OLDS RESEARCH

The Real Cost of AI Infrastructure

A TCO Analysis of On-Premises vs Cloud

May 2026

Executive Summary

There's been a tipping point over the last couple of years as AI moved from experiments to pilot projects to the point where organizations have decided that there are enough benefits from AI to justify spending big money on it. That's when AI becomes infrastructure.

Over the next several years, we believe that AI won't be discussed as a separate "thing." AI features like machine learning, deep learning, and generative AI will be folded into any application or workflow where it makes sense (and plenty where it probably doesn't make sense as well).

In 2025 alone, the global AI infrastructure market (AI hardware, not facilities) was $318 billion according to IDC, more than double the $153 billion spent in 2024. This is expected to continue unabated for the foreseeable future.

The problem that many organizations are wrestling with right now is not the "if" they're going to need an AI infrastructure, but the "what" and "where" questions.

The "what" question is highly dependent on unique organizational factors like what they see as the low hanging AI fruit for their business and how big a move they want to make.

In this report, we are addressing the "where" question in terms of on-premises vs. public cloud options and the respective costs.

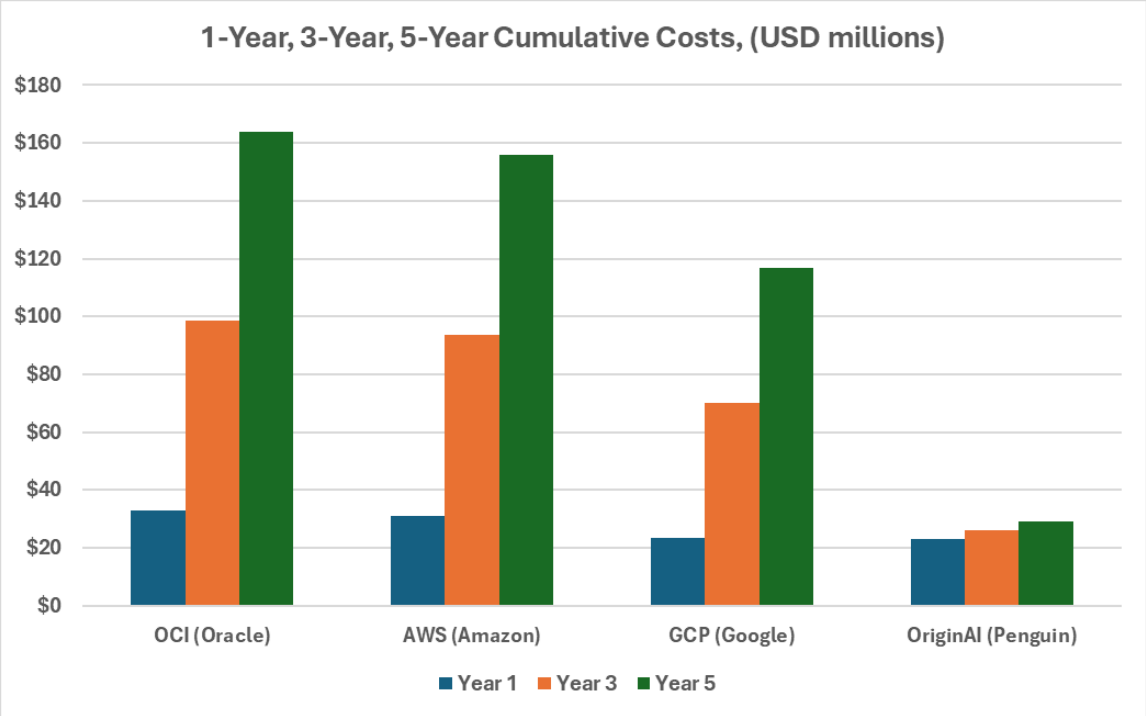
We compared the estimated cost of deploying and operating a production-scale AI infrastructure on-premises versus in three major cloud platforms: Amazon Web Services (AWS), Google Cloud Platform (GCP), and Oracle Cloud Infrastructure (OCI). The comparison is apples-to-apples, based on publicly available pricing and clearly defined workload assumptions. The model assumes steady-state, 24×7 operation of a 248-GPU enterprise AI cluster.

The scenario examines a mature, diversified manufacturing organization that has already completed extensive AI pilot programs across multiple business units.

Those efforts have demonstrated measurable operational improvements in areas such as defect reduction, scheduling, supply chain optimization, and AI-augmented HPC workloads. Based on validated demand, the organization is evaluating a

centralized, shared AI infrastructure to support ongoing model training, testing, and development.

Cloud configurations are provisioned to deliver comparable performance and continuous enterprise-scale capacity under committed pricing. All costs are derived from publicly available price sources at the time of analysis.



Under these conditions, the cost difference is considerable. Over a three-year period, the cloud options are estimated to cost an average of over 3.5× more than the on-premises deployment. Over five years, that differential widens further.

The cost spread is an illustration of rent vs buy economics. Simply put, if you use something a lot and it's a significant cost, in most cases you're better off purchasing it than renting it.

If you change the requirements and assumptions, the numbers will change and sometimes the decision will change as well. In this report, we clearly state the requirements/assumptions, configure on-premises and Cloud Service Provider (CSP) environments to best satisfy the conditions, then cost them out.

OLDS Research

Table of Contents

Why This Report, Why Now?

Everyone who has even casually kept up with business or technology knows that we're in the midst of a huge move to add AI functionality to nearly every aspect of business.

The key question that many organizations are wrestling with now is how to add the necessary technology to enable their AI initiatives. Do they add new servers, GPUs, and infrastructure to their existing data center? Or do they rent the capacity in the cloud? And what can they expect to pay in either case?

I've been curious for a while about what it really costs to run serious IT (and now AI) infrastructure on-premises versus in the public cloud.

Cloud pricing is public. Anyone can go into Amazon Web Services (AWS) or Google Cloud pricing tools and model detailed configurations. But it's much harder to obtain a fully built-out, enterprise-sized on-premises with real pricing attached. Without that, an apples-to-apples comparison is largely speculation.

