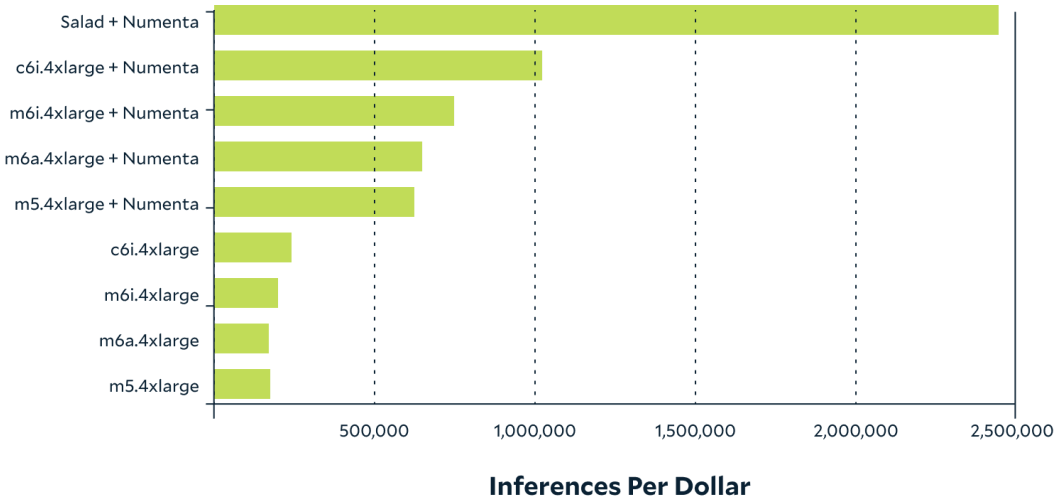
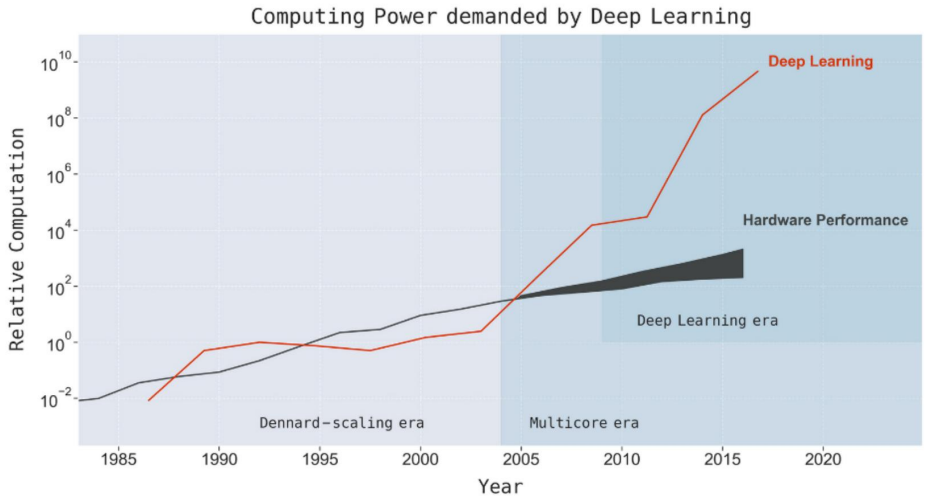
Optimizing AI Inference Costs: Numenta on Salad's Cloud Infrastructure (2022)



Abstract

This study benchmarks the price performance of BERT standard transformer networks and core technologies from Numenta's novel brain-based platform for practical AI systems. To establish a baseline and assess the potential for further cost optimizations in conjunction with decentralized cloud infrastructure, Numenta engineers deployed container workloads to both Amazon Web Services (AWS) and Salad Container Engine (SCE). Numenta's optimized BERT framework consistently achieved up to 6.5x more queries per second than conventional transformer networks running on AWS infrastructure. Deployment via SCE proved to be the most efficient configuration, during which Numenta attained 2,450,700 inferences per dollar, or a 10x improvement in cost efficiency over the nearest BERT base workload running on AWS.

1. Problem: Cost of Applied AI



"The Computational Limits of Deep Learning" by N. Thompson, et al. (2020)

The majority of today's AI solutions are single purpose, brittle, and prone to spiraling costs and unsustainable energy utilization. Despite the broad utility of many commercially available and open-source deep learning models, few have managed to make more than incremental optimizations to the cost, energy efficiency, and compilation time of generic inference.

