CAST AI

2025 Kubernetes Cost Benchmark Report

An in-depth analysis of cloud resource overprovisioning and utilization in Kubernetes applications, with key insights into CPU and GPU Spot Instance price trends.

https://cast.ai

Introduction

Despite widespread adoption within a mature cloud-native landscape, Kubernetes requires significant manual effort to manage cloud resources. When teams are stuck managing repetitive tasks and manually overseeing cloud infrastructure, they often overprovision clusters, misallocate resources for applications, and leave excessive headroom in workload requests – leading to cloud waste. This is largely due to the need to configure Kubernetes resources and parameters manually, along with the absence of native Kubernetes tooling for cost reporting and automated resource management.

Last year, our 2024 Kubernetes Cost Benchmark Report highlighted the disparity between provisioned and utilized cloud resources in Kubernetes applications. The report identified several reasons companies overspend and suggested optimization best practices DevOps teams could use to reduce costs while maintaining high performance and availability.

We are excited to share the 2025 edition of the report. It brings fresh insights into Kubernetes resource utilization and price trends for CPU- and GPU-based instances across the three major cloud providers – Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure (Azure) – along with best practices for reducing waste based on real-life examples.

Methodology

The 2025 Kubernetes Cost Benchmark Report is based on our analysis of 2,100+ organizations across AWS, GCP, and Azure between January 1 and December 31, 2024. This analysis excludes clusters with fewer than 50 CPUs and focuses on data collected **before these organizations used Cast AI's automation**.

The pricing data comes from publicly available inventory APIs provided by AWS, GCP, and Azure: AWS Price List Query API, Cloud Billing Pricing API, and Azure Retail Prices. We have recorded this data over the same period for analysis, with sampling at a rate of less than 60 seconds.

2,100+

ORGANIZATIONS ANALYZED

50+

MINIMUM CPU COUNT IN CLUSTERS ANALYZED

Jan-Dec, 2024

ANALYSIS TIMELINE

CLOUD PROVIDERS ANALYZED

Table of contents

1. Key findings
2. CPU and memory utilization in Kubernetes applications
3. Boosting CPU and memory utilization: six best practices with real-world examples
4. Spot Instances: potential savings and price trends
5. Maximizing the value of Spot Instances: four best practices with real-life examples
6. GPU availability and price trends
7. How to increase GPU utilization

Key findings

CPU and memory utilization

The average CPU utilization across clusters remained low at 10% (-23% YoY), while average memory utilization was marginally better at 23% (+15% YoY), indicating no significant year-over-year improvement in resource efficiency across cloud platforms compared to our previous report from 2024.

10%

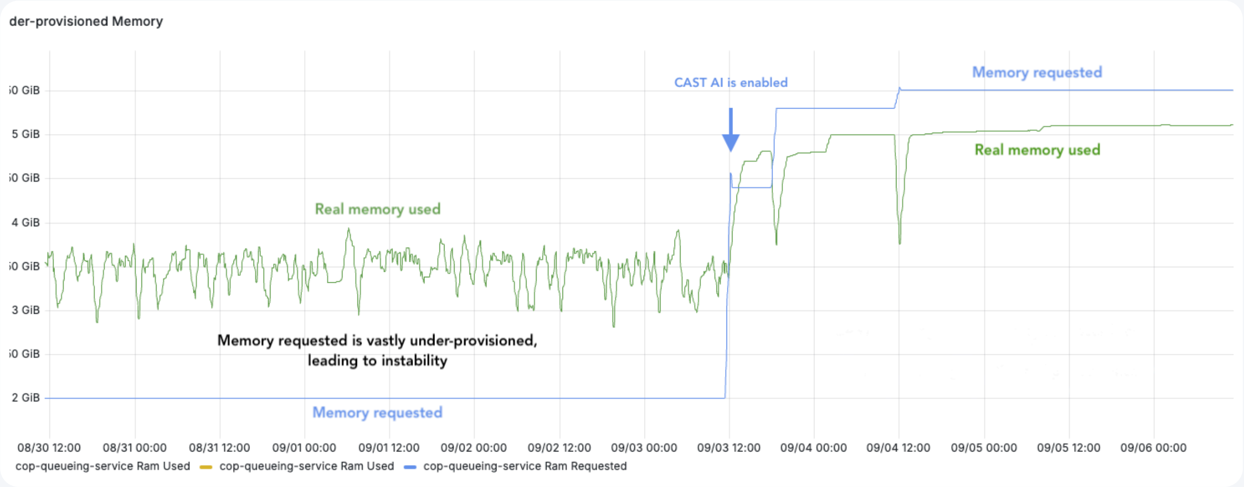
AVERAGE CPU UTILIZATION

23%

AVERAGE MEMORY UTILIZATION

Despite CPU being vastly overprovisioned, occasional memory underprovisioning is a major issue and it's more common than expected

Over a 24-hour period, 5.7% of containers exceed their requested memory at some point, indicating they are underprovisioned. This leads to instability, out-of-memory errors, and frequent restarts—overall significantly more disruption than anticipated—because workloads simply lack the resources they need to run reliably.



Automated change of memory request for a stability and performance increase

Spot Instances: potential savings and price trends

Our research shows teams can significantly reduce their cloud costs using Spot Instances. One indication is the level of potential savings the Cast AI agent identified.

For clusters with a mix of On-Demand and Spot Instances, we recorded an average of 59% compute cost savings. This means that organizations could cut their costs 2.5x if Spot-friendly workloads actually leveraged Spot Instances. Clusters running only Spot Instances recorded an average of 77% reduction in cost savings.

59%

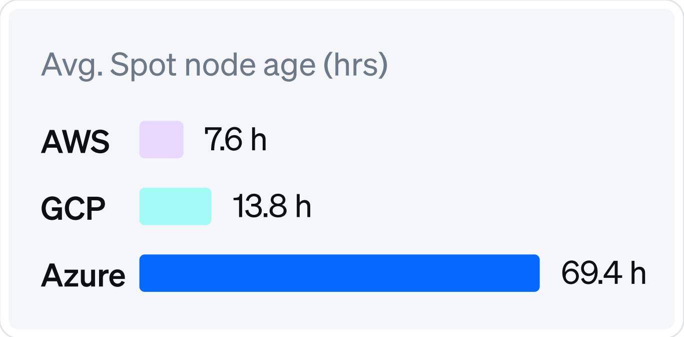
AVERAGE SAVINGS FOR MIX OF ON-DEMAND AND SPOT

77%

AVERAGE SAVINGS FOR SPOT ONLY CLUSTERS

Many teams hesitate to use Spot Instances due to interruptions. Our research shows significant differences between cloud providers regarding Spot Instance interruptions.

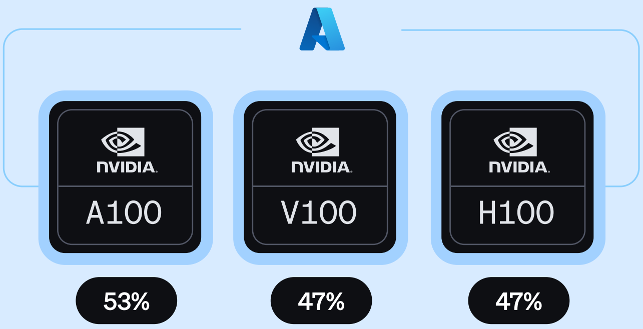
- AWS exhibits the highest overall interruption rate across shorter timeframes, with **50%+ of interruptions** occurring in the first hour of a node's lifetime.
- Azure demonstrates more stability, with much lower percentages of interruptions across all intervals, especially **within the first 12 hours.** GCP falls in the middle.