Penguin Solutions was interested in the same question. The surge in AI investment has created a wave of infrastructure decision points, and organizations are actively debating where their long-term AI backbone should reside. Penguin funded the time required for me to conduct a detailed cost analysis of a 248-GPU on-premises cluster and compare it against AWS, Google Cloud, and Oracle Cloud.

To be clear, Penguin provided a detailed hardware configuration and associated pricing for the on-premises system. I built the cloud models by using publicly available pricing tools. I defined the workload assumptions, utilization model, storage and network requirements, and ran the cost comparisons. The methodology and conclusions in this report are mine.

Prior work in HPC and infrastructure economics has given me a general sense of how this might turn out. The underlying economics aren't mysterious. If you're visiting a city for a week, you book a hotel. If you're staying for a year, you rent an apartment. If you're settling in for the long term, you buy a house. The longer and steadier the demand, the stronger the economic case for ownership.

That doesn't mean hotels are wrong. We absolutely need hotels. We also need apartments and houses. But one model doesn't fit every situation. The same is true for AI infrastructure.

But analogies aren't analysis. The point of this report is to move beyond intuition and run the numbers under clearly defined real-world conditions.

Building the Model

We approached this comparison the same way we would approach any infrastructure evaluation: define the workload first, then build systems to support it.

The model assumes persistent usage and high utilization plus 24×7 operation of a centralized 248-GPU AI cluster serving multiple business units within a mature, global manufacturing organization. Utilization is modeled at approximately 80%, reflecting aggregate demand across a range of ongoing AI initiatives.

Cloud configurations were built to mirror the same GPU count, storage capacity, and overall performance envelope. No spot instances or preemptible capacity are used.

I selected Amazon AWS, Google Cloud Platform, and Oracle Cloud Infrastructure because they had instances with x86 CPUs coupled with NVIDIA B200 GPUs. Microsoft Azure did not have these instances available during our research period. We also did not analyze the smaller AI specialist cloud providers due to their relatively small size and newness to the market.

Where published three-year commitment pricing was not available (AWS B200 and OCI), we applied conservative discount assumptions based on historical precedent and typical enterprise negotiation patterns. Those assumptions are called out explicitly in the relevant sections.

Five-year totals are calculated by extending monthly pricing linearly. Why five years instead of three? Because major CSPs have extended their own server depreciation schedules to five years, and the Penguin Solutions on-premises configuration includes five-year break/fix service coverage.

No assumptions are made about renegotiation, future price reductions, or hardware refresh cycles. Vendors negotiate pricing all the time based on deal size, timing, and internal revenue targets. That's real, but it's outside the scope of this model.

Global Assumptions and Modeling Boundaries

Continuous 24×7 steady-state operation:

Large global organizations with multiple lines of business tend to have many AI efforts underway at the same time. Some are early pilots, others are testing phases, and some are full-scale training or retraining workloads. This cluster is designed primarily for AI development and training. It could handle inference workloads as well, but inference demand varies widely depending on the model and deployment pattern, so we're not looking at it specifically here.

Cluster Utilization Assumption:

Utilization is modeled at the aggregate cluster level, not per individual node. We are using 80% as the average cluster utilization rate. The idea is to reflect total demand across multiple concurrent initiatives rather than assume every GPU is fully saturated at all times. In centralized HPC and enterprise AI environments, sustained utilization in the 75–85% range isn't unusual when infrastructure is sized to the validated demand instead of speculative growth. This model shows operating reality.

High Performance Infrastructure Configuration:

AI training workloads are computationally intensive and require tightly integrated, high-performance infrastructure — essentially supercomputing-class systems. Equivalent configurations are available from both on-premises vendors and CSPs, and the cloud models were built accordingly.

No Spot or Preemptible Capacity:

The on-premises system is available to users 24×7, so the CSP configurations are held to the same standard. Many of the workloads modeled here run for days or even weeks. Allowing interruptions would require frequent checkpointing, which adds overhead and cost. Spot market pricing and availability also fluctuate significantly, making them better suited for short-term or burst workloads rather than production environments with ongoing high utilization.

Three-Year Commitment Pricing Where Available:

Committed pricing improves cost predictability and reduces cloud expense. In some cases, high-demand GPU instances are only offered under reserved or committed agreements, so this assumption mimics how large enterprises actually procure cloud infrastructure.

What This Model Does Not Attempt to Capture

- Software or application licensing beyond baseline system management tools.
- Managed services. These are widely available and customizable, but they are outside the scope of this infrastructure cost comparison. Break/fix service is included for the on-premises system since CSP infrastructure inherently includes hardware support.
- Financial treatment (CapEx vs OpEx). Depreciation schedules, tax treatment, and capital structuring vary by organization and aren't presented in depth here.
- Future pricing changes. All pricing is from publicly available data as of 1Q 2026.

AI Infrastructure Configuration Baseline

Before comparing specific vendor implementations, it's important to define the common configuration baseline used across all scenarios. The goal is not to model a hypothetical hyperscale environment or a small experimental cluster, but a serious, large-scale enterprise deployment sized to the validated demand.

This configuration assumes a large, mature organization consolidating AI development and training workloads into a centralized shared resource. It is intentionally large enough to expose meaningful economic differences, but not so large as to represent hyperscaler-scale infrastructure.

Cluster Size and GPU Generation

On Premises & CSP Configuration Summary				
Specification	**OriginAI (Penguin Solutions)**	**AWS**	**Google Cloud**	**OCI**
Compute Platform	Custom GPU cluster	p6-b200.48xlarge	a4-highgpu-8g	BM.GPU.B200.8
# Compute Nodes	31	31	31	31
GPUs (Total / Node)	248 / 8	248 / 8	248 / 8	248 / 8
GPU Type	NVIDIA B200	NVIDIA B200	NVIDIA B200	NVIDIA B200
CPU per Node	Intel Xeon (128 cores)	Intel (192 vCPUs)	Intel (224 vCPUs)	AMD EPYC (256 cores)
Memory per Node (GB)	2048	2048	3968	2048
Local Storage per Node	30.72 TB NVMe	30.72 TB SSD	28 TB SSD	54 TB SSD
Interconnect	400 Gb/s NDR InfiniBand	3200 Gb/s (EFA, RDMA)	3600 Gb/s (RDMA fabric)	3200 Gb/s (RoCE v2 RDMA)
Shared Storage	VAST Data, 2.5 PB usable	FSx for Lustre, 2.5 PB usable	Managed Lustre, 2.5 PB usable	OCI Lustre, 2.5 PB usable
Admin / Mgmt Nodes	8 × dual Intel 6767P	8 × i4i.32xlarge	8 × m3-megamem-128	8 × VM.Standard.E5.Flex
Deployment Model	On-prem cluster	Persistent (24x7)	Persistent (24x7)	Persistent (24x7)

(*detailed configurations included below)

The configuration is built around 248 NVIDIA B200 GPUs, deployed as 31 nodes with 8 GPUs per node.

