

AWS COST · SAMPLE DELIVERABLE

Rightsizing and Commitments

Week 4 of 6: Rightsizing and commitments · prepared for
Seatlark, Inc.

Prepared by Charles Sieg · charles.sieg@vantalect.com · vantalect.com

June 5, 2026 · AWS Cost Analysis & Optimization: The Optimization (six weeks)

SAMPLE DELIVERABLE

This is a sample deliverable. Seatlark, Inc. is a fictional company: every account, resource, finding, figure, and person in this report is simulated to show the form and depth of a Vantalect engagement. AWS prices are list prices from the AWS Price List API retrieved on October 13 and 16, 2026, and AWS Business Support+ pricing is from the AWS Support pricing page on October 16, 2026.

Contents

Summary	2
What went in	5
Production rightsizing on June 2, 2026	6
The June 5 on-sale was the test	9
The ticket-scan lifecycle rule	10
The CloudWatch agent	11
The commitment decision memo	12
The numbers	12
Why \$5.00 and not more	13
What was declined, and why	15
The risk, plainly	15
Approval and purchase	16
What was deferred, and why	17
Honest limits, as of June 5, 2026	18
What comes next	19

Summary

In week 4 Seatlark put **\$56,316 a year** of savings in place, in 6 changes: the production API tier and the partner APIs rightsized before the freeze, the staging and partner non-production hosts, the ticket-scan lifecycle rule, the CloudWatch agent change, and a one-year Compute Savings Plan. Added to week 3's \$98,760, that makes 20 changes and **\$155,076 a year in place, estimated**: 1.48 times the \$105,000 gate at the estimates. The gate is a count of measured savings, and nothing is measured until week 6. Every figure here is the Analysis's price for the change at list, annualized.

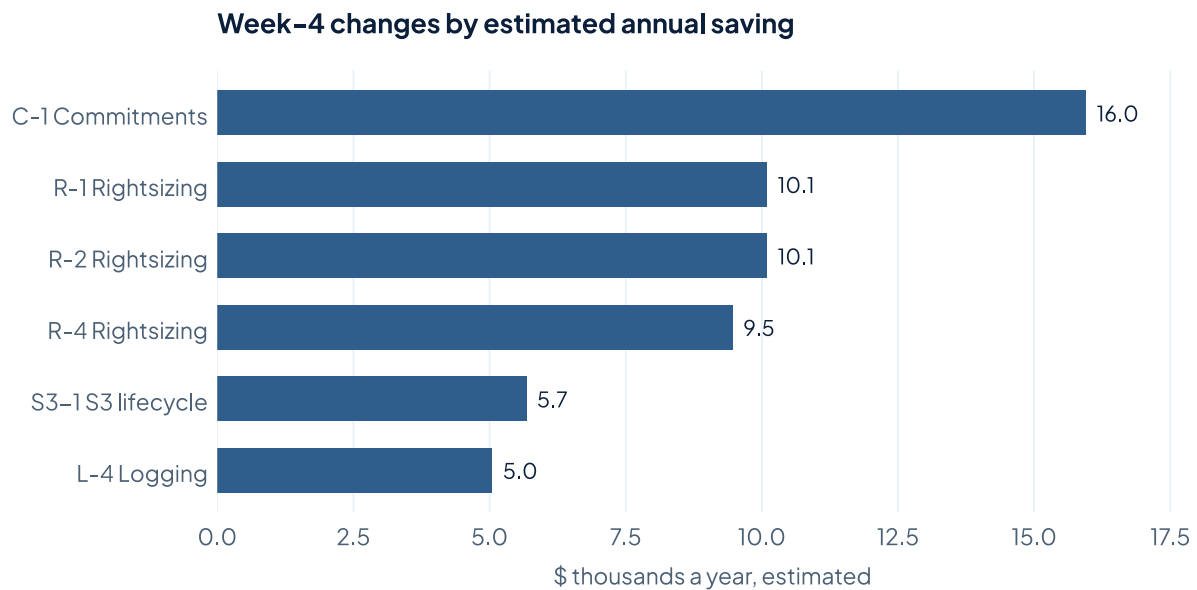
Production rightsizing went in on June 2, 2026, before the freeze. The marketplace API tier (6 hosts) and the partner APIs (6 hosts) moved from m5.2xlarge to m5.xlarge, one host at a time behind the load balancer. The Friday, June 5, 2026 on-sale was the live test: the API tier peaked at **61% CPU on the smaller size**, three points above the 58% we predicted from April's 29%, and 1 point over the 60% ceiling the Analysis set for a production rightsizing ([The June 5 on-sale was the test](#)). By our own rule that is over the line, not a pass: the tier stays on the smaller size with an alarm at 75% CPU, and the second on-sale, on June 10, 2026, decides whether it stays.