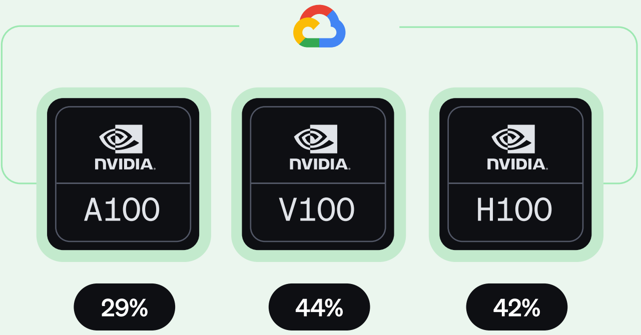
To mitigate these interruptions, teams can use automation solutions that provision, scale, and decommission Spot Instances to match real-time interruption rates.

GPU availability and price trends

GPU chip availability across regions

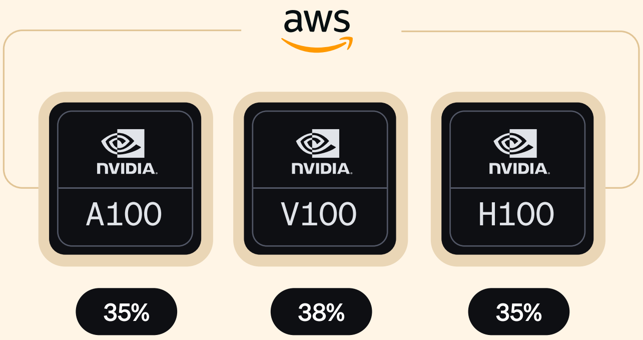


Azure leads the pack for A100, V100, and H100 availability, offering these instances in 53%, 47%, and 47% of its regions, respectively.



By contrast, GCP offers these instances in only 29%, 44%, and 42% of its regions.

Google Cloud provides good coverage for V100 and H100 GPUs.



AWS has even lower availability, offering these instance types in 35%, 38%, and 35% of its regions.

Price trends for GPU compute instances across AWS, GCP, and Azure

T4

15%

CHEAPER THAN EQUIVALENT IN AWS

82%

DISCOUNT FOR SPOT INSTANCES

A100

28%

CHEAPER THAN EQUIVALENT IN AWS

18%

CHEAPER THAN EQIVALENT IN AZURE

For machines running on Nvidia T4 chips, Azure offers a better deal than AWS. On average, the Azure instance is 15% cheaper than the AWS instance running 4 GPUs of the same type.

Google offers the lowest average On-Demand price for a compute instance running 8 NVIDIA A100 chips – 28% cheaper than an analogous instance in AWS and 18% cheaper than Azure.

Running GPU workloads on Spot Instances

90%

AVERAGE SPOT DISCOUNT

67%

AVERAGE SPOT DISCOUNT

aws

66%

AVERAGE SPOT DISCOUNT

Note: The listed discounts are for instances with A100, T4, V100, L4, and M60 chips.

Azure consistently offers the biggest discount on GPU-powered Spot Instances, regardless of the GPU chip type – ~90% on average.

Azure and Google Spot GPU prices are more stable, changing only a few times per month, while AWS adjusts Spot prices frequently for both GPU and non-GPU instances.

The average monthly number of distinct prices set by AWS is 197. In GCP, a new price is set every three months; in Azure less than once a month.

It's slightly more challenging to manage Spot fleets manually on AWS because of the price fluctuation, which is why automation is welcome there.

Best region and availability zone (AZ) to run your GPU workloads: some regions and AZs cost 6 times less compared to the average

GPU pricing and availability pressure change frequently over time. We analyzed the regions and AZs where GPU availability is the highest and compared the cost of running workloads in some of the hardest-to-find GPUs. The study was done for AWS instance p4d.24xlarge with 8 NVIDIA A100 GPU chips, spanning January 2024 until February 2025. Here is what we found:

AI workloads could cost six times less if they were moved based on the most cost-efficient regions and AZs instead of defaulting to us-east-1a, for example.

This would increase average cost savings by:

2x-7x

COMPARED WITH THE AVERAGE SPOT INSTANCE PRICE

3x-10x

COMPARED WITH THE AVERAGE ON-DEMAND INSTANCE PRICE

Cost savings achieved in specific regions/AZs for p4d.24xlarge instance

Time period	Best AWS region to run AI workloads	Cost savings vs. Spot Instance average	Cost savings vs. On-Demand average	% of all GPU instance types offered in the given region
Jan 2024	us-west-2a	7x	10x	93.7%
Feb-May 2024	us-east-2b	2.1x	7.7x	82.5%
Jun-Jul 2024	ap-northeast-1a	2x	4.8x	92.1%
Aug-Sep 2024	ap-northeast-2d	2.7x	5.7x	66.7%
Oct-Nov 2024	us-west-2b	1.9x	3.7x	93.7%
Dec 2024	us-west-2c	1.8x	3.2x	93.7%
Jan-Feb 2025	ap-northeast-2d	3x	4.7x	66.7%

CPU and memory utilization in Kubernetes applications: utilization rates remain low YoY

The average CPU utilization in our cluster sample **before optimization was 10%, a small decrease from last year's finding of 13%.** While there is a slight variation in utilization across the three major cloud platforms, the overall range remains comparable, indicating that resource efficiency practices have not significantly changed YoY. It is equally poor across all cloud providers, with no statistically significant differences between regions.

The average gap between provisioned and requested CPUs was 40% (compared to 43% in the 2024 edition of this report), showing that teams are not deploying as many workloads onto clusters as there is capacity for. What about memory utilization?

The average memory utilization level before optimization was 23%.

Compared to last year's number (20%), it shows that teams face the same challenge, especially given that the gap between provisioned and requested memory was 57%. Memory overprovisioning was slightly different across cloud providers, with the highest rate noted for Azure (65%), while AWS and GCP overprovisioned by 58% and 53%, respectively.

Cloud provider	aws	☁	A
CPU utilization (resources used vs. provisioned)	10%	12%	8%
Memory utilization (resources used vs. provisioned)	24%	23%	21%
CPU overprovisioning (resources requested vs. provisioned)	39%	39%	41%
Memory overprovisioning (resources requested vs. provisioned)	58%	53%	65%

Two key drivers of low CPU and memory utilization

Cloud waste in Kubernetes applications primarily stems from two factors:

1. Overprovisioning compute resources

One of the leading causes of inefficiency is deploying Kubernetes applications on virtual machine (VM) instances that are oversized for the actual workload. This issue is further exacerbated by non-production environments often left running during off-peak times.

