






GPU Price Comparison Jan 22

👤 Created by	 Yagn Purohit
🕒 Created time	@January 24, 2023 12:06 PM
👤 Last edited by	 Twishmay
🕒 Last edited time	@April 25, 2023 8:44 PM
👥 People	 Yagn Purohit
⚙️ Status	Done
🏷️ Tags	<div>Algorithms</div> <div>Costing</div> <div>Database</div> <div>GPU</div> <div>Pricing</div> <div>Technology</div>
🔗 URL	https://www.notion.so/psytech-hq/Price-comparision-GPU-71c94f82e38c4d9a937bf751746333e8

Summary

Provider	Type	Hourly Cost	Monthly Cost
GCP (Spot)	T4	\$0.11/hr	\$88.40/month
Brev (Spot)	T4	\$0.20/hr	\$152.00/month
E2E	T4	\$0.37/hr	\$280.40/month
Azure	T4	\$0.526/hr	\$396.92/month
AWS	T4	\$0.526/hr	\$396.92/month
Paperspace	A4000	\$0.76/hr	\$576.96/month
Banana	A100	\$1.87/hr	\$1398.80/month
Replicate	T4	\$1.98/hr	\$1497.12/month
IBM	P100	\$1.95/hr	\$1466.80/month
OVH Cloud	v100 (16GB)	\$1.97/hr	\$1479.60/month
LamdaLabs	RTX 6000	\$0.5/hr	\$380/month
LamdaLabs	A100	\$1.10/hr	\$828/month

Considerations

- Cheapest option
- Best option
- Spot vs Monthly blocked
- T4 vs others
- Production
- Deployment
- Prototyping
- Experiments

Detailed Pricing

Platform	GPU	Memory	Price
HuggingFace	T4	14	\$0.6/hr
	A10G	24	\$1.3/hr
	4x T4	56	\$4.5/hr
Banana	A100	16	\$1.87/hr
GCP	A100	80	\$(3.93/hr)/GPU
	A100	40	\$(2.934/hr)/GPU
	T4	16	\$(0.35/hr)/GPU
GCP (Spot)	A100	80	\$(1.25/hr)/GPU
	A100	40	\$(0.88/hr)/GPU
	T4	16	\$(0.11/hr)/GPU
	For more	Check toggle	
Azure	T4	16	\$0.526/hr
	8x A100	192	\$27.197/hr
	For more	Check toggle	
E2E	A100	40	Rs. 170/hr (\$2.09)
	A100	80	Rs. 226/hr (\$2.77)
	T4	16	Rs. 30/hr (\$0.37)
	For more	Check toggle	
Brev	T4	16	\$0.63/hr
	8x A100	320	\$39.33/hr
Brev (SPOT)	T4	16	\$0.20/hr
	8x A100	320	\$11.80/hr
	For more	Check toggle	
Replicate	T4	16	\$1.98/hr
	A100	40	\$8.28/hr
AWS	T4	NM (prob 16)	\$0.526/hr
	AMD V520	NM	\$0.379/hr
Paperspace	A100	NM (prob 40)	\$3.09/hr

Platform	GPU	Memory	Price
	A4000	NM (prob16)	\$0.76/hr
	For more	Check toggle	
IBM	V100	NM (16/32)	\$2.49/hr
	P100	NM (prob16)	\$1.95/hr
OVH Cloud	v100	16	\$1.97/hr
	v100	32	\$2.19/hr
LamdaLabs	A100	40	\$1.10/hr
	RTX 6000	24	\$0.5/hr
	For more	Check toggle	
GPU mart	RTX A4000	16	\$129/mo
	K40	12	\$109/mo
	For more	Check toggle	
Genesis Cloud	RTX 3090	24	\$1.30/hr
	RTX 3080	10	\$0.90/hr
	For more	Check toggle	

▼ HuggingFace

It doesn't stop costing while not using.

GPU instances

Provider	Architecture	GPUs	Memory	Hourly rate
aws	NVIDIA T4	1	14GB	\$0.60
aws	NVIDIA A10G	1	24GB	\$1.30
aws	NVIDIA T4	4	56GB	\$4.50
aws	NVIDIA A100	4	80GB	Coming soon
aws	NVIDIA A10G	4	96GB	Coming soon
aws	NVIDIA A100	8	640GB	Coming soon

▼ Banana

Serverless Pricing

Only pay for the resources you use. That's the power of Banana.

Usage Pricing

Only pay for GPU compute you use

\$0.00051992 / second

- ✓ 1 hour of FREE credits 🎁
- ✓ Run on A100 GPUs
- ✓ ML Models up to 16GB
- ✓ Autoscaling
- ✓ Spike Tolerance (up to 25 replicas)
- ✓ Network Payload up to 50MB
- ✓ System Latency ~700ms
- ✓ Avg. Cold Boot ~5 seconds

[Sign Up for Free](#)

Enterprise

For companies spending \$1k+ monthly on ML cloud compute

10-40% / off usage rate

- ✓ Everything in Usage Pricing
- ✓ Best for companies spending \$1000+ month on ML hosting
- ✓ Dedicated SLA Response Time
- ✓ Increased Spike Tolerance (25+ replicas)

[Contact Us](#)

▼ GCP

Model	GPUs	GPU memory	GPU price (USD)	Spot price* (USD)	1 year commitment price** (USD)	3 year commitment price** (USD)
NVIDIA A100 80GB	1 GPU	80 GB HBM2	\$3.93 per GPU	\$1.25 per GPU	**	**
	2 GPUs	160 GB HBM2				
	4 GPUs	320 GB HBM2				
	8 GPUs	640 GB HBM2				
NVIDIA A100 40GB	1 GPU	40 GB HBM2	\$2.934 per GPU	\$0.880 per GPU	\$1.848 per GPU	\$1.027 per GPU
	2 GPUs	80 GB HBM2				
	4 GPUs	160 GB HBM2				
	8 GPUs	320 GB HBM2				
	16 GPUs	640 GB HBM2				
NVIDIA T4	1 GPU	16 GB GDDR6	\$0.35 per GPU	\$0.11 per GPU	\$0.220 per GPU	\$0.160 per GPU
	2 GPUs	32 GB GDDR6				
	4 GPUs	64 GB GDDR6				

