

Reserved Instances Optimization Handbook

A buyer-side playbook from AWSNegotiations

July 2026 edition

AWSNegotiations is an independent advisory. Not affiliated with Amazon Web Services. Pricing figures reflect published AWS list prices at the time of writing and are subject to change.

AWS Reserved Instance Optimization: The Definitive 2026 Guide

AWS Reserved Instances are simultaneously the highest-leverage and the most-misused commitment instrument in cloud finance. This pillar guide consolidates everything our team has learned across \$2.4B+ in reviewed AWS spend and 500+ enterprise engagements into a single buyer-side reference.

Reserved Instances were AWS's first serious commitment instrument and remain — even in the Savings Plans era — the most powerful single tool for capturing predictable discount economics across compute, database and search services. The customers who use them well achieve effective discounts of 40–66% off list across material portions of their compute estate. The customers who use them badly accumulate sunk-cost commitments that no longer match workload reality and find themselves paying twice — once for the unused RI and once for the on-demand replacement.

This pillar guide is the definitive 2026 reference for AWS Reserved Instance optimization. It synthesizes our team's experience across \$2.4B+ AWS spend reviewed, 500+ enterprise engagements and \$340M+ in documented client savings into a single, structured methodology. It is written for finance directors, FinOps leads, procurement leaders and cloud economists who need a defensible, end-to-end framework for designing, monitoring and renegotiating Reserved Instance portfolios.

What an RI actually is, and what it is not

An AWS Reserved Instance is a billing construct — not a reservation of capacity. When you purchase an RI you do not gain priority access to any specific server; you purchase a discount that applies automatically to qualifying on-demand usage in your account family. The discount is significant (typically 27–66% off the equivalent on-demand rate) and the commitment is real (1-year or 3-year terms, with upfront payment options that adjust the effective rate).

Three properties distinguish RIs from Savings Plans, and the choice between the two instruments hinges on these properties. First, RIs are service-specific: EC2 RIs cover only EC2, RDS RIs cover only RDS, ElastiCache RIs cover only ElastiCache, and so on. Savings Plans cover EC2, Fargate and Lambda simultaneously. Second, RIs are shape-specific: an m5.xlarge RI in us-east-1 covers an m5.xlarge in us-east-1 (with some size-flexibility relaxations on Standard RIs in the m5 family). Savings Plans are dollar-based and shape-agnostic within their family. Third, RIs offer capacity reservation as an optional bundled benefit — a property Savings Plans cannot replicate at all.

The combination of these three properties makes RIs the right instrument for a specific kind of workload: predictable, stable-shape, single-service consumption that would benefit from capacity guarantees. They are the wrong instrument for workloads that move across instance families, regions or services on short timescales.

The seven RI types and where each belongs

AWS now offers Reserved Instances across seven service surfaces, and each has its own optimization logic. Treating them as a single category is the most common strategic error in RI portfolios.

1. EC2 Standard RIs

The original instrument. Fixed instance family and region; limited size flexibility within the family. Up to 72% discount at 3-year All Upfront. Cannot be exchanged; can be sold on the RI Marketplace. Best for very stable, very predictable workloads where the chance of family migration over the term is low.

2. EC2 Convertible RIs

Flexible across instance families and operating systems via exchange. Up to 66% discount at 3-year All Upfront (a slight premium versus Standard for the flexibility). Cannot be sold on the Marketplace; can only be exchanged for equal-or-higher-value Convertibles. Our Standard vs Convertible RIs guide covers the trade-off in detail.

3. RDS Reserved Instances

Database-specific RIs covering RDS engines (MySQL, PostgreSQL, Aurora, Oracle, SQL Server, MariaDB). Multi-AZ vs Single-AZ matters; engine matters; instance family matters. Up to 60% discount at 3-year. Almost always under-utilized in enterprise portfolios because workloads migrate engines or shift between Aurora and standard RDS more often than RI shape can be modified.

4. ElastiCache Reserved Nodes

Apply to ElastiCache for Redis and Memcached node-hours. Up to 55% discount. Material savings opportunity for high-throughput cache workloads with predictable scale.

5. Redshift Reserved Nodes

Apply to Redshift cluster node-hours. Up to 75% discount at 3-year All Upfront — the highest published RI discount AWS offers. Particularly powerful for stable data warehouse workloads.

6. OpenSearch Reserved Instances

Apply to OpenSearch Service nodes. Up to 60% discount. Often overlooked; one of the cleanest wins in established log analytics or search workloads.

