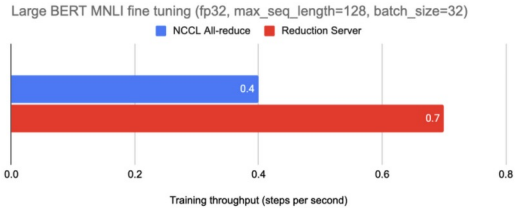
Vertex AI Training Reduction Server

Optimize distributed GPU training for synchronous data parallel algorithms



Adding 20 reduction server nodes increased the training throughput by **75%**.

Vertex Job Configuration	GPU workers	GPU workers per hour cost	RS workers per hour cost	Cost per step
NCCL All-reduce	8 x a2-highgpu-8g	$245.73	N/A	$0.17
Reduction Server	8 x a2-highgpu-8g	$245.73	$11.34	$0.10

Reduced cost per step by **42%** even with additional nodes.

Training time is a key bottleneck

The exponential growth of datasets and model sizes has caused training time to become one of the key bottlenecks in the development and deployment of ML systems.

Distributed training at scale is difficult

Limited network bandwidth between nodes makes optimizing performance of distributed training inherently difficult, particularly for large cluster configurations.

Reduction Server allows for faster distributed training

A faster GPU all-reduce algorithm developed at Google optimizes bandwidth and latency of multi-node distributed training on NVIDIA GPUs for synchronous data parallel algos.

Framework agnostic

Support mainstream deep learning frameworks including TensorFlow, PyTorch, JAX.

Seamless integration with existing training jobs

Zero-touch enablement for your multi-worker GPU training jobs on Vertex AI Training with optimized training bandwidth and latency.