



AWS COST · SAMPLE DELIVERABLE

Rightsizing, Moderation, and Commitments

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

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

May 1, 2026 · AWS Cost Analysis & Optimization: The Optimization (six weeks)

SAMPLE DELIVERABLE

This is a sample deliverable. Tidepost, Inc. is a fictional company: its people, products, users, accounts, investors, and figures are invented for this sample, and any resemblance to a real company, person, product, or publication is coincidental. AWS prices are list prices from the AWS Price List API retrieved on October 13, 2026, and from the public AWS Price List bulk API on October 18, 2026.

Contents

Summary	3
What went in	4
Change by change	6
Order matters: size first, then commit	8
Moderation: MOD-1, and why MOD-2 was declined	8
The Savings Plan: SP-1	9
The second plan: SP-2	11
The evidence the auditor can read	12
The changes, one at a time	13
SM-1a: The embedding endpoint on one GPU	13
EKS-1: Karpenter consolidates the production nodes	14
MOD-1: Rekognition Video where it earns its cost	14
EC2-5: Transcoders scale on queue depth	14
CACHE-1: Smaller ElastiCache nodes	15
CDN-1: A leaner rendition ladder	15
DDB-1: Provisioned capacity on the two steady tables	16
KIN-2: Fewer Kinesis shards	16
RDS-1: Smaller read replicas	17
SP-1: The Compute Savings Plan	17
EC2-4: The API tier on Graviton	18
The Well-Architected questions	18
Decided at the readout, and not made	19
Next week	20

Summary

Dev Raman and Kofi Brandvold: week 4 of the Optimization, Monday, April 27, to Friday, May 1, was the heaviest week by value, \$400,334 of estimates against week 3’s \$228,238. Tidepost’s engineers put **eleven changes in place** across two accounts (tp-management and tp-prod), estimated to save **\$400,334 a year**, 19.0% of the April bill annualized (\$2,110,762). With week 3’s thirteen changes, 24 are in place at an estimated **\$628,572 a year, 29.8% of the bill**.

The week had three parts. The **sizing** changes made the hardware smaller or fewer: the embedding endpoint, the production nodes, the transcoders, the cache nodes, the read replicas, and the Kinesis shards. The **demand** changes made the work smaller: the moderation routing (MOD-1), the leaner rendition ladder, and the API tier on Graviton. The **commitment**, the Compute Savings Plan, came last on purpose: it is sized on the usage that remains after every change on the roadmap, so it could be bought only after most of the sizing was done ([The Savings Plan](#)).

Every dollar here is an estimate. It is the Analysis’s price for the change at the April 6 usage, annualized, at list price, with the change’s own basis printed beside it. The week-six measurement report, on May 15, measures each change on the Cost and Usage Report, and the figure it prints will differ from the one here for a reason it will name.

Week 4 in numbers. Savings are annualized estimates at list price; the measurement is week six.

Week 4 at a glance	Result
Changes in place, April 27 to May 1	11
Estimated savings, a year	\$400,334
The same, a month	\$33,361
Share of the April bill annualized	19.0%
In place so far (weeks 3 and 4)	24
Estimated savings so far, a year	\$628,572
The three largest changes (SM-1a, MOD-1, CDN-1)	\$236,277
Well-Architected questions closed this week	3 of 5
Measured savings	none until week six

What the sponsor needs to do

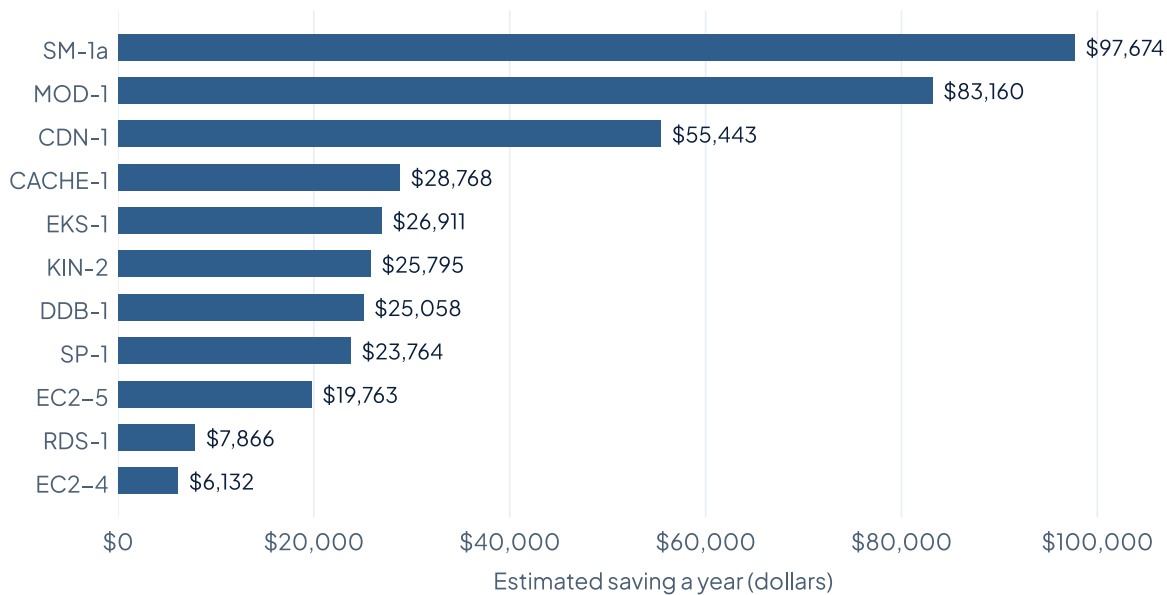
Dev Raman: nothing is needed to keep this week's savings. Gwen Tallis decides the SageMaker Savings Plan (SP-2) on Tuesday, May 5. She will ask Yuki Sorensen's team whether the moderation classifier's move to Inferentia2 (SM-3, deferred to July 13) stands, because it changes the floor SP-2 is sized on ([The second plan](#)). No decision is needed from you.