The B200 generation was selected because it is a widely available, production-proven GPU architecture offered across the selected cloud providers and the on-premises vendor.

At the time of modeling, B300-class systems were emerging but not yet broadly available across all providers in comparable configurations. Using B200 GPUs allows for a consistent, apples-to-apples comparison.

Future GPU generations will improve performance per watt and per dollar. That improvement benefits both rental and ownership models. The focus here is the cost structure under sustained utilization, not generational benchmarking.

The 248-GPU scale was chosen to represent a substantial but realistic enterprise deployment. It is large enough to support multiple concurrent AI initiatives across business

units and to stress the economic model meaningfully, but it does not approach hyperscale data center magnitude.

Storage Performance Baseline

AI training workloads require high-throughput, parallel storage capable of sustaining large data flows to and from GPU nodes. Storage needs to deliver enough data to keep GPUs highly utilized and avoid idle time. The baseline configuration assumes storage performance consistent with approximately 250 MB/s per TiB in the cloud environments, aligned with enterprise-class managed parallel file system tiers.

On-premises storage is configured to deliver comparable or higher aggregate read/write throughput using a scale-out architecture. The intent is not to optimize for any single vendor's peak performance, but to ensure that all configurations meet a consistent performance envelope appropriate for large-scale AI training workloads.

Capacity and throughput were sized to avoid bottlenecks in either environment. Storage capacity was set at 2.5 petabytes of usable space.

This baseline intentionally focuses on AI development and training workloads. Inference environments vary significantly depending on model size, deployment pattern, and latency requirements, making them highly workload-specific and less suitable for generalized cost modeling at this scale.

Networking Requirements

Large-scale AI training depends on high-bandwidth, low-latency interconnects between GPU nodes. The baseline assumes a high-performance fabric appropriate for distributed training workloads.

On-premises configurations utilize modern high-speed interconnect technologies common in HPC-class deployments. Cloud configurations were selected to provide comparable intra-cluster networking capabilities within a single region.

All scenarios assume that the full cluster operates within a single region or data center environment to avoid cross-region latency penalties and to maintain training efficiency.

This configuration is optimized for distributed, multi-node training workloads rather than single-node fine-tuning tasks. While smaller fine-tuning jobs can be accommodated, the cluster is sized and interconnected to support large-scale synchronized training runs across many GPUs.

Performance Equivalence

All vendor configurations are constructed to meet the same general performance and capacity targets. The goal is not to claim architectural superiority for any deployment model, but to define a consistent baseline so that cost comparisons reflect economic structure rather than configuration differences.

This baseline is a realistic, enterprise-scale AI infrastructure for a mature company where AI models will need to be tested and trained at scale.

Configuring equally capable systems across on-premises and cloud environments is not trivial. The terminology, specifications, and packaging differ significantly between physical infrastructure and cloud-defined instances. Even among CSPs, there is limited common ground in how instances and performance characteristics are described.

For that reason, the following sections explicitly outline how each configuration was constructed. While the implementations differ in detail, each is designed to meet comparable performance expectations.

On-Premises Configuration: Penguin Solutions OriginAI Infrastructure

The on-prem configuration in this analysis is based on a detailed system specification and pricing provided by Penguin Solutions. The system is a centralized 248-GPU AI cluster intended to support sustained training workloads across multiple business units. The specifications below summarize the compute, storage, networking, and supporting infrastructure components included in the quoted configuration.

Compute Infrastructure (Summarized)

Component	Specification
GPU Node Configuration	31 nodes; 248 GPUs total (8 × NVIDIA B200 (192 GB) per node); Dual Intel Xeon CPUs; 2 TB RAM per node; 30.73 TB NVMe per node; NVIDIA 400 Gb/s InfiniBand

The cluster consists of 31 GPU nodes configured for distributed multi-node training workloads. Each node combines eight B200 GPUs with balanced CPU, memory, and high-speed interconnect resources.

Detailed system specifications are provided in Appendix A.

Storage Configuration (Summarized)

Storage Configuration

Component	Specification
Architecture	VAST Data scale-out storage
Configuration	64 VAST C boxes; 10 VAST D boxes
Usable Capacity	~2.5 PB
Aggregate Throughput	384 GB/s write; 600 GB/s read
Software Subscription	$354,000 per year

Storage capacity and throughput were sized to meet the cloud storage performance tier at approximately 250 MB/s per TiB. The configuration supports sustained training workloads and large dataset iteration cycles.

Detailed storage specifications are provided in Appendix A.

Networking Fabric (Summarized)

Component	Specification
Back-End Network	NDR InfiniBand (400 Gb/s per node); non-blocking, fully redundant
Front-End Network	400 Gb Ethernet; non-blocking GPU/storage connectivity; redundant architecture
Out-of-Band Network	Dedicated Gigabit Ethernet management network

The architecture separates GPU interconnect traffic, storage and in-band communications, and management functions. The GPU fabric operates in a fully redundant, non-blocking topology optimized for distributed training.

InfiniBand is used by 55% of the systems on the Top500 list of largest supercomputers and 68% of the top 100 supercomputers.

Detailed network topology and switch specifications are provided in Appendix A.

Integration, Support, and Cluster Management

The on-premises purchase price includes all rack hardware, cabling, and related infrastructure required to deliver a complete and fully functioning system. The quoted configuration is an integrated, deployable cluster rather than a partial bill of materials.