The commitment decision was made on Wednesday, June 3, 2026. Lydia Brandt, Vice President, Finance, with Renee Castellano, approved a **one-year No Upfront Compute Savings Plan at \$5.00 an hour**, and declined a three-year plan and All Upfront for now: "we don't know what checkout runs on in a year, and I'd rather keep the cash." Seatlark's finance team bought it in the management account that afternoon. It covers 58.6% of the \$11.63 an hour the always-on floor costs On-Demand, saves an estimated \$15,950 a year, and stays fully used after every deferred change. The memo is in [The commitment decision memo](#).

Two production changes stay deferred, as decided at the readout, with dates and owners: **D-2**, the partner database, to Saturday, June 27, 2026, because it waits on a load test against a restored copy at the new class and then on the 10 business days' notice of a maintenance window that Seatlark Connect's partner contracts require (the notice goes out when the test passes, by June 11, 2026; a notice sent at the readout would have put the first window on June 8, 2026, inside the freeze), and **E-2b**, the production search volumes, to June 22, 2026 because Imani Brooks will not touch the search cluster in the week before the biggest on-sale of the summer. Together they are \$20,094 a year and count for nothing toward the gate ([What was deferred, and why](#)).

Week 4 in numbers. Savings are estimates until the week-6 measurement.

Week 4 at a glance	Result
Changes put in place	6
Resources changed, not counting the Savings Plan	19
Annualized savings, estimated	\$56,316
Of which the Savings Plan	\$15,950
In place after weeks 3 and 4, estimated	\$155,076
Made by pull request	5
Hourly commitment bought	\$5.00
Committed over the one-year term	\$43,800
Moved to deferred this week	\$20,094 a year



The week-4 changes, estimated and annualized.

Decision for the sponsor

Renee Castellano: no decision is needed to keep this week's savings. The freeze runs from June 4, 2026 to June 14, 2026, and no production change goes in during it. Two dates need an owner each and are already set: Imani Brooks on June 22, 2026 for the search volumes, and Diego Saldana on June 27, 2026 for the partner database, with a load test against a restored copy first and the partner notice going out when it passes, by June 11, 2026.

What went in

The week-4 changes in date order. The Savings Plan row counts the 30 floor hosts it covers, so its resources are left out of the total.

ID	Change	Owner	Channel	Date	Resources	Monthly	Annualized
L-4	CloudWatch agent: per-mount and per-process metrics dropped (2,400 custom metrics to 1,000)	Beth Morrow	seatlark-landing-zone PR #418	Jun 1	1	\$421	\$5,047
R-4	Non-production: staging search nodes to r5.xlarge, partner non-production to m5.xlarge	Beth Morrow	marketplace-infra PR #1812	Jun 1	5	\$788	\$9,461
S3-1	Ticket-scan uploads to S3 Glacier Instant Retrieval after 120 days (26 TB)	Kara Lindqvist	data-content-infra PR #234	Jun 1	1	\$473	\$5,674
R-1	Marketplace API tier: 6 of 12 m5.2xlarge to m5.xlarge; the web tier keeps its on-sale headroom	Aaron Feld	marketplace-infra PR #1813	Jun 2	6	\$841	\$10,092
R-2	Partner APIs: 6 x m5.2xlarge to m5.xlarge	Diego Saldana	partner-infra PR #641	Jun 2	6	\$841	\$10,092
C-1	Compute Savings Plan, one year, No Upfront, \$5.00 an hour	Lydia Brandt	purchase approved by the VP of Finance, bought in the management account	Jun 3	30	\$1,329	\$15,950
Total					19	\$4,693	\$56,316

5 of the 6 are rated High confidence. **S3-1**, the ticket-scan lifecycle rule, is Medium: its saving depends on how many objects are small and how often old ones are read ([The ticket-scan lifecycle rule](#)).

How the roadmap reconciles. Deferred and declined changes count for nothing toward the gate.

Where the Analysis's \$246,839 stands on Jun 5	Changes	Annualized, estimated
In place: week 3	14	\$98,760
In place: week 4	6	\$56,316
In place, counted when measured	20	\$155,076
Approved and deferred, with dates	6	\$52,808
Declined at the readout	3	\$35,846
Approved at the readout, not yet run (S3-3)	1	\$3,109
Identified by the Analysis		\$246,839

Production rightsizing on June 2, 2026

The Analysis recommended resizing the marketplace API tier and the partner APIs. Both ran at 9% CPU at the 95th percentile, on hosts that MGN launched at the size of their VMware allocation. The evidence is 93 days of CloudWatch agent metrics (the agent was installed in each wave, so the data exists), not Compute Optimizer alone: its default 14-day window never saw April's on-sale, which is why R-3, the ingestion workers, stays rated Medium and declined.

