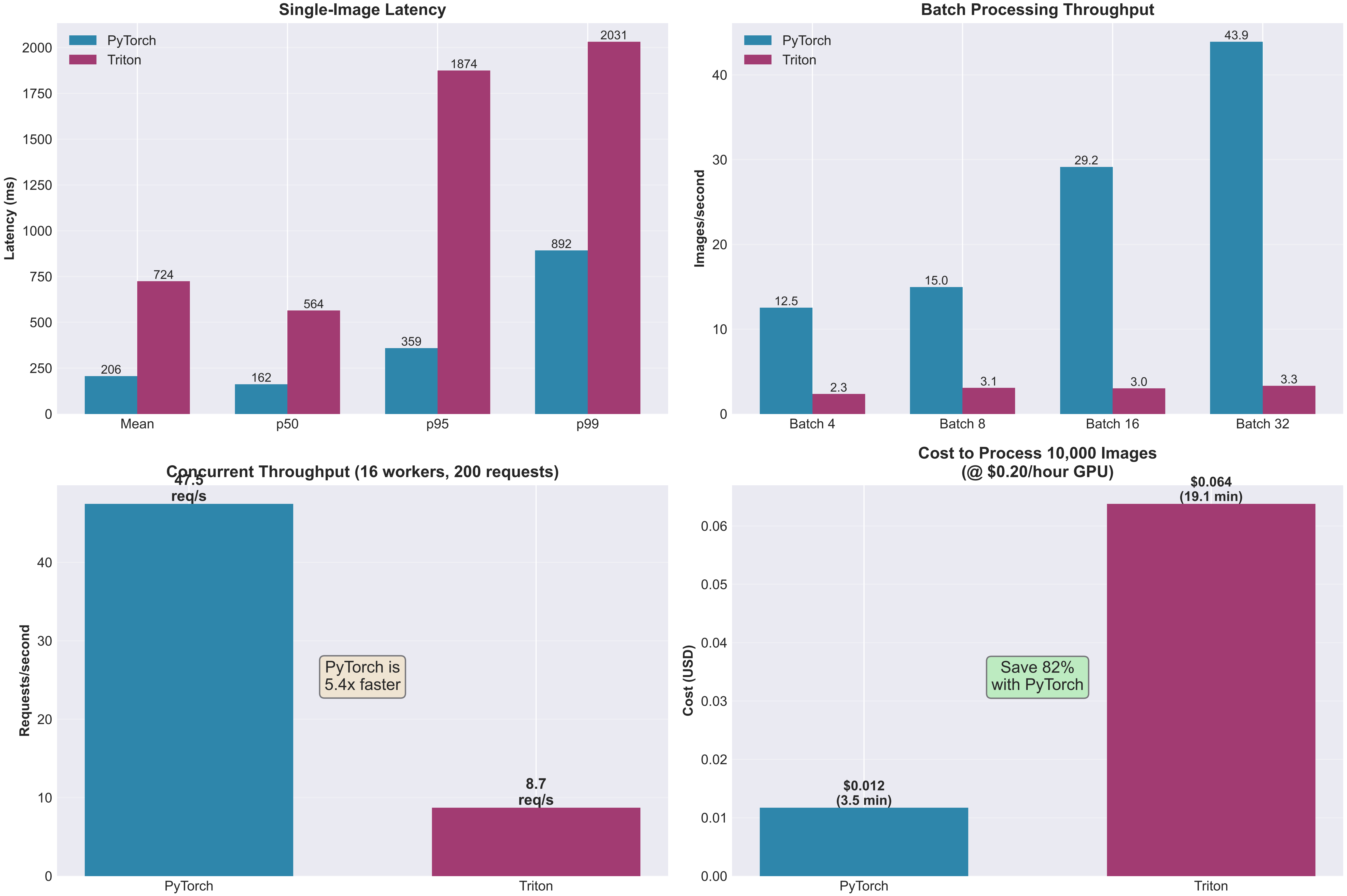
PyTorch vs Triton Inference Server
OpenCLIP ViT-B-32 on RTX 3070



Single-Image Latency

Batch Processing Throughput

Concurrent Throughput (16 workers, 200 requests)

Cost to Process 10,000 Images
(@ $0.20/hour GPU)