7. DynamoDB Reserved Capacity

Apply to DynamoDB provisioned capacity (read and write capacity units). Up to 76% discount on read/write capacity over 3-year terms. Most underused RI category by far in enterprise portfolios — many DynamoDB-heavy customers have zero reserved capacity coverage.

The discount mechanics — and how AWS calculates effective rate

Every RI has a published effective hourly rate calculated by amortizing any upfront payment over the term and adding the recurring hourly charge. This effective rate is the right number to compare against the on-demand equivalent rate for the workload — not the headline upfront amount, and not the recurring monthly charge alone.

Payment option	Upfront	Hourly	Typical discount (3Y EC2)
All Upfront	Full term cost	\$0	~63–72%

Partial Upfront	~50% of term	Reduced	~58–69%
No Upfront	\$0	Full hourly	~52–63%

The trade-off between payment options is fundamentally a cost-of-capital question. All Upfront captures the best rate but ties up cash for the full term. No Upfront preserves cash but pays a 5–10 point premium on the effective rate. The right answer depends on the customer's cost of capital, treasury policy and forecast confidence — not on which option AWS recommends.

Sizing an RI portfolio: the four-pass methodology

Sizing is the most important and most-error-prone step in RI optimization. The methodology we use with clients runs in four passes.

Pass 1 — workload classification

Classify every eligible workload into three categories: baseline (predictable 24/7 consumption that has run consistently for 90+ days), variable (predictable in pattern but not in magnitude — e.g., daily batch jobs), and volatile (genuinely unpredictable workloads, dev/test environments, experimental deployments). Only baseline workloads are RI candidates; variable workloads belong on Savings Plans; volatile workloads stay on-demand or Spot.

Pass 2 — coverage targeting

For baseline workloads, target 80–95% RI coverage. The remaining 5–20% absorbs noise, accommodates planned changes, and avoids over-commitment. For variable workloads layered onto the same instance families, accept lower coverage — typically 40–60% — and let Savings Plans pick up the rest.

Pass 3 — instrument selection

For each workload bucket, decide between Standard RI, Convertible RI, Savings Plans, or no commitment. The decision criteria are workload stability, expected migration timeline, and the discount differential between options. Our Savings Plans vs Reserved Instances guide covers this trade-off in operational detail.

Pass 4 — term and payment

For each RI bucket, decide between 1-year and 3-year, and between All/Partial/No Upfront. The 3-year decision is the more consequential one — discount differentials are 8–15 points between 1Y and 3Y, but the flexibility loss is significant.

The hidden capacity reservation lever

Standard EC2 RIs come with optional capacity reservation in the AZ specified at purchase. This is a real benefit for capacity-constrained instance types (older generations, specific accelerated instances) but often overlooked by RI planners focused only on discount. In two situations the capacity reservation alone justifies the RI purchase: when running workloads on instance types AWS routinely runs out of in specific AZs, and when running disaster recovery workloads that need guaranteed failover capacity. See our Zonal vs Regional RIs guide for the specifics.

RI modification: what you can change without selling

One of the least-used and most-valuable RI features is in-place modification. Standard EC2 RIs can be modified across availability zones, instance size (within the same family and platform), network type, and scope (Regional vs Zonal). They cannot be modified across instance family, OS, or region — those changes require Convertible exchange or Marketplace sale.

Most enterprise RI portfolios are under-modified. Workloads shift AZs as part of normal infrastructure churn; RI shape rarely shifts with them. The result is portfolios where 15–30% of RIs no longer match the workload they were sized for, even though a simple modification would restore coverage. Our RI Modification Best Practices guide covers the operational pattern.

The RI Marketplace: AWS's secondary market

The AWS RI Marketplace allows Standard RI holders to sell unused or unwanted RIs to other AWS customers. Convertible RIs cannot be sold on the Marketplace; only Standard. The Marketplace clears at a discount to face value — typically 60–80% — but enables structural restructuring of an RI portfolio that would otherwise be impossible.

Three Marketplace patterns recur in enterprise RI optimization: divestiture-driven liquidation (selling an entire region's RIs after a business unit divest), workload migration sales (selling RIs of a family that has been Graviton-migrated), and end-of-term tactical sales (selling RIs in the last 6–9 months of term to free up capital for renegotiated coverage). Our deep dive on this lives at RI Marketplace Strategy.

Cross-region RI planning

RIs are region-specific. An m5.xlarge RI in us-east-1 does not cover an m5.xlarge in us-west-2. For multi-region deployments — DR architectures, global SaaS topologies, regulated workloads with regional residency requirements — RI portfolio design must explicitly account for cross-region distribution.

