GOOGLE CLOUD: GEMINI ENTERPRISE COMMIT ECONOMICS AND THE CUD SCOPE CHANGE

What Google changed on June 16 and August 26, how the new commitment levers work, what they do to the enterprise agreement and the chargeback model, the dates that matter, and how customers should respond

> *A research note for CIOs, CFOs, FinOps leads, cloud platform owners, and procurement teams running Google Cloud with a Gemini Enterprise deployment, an enterprise agreement, or Compute Engine committed use discounts, who have been offered a Flexible Savings Plan, seen a change in how commitments are attributed across projects, or been asked to size an AI commitment for 2027. Written from the buyer's side of the table. Redress Compliance, 2026.*

Key takeaways

- On August 26, 2026 Google added three ways to pay for Gemini Enterprise: pay as you go at standard token and compute rates, Flexible Savings Plans at 10 percent off for one year and 20 percent off for three, and deferred execution at up to 50 percent off inference for workloads that can wait. Spend caps, anomaly alerts, and a pooled quota across Gemini Enterprise, Antigravity, and Android Studio came with it.
- The savings plan is a spend commitment with the usual mechanics: usage below the commit is still billed in full, unused commitment does not roll over, and usage above it is billed at on demand rates with no discount. The 3 year rate is thin for infrastructure and expensive as a bet on a market where model prices, model versions, and your own architecture change every quarter.
- Separately, on June 16, 2026 Google changed the default scope of resource based committed use discounts from project to billing account, with sharing enabled. New accounts and existing accounts without active commitments were switched automatically. A business unit that bought commitments for its own projects may now be subsidizing others under proportional attribution, and the chargeback model most enterprises built no longer describes what the bill does.
- A published price step is already in the calendar: Gemini 3.6 and 3.7 Flash double from USD 0.75 and 3.75 to USD 1.50 and 7.50 per million tokens on January 1, 2027. A dollar commitment sized on 2026 token prices buys half the volume in 2027; a commitment sized on 2027 prices is overcommitted in 2026.
- Treat every commitment as a negotiation, not a checkbox. Pilot on pay as you go, size the plan on observed run rate after the price step, keep the term at one year, route batch work to deferred execution, and fix the CUD scope, attribution model, and enterprise agreement burn down in writing before the next commitment is bought.

1. WHAT GOOGLE IS DOING

Two changes, three months apart, both to how Google Cloud converts consumption into commitment. Neither is a price increase. Both move risk toward the customer while presenting as flexibility.

- **Gemini Enterprise billing, August 26.** Agent workloads broke the assumptions behind per seat pricing: a single user request can trigger a chain of model calls, reasoning steps, subagent launches, and retries that scale with nothing a seat count predicts. Google's response is to unbundle the price into levers. Pay as you go removes the base subscription and bills tokens and compute at standard rates, initially for select customers. Flexible Savings Plans are spend based commitments on Gemini Enterprise and the broader AI platform stack, with no stated minimum or maximum, that count against an existing enterprise agreement. Deferred execution marks eligible agent jobs for off peak processing at up to half price. Spend caps and anomaly alerts arrive alongside, as does a single project wide quota pool replacing separate licenses for Antigravity and Android Studio.
- **CUD scope, June 16.** Resource based committed use discounts on Compute Engine were, by default, scoped to the project that bought them. Google changed the default to billing account scope with sharing enabled. New billing accounts get it automatically. Existing accounts without active commitments were switched automatically. Existing accounts with active

commitments were left as they were. Under billing account scope a commitment bought in one project covers eligible usage in any project on the same billing account, allocated by proportional attribution unless the customer configures prioritized attribution. Regional and machine family eligibility still apply.

- **The billing schema change, January 21.** Spend based CUDs moved to a consumption model schema where the bill shows the discounted rate directly rather than list price plus a credit line. Google states total cost is unchanged. What changed is how savings are reported, which is what every finance model and chargeback report was built on.
- **The price step, January 1, 2027.** Gemini 3.6 Flash and 3.7 Flash launched at introductory rates through December 31, 2026, with standard rates double that from the first day of 2027. Pro tier prompts above 200,000 tokens already bill at roughly twice the base rate. Both are published, not speculative.