What went in

Six kinds of change went in. The largest by value is rightsizing, \$206,777 a year across six changes. The three largest changes, SM-1a, MOD-1, CDN-1, are \$236,277 of the week's \$400,334 ([the figure below](#)).

The week's changes by kind. The estimates add up to the week total to the dollar.

Kind of change	IDs	Changes	Estimate, a year	Share of week 4
Rightsizing	SM-1a, EKS-1, EC2-5, CACHE-1, KIN-2, RDS-1	6	\$206,777	51.7%
Moderation	MOD-1	1	\$83,160	20.8%
Architecture	CDN-1	1	\$55,443	13.8%
Pricing model	DDB-1	1	\$25,058	6.3%
Commitments	SP-1	1	\$23,764	5.9%
Graviton	EC2-4	1	\$6,132	1.5%
Total		11	\$400,334	100.0%



Estimated saving a year for each change in place in week 4.

When each change went in. The Savings Plan was bought on Thursday, the day before the API tier moved to Graviton; a Compute Savings Plan covers Graviton, so the order costs nothing.

Day	In place	Changes	Estimate, a year
Monday, April 27	SM-1a, EKS-1	2	\$124,585
Tuesday, April 28	MOD-1, EC2-5, CACHE-1	3	\$131,691
Wednesday, April 29	CDN-1	1	\$55,443
Thursday, April 30	DDB-1, KIN-2, RDS-1, SP-1	4	\$82,483
Friday, May 1	EC2-4	1	\$6,132
Total		11	\$400,334

Change by change

Each row names the change, the account, the owner, and the date, then what the resource looked like before, what it looks like now, the estimate, and the way back. "Before" is the line of the April 6 bill that the change touches; for SP-1 there is no line, because nothing was committed.

Every change in place in week 4, in the order it went in. Estimates are annualized at list price; nothing is measured.

Change	Before	After	Way back	Estimate, a year
SM-1a The embedding endpoint on one GPU. tp-prod; Yuki Sorensen; Monday, April 27.	Embedding endpoint (asynchronous inference): 2 ml.g5.12xlarge, one GPU of four used, at \$7.090 an hour.	Embedding endpoint from ml.g5.12xlarge (one GPU of four used) to ml.g5.2xlarge	Update the endpoint to the previous endpoint configuration. The same blue/green swap runs in the other direction.	\$97,674
EKS-1 Karpenter consolidates the production nodes. tp-prod; Kofi Brandvold; Monday, April 27.	EKS production nodes: 12 m5.4xlarge, On-Demand, always on, at \$0.768 an hour.	Karpenter consolidation of the production nodes (12 to 8)	Set the NodePool's consolidation policy back to when-empty, or set a node floor. Karpenter stops removing nodes at its next pass, and new ones launch as pods need them.	\$26,911
MOD-1 Rekognition Video where it earns its cost. tp-prod; Amara Delacroix; Tuesday, April 28.	Stored-video content moderation: 173,250 minutes a month, at \$0.10 a minute.	Rekognition Video only for accounts under 90 days old and for reported videos; the frame path covers the rest	Set the routing flag to send every video to Rekognition Video. It takes effect on the next upload.	\$83,160
EC2-5 Transcoders scale on queue depth. tp-prod; Kofi Brandvold; Tuesday, April 28.	Transcoder fleet: 6 g4dn.2xlarge, On-Demand, always on, at \$0.752 an hour.	Transcoders scale on queue depth (6 always on to an average of 3)	Set the group's minimum back to its old fixed size. The group stops scaling in and holds that size.	\$19,763
CACHE-1 Smaller ElastiCache nodes. tp-prod; Kofi Brandvold; Tuesday, April 28.	ElastiCache for Redis OSS, two shards with replicas: 4 cache.r6g.4xlarge, at \$1.642 an hour.	ElastiCache nodes from cache.r6g.4xlarge to cache.r6g.2xlarge (peak 29 GB used of 105 a shard)	Modify the replication group back to the larger node type. It takes the same rolling steps, and the data stays in memory on the replicas.	\$28,768