Integration and delivery services are included as part of the system price. This covers system assembly, configuration, validation, and deployment within the target data center environment.

Five-year Extended Platform Support is included and provides:

- 24×7 technical support availability
- Severity 1: 24×7 response, one-hour target

- Severity 2: 9×5 response, four-hour target
- Severity 3: 9×5 response, eight-hour target

This support coverage spans the full five-year modeling horizon used in this analysis. The configuration also includes Penguin's ClusterWareAI software platform, which provides centralized cluster monitoring, management, and operational tooling across compute, storage, and networking components. ClusterWareAI software is billed at $128,000 per year (invoiced monthly) and is included in the total cost of ownership calculations across the five-year period.

All integration, support, and management software costs are included in the total cost of ownership model and are not treated as optional add-ons. Detailed support terms, service scope, and software inclusions are provided in Appendix A.

Energy and Facilities Assumptions

Penguin provided measured power data for the cluster under maximum system load: **538 kW of IT load**. For the continuous usage model, we applied our **80% average utilization** assumption to convert peak test power into an estimated average operating load.

Annual Energy Calculation (Base Case)

- Peak tested IT load: **538 kW**
- Modeled average IT load (80%): **538 × 0.80 = 430 kW**
- Marginal PUE (air-cooled): **1.30**
- Incremental facility load: **430 × 1.30 = 559 kW**
- Annual hours: **8,760**
- Annual energy: **559 × 8,760 = 4,896,840 kWh**
- Electricity rate: **$0.12/kWh (US commercial rate average, Ohio)**
- **Estimated additional annual electricity cost: $587,621**

This approach treats the cluster as an incremental load within an existing data center and applies a marginal PUE to capture cooling and power overhead attributable to the added IT load.

On the facilities side, this model assumes that there is sufficient electrical service and data center floor space to accommodate the AI infrastructure cluster. This isn't true in all situations, of course. But there are some routes to pursue before looking for a co-location

deal. Large data centers nearly always have 'ghost' systems that are powered on, take up floor space, and yet have few or no users. A data center assessment uncovers these systems so they can be decommissioned, and their power/footprint devoted to newer and more efficient systems.

If this doesn't free up enough resources for the new cluster, then co-location is the best option, and there is a dizzying array of co-location vendors and plans available today.

Personnel Assumptions

Operating a centralized 248-GPU AI cluster requires dedicated infrastructure oversight. The model assumes three incremental full-time equivalents (FTEs) associated with ongoing cluster operations.

This includes:

- One senior cluster/AI architect at an annual fully loaded cost of **$260,000**
- Two cluster support engineers at **$125,000 each per year**

These roles cover AI infrastructure architecture, system management, administration, and user support. The additional staff will help support a centralized shared environment serving multiple business units and supporting concurrent training workloads across the organization.

Total incremental personnel cost is **$510,000 per year**, included in the total cost of ownership calculations and extended across the five-year modeling horizon.

On-Premises Cost Summary: Penguin Solutions OriginAI Infrastructure

Capital Investment

Component	Cost
GPU Node Infrastructure (31 nodes, 248 GPUs)	$18,833,366
Administrative Nodes (8 nodes)	$336,940
Storage Hardware (VAST Configuration)	$2,138,000
Total Capital Investment	**$21,308,306**

The capital investment includes all networking infrastructure, racks, cabling, integration services, and five-year extended platform support. Detailed specifications are provided in Appendix A.

Annual Operating Costs

Component	Annual Cost
Energy (Incremental)	$587,621
VAST Software Subscription	$354,000
ClusterWareAI	$128,000
Personnel (3 FTE)	$510,000
Total Annual Operating Cost	**$1,579,621**

Total Cost of Ownership (OpEx + CapEx)

3-Year TCO $26,047,168

5-Year TCO $29,206,410

Annual Costs with Depreciation and Cost Share:

	Annual Costs, 3- Year Straight Line Depreciation			Annual Costs, 5- Year Straight Line Depreciation	
	Cost	% Total		Cost	% Total
Purchase ($21,308,306)	$7,102,769	82%		$4,261,661	73%
Vast Storage SW cost	$354,000	4%		$354,000	6%
ClusterWareAI Software	$128,000	1%		$128,000	2%
Power & Cooling costs	$587,621	7%		$587,621	10%
Additional Personnel costs	$510,000	6%		$510,000	9%
	$8,682,389.47			**$5,841,282.00**	

It's interesting to note that hardware depreciation is by far the largest portion of annual costs both on a three- and five-year depreciation basis. We've heard a lot in the press about electricity demand radically increasing with the advent of AI, which has driven up prices. While this is true, notice that the cost to both power and cool this cluster (calculations on page 12) is only 7% and 10% of the overall cluster cost. Even if electricity costs increased by 50%, it wouldn't have much of an impact on the results of this TCO study.

Amazon Web Services (AWS) Configuration and Cost Summary

For AWS, we configured 31 **p6-b200.48xlarge** instances to match the 248-GPU baseline (8 GPUs per instance).

We selected Dedicated Instances rather than shared tenancy. If the goal is to compare this to a fully owned, dedicated on-premises system, then shared multi-tenant infrastructure isn't really the same. Dedicated instances give hardware isolation and more predictable performance. AWS is the only CSP in our study that offers Dedicated Instances, which gives them an advantage, although one that is difficult to quantify.

AWS does not currently publish three-year reserved pricing for B200 instances. To model committed pricing, we applied a 26.55% discount to on-demand rates. That percentage is derived from the historical difference between on-demand and three-year reserved pricing for earlier GPU generations. It's an estimate, and real-world enterprise agreements can vary.

All instances are placed within a single region to avoid cross-region latency and distributed training penalties.

Storage Configuration

Storage is modeled using Amazon FSx for Lustre at the 250 MB/s per TiB performance tier.

At roughly 2.5 PB of capacity, that tier provides enough aggregate throughput to keep the GPUs fed under distributed training workloads. Managed Lustre scales performance with provisioned capacity. It is architecturally different from the VAST scale-out storage used on-prem, but both were sized to avoid creating storage bottlenecks in the model.