2. Excessive headroom in workload requests

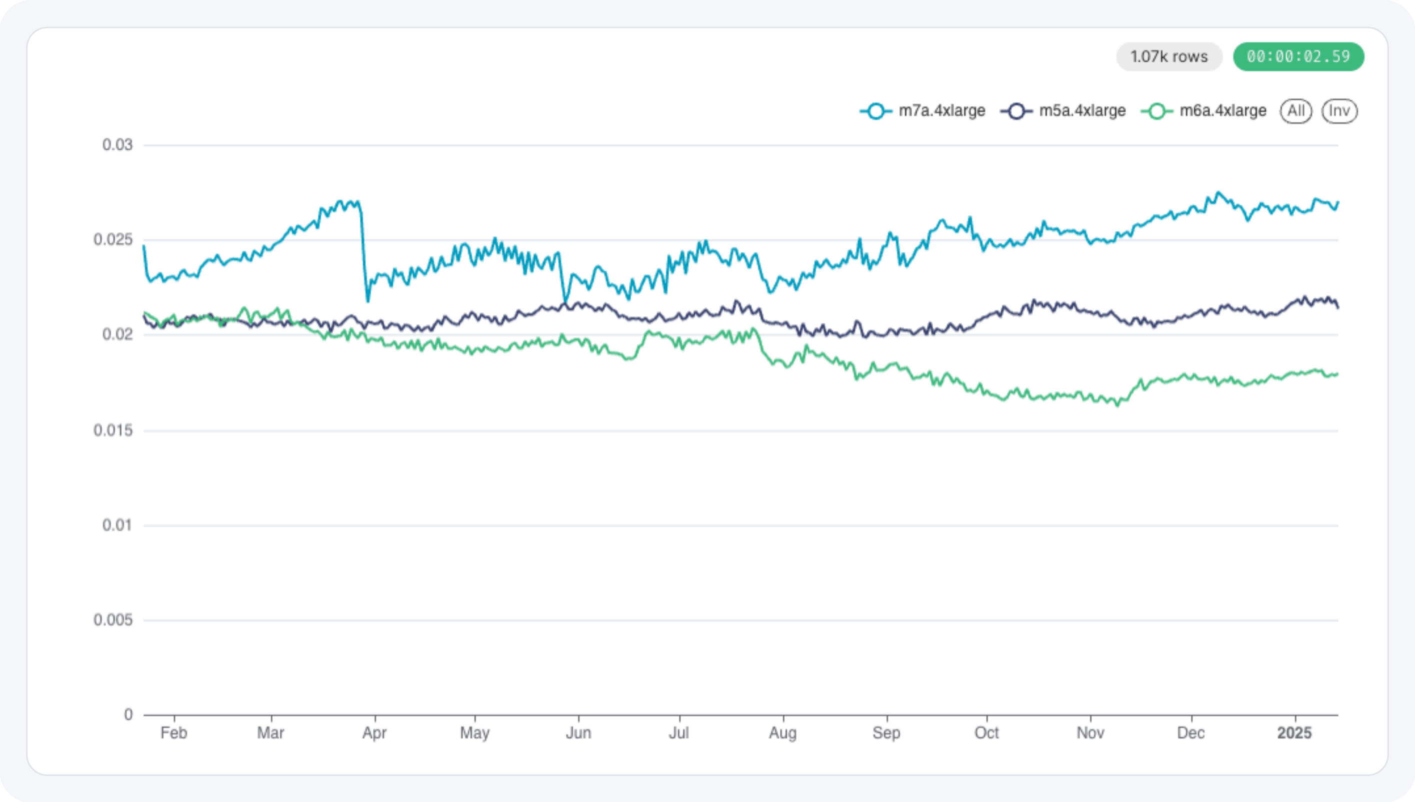
Kubernetes allows users to define resource requests and limits for workloads, ensuring that the specified amount of CPU and memory is reserved. However, if the workload fails to utilize the allocated resources fully, this reserved capacity goes unused, contributing to inefficiency.

How can teams address overprovisioning that occurs at the VM and workload levels?

Boosting CPU and memory utilization: six best practices with real-world examples

1. Be flexible when it comes to compute generation choice

The graph below illustrates the price evolution of three compute instances representing different generations. Being able to choose from various generations of compute is essential for users looking to benefit from the latest hardware advancements or more cost-effective, previous options.



2. Consider processor architecture (x86 vs. Arm)

This table provides a comparative overview of the hourly Spot and On-Demand (OD) pricing for x86 and Arm CPUs on AWS, GCP, and Azure. The focus is on the Spot-to-OD price ratio, which shows the percentage of the OD price a customer pays for Spot Instances.

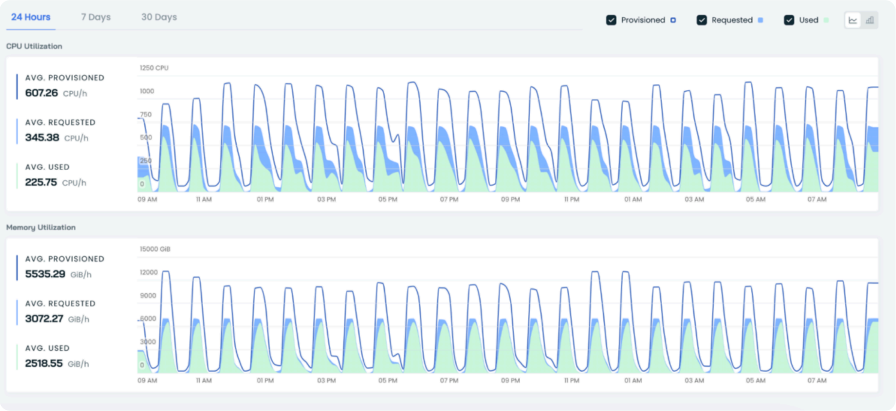
Cloud provider	aws		
Avg. x86 Spot price per CPU per hour	$0.039	$0.021	$0.025
Avg. x86 OD price per CPU per hour	$0.078	$0.066	$0.135
Avg. Arm Spot price per CPU per hour	$0.020	$0.016	$0.008
Avg. Arm OD price per CPU per hour	$0.049	$0.041	$0.047

- **Arm CPUs are consistently more cost-effective than x86 CPUs** across both On-Demand and Spot pricing, making them a more cost-effective choice for users who can take advantage of the Arm architecture.
- **Azure consistently has the largest relative difference in pricing between x86 and Arm CPUs (65%).** Spot prices for Arm are far lower than for x86 (69%), making it a compelling option for flexible, cost-conscious workloads.

3. Use custom agentic autoscalers to reduce waste

Akamai

One of the world's largest and most trusted cloud delivery platforms – was looking to optimize the costs of running its core infrastructure. The graph below shows how the Cast AI Autoscaler scales cloud resources up and down (both node number and size) with real-time demand, giving enough headroom to meet the application's requirements.
[Learn more]



CPU Utilization

AVG. PROVISIONED 607.26 CPU/h

AVG. REQUESTED 345.38 CPU/h

AVG. USED 225.75 CPU/h

Memory Utilization

AVG. PROVISIONED 5535.29 GiB/h

AVG. REQUESTED 3072.27 GiB/h

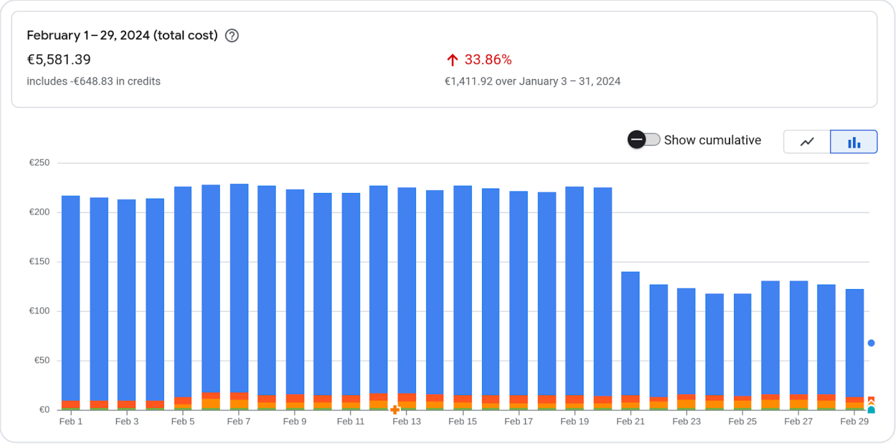
AVG. USED 2518.55 GiB/h

Autoscaling compute resources based on real-time demand

4. Bin-packing workloads brings substantial resource efficiency increase

heureka!group

The company has seen a 30% drop in compute costs in its Dev cluster achieved by bin-packing Spot-friendly workloads (stateless workloads that tolerate interruptions) and quickly removing the empty Spot nodes to drive down the number of provisioned CPUs.
[Learn more]


February 1 – 29, 2024 (total cost)

