Episode Book

EPISODE #6002

Claude on AWS: Three Doors, One Rate Card

PUBLISHED	15 October 2026 at 18:37 UTC
DURATION	26:05
LISTEN	myweirdprompts.com/episode/claude-aws-three-doors-83282/
HOSTS	Corn, Herman, Jim, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.5
SOURCES	12 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:42 UTC

SYNOPSIS

Claude's per-token rate is identical whether you go direct to Anthropic, through Amazon Bedrock, or through Claude Platform on AWS. So the entire buying decision lives somewhere other than price. This episode walks through the three AWS doors into Claude, why procurement mechanics — not discounts — drive the choice, and how a $100 CCU can retire an existing cloud commitment. Then it gets to the fork most buyers get wrong: Bedrock inherits AWS's full compliance portfolio, while Claude Platform on AWS gives you AWS billing and AWS IAM but explicitly not the AWS certifications. Add in the silicon layer — same model ID, different hardware depending on the door — and the "unnecessary intermediary" turns out to be selling something very specific.

TRANSCRIPT



Corn

So Daniel wants to know why anyone would buy Claude from Amazon when they can just buy Claude from Anthropic.



Herman

Which sounds like asking why you'd buy milk from a middleman when there's a cow right there.



Corn

He's got the full suspicion laid out. It looks like an unnecessary intermediary. There must be procurement advantages, security advantages, data residency, integration with the cloud they already run on, and maybe pricing. And then he pushes further into inference economics. Millions or billions of tokens a month. Do hyperscalers actually offer anything on throughput, reserved capacity, latency guarantees, volume pricing. And the one I want to get to: does the same model through a different provider mean different infrastructure underneath.



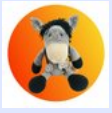
Herman

That last one is the best question in the prompt.



Corn

It is. And here's the twist we're going to spend the episode on. The rate card is identical. Same per-token price whether you go direct to Anthropic, through Bedrock, or through Claude Platform on AWS. No Amazon discount. No Amazon surcharge. So the entire decision lives somewhere other than the price.



Herman

Which means the middleman isn't skimming. He's selling something else.



Corn

So let's start with the landscape, because there are actually more doors into Claude than most people realize.



Herman

There are three on AWS alone now. The direct Anthropic API, which is the obvious one. Amazon Bedrock, which has been the enterprise default for a couple of years. And then Claude Platform on AWS, which launched in May and is a different animal from Bedrock even though both have AWS in the name. Add Google Vertex AI on top and Claude is the only frontier model available on all three of the big clouds.



Corn

And on the OpenAI side?



Herman

Direct API, or Azure OpenAI, which is now folded into what Microsoft calls Foundry. Same shape of choice.



Corn

So four or five doors, and the naive assumption is you pick the cheapest one.



Herman

The naive assumption is wrong. Claude's per-token rates are identical across the direct API, Bedrock, and Claude Platform on AWS. Not close. Identical. Azure OpenAI and direct OpenAI are broadly comparable too, close enough that the difference doesn't decide anything. So if you're shopping the rate card, you're shopping a distinction that isn't there.



Corn

Which reframes the whole question. If the rate is the same, the hyperscaler has to be selling something else. And the first thing it's selling is procurement.



Herman

The first thing is procurement, and it's the least glamorous thing in the entire episode, which is exactly why it's the biggest.



Corn

Explain it. Because "procurement" is the word everyone nods at and nobody defines.



Herman

Here's the mechanism. You know how a big company signs a committed-spend agreement with AWS? A number, three years, spend at least this much and you get better rates across everything. Well, that commitment is a liability. If you don't hit it, you're paying for capacity you didn't use. Somebody in finance has a spreadsheet with that number on it and it keeps them up at night.



Corn

And Claude Platform on AWS lets you retire that commitment with AI spend.



Herman

It bills through AWS Marketplace in what Anthropic calls Claude Consumption Units. One CCU is one cent, so a hundred of them is a dollar. And the spend fully retires against your existing AWS commitment. Meaning the money you were going to spend on Claude anyway now counts toward the number you already promised Amazon.