What was resized in week 4. R-1 and R-2 are production; R-4 is staging search and partner non-production.

ID	Tier	HostsFrom	To	vCPU and memory	Per host-hour	Monthly, before to after	CPU at 95th percentile
R-1	mkt-api in marketplace-prod	6m5.2xlarge	m5.xlarge	8 and 32 GiB to 4 and 16 GiB	\$0.384 to \$0.192	\$1,682 to \$841	9%
R-2	ptr-api in partner-prod	6m5.2xlarge	m5.xlarge	8 and 32 GiB to 4 and 16 GiB	\$0.384 to \$0.192	\$1,682 to \$841	9%
R-4	stg-search in marketplace-staging	2r5.2xlarge	r5.xlarge	8 and 64 GiB to 4 and 32 GiB	\$0.504 to \$0.252	\$736 to \$368	14%
R-4	ptr-np in partner-nonprod	3m5.2xlarge	m5.xlarge	8 and 32 GiB to 4 and 16 GiB	\$0.384 to \$0.192	\$841 to \$420	7%

How it was done. Each tier is a fixed-size Auto Scaling group behind its load balancer. The pull request changes the instance type in the launch template, and CloudFormation's rolling update replaces one host at a time, keeping all but one in service. Nobody replaces a host by hand. The owners approved the pull requests (marketplace-infra PR #1813 for Aaron Feld, partner-infra PR #641 for Diego Saldana) and the change went in on Tuesday, June 2, 2026, two days before the freeze.

YAML

```
# marketplace-infra/stacks/api-tier.yaml (excerpt)
Parameters:
  ApiInstanceType:
    Type: String
    Default: m5.xlarge          # was m5.2xlarge
Resources:
  ApiLaunchTemplate:
    Type: AWS::EC2::LaunchTemplate
    Properties:
      LaunchTemplateData:
        ImageId: !Ref ApiImageId
        InstanceType: !Ref ApiInstanceType
  ApiGroup:
    Type: AWS::AutoScaling::AutoScalingGroup
    UpdatePolicy:
      AutoScalingRollingUpdate:
        MaxBatchSize: 1
        MinInstancesInService: 5
        PauseTime: PT10M
    Properties:
      MinSize: "6"
      MaxSize: "6"
      DesiredCapacity: "6"
      LaunchTemplate:
        LaunchTemplateId: !Ref ApiLaunchTemplate
        Version: !GetAtt ApiLaunchTemplate.LatestVersionNumber
      TargetGroupARNs: [!Ref ApiTargetGroup]
      HealthCheckType: ELB
      HealthCheckGracePeriod: 300
```

How it is undone. The previous launch template version is still there, and nothing was lost or migrated, so the rollback is the same rolling update in the other direction: revert the pull request, or pass the old size to the stack. Each batch is one host, so a bad first host is found before the second is touched.

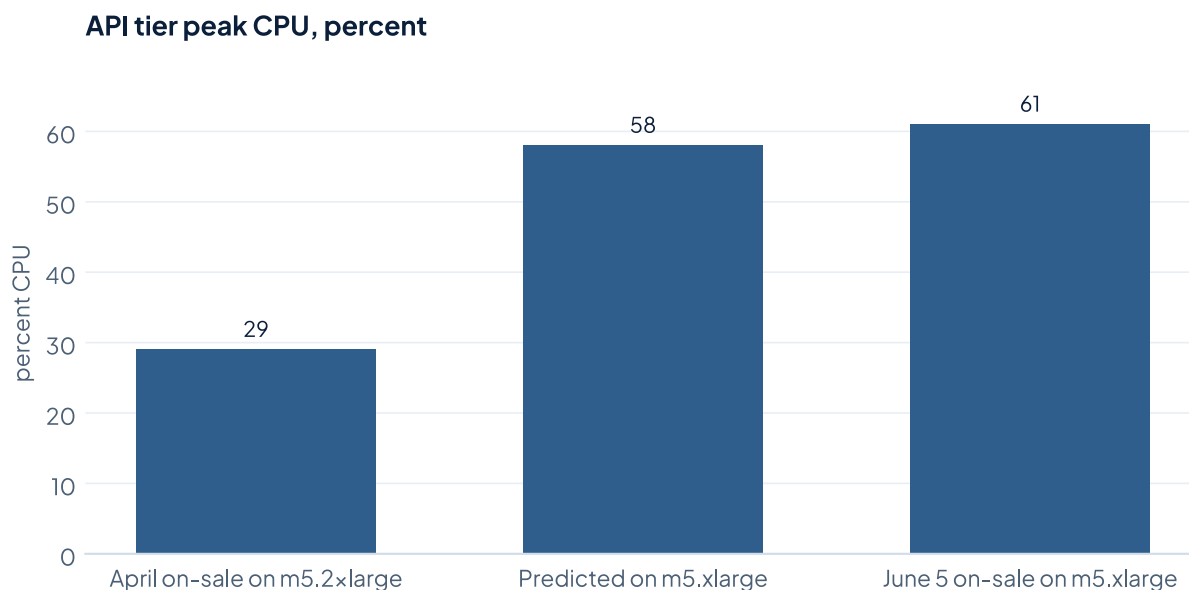
BASH

```
aws cloudformation update-stack --stack-name marketplace-api-tier --use-previous-template \
  --parameters ParameterKey=ApiInstanceType,ParameterValue=m5.2xlarge
ParameterKey=ApiImageId,UsePreviousValue=true
```