Networking

The p6-b200 instances use Elastic Fabric Adapter (EFA) networking with up to 3,200 Gbps aggregate bandwidth per instance.

This is not native InfiniBand like the on-prem system. EFA provides RDMA capabilities over AWS's high-performance Ethernet network. InfiniBand and EFA are architecturally different, but both are intended to support large-scale distributed AI training.

I modeled all nodes within the same placement configuration to preserve network performance.

Personnel

Cloud does not eliminate operational responsibility. In fact, a data center that is primarily on-prem based and is contemplating a 248 GPU shared AI cluster in the cloud will need some additional personnel to help architect, manage, and provide user support. With this in mind, we assumed three incremental roles:

- One senior cloud architect
- Two cloud administrators

These roles cover architecture oversight, cost management, monitoring, and user support for a centralized AI environment.

Total personnel cost is estimated at **$510,000 annually.**

AWS Cost Roll-Up (Including Personnel)

Component	Monthly	Annual	Three-Year	Five-Year
GPU Instances	**$2,007,507**	**$24,090,080**	**$72,270,240**	**$120,450,400**
Admin Instances	**$28,482**	**$341,787**	**$1,025,362**	**$1,708,937**
FSx for Lustre Storage	**$538,383**	**$6,460,591**	**$19,381,773**	**$32,302,956**
Personnel (3 FTE)	**$42,500**	**$510,000**	**$1,530,000**	**$2,550,000**
Total AWS Cost	**$2,616,871**	**$31,402,459**	**$94,207,376**	**$157,012,293**

Three-year totals reflect committed pricing less reserved discount assumption discussed above. Five-year totals extend the same CSP monthly cost over the next two years. We did not try to guess future discounts or price changes.

Detailed costs for AWS configuration are provided in Appendix B

Google Cloud Platform (GCP) Configuration and Cost Summary

For GCP, the configuration was 31 **A4-highgpu-8g** instances to match the 248-GPU baseline (8 NVIDIA B200 GPUs per instance).

Unlike AWS, Google does not offer dedicated host options for this instance class. These are shared instances operating within Google's multi-tenant infrastructure. We configured them as closely as possible to the on-prem and AWS environments, but true hardware isolation is not available in the same way.

The pricing used here shows Google's published **three-year committed use discount (CUD)** rates. These are not estimated discounts; they are taken directly from Google's pricing tools.

All instances are deployed within a single region.

GCI Compute Configuration Notes

There are several hardware differences worth noting:

- Each GPU node provides **3,968 GB of system memory**, nearly double the memory in the on-prem and AWS configurations.
- Local NVMe storage per GPU node is **12.88 TB**.
  To bring this closer to the on-prem baseline of 30.73 TB per node, we provisioned an additional **16.88 TB of Hyperdisk** per node.

For the administrative nodes:

- System memory is approximately **21% higher** than the on-prem configuration.
- Local SSD capacity is **3 TB**. We added **20 TB of Hyperdisk**, bringing total per-node storage to approximately **23 TB**, which is still roughly 23% lower than the 46 TB available per on-prem admin node.

Whether these differences represent meaningful performance advantages or disadvantages depends on workload characteristics. For purposes of this economic comparison, the systems were configured to approximate functional parity rather than achieve exact hardware symmetry.

Networking

Google provides up to 3,600 Gbps maximum RDMA bandwidth via its custom network fabric.

This differs architecturally from both AWS EFA and on-prem InfiniBand but is designed to support large-scale distributed AI training. All instances need to be run within a single region to preserve network efficiency.

Storage

GCP storage utilizes **Google Managed Lustre** at a monthly cost of **$551,131**.

As with AWS FSx for Lustre, performance scales with provisioned capacity. The configuration was sized to provide sufficient throughput to sustain distributed GPU training without introducing storage bottlenecks.

Architectural differences between managed Lustre and the on-prem VAST configuration remain, but each was provisioned to support the same performance envelope.

Personnel:

As with AWS, we assumed three incremental full-time equivalents at an annual cost of $510,000.

GCP Cost Summary

Component	Monthly	Annual	Three-Year	Five-Year
GPU Instances	$1,345,073	$16,140,876	$48,422,628	$80,704,380
Admin Instances	$48,275	$579,300	$1,737,900	$2,896,500
Managed Lustre Storage	$551,131	$6,613,572	$19,840,716	$33,067,860
Personnel (3 FTE)	$42,500	$510,000	$1,530,000	$2,550,000
Total GCP Cost	**$1,986,979**	**$23,843,748**	**$71,531,244**	**$119,218,740**

Three-year totals reflect Google's published committed-use pricing. Five-year totals simply extend the same CPS monthly cost forward. Again, we did not try to guess future discounts or price changes.

Detailed costs for GCP configuration are provided in Appendix B.

Oracle Cloud Infrastructure (OCI) Configuration and Cost Summary

For OCI, the configuration was GPU instances to match the 248-GPU baseline across 31 nodes.

As with Google Cloud, OCI does not offer dedicated host options for this instance class. These are shared instances operating within OCI's multi-tenant infrastructure. We configured them to approximate the same functional capacity as the on-prem and AWS models, but true hardware isolation is not available.

There are some hardware differences worth noting. OCI's GPU nodes use **AMD EPYC processors (128 cores at 2.25 GHz)**, and each node includes **54.5 TB of local SSD storage**, substantially more on-node local storage than either AWS or Google in this model. Whether that translates into a performance advantage depends heavily on workload characteristics and how data is staged and accessed.

OCI Pricing Assumptions

OCI does not publish multi-year committed pricing for these GPU instances in a tool-based format comparable to Google's committed-use discounts. List pricing for the GPU nodes is **$2,534,560 per month.**

OCI's pricing model places greater emphasis on direct enterprise negotiation rather than standardized, tool-based multi-year committed pricing tiers. As a result, committed-use discounts are not transparently published in the same way as Google Cloud's CUD rates. The 15% discount applied here is intended to approximate a reasonable negotiated agreement for modeling purposes. Actual discounts will vary by customer and contract terms.