Change	Before	After	Way back	Estimate, a year
CDN-1 A leaner rendition ladder. tp-prod; Bea Okafor; Wednesday, April 29.	Video and images delivered to North America: 470 TB a month, at the published North America tiers.	Rendition ladder: 720p on cellular, no 1080p for vertical clips (app release April 29)	Change the manifest to offer 1080p again; apps that have already updated follow it. Apps that never updated never changed.	\$55,443
DDB-1 Provisioned capacity on the two steady tables. tp-prod; Lior Haddad; Thursday, April 30.	On-demand requests: 3.3 billion write and 16.5 billion read request units, at \$0.625 and \$0.125 per million.	Provisioned capacity with auto scaling on the two steady tables (85% of the traffic)	Switch the table back to on-demand (AWS allows it once a day). The data and the keys do not change.	\$25,058
KIN-2 Fewer Kinesis shards. tp-prod; Lior Haddad; Thursday, April 30.	Kinesis shards: 100 shards, at \$0.015 a shard-hour.	Kinesis from 100 shards to 20, records aggregated in the collector	UpdateShardCount back up, in steps, and switch the collector's aggregation off with its flag.	\$25,795
RDS-1 Smaller read replicas. tp-prod; Dev Raman; Thursday, April 30.	Read replicas: 2 db.r6g.2xlarge, at \$0.899 an hour.	Read replicas from db.r6g.2xlarge to db.r6g.xlarge (working set 21 GB)	Modify the replica's class back, which restarts it again.	\$7,866
SP-1 The Compute Savings Plan. tp-management; Gwen Tallis; Thursday, April 30.	No Reserved Instance or Savings Plan exists; everything bills On-Demand (the Foundation's FR-2 found none).	Compute Savings Plan, one year, No Upfront, \$7.27 an hour (70% of the floor's plan cost), bought in tp-management with sharing on	Returnable only through April 30; after that it runs to April 30, 2027 (The Savings Plan)	\$23,764
EC2-4 The API tier on Graviton. tp-prod; Marcus Oyelaran; Friday, May 1.	The API tier as EC2-3 leaves it, on c6i.4xlarge instances.	API tier on Graviton (c7g.4xlarge)	Point the launch template back to the c6i type and run the refresh. Both images stay built.	\$6,132
Total		11 changes		\$400,334

Order matters: size first, then commit

The week's changes are in the order they are because each one changes what the next can safely assume. The API tier shows it best. In week 3, EC2-3, in place April 22, took the tier's minimum down: API tier minimum from 9 to 6 (two a zone), with target tracking. This week, EC2-4 moved that smaller tier to Graviton, and the estimate for EC2-4 is priced on the tier as EC2-3 leaves it, so the two do not count the same instance twice. The same pairs run through the rest of the week: EC2-5 (transcoders that scale) comes before CDN-1b (one rendition fewer to transcode), and SM-1a (a smaller endpoint) comes before SM-1b (an endpoint that scales).

A commitment is the one change that cannot be priced after the fact, because it is a promise to spend. That is why SP-1 was bought last this week, on the floor that every other change, in place or deferred, leaves behind.

Moderation: MOD-1, and why MOD-2 was declined

MOD-1 is the week's second-largest change, \$83,160 a year, and the one with the most at stake, because it touches what reaches your users' timelines. The bill's history says why the second path exists. Rekognition Video added on July 14 beside the frame path, after a graphic video reached timelines. MOD-1 keeps that second path exactly where the risk is.

What each moderation path checks, before and after. The frame path is untouched.

Check	Before MOD-1	After MOD-1
The frame path (Rekognition on sampled frames and uploaded images)	Every video	Every video
Rekognition Video on the whole video	Every video	Rekognition Video only for accounts under 90 days old and for reported videos; the frame path covers the rest
Transcribe and the toxicity check on the transcript	Every video	Only videos with detected speech (TR-1, in place April 24)

What made it safe. The moderation log for March: every video already passes the frame path, and the second path flagged 0.02% of older accounts' videos the frame path missed. Amara Delacroix owns the decision and the rollback: one routing flag sends every video to Rekognition

Video again, effective on the next upload, and her team reviews the reported-video queue daily. **The estimate** is 173,250 minutes a month at \$0.10, times the 40% of videos from older accounts that nobody reports.

MOD-2 was declined, on purpose. After MOD-1 the frame path alone carries most videos, and halving it halves what it sees. It would have saved \$37,422 a year at the Analysis's price, and it is not counted. The two changes are alternatives that interact: halving the frame path's sampling after MOD-1 would have left the frame path as the only check on most videos, at half the frames.