The web tier keeps its size, on purpose. The 6 web hosts (m5.2xlarge) peak at 41% CPU in an on-sale and have no autoscaling yet. Halving them would put that peak near 82%, which is not headroom for traffic that jumps 8 to 15 times in the first 10 minutes of a big on-sale. It is the first change to revisit when autoscaling exists, and G-1 is the more likely one to touch it.

The June 5 on-sale was the test

A rightsizing estimate is a prediction until a real peak runs through it. Halving a host's vCPUs, from 8 to 4, should double the CPU percentage for the same work (that is the assumption, and the memory and network are not modeled). April's on-sale peaked at 29% on the larger size, so the smaller size should peak at 58%.



The API tier's peak CPU: April on the old size, the prediction for the new size, and the Friday, June 5, 2026 on-sale.

On Friday, June 5, 2026, the API tier peaked at **61%**: 3 points above the prediction, which is a 5% bigger peak than April's after allowing for the size, and 39 points of headroom left. The prediction did not hold: 61% is 1 point over the 60% line the Analysis set, so by the rule we wrote, this size would not have been recommended on that peak. We did not roll back, because 39 points of headroom remain and the way back is cheap: one host at a time behind the load balancer, in minutes. Aaron Feld added a CloudWatch alarm on the API tier at 75% CPU, and a sustained breach rolls the tier back to m5.2xlarge. The next test is Wednesday, June 10, 2026, inside the freeze; the week-5 report records the result. The partner APIs have not yet run their busiest feed on the smaller size: April's busiest was 21% on the larger size, so we expect 42%, and the week-5 report gives what it was.

BASH

```
aws cloudwatch get-metric-statistics --namespace AWS/EC2 --metric-name CPUUtilization \
--dimensions Name=AutoScalingGroupName,Value=marketplace-api --statistics Maximum --period
300 \
--start-time <on-sale-start> --end-time <on-sale-end> --query 'max(Datapoints[].Maximum)'
```

The ticket-scan lifecycle rule

S3-1 moves ticket-scan uploads older than 120 days to S3 Glacier Instant Retrieval. Kara Lindqvist approved it: sellers upload ticket PDFs, barcode images, and transfer proofs, read them heavily until the event and through the 120-day chargeback window, and almost never afterward. The bucket holds 40 TB in 238,615,385 objects, and 26 TB of it (65%) is older than the rule. Glacier Instant Retrieval stores a gigabyte for \$0.004 a month against \$0.022 in S3 Standard, the top tier of your combined volume, which is where bytes that leave Standard come off, and still serves a read in milliseconds, which a dispute needs.

The ticket-scan rule in numbers. The retrieval charge is not netted into the estimate.

S3-1	Figure
Objects that move	155,100,000
One-time transition charge (\$0.02 per 1,000 objects)	\$3,102
Monthly saving, estimated	\$473
Payback of the transition charge	6.6 months
Allowance for objects billed at the 128 KB minimum	6%
Read from objects past 90 days, a month	40 GB
Retrieval charge on those reads (\$0.03 a GB)	\$1.20 a month

The one-time cost is real and is reported as one: \$3,102 for 155,100,000 transitions, repaid in 6.6 months by this rule's own saving. It is never annualized, and the week-6 gate is shown net of it. The 6% allowance is for objects under 128 KB, which Glacier Instant Retrieval bills as 128 KB; S3's default lifecycle behavior leaves objects that small in Standard, so the transition count above is a ceiling (it treats every aged object as moving) and the allowance stands in for the small ones; the week-6 measurement shows how they net out. The average object is 175 KB.