For modeling purposes, we applied an estimated 15% discount to both GPU and admin instance pricing to approximate a negotiated enterprise agreement. The resulting net monthly pricing is:

- GPU Instances: **$2,154,376**
- Admin Instances: **$30,773**

Storage is at published list pricing to maintain consistency with the AWS and Google models.

OCI Storage

OCI Lustre storage is calculated using:

- **$0.086 per GB** (capacity charge)
- **$0.125 per GB** (performance charge)

For approximately 2.5 PB usable storage (2,621,444 GB), this produces a monthly storage cost of: **$553,125** (No discount was applied to storage pricing.)

Networking

OCI provides high-performance (presented as 3200 Gbs) RDMA networking designed to support distributed AI training workloads. All instances are within a single region to avoid cross-region latency penalties and preserve cluster efficiency.

Personnel

As with AWS and Google Cloud, we assumed three incremental full-time equivalents:

- One senior cloud architect
- Two cloud administrators

These roles support architecture oversight, cost management, monitoring, and user support for a centralized AI deployment.

Total annual personnel cost is estimated at **$510,000**.

OCI Cost Roll-Up

Component	Monthly	Annual	Three-Year	Five-Year
GPU Instances (net)	$2,154,376	$25,852,512	$77,557,536	$129,262,560
Admin Instances (net)	$30,773	$369,273	$1,107,820	$1,846,367
OCI Lustre Storage	$553,125	$6,637,496	$19,912,489	$33,187,481
Personnel (3 FTE)	$42,500	$510,000	$1,530,000	$2,550,000
Total OCI Cost	**$2,780,773**	**$33,369,282**	**$100,107,845**	**$166,846,408**

Three-year totals reflect the 15% discount applied to compute instances as mentioned above. Five-year totals simply extend the monthly costs forward. Also, like before, I'm not going to guess future discounts or prices.

Detailed costs for OCI configuration are provided in Appendix B.

What the Numbers Show

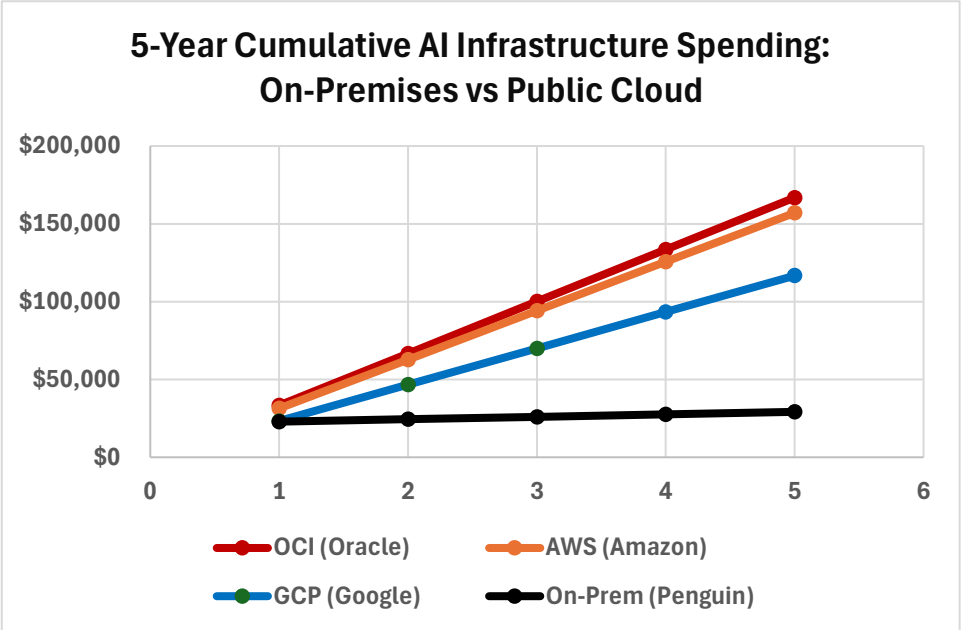
To make the comparison clear, the fully loaded costs for each deployment model are shown below. These totals include compute, storage, network, facilities (if applicable) and personnel, and reflect the assumptions described earlier.

Deployment Model	First Year	Three-Year Total	Five-Year Total
On-Prem (Penguin)	$22,887,927	$26,047,168	$29,206,410
Google Cloud	$23,843,748	$71,531,244	$119,218,740
AWS	$31,402,459	$94,207,376	$157,012,293
Oracle Cloud (OCI)	$33,369,282	$100,107,845	$166,846,408

With the numbers presented side by side, the relative scale becomes clear.

Over three years, the cloud configurations range from approximately $71.5M to $100.1M. The comparable on-premises deployment is $26.0M.

Over five years, the full on-premises system cost $29.2M, while the cloud configurations range from $119M to $167M.



Costs include infrastructure and operational staffing under the assumptions described in this report.

The annual cloud spend ranges from roughly $24M to $33M per year; it doesn't change from the initial first year spending due to the three-year agreements. These contracts were negotiated to both ensure capacity availability and to reduce costs.

During that three-year period, the CSP configurations would remain static as per the contract. In other words, the GPUs wouldn't be upgraded by the CSP unless there is some mechanism in the contract to require it, probably at a higher cost.

By comparison, the entire five-year cost of the on-premises system is $29.2M. Its configuration would remain static as well, although an upgrade of all or some of the B200 GPUs is technically possible, but the customer would be assuming the cost of the new GPUs.

We believe this cost disparity between CSPs and on-premises systems has always existed and will always exist. The CSPs can and do negotiate deals with customers to reduce costs, but it's highly unlikely that a special deal would bring cloud TCO to parity with an on-premises deployment. Plus, on-prem vendors also negotiate their prices.

Key considerations for customers trying to decide whether to buy or rent an AI infrastructure (or any other sizable compute project) are:

- Current and future demand patterns for the new system or workflow. How many concurrent users, demand schedules, etc. Constant demand skews the economic decision toward on-prem.
- System sizing and utilization are also important. System utilization under 50% a year edges into cloud territory, unless there is slack space to fit the workload onto another on-prem system. When system utilization climbs above 60%, it skews cost wise toward on-prem.

The decisions here aren't technical, they're firmly rooted in buy vs rent economics.