What has not changed matters too. Per seat Gemini Enterprise subscriptions remain available with a daily quota pool. Compute Engine Flex CUDs at roughly 28 percent for one year and resource based CUDs at 37 percent and above remain the benchmark a 10 percent AI savings plan is measured against. Existing enterprise agreements are honored; the new plans burn down against them.

2. WHAT IT CHANGES FOR CUSTOMERS

- **Agent spend is now a commitment decision.** Until August the only way to buy Gemini Enterprise was a seat. Now there is a commit, and Google's account team will size it. The customer who has no usage baseline is being asked to forecast a workload whose cost per outcome changes with every model release.
- **The savings plan's downside is symmetric and the upside is not.** Commit USD 200,000 a month and use USD 150,000: the bill is USD 200,000. Use USD 250,000: the extra USD 50,000 is billed at on demand rates with no discount. The plan protects Google's revenue floor in both directions and gives the customer 10 or 20 percent in exchange.
- **Three years is a bet on the market, not the workload.** A three year compute commitment assumes the workload stays. A three year Gemini commitment assumes the workload, the model roadmap, Google's pricing structure, and the competitive landscape all stay favorable while model versions turn over every few months. Advisors across the industry have said the same thing: pilot, then one year, then reconsider.
- **The chargeback model no longer matches the bill.** Under project scope a business unit that bought a commitment got the discount. Under billing account scope with proportional attribution the discount is spread across every eligible project on the account. The unit that paid for the commitment sees a higher effective rate; the units that did not see a lower one. Nobody was told.
- **The price step invalidates the sizing.** A savings plan sized on the 2026 Flash rate covers half the tokens in 2027. A plan sized on the 2027 rate is a 100 percent overcommit in 2026. The only correct sizing accounts for the step explicitly, and the account team's model may not.
- **The enterprise agreement is the real contract.** Savings plan spend counts against an enterprise agreement commit. That is useful only if the agreement's eligible services list includes the AI platform, the burn down treats plan spend at full value, and the agreement's own uplift and true up terms have not just been triggered by a new AI commitment layered on top.
- **The buyer is the platform team, not procurement.** Spend caps, deferred execution flags, and the CUD attribution model are configured in the console. The commitment, the term, and the enterprise agreement amendment are signed by procurement. The two decisions are made by different people on different weeks.

3. THE GEMINI ENTERPRISE PRICING LEVERS COMPARED

Lever	Discount	Mechanics	Where the risk sits
Per seat subscription	Fixed monthly	Fixed price per user with a daily quota pool shared across a project. Unchanged.	Paying for idle seats and quota. Predictable but decoupled from value.
Pay as you go	None	No base fee. Tokens, compute, memory, and storage at standard	Full list price. Uncapped unless caps are configured. The right place to start, the wrong place to stay.

		rates. Select customers now, broader rollout undated.	
Flexible Savings Plan, 1 year	10 percent	Monthly spend commitment, no minimum or maximum, applies automatically to eligible usage. Counts against an enterprise agreement.	Shortfall billed in full, no rollover, overage at on demand. Moderate: a twelve month forecast in a market that reprices quarterly.
Flexible Savings Plan, 3 year	20 percent	Same mechanics, thirty six month term.	High: the discount is thin and the term outlasts the model, the price list, and probably the architecture.
Deferred execution	Up to 50 percent on inference	Eligible agent workloads flagged for off peak capacity windows. Select workloads only at launch.	Low financial risk. Engineering cost to make workflows tolerate delay. Best return of the four for batch, evaluation, indexing, and document processing.

For comparison, Compute Engine Flex CUDs run roughly 28 percent for one year and resource based CUDs roughly 37 percent and above. The AI plan's rate reflects how little usage history Google has to price against, and how much it wants full price experimentation revenue in the meantime.

4. THE CUD SCOPE CHANGE IN DETAIL

Element	Before June 16, 2026	After June 16, 2026
Default scope	Project. A commitment covered usage in the project that bought it.	Billing account, with sharing enabled. A commitment covers eligible usage in any project on the account.
Who was switched	Not applicable	All new billing accounts. Existing accounts with no active resource based commitments, automatically. Existing accounts with active commitments, unchanged until they act.
Attribution	To the purchasing project.	Proportional across eligible projects by default. Prioritized attribution available if configured.
Chargeback effect	Discount lands where the commitment was bought.	Discount spreads. The buying unit's effective rate rises; other units' rates fall; total is unchanged.
What did not change	Commitments remain regional and resource specific, cannot be cancelled, and bill for the full term regardless of utilization. GPU and Local SSD commitments generally still require attached reservations.	