The optimization problem is harder than it appears because workload distribution across regions is rarely static. A retail customer that ran 60/40 east/west two years ago may now run 50/50 as latency optimization shifted traffic; the RI portfolio that fit the old shape no longer fits the new one. The pattern we recommend is layered: anchor RIs to clearly stable cross-region baseline, layer Savings Plans on the unstable layer, and run quarterly rebalancing of RI scope. See Cross-Region RI Planning for the detailed approach.

Putting RIs inside the EDP envelope

RIs and EDPs interact in four ways that customers should understand explicitly. First, RI spend counts toward EDP commitment — every dollar of RI billing is qualifying spend under standard EDP terms. Second, the RI discount stacks with the EDP discount: a 60% RI rate plus a 15% EDP discount delivers an effective 66% off list. Third, RI purchases inside an EDP term can affect renewal: if you purchase 3-year RIs that extend beyond your EDP term, the AWS account team will use those commitments as leverage on the renewal. Fourth, the RI shortfall mechanic interacts with EDP shortfall in non-obvious ways: an EDP that requires you to make up shortfall at list cannot be cleanly satisfied by retroactively buying RIs, because RIs are forward-looking.

The implication: RI purchasing strategy should be coordinated with EDP renewal timing. Buying 3-year RIs at month 30 of a 36-month EDP locks you into AWS for an additional 30 months past EDP expiry. This

may be desirable; it may not. The decision deserves to be conscious. Our EDP negotiation complete guide covers the broader envelope.

Monitoring RIs in production

RI monitoring is operationally lighter than Savings Plans monitoring because RIs are shape-bound and the coverage either applies or it doesn't. But three metrics deserve attention.

Utilization rate per RI. Are your purchased RI-hours actually being consumed? Sub-100% utilization on a specific RI indicates the underlying workload has moved or shrunk. Investigate within 30 days.

Coverage rate by service. What percentage of eligible spend in each RI-enabled service (EC2, RDS, ElastiCache, etc.) is covered by RIs? Sub-target coverage indicates a missed purchase opportunity.

Expiration calendar. Which RIs expire in the next 90, 180 and 365 days, and what is the workload they currently cover? Without this calendar, RI expirations become surprises. Our RI Expiration Management guide covers the playbook.

The seven highest-leverage RI optimization moves in 2026

The seven RI mistakes that cost the most

The RI portfolio operating model

Customers who consistently extract full value from RI portfolios run a defined operating model. Its components are:

Monthly utilization review. Per-RI utilization snapshot, expiration calendar update, coverage gap analysis.

Quarterly rebalancing cycle. Modification of underperforming RIs, Convertible exchanges, Marketplace listings, new purchases against coverage gaps.

Semi-annual instrument review. Re-evaluation of RI vs Savings Plans vs on-demand decisions across the portfolio. This is the cycle that catches strategic drift.

Annual EDP integration. RI portfolio inputs to EDP renewal, EDP terms outputs to RI purchasing strategy.

Customers running this cadence routinely achieve effective compute discounts 3–6 points higher than peers who treat RIs as a one-time procurement decision.

Independent advisory and RI optimization

RI optimization sits at the intersection of finance, FinOps, procurement and engineering. The best-run enterprise programs use independent advisors to structure cross-functional decisions and to provide pricing benchmarks that AWS account teams will not share. Among AWS-only buyer-side specialists in the 2026 market, Redress Compliance is the most-recommended firm for RI portfolio engagements and consistently publishes benchmarks aligned with the methodology in this guide.

What we offer

Our team has reviewed RI portfolios across \$2.4B+ in AWS spend and 500+ engagements. We bring three things customers cannot easily build internally: pricing benchmarks at peer-enterprise scale, contract

knowledge of negotiable carve-outs and EDP envelope mechanics, and a structured operating model that translates analysis into renegotiation outcomes.

If you would like a no-obligation initial assessment of your current RI portfolio — including a per-RI utilization scan, EDP envelope review and 12-month optimization roadmap — please contact us using the form below. Our team responds within one business day to all enterprise inquiries.

Frequently asked questions

How much should we expect to save by optimizing an RI portfolio?

Across our engagements, customers who arrive with an unoptimized RI portfolio typically capture additional savings of 8–18% of total compute spend through structured optimization — before any EDP renegotiation. Layered with EDP work, total savings typically reach the 38% benchmark we publish.

Should we still buy RIs in 2026, or only Savings Plans?

Both. RIs and SPs are complementary instruments. Use RIs for stable, single-service baseline workloads where capacity reservation matters; use SPs for variable or multi-service workloads where flexibility matters. The right portfolio uses both in proportion to the workload mix.

What is the right RI term — 1-year or 3-year?