Appendix A: Detailed Penguin OriginAI System Configuration & Costs 1

		Configuration	Memory (GB)	On Node Storage (NVMe)	Network Performance (Gigabits)	Purchase Price
Compute	GPU Node configuration	2 x Intel 8592+, 64 cores, 1.9 GHz, 8 x Nvidia B200 SXM GPUs, NVLink,Nvidia ConnectX-7 1-Port OSFP/400Gb NDR IB x 8, 2 x 2 TB M.2 SSD (boot)	2048	30.73 NVMe, persistent	400 Gbs Infiniband	$18,833,366
	Admin Node Configuration	2 x Intel 6767P 64C, 2.4Hz, 6400MHz, 350W, 12 x U.2 3.84 TB SSD, 2 x Nvidia ConnectX-7 2-Port QSFP112/400Gb VPI, 2 x 1 TB M.2 SSD (boot)	1536	46 TB NVMe, persistent	400 GbE	$336,940

Storage	Vast Data Universal Storage, C + D boxes, 2.5 PB usable capacity 64 x Vast C Boxes, 10 x Vast D boxes	384 GB/s aggregated write throughput 600 GB/s aggregated read throughput	Storage Purchase Price $2,138,000
	Annual software subscriptiong cost		$354,000

Networks	Back-End Network	GPU-to-GPU Communication. Non-blocking 400Gb NDR InfiniBand.	12 x Nvidia QM9700 64-port core switches = 4 x spine switches and 8 x leaf switches (GPU+UFM)	Included
	Front-End Network	In-Band and Storage Communication. Non-blocking 400Gb Ethernet.	Edgecore Switches (SONiC O/S): 4 x 64-port core, 4 x 64-port edge, 2 x 32-port edge (admin), 2 x 32-port edge (storage)	Included
	Out-of-Band Network	IPMI and Serial Console Communication for Remote Management.	Edgecore Switches (SONiC O/S) GbE, 2 x 32-port core switches, Edge switches: 2 x 48-port (admin), 1 x 48 (BE spine/FE core), 2 x 48-port (BE leaf/FE Edge, 1 x 48-port	Included

Systems Management & Monitoring Software	ClusterWareAI: cluster monitoring and management platform with tool integration and secure operations	Handles provisioning/deployment plus Ansible-based orchestration. Monitors status, utilization, and performance at cluster and node level. Slurm, Kubernetes, OpenPBS integration. performance, status			
		ICE ClusterWare Cost	1 year	3 year	5 year
			$128,000	$348,000	$640,000

Appendix A: Detailed Penguin Solutions Origin AI (On-Premises)

		Cost
Additional Components, Integraton/Delivery Services, Warranty	Purchase price includes all rack hardware, cabling, etc., for complete and functioning system.	Included
	Integration services: advanced rack level integraton of nodes,	Included
	Delivery Services: Engineers position racks into data center	Included
	Five Year Extended Platform Support: Technical support availability 24x7, Support team response times: Severity 1 - 24x7, one hour, Severity 2, 9x5, four hours, Severity 3, 9x5, 8 hours	Included

			Annual
Additional User Side Costs, Personnel	AI architecture, management administration, user support	1x Senior cluster/AI architect	$260,000
		2x junior cluster support (125k each)	$250,000
		Total annual	$510,000

		Year 1 Costs	Year 2 Costs	Year 3 Costs	Year 4 Costs	Year 5 Costs
Total Penguin Solutions Costs, Annual 1 - 5 years	Net GPU node costs	$18,833,366	-	-	-	-
	Net admin nodes	$336,940	-	-	-	-
	Storage	$2,138,000	-	-	-	-
	Storage SW Cost	$354,000	$354,000	$354,000	$354,000	$354,000
	ICE ClusterWare Systems Management SW	$128,000	$128,000	$128,000	$128,000	$128,000
	Power & Cooling costs	$587,621	$587,621	$587,621	$587,620.80	$587,621
	Additional Personnel costs	$510,000	$510,000	$510,000	$510,000	$510,000
	Total	**$22,887,927**	**$1,579,621**	**$1,579,621**	**$1,579,621**	**$1,579,621**
	Total 3-year Costs	**$26,047,168**				
	Total 5-Year Costs	**$29,206,410**				

Appendix B: Detailed CSP Configurations & Costs, Amazon AWS

	Item	Notes	Instance/Service	Memory (GB)	Storage (NVMe on node, TB)	Network Performance (Gigabits, agg. fabric)
Compute	GPU Node configuration	*Dedicated instance, 24 x 7 x 365 availability, three year reserved, 248 Nvidia B200 GPUs*	p6-b200.48.xlarge x 31 nodes, 8 x Nvidia B200 GPUs, dual Intel 48 core, 2.3 GHz base clock, 192 vCPUs, NVLink	2048	30.72 TB ephemeral	3200, RDMA over custom EFA fabric
	Admin Node Configuration	*Dedicated instance, 24 x 7 x 365 availability, three year reserved, 8x admin nodes to match on-premises configuration*	i4i.32xlarge x 8 instances, 128 vCPU	1952	30 (35% less than 46 TB on -prem)	75 Gb (vs 400Gb on-prem)

Monthly Costs	GPU Node(s) Cost Calculations		Admin Nodes (8 instances) Cost Calculations	
	Base GPU Server On Demand Monthly cost (31 x $120.776 x 730 hours a month)	Estimated AWS Discounted Monthly Cost based on three year reserve (26.55%*)	Admin server monthly cost (8 x $ 6.09638 x 730)	Est. AWS Discounted Monthly Cost based on three year reserve (20%)
	$2,733,161	$2,007,507	$35,603	$28,482

*there isn't a published 3-year p6-b200.48xlarge instance reserve rate at this time, so using same On-Demand vs 3-year reserve rate as on p5.48xlarge GPU instance

Storage				Throughput Capacity	Monthly storage cost (2.560 petabytes provisioned x $0.000288
	GPU cluster storage	*FSx for Lustre (also requires Elastic Fabric Adapter and Nvidia GPUDirect Storage)*	AWS FSx for Lustre (2.5PB usable capacity, SSD Persistent Storage)	250 MB/s/TiB	$538,383