€5,581.39

includes -€648.83 in credits

↑ 33.86%

€1,411.92 over January 3 – 31, 2024

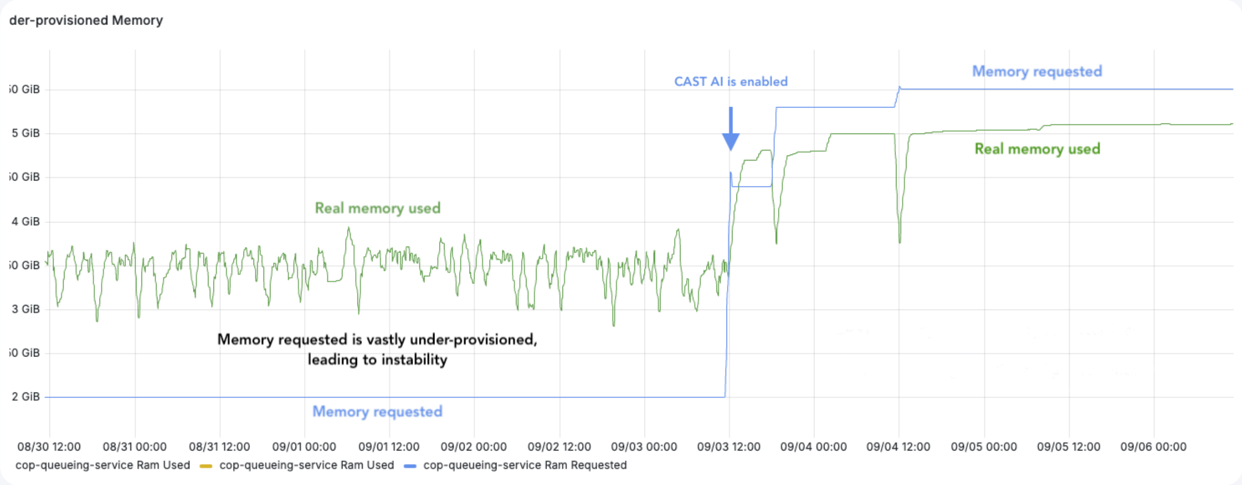
Dramatic cost drop thanks to automated workload bin-packing

5. Automate workload rightsizing to improve performance

Underprovisioning is more common than you might expect. **In 5.7% of analyzed containers, memory usage exceeded the requested allocation** at some point over a 24-hour period. This shortage leads to instability, as applications struggle to run with insufficient resources.

When memory is under-requested, performance takes a hit. Applications face instability, out-of-memory errors, and frequent restarts—all of which disrupt operations.

Automation eliminates this issue without adding to the engineering workload. In the example below, a cluster was running with approximately 3.5GiB of memory until September 3, 2024. Once an AI agent was activated, it optimized the memory request to 5.5GiB—ensuring stable performance without overprovisioning.

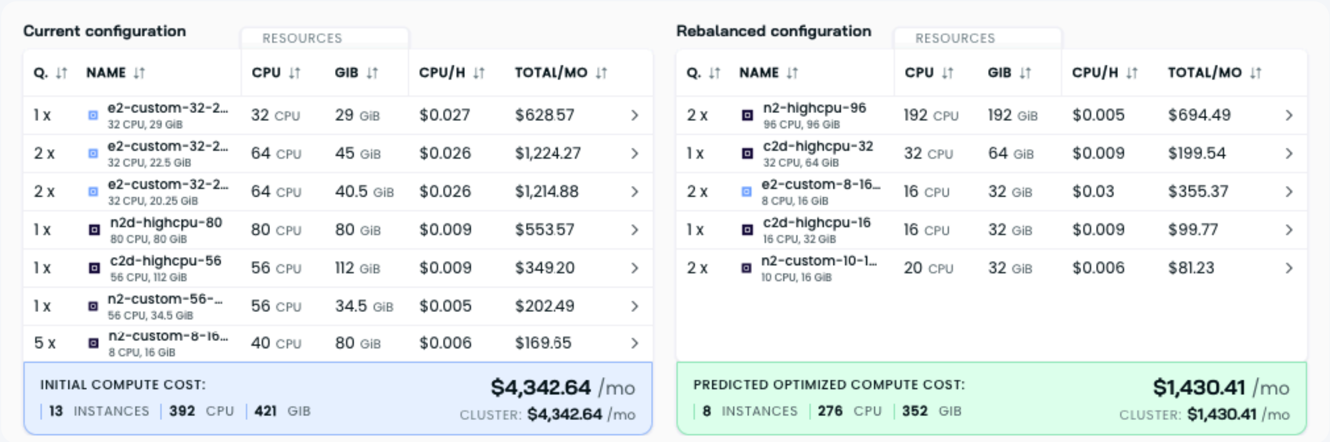


Automated change of memory request for a stability and performance increase

6. Select and run cost-effective nodes automatically

PlayPlay

The video creation platform, used automation to move its Kubernetes workloads to more cost-efficient VMs. The screenshot below illustrates a rebalancing operation in which Cast AI replaced 13 nodes without impacting service availability, saving PlayPlay $1,430 monthly. Learn more



Current configuration

		RESOURCES			
Q.	NAME	CPU	GIB	CPU/H	TOTAL/MO
1 x	e2-custom-32-2... 32 CPU, 29 GiB	32 CPU	29 GiB	$0.027	$628.57
2 x	e2-custom-32-2... 32 CPU, 22.5 GiB	64 CPU	45 GiB	$0.026	$1,224.27
2 x	e2-custom-32-2... 32 CPU, 20.25 GiB	64 CPU	40.5 GiB	$0.026	$1,214.88
1 x	n2d-highcpu-80 80 CPU, 80 GiB	80 CPU	80 GiB	$0.009	$553.57
1 x	c2d-highcpu-56 56 CPU, 112 GiB	56 CPU	112 GiB	$0.009	$349.20
1 x	n2-custom-56-... 56 CPU, 34.5 GiB	56 CPU	34.5 GiB	$0.005	$202.49
5 x	n2-custom-8-16... 8 CPU, 16 GiB	40 CPU	80 GiB	$0.006	$169.65

INITIAL COMPUTE COST: $4,342.64 /mo
13 INSTANCES | 392 CPU | 421 GIB
CLUSTER: $4,342.64 /mo

Rebalanced configuration

		RESOURCES			
Q.	NAME	CPU	GIB	CPU/H	TOTAL/MO
2 x	n2-highcpu-96 96 CPU, 96 GiB	192 CPU	192 GiB	$0.005	$694.49
1 x	c2d-highcpu-32 32 CPU, 64 GiB	32 CPU	64 GiB	$0.009	$199.54
2 x	e2-custom-8-16... 8 CPU, 16 GiB	16 CPU	32 GiB	$0.03	$355.37
1 x	c2d-highcpu-16 16 CPU, 32 GiB	16 CPU	32 GiB	$0.009	$99.77
2 x	n2-custom-10-1... 10 CPU, 16 GiB	20 CPU	32 GiB	$0.006	$81.23

PREDICTED OPTIMIZED COMPUTE COST: $1,430.41 /mo
8 INSTANCES | 276 CPU | 352 GIB
CLUSTER: $1,430.41 /mo

Rebalancing operation where workloads are moved to optimized instances

Even though Spot Instances offer significant discounts, their adoption remains relatively low due to the risk of interruption. We decided to investigate Spot Instances to shed more light on the real interruption rates and give you a realistic outlook on Spot Instance price fluctuations in the next chapter.

Spot Instances: potential savings and price trends

Our research shows that teams can significantly reduce their cloud costs using Spot Instances. **Clusters optimized with partial usage of Spot Instances recorded an average of 59% cost savings. Ones running only on Spot Instances recorded an average of 77% reduction in compute costs.**

Teams looking to leverage Spot Instances face two challenges: frequently changing prices and potential interruptions.