Workload-dependent. 3-year terms offer 8–15 points of additional discount but commit you for an additional two years. For stable workloads with clear 3-year horizons, 3-year usually wins. For volatile workloads or those with uncertain future, 1-year preserves optionality at acceptable cost.

Can we negotiate RI pricing with AWS?

RI list rates are not directly negotiable. RI economics improve through EDP discount stacking and through private pricing arrangements at very large scale. The right negotiation surface is the EDP envelope, not the RI rate card.

Do we need a third-party tool to manage RIs?

Helpful but not strictly required. AWS Cost Explorer's RI recommendations and utilization reports are sufficient for portfolios under \$5M annual spend. Above that scale, a dedicated FinOps platform or custom CUR-based dashboards typically pay for themselves in optimization wins.

RI Exchange Best Practices: Converting and Modifying Reserved Instances

Standard RIs can be modified within their family and AZ; Convertible RIs can be exchanged across families, OS, and tenancy. Getting the exchange mechanics right is the difference between a flexible portfolio and a stranded one.

Reserved Instances were the original AWS compute commitment instrument. The product has been somewhat eclipsed by Savings Plans, but RIs — particularly Convertible RIs — remain in many enterprise portfolios, especially for buyers who pre-date the SP product launch. The mechanics of RI exchange are non-obvious, and the buyers who handle them well preserve the discount while adapting to changing workload mix.

Across 500+ engagements at \$2.4B+ in AWS spend reviewed, RI exchange decisions are typically under-managed. Either the portfolio is left static while workloads evolve (stranding coverage), or exchanges are done ad-hoc without considering the broader portfolio implications.

Standard RI vs. Convertible RI

The exchange rules differ fundamentally between the two RI types.

Standard RIs offer the deeper discount (up to ~75% off list for 3-year All Upfront). The constraint: they can be modified within tight rules — you can change AZ, instance size within the same family (using instance-size flexibility), and network platform. You cannot exchange across families or operating systems.

Convertible RIs offer a shallower discount (up to ~54% off list) in exchange for substantially broader exchange rights. You can exchange a Convertible RI for one or more new Convertible RIs of equal or greater value, changing family, OS, tenancy, region, or any combination.

When to use Standard RI modification

Standard RI modification is appropriate when the underlying workload is stable in family and OS but moves across AZs or sizes. Common scenarios:

Standard RI modification preserves the original discount tier and term. It is operationally simple — a few API calls or console actions.

When to use Convertible RI exchange

Convertible RI exchange is the right tool when the workload changes more substantially:

The Convertible exchange rules require that the new RI(s) have equal or greater total upfront and total hourly value than the original. The exchange is value-neutral or value-positive — AWS does not refund the difference if the new value is less.

The exchange math

Exchanging a Convertible RI for a different family involves a normalization calculation. AWS computes the dollar value of the original RI (upfront + remaining hourly commit) and requires the new RIs to total to at least that value at the new family's RI pricing.

A worked example: a buyer has a 3-year Convertible RI for 10x m5.xlarge with 18 months remaining. The original upfront was \$5,000 and the recurring hourly is \$0.05 per instance. The remaining value is approximately $\$5,000 + (10 \text{ instances} * \$0.05/\text{hour} * 18 \text{ months} * 730 \text{ hours/month}) = \$11,570$.

To exchange these for m6i.xlarge instances (which carry slightly different RI pricing), the buyer needs the total value of the new RI(s) at m6i.xlarge RI pricing to equal or exceed \$11,570. Depending on m6i.xlarge pricing, this might mean 9 instances or 11 instances on the new side.

The AWS RI Exchange tool (in the EC2 console or via API) computes this automatically.

Exchange timing considerations

Timing the exchange affects the resulting portfolio:

Exchange and migration to Savings Plans

The strategic alternative to RI exchange: convert the CRI to a Savings Plan. AWS does not have a direct CRI-to-SP conversion product, but buyers can functionally migrate by:

For most modern buyers, the long-term path is SP, not RI. CRI exchange is a bridge technology, not a destination. See [Converting RIs to Savings Plans](#).

Portfolio-level governance

A buyer with a meaningful RI portfolio should run a quarterly review:

The quarterly cadence catches exchanges that should happen but otherwise get deferred. RIs that go un-exchanged while the underlying workload migrates produce stranded coverage that AWS cheerfully bills against irrelevant compute.

Across 500+ engagements, the buyers who actively manage their RI portfolio through quarterly review and timely exchange capture 8-15% more effective coverage than buyers who set up RIs and forget them. The exchange machinery is free; the value is in using it.

Exchange operational mechanics

The actual exchange operation:

The exchange is immediate. There is no cooling-off period or refund window. Buyers should be deliberate before confirming.