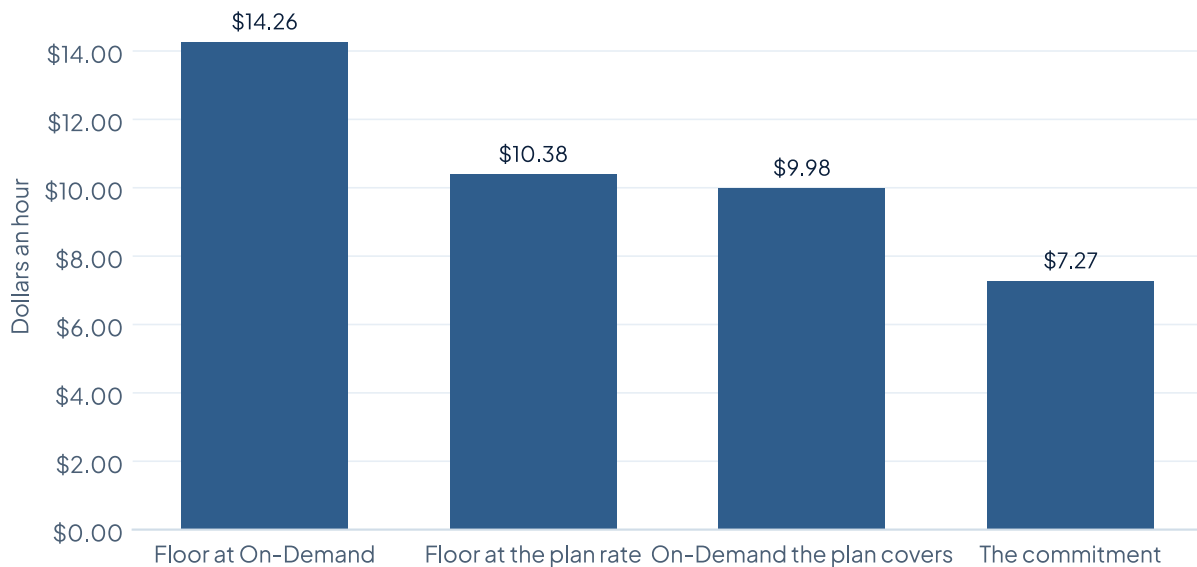
The Savings Plan: SP-1

Gwen Tallis owns commitments, and SP-1 is hers. It is a Compute Savings Plan, one year, No Upfront, **\$7.27 an hour**, bought in tp-management on Thursday, April 30, with discount sharing left on so that it covers eligible compute in every account of the organization. It is a commitment of \$63,685 over its term, billed every hour whether it is used or not. Before this week, Tidepost had no Reserved Instance and no Savings Plan, and every compute hour billed at On-Demand.

How it was sized. The floor is the compute that runs in every hour after every change on the roadmap, in place or deferred: the lowest usage the organization's 9 accounts will see together, because the plan is bought in the management account with sharing on. At On-Demand it costs **\$14.26 an hour**. At the one-year No Upfront Savings Plan rates for the same mix it costs \$10.38, which is 27.2% off. The plan commits 70% of that, \$7.27, and covers \$9.98 of On-Demand usage at the discount. The 30% left uncommitted is the safety margin: if the floor is smaller than the roadmap says, the plan is still fully used.

How SP-1 was sized. Every row is derived from the floor; the plan's utilization is read daily.

Step	How it is derived	Dollars an hour
The floor at On-Demand	The usage that remains after every change, priced at On-Demand	\$14.26
The same usage at the plan rate	The one-year No Upfront rate for each instance type in the floor	\$10.38
The commitment	70% of the plan-rate cost	\$7.27
On-Demand usage the commitment covers	The commitment divided by 72.8%, the plan rate as a share of On-Demand	\$9.98
Saving, an hour	Covered On-Demand usage less the commitment	\$2.71
Saving, a year	The hourly saving over every hour of the year; this is the estimate in the register	\$23,764



The floor and the commitment, in dollars an hour.

The return rule, said exactly. AWS lets a Savings Plan of \$100 an hour or less be returned within seven days of purchase, and only inside the calendar month of the purchase. SP-1 is \$7.27 an hour, under the limit, and it was bought on April 30, the last day of the month, so its return window ran through April 30: that one day. After that the plan stays until April 30, 2027. A plan is not a contract with no exit: it was returnable on the day. After that, the protection is the floor, built to hold even if every change on the roadmap lands.

What it covers and what it does not. Usage above the floor bills at On-Demand; usage below it, if a change removes more than the roadmap says, leaves part of the commitment unused. The deferred staging rebuild (EC2-1a and EC2-1b) is the one deferred change to usage the plan covers, and the floor is already sized as if it were done. Valkey (CACHE-2) and Inferentia2 (SM-3) change ElastiCache and SageMaker, which a Compute Savings Plan does not cover, so they cannot move its floor. The part of the bill a Compute Savings Plan does not cover, such as SageMaker, RDS, ElastiCache, and OpenSearch, is outside it.

Evidence. There is no pull request for a purchase. The record is the purchase in tp-management (the Savings Plans console), the CreateSavingsPlan call in the organization trail in tp-log-archive, and Gwen Tallis’s written approval before the purchase. The Well-Architected question this closes is COST07, which she accepted in writing on March 10; it closes on April 30, when SP-1 lands.