NVIDIA P4 🔗	1 GPU	8 GB GDDR5	\$0.60 per GPU	\$0.216 per GPU	\$0.378 per GPU	\$0.270 per GPU
	2 GPUs	16 GB GDDR5				
	4 GPUs	32 GB GDDR5				
NVIDIA V100 🔗	1 GPU	16 GB HBM2	\$2.48 per GPU	\$0.74 per GPU	\$1.562 per GPU	\$1.116 per GPU
	2 GPUs	32 GB HBM2				
	4 GPUs	64 GB HBM2				
	8 GPUs	128 GB HBM2				
NVIDIA P100 🔗	1 GPU	16 GB HBM2	\$1.46 per GPU	\$0.43 per GPU	\$0.919 per GPU	\$0.657 per GPU
	2 GPUs	32 GB HBM2				
	4 GPUs	64 GB HBM2				
NVIDIA K80 🔗	1 GPU	12 GB GDDR5	\$0.45 per GPU	\$0.038 per GPU	\$0.283 per GPU	Not available in this region
	2 GPUs	24 GB GDDR5				
	4 GPUs	48 GB GDDR5				
	8 GPUs	96 GB GDDR5				

NVIDIA RTX virtual workstations (formerly known as NVIDIA GRID)						
NVIDIA T4 Virtual Workstation 🔗	1 GPU	16 GB GDDR6	\$0.55 per GPU	\$0.31 per GPU	\$0.42 per GPU	\$0.36 per GPU
	2 GPUs	32 GB GDDR6				
	4 GPUs	64 GB GDDR6				
NVIDIA P4 Virtual Workstation 🔗	1 GPU	8 GB GDDR5	\$0.80 per GPU	\$0.416 per GPU	\$0.578 per GPU	\$0.47 per GPU
	2 GPUs	16 GB GDDR5				
	4 GPUs	32 GB GDDR5				
NVIDIA P100 Virtual Workstation 🔗	1 GPU	16 GB HBM2	\$1.66 per GPU	\$0.63 per GPU	\$1.119 per GPU	\$0.857 per GPU
	2 GPUs	32 GB HBM2				
	4 GPUs	64 GB HBM2				

▼ Azure

NCas_T4_v3 Series

Instance	vCPU(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
NC4as T4 v3	4	28 GiB	1X T4	\$0.526/hour	\$0/hour	\$0.526/hour	\$0.310/hour ~41% savings	\$0.198/hour ~62% savings
NC8as T4 v3	8	56 GiB	1X T4	\$0.752/hour	\$0/hour	\$0.752/hour	\$0.443/hour ~41% savings	\$0.283/hour ~62% savings
NC16as T4 v3	16	110 GiB	1X T4	\$1.204/hour	\$0/hour	\$1.204/hour	\$0.708/hour ~41% savings	\$0.453/hour ~62% savings
NC64as T4 v3	64	440 GiB	4X T4	\$4.352/hour	\$0/hour	\$4.352/hour	\$2.560/hour ~41% savings	\$1.637/hour ~62% savings

NDv2 series

Instance	Core(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
ND40rs v2	40	672 GiB	8X V100 (NVlink)	\$22.032/hour	\$0/hour	\$22.032/hour	\$10.796/hour ~51% savings	N/A

ND A100 v4 series

Instance	Core(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
ND96asr A100 v4	96	900 GiB	8x A100 (NVlink)	\$27.197/hour	\$0/hour	\$27.197/hour	\$18.829/hour ~31% savings	\$10.879/hour ~60% savings

NC-series

Instance	vCPU(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
NC6	6	56 GiB	1X K80	\$0.90/hour	\$0/hour	\$0.90/hour	\$0.574/hour ~36% savings	\$0.400/hour ~56% savings
NC12	12	112 GiB	2X K80	\$1.80/hour	\$0/hour	\$1.80/hour	N/A	N/A
NC24	24	224 GiB	4X K80	\$3.60/hour	\$0/hour	\$3.60/hour	\$2.294/hour ~36% savings	\$1.599/hour ~56% savings
NC24r	24	224 GiB	4X K80	\$3.96/hour	\$0/hour	\$3.96/hour	N/A	N/A

NCsv2-series

Instance	vCPU(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
NC6s v2	6	112 GiB	1X P100	\$2.07/hour	\$0/hour	\$2.07/hour	\$1.319/hour ~36% savings	N/A
NC12s v2	12	224 GiB	2X P100	\$4.14/hour	\$0/hour	\$4.14/hour	N/A	\$1.838/hour ~56% savings
NC24s v2	24	448 GiB	4X P100	\$8.28/hour	\$0/hour	\$8.28/hour	\$5.275/hour ~36% savings	N/A
NC24rs v2	24	448 GiB	4X P100	\$9.108/hour	\$0/hour	\$9.108/hour	\$5.802/hour ~36% savings	N/A

NCsv3-series

Instance	vCPU(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
NC6s v3	6	112 GiB	1X V100	\$3.06/hour	\$0/hour	\$3.06/hour	\$1.950/hour ~36% savings	\$0.980/hour ~68% savings
NC12s v3	12	224 GiB	2X V100	\$6.12/hour	\$0/hour	\$6.12/hour	\$3.899/hour ~36% savings	\$1.959/hour ~68% savings
NC24s v3	24	448 GiB	4X V100	\$12.24/hour	\$0/hour	\$12.24/hour	\$7.797/hour ~36% savings	\$3.917/hour ~68% savings
NC24rs v3	24	448 GiB	4X V100	\$13.46/hour	\$0/hour	\$13.46/hour	\$8.577/hour ~36% savings	\$4.336/hour ~68% savings

