Otium

AI COMPUTE MARKET REPORT

GPU prices, September 2026

5.52 million AWS spot prices across 17 regions and 41 machine types.



Inside

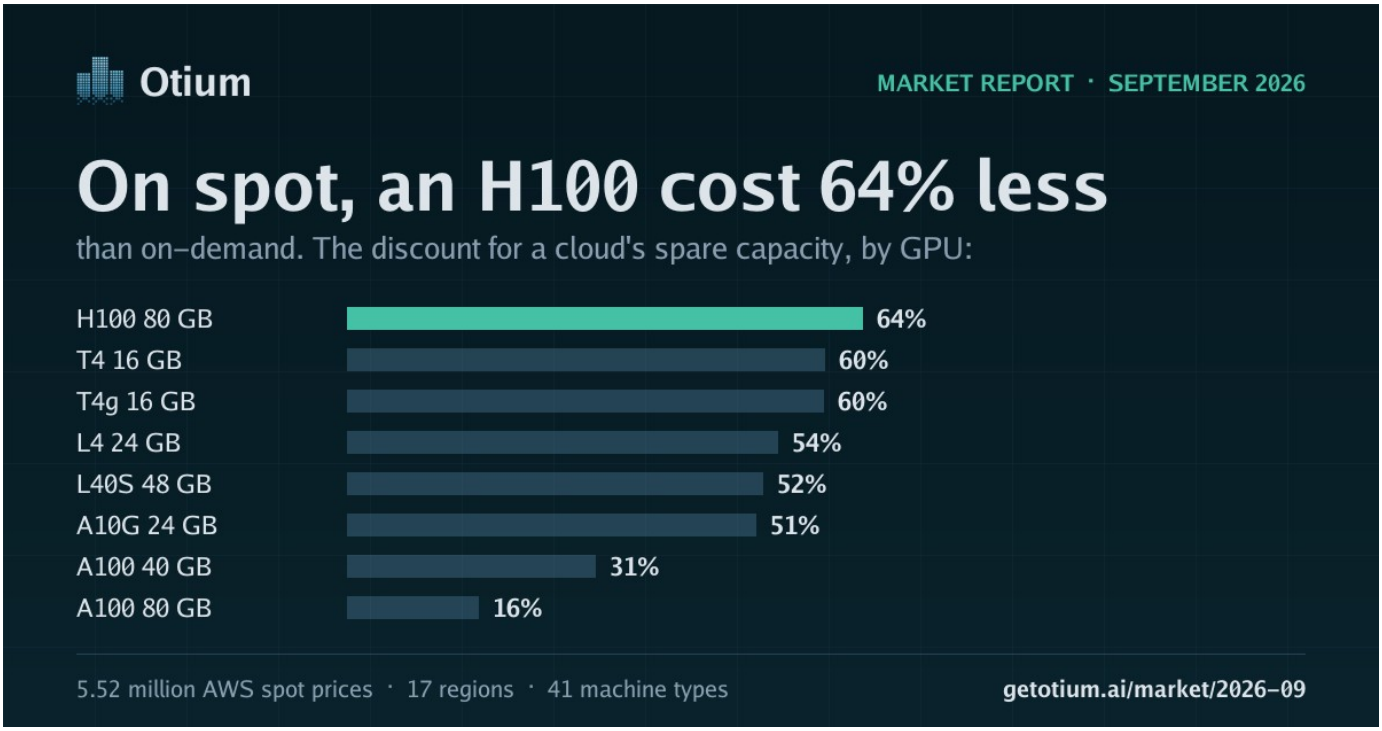
- What stood out: the month's findings, one to a chart
- GPU prices: spot and on–demand per GPU–hour, and the change since last month
- Regions: where the same machines cost more or less
- Renting against owning a GPU: the hours of use before buying costs less
- How to read this, and how to use and cite the data

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

getotium.ai/market/2026–09

What stood out





Otium

MARKET REPORT · SEPTEMBER 2026

Same machine, 10.2× the price

g5.16xlarge (1× A10G), spot, typical across the month

Stockholm — $0.58 / hr

eu-north-1

São Paulo — $5.91 / hr

sa-east-1

5.52 million AWS spot prices · 17 regions · 41 machine types

getotium.ai/market/2026-09

Otium

MARKET REPORT · SEPTEMBER 2026

GPU prices, September 2026

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

T4g 16 GB	$0.19	(on-demand $0.52)
T4 16 GB	$0.28	(on-demand $0.61)
L4 24 GB	$0.48	(on-demand $0.99)
A10G 24 GB	$0.56	(on-demand $1.22)
L40S 48 GB	$1.38	(on-demand $2.69)
A100 40 GB	$2.05	(on-demand $3.29)
H100 80 GB	$2.57	(on-demand $8.43)
A100 80 GB	$3.85	(on-demand $4.12)

5.52 million AWS spot prices · 17 regions · 41 machine types

getotium.ai/market/2026-09

GPU prices

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

GPU	Spot	On–demand	Spot discount	Since August	Cheapest machine
T4g 16 GB	$0.19	$0.52	60%	+18%	g5g.xlarge
T4 16 GB	$0.28	$0.61	60%	+3%	g4dn.xlarge
L4 24 GB	$0.48	$0.99	54%	—	g6.xlarge
A10G 24 GB	$0.56	$1.22	51%	+5%	g5.xlarge
L40S 48 GB	$1.38	$2.69	52%	+15%	g6e.2xlarge
A100 40 GB	$2.05	$3.29	31%	–4%	p4d.24xlarge
H100 80 GB	$2.57	$8.43	64%	—	p5.48xlarge
A100 80 GB	$3.85	$4.12	16%	—	p4de.24xlarge

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	Code	Against typical	Machines compared
Stockholm	eu–north–1	–54%	34
São Paulo	sa–east–1	–44%	26
Paris	eu–west–3	–21%	23
Osaka	ap–northeast–3	–13%	15
London	eu–west–2	–8%	26
Mumbai	ap–south–1	–7%	34
Oregon	us–west–2	–4%	41
Ohio	us–east–2	–3%	34
N. Virginia	us–east–1	+1%	41
Frankfurt	eu–central–1	+5%	39
Montréal	ca–central–1	+6%	25
Seoul	ap–northeast–2	+8%	39
Singapore	ap–southeast–1	+9%	14
Ireland	eu–west–1	+9%	16
N. California	us–west–1	+17%	8
Sydney	ap–southeast–2	+22%	25
Tokyo	ap–northeast–1	+29%	41

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 July 2026 (18.31¢ 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	1,102	~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	$2.56	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	4× L4 24 GB	$5.66	$5.55	180	~1.5x slower
ASUS Ascent GX10	128 GB	4× L40S 48 GB	$12.60	$12.56	80	~13x 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: 5.52 million AWS spot prices across 17 regions and 41 machine types, recorded every ten minutes. 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: September 2026", getotium.ai/market/2026–09

getotium.ai/market/2026–09