Otium

AI COMPUTE MARKET REPORT

GPU prices, October 2026

1.35 million AWS, Google Cloud spot prices across 55 regions and 64 machine types.



In brief

Prices held steady overall: all but three of the 13 GPUs we track moved less than 5% on spot since September. The two biggest moves went in opposite directions, both on AWS. The H100 rose 13% to $3.00 per GPU–hour and the A100 80 GB fell 12% to $2.82, so September's headline, an H100 that cost less than an A100, no longer held. The third was the RTX PRO 6000 on Google Cloud, up 5%. October's headline is an L40S, with 48 GB of memory, at $1.53 against $2.22 for an A100 with 40 GB.

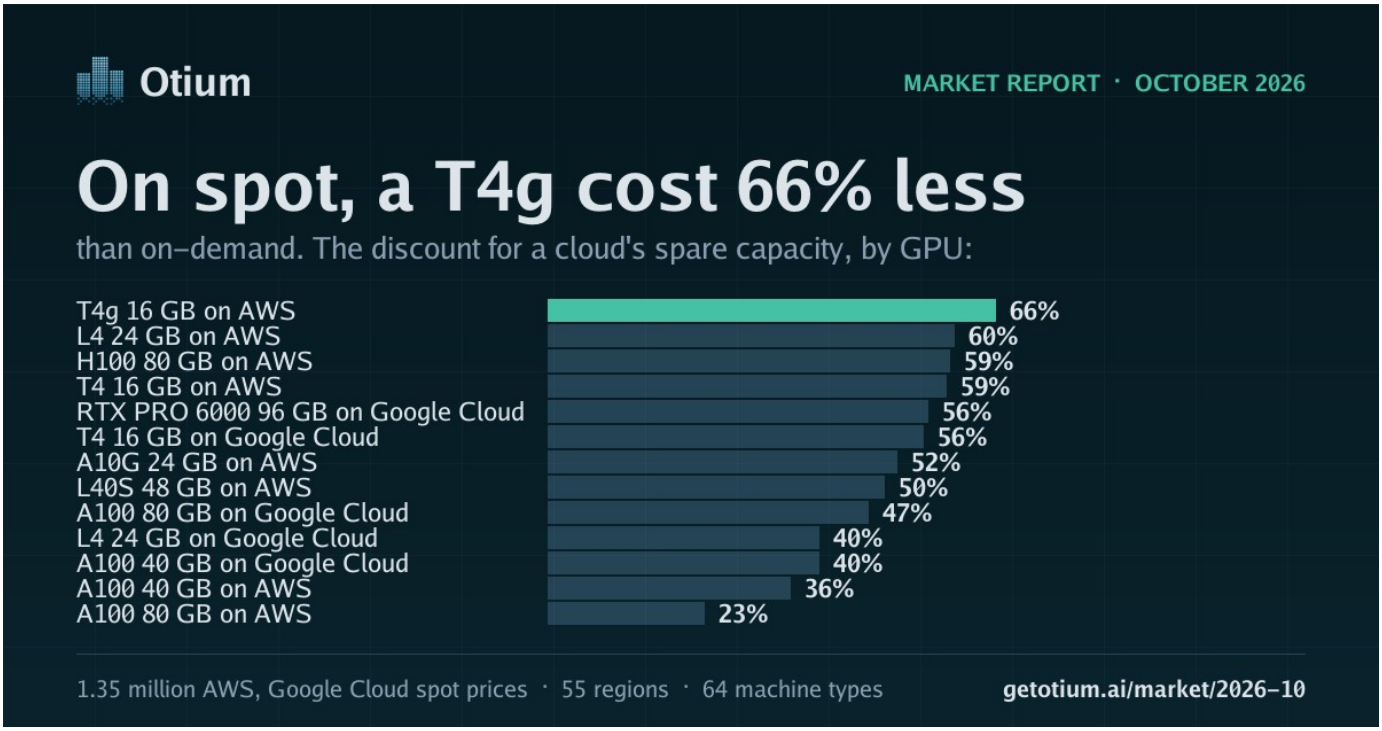
Where you rent still mattered a great deal. The same four–GPU L4 machine cost 10.8 times as much in Sydney as in Hyderabad, one of six AWS regions new to the report. Stockholm stayed the cheapest AWS region, though its prices rose 12%. This is also the first report to include Google Cloud, in 32 regions, and the two clouds priced the GPUs they both sell close together: an L4 was $0.42 on AWS and $0.47 on Google Cloud. Because of the new regions, a price here will not match the one September's report printed plus the change shown: the change compares only regions in both reports.