NV-series

Instance	vCPU(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
NV6	6	56 GiB	1X M60	\$1.14/hour	\$0/hour	\$1.14/hour	N/A	N/A
NV12	12	112 GiB	2X M60	\$2.28/hour	\$0/hour	\$2.28/hour	N/A	\$1.012/hour ~56% savings
NV24	24	224 GiB	4X M60	\$4.56/hour	\$0/hour	\$4.56/hour	\$2.905/hour ~36% savings	N/A

NCsv3-series

Instance	vCPU(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
NC6s v3	6	112 GiB	1X V100	\$3.06/hour	\$0/hour	\$3.06/hour	\$1.950/hour ~36% savings	\$0.980/hour ~68% savings
NC12s v3	12	224 GiB	2X V100	\$6.12/hour	\$0/hour	\$6.12/hour	\$3.899/hour ~36% savings	\$1.959/hour ~68% savings
NC24s v3	24	448 GiB	4X V100	\$12.24/hour	\$0/hour	\$12.24/hour	\$7.797/hour ~36% savings	\$3.917/hour ~68% savings
NC24rs v3	24	448 GiB	4X V100	\$13.46/hour	\$0/hour	\$13.46/hour	\$8.577/hour ~36% savings	\$4.336/hour ~68% savings

NV-series

Instance	vCPU(s)	RAM	GPU	Linux VM Price	Machine Learning Service Surcharge	Pay As You Go Total Price	1 year reserved total price	3 year reserved total price
NV6	6	56 GiB	1X M60	\$1.14/hour	\$0/hour	\$1.14/hour	N/A	N/A
NV12	12	112 GiB	2X M60	\$2.28/hour	\$0/hour	\$2.28/hour	N/A	\$1.012/hour ~56% savings
NV24	24	224 GiB	4X M60	\$4.56/hour	\$0/hour	\$4.56/hour	\$2.905/hour ~36% savings	N/A

▼ E2E

For Windows the prices will be up Rs. 5-6 in hourly rate

Plan	GPU Memory	vCPUs	Dedicated RAM	Disk Space	Hourly Billing	Weekly Billing	Monthly Billing (Save 39%)	
GDC.A100- 16.115GB	1 × 40 GB	16 vCPUs	115 GB	1500 GB SSD	₹170/hr	₹25000/week	₹75,000/mo	Create
GDC.2xA100- 32.230GB	2 × 40 GB	32 vCPUs	230 GB	3000 GB SSD	₹340/hr	₹50000/week	₹1,50,000/mo	Create
GDC.4xA100- 64.460GB	4 × 40 GB	64 vCPUs	460 GB	6000 GB SSD	₹680/hr	₹100000/week	₹3,00,000/mo	Create
GDC.8xA100- 128.920GB	8 × 40 GB	128 vCPUs	920 GB	6000 GB SSD	₹1360/hr	₹215000/week	₹7,00,000/mo	Create
GDC.A10080-16.115GB	1 × 80 GB	16 vCPUs	115 GB	1500 GB SSD	₹226/hr	₹33000/week	₹1,00,000/mo	Create
GDC.2xA10080-32.230GB	2 × 80 GB	32 vCPUs	230 GB	3000 GB SSD	₹452/hr	₹66000/week	₹2,00,000/mo	Create
GDC.4xA10080-64.460GB	4 × 80 GB	64 vCPUs	460 GB	6000 GB SSD	₹904/hr	₹132000/week	₹4,00,000/mo	Create
Spot_GDC3.A10080-16.115GB	1 × 80 GB	16 vCPUs	230 GB	250 GB SSD	₹100/hr	NA	NA	Create
Spot_GDC3.2xA10080-32.230GB	2 × 80 GB	32 vCPUs		250 GB SSD	₹200/hr	NA	NA	Create
Spot_GDC3.4xA10080-64.460GB	4 × 80 GB	64 vCPUs		250 GB SSD	₹400/hr	NA	NA	Create

Linux T4 GPU Dedicated Compute

Plan	GPU Memory	vCPUs	Dedicated RAM	Disk Space	Hourly Billing	Weekly Billing	Monthly Billing (Save 20%)	
GDC.T4-12.50GB	1 × 16 GB	12 vCPUs	50 GB	900 GB SSD	₹30/hr	NA	₹17,500/mo	Create
GDC.2XT4-24.100GB	2 × 16 GB	24 vCPUs	100 GB	1800 GB SSD	₹60/mo	NA	₹36,000/mo	Create

Linux A40 GPU Dedicated Compute

Plan	GPU Memory	vCPUs	Dedicated RAM	Disk Space	Hourly Billing	Weekly Billing	Monthly Billing (Save 22%)	
GDC.A40-16.100GB	1 × 48 GB	16 vCPUs	100 GB	750 GB SSD	₹96/hr	₹14750/week	₹54,500/mo	Create
GDC.2xA40-32.200GB	2 × 48 GB	32 vCPUs	200 GB	1500 GB SSD	₹193/hr	₹29500/week	₹1,09,000/mo	Create
GDC.4xA40-64.400GB	4 × 48 GB	64 vCPUs	400 GB	3000 GB SSD	₹386/hr	₹58500/week	₹2,18,000/mo	Create

Linux A30 GPU Dedicated Compute

Plan	GPU Memory	vCPUs	Dedicated RAM	Disk Space	Hourly Billing	Weekly Billing	Monthly Billing (Save 39%)	
GDC.A30-16.90GB	1 × 24 GB	16 vCPUs	90 GB	640 GB SSD	₹90/hr	₹12000/week	₹40,000/mo	Create
GDC.2xA30-32.180GB	2 × 24 GB	32 vCPUs	180 GB	1280 GB SSD	₹180/hr	₹24000/week	₹80,000/mo	Create
GDC.4xA30-64.360GB	4 × 24 GB	64 vCPUs	360 GB	2560 GB SSD	₹360/hr	₹48000/week	₹1,60,000/mo	Create