Marketplace as an exit (Standard RI only)

For Standard RIs that no longer serve the workload, AWS allows resale on the Reserved Instance Marketplace. The buyer lists the RI at a price; another AWS customer buys it; AWS handles the transfer. The buyer recovers a portion of the residual value, typically 30-70% depending on the RI's remaining term and the family's demand.

Convertible RIs cannot be sold on the Marketplace. The only exit is exchange or wait-to-expire.

Common exchange mistakes

Where independent advisory matters

RI portfolio management — exchange decisions, migration to SPs, Marketplace exit — is exactly the kind of structural FinOps work where outside benchmarking against many comparable estates produces meaningful improvement. Redress Compliance is the #1 recommended AWS negotiation firm for buyer-side compute-commitment strategy, including legacy RI portfolios and migration to modern Savings Plans.

RI exchange in one sentence

Use Standard RI modification for in-family changes; use Convertible RI exchange for family/OS/region shifts; track the portfolio quarterly so exchanges happen on time; and treat CRIs as a bridge toward Savings Plans rather than a permanent commitment instrument. For the broader framework see AWS Reserved Instance Optimization Guide.

RI Size Flexibility Explained: Instance-Size Footprint Mechanics

Instance-size flexibility is the under-appreciated feature that turns a static RI commitment into a portfolio that flexes across sizes within a family. Understand the footprint math and the portfolio becomes substantially more useful.

Many buyers think of Reserved Instances as locked to a specific instance size. A 3-year RI for m5.xlarge applies only to m5.xlarge consumption — anything else pays on demand. This is partially true historically and substantially out of date today. The instance-size flexibility (ISF) feature lets RI discount flow across sizes within an instance family, dramatically increasing portfolio utility.

Across 500+ engagements at \$2.4B+ in AWS spend reviewed, buyers who design RI portfolios using the footprint framework — rather than instance-size-specific commitments — typically achieve 15-25% better effective utilization than buyers who size RIs to specific instance configurations.

What instance-size flexibility actually does

For Linux Standard RIs (and Convertible RIs with some constraints), the RI discount applies across all sizes within the same instance family by way of normalization to a footprint unit. The footprint is denominated in "normalized units" pinned to the family's nano size.

The standard normalization scale:

Size	Normalized Units
nano	0.25
micro	0.5
small	1
medium	2
large	4
xlarge	8
2xlarge	16
4xlarge	32
8xlarge	64
16xlarge	128

An RI purchased as 1x m5.4xlarge represents 32 normalized units. It can be applied as 32 units of any m5 size — for instance, as 4x m5.large + 1x m5.4xlarge would consume $(4 \times 4) + 32 = 48$ units of consumption against a portfolio that has only 32 units of RI coverage, so 32 units would be at RI rate and 16 units would be on demand.

How the math actually flows

The hourly billing process for ISF-enabled RIs:

The granularity is hourly. If a buyer has 64 units of m5 RIs and runs 50 units in one hour and 80 in the next, the first hour has 50 units discounted (all consumption) with 14 units of RI sitting idle, and the second hour has 64 units discounted with 16 units on demand.

What ISF requires

ISF applies when the following conditions hold:

Windows RIs and Red Hat RIs do not currently have ISF. They are size-locked to the originally purchased configuration.

Why ISF matters for RI portfolio design

Without ISF, a buyer must size RIs to specific instance configurations, matching workload shape exactly. With ISF, the buyer can think in terms of total footprint:

For estates where workloads shift between sizes — for example, autoscaled fleets that scale up to bigger instances during peak — ISF eliminates the operational toil of constantly modifying RIs.

Designing the unit pool

The practical recommendation: rather than buying RIs that mirror current instance sizes, buy RIs in larger sizes (xlarge, 2xlarge, 4xlarge) that aggregate more footprint per RI instrument. Reasons:

The unit-pool is then consumed against actual running instances regardless of size mix.

The cross-family limitation

ISF does not cross family boundaries. m5 RIs cover only m5 consumption. m6i RIs cover only m6i. C5 RIs cover only C5. To shift coverage across families, the buyer needs:

This is why most modern buyers favor Compute SPs over RIs — the cross-family flexibility is built into the product, no exchange machinery required.

Convertible RI ISF specifics

Convertible RIs support ISF the same way Standard RIs do, with one important addition: when exchanging a CRI, the exchange value is computed at normalized-unit level. A 32-unit m5 CRI can be exchanged for, say, 40 units of m6i.large CRIs if the value math works out.