The full report, with every table, the data as CSV and a calculator for renting against owning a GPU:

getotium.ai/market/2026–10

What stood out





Otium

MARKET REPORT · OCTOBER 2026

Same machine, 10.8× the price

g6.24xlarge (4× L4) on AWS, spot, typical across the month

Hyderabad — $0.80 / hr

ap-south-2

Sydney — $8.68 / hr

ap-southeast-2

1.35 million AWS, Google Cloud spot prices · 55 regions · 64 machine types

getotium.ai/market/2026-10

Otium

MARKET REPORT · OCTOBER 2026

GPU prices, October 2026

Typical spot price per GPU-hour (AWS, Google Cloud); the faint bar is on-demand

T4g 16 GB on AWS	$0.13	(on-demand $0.52)
T4 16 GB on Google Cloud	$0.26	(on-demand $0.61)
T4 16 GB on AWS	$0.27	(on-demand $0.62)
L4 24 GB on AWS	$0.42	(on-demand $0.99)
L4 24 GB on Google Cloud	$0.47	(on-demand $0.79)
A10G 24 GB on AWS	$0.54	(on-demand $1.22)
L40S 48 GB on AWS	$1.53	(on-demand $2.69)
A100 40 GB on Google Cloud	$2.09	(on-demand $3.54)

1.35 million AWS, Google Cloud spot prices · 55 regions · 64 machine types

getotium.ai/market/2026-10

GPU prices

The cheapest machine for each GPU, per GPU-hour, typical (median) across the regions that sold it in October 2026.

GPU	Cloud	Spot	On-demand	Spot discount	Since September	Cheapest machine
T4g 16 GB	AWS	$0.13	$0.52	66%	—	g5g.xlarge
T4 16 GB	Google Cloud	$0.26	$0.61	56%	—	n1-standard-4
T4 16 GB	AWS	$0.27	$0.62	59%	-1%	g4dn.xlarge
L4 24 GB	AWS	$0.42	$0.99	60%	+1%	g6.xlarge
L4 24 GB	Google Cloud	$0.47	$0.79	40%	—	g2-standard-4
A10G 24 GB	AWS	$0.54	$1.22	52%	+1%	g5.xlarge
L40S 48 GB	AWS	$1.53	$2.69	50%	+2%	g6e.2xlarge
A100 40 GB	Google Cloud	$2.09	$3.54	40%	—	a2-megagpu-16g
A100 40 GB	AWS	$2.22	$3.29	36%	—	p4d.24xlarge
RTX PRO 6000 96 GB	Google Cloud	$2.35	$5.17	56%	+5%	g4-standard-192
A100 80 GB	AWS	$2.82	$4.12	23%	-12%	p4de.24xlarge
A100 80 GB	Google Cloud	$2.91	$5.54	47%	+2%	a2-ultragpu-1g
H100 80 GB	AWS	$3.00	$8.60	59%	+13%	p5.48xlarge

Regions

Each region's spot prices against the same machines' typical price everywhere, so a region that only sells small cards does not look cheap. Below zero is cheaper than typical.

Region	Cloud	Code	Against typical	Machines	Since September
Stockholm	AWS	eu-north-1	-48%	34	+12%
Hyderabad	AWS	ap-south-2	-44%	16	—
Milan	AWS	eu-south-1	-38%	7	—
São Paulo	AWS	sa-east-1	-37%	26	+1%
Spain	AWS	eu-south-2	-29%	30	—
Paris	AWS	eu-west-3	-24%	22	-2%
Zurich	AWS	eu-central-2	-17%	8	—
Osaka	AWS	ap-northeast-3	-8%	15	-1%
London	AWS	eu-west-2	-7%	26	0%
Jakarta	AWS	ap-southeast-3	-3%	11	—
Mumbai	AWS	ap-south-1	-1%	34	+9%
Ohio	AWS	us-east-2	-1%	34	0%
Malaysia	AWS	ap-southeast-5	0%	8	—
N. Virginia	AWS	us-east-1	0%	41	-3%
Oregon	AWS	us-west-2	0%	41	+2%
Ireland	AWS	eu-west-1	+9%	16	-1%