How often do cloud providers change Spot Instance prices?

In general, Azure and GCP Spot prices tend to be more stable and predictable, changing only a couple of times a month. On the other hand, AWS continuously changes its Spot prices, which is true for both GPU and non-GPU instances.

The average monthly number of distinct prices is **197 for AWS, 0.3 for GCP** (a new price is set every three months), and **0.8 for Azure** (a new price is set less than once a month).

Cloud Provider	aws	☁	A
Average number of distinct Spot prices per instance type per month	197.5	0.3	0.8

Average number of times cloud providers changed Spot prices per month in 2024

How long can you expect your Spot Instance to run without interruption?

Cloud provider	aws	☁	A
Avg. Spot node age	7.6 hours	13.8 hours	69.4 hours
% of interruptions within the first hr	51%	32%	18%
% of interruption in 2nd hour	13%	12%	10%
% of interruptions that occur on 2nd day	3%	2%	10%
% of interruptions that occur on 3rd and 4th day	2%	2%	9%

Average Spot Instance interruptions in 2024

Azure stands out with a higher average node age of 69.4 hours. This could indicate a lower volatility or better management of Spot Instances.

GCP also has a relatively long node age compared to AWS, with instances lasting 13.8 hours on average versus 7.6 hours for AWS. AWS has the shortest node lifespan of three cloud providers, indicating higher interruption rates or shorter availability for Spot Instances.

Interruptions within one hour are the most frequent, with an average of 34% occurring within this time frame across all providers. The second most common interruption timeframe is between 1-2 hours (11% on average).

Which cloud provider interrupts Spot Instances most often?

AWS exhibits the highest overall interruption rate across shorter timeframes, with 50%+ of interruptions occurring in the first hour of a node's lifetime and 9%+ of Spot nodes suffering interruptions within a week. Azure demonstrates more stability, with much lower percentages of disruption across all intervals, especially within the first 12 hours. GCP falls in the middle.

Here's a deeper dive into interruption data per time interval:

Interruptions within 1 hour

- AWS has the highest percentage of interruptions within one hour (51%), suggesting that their Spot Instances are highly volatile.
- GCP follows with 32% of interruptions in the same time frame.
- Azure has the least volatility in this category, with only 18% of interruptions happening within one hour.

Interruptions between 1-2 hours

- The percentage of interruptions between 1-2 hours is relatively close for AWS (13%) and GCP (12%), while Azure has a slightly lower rate (10%).
- This data shows a significant difference between interruptions occurring within the first and second hour.

Interruptions between 24-96 hours

- Azure shows significantly higher percentages for interruptions after 24-96 hours, suggesting instances are stable over longer periods (9% for 24-48 hours and 6% for 48-96 hours).
- AWS and GCP have lower percentages for these longer interruptions.

Maximizing the value of Spot Instances: four best practices with real-life examples

1. Remember that selecting a Spot Instance isn't a one-time exercise

The graph on page seven shows the price evolution of three generations of Spot Instances from the same family: m5a, m6a, and m7a.

The growing difference in prices demonstrates that choosing a Spot Instance should not be a single point-in-time decision. In February, all three instances were priced at a similar level, with the latest generation m7a being the most expensive. But the difference became much more pronounced during the last quarter of the year.

If a team picked m7a, thinking the slight price bump was worth the performance, they could actually mix these instances to be more cost-effective by year's end. For example, using m6a would have been more cost-effective in a dev environment since performance isn't critical, while running production clusters on the more expensive m7a will help with high performance.

This is why keeping track of price trends and being able to revisit earlier decisions is important. Teams tend to stick to familiar and common instances, missing the opportunity to explore alternatives.

The analysis of clusters connected to the Cast AI platform revealed that optimized clusters rely significantly less on previous-generation instance types. Only 5% of clusters running are provisioned with older-generation instances when optimized by Cast AI, compared to 30% in non-optimized clusters.

This highlights Cast AI's focus on selecting the latest and most efficient instance types for cost and performance optimization.

m5a

PRICE DECREASE ACROSS 2024

m6a

PRICE SIMILAR ACROSS 2024

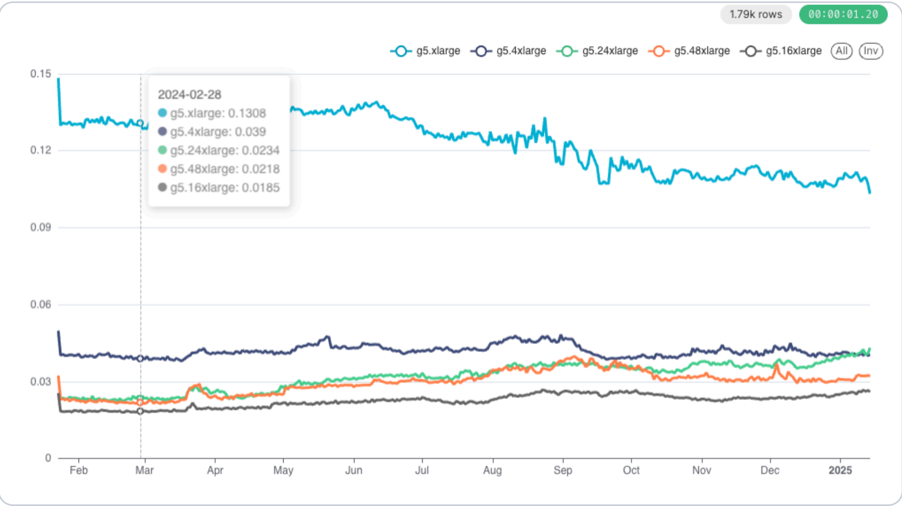
m7a

PRICE INCREASE ACROSS 2024

2. Consolidate clusters on a larger Spot Instance

Different GPUs carry different Spot Instance prices, potentially unlocking new savings.

Here's an example of the AWS G5 instance types offering varying GPU configurations depending on the size of the instance. Specifically, the g5.xlarge, g5.4xlarge, and g5.16xlarge (NVIDIA A10G Tensor Core GPU) instances are each equipped with one GPU. The g5.24xl has four GPUs, while the g5.48xl provides eight GPUs.



When running the g5.16xl instance, teams can use the single GPU while gaining access to significant additional compute resources (CPU). Workloads that don't require the GPU can utilize these additional resources, leading to cost-effective computation.

In terms of price per GPU, larger multi-GPU instances provide better value than several smaller, single-GPU instances. When you look at the prices per GPU, there isn't much difference between the 4-GPU (g5.24xl) and 8-GPU (g5.48xl) instances. This means that larger instances are cheaper for tasks that use a lot of GPUs.

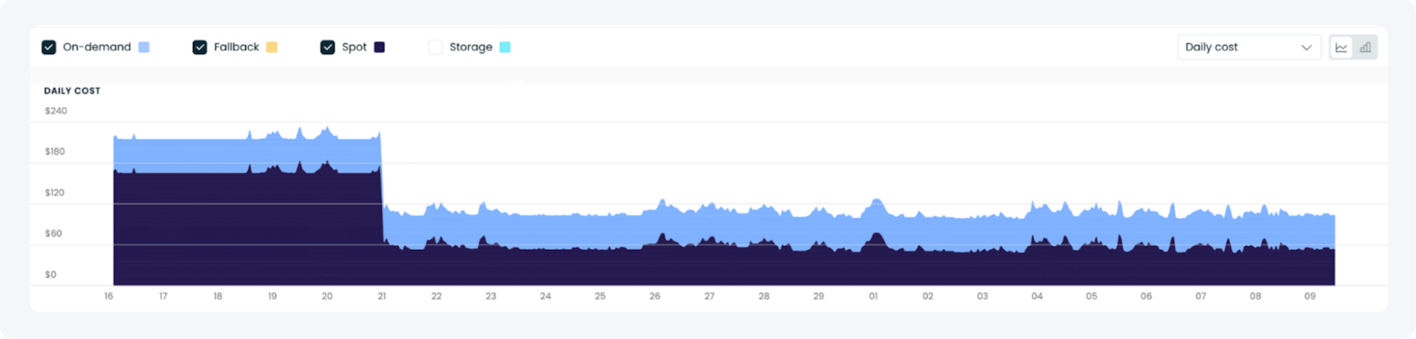
3. Choosing the right instance family can make a big difference

heureka!group

The company uses automated scaling and bin-packing mechanisms to provision just enough resources to ensure excellent performance.