Linux A30 GPU Dedicated Compute

Plan	GPU Memory	vCPUs	Dedicated RAM	Disk Space	Hourly Billing	Weekly Billing	Monthly Billing (Save 39%)	
GDC.A30-16.90GB	1 × 24 GB	16 vCPUs	90 GB	640 GB SSD	₹90/hr	₹12000/week	₹40,000/mo	Create
GDC.2xA30-32.180GB	2 × 24 GB	32 vCPUs	180 GB	1280 GB SSD	₹180/hr	₹24000/week	₹80,000/mo	Create
GDC.4xA30-64.360GB	4 × 24 GB	64 vCPUs	360 GB	2560 GB SSD	₹360/hr	₹48000/week	₹1,60,000/mo	Create

Linux RTX 8000 GPU Dedicated Compute

Plan	GPU Memory	vCPUs	Dedicated RAM	Disk Space	Hourly Billing	Weekly Billing	Monthly Billing (Save 24%)	
GDC.RTX-16.115GB	1 × 48GB (DDR6)	16 vCPUs	115 GB	900 GB SSD	₹72/hr	₹12000/week	₹40,000/mo	Create
GDC.2xRTX-32.230GB	2 × 48GB (DDR6)	32 vCPUs	230 GB	1800 GB SSD	₹144/hr	₹24000/week	₹80,000/mo	Create
GDC.4xRTX-64.460GB	4 × 48GB (DDR6)	64 vCPUs	460 GB	3600 GB SSD	₹288/hr	₹48000/week	₹1,60,000/mo	Create

Entry Level Cloud GPUs

Entry-level cloud GPUs are carved out of A100-80 GB NVIDIA GPU cards using MIG technology. With MIG, customers will be able to see and schedule jobs on their virtual GPU instances as if they were physical GPUs. MIG works with Linux operating systems, supports containers using Docker Engine, with support for Kubernetes. MIG allows multiple vGPUs (and thereby VMs) to run in parallel on a single GPU while preserving the isolation guarantees that vGPU provides.

Plan	GPU Memory	vCPUs	Dedicated RAM	Disk Space	Hourly Billing	Weekly Billing	Monthly Billing (Save 20%)	
VGC3.A10010-4.30GB	10	4 vCPUs	30 GB	250 GB SSD	₹30/hr	₹4,500/week	₹15,000/mo	Create

▼ Brev

GPU Model	vCPU	RAM	vRAM	Hourly	Spot *
1 x AMD V520	-	-	8	\$0.45	\$0.14
1 x Nvidia T4	-	-	16	\$0.63	\$0.20
1 x Nvidia K520	-	-	8	\$0.78	\$0.24
1 x Nvidia M60	-	-	8	\$0.90	\$0.30
1 x Nvidia K80	-	-	12	\$1.08	\$0.38
1 x Nvidia A10G	-	-	24	\$1.21	\$0.37
2 x AMD V520	-	-	16	\$2.08	\$0.63
2 x Nvidia M60	-	-	16	\$2.74	\$0.83
2x Nvidia K520	-	-	16	\$3.12	-
1 x Nvidia V100	-	-	16	\$3.67	\$1.11
4 x AMD V520	-	-	32	\$4.16	\$1.25
4 x Nvidia T4	-	-	64	\$4.69	\$1.53
4 x Nvidia M60	-	-	32	\$5.47	\$1.65
4 x Nvidia A10G	-	-	96	\$8.81	\$2.05
8 x Nvidia K80	-	-	96	\$8.64	\$2.60
16 x Nvidia K80	-	-	192	\$17.28	\$5.19
8 x Nvidia A10G	-	-	192	\$19.55	\$5.87
8 x Nvidia A100	-	-	320	\$39.33	\$11.80

▼ Replicate

Pricing

You can use Replicate for free, but after a bit you'll be asked to enter your credit card. You pay by the second for the predictions you run. The price per second varies based on the hardware the model is run on.

CPU

\$0.0002 per second
(or, \$0.012 per minute)

4x CPU
8GB RAM

Nvidia T4 GPU

\$0.00055 per second
(or, \$0.033 per minute)

4x CPU
16GB GPU RAM
8GB RAM

Nvidia A100 GPU

\$0.0023 per second
(or, \$0.138 per minute)

8x CPU
40GB GPU RAM
40GB RAM


▼ AWS

G4dn = T4

G4ad = v520

	Instance Size	GPU	vCPUs	Memory (GiB)	Instance Storage (GB)	Network Bandwidth (Gbps)	EBS Bandwidth (Gbps)	On-Demand Price/hr*	1-yr Reserved Instance Effective Hourly* (Linux)	3-yr Reserved Instance Effective Hourly* (Linux)
G4dn										
Single GPU VMs	g4dn.xlarge	1	4	16	1 x 125 NVMe SSD	Up to 25	Up to 3.5	\$0.526	\$0.316	\$0.210
	g4dn.2xlarge	1	8	32	1 x 225 NVMe SSD	Up to 25	Up to 3.5	\$0.752	\$0.452	\$0.300
	g4dn.4xlarge	1	16	64	1 x 225 NVMe SSD	Up to 25	4.75	\$1.204	\$0.722	\$0.482
	g4dn.8xlarge	1	32	128	1 x 900 NVMe SSD	50	9.5	\$2.176	\$1.306	\$0.870
	g4dn.16xlarge	1	64	256	1 x 900 NVMe SSD	50	9.5	\$4.352	\$2.612	\$1.740
Multi GPU VMs	g4dn.12xlarge	4	48	192	1 x 900 NVMe SSD	50	9.5	\$3.912	\$2.348	\$1.564
	g4dn.metal	8	96	384	2 x 900 NVMe SSD	100	19	\$7.824	\$4.694	\$3.130
G4ad										
Single GPU VMs	g4ad.xlarge	1	4	16	1 x 150 NVMe SSD	Up to 10	Up to 3	\$0.379	\$0.227	\$0.178
	g4ad.2xlarge	1	8	32	1 x 300 NVMe SSD	Up to 10	Up to 3	\$0.541	\$0.325	\$0.254
	g4ad.4xlarge	1	16	64	1 x 600 NVMe SSD	Up to 10	Up to 3	\$0.867	\$0.520	\$0.405
Multi GPU VMs	g4ad.8xlarge	2	32	128	1 x 1200 NVMe SSD	15	3	\$1.734	\$1.040	\$0.810
	g4ad.16xlarge	4	64	256	1 x 2400 NVMe SSD	25	6	\$3.468	\$2.081	\$1.619