Total AWS Costs	Monthly	Annually	Three Year	Five Year
Net GPU node costs	$2,007,507	$24,090,080	$72,270,240	$120,450,400
Net admin nodes	$28,482	$341,787	$1,025,362	$1,708,937
Storage	$538,383	$6,460,591	$19,381,773	$32,302,956
AWS Total	**$2,574,372**	**$30,892,459**	**$92,677,376**	**$154,462,293**

Additional User Side Costs			Annual Expense
	Cloud architecture, management administration, user support	1x Senior cloud architect/admin	$260,000
		2x junior cloud admin (125k)	$250,000
		Total annual	$510,000

Total Infrastructure: AWS	Monthly	Annually	Three Year	Five Year
	$2,616,871.55	$31,402,459	$94,207,376	$157,012,293

Configurations & Costs, Google Cloud Platform (GCP)

	Item	Notes	Instance/Service	Memory (GB)	Storage (SSD on node, TB)	Network Performance (Gigabits, agg. fabric)
Compute	GPU Node configuration	*Shared instance, dedicated instance not available, 24 x 7 x 365 availability, three year reserved, 248 Nvidia B200 GPUs*	A4 highgpu-8g, Intel Emerald Rapids, 56c,2.1 GHz base frequency, 224 vCPUs, NVLink	3968 (93% more than on-prem)	12.88 TB (ephemeral) plus 16 TB hyperdisk (persistent) to bring parity with on-prem	3,600 max, RDMA over custom NIC/fabric
	Admin Node Configuration	*Shared instance, 24 x 7 x 365 availability, three year reserved, 8x admin nodes to match on-premises configuration*	m3-megamem-128 x 8 instances, 128 vCPUs	1952 (21% more than on-prem)	3 TB plus 20 TB hyperdisk extreme = 23 TB, (23% less than on-prem)	200 Gb (vs 400Gb on-prem)
Monthly Costs	GPU Node(s) Cost Calculations					
	Monthly 31 x GPU Servers 3-year committed cost $1,283,309 plus hyperdisk cost $61,764	$1,345,073	Admin Nodes (8 instances) Cost Calculations			
	Monthly 8x admin servers, 3-year committed cost $26,124 plus 20 TB hyperdisk @ $22,152 per month	$48,275				
Storage	GPU cluster storage	*GCP Managed Lustre*	2.5PB usable capacity, SSD Persistent Storage	Throughput Capacity 250 MB/s/TiB	Monthly storage cost (2.621 petabytes provisioned x $0.000288 x 730 hours) $551,131	

Total GCP Costs	Monthly	Annually	Three Year	Five Year
Net GPU node costs	$1,345,073	$16,140,876	$48,422,628	$80,704,380
Net admin nodes	$48,275	$579,300	$1,737,900	$2,896,500
Storage	$551,131	$6,613,572	$19,840,716	$33,067,860
GCP Total	**$1,944,479**	**$23,333,748**	**$70,001,244**	**$116,668,740**

Additional User Side Costs			Annual Expense
Cloud architecture, management administration, user support	1x Senior cloud architect/admin, annual		$260,000
	2x junior cloud admin (125k each)		$250,000
		Total annual	$510,000

Total Infrastructure: Google Cloud Platform	Monthly	Annually	Three Year	Five Year
	$1,986,979.00	**$23,843,748**	**$71,531,244**	**$119,218,740**

Appendix B: Detailed CSP Configurations & Costs, Oracle Cloud Infrastructure (OCI)

	Item	Notes	Instance/Service	Memory (GB)	Storage (SSD on node, TB)	Network Performance (Gigabits, agg. fabric)
Compute	GPU Node configuration	*Shared instance, dedicated instance not available, 24 x 7 x 365 availability, three year reserved, 248 Nvidia B200 GPUs*	BM.GPU.B200.8 x31 instances, 8 x Nvidia B200 GPUs, 2 x AMD EPYC CPU, 128 cores @2.25 GHz, NVLink	2048	54.5 TB (vs 30.72 on-prem, ephemeral)	3,200 (RDMA over RoCE v2)
	Admin Node Configuration	*Shared instance, 24 x 7 x 365 availability, three year reserved, 8x admin nodes to match on-premises configuration*	VM.Standard.E5.Flex, 64 OCPUs	1536	add 23 TB local storage @ $881 x 8 nodes per month = $7,050	50-100 Gb (vs 400Gb on-prem)

Monthly Costs	GPU Node(s) Cost Calculations		Admin Nodes (8 instances) Cost Calculations		
	Monthly 31 x GPU Servers 3-year committed cost $14 per GPU hour x 248 GPUs x 730 hours	Estimated 15% discount based on size/commitment*	Monthly 8x admin servers, 3-year committed cost, $4.992 x 730 x 8 plus $7,050 local storage	Estimated 15% discount based on size/commitment*	*there isn't a published 3-year commitment discount rate at this time, estimating 15%
	$2,534,560	$2,154,376	$36,203	$30,773	

Storage				Throughput Capacity	Monthly storage cost (2,621,444 GiB provisioned)
	GPU cluster storage	*OCI Lustre shared storage, monthly: $0.086 per GB storage + $0.125 performance per GB*	2.5PB usable capacity, SSD Persistent Storage	250 MB/s/TiB	$553,125

Total OCI Costs		Monthly	Annually	Three Year	Five Year
	Net GPU node costs	$2,154,376	$25,852,512	$77,557,536	$129,262,560
	Net admin nodes	$30,773	$369,273	$1,107,820	$1,846,367
	Storage	$553,125	$6,637,496	$19,912,489	$33,187,481
	OCI Total	**$2,738,273**	**$32,859,282**	**$98,577,845**	**$164,296,408**

Additional User Side Costs			Annual Expense
	Cloud architecture, management administration, user support	1x Senior cloud architect/admin, annual	$260,000
		2x junior cloud admin (125k each)	$250,000
		Total annual	$510,000

Total Infrastructure: OCI	Monthly	Annually	Three Year	Five Year
	$2,780,773	$33,369,282	$100,107,845	$166,846,408