The utilization argument cuts both ways. Sharing improves coverage when workloads move between projects, which is Google's stated reason. It also removes the discipline that project scope imposed: a unit that overbought commitments under project scope carried its own shortfall. Under billing account scope the shortfall is absorbed by whichever project happens to have eligible usage, and the signal that someone overcommitted disappears from the report.

5. KEY DATES

Date	Event
January 21, 2026	Spend based CUDs move to the consumption model billing schema. Discounted rates shown directly; list plus credit reporting retired. Eligible usage for some plans widened.
May 1, 2026	A3 Ultra GPU VM pricing increased in Europe and Asia.
June 16, 2026	Default resource based CUD scope changed from project to billing account with sharing enabled. Existing accounts without active commitments switched automatically. CUD Analysis reaches general availability.
June 22, 2026	

	Resource based CUD recommendations extended to GPUs, Local SSD, and premium OS licenses, based on the prior thirty days of usage.
July 1, 2026	Non global endpoint pricing takes effect for Gemini 3 and later model families.
August 26, 2026	Gemini Enterprise pay as you go, Flexible Savings Plans, deferred execution, spend caps, anomaly alerts, and pooled developer tool quota announced. Savings plans live for self serve and enterprise agreement customers.
December 31, 2026	Introductory pricing for Gemini 3.6 Flash and 3.7 Flash ends.
January 1, 2027	Flash standard pricing takes effect at USD 1.50 input and USD 7.50 output per million tokens, double the 2026 rate.
Undated	Broader rollout of pay as you go. Deferred execution beyond select workloads. Whether Flexible Savings Plans cover the pooled Antigravity and Android Studio quota. Confirm each with the account team in writing.

6. WHERE THE MONEY SITS

Unlike an audit, a commitment does not produce a claim. It produces a bill for capacity that was not used, at a discount that was never large enough to justify the risk. The illustration below is an enterprise with a growing Gemini Enterprise and AI platform footprint, offered a three year plan by its account team in the last quarter of 2026. Figures are for structure, not quotation.

Item	Illustrative figures	Effect
Account team's sizing	Run rate of USD 250,000 a month in Q4 2026, projected flat. Three year Flexible Savings Plan at USD 250,000 a month, 20 percent off.	Total committed: USD 9.0M over thirty six months.
What actually happens in year one	Usage holds at USD 250,000 at list. Flash price doubles in January, but the team moves 40 percent of tokens to smaller models and batches evaluations.	Year one spend at list: USD 3.1M. Commitment covers it. Discount earned: about USD 0.6M.
Years two and three	Architecture matures, caching and routing cut tokens per outcome, one agent program is cancelled. List spend settles at USD 150,000 a month.	Spend at list: USD 3.6M across both years. Commitment still bills USD 6.0M. Shortfall of USD 2.4M paid for nothing; no rollover.
Overage months	Two launch months in year one exceed the commit by USD 60,000 each.	Billed at on demand rates, no discount. The plan does not cover the peaks it was sold to smooth.
CUD scope side effect	The AI platform team's Compute Engine commitments, bought under project scope, were switched to billing account scope in June. Two other units now draw on them.	The AI team's effective compute rate rises 8 to 12 percent; the units it subsidizes report savings they did not buy. Finance cannot reconcile either number to a commitment.
Three year cost of the plan as sold	**USD 9.0M paid against USD 6.7M of usage at list. Pay as you go would have cost USD 6.7M. The commitment cost USD 2.3M more than no commitment, in exchange for a discount worth USD 1.3M if usage had held.**	
Same estate after a defended approach	**Six months on pay as you go with spend caps to establish the baseline. One year plan sized at 70 percent of the observed run rate after the January price step, renewed annually. Batch, evaluation, and indexing workloads on deferred execution at up to half price. CUD attribution set to prioritized so the buying unit**	

	keeps its discount. Roughly USD 5.0M to 5.5M over three years, with the variable cost capped and no forfeited commitment.