Region	Cloud	Code	Against typical	Machines	Since September
Frankfurt	AWS	eu-central-1	+13%	39	+1%
Seoul	AWS	ap-northeast-2	+14%	39	0%
Montréal	AWS	ca-central-1	+17%	24	0%
Singapore	AWS	ap-southeast-1	+20%	14	0%
N. California	AWS	us-west-1	+20%	8	+1%
Sydney	AWS	ap-southeast-2	+25%	25	−1%
Tokyo	AWS	ap-northeast-1	+36%	41	0%
Seoul	Google Cloud	asia-northeast3	−27%	14	+10%
Mumbai	Google Cloud	asia-south1	−20%	14	0%
N. Virginia	Google Cloud	us-east4	−19%	18	0%
Columbus	Google Cloud	us-east5	−14%	8	+7%
Iowa	Google Cloud	us-central1	−9%	23	0%
South Carolina	Google Cloud	us-east1	−9%	18	0%
Oregon	Google Cloud	us-west1	−9%	18	0%
Salt Lake City	Google Cloud	us-west3	−8%	10	−1%
Netherlands	Google Cloud	europe-west4	−4%	23	0%
Montréal	Google Cloud	northamerica-northeast1	0%	10	0%
Toronto	Google Cloud	northamerica-northeast2	0%	8	0%
Belgium	Google Cloud	europe-west1	0%	14	+8%
Las Vegas	Google Cloud	us-west4	+2%	18	+6%
London	Google Cloud	europe-west2	+4%	14	0%
Tokyo	Google Cloud	asia-northeast1	+4%	18	−4%
Tel Aviv	Google Cloud	me-west1	+4%	6	0%
Taiwan	Google Cloud	asia-east1	+5%	14	+5%
Frankfurt	Google Cloud	europe-west3	+7%	10	0%
São Paulo	Google Cloud	southamerica-east1	+9%	6	—
Singapore	Google Cloud	asia-southeast1	+12%	23	0%
Jakarta	Google Cloud	asia-southeast2	+12%	6	+5%
Zurich	Google Cloud	europe-west6	+23%	8	0%
Dammam	Google Cloud	me-central2	+37%	12	−1%

Renting against owning a GPU

Owning a card means paying for it once, then for its electricity. Renting means paying by the hour. Each home setup is paired with the rented machine that holds the same GPU memory, which decides whether a model fits. The extra to rent is the rented price, on–demand, less the setup's electricity at the US average home rate for August 2026 (18.16¢ per kWh) at its maker's rated maximum power. Divide what the setup would cost you by it for the hours of use before buying costs less than renting would have.

Home setup	Memory	Rented instead	On–demand	Extra / hr	Hrs / $1,000	VRAM speed
RTX 3090	24 GB	1× L4 24 GB	$0.99	$0.93	1,080	~3x slower
RTX 4090	24 GB	1× L4 24 GB	$0.99	$0.91	1,101	~3x slower
RTX 5090	32 GB	1× L40S 48 GB	$2.69	$2.59	386	~2.1x slower
2× RTX 3090	48 GB	1× L40S 48 GB	$2.69	$2.57	390	~2.2x slower
RTX 6000 Ada	48 GB	1× L40S 48 GB	$2.69	$2.64	379	about the same
RTX PRO 6000 Blackwell	96 GB	1× RTX PRO 6000 96 GB	$5.17	$5.07	197	about the same
ASUS Ascent GX10	128 GB	2× RTX PRO 6000 96 GB	$10.35	$10.31	97	~12x faster

It pairs by memory, so read the speed column with it: memory speed (bandwidth, the makers' figures) sets how fast text comes out once a model fits. Where the rented machine is many times faster, it does the same work in far fewer hours than the break–even counts. The break–even counts the card as worth nothing at the end, so a reader who expects to sell it takes that off the price, and leaves out the rest of the computer (except for a whole computer). On spot the hours roughly double, for work that can stop and resume.



How to read this

Spot is spare capacity a cloud sells at a discount, on one condition: it can take the machine back at short notice, so it suits work that can pause and resume. On–demand is the standard price; the machine is yours until you stop it.

A GPU–hour is one GPU for one hour. Prices are divided by the number of GPUs in a machine, so an eight–GPU server and a single card compare fairly. A median is the middle value: half the prices were lower and half higher, so one odd reading barely moves it.

Where the numbers come from: 1.35 million AWS, Google Cloud spot prices across 55 regions and 64 machine types, recorded at least every half hour on AWS and at least once a day on Google Cloud. Each figure is the median of those readings. Discounts compare spot and on–demand in the same region only. Prices are for Linux, before tax.

A price is not a guarantee. These are the prices the clouds published. They do not prove a machine could actually be rented at that price. A region that looks cheap may also be one where spot machines ran out.

Use and cite

The figures are free to use, chart and republish under CC BY 4.0, with credit. Every row is a CSV download on the report page.

Cite as: Otium, "AI compute market report: October 2026", getotium.ai/market/2026–10

getotium.ai/market/2026–10