Why not Intelligent-Tiering. Intelligent-Tiering moves objects by access, and for data nobody reads it ends in its Archive Instant Access tier at the same \$0.004 a GB. But it charges a monitoring fee of \$0.0025 per 1,000 objects a month, and on the 155,100,000 objects that

move that is **\$388 a month, 82% of S3-1's saving**, before any storage saving. What it buys is no retrieval charge, and retrieval here costs \$1.20 a month. The monitoring fee would be repaid only by about 12,925 GB of reads a month, against 40. Intelligent-Tiering was considered and rejected.

YAML

```
# data-content-infra/stacks/ticket-scans.yaml (excerpt)
LifecycleConfiguration:
  Rules:
    - Id: scans-to-glacier-ir
      Status: Enabled
      Transitions:
        - StorageClass: GLACIER_IR
          TransitionInDays: 120
```

The rule applies to objects already past 120 days at its next daily run, and S3 moves them asynchronously, so the first days can still bill Standard for objects waiting their turn. **The way back is not a lifecycle rule.** Lifecycle transitions run one way, down; to bring objects back to Standard you copy them in place with S3 Batch Operations, which pays the retrieval charge and request charges again. Objects in Glacier Instant Retrieval bill a 90-day minimum, so moving one back early bills the rest of it.

The CloudWatch agent

L-4 cut the CloudWatch agent's custom metrics from 2,400 to 1,000: 1,400 per-mount and per-process series that nobody alarms on. Each custom metric costs \$0.30 a month, so the change saves about \$421 a month. One shared configuration does it. Ansible installs it (`ansible.builtin.copy to /opt/aws/amazon-cloudwatch-agent/etc/amazon-cloudwatch-agent.json`, then `amazon-cloudwatch-agent-ctl -a fetch-config -m ec2 -s -c file:... as a handler`):

JSON

```
{
  "agent": { "metrics_collection_interval": 60 },
  "metrics": {
    "append_dimensions": { "InstanceId": "${aws:InstanceId}" },
    "aggregation_dimensions": [ ["InstanceId"] ],
    "metrics_collected": {
      "mem": { "measurement": ["mem_used_percent"] },
      "disk": { "measurement": ["used_percent"], "resources": ["/"], "drop_device": true }
    }
  }
}
```

To roll back, restore the previous version of the file and run the handler again. The old series stop being published at once and age out of CloudWatch on its own schedule.

The commitment decision memo

The decision

To: Lydia Brandt, Vice President, Finance. **Copy:** Renee Castellano, Owen Pradhan, Beth Morrow.

Decided Wednesday, June 3, 2026: buy a one-year Compute Savings Plan, No Upfront, at **\$5.00 an hour**, a commitment of \$43,800 over the year. Do not buy a three-year plan or pay All Upfront yet. Seatlark’s finance team bought the plan in the management account (s1-management) that afternoon.

The numbers

Seatlark has held no commitments of any kind. The team deferred every one "until the migration completes," because the estate keeps changing. That is half right: a three-year plan or a reserved instance sized on today's estate would strand money as the remaining waves change it. But a floor of hosts runs every hour of every day and will still be there: the marketplace web and API tiers, the search cluster, ingestion, the partner APIs, the CI runners, and the shared services hosts. After the changes in place, those 30 hosts cost **\$11.63 an hour On-Demand**. A Compute Savings Plan covers any instance family, size, or Region, and Fargate and Lambda, so the July container replatform does not strand it.

The commitment in numbers, from the committed price table. A year is 8,760 hours.

The plan	Figure
The always-on floor, On-Demand, an hour	\$11.63
The same floor at the plan’s rates, an hour	\$8.53
Discount off On-Demand at the plan’s rates	26.7%
Hourly commitment	\$5.00
On-Demand usage the commitment covers, an hour	\$6.82
Coverage of the floor	58.6%
What it costs a month	\$3,650

The plan	Figure
What the covered usage would cost On-Demand a month	\$4,979
Saving a month, estimated	\$1,329
Saving a year, estimated	\$15,950
Committed over the one-year term	\$43,800

Why \$5.00 and not more

The plan is sized below the floor on purpose. It pays \$5.00 every hour whether or not the usage exists, so it must stay fully used in the worst case, which is after every deferred change lands. Three of them move compute hours: R-3 (the ingestion workers, declined for now), G-2 (the partner APIs on Graviton), and G-1 (the marketplace on Graviton). The others, E-2b, D-2, D-3, N-3, change a database, a volume, or the network and cannot move the EC2 floor.