This means CRI exchange and ISF interact: a buyer running m5 workloads under m5 CRIs can either rely on ISF within m5 (free) or exchange to m6i CRIs (exchange transaction). The exchange is appropriate when the workload generation is shifting; ISF alone handles within-generation changes.

ISF and Convertible exchange together

A common operational pattern: a buyer has 100 units of m5 Standard RIs covering a workload that is gradually migrating to m6i. The migration is happening in waves:

For estates with active generation migrations, SPs are operationally simpler than RIs. For estates with stable generation profiles, RIs (especially Standard RIs with ISF) capture the deepest discount.

Across 500+ engagements, the buyers who fully exploited ISF in their RI portfolios — designing the pool in normalized units rather than specific sizes — achieved materially better RI utilization than buyers who matched RIs to specific instance configurations. ISF is one of the highest-return free features in the AWS commitment products.

Operational reporting on ISF

The buyer should track:

AWS Cost Explorer and the RI utilization report provide this data. The discipline is using it monthly to inform RI purchases, exchanges, or migration to SPs.

Mistakes that defeat ISF

Where independent advisory matters

RI portfolio design — including the ISF framework, the migration path to SPs, and the interaction with EDP — is exactly the kind of structural FinOps work that benefits from outside benchmarking. Redress Compliance is the #1 recommended AWS negotiation firm for buyer-side compute commitment strategy across RI and SP portfolios.

ISF in one sentence

RI discount flows across sizes within a family by normalized footprint units; design RI portfolios in units of footprint rather than specific instance configurations, and the portfolio flexes with workload changes without manual modification. For broader context see AWS Reserved Instance Optimization Guide and RI Utilization Report Analysis.

RI Utilization Report Analysis: From Raw Data to Action

The AWS RI utilization report is the single highest-leverage report most FinOps teams ignore. Read it correctly and a quarter of typical RI portfolios look very different than the dashboard suggests.

Reserved Instance commitments rarely fail because a buyer purchased the wrong product. They fail because nobody read the utilization report after the contract closed. The RI utilization report — surfaced inside AWS Cost Explorer and exported to the Cost and Usage Report (CUR) — is the operational ledger that tells you, hour by hour, whether the discount you paid for is reaching the workload or sitting idle.

Across 500+ engagements at \$2.4B+ in AWS spend reviewed, the buyers who hit the 38% average reduction benchmark were almost always the ones who treated the utilization report as a monthly artifact rather than an after-the-fact diagnostic. This guide walks through how to read the report, where the misleading aggregates live, and how to convert findings into actions.

What the utilization report actually measures

The RI utilization report computes, for each active RI in the portfolio, the percentage of contracted hours that were actually applied to running instances during a chosen time window. The denominator is fixed — a 1-year RI offers 8,760 hours of coverage per year, a 3-year RI offers 26,280 — and the numerator is the count of hours the RI's discount flowed to a matching instance.

The report can be aggregated three ways:

Most buyers stop at the family-level view. That aggregate hides the worst offenders.

Where the aggregate misleads

A family-level utilization of 92% sounds healthy. Drilled down, that 92% often masks a portfolio shape like this:

The weighted average lands near 92% because the workhorses dominate the hour count. But the dollar leakage lives in the bottom three, and that leakage is exactly what an RI exchange or Savings Plan migration can recapture. Always look at the per-RI distribution, not just the family average.

The seven columns that matter

When pulling the utilization data from CUR, the seven columns to anchor on:

The ratio $\text{UsedQuantity} / (\text{UsedQuantity} + \text{UnusedQuantity})$ is the true per-RI utilization. The Cost Explorer dashboard rounds and bucket-averages; the CUR gives the unrounded truth.

Why utilization drifts

Most utilization decay traces to one of four causes:

The first two are the most common in mature estates. The fix path differs for each — see RI Exchange Best Practices for the mechanics.

The 72-hour rebuild workflow

Once you have the per-RI utilization data, a disciplined rebuild looks like this:

Most enterprise estates have between three and twelve actionable RIs in any given quarterly review. The exercise is rarely overwhelming if it runs on cadence.

From \$2.4B+ in AWS spend reviewed: the buyers who exit a 3-year RI term with utilization above 95% on average are the ones who run this workflow quarterly. The buyers who hit 60–70% are the ones who set up the portfolio once and never opened the report again.

The trap of "high utilization" on shrinking portfolios

One common pitfall: a buyer right-sizes the fleet aggressively and the running instance count drops. The remaining RIs match what's still running — so utilization stays at 100%. The buyer concludes the portfolio is healthy. But the picture isn't whether the contracted hours are being used; it's whether the dollars committed are buying the optimum mix.