- **The forecast is the whole risk.** A spend plan is only as good as the run rate it was sized on. Agent workloads in their first year are the least stable spend line on the bill: they grow with adoption, shrink with optimization, and reprice with every model release.
- **The discount is not the point.** Ten percent on a workload you forecast correctly is worth less than the shortfall on a workload you forecast wrong by a quarter. The break even on a one year plan is roughly 90 percent utilization; on a three year plan the forecast horizon makes the number meaningless.
- **Deferred execution is the lever with real money in it.** Half off inference on workloads that can wait, with no term and no commitment, outperforms both savings plans for any estate with meaningful batch or evaluation volume. It costs engineering time, not commitment risk.
- **The scope change is a governance event.** Nothing about it costs more in total. Everything about it changes who appears to pay, which is what every internal budget, showback, and commitment approval was built on.

7. HOW CUSTOMERS SHOULD RESPOND

The response is a sequence. Its purpose is to size any commitment on evidence rather than a sales forecast, to keep the term short while the market reprices, and to put the attribution model, the plan terms, and the enterprise agreement treatment in writing before the next commitment is signed.

- **1. Establish the baseline on pay as you go.** Request pay as you go access if not already granted. Run Gemini Enterprise for at least one full quarter with spend caps and anomaly alerts configured. Export usage by model, by agent, and by project. This is the only forecast worth committing against.
- **2. Reprice the baseline for January 1.** Apply the published Flash step to the observed token mix. Identify which workloads can move to smaller models before the step. Size any commitment on the post step, post optimization run rate, not the Q4 2026 bill.
- **3. Route batch work to deferred execution first.** Evaluations, indexing, document processing, summarization, and code analysis belong on the half price lever, not inside a commitment. Get the eligible workload list from the account team in writing and re architect the obvious candidates.
- **4. Take one year, not three.** Size the plan at 70 to 80 percent of the repriced run rate. Renew annually. Decline the three year term until the workload, the model mix, and the price list have been stable for a year, which has not yet happened for anyone.
- **5. Audit the CUD scope and set attribution deliberately.** Check every billing account for the June 16 switch. Decide whether proportional or prioritized attribution matches the chargeback model. Confirm which projects are eligible for shared coverage. Document the decision with finance before buying another commitment. Use CUD Analysis to test whether existing commitments are still matched to workload.
- **6. Reconcile the billing schema.** Update finance models and showback reports for the consumption model schema. A report that still expects list plus credit lines will understate or misattribute savings.
- **7. Fix the enterprise agreement treatment in writing.** Confirm that Flexible Savings Plan spend counts toward the agreement commit at full value, that the AI platform services are on the eligible list, and that adding an AI commitment does not trigger an uplift, a true up, or a term extension elsewhere in the agreement. Ask for price protection on the model rates the plan will consume.
- **8. Negotiate, do not configure.** The plan has no stated minimum or maximum. That makes the commitment amount, the term, the rate, and the treatment of overage negotiable. Take them into the enterprise agreement renewal on your calendar, with the pay as you go baseline and the deferred execution savings as leverage, not at Google's quarter end.

8. MISTAKES THAT COST THE MOST

- Signing a three year savings plan in Q4 2026 on a run rate that includes launch month peaks and 2026 Flash prices.

- Letting the account team size the commitment. Their forecast is a quota, not a model of your workload.
- Discovering the CUD scope change from a business unit's budget variance rather than from a billing account audit in June.
- Leaving attribution on proportional because nobody knew there was a choice, and defending the chargeback to the unit that bought the commitment.
- Putting batch and evaluation workloads inside a savings plan when deferred execution would have taken half off with no term.
- Assuming savings plan spend burns down the enterprise agreement without checking the eligible services list or the uplift clause.
- Treating spend caps as the control. Caps stop the bill. They do not stop the commitment.

9. HOW REDRESS HELPS

Redress Compliance runs the exercise from the buyer's side: the pay as you go baseline and workload classification, the repricing for the January step, the deferred execution candidate list, the CUD scope and attribution audit with finance, the enterprise agreement read, and the negotiation of the plan amount, term, rate, overage treatment, and price protection inside the client's Google Cloud agreement strategy. Our fees are fixed, and every dollar kept out of a forfeited commitment belongs to the client.

REDRESS COMPLIANCE LLC

1314 E Las Olas Blvd, Fort Lauderdale, FL 33301 | redresscompliance.com | +1 (239) 402-7397

100 percent buyer side. Zero vendor affiliations. No reseller agreements. No referral fees.