The second plan: SP-2

SP-2 is a SageMaker Savings Plan, one year, No Upfront, **\$4.14 an hour**, 80% of the hosting floor’s plan cost. It was approved at the April 17 readout and has not been bought. Gwen Tallis decides it on Tuesday, May 5.

How SP-2 is sized, on the hosting floor after SM-1a.

Step	Dollars an hour
The SageMaker hosting floor at On-Demand	\$6.78
The same usage at the plan rate (23.6% off)	\$5.18
The commitment (80% of the plan-rate cost)	\$4.14
On-Demand usage it covers	\$5.42
Estimated saving, a year	\$11,192
Commitment over the term	\$36,266

The floor is the question. It is sized after SM-1a (the smaller embedding endpoint), which landed this week, and it still holds the moderation classifier on its current GPU instances. SM-3 would move the classifier to Inferentia2. Deferred to July 13: it needs a Neuron compile and a precision check against the labeled March set. If it lands, the floor under SP-2 changes. That is the trade Gwen Tallis weighs on May 5, and it is hers to make; either answer is a good one if it is made with the floor in front of her.

The evidence the auditor can read

Tidepost's SOC 2 Type 2 observation window opened April 1 and runs to September 30, so every change in this report was made inside it. The change-management control (CC8.1) has been operating since March 27: Production changes through a pull request and the pipeline: read-only permission sets in tp-prod for developers, machine learning, and data, console-built resources imported into the CDK, and drift reports on console changes. The auditor tests a sample of the changes in the window; Vantalect does not select the sample and gives no opinion on what the auditor finds. What follows is where each record is, so that Sasha Ilyin can produce it. The pull request, pipeline run, drift report, API call, and window are the same five records as week 3's report.

The repository, date, and window class for every change in week 4. SP-1 has a purchase record in place of a pull request.

Change	Where the record lives	Date	Window	Approved by
SM-1a	tidepost-cdk	Monday, April 27	Standard	Yuki Sorensen, at the April 17 readout
EKS-1	tidepost-cdk	Monday, April 27	Off-peak, rolling	Kofi Brandvold, at the April 17 readout
MOD-1	the moderation service's repository	Tuesday, April 28	Standard	Amara Delacroix, at the April 17 readout
EC2-5	tidepost-cdk	Tuesday, April 28	Standard	Kofi Brandvold, at the April 17 readout
CACHE-1	tidepost-cdk	Tuesday, April 28	Off-peak, rolling	Kofi Brandvold, at the April 17 readout
CDN-1	the mobile app's repository and the playback manifest in tidepost-cdk	Wednesday, April 29	App release	Bea Okafor, at the April 17 readout
DDB-1	tidepost-cdk	Thursday, April 30	Standard	Lior Haddad, at the April 17 readout
KIN-2	tidepost-cdk	Thursday, April 30	Off-peak, rolling	Lior Haddad, at the April 17 readout
RDS-1	tidepost-cdk	Thursday, April 30	Off-peak, rolling	Dev Raman, at the April 17 readout
SP-1	the Savings Plans purchase record in tp-management	Thursday, April 30	Purchase	Gwen Tallis, at the April 17 readout

Change	Where the record lives	Date	Window	Approved by
EC2-4	tidepost-cdk	Friday, May 1	Off-peak, rolling	Marcus Oyelaran, at the April 17 readout

Standard. Nothing users see is interrupted, so the day the pipeline made the change is its window. **Off-peak, rolling.** Announced to the platform team the day before and made off-peak, one replica, shard, or zone at a time, so a failure stops after the first. **App release.** Made by a release through the app stores on the day, and reaching viewers as they update. **Purchase.** A purchase, not a deployment: made on the day, with the finance owner's written approval before it.

The changes, one at a time

SM-1a: The embedding endpoint on one GPU

Owner: Yuki Sorensen. **Account:** tp-prod. **In place:** Monday, April 27. **Estimate:** \$97,674 a year (\$8,140 a month). **What changed:** Embedding endpoint from ml.g5.12xlarge (one GPU of four used) to ml.g5.2xlarge.

How it was made. A new endpoint configuration names the smaller instance type with the same instance count. SageMaker brings the new instances up beside the old ones and shifts the asynchronous queue across, so no request is dropped. The previous configuration stays registered. **The estimate** is 2 instances $\times$ (\$7.090 – \$1.515) $\times$ 730 hours. **Why it was safe:** GPU utilization by device: device 0 at 31% on average, devices 1 to 3 idle.

Way back. Update the endpoint to the previous endpoint configuration. The same blue/green swap runs in the other direction. **What the owner watches.** Yuki Sorensen's team watches the queue's backlog and the endpoint's invocation errors through the swap. **Well-Architected.** This change is on the list for COST06 (closes April 27, when SM-1a lands) and SUS02 (closes April 27, when SM-1a lands), accepted in writing and taken up by the cost engagement.

Next.