The Cast AI Autoscaler unlocked more savings by moving Spot-friendly workloads that have already been deployed on Spot VMs to cheaper families while maintaining service uptime and application performance. Learn more

> "
>
> Cast AI provisions nodes with just the right amount of resources needed at a given time. This also includes handling Spot VM interruptions without headaches.
>
>
>
> **Martin Petak**
> Infrastructure Team Lead at Heureka Group



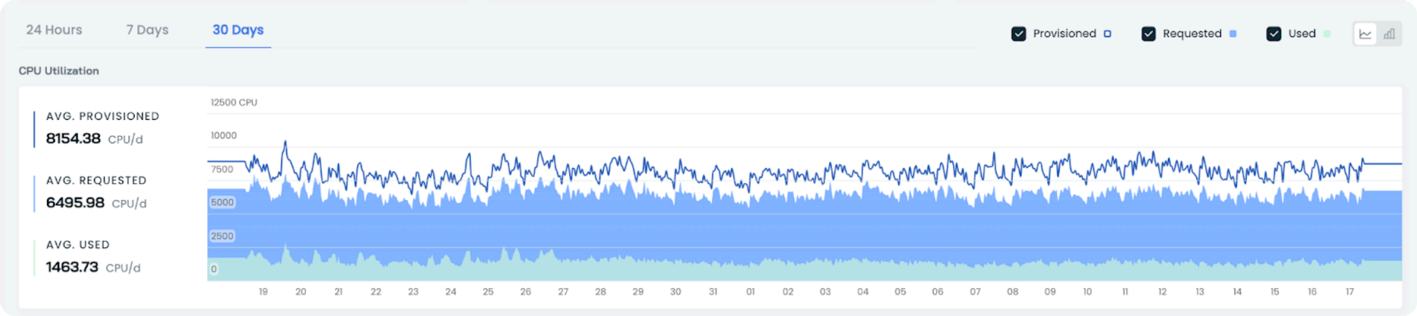
4. Scaling Spot Instances boosts savings, but you need automation to manage interruptions

yotpo.

The company runs at least 80% of workloads on Spot Instances and integrated Cast AI to manage the entire Spot Instance lifecycle, including:

- Moving workloads back and forth between Spots and On-Demand instances when availability changes
- Provisioning the most cost-effective instance type and size
- Autoscaling instances in line with fluctuating demand

The graph below shows the efficiency of the production workload when running on Cast AI over 30 days. The autoscaler closely follows the workload's changing demands, increasing and decreasing provisioned CPUs. Learn more



AVG. PROVISIONED 8154.38 CPU/d

AVG. REQUESTED 6495.98 CPU/d

AVG. USED 1463.73 CPU/d

> After integrating Cast AI, we didn't have to do anything during Black Friday, which is amazing. We gained not just compute cost reduction but also a reduction in engineer workload.
>
> No Spot Instances are available? No problem; the workload is automatically moved to On-Demand instances – and Cast AI makes sure to select the cheapest instance that matches the resource consumption of that workload. And once Spots are available again, the workload is moved back there. The whole spike on Black Friday was much lower, which is cool

Achi Solomon
Director of DevOps at Yotpo

Are Spot Instances a good choice for teams running ML workloads on GPU-powered instances?

Our research into GPU price trends and coverage shows that GPU Spot Instances can deliver **massive cost savings from 55% to almost 90%, depending on the cloud provider.** The next chapter shows which providers offer the best prices for GPUs across the most popular versions and where teams can find the biggest GPU instance coverage.

GPU availability and price trends

Which cloud provider offers the biggest discounts for GPU-powered Spot Instances?

We compared the Spot price as a percentage of the On-Demand hourly price for various GPU instance types across the three major cloud providers: AWS, GCP, and Azure.

GPU Spot Instance price across all different chip types

A100

Cloud Provider	Spot Instance discount
GCP	61%
AWS	65%
Azure	90%

T4

Cloud Provider	Spot Instance discount
GCP	64%
AWS	67%
Azure	89%

V100

Cloud Provider	Spot Instance discount
GCP	69%
AWS	55%
Azure	89%

L4

Cloud Provider	Spot Instance discount
GCP	70%
AWS	78%
Azure	N/A (Not offered)

M60

Cloud Provider	Spot Instance discount
GCP	N/A (Not offered)
AWS	69%
Azure	90%

Azure consistently offers the biggest discount on Spot VMs, regardless of the chip type. **Spot price constitutes around 10-11.5% of the On-Demand price,** often by a significant margin compared to GCP and AWS, which suggests Azure is aggressively discounting its Spot Instances to drive more usage or better utilize its GPU capacity.

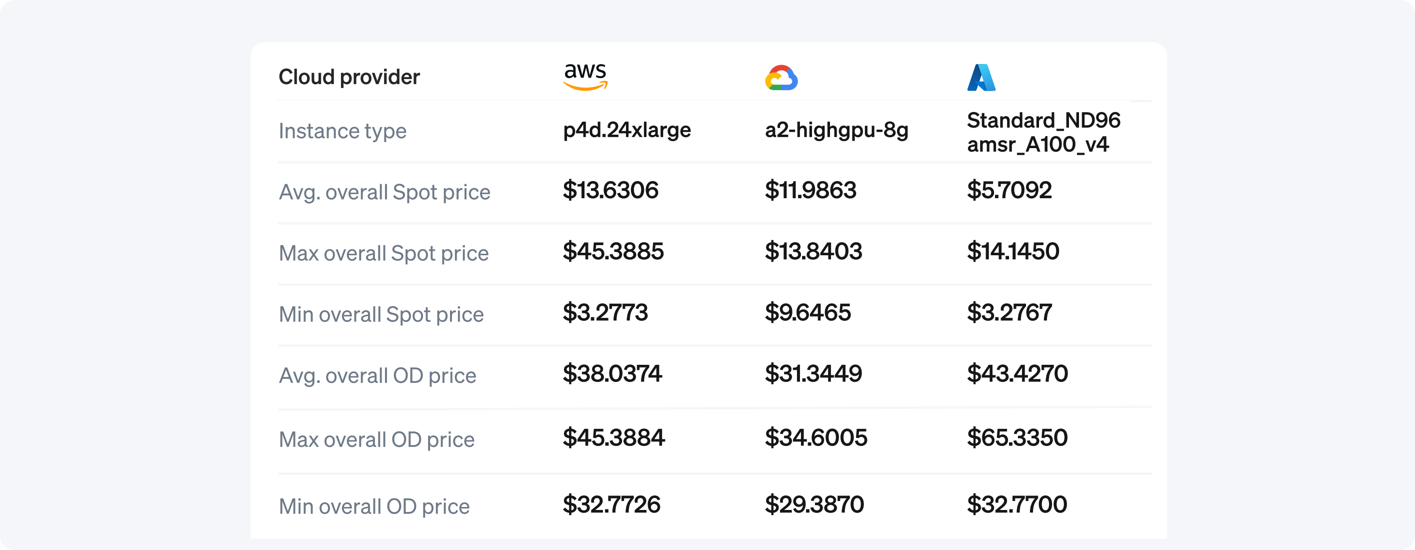
Note: Microsoft has been prioritizing inference workloads over training.