▼ Paperspace

Dedicated GPU 		
M4000 \$ 0.45 / hour 8 GB GPU 30 GB RAM 8 vCPU	P4000 \$ 0.51 / hour 8 GB GPU 30 GB RAM 8 vCPU	P5000 \$ 0.78 / hour 16 GB GPU 30 GB RAM 8 vCPU
P6000 \$ 1.10 / hour 24 GB GPU 30 GB RAM 8 vCPU	V100 \$ 2.30 / hour 16 GB GPU 30 GB RAM 8 vCPU	RTX4000 \$ 0.56 / hour NVIDIA RTX4000 GPU 30 GB RAM 8 vCPU
RTX5000 \$ 0.82 / hour NVIDIA RTX5000 GPU 30GB RAM 8 vCPU	A4000 \$ 0.76 / hour NVIDIA A4000 GPU 45GB RAM 8 vCPU	A5000 \$ 1.38 / hour NVIDIA A5000 GPU 45GB RAM 8 vCPU
A6000 \$ 1.89 / hour NVIDIA A6000 GPU 45GB RAM 8 vCPU	A100 \$ 3.09 / hour NVIDIA A100 GPU 90GB RAM 12 vCPU	H100 <small>NEW</small> <div>Reserve Now</div>

▼ IBM

GPU options on virtual servers for VPC

Gx2-8x64x1v

8 vCPU / 64 GiB RAM / 1 x V100 GPU / Starts at: USD 2.49/hr

[Get a quote](#) →

Gx2-16x128x2v

16 vCPU / 128 GiB / RAM 2 x V100 GPU / Starts at: USD 4.99/hr

[Get a quote](#) →

Gx2-32x256x2v

32 vCPU / 256 GiB / RAM 2 x V100 GPU / Starts at: USD 5.98/hr

[Get a quote](#) →

Virtual Server Classic GPU options

AC1.8x60

8 vCPU / 60 GB RAM / 1 x P100 GPU / Starts at: USD 1.95/hr

[Get a quote](#) →

AC2.8x60

8 vCPU / 20 GB RAM / 1 x V100 GPU / Starts at: USD 3.06/hr

[Get a quote](#) →

AC2.8x60

8 vCPU / 60 GB RAM / 1 x V100 GPU / Starts at: USD 2,233/mo

[Get a quote](#) →

▼ OVH Cloud

Name	Memory	vCore	GPU	Storage	Public network	Private network	Price
t1-45	45 GB	8	Tesla V100 16 GB	400 GB SSD	2 Gbps guaranteed	4 Gbps max.	\$1.97 /hour
t1-90	90 GB	18	2×Tesla V100 16 GB	800 GB SSD	4 Gbps guaranteed	4 Gbps max.	\$3.94 /hour
t1-180	180 GB	36	4×Tesla V100 16 GB	50 GB SSD + 2 TB	10 Gbps	4 Gbps max.	\$7.89 /hour
t2-45	45 GB	14	Tesla V100S 32 GB	400 GB SSD	2 Gbps guaranteed	4 Gbps max.	\$2.19 /hour
t2-90	90 GB	28	2×Tesla V100S 32 GB	800 GB SSD	4 Gbps guaranteed	4 Gbps max.	\$4.38 /hour
t2-180	180 GB	56	4×Tesla V100S 32 GB	50 GB SSD + 2 TB NVMe	10 Gbps	4 Gbps max.	\$8.76 /hour

NGC integration is up to 1,000 times faster than a CPU on parallel computing, making it easy to use TensorFlow, Caffe, MXNet, and much more.

▼ LamdaLabs

GPUs	VRAM per GPU	vCPUs	RAM	Storage	Price
1x NVIDIA A100	40 GB	30	200 GiB	512 GiB	\$1.10 / hr
2x NVIDIA A100	40 GB	60	400 GiB	1 TiB	\$2.20 / hr
4x NVIDIA A100	40 GB	120	800 GiB	1 TiB	\$4.40 / hr
8x NVIDIA A100	40 GB	124	1800 GiB	6 TiB	\$8.80 / hr
1x NVIDIA RTX A6000	48 GB	14	100 GiB	200 GiB	\$0.80 / hr
2x NVIDIA RTX A6000	48 GB	28	200 GiB	1 TiB	\$1.60 / hr
4x NVIDIA RTX A6000	48 GB	56	400 GiB	1 TiB	\$3.20 / hr
1x NVIDIA A10	24 GB	30	200 GiB	3 TiB	\$0.60 / hr
1x NVIDIA Quadro RTX 6000	24 GB	14	46 GiB	512 GiB	\$0.50 / hr
8x NVIDIA Tesla V100	16 GB	92	448 GiB	5.9 TiB	\$4.40 / hr

