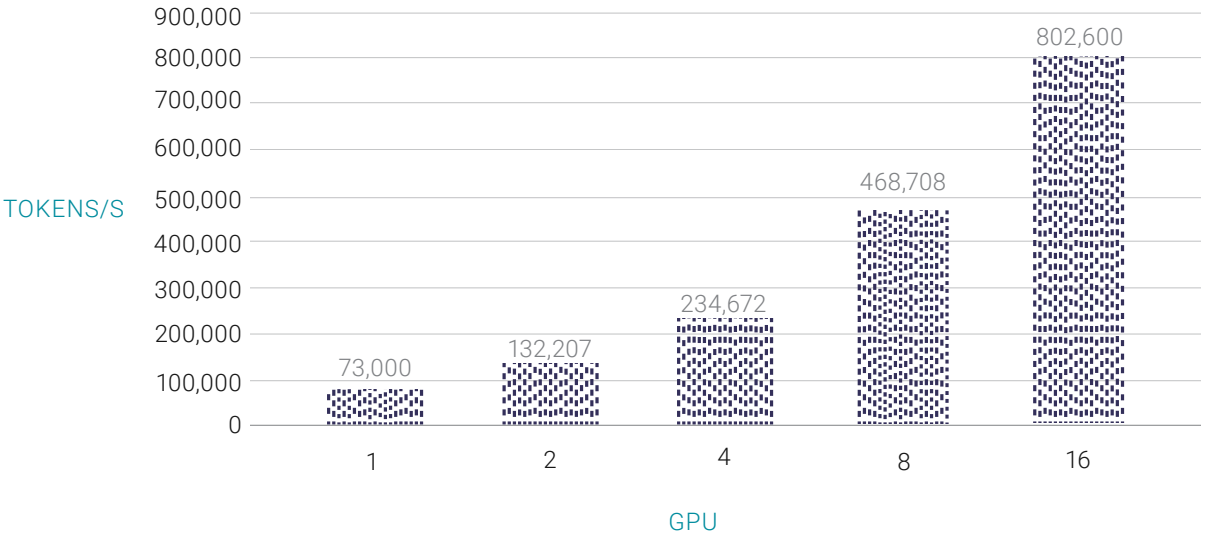
NVIDIA A100 running ImageNet training using ResNet 50 v1.5 over TensorFlow, Batch size of 768.



Next, OrangeSV tested NLP Via Transformer for pyTorch. Transformer is currently the backbone for neural network architecture for training complex and structured NLP use cases. They executed the standard WMT14_en_de benchmark to collect a standard baseline:

ImageNet Results

Nvidia A100 GPU running WMT14_en_de training throughput benchmark using transformer over PyTorch; Batch Size = 10240.



Additionally, OrangeSV ran full training to achieve minimal validation loss, which was reached within two hours with a batch size of 10,240 tokens (as part of words split of the SentencePiece software). They attempted to increase the batch size, and we were able to squeeze in a maximum batch size of 16,000. This allowed them to achieve a throughput of 935,343 tokens/second and were able to reach the minimal validation loss (the objective function the Neural Net is trying to minimize) under 1 hour and 49 mins.

The Orange Silicon Valley (OrangeSV) case study demonstrates the power of Liqid Matrix Composable Disaggregated Infrastructure (CDI) in addressing the GPU demands of modern AI/ML workloads. By using Liqid Matrix to compose 16 A100 GPUs and Liqid's 16TB IO Accelerator card into a single Dell server, OrangeSV achieved remarkable performance and efficiency for ImageNet training and natural language processing tasks. This groundbreaking approach enabled OrangeSV to transform a standard server into a multi-GPU single-node supercomputer that can scale from 1 to 16 GPUs linearly, proving the immense potential of Liqid's CDI solution for optimizing GPU resource utilization and meeting the most demanding AI/ML workloads.