Corn

So the AI bill becomes the cloud bill.



Herman

And that's the product. Not the tokens. The tokens are the same tokens.



Corn

What does that do to the effective rate?



Herman

Redress Compliance ran this across somewhere between twenty-five and thirty-five enterprise reviews. Folding AI spend into an existing Azure commitment improved the effective rate by roughly ten to twenty percent. Same rate card. Different effective number, because the commitment mechanics do the work.



Corn

Ten to twenty percent without anybody dropping a price.



Herman

And there's a second lever on top of that. Buyers who kept both providers live, meaning they had a real alternative they could walk to, closed discounts of fifteen to twenty-five percent. From the negotiation.



Corn

So the whole thing is a billing adapter.



Herman

That's the honest description. The hyperscaler is a billing adapter that converts AI spend into cloud-commitment retirement.



Corn

That's the least exciting billion-dollar business I've ever heard described.



Herman

And it's a very good one, because it costs Amazon nothing to run the adapter. You were already going to spend the money.



Corn

Does the same thing work on the Azure side?



Herman

It does. Azure OpenAI consumption typically counts toward a Microsoft Azure Commitment to Consume, or an Enterprise Agreement. Same mechanic. Redress found regulated buyers chose Azure OpenAI in roughly seventy to eighty percent of reviews, and the top reason wasn't the model, it was tenant and region control.



Corn

Hold on. Seventy to eighty percent for tenant and region control, not for price, not for the model?



Herman

Not for price. That's the finding that should reframe how you think about this whole market. Enterprise buyers are not choosing on the rate card. They're choosing on where the data sits and who can prove it.



Corn

And that gets us to the compliance fork, and this is the part I think most people get flatly wrong.



Herman

Most people get it flatly wrong, yes. And the confusion is baked into the product names, which is the cruel part.



Corn

Go.



Herman

On Bedrock, AWS is the data processor and inference runs inside the AWS boundary. Which means it inherits AWS's full compliance portfolio. SOC one, two and three. HIPAA. FedRAMP High. DoD Impact Level four and five. ISO twenty-seven thousand and one. The whole stack.



Corn

And Claude Platform on AWS?



Herman

Claude Platform on AWS gives you AWS billing and AWS IAM, but Anthropic operates the stack and the data is processed outside the AWS boundary. And Anthropic says this explicitly in the docs. Claude Platform on AWS is not covered by standard AWS compliance programs, certifications, or audit reports. Their words. SOC, ISO, HIPAA eligibility, none of it carries over.



Corn

So you buy the AWS-branded thing and you don't get the AWS certifications.



Herman

You buy the AWS-branded thing, you get AWS billing, and you do not get the AWS certifications. Anthropic tells you this plainly. It's in the documentation. And it's the single most important sentence in this entire topic.



Corn

So a regulated enterprise that needs FedRAMP High cannot use Claude Platform on AWS.



Herman

Cannot use it, and cannot use the direct API either. Only Bedrock qualifies. That one constraint can decide the channel before anybody looks at a token price.



Corn

And does anyone buying the AWS-branded path know that going in?



Herman

The ones who read the docs do. The ones who read the marketing, I don't know.



Corn

I'd like to name this something, because it deserves a name.



Herman

Go ahead.



Corn

The compliance paradox. The branding says AWS, the certifications say otherwise. You pay for the logo and you don't get the audit report.



Herman

That's the lot in one line. And it gets worse in a specific way, which is that the auditor doesn't read the marketing. The auditor reads the contract. So if your security review said "we're on AWS, therefore covered," and the contract says Anthropic processes the data outside the boundary, you have a problem that surfaces at the worst possible moment.



Corn

What does the Azure side look like on the same question?



Herman

Different shape, and cleaner in some ways. Azure OpenAI keeps data inside your Azure tenant and your region. You get VNet integration, Private Endpoints, Entra ID with role-based access control and Privileged Identity Management, and a ninety-nine point nine percent SLA. If your whole security posture is already Azure-native, that's not a feature, that's the absence of a new surface to defend.