- **SM-1b** (planned for Monday, May 4): scales the smaller endpoint on its backlog.
- **SM-3** (deferred to Monday, July 13): moves the moderation classifier to Inferentia2.
- **SP-2** (decided on Tuesday, May 5): the SageMaker plan is sized on the hosting floor this change leaves.

EKS-1: Karpenter consolidates the production nodes

Owner: Kofi Brandvold. **Account:** tp-prod. **In place:** Monday, April 27. **Estimate:** \$26,911 a year (\$2,243 a month). **What changed:** Karpenter consolidation of the production nodes (12 to 8).

How it was made. A Karpenter NodePool allows consolidation of underutilized nodes. The pod disruption budgets and zone-spread constraints stay, so Karpenter drains one node at a time and never below what the budgets allow. Pods with zone constraints limit how far consolidation can go. **The estimate** is 4 m5.4xlarge at \$0.768 × 730 hours. **Why it was safe:** Pod requests sum to 58% of node capacity.

Way back. Set the NodePool's consolidation policy back to when-empty, or set a node floor. Karpenter stops removing nodes at its next pass, and new ones launch as pods need them. **What the owner watches.** Kofi Brandvold's team watches pending pods and node count through the first full day. The estimate assumes the node count in the title; the Cost and Usage Report shows what the cluster holds.

MOD-1: Rekognition Video where it earns its cost

Owner: Amara Delacroix. **Account:** tp-prod. **In place:** Tuesday, April 28. **Estimate:** \$83,160 a year (\$6,930 a month). **What changed:** Rekognition Video only for accounts under 90 days old and for reported videos; the frame path covers the rest.

How it was made. The moderation service sends a video to Rekognition Video only when the uploading account is newer than the age in the title, or when the video has been reported. Every video still passes the frame path. Rekognition Video was added beside the frame path in July after a graphic video reached timelines; this change keeps it where that risk is. **The estimate** is 173,250 minutes a month at \$0.10, times the 40% of videos from older accounts that nobody reports.

Why it was safe: The moderation log for March: every video already passes the frame path, and the second path flagged 0.02% of older accounts' videos the frame path missed.

Way back. Set the routing flag to send every video to Rekognition Video. It takes effect on the next upload. **What the owner watches.** Amara Delacroix's team reviews the reported-video queue daily and the videos the frame path alone passed each week.

EC2-5: Transcoders scale on queue depth

Owner: Kofi Brandvold. **Account:** tp-prod. **In place:** Tuesday, April 28. **Estimate:** \$19,763 a year (\$1,647 a month). **What changed:** Transcoders scale on queue depth (6 always on to an average of 3).

How it was made. The transcoder Auto Scaling group scales on queue depth instead of staying fixed. It adds instances when jobs wait and, through a lifecycle hook, lets a job in flight finish before an instance is terminated. **The estimate** is From 6 to 3 g4dn.2xlarge on average at \$0.752.

Why it was safe: Queue depth: zero for 61% of the day.

Way back. Set the group's minimum back to its old fixed size. The group stops scaling in and holds that size. **What the owner watches.** Kofi Brandvold's team watches the age of the oldest job in the queue.

Next.

- **CDN-1b** (planned for Wednesday, May 6): removes a rendition the transcoders no longer need to make.

CACHE-1: Smaller ElastiCache nodes

Owner: Kofi Brandvold. **Account:** tp-prod. **In place:** Tuesday, April 28. **Estimate:** \$28,768 a year (\$2,397 a month). **What changed:** ElastiCache nodes from cache.r6g.4xlarge to cache.r6g.2xlarge (peak 29 GB used of 105 a shard).

How it was made. Each shard's nodes move to the smaller type through ModifyReplicationGroup, one replica at a time. Each shard fails over to a resized replica, and the old primary is resized last. Applications reconnect through the cluster endpoint. **The estimate** is 4 nodes $\times$ (\$1.642 - \$0.821) $\times$ 730 hours. **Why it was safe:** BytesUsedForCache peaks at 29 GB a shard.

Way back. Modify the replication group back to the larger node type. It takes the same rolling steps, and the data stays in memory on the replicas. **What the owner watches.** BytesUsedForCache and evictions for each shard, which the evidence read at their peak before the change.

Next.

- **CACHE-2** (deferred to Tuesday, May 19): moves the same nodes from Redis OSS to Valkey.

CDN-1: A leaner rendition ladder

Owner: Bea Okafor. **Account:** tp-prod. **In place:** Wednesday, April 29. **Estimate:** \$55,443 a year (\$4,620 a month). **What changed:** Rendition ladder: 720p on cellular, no 1080p for vertical clips (app release April 29).

How it was made. The player picks its rendition from the manifest the server sends. The new app version asks for 720p on cellular and does not ask for 1080p on vertical clips. The change reaches viewers as they update the app, so the saving arrives gradually and the estimate is for every viewer

on the new version. **The estimate** is 470 TB less 24% at the published North America tiers, once every viewer has the new app. **Why it was safe:** Bytes per view by rendition: 1080p on a phone screen is 41% of bytes.