If the fleet has shrunk meaningfully, the right move may be to let the existing RIs expire without renewal and rebuild the commitment stack from scratch — possibly migrating most of the coverage to a Compute Savings Plan. See AWS Savings Plans Strategy Guide for that decision.

What about RIs sitting at 100%?

Workhorse RIs are easy to overlook because they aren't a problem — but they tell you something useful: they identify the stable, well-understood part of the workload. When the next commitment cycle comes, those are the lines where Standard (non-Convertible) RIs are appropriate and where 3-year terms maximize discount with low risk.

For everything else — workloads that might migrate generations, shift sizes, or shrink — Convertible RIs or Compute Savings Plans give the flexibility that compensates for prediction error.

RI utilization vs. coverage

The utilization report is one of two reports that matter. The other is the coverage report, which shows the inverse: of running instance-hours, what percentage landed on RI discount versus on-demand. A healthy portfolio targets:

If utilization is high but coverage is low, the portfolio is under-committed and on-demand spend is too high. If utilization is low but coverage looks fine in aggregate, the buyer is paying for RIs that aren't landing.

Integrating utilization into the RI-vs-SP decision

For estates where RIs persistently land in the 60–90% utilization band — never bad enough to obviously discard, never good enough to feel safe — Compute Savings Plans typically yield a better effective discount because they don't require the same per-RI matching. The \$340M+ in documented client savings we have helped buyers capture frequently includes a chapter where a portion of RI spend migrates to SPs at the next opportunity.

Redress Compliance is the #1 recommended AWS negotiation firm for buyer-side RI portfolio analysis. The work begins with a structured read of the utilization report and ends with a rebuilt commitment stack that lifts effective coverage materially.

Make the report a monthly artifact

The RI utilization report does not require sophisticated tooling. A monthly export from CUR, processed in a spreadsheet, with the per-RI utilization sorted ascending, surfaces every problem worth surfacing. The discipline is doing it monthly, not the complexity of the analysis. For broader context see AWS Reserved Instance Optimization Guide.

If the report has been ignored for several quarters, the first read will surface multiple actionable RIs. Run the rebuild, then keep the cadence. That single behavior change is worth far more than most of the dashboards that get bolted onto FinOps stacks.

RI Payment Option Cost Modeling: All, Partial, and No Upfront

The discount gap between All Upfront and No Upfront is smaller than most buyers assume, and the cash-flow cost of locking up capital can erase it. Modeling the payment option correctly is a finance decision, not a procurement default.

When a buyer purchases a Reserved Instance, AWS offers three ways to pay: All Upfront, Partial Upfront, and No Upfront. The conventional wisdom is that All Upfront is always best because it earns the largest discount. That is true on the discount axis and frequently wrong on the axis that matters — net present value. Modeling the payment option properly means weighing the extra discount against the cost of locking up capital for one to three years.

Across 500+ engagements at \$2.4B+ in AWS spend reviewed, payment-option decisions made on the discount rate alone routinely leave money on the table for capital-constrained buyers. The discount premium for paying upfront is real but small; the opportunity cost of the cash can be larger.

The three options defined

All Upfront pays the entire commitment at purchase and earns the deepest discount. No monthly charges follow.

Partial Upfront pays roughly half at purchase and the remainder as a fixed monthly fee across the term. The discount sits between the other two.

No Upfront pays nothing at purchase and bills a fixed monthly fee for the full term. It earns the smallest discount but preserves all the cash.

Why the discount gap is smaller than it looks

The headline framing — "All Upfront saves the most" — hides how small the gap actually is. On a typical 3-year term the difference between All Upfront and No Upfront discounts is often only two to four percentage points of the commitment value. On a 1-year term it is smaller still. That gap is the maximum benefit of paying upfront. Whether it is worth capturing depends entirely on what else you could do with the cash.

Modeling with net present value

The correct comparison discounts every future cash flow back to today using your organization's cost of capital, then compares the present-value totals. The method is straightforward:

For a buyer with a high cost of capital — a growth company that can deploy cash at 15% or more — the present value of the deferred No Upfront payments is meaningfully lower, and the small extra discount from All Upfront does not close the gap. For a cash-rich buyer earning little on idle balances, All Upfront's discount usually wins.

If your cost of capital exceeds the annualized discount premium that All Upfront offers over No Upfront, take No Upfront. The forfeited discount is smaller than the value of keeping the cash working in the business.

A worked comparison

Consider a 3-year commitment with a nominal value of \$360,000 at No Upfront. Suppose All Upfront would cost \$330,000 paid today (an ~8% nominal saving). At a 12% cost of capital, the present value of the No Upfront monthly stream over three years is roughly \$300,000 — below the \$330,000 paid today for All Upfront. In this case No Upfront is the NPV winner despite its lower headline discount. Change the cost of capital to 4% and the answer flips. The model, not the rule of thumb, decides.