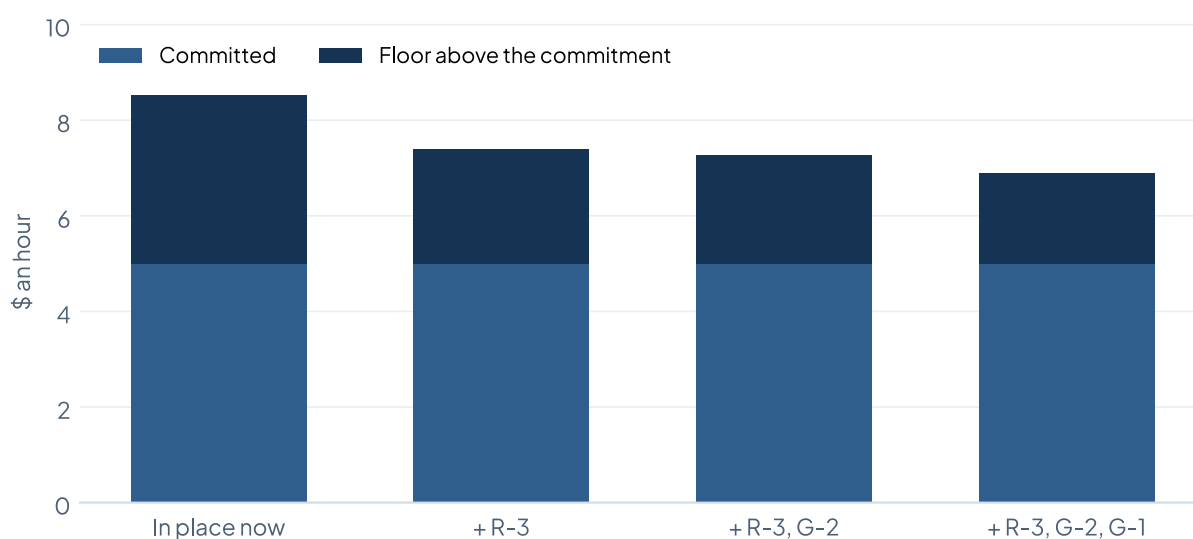
What the always-on floor is made of, now and after every deferred or declined compute change.

Workload in the floor	HostsTypes	On-Demand an hour	At the plan's rates	After all deferred changes, at the plan's rates
Marketplace web tier	6m5.2xlarge	\$2.30	\$1.69	\$1.44
Marketplace API tier	6m5.xlarge	\$1.15	\$0.85	\$0.72
Listing ingestion workers	4m5.4xlarge	\$3.07	\$2.26	\$1.13
Partner APIs	6m5.xlarge	\$1.15	\$0.85	\$0.72
Search cluster	3r5.2xlarge	\$1.51	\$1.09	\$1.09
CI runners	3m5.4xlarge	\$2.30	\$1.69	\$1.69
Shared services hosts	2m5.large, t3.medium	\$0.14	\$0.10	\$0.10
Total	30	\$11.63	\$8.53	\$6.89

The commitment's utilization as every deferred or declined compute change is added, one at a time. Declined at the readout; counted here in case the load test changes the answer.

Estate	Floor, On-Demand	Floor at the plan's rates	Utilization of the \$5.00	Margin above the commitment
In place now	\$11.63	\$8.53	100%	\$3.53 (41%)
And R-3, the ingestion workers	\$10.10	\$7.40	100%	\$2.40 (32%)
And G-2, the partner APIs on Graviton	\$9.92	\$7.27	100%	\$2.27 (31%)
And G-1, the marketplace on Graviton	\$9.41	\$6.89	100%	\$1.89 (27%)

The 24-hour floor at the plan's rates against the \$5.00 commitment



The floor at the plan's rates against the commitment, as each deferred change is added.

After every deferred change, and the declined ingestion resize too, the floor is still \$9.41 an hour On-Demand, or \$6.89 at the plan's rates: the commitment is used 100% of the time, with 27% of the floor to spare. The plan would earn less than it costs only if the covered usage fell below \$6.82 an hour On-Demand, and it would have to fall that far for a year. There is no break-even date: a No Upfront plan saves from its first hour for as long as the usage it covers runs.

The order mattered

We sized the plan after rightsizing, not before. The same floor cost \$13.94 an hour On-Demand before R-1 and R-2 halved the API tiers, and \$11.63 after. A plan sized on the larger number would have committed to hours that the Jun 2 resize removed.

What was declined, and why

The same covered usage bought four ways, from the committed price table. The three-year and All Upfront options save more and commit more.