Deploying practical AI at scale typically requires the distribution of large data sets to complex networks of specialized hardware. Due to the inherently persistent nature of such applications, many researchers and data scientists choose to deploy their transformer networks to the public cloud.

As AI models become larger and practical AI systems become increasingly complex, cloud customers will need to provision increasingly expensive resources to fit the stringent technical requirements of such high-performance applications. Though deep neural networks have facilitated significant advancements in recent years, their fundamental reliance on highly available processing resources and their tendency toward rapid expansion continue to render them costly and inefficient to run on public cloud infrastructure.

Progress and iteration are therefore limited by the implicit constraints of both cloud spending and conventional AI modeling frameworks.

2. Numenta Intelligent Computing

Numenta solutions deliver dramatic improvements in throughput, latency, and price performance through an innovative AI technology platform modeled on the efficient mechanisms and structures of the human brain.

Using hardware-aware optimizations and neuroscience-based acceleration techniques, Numenta's solutions take advantage of the cache hierarchies and SIMD instruction sets prevalent in today's CPU hardware. With their AI expertise, they're able to train deep learning networks that can be highly performant.

Grounded in the sensorimotor framework of intelligence described in co-founder Jeff Hawkins's bestselling book, *A Thousand Brains*, Numenta's optimized AI solutions accelerate inference, minimize overall resource usage, and reduce the number of cycles needed to perform predictions.

3. SaladCloud Distributed Infrastructure

Distributed computesharing networks such as SETI@home, Folding@home, and the BOINC platform have demonstrated that there is a continuously replenished supply of performant and highly available consumer hardware that could readily function as affordable AI accelerators. Salad designed its solutions to ameliorate the dual issues of workload specificity and intrinsic motivation that have heretofore prevented private enterprises from following their example.

Every Salad node instance represents a real, actually available, privately owned computer. Individuals around the world use Salad's open-source desktop software to share idle compute resources in exchange for redemption credits and digital rewards on a community platform. In return, Salad sustainably activates an underutilized supply of latent processing power, bandwidth, and other compute resources as a globally distributed cloud infrastructure layer.

This mutually beneficial relationship permits Salad to offer more affordable on-demand pricing than centralized cloud service providers for analogous services. By virtue of this voluntary computesharing arrangement, SaladCloud customers gain secure access to dedicated CPU processing, cutting-edge GPUs, high-speed internet connections, and non-sequential public IP addresses from residences worldwide, for far less than the price of a conventional data center.

Salad Container Engine is a fully managed orchestration platform built to facilitate container deployments to ideal machine cohorts on the Salad network. In order to serve SCE workloads, host hardware must earn a sufficient eligibility score based on factors that include technical specifications, connection quality, typical uptime, and trustworthiness based on its unique performance record in completing trustless computing tasks. On host machines, the Salad node client creates secure, stateless execution contexts in virtual environments for every container instance.

4. Case Study

In partnership with Salad, engineers from Numenta benchmarked their optimized BERT framework against a standard BERT base distribution using both Salad Container Engine (SCE) and premium cloud services from Amazon Web Services (AWS).

The study included four (4) benchmark trials running a conventional BERT base transformer network on various AWS configurations (Trials 1–4, below), four (4) optimized trials running Numenta's optimized BERT framework on the same configuration (Trials 5–8), and a final trial (Trial 9) that assessed the net price-performance benefits of deploying Numenta technologies across distributed SCE instances.

In side-by-side comparison trials, Numenta technologies consistently performed 5.4–6.5x more efficiently than standard transformer networks. When deploying on Salad's affordable infrastructure, Numenta engineers achieved 10x more inferences per dollar spent.

Analysis: Throughput Improvements

A summary of the instances, workloads, and host hardware used in this benchmark is provided below. For the purposes of accuracy, the administrators selected a representative processor cohort from Salad's network of heterogeneous hardware to compare with a premium c6i.4xlarge AWS instance. All trials featured a batch size of one, and a sequence length of 128.

Numenta technologies consistently delivered **5.4–6.5x more queries per second** than a standard BERT transformer network.