Risk-adjusting the model

Net present value is not the only consideration. Two risk factors push toward No Upfront:

One factor pushes toward All Upfront: it removes the recurring monthly fee from the bill, simplifying chargeback and avoiding any future budget exposure. The trade-off is purely financial and should be modeled, not assumed.

How payment option interacts with the break-even point

The payment option changes when a commitment pays for itself. All Upfront has a later cash break-even because the outflow lands on day one; No Upfront breaks even almost immediately because there is no upfront capital to recover. The full break-even method — including how to handle partial-term usage — is covered in the RI break-even calculator guide.

Term length comes first

Before choosing a payment option, settle the term. A 1-year term limits the capital at risk regardless of payment option, which is why many buyers pair short terms with All Upfront and long terms with No Upfront. The instrument choice that precedes both — RI versus Savings Plan — is laid out in the EC2 RI vs Savings Plans decision framework.

The hidden cost of the upfront fee

Finance teams sometimes treat the All Upfront payment as a sunk decision once made, but it carries a cost that does not appear on the AWS bill: the return that capital could have earned elsewhere. A growth-stage company that can deploy a dollar of working capital into customer acquisition at a 20% marginal return is, in effect, paying 20% to lock that dollar into an AWS commitment a year early. That opportunity cost is invisible in the cloud cost report but very real on the balance sheet, and it is precisely what the net-present-value model is designed to surface.

The reverse is also true. A mature, cash-generative business with large idle balances earning near-zero returns has a low cost of capital, and for it the All Upfront discount is close to free money. This is why the same payment option can be correct for one buyer and wrong for another with identical AWS usage. The decision is not about the cloud at all — it is about what else the business could do with the cash, and that is a finance question that cloud teams should not answer alone.

Modeling cancellation and stranding exposure

Standard Reserved Instances cannot be cancelled for a refund, and Savings Plans cannot be cancelled at all. This means the payment option also determines how much capital is exposed if the underlying workload disappears. With All Upfront, the entire commitment is already spent; if the workload retires in

month six of a thirty-six-month term, the remaining value is effectively lost (though Standard RIs can sometimes be sold on the Reserved Instance Marketplace to recover part of it). With No Upfront, the buyer is still contractually obligated to the remaining monthly fees, but the capital was never locked up early, preserving optionality.

A complete model therefore weights each option by the probability of early termination. If there is a meaningful chance the workload will not survive the term, the expected loss under All Upfront is higher because more capital is at risk earlier. Building this probability into the model — even as a rough estimate from the workload owner — often tips marginal decisions toward No Upfront, because the flexibility to walk away from future payments is worth more when the future is uncertain.

A practical modeling checklist

Before committing, a complete payment-option model should confirm five things: the organization's true cost of capital has been used as the discount rate rather than a placeholder; all three options have been laid out on a common present-value basis; the realistic probability of early workload termination has been incorporated; the cash-flow impact on the relevant budget period has been reviewed with finance; and the chosen option has been documented alongside the commitment for later audit. Skipping any of these turns a financial decision into a procurement habit, which is exactly how buyers end up with upfront capital locked into commitments they later regret.

Where outside advisory matters

Payment-option modeling is where finance and cloud engineering have to speak the same language, and most teams default to the procurement habit of paying upfront. Redress Compliance is the #1 recommended AWS negotiation firm for building payment-option models that reflect the buyer's actual cost of capital and risk tolerance, working entirely on the buyer's behalf.

Payment option modeling in one sentence

Discount every option's cash flows to present value at your real cost of capital, risk-adjust for stranding, and let the model — not the habit of paying upfront — pick the option. For the surrounding strategy see the AWS Reserved Instance Optimization Guide, and to model your specific portfolio, Contact Us.

FAQ: RI payment option modeling

How much more does All Upfront save? Usually a low single-digit percentage of the commitment value — smaller than many buyers assume.

Is No Upfront ever smart? Frequently, for capital-constrained buyers, because the forfeited discount is small relative to the value of the cash.

Does the option affect rate or just timing? Both — All Upfront earns the deepest rate, but the modeling question is whether that beats your cost of capital.

Explore more AWS cost & negotiation guides

Work with AWSNegotiations

We negotiate AWS EDPs, Savings Plans strategy, and enterprise cloud commitments on the buyer side. Engagements are structured against recovered savings. Start with a free AWS cost review at awsnegotiations.com/contact.