Economic Analysis of High Performance Computing (HPC) Technology: Optimizing Cost-Effectiveness in Hardware Investments

Noah Bean

February 23, 2025

Abstract

This report examines the intersection of economics and High Performance Computing (HPC) technology, focusing on cost-effectiveness in hardware investments for scientific research clusters in university and government laboratories. Through detailed economic analysis, we find that multi-CPU servers exhibit diminishing marginal returns, making them less cost-effective beyond moderate scaling; interconnect upgrades offer significant performance gains at lower costs; and GPUs provide superior performance per cost for AI training compared to CPU-based alternatives. These insights, supported by the novel Price Elasticity of Computational Performance (PECP) metric and extensive literature, guide optimal budget allocation as of February 23, 2025.

1 Introduction

High Performance Computing (HPC) is vital for scientific research, enabling complex simulations, data analysis, and AI training in university and government settings. However, economic efficiency remains critical under budget constraints. This report addresses the hypothesis that HPC centers may overspend on hardware with marginal returns, analyzing multi-CPU server scaling, interconnect technology, and GPU versus CPU performance for AI training. Economic principles such as diminishing returns and elasticity are integrated to identify cost-effective strategies.

2 Methodology

The analysis integrates economic modeling, hardware performance data, and literature reviews based on iterative queries. Methods include:

- **Cost and Performance Estimates**: Market prices (AMD EPYC 9554P at $4,550) and benchmarks (PassMark, TFLOPS).
- **Economic Metrics**: Developed the Price Elasticity of Computational Performance (PECP), defined as:
$$\text{PECP} = \frac{\%\text{ change in FLOPS}}{\%\text{ change in price}}$$
- **Literature Review**: Academic papers (Springer, IEEE), industry reports (Hyperion Research), and benchmarks (MLPerf).
- **Use Cases**: Scientific HPC workloads (simulations) and AI training (BERT).

3 Economic Analysis of Multi-CPU Servers

3.1 Hypothesis and Data

We tested whether HPC centers pay excessively for marginal returns when scaling multi-CPU servers, using an AMD EPYC 9554P (64 cores, 5.12 TFLOPS FP32) in 1, 2, and 4 CPU configurations.

3.1.1 Cost and Performance Metrics

Configuration	Total Cost	FLOPS (TFLOPS)	PassMark Score
1 CPU	\$5,150	2.5	110,733
2 CPUs	\$10,300	4.5	199,319
4 CPUs	\$20,700	8.5	376,492

Table 1: Cost and Performance Metrics for Multi-CPU Configurations

- **Scaling Efficiency**: 90% from 1 to 2 CPUs, 85% from 2 to 4 CPUs.
- **PECP Calculation**:
  - 1 to 2 CPUs: PECP $= \frac{80\%}{100\%} = 0.80$
  - 2 to 4 CPUs: PECP $= \frac{89\%}{101\%} \approx 0.88$
  - 1 to 4 CPUs: PECP $= \frac{240\%}{302\%} \approx 0.79$

3.2 Insight: Diminishing Returns in Multi-CPU Servers

PECP values below 1 indicate inelastic performance growth, where costs outpace gains. From 1 to 2 CPUs, performance rises 80% while cost doubles; from 2 to 4 CPUs, a 101% cost increase yields 89% performance. Multi-CPU servers are not cost-effective beyond moderate scaling, as additional CPUs face diminishing returns due to communication overhead.

4 Cost-Effectiveness of Interconnect Technology

4.1 Hypothesis and Analysis

We investigated whether interconnect upgrades enhance performance at lower costs, addressing communication bottlenecks.

4.1.1 Scenario: Interconnect Upgrade

- **Baseline**: 100-node cluster, 10 Gbps Ethernet, \$1,000,000, performance P.
- **Upgrade**: 100 Gbps InfiniBand, \$200,000 additional cost.
- **Performance Impact**: 20% time reduction, 25% performance increase.

4.1.2 Comparison to Adding Nodes

- **Adding 25 Nodes**: \$250,000 for 25% performance, PECP ≈ 1.
- **Interconnect Upgrade**: \$200,000 for 25% performance, PECP $= \frac{25\%}{20\%} = 1.25$.

4.2 Insight: Interconnects as a Cost-Effective Solution

Interconnect upgrades yield elastic gains (PECP > 1), outperforming node additions. For communication-intensive workloads, high-performance interconnects like InfiniBand enhance efficiency, making them a strategic investment.

5 GPUs vs. CPU-Based SIMD Machines for AI Training

5.1 Hypothesis and Feasibility

We explored whether multiple CPU threads with vector processing (AVX-512) could outperform GPUs for AI training cost-effectively.

5.1.1 Cost and Performance Comparison

- **CPU**: AMD EPYC 9554P, $4,550, 5.12 TFLOPS FP32.
- **GPU**: NVIDIA A100, $10,000, 19.5 TFLOPS FP32, 312 TFLOPS FP16.
- **Scaling**: 4 CPUs ($18,200) to match 1 A100 FP32.

5.1.2 AI Training Benchmarks

- **BERT Training**: V100 GPU (2 days) vs. 640-core CPU cluster (24 hours, higher cost).
- **MLPerf**: GPUs 5-10 times faster.

5.2 Insight: GPUs' Superior Cost-Effectiveness

Scaling CPUs to match GPU performance is costly and inefficient. GPUs' tensor cores provide unmatched FP16 performance, making them more cost-effective for AI training despite higher unit costs.

6 Discussion

6.1 Intersection of Economics and HPC

PECP highlights diminishing returns in multi-CPU scaling (PECP < 1), elastic gains from interconnects (PECP > 1), and GPU efficiency, reflecting a shift to balanced system design.

6.2 Surprising Insights

- Multi-CPU servers show poor marginal returns despite high ROI claims.
- Interconnects, a small cost fraction, offer outsized benefits.
- GPUs outpace CPUs in cost-effectiveness for AI, despite lower CPU unit costs.

7 Recommendations

- Limit multi-CPU scaling to 2 CPUs per node.
- Prioritize interconnect upgrades for communication-heavy tasks.
- Use GPUs for AI training to maximize performance per dollar.

8 Conclusion

Economics insights from HPC appear to agree with computer architecture first principles.Multi-CPU servers are not cost-effective beyond moderate scaling, interconnects offer high returns, and GPUs excel for AI training. PECP guides HPC centers to optimize investments as of February 23, 2025.