Trial	Instance	Workload	Processor	RAM (GB)	Cores	vCPU	Throughput (QPS)	Inference Optimization
1	m5.4xlarge	BERT base	Intel Xeon Platinum 8259CL	64	8	16	37.5	—
2	m6a.4xlarge	BERT base	AMD EPYC 7R13	64	8	16	32.6	—
3	m6i.4xlarge	BERT base	Intel Xeon Platinum 8375C	64	8	16	42.9	—
4	c6i.4xlarge	BERT base	Intel Xeon Platinum 8375C	32	8	16	45.8	—
5	m5.4xlarge	Numenta Optimized BERT	Intel Xeon Platinum 8259CL	64	8	16	**203.4**	**5.4x**
6	m6a.4xlarge	Numenta Optimized BERT	AMD EPYC 7R13	64	8	16	**185.8**	**5.7x**
7	m6i.4xlarge	Numenta Optimized BERT	Intel Xeon Platinum 8375C	64	8	16	**279.7**	**6.5x**
8	c6i.4xlarge	Numenta Optimized BERT	Intel Xeon Platinum 8375C	32	8	16	**288.8**	**6.3x**

Analysis: On-Demand Price Performance

When deployed on SCE, Numenta's customers can attain **10x better price performance** over BERT on AWS.

Trial	Instance	Workload	Hourly Cost	Inferences per Hour	Inferences per Dollar	Cost per 1M Inferences	Cost Efficiency vs. AWS
1	m5.4xlarge	BERT base	$0.77	135,000	175,325	$6.00	—-
2	m6a.4xlarge	BERT base	$0.69	117,360	170,087	$5.88	—-
3	m6i.4xlarge	BERT base	$0.77	154,440	200,571	$4.99	—-
4	c6i.4xlarge	BERT base	$0.68	164,880	242,471	$4.12	—-
9	SCE	Numenta Optimized BERT	**$0.40**	**980,280**	**2,450,700**	**$0.41***	**10.11x**

**Pricing is subject to change.*

Analysis: Cost Efficiency vs. Spot Pricing

This table compares the price performance of on-demand SCE instances and spot instances from AWS. On SCE, Numenta's optimized BERT server outperformed the cost efficiency of the nearest spot-basis AWS instance by **2.39x**.

Instance	Pricing	Workload	Throughput (QPS)	$/hr	Inferences /hr	Inferences/ $	Cost Per 1M Inferences	Cost Efficiency vs. AWS
m5.4xlarge	On Demand	BERT base	37.5	$0.77	135,000	175,325	$5.70	
m6a.4xlarge	On Demand	BERT base	32.6	$0.69	117,360	170,087	$5.88	
m6i.4xlarge	On Demand	BERT base	42.9	$0.77	154,440	200,571	$4.99	
c6i.4xlarge	On Demand	BERT base	45.8	$0.68	164,880	242,471	$4.12	
m5.4xlarge	Spot Instance	BERT base	37.5	$0.21	135,000	635,294	$1.57	
m6a.4xlarge	Spot Instance	BERT base	32.6	$0.18	117,360	637,134	$1.57	
m6i.4xlarge	Spot Instance	BERT base	42.9	$0.20	154,440	760,414	$1.32	
c6i.4xlarge	Spot Instance	BERT base	45.8	$0.16	164,880	1,026,650	$0.97	
SCE	Flat Rate	Numenta Optimized BERT	**272.3**	**$0.40**	**980,280**	**2,450,700**	**$0.41**	**2.39x**

5. Conclusion

Salad and Numenta have partnered to support AI innovation. In a marketplace where the cost of conventional cloud infrastructure can hinder inference at scale, optimized technologies and novel orchestration systems present the best opportunity to test the bounds of possibility, derive more meaningful inferences, and launch never-before-seen categories of discrete applications.

To apply for access to Salad Container Engine, please visit our beta onboarding guide. If you would like to explore the performance improvements of Numenta's unique technologies, please apply for private beta access.

About Numenta

Numenta has developed new artificial intelligence technologies that deliver breakthrough performance in AI/ML applications such as natural language processing and computer vision. Backed by two decades of neuroscience research, Numenta's novel architectures, data structures, and algorithms deliver disruptive performance improvements. Numenta is currently engaged in a private beta with several Global 100 companies and startups to apply its platform technology across the full spectrum of AI, from model development to deployment—and ultimately enable novel hardware architectures and whole new categories of applications.

About Salad Technologies

Salad is the world's largest distributed cloud platform—and the easiest, most trusted way to profit from your PC. Private individuals share latent compute resources through the Salad desktop application in exchange for personalized rewards from our community platform. With integrated support for standard container development workflows and a high-trust network of GPU-enabled machines, Salad Container Engine powers cutting-edge compute applications for a fraction of the cost of conventional data center resources. Salad hopes to build a sustainable, more equitable Internet in which everyone can support innovation right from home by activating underutilized compute capacity as affordable cloud infrastructure.