AWS pricing varies significantly across GPU types. While the provider offers competitive pricing for certain GPUs (like A100 and T4), it also charges considerably more for others (like V100). In most cases, GCP sits between Azure and AWS, with moderately higher discounts than AWS but lower than Azure.

Which cloud provider offers the best rates for GPU instances?

Here's a comparison of equivalent GPU instance types, with hourly prices across 2024 for three common GPU chips: NVIDIA A100, L4, and T4.

GPU A100 equivalent instance types - 96 VCPU, 8 NVIDIA A100s



Cloud provider	aws		
Instance type	p4d.24xlarge	a2-highgpu-8g	Standard_ND96 amsr_A100_v4
Avg. overall Spot price	$13.6306	$11.9863	$5.7092
Max overall Spot price	$45.3885	$13.8403	$14.1450
Min overall Spot price	$3.2773	$9.6465	$3.2767
Avg. overall OD price	$38.0374	$31.3449	$43.4270
Max overall OD price	$45.3884	$34.6005	$65.3350
Min overall OD price	$32.7726	$29.3870	$32.7700

Google offers the lowest average On-Demand price for a compute instance running 8 NVIDIA A100 chips. Azure's instance not only comes with the biggest price tag, but also saw the biggest fluctuation throughout 2024. However, note that the cloud provider offers the largest discount for Spot (87%) compared to others.

The GCP Spot price for this instance type saw the smallest price fluctuation ($10.86 in Azure, $42.11 in AWS, $4.19 in Google). The same is true for On-Demand ($32.56 in Azure, $12.61 in AWS, $5.22 in Google).

GPU L4 equivalent instance types - 4 VCPU, 1 NVIDIA L4 GPU

For these two machines powered by the NVIDIA L4 chip, AWS and Google set similar prices, with Google's instance averaging 8% less than AWS.

Google also provides the largest discount on Spot Instances (71%), with prices fluctuating less compared to AWS ($0.259 vs. $1.2875).

Cloud provider	aws	
Instance typet	g6.xlarge	g2-standard-4
Avg. overall Spot price	$0.2759	$0.2589
Max overall Spot price	$1.3680	$0.4325
Min overall Spot price	$0.0805	$0.1732
Avg. overall OD price	$0.9656	$0.8898
Max overall OD price	$1.3680	$1.1309
Min overall OD price	$0.8048	$0.7045

Note: Azure doesn't offer compute instances running on L4 chips.

GPU T4 equivalent instance types - 4 VCPU, 1 NVIDIA T4 GPU

For these machines running on 1 NVIDIA T4 chip, Azure offers a better deal than AWS. On average, the Azure instance is 15% cheaper than the AWS instance running 4 GPUs of the same type.

Azure wins in terms of Spot discount (82%).

Azure Spot Instances for this machine type also see a smaller price fluctuation ($0.185 for Azure vs. $0.648 for AWS). Both Azure and AWS instances had the same minimum price in 2024.

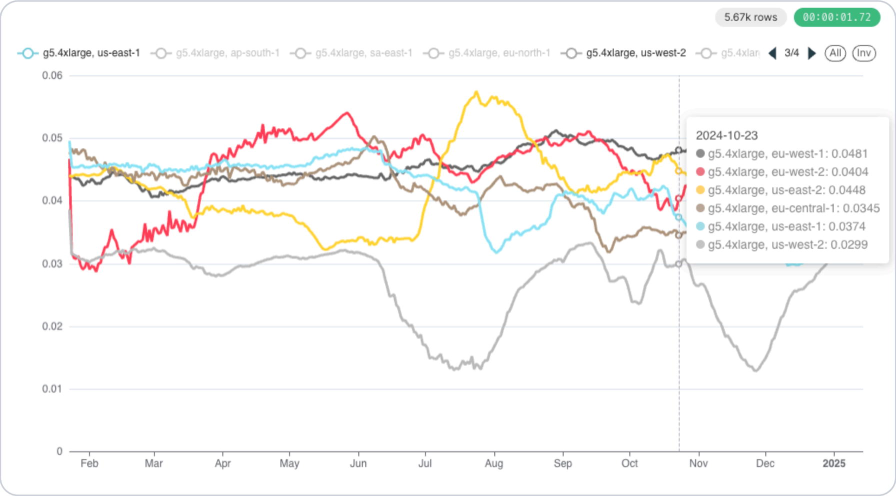
Cloud provider		aws
Instance typet	Standard_NC4as_T4_v3	g4dn.xlarge
Avg. overall Spot price	$0.1435	$0.2083
Max overall Spot price	$0.2374	$0.7080
Min overall Spot price	$0.0526	$0.0619
Avg. overall OD price	$0.7678	$0.8940
Max overall OD price	$1.0780	$0.8940
Min overall OD price	$0.5260	$0.5260

Note: Google offers T4-based instances as GPU chips attached to N1 instance types, meaning no instance type matches those offered by Azure or AWS.

Which cloud provider offers the best rates for GPU instances?

Being flexible regarding where you can schedule your GPU workloads is beneficial in terms of cost-effectiveness.

The graph below illustrates the price of the g5.4xlarge instance across various regions over time. This comparison shows that the us-west-2 region is generally the cheapest, while the other areas have high volatility.



Which cloud provider has the broadest regional coverage for GPUs?

Cloud Provider	aws	☁	A
Avg. percentage of regions where any given GPU chip type is offered	31%	21%	33%
% of regions that offer any GPU instance	85%	58%	76%

AWS is the clear winner with GPU-powered instances available in 85% of its regions. However, Azure stands out by offering broader coverage of individual GPU chip types, making each chip available in a larger share of its regions—despite supporting slightly fewer chip types overall. Notably, Azure also has the most regions, with 60 compared to AWS's 36 and GCP's 41.

What about coverage of specific GPU chips?

The table on the right represents the percentage of a cloud provider's regions where compute instances running on a specific GPU chip are available.

AWS offers the widest availability of T4 chips, with instances in 68% of its regions. However, while AWS leads in T4 GPU coverage, it falls behind Azure and Google in supporting newer GPU models like A100 and H100.

Cloud Provider	☁	aws	A
T4	56%	68%	49%
L4	41%	50%	N/A
A100	29%	35%	53%
V100	44%	38%	47%
H100	41%	35%	47%

Azure leads the pack for A100, V100, and H100 availability, offering these instances in 53%, 47%, and 47% of its regions, respectively. Google Cloud provides good coverage for V100 and H100 GPUs. While it trails behind Azure in A100 availability, it can still be a viable option, especially if pricing and other factors are more important than GPU availability.

What is the best region or AZ to run your GPU workloads?

GPU pricing and availability pressure change frequently over time. We analyzed the regions and AZs where GPU availability is the highest. We then compared the cost of running workloads in all regions and AZs for some of the hardest-to-find GPUs. The study was done for AWS instance p4d.24xlarge with 8 NVIDIA A100 GPU chips from January 2024 to February 2025.

Moving AI workloads to the most cost-efficient regions and availability zones, rather than defaulting to a specific one like us-east-1a, can reduce costs by a factor of six.

