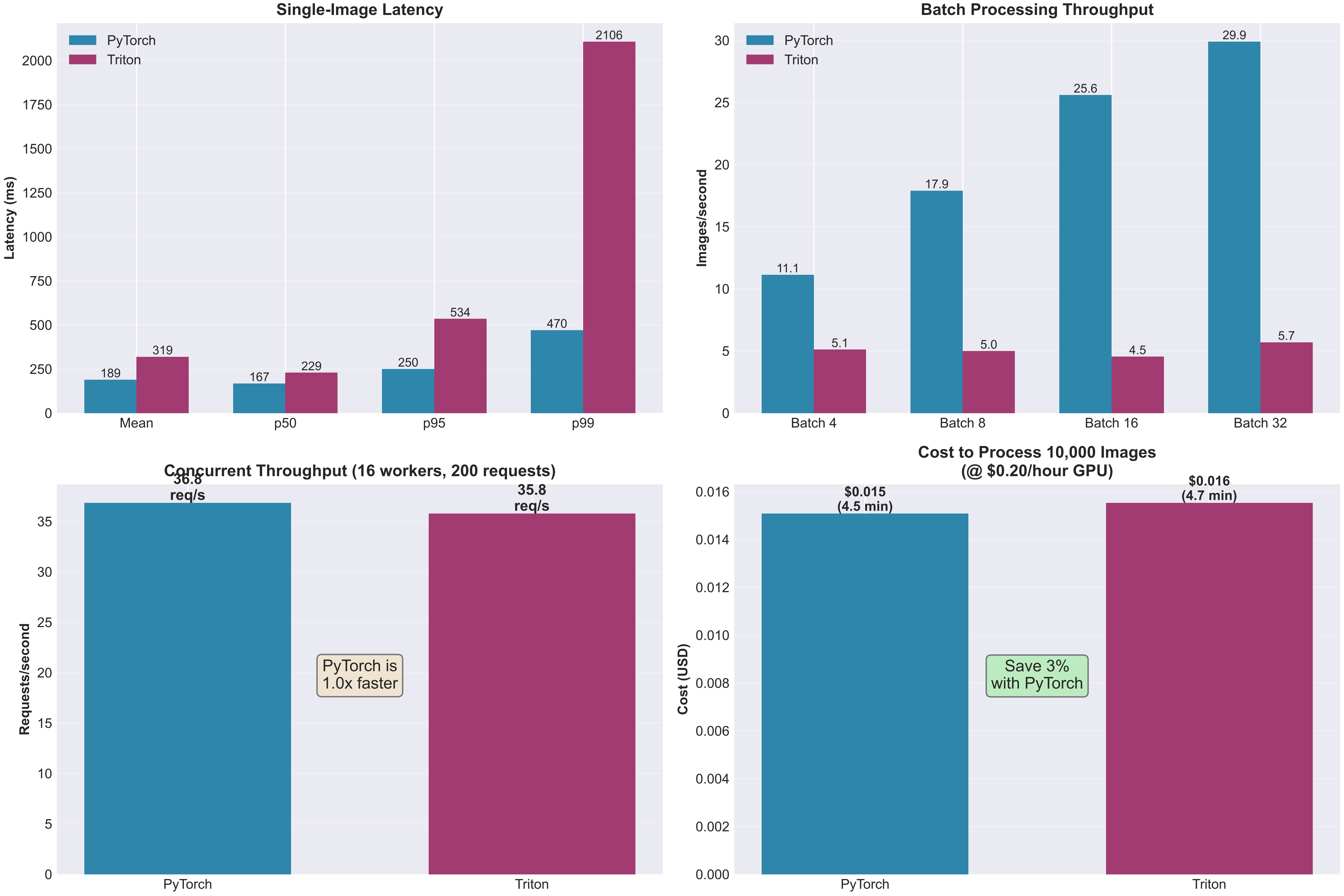
PyTorch vs Triton Inference Server
OpenCLIP ViT-B-32 on RTX A4000


Single-Image Latency
PyTorch
Triton
Latency (ms)
Mean
p50
p95
p99
Batch Processing Throughput
PyTorch
Triton
Images/second
Batch 4
Batch 8
Batch 16
Batch 32
Concurrent Throughput (16 workers, 200 requests)
Requests/second
req/s
req/s
PyTorch is
1.0x faster
PyTorch
Triton
Cost to Process 10,000 Images
(@ $0.20/hour GPU)
Cost (USD)
$0.015
(4.5 min)
$0.016
(4.7 min)
Save 3%
with PyTorch
PyTorch
Triton