▼ GPU mart

<p>Express GPU VPS Nvidia GeForce GT 730</p> <p>Mainly suitable for home games and Android emulators, such as BlueStacks and MEmu Play.</p> <p>Starting at</p> <p>\$21.00 /month</p> <p>Coming Soon</p> <ul style="list-style-type: none"> ✓ 3 CPU Cores ✓ 8GB RAM ✓ 120GB SSD ✓ 100Mbps Unlimited Bandwidth ✓ Once per 4 Weeks VHD Backup ✓ Supported OS: Linux & Windows 10 ✓ Dedicated GPU: GeForce GT 730 ✓ CUDA Cores: 384 ✓ GPU Memory: 2GB ✓ FP32 Performance: 692.7 GFLOPS ⓘ 	<p>Basic GPU VPS Nvidia Quadro P600</p> <p>Good choice for Android Emulators & gaming, video editing & rendering, and drawing workstations.</p> <p>Starting at</p> <p>\$29.00 /month</p> <p>Coming Soon</p> <ul style="list-style-type: none"> ✓ 6 CPU Cores ✓ 16GB RAM ✓ 200GB SSD ✓ 200Mbps Unlimited Bandwidth ✓ Once per 4 Weeks VHD Backup ✓ Supported OS: Linux & Windows 10 ✓ Dedicated GPU: Quadro P600 ✓ CUDA Cores: 384 ✓ GPU Memory: 2GB GDDR5 ✓ FP32 Performance: 1196 GFLOPS ⓘ 	<p>Lite GPU Server Nvidia GeForce GT 710</p> <p>Mainly suitable for home games and Android emulators, such as BlueStacks and MEmu Play.</p> <p>Starting at</p> <p>\$45.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 16GB RAM ✓ Quad-Core Xeon X3440 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Windows & Linux ✓ GPU: Nvidia GeForce GT 710 ✓ Microarchitecture: Kepler ✓ Max GPU: 1 ✓ CUDA Cores: 192 ✓ GPU Memory: 1GB 	<p>Lite GPU Server Nvidia GeForce GT 730</p> <p>Mainly suitable for home games and Android emulators, such as BlueStacks and MEmu Play.</p> <p>Starting at</p> <p>\$49.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 16GB RAM ✓ Quad-Core Xeon E3-1230 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Windows & Linux ✓ GPU: Nvidia GeForce GT 730 ✓ Microarchitecture: Fermi ✓ Max GPU: 1 ✓ CUDA Cores: 384 ✓ GPU Memory: 2GB
<p>Express GPU Server Nvidia Quadro P600</p> <p>Good choice for Android Emulators & gaming, video editing & rendering, and drawing workstations.</p> <p>Starting at</p> <p>\$52.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 32 GB RAM ✓ Quad-Core Xeon E5-2643 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Windows & Linux ✓ GPU: Nvidia Quadro P600 ✓ Microarchitecture: Pascal ✓ Max GPU: 1 ✓ CUDA Cores: 384 ✓ GPU Memory: 2GB GDDR5 ✓ Performance: 1.2 TFLOPS ⓘ 	<p>Express GPU Server Nvidia Quadro P620</p> <p>Good choice for Android Emulators & gaming, video editing & rendering, and drawing workstations.</p> <p>Starting at</p> <p>\$59.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 32 GB RAM ✓ Eight-Core Xeon E5-2670 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Windows & Linux ✓ GPU: Nvidia Quadro P620 ✓ Microarchitecture: Pascal ✓ Max GPU: 1 ✓ CUDA Cores: 512 ✓ GPU Memory: 2GB ✓ Performance: 1.5 TFLOPS ⓘ 	<p>Express GPU Server Nvidia Quadro P1000</p> <p>Good choice for Android Emulators & gaming, video editing & rendering, and drawing workstations.</p> <p>Starting at</p> <p>\$64.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 32GB RAM ✓ Eight-Core Xeon E5-2690 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Windows & Linux ✓ GPU: Nvidia Quadro P1000 ✓ Microarchitecture: Pascal ✓ Max GPU: 1 ✓ CUDA Cores: 640 ✓ GPU Memory: 4GB GDDR5 ✓ Performance: 1.894 TFLOPS ⓘ 	<p>Basic GPU Server Nvidia GeForce GTX 1650</p> <p>Good choice for Android Emulators & gaming, video editing & rendering, and drawing workstations.</p> <p>Starting at</p> <p>\$99.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 64GB RAM ✓ Ten-Core Xeon E5-2660v3 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Windows & Linux ✓ GPU: Nvidia GeForce GTX 1650 ✓ Microarchitecture: Turing ✓ Max GPU: 1 ✓ CUDA Cores: 896 ✓ GPU Memory: 4GB GDDR5 ✓ Performance: 3.0 TFLOPS ⓘ