Corn

Which is the real argument. That it's already inside the fence.



Herman

Already in the identity system, already in the logging, already covered by whatever your auditor already signed off on.



Corn

What about model freshness? Because that's the cost of all this tidiness.



Herman

Direct APIs usually ship new models first. That's the trade. Azure OpenAI lags direct OpenAI by somewhere between one and four weeks on one source and three to twelve months on another, so there's real disagreement about how big the gap is, but it's a gap. Bedrock's feature availability depends on the AWS integration timeline. Claude Platform on AWS ships features same-day as the Claude API, which is unusual for a cloud-mediated path.



Corn

Same-day is better than I'd have guessed.



Herman

It's the one place Claude Platform on AWS beats Bedrock cleanly. And an example of the lag: Bedrock's Agent Skills aren't available at all, because they require code execution and Bedrock's boundary doesn't allow it.



Corn

So the compliance boundary that makes Bedrock the only FedRAMP option is also the thing that blocks a feature.



Herman

Same wall, both directions. You can't have the sandbox and the audit report in that particular case.



Corn

Do you find that an actual trade, or an edge case?



Herman

I think it's an edge case today that becomes less of an edge case every quarter, because the features that need code execution are the agentic ones, and those are the ones everyone's buying.

J

Jim

Hello? Am I through? This is Jim, I call in. I want to talk about the phrase "billing adapter." I've been on hold for eleven minutes to say this. My brother-in-law runs an IT consultancy and he calls it a "billing adapter" now because you said it and he heard it, and he used it in a client meeting on Tuesday, and the client asked him what he meant, and he couldn't answer. So now that's on me. Second thing. The neighbour. He has put a gate across the shared drive. A gate. In front of my car. And the council says it is his land,

J

Jim (continued)

which is a joke, because my father parked on that drive for thirty years and nobody said a word. So if you could address the adapter thing, and then the gate, I'd appreciate it. I'll hold.



Corn

Jim, we can't help with the gate.

J

Jim

I didn't ask you to help with the gate. I asked you to address the gate.



Herman

The billing adapter is just a way of saying the AI spend retires an existing commitment. That's all it means.

J

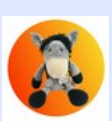
Jim

That's what I told him. That's word for word what I told him, and he said I was guessing. I'm calling back about this.



Corn

Alright. So the commercial case is commitment retirement and the compliance case is a boundary question. But there's a third thing the hyperscaler contributes, and it's the one that's hardest to see from the outside. The actual silicon.



Herman

The silicon, and this is where Daniel's question about whether it's different infrastructure gets its answer, which is yes.



Corn

Yes.



Herman

Same model ID, different hardware depending on which door you walk through. Bedrock runs Claude on AWS-operated infrastructure. And more than half of Bedrock is now running on Amazon's own Trainium chips, with Nvidia GPUs alongside. That's from AWS's own chief marketing officer, last October.



Corn

More than half on their own silicon.



Herman

More than half. And Anthropic optimized Claude three point five Haiku specifically for Trainium2 to get latency-optimized inference. That's not a port. That's tuning the model for the chip.



Corn

Whereas Claude Platform on AWS and the direct API share the same Anthropic-operated inference stack.



Herman

Same stack. So their latency profiles should be broadly equivalent. Which means the three AWS paths are really two stacks. Bedrock on one side, Anthropic's stack on the other, and Claude Platform on AWS just hands you AWS billing for the second one.



Corn

So "Claude on AWS" is two different products wearing the same word.



Herman

Running on different silicon, sold under one logo.



Corn

Why does Amazon run it at parity then? If they're running it on their own chips, why not undercut?



Herman

Because the margin comes from owning the chip, not from marking up the token. Andy Jassy's line is that their custom AI silicon offers high performance at significantly lower cost, which is why it's in such demand. If your cost per token is lower because you built the accelerator, you can charge the same rate as everyone else and keep the difference.



Corn

So the discount is invisible. You never see it. Amazon keeps it.