Option	Outcome	Discount on the floor	Commitment for the same coverage, an hour	Saving a month	Committed over the term	Paid on day one
One year, No Upfront	Bought	26.7%	\$5.00	\$1,329	\$43,800	\$0
One year, All Upfront	Declined for now	31.5%	\$4.67	\$1,569	\$40,924	\$40,924
Three years, No Upfront	Declined for now	49.1%	\$3.47	\$2,447	\$91,148	\$0
Three years, All Upfront	Declined for now	53.8%	\$3.15	\$2,681	\$82,721	\$82,721

Lydia Brandt declined the three-year plan and All Upfront for now, in her words: "we don't know what checkout runs on in a year, and I'd rather keep the cash." The numbers say what that costs. A three-year plan would take 49.1% off the floor instead of 26.7%, about \$1,118 a month more, and it is the right buy at colo exit, when the end-state estate is known: the migration business case, not this report, decides what checkout runs on in a year. All Upfront on one year would pay \$40,924 on day one to save \$2,876 a year more, which is a decision about cash, and hers. Database commitments wait too: D-3, the catalog database resize, is deferred, and the orders database arrives in July.

The risk, plainly

- **A plan can be returned only for 7 days.** AWS allows a return within 7 days of purchase, in the same calendar month, for a plan with an hourly commitment of \$100 or less; this one qualifies, so it could be returned through June 10, 2026. After that it cannot be canceled, and the \$43,800 is owed for the year, used or not. The sizing is the only protection, and it holds 27% of the post-change floor in reserve.

- **A new floor could be smaller.** If the always-on estate fell below \$6.82 an hour On-Demand, the plan would pay for unused hours. Retiring a workload, or moving it to a service the plan does not cover, would do it. The remaining waves add compute; none removes the web, API, search, ingestion, or partner tiers.
- **Graviton helps, not hurts.** A Compute plan applies to Graviton as well as x86, so G-1 and G-2 lower the bill without stranding it. That is why it is a Compute plan and not an instance-family plan.
- **Support follows usage.** AWS Business Support+ is a percentage of usage, so the saving also lowers it a little. That is not counted.

Approval and purchase

Nothing was bought without Lydia Brandt's approval, and Vantalect buys nothing. Seatlark's finance team made the purchase in the management account, where a Savings Plan applies across every account in the organization by default. The steps: read AWS's own recommendation to compare with ours, find the offering for the term and payment option, buy at the approved hourly rate, and confirm.

BASH

```
aws ce get-savings-plans-purchase-recommendation --savings-plans-type COMPUTE_SP \
  --term-in-years ONE_YEAR --payment-option NO_UPFRONT --lookback-period-in-days SIXTY_DAYS
aws savingsplans describe-savings-plans-offerings --plan-types Compute --durations 31536000 \
  --payment-options "No Upfront" --currencies USD
aws savingsplans create-savings-plan --savings-plan-offering-id <offering-id> \
  --commitment 5.00 --client-token <uuid>
aws savingsplans describe-savings-plans --states active
aws ce get-savings-plans-utilization --time-period Start=<first-day>,End=<last-day> --
granularity DAILY
```

The last command is how utilization is watched: it should read 100% for as long as the floor holds. Record the purchase, the approver, and the date in the commitment register that the monthly cost review keeps.

What was deferred, and why

Two production changes were approved and then deferred this week, each for a reason its owner gave. Together they are \$20,094 a year, and they count for nothing toward the gate until they are in place and measured.

The two changes deferred this week, then the four deferred at the readout. Deferred changes count for nothing toward the gate.

ID	Change	Owner	Annualized	Moved to	Reason, as the owner gave it
E-2b	Search cluster io1 to gp3, production (3 volumes at 4,000 IOPS)	Imani Brooks	\$10,800	Monday, June 22, 2026	the production freeze (June 4 to 14) and the biggest on-sale of the summer: Imani Brooks will not touch the search cluster in the week before it
D-2	Partner database: downsize db.r5.2xlarge to db.r6g.xlarge (64 to 32 GiB) and move to Graviton, Multi-AZ	Diego Saldana	\$9,294	Saturday, June 27, 2026	a load test against a restored copy at the new class, then ten business days' notice of a maintenance window, which Seatlark Connect's partner contracts require
D-3	Catalog database: downsize db.r5.4xlarge to db.r6g.2xlarge (128 to 64 GiB) and move to Graviton, Multi-AZ	Imani Brooks	\$19,290	July 9, 2026	a load test that replays an on-sale, then the July maintenance window
G-2	Partner APIs to m7g.xlarge, once the arm64 build is validated	Diego Saldana	\$1,514	July 10, 2026	an arm64 build of the partner APIs, validated in staging
G-1	Marketplace web and API to Graviton (m7g), with the July container replatform	Aaron Feld	\$4,541	July 23, 2026	the July container replatform, which moves the marketplace to arm64 images
N-3	Cache the colo order-status and held-inventory calls, and read in-AZ (Transit Gateway, Direct Connect, and cross-AZ transfer down 30%)	Aaron Feld	\$7,369	August 14, 2026	an application change (the order-status cache), approved for the Q3 sprint