Way back. Change the manifest to offer 1080p again; apps that have already updated follow it. Apps that never updated never changed. **What the owner watches.** Bea Okafor's team watches playback start time and rebuffering. Week six normalizes CloudFront bytes to bytes per view so that growth in views is not counted as a saving.

Next.

- **CDN-1b** (planned for Wednesday, May 6): stops transcoding the rendition the new ladder never serves.

DDB-1: Provisioned capacity on the two steady tables

Owner: Lior Haddad. **Account:** tp-prod. **In place:** Thursday, April 30. **Estimate:** \$25,058 a year (\$2,088 a month). **What changed:** Provisioned capacity with auto scaling on the two steady tables (85% of the traffic).

How it was made. The two steady tables switch from on-demand to provisioned capacity with Application Auto Scaling, a target utilization and a floor set from the daily curve. The other tables stay on-demand. AWS allows one switch of billing mode per table a day. **The estimate** is 85% of \$4,125.00 on demand, against provisioned units at 140% of the mean. **Why it was safe:** Request rate by minute: the two feature tables vary 30% across the day.

Way back. Switch the table back to on-demand (AWS allows it once a day). The data and the keys do not change. **What the owner watches.** Lior Haddad's team watches throttled requests, with an alarm on any throttle on the two tables.

KIN-2: Fewer Kinesis shards

Owner: Lior Haddad. **Account:** tp-prod. **In place:** Thursday, April 30. **Estimate:** \$25,795 a year (\$2,150 a month). **What changed:** Kinesis from 100 shards to 20, records aggregated in the collector.

How it was made. The collector aggregates records into one payload unit, and the stream is resharded in steps with UpdateShardCount, which limits each step to halving the shard count. **The estimate** is $80 \text{ shards} \times (\$0.015 + \$0.02) \times 730 \text{ hours}$, and $\text{PUT units} \div 25$. **Why it was safe:** Peak ingest is 3.1 MB a second; each shard takes 1 MB.

Way back. UpdateShardCount back up, in steps, and switch the collector's aggregation off with its flag. **What the owner watches.** Lior Haddad's team watches write throttling and consumer iterator age through the steps.

Next.

- **GR-6** (built in week 5): caps retention on every stream.

RDS-1: Smaller read replicas

Owner: Dev Raman. **Account:** tp-prod. **In place:** Thursday, April 30. **Estimate:** \$7,866 a year (\$656 a month). **What changed:** Read replicas from db.r6g.2xlarge to db.r6g.xlarge (working set 21 GB).

How it was made. Each read replica is modified in turn to the smaller class, which restarts it. The API's read pool uses the other replica while one restarts. The writer is not touched: Dev Raman declined that change at the readout. **The estimate** is $2 \times (\$0.899 - \$0.450) \times 730$ hours. **Why it was safe:** Buffer cache hit 99.6% with a 21 GB working set; CPU peaks at 31%.

Way back. Modify the replica's class back, which restarts it again. **What the owner watches.** Replica lag and CPU on both replicas, which the evidence read at their peak before the change.

SP-1: The Compute Savings Plan

Owner: Gwen Tallis. **Account:** tp-management. **In place:** Thursday, April 30. **Estimate:** \$23,764 a year (\$1,980 a month). **What changed:** Compute Savings Plan, one year, No Upfront, \$7.27 an hour (70% of the floor's plan cost), bought in tp-management with sharing on.

How it was made. Bought once in tp-management with discount sharing left on, so it covers Compute Savings Plan eligible usage in every account of the organization. No resource is touched or tagged. Usage above the commitment bills at On-Demand, and the commitment is billed every hour whether it is used or not. **The estimate** is \$7.27 an hour covers \$9.9828 of On-Demand usage at the floor's blended one-year rate (27.2% off). **Why it was safe:** The floor after every change on the roadmap is \$14.260 an hour On-Demand; a plan of \$100 an hour or less can be returned within seven days in the same month.

Way back. See [The Savings Plan](#): returnable through April 30 only, then it runs to April 30, 2027.

What the owner watches. Gwen Tallis reads the plan's utilization every day in the first month. The floor is the usage that remains after every change on the roadmap, in place or deferred, so utilization should stay full. **Well-Architected.** This change is on the list for COST07 (closes April 30, when SP-1 lands), accepted in writing and taken up by the cost engagement.

Next.

- **SP-2** (decided on Tuesday, May 5): the second commitment, on SageMaker hosting.

EC2-4: The API tier on Graviton

Owner: Marcus Oyelaran. **Account:** tp-prod. **In place:** Friday, May 1. **Estimate:** \$6,132 a year (\$511 a month). **What changed:** API tier on Graviton (c7g.4xlarge).

How it was made. The API tier's launch template moves from c6i.4xlarge to c7g.4xlarge, with the arm64 container image the pipeline now builds beside the amd64 one. An instance refresh replaces the instances one zone at a time. The API is written in Go, and the arm64 build passed the load test. **The estimate** is 7 instances $\times$ (\$0.680 – \$0.580) $\times$ 730 hours. **Why it was safe:** The API is Go; its arm64 build passed the load test.