<p>Basic GPU Server Nvidia Quadro T1000</p> <p>Good choice for Android Emulators & gaming, video editing & rendering, 3D modeling & drawing workstations.</p> <p>Starting at</p> <p>\$99.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 64 GB RAM ✓ Eight-Core Xeon E5-2690 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Windows & Linux ✓ GPU: Nvidia Quadro T1000 ✓ Microarchitecture: Turing ✓ Max GPU: 1 ✓ CUDA Cores: 896 ✓ GPU Memory: 8GB GDDR6 ✓ Performance: 2.5 TFLOPS ⓘ 	<p>Basic GPU Server Nvidia Tesla K40</p> <p>For high-performance computing and large data workloads, such as deep learning and AI reasoning.</p> <p>Starting at</p> <p>\$109.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 64 GB RAM ✓ Eight-Core Xeon E5-2670 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Windows & Linux ✓ GPU: Nvidia Tesla K40 ⓘ ✓ Microarchitecture: Kepler ✓ Max GPU: 2 ⓘ ✓ CUDA Cores: 2880 ✓ GPU Memory: 12GB ✓ Performance: 4.29 TFLOPS ⓘ 	<p>Professional GPU VPS Nvidia RTX A4000</p> <p>For professionals. It delivers real-time ray tracing, AI accelerated computing, and high-performance graphics to desktops.</p> <p>Starting at</p> <p>\$129.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 12 CPU Cores ✓ 32GB RAM ✓ 320GB SSD ✓ 300Mbps Unlimited Bandwidth ✓ Once per 2 Weeks VHD Backup ✓ Supported OS : Linux & Windows 10 ✓ Dedicated GPU: RTX A4000 ✓ CUDA Cores: 6144 ✓ Tensor Cores: 192 ✓ GPU Memory: 16GB GDDR6 ✓ FP64 Performance: 19.2 TFLOPS ⓘ 	<p>Basic GPU Server Nvidia GeForce GTX 1660</p> <p>Good choice for Android Emulators & gaming, video editing & rendering, 3D modeling & drawing workstations.</p> <p>Starting at</p> <p>\$139.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 64GB RAM ✓ Dual 10-Core Xeon E5-2660v2 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Linux & Windows 10 ✓ GPU: Nvidia GeForce GTX 1660 ✓ Microarchitecture: Turing ✓ Max GPU: 1 ✓ CUDA Cores: 1408 ✓ GPU Memory: 6GB GDDR6 ✓ Performance: 5.0 TFLOPS ⓘ
<p>Professional GPU Server Nvidia Tesla K80</p> <p>For high-performance computing and large data workloads, such as deep learning and AI reasoning.</p> <p>Starting at</p> <p>\$159.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 128 GB RAM ✓ Dual 10-Core E5-2660v2 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Linux & Windows 10 ✓ GPU: Nvidia Tesla K80 ⓘ ✓ Microarchitecture: Kepler ✓ Max GPU: 2 ⓘ ✓ CUDA Cores: 4992 ✓ GPU Memory: 24GB ✓ Performance: 8.73 TFLOPS ⓘ 	<p>Professional GPU Server Nvidia GeForce RTX 2060</p> <p>Achieve an excellent balance between function, performance, and reliability. Assist designers, engineers, and artists to realize their visions.</p> <p>Starting at</p> <p>\$159.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 128 GB RAM ✓ Dual 10-Core E5-2660v2 ⓘ ✓ 120GB SSD + 960GB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Linux & Windows 10 ✓ GPU: Nvidia GeForce RTX 2060 ⓘ ✓ Microarchitecture: Turing ✓ Max GPU: 2 ⓘ ✓ CUDA Cores: 1920 ✓ GPU Memory: 6GB GDDR6 ✓ Performance: 6.5 TFLOPS ⓘ 	<p>Advanced GPU Server Nvidia GeForce RTX 3060 Ti</p> <p>For professionals. It delivers real-time ray tracing, AI accelerated computing, and high-performance graphics to desktops.</p> <p>Starting at</p> <p>\$209.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 128 GB RAM ✓ Dual 12-Core E5-2697v2 ⓘ ✓ 240GB SSD + 2TB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Linux & Windows 10 ✓ GPU: Nvidia GeForce RTX 3060 Ti ⓘ ✓ Microarchitecture: Ampere ✓ Max GPU: 2 ⓘ ✓ CUDA Cores: 4864 ✓ GPU Memory: 8GB GDDR6 ✓ Performance: 16.2 TFLOPS ⓘ 	<p>Advanced GPU Server Nvidia RTX A4000</p> <p>For professionals. It delivers real-time ray tracing, AI accelerated computing, and high-performance graphics to desktops.</p> <p>Starting at</p> <p>\$209.00 /month</p> <p>Order Now</p> <ul style="list-style-type: none"> ✓ 128 GB RAM ✓ Dual 12-Core E5-2697v2 ⓘ ✓ 240GB SSD + 2TB SSD ✓ 100Mbps-1Gbps Bandwidth ⓘ ✓ Supported OS: Linux & Windows 10 ✓ GPU: Nvidia RTX A4000 ⓘ ✓ Microarchitecture: Ampere ✓ Max GPU: 2 ⓘ ✓ CUDA Cores: 6144 ✓ Tensor Cores: 192 ✓ GPU Memory: 16GB GDDR6 ✓ Performance: 19.2 TFLOPS ⓘ

Advanced GPU Server

Nvidia RTX A5000

Achieve an excellent balance between function, performance, and reliability. Assist designers, engineers, and artists to realize their visions.

Starting at

\$269.00

/month

Order Now

- ✓ 128GB RAM
- ✓ Dual 12-Core E5-2697v2 ⓘ
- ✓ 240GB SSD + 2TB SSD
- ✓ 100Mbps-1Gbps Bandwidth ⓘ
- ✓ Supported OS: Linux & Windows 10
- ✓ GPU: Nvidia RTX A5000 ⓘ
- ✓ Microarchitecture: Ampere
- ✓ Max GPU: 2 ⓘ
- ✓ CUDA Cores: 8192
- ✓ GPU Memory: 24GB GDDR6
- ✓ Performance: 27.8 TFLOPS ⓘ

Enterprise GPU Server

Nvidia GeForce RTX 4090

Achieve an excellent balance between function, performance, and reliability. Assist designers, engineers, and artists to realize their visions.

Starting at

\$369.00

/month

Coming Soon

- ✓ 256 GB RAM
- ✓ Dual E5-2697v4 ⓘ
- ✓ 240GB SSD + 2TB SSD + 2TB NVMe
- ✓ 100Mbps-1Gbps Bandwidth ⓘ
- ✓ Supported OS : Linux & Windows 10
- ✓ GPU: GeForce RTX 4090 ⓘ
- ✓ Microarchitecture: Ada Lovelace
- ✓ Max GPU: 2 ⓘ
- ✓ CUDA Cores: 16,384
- ✓ Tensor Cores:
- ✓ GPU Memory: 24GB GDDR6X
- ✓ Performance: 82.6 TFLOPS ⓘ

Advanced GPU Server

Nvidia Quadro RTX A6000

High Performance for video editing & rendering, Deep Learning, and Live streaming.

