



AWS Pricing Models and the Bill

AWS pricing pages, Billing and Cost Management user guide, as of 2026 | Pricing dimensions, On-Demand, Savings Plans, Reserved Instances, Spot, capacity reservations, Dedicated Hosts, free tier, data transfer, consolidated billing, how to read a bill

FinOps Pricing Savings Plans Spot CLF SAA

How AWS charges

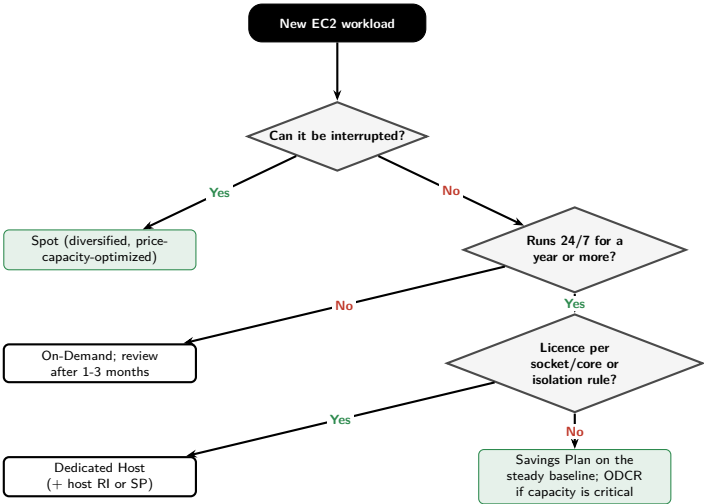
Dimension	Examples	Unit
Compute time	EC2, Fargate (vCPU + GB), Lambda (GB-seconds), RDS instance hours	second / hour
Storage	S3, EBS, snapshots, backups, log storage	GB-month
Requests and operations	S3 PUT/GET, API Gateway calls, DynamoDB on-demand RRU/WRU, KMS calls	per 1,000 / million
Data transfer and processing	internet egress, cross-AZ, cross-Region, NAT gateway and TGW processing	GB
Provisioned capacity	provisioned IOPS, DynamoDB RCU/WCU, Kinesis shards, MSK brokers	per hour
Fixed fees	public IPv4 (0.005 USD/h each), load balancer hours, endpoint hours, support plans	hour / month

	On-Demand	Savings Plans	Reserved Instances	Spot
Commitment	✗	USD/hour, 1 or 3 years	instance config, 1 or 3 years	✗
Discount vs On-Demand (max, as of 2026)	0 %	Compute ~66 %, EC2 Instance ~72 %, SageMaker AI ~64 %	Standard ~72 %, Convertible ~66 %	up to 90 %
Flexibility	full	high (Compute SP: any Region, family, Fargate, Lambda)	low to medium	interruptible
Capacity guarantee	no (use ODCR)	✗	zonal RI only	✗
Best for	spiky, new, short-lived	steady baseline of compute	RDS, ElastiCache, OpenSearch, Redshift baselines	fault-tolerant, stateless, batch, CI

INFO

Payment options for Savings Plans and RIs: **No Upfront, Partial Upfront, All Upfront** (deepest discount). Discounts are percentages from the pricing pages and differ by service, Region and instance; never assume the maximum. Depth on commitments: next sheet.

Purchase options beyond the discount



Option	What you get	Watch out
On-Demand Capacity Reservation	guaranteed capacity in one AZ, any duration	billed at On-Demand rate even when idle; apply SP/regional RI for the discount
Capacity Blocks for ML	GPU instances (P5, Trn) reserved for 1-182 days at a future date	price per block, paid upfront
Dedicated Instances	hardware not shared with other accounts	per-Region fee + higher hourly rate
Dedicated Hosts	a whole physical server, socket/core visibility	BYOL Windows Server, SQL Server, Oracle
Spot	2-minute notice, rebalance recommendation earlier	Spot price changes slowly; no bidding needed (max price optional)
Free tier credits (2025+)	up to 200 USD credits for new accounts	free plan closes after 6 months or when credits end

Data transfer basics

Path	Charge (as of 2026)	Design tip
Internet → AWS	free	ingest freely
AWS → internet	per GB, tiered; first 100 GB/month free (all services combined)	serve through CloudFront: AWS origin → CloudFront is free
Between AZs (same Region)	per GB in each direction (approx. 0.01 USD/GB each way)	keep chatty services AZ-local; NLB cross-zone off saves it
Between Regions	per GB out of the source Region	replicate only what DR needs
Same AZ, private IP	free	use private IPs, not public/Elastic IPs, inside a Region
NAT gateway	per hour + per GB processed	gateway endpoints for S3/DynamoDB are free
Transit Gateway, PrivateLink	per attachment/endpoint hour + per GB	peering has no hourly fee
Direct Connect	port hours + lower per-GB egress rate	cheaper than internet egress at high volume

Read the bill

Estimate one small web app's monthly bill

PROBLEM

Estimate the monthly cost of a web app: 4 instances 24/7, 500 GB of gp3, 2,000 GB served to the internet, one NAT gateway processing 1,000 GB. Illustrative prices, check the pricing pages.