Way back. Point the launch template back to the c6i type and run the refresh. Both images stay built. **What the owner watches.** Marcus Oyelaran's team watches latency and error rate through the refresh. SP-1 covers Graviton instances, so the plan bought the day before covers them from the first hour.

The Well-Architected questions

Five Well-Architected questions were accepted in writing because their fixes are cost changes, and a question closes when the last change on its list is in place. Three of the five close this week; two wait on week 5.

The five questions accepted in writing. Tidepost saves its answers in the Well-Architected Tool at the May 14 milestone.

Question	Signed by	Changes on its list	In place at May 1	State
COST06 How do you meet cost targets when you select resource type, size and number?	Gwen Tallis, March 10	EC2-2, SM-1a, SM-2	EC2-2, SM-1a, SM-2	The fix is in place since April 27, with SM-1a
SUS02 How do you align cloud resources to your demand?	Yuki Sorensen, March 11	SM-1a, SM-2	SM-1a, SM-2	The fix is in place since April 27, with SM-1a
COST07 How do you use pricing models to reduce cost?	Gwen Tallis, March 10	SP-1	SP-1	The fix is in place since April 30, with SP-1
PERF03 How do you store, manage, and access data in your workload?	Lior Haddad, March 9	S3-1, ATH-1	none	Closes May 6, when S3-1 lands
SUS04 How do you take advantage of data management policies and patterns to support your sustainability goals?	Lior Haddad, March 9	KIN-1, LOG-2, S3-1, S3-3	KIN-1	Closes May 6, when S3-1 lands

Decided at the readout, and not made

Three changes were priced and then declined at the April 17 readout. They are not in any estimate here and they are not counted. Two bear on this week: MOD-2 ([above](#)) and RDS-2, the database writer. Dev Raman declined RDS-2 because growth after the round, and a resize is a failover inside the observation window. RDS-1, the read replicas, went ahead on the same database because replicas restart one at a time without a failover.

The changes declined at the readout. They stay in the register and are not counted.

Change	Decided by	Reason	Estimate, a year
MOD-2 Frame sampling from one frame a second to one every two seconds	Amara Delacroix	after MOD-1 the frame path alone carries most videos, and halving it halves what it sees	\$37,422
RDS-2 Writer from db.r6g.4xlarge to db.r6g.2xlarge (Multi-AZ)	Dev Raman	growth after the round, and a resize is a failover inside the observation window	\$15,759
OS-1 OpenSearch data nodes from r6g.2xlarge.search to r6g.xlarge.search	Lior Haddad	the search team is midway through a k-NN migration	\$8,778

Next week

Week 5 runs May 4 to May 8: data lifecycle, telemetry, and the guardrails. The facts give these dates.

The changes planned for week 5, and the Savings Plan decision on May 5.

Change	Planned	Owner	Estimate, a year
EBS-3 Snapshot chains past their keep period deleted	Monday, May 4	Kofi Brandvold	\$31,949
SM-1b The embedding endpoint scales on its backlog	Monday, May 4	Yuki Sorensen	\$6,636
LOG-2 Shorter application log retention, counsel's copy first	Tuesday, May 5	Sasha Ilyin	\$32,197
S3-3 Old renditions to Glacier Instant Retrieval	Tuesday, May 5	Amara Delacroix	\$30,734
SP-2 The SageMaker Savings Plan	Tuesday, May 5	Gwen Tallis	\$11,192
ATH-1 Athena on partitioned Parquet	Wednesday, May 6	Lior Haddad	\$17,340
CDN-1b One rendition fewer to transcode	Wednesday, May 6	Kofi Brandvold	\$2,635
S3-1 Telemetry to Parquet at ingest	Wednesday, May 6	Lior Haddad	\$30,640

Five more have dates after week 5: CACHE-2 (Redis OSS to Valkey on the same nodes) is deferred to May 19; S3-2 (Originals to S3 Glacier Instant Retrieval) is deferred to May 20; EC2-1a (Staging rebuilt smaller in tp-staging) is deferred to June 1; EC2-1b (Staging on weekdays only) is deferred to June 8; SM-3 (The moderation classifier on Inferentia2) is deferred to July 13. Each is deferred for a reason the facts give, and none is dropped.

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

AWS, Amazon Web Services, AWS Graviton, AWS Inferentia, and the AWS service names are trademarks of Amazon.com, Inc. or its affiliates. PostgreSQL is a registered trademark of the PostgreSQL Community Association of Canada. Kubernetes is a registered trademark of The Linux Foundation. Redis is a registered trademark of Redis Ltd. Valkey is a trademark of the Linux Foundation. OpenSearch is a registered trademark of Amazon.com, Inc. SOC 2 is a registered trademark of the American Institute of Certified Public Accountants. CIS Benchmarks is a trademark of the Center for Internet Security, Inc. GitHub is a trademark of GitHub, Inc. Jupyter is a trademark of LF Charities. Vantalect is not affiliated with or endorsed by any of them. Tidepost, Inc. is fictional.