This would increase average cost savings by 2x-7x compared with the average Spot Instance price worldwide and 3x-10x compared with the average On-Demand Instance price. One reason these regions are significantly cheaper than average is that some regions are more GPU-friendly than others. These regions offer a greater variety of GPU types and tend to have more GPU instances on offer. One measure of GPU-friendliness is the availability ratio, the percentage of all GPU instance types that are offered in a given region. **Each of these cheap regions has a significantly higher availability ratio than the AWS average of 40.7% and is therefore more GPU-friendly than average.**

2x-7x

COMPARED WITH THE AVERAGE SPOT INSTANCE PRICE

3x-10x

COMPARED WITH THE AVERAGE ON-DEMAND INSTANCE PRICE

Cost savings achieved in specific regions/AZs for p4d.24xlarge instance

Time period	Best AWS region to run AI workloads	Cost savings vs. Spot Instance average	Cost savings vs. On-Demand average	% of all GPU instance types offered in the given region
Jan 2024	us-west-2a	7x	10x	93.7%
Feb-May 2024	us-east-2b	2.1x	7.7x	82.5%
Jun-Jul 2024	ap-northeast-1a	2x	4.8x	92.1%
Aug-Sep 2024	ap-northeast-2d	2.7x	5.7x	66.7%
Oct-Nov 2024	us-west-2b	1.9x	3.7x	93.7%
Dec 2024	us-west-2c	1.8x	3.2x	93.7%
Jan-Feb 2025	ap-northeast-2d	3x	4.7x	66.7%

How to increase GPU utilization

Why is GPU utilization so difficult to measure?

Measuring GPU utilization is significantly more complex than tracking CPU usage, as it involves multiple metrics, including memory usage, data transfer efficiency, and the utilization of compute cores and streaming multiprocessors.

Monitor the right metrics

To maximize your utilization of GPUs, it's essential to monitor several key metrics to ensure optimal performance and cost-efficiency:

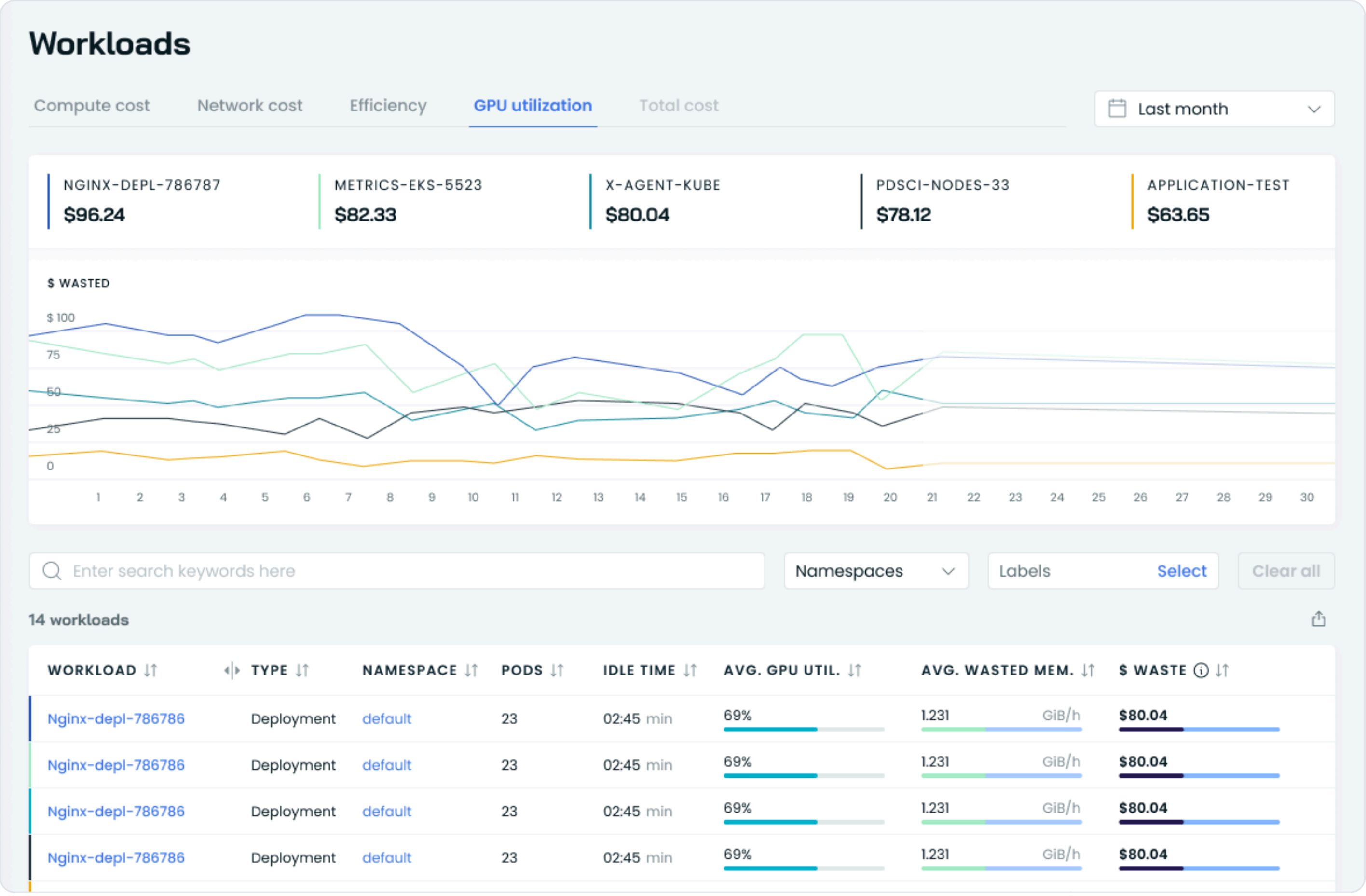
Idle time – High idle times suggest that your GPU is underutilized and resources are wasted. Reduce idle time by efficiently scheduling jobs and ensuring continuous workloads can significantly boost performance.

Average wasted memory – Reflects unused memory capacity during operations, which can directly affect overall efficiency, as memory bottlenecks may hinder performance.

Waste in dollar amount – Translating these inefficiencies, such as idle time and wasted memory, into dollar amounts helps quantify the financial impact.

Average GPU utilization – Provides insight into how effectively the GPU is being used; higher values indicate better use, while low utilization implies inefficient resource allocation.

Here's an example of a dashboard available in Cast AI that lets you monitor these metrics for every workload running on your GPU instance.



Workloads

NGINX-DEPL-786787	METRICS-EKS-5523	X-AGENT-KUBE	PDSCI-NODES-33	APPLICATION-TEST
$96.24	$82.33	$80.04	$78.12	$63.65

14 workloads

WORKLOAD	TYPE	NAMESPACE	PODS	IDLE TIME	AVG. GPU UTIL.	AVG. WASTED MEM.	$ WASTE
Nginx-depl-786786	Deployment	default	23	02:45 min	69%	1.231 GiB/h	$80.04
Nginx-depl-786786	Deployment	default	23	02:45 min	69%	1.231 GiB/h	$80.04
Nginx-depl-786786	Deployment	default	23	02:45 min	69%	1.231 GiB/h	$80.04
Nginx-depl-786786	Deployment	default	23	02:45 min	69%	1.231 GiB/h	$80.04

By calculating the cost of wasted GPU resources, you can make data-driven decisions to optimize workloads, find opportunities to improve parallelization, and identify applications that are not compute-intensive enough or have synchronization and blocking conditions that limit GPU utilization.

CAST AI

Cut cloud costs and boost DevOps efficiency with automation

Cast AI is the leading Kubernetes automation platform. Unlike traditional solutions that merely monitor clusters and provide recommendations, Cast AI leverages advanced machine learning algorithms to continuously analyze and automatically optimize clusters in real time, cutting cloud costs, securing Kubernetes applications, and boosting DevOps efficiency.

Start free →

Momentum Leader WINTER 2025

Leader WINTER 2025

Users Most Likely To Recommend Mid-Market WINTER 2025

Learn more about Cast AI

TRUSTED BY 800+ COMPANIES GLOBALLY

Akamai

BMW GROUP

FICO

Hugging Face

ShareChat