- **D-2, the partner database.** It halves the database's memory, so a load test against a restored copy at the new class comes first (it touches no production system and can run during the freeze). Diego Saldana's Seatlark Connect contracts then require 10 business days' notice of a maintenance window: the notice goes out when the test passes, by June 11, 2026, because a notice sent at the readout would have put the first window on June 8, 2026, inside the freeze. The class change fails a Multi-AZ database over once, and the window is Saturday, June 27, 2026.
- **E-2b, the production search volumes.** The change from io1 to gp3 is online, one volume at a time. Measured peak is 1,900 IOPS, including the reindex after the April on-sale. Imani Brooks will not change the search cluster in the week before the biggest on-sale of the summer, and the freeze runs June 4, 2026 to June 14, 2026. It is scheduled for Monday, June 22, 2026.

Three findings stay declined: **N-4** (Owen Pradhan, \$17,082 a year), **R-3** (Imani Brooks, \$13,455 a year), **S-2** (Aaron Feld, \$5,308 a year), and they count for nothing. R-3 is the one this week bears on: the ingestion workers run at 18% CPU at the 95th percentile, and Imani Brooks's answer was a load test first, which becomes a perf-test exercise after the freeze.

Honest limits, as of June 5, 2026

- **Nothing is measured.** Every dollar is an estimate at list. The week-6 measurement uses the 7 days before and after each change, on data through June 15, 2026; a change made after June 8, 2026 would not have 7 full days after it. All six of this week's changes were in place by June 3, 2026.
- **The June 5 on-sale is one observation, and it went over the line.** The prediction rests on the assumption that CPU doubles when vCPUs halve; the peak was 61%, over the 60% ceiling. The second on-sale is June 10, 2026, and the partner APIs have not yet run their busiest feed.
- **Compute Optimizer's default window missed April.** Its 14-day lookback never saw the on-sale, so we used 93 days of agent metrics. It is why the ingestion workers' recommendation is rated Medium.
- **The web tier keeps its size.** Resizing it like the API tier would save about \$841 a month. It is not a finding until autoscaling exists.
- **S3-1 can be off.** The saving depends on object sizes and on how often old objects are read, and S3 moves objects over days, not at once.
- **The Savings Plan estimate holds only while the floor holds.** It is measured by the plan's ARN in week 6: its fee, the usage it covered, and its utilization.

What comes next

The freeze starts June 4, 2026 and runs to June 14, 2026: no production change goes in. Week 5 builds what keeps the bill down and lands the remaining migration waves right: budgets and anomaly monitors per account, a tag policy, an AWS Config rule and a service control policy that refuse new gp2 and io1 volumes without an exception tag, Instance Scheduler as the non-production default, log retention in the account baseline, the VPC template change, and the wave runbook's sizing and MGN cleanup steps. What those guardrails prevent on waves 7 to 11 is cost avoidance: **not measured, not counted toward the gate, and not a saving on the bill.**

The Wednesday, June 10, 2026 on-sale is the second test of the smaller API tier, and the week-5 report records it. Diego Saldana's team is still checking that nothing links to the S3-3 renditions before it runs. Production changes resume after June 14, 2026: the search volumes on June 22, 2026 and the partner database on June 27, 2026.

© 2026 Vantalect, LLC. Sample deliverable; Seatlark, Inc. is fictional.

AWS, Amazon Web Services, AWS Graviton, and the AWS service names are trademarks of Amazon.com, Inc. or its affiliates. Ansible is a trademark of Red Hat, Inc. Elasticsearch is a trademark of Elasticsearch B.V. PostgreSQL is a registered trademark of the PostgreSQL Community Association of Canada. MySQL is a trademark of Oracle Corporation and its affiliates. Redis is a trademark of Redis Ltd. Microsoft and SQL Server are trademarks of the Microsoft group of companies. VMware and vCenter are trademarks of Broadcom Inc. NetApp is a trademark of NetApp, Inc. Java is a trademark of Oracle Corporation and its affiliates. Vantalect is not affiliated with or endorsed by any of them. Seatlark, Inc. is fictional.