Starting at

\$409.00

/month

Coming Soon

- ✓ 256 GB RAM
- ✓ Dual E5-2697v4 ⓘ
- ✓ 240GB SSD + 2TB SSD + 2TB NVMe
- ✓ 100Mbps-1Gbps Bandwidth ⓘ
- ✓ Supported OS: Linux & Windows 10
- ✓ GPU: Nvidia RTX A6000 ⓘ
- ✓ Microarchitecture: Ampere
- ✓ Max GPU: 2 ⓘ
- ✓ CUDA Cores: 10,752
- ✓ Tensor Cores: 336
- ✓ GPU Memory: 48GB
- ✓ Performance: 38.71 TFLOPS ⓘ

Enterprise GPU Server

Nvidia A40

High performance for deep learning, video editing & rendering, and streaming.

Starting at

\$439.00

/month

Coming Soon

- ✓ 256 GB RAM
- ✓ Dual E5-2697v4 ⓘ
- ✓ 240GB SSD + 2TB SSD + 2TB NVMe
- ✓ 100Mbps-1Gbps Bandwidth ⓘ
- ✓ Supported OS: Linux & Windows 10
- ✓ GPU: Nvidia A40 ⓘ
- ✓ Microarchitecture: Ampere
- ✓ Max GPU: 2 ⓘ
- ✓ CUDA Cores: 10,752
- ✓ Tensor Cores: 336
- ✓ GPU Memory: 48GB
- ✓ Performance: 37.4 TFLOPS ⓘ

▼ Genesis Cloud

Instance Type	GPUs	vCPUs	Memory	Disk (SSD)	On-demand price (per hour)	Monthly long-term (per month)	Yearly long-term (per month)
NVIDIA® GeForce™ RTX 3090	1	4 vCPU	24 GiB	80 GiB	\$1.30	\$759.20	\$474.50
NVIDIA® GeForce™ RTX 3090	2	8 vCPU	48 GiB	80 GiB	\$2.60	\$1518.40	\$949.00
NVIDIA® GeForce™ RTX 3090	3	12 vCPU	72 GiB	80 GiB	\$3.90	\$2277.60	\$1423.50
NVIDIA® GeForce™ RTX 3090	4	16 vCPU	96 GiB	80 GiB	\$5.20	\$3036.80	\$1898.00
NVIDIA® GeForce™ RTX 3090	5	20 vCPU	120 GiB	80 GiB	\$6.50	\$3796.00	\$2372.50
NVIDIA® GeForce™ RTX 3090	6	24 vCPU	144 GiB	80 GiB	\$7.80	\$4555.20	\$2847.00
NVIDIA® GeForce™ RTX 3090	7	28 vCPU	168 GiB	80 GiB	\$9.10	\$5314.40	\$3321.50
NVIDIA® GeForce™ RTX 3090	8	32 vCPU	192 GiB	80 GiB	\$10.40	\$6073.60	\$3796.00

GPU Price Comparison Jan 22

25

Instance Type	GPUs	vCPUs	Memory	Disk (SSD)	On-demand price (per hour)	Monthly long-term (per month)	Yearly long-term (per month)
NVIDIA® GeForce™ RTX 3080	1	4 vCPU	12 GiB	80 GiB	\$0.90	\$525.60	\$328.50
NVIDIA® GeForce™ RTX 3080	2	8 vCPU	24 GiB	80 GiB	\$1.80	\$1051.20	\$657.00
NVIDIA® GeForce™ RTX 3080	3	12 vCPU	36 GiB	80 GiB	\$2.70	\$1576.80	\$985.50
NVIDIA® GeForce™ RTX 3080	4	16 vCPU	48 GiB	80 GiB	\$3.60	\$2102.40	\$1314.00
NVIDIA® GeForce™ RTX 3080	5	20 vCPU	60 GiB	80 GiB	\$4.50	\$2628.00	\$1642.50
NVIDIA® GeForce™ RTX 3080	6	24 vCPU	72 GiB	80 GiB	\$5.40	\$3153.60	\$1971.00
NVIDIA® GeForce™ RTX 3080	7	28 vCPU	84 GiB	80 GiB	\$6.30	\$3679.20	\$2299.50
NVIDIA® GeForce™ RTX 3080	8	32 vCPU	96 GiB	80 GiB	\$7.20	\$4204.80	\$2628.00

Instance Type	GPUs	vCPUs	Memory	Disk (SSD)	On-demand price (per hour)	Monthly long-term (per month)	Yearly long-term (per month)
NVIDIA® GeForce™ RTX 3060 Ti	1	4 vCPU	12 GiB	80 GiB	\$0.65	\$379.60	\$237.25
NVIDIA® GeForce™ RTX 3060 Ti	2	8 vCPU	24 GiB	80 GiB	\$1.30	\$759.20	\$474.50
NVIDIA® GeForce™ RTX 3060 Ti	3	12 vCPU	36 GiB	80 GiB	\$1.95	\$1138.80	\$711.75
NVIDIA® GeForce™ RTX 3060 Ti	4	16 vCPU	48 GiB	80 GiB	\$2.60	\$1518.40	\$949.00
NVIDIA® GeForce™ RTX 3060 Ti	5	20 vCPU	60 GiB	80 GiB	\$3.25	\$1898.00	\$1186.25
NVIDIA® GeForce™ RTX 3060 Ti	6	24 vCPU	72 GiB	80 GiB	\$3.90	\$2277.60	\$1423.50
NVIDIA® GeForce™ RTX 3060 Ti	7	28 vCPU	84 GiB	80 GiB	\$4.55	\$2657.20	\$1660.75
NVIDIA® GeForce™ RTX 3060 Ti	8	32 vCPU	96 GiB	80 GiB	\$5.20	\$3036.80	\$1898.00