GIVEN

- Instance approx. 0.10 USD/h, 730 h/month
- gp3 approx. 0.08 USD per GB-month
- Internet egress approx. 0.09 USD/GB after 100 GB free
- NAT approx. 0.045 USD/h and 0.045 USD/GB

1

Compute

$$4 \times 730 \times 0.10 = 292 \text{ USD}$$

Why: *instance-hours x rate*

2

Storage

$$500 \times 0.08 = 40 \text{ USD}$$

Why: *GB-month x rate*

3

Egress

$$(2,000 - 100) \times 0.09 = 171 \text{ USD}$$

Why: *first 100 GB free*

4

NAT

$$730 \times 0.045 + 1,000 \times 0.045 = 32.85 + 45 = 77.85 \text{ USD}$$

Why: *hourly + per GB processed*

5

Total

$$292 + 40 + 171 + 77.85 = 580.85 \text{ USD}$$

Why: *before tax, support and credits*

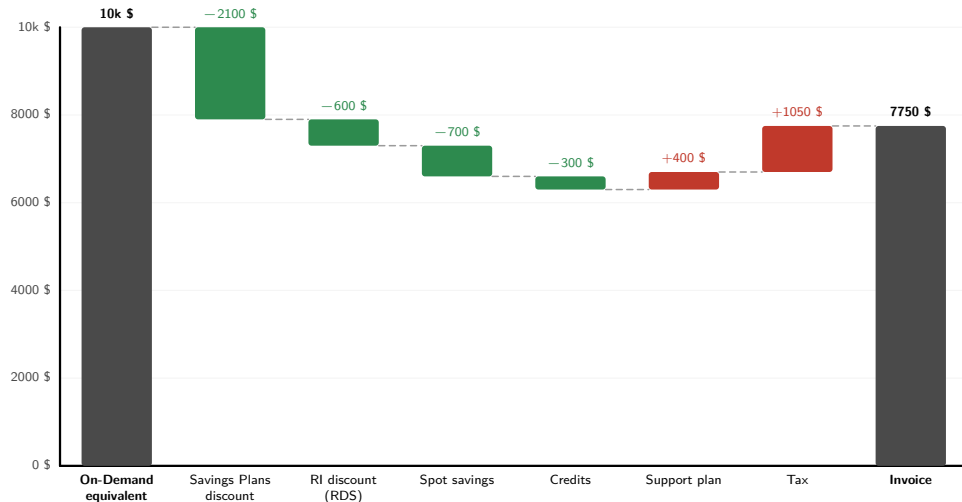
ANSWER

About 581 USD/month; egress and NAT are 43 % of it, not only the instances.

CHECK

- ✓ A Compute Savings Plan covering the instances (say ~30 % off) saves about 88 USD; CloudFront and an S3 gateway endpoint attack the other 43 %.

From list price to net bill (illustrative, USD)



Effective savings rate (ESR)

FORMULA $ESR = (ODE - C) / ODE$

WHERE

- ESR = share of the On-Demand cost avoided by all rate optimisations (0-1)
- ODE = On-Demand equivalent cost of the usage (what it would cost with no commitments or Spot) (USD per month)
- C = actual amortised cost of the same usage (commitment fees + On-Demand + Spot) (USD per month)

WORKED

ODE = 10 000
C = 6 600
 $ESR = (10\ 000 - 6\ 600) / 10\ 000$
 $ESR = 0.34$ (34 %)

RESULT ESR = 0.34 (34 %)

READ IT 34 % of what On-Demand would have cost is saved; track it monthly as the headline rate-optimisation KPI.

MISTAKES

- Using unblended cost: upfront payments distort the month; use amortised cost
- Credits and refunds inflate ESR: exclude them
- A high ESR with low utilisation hides waste: track commitment utilisation too

Recap and self-test

YOU MAY HEAR
"Reserved Instances reserve capacity"

BUSTED

ACTUALLY
Only zonal RIs (and ODCRs) reserve capacity
Regional RIs and Savings Plans are pure billing discounts.

YOU MAY HEAR
"Spot needs a bidding strategy"

BUSTED

ACTUALLY
You pay the current Spot price; the max price is optional
Interruptions come from capacity, not from being outbid.

YOU MAY HEAR
"Data transfer in AWS is free"

BUSTED

ACTUALLY
Only inbound and same-AZ private traffic are free
Cross-AZ, cross-Region, NAT and internet egress are billed per GB.

YOU MAY HEAR
"Savings Plans cover RDS"

BUSTED

ACTUALLY
Compute SP covers EC2, Fargate and Lambda
RDS, ElastiCache, OpenSearch, Redshift use RIs (and Database Savings Plans announced late 2025, check coverage).

Key takeaways

01

Every bill = compute + storage + requests + transfer + fixed fees
look beyond instances

02

On-Demand → Savings Plans → Spot
by predictability and interruption tolerance

03

Capacity and discount are separate
ODCR/zonal RI for capacity, SP for price

04

Data transfer is a design choice
AZ locality, endpoints, CloudFront

05

Measure ESR on amortised cost
the rate-optimisation KPI

Q1 Which Savings Plan also covers Fargate and Lambda?

A Compute Savings Plan.

Q2 Maximum Spot discount and interruption notice?

A Up to 90 % off On-Demand; 2-minute notice (plus earlier rebalance recommendations).

Q3 Which RI type can be sold on the RI Marketplace?

A Standard RIs (not Convertible).

Q4 Need guaranteed capacity in one AZ for a 3-day event?

A On-Demand Capacity Reservation (any duration, cancel anytime).

Q5 Is data from EC2 to S3 in the same Region charged?

A No transfer charge (requests and storage still bill); via a NAT gateway it pays NAT processing.

Q6 How is the first 100 GB of internet egress treated?

A Free each month, aggregated across services and Regions (except China and GovCloud).

Q7 What does consolidated billing add to discounts?

A Aggregated volume tiers and RI/SP sharing across member accounts (sharing can be turned off per account).

Q8 BYOL Windows Server licensed per physical core?

A Dedicated Hosts.

Q9 EC2 Linux billing granularity?

A Per second with a 60-second minimum.