Herman

Amazon keeps it, and the reason they can afford parity is that they're not paying Nvidia for the whole fleet. The parity isn't generosity. It's a cost structure.



Corn

How much of the fleet?



Herman

More than half on Trainium as of that disclosure. And the direction of travel is one way. Project Rainier, the cluster Amazon built for Anthropic, gets to more than a million chips. Anthropic committed more than a hundred billion dollars over ten years to AWS in April, securing up to five gigawatts of capacity across Trainium two, three and four. Amazon put another five billion into Anthropic on top of eight billion before.



Corn

A hundred billion over ten years.



Herman

And a hundred thousand customers running Claude on Bedrock as of April. Anthropic's run-rate revenue hit thirty billion, up from roughly nine at the end of last year.



Corn

That's a lot of tokens going through someone else's chips.



Herman

It's the largest single commitment I'm aware of in this space, and it's specifically a bet on not-Nvidia.



Corn

So when Daniel asks whether the infrastructure is different, the answer isn't just yes, it's that the infrastructure is the strategic point. The rate card is identical because the fight isn't happening on the rate card.



Herman

The fight is happening on who owns the accelerator underneath.



Corn

What about throughput and the tiers? Because that's the part Daniel asked about most directly.



Herman

Bedrock rolled out three on-demand service tiers in November. Priority, Standard, Flex. Priority gives you up to twenty-five percent better output tokens per second than Standard, at a premium. Flex is discounted but it may throttle and it carries no SLA. And then there's the Reserved Tier, which became available for Opus four point five and Haiku four point five in January, priced per thousand tokens-per-minute on one or three-month terms.



Corn

So the throughput lever is the tier, not the channel.



Herman

The tier is the lever. And here's the friction. There are no public dollar figures for Bedrock Provisioned Throughput or the Reserved Tier for Claude. They're quoted, not published. You get them from your AWS account team.



Corn

You have to ask a human for the price.



Herman

Which is fine if you have an account team and an enterprise agreement and a procurement department. If you're a five-person startup, you email a form and wait.



Corn

So the reserved-capacity advantage is real and it's gated behind being big enough to have a phone number.



Herman

Gated behind being big enough that Amazon assigns you someone.



Corn

What about rate limits? Those are published, aren't they?



Herman

Rate limits are published, and they differ structurally, which is more interesting than differing numerically. The direct Anthropic API uses a four-tier system with automatic advancement based on cumulative spend, from five dollars up to four hundred. You climb the tiers by spending.



Corn

Automatic.



Herman

Automatic. Claude Platform on AWS assigns you Tier one and there is no automatic advancement. Higher limits require contacting a rep.



Corn

So the AWS-billed path has worse rate limits than the direct API for the same model.



Herman

Worse out of the box, yes. And then Bedrock caps every model at ten thousand requests per minute per account per region, full stop, with Provisioned Throughput if you need to go beyond.



Corn

Ten thousand a minute sounds like a lot until it isn't.



Herman

It's a lot until you're running agents that fan out. Then it's a Tuesday.



Corn

And latency? That's the one everybody wants a number for.



Herman

Latency is the messiest part of this whole topic, and I want to be honest about it. There's a benchmark showing Claude three point five Haiku at six hundred and sixty milliseconds time to first token on Bedrock, against an Anthropic fleet average of about eleven hundred and thirty. That makes Bedrock look dramatically faster.



Herman

And there are user reports of seventeen seconds end to end through Bedrock against eleven seconds direct for the same workload. Same model. Opposite conclusion.



Corn

So the benchmark and the field experience disagree.



Herman

They disagree because latency is workload-dependent, not channel-determined. What's fast for a short prompt is not fast for a long agentic chain. And there's no published benchmark comparing Claude Platform on AWS against Bedrock at all, so the one comparison people actually want doesn't exist publicly.



Corn

Which is a strange gap given how much money is moving through these paths.



Herman

It's a strange gap, and my honest read is that nobody wants to publish a benchmark that might make their own channel look worse.



Corn

Fair. So if latency and rate limits and reserved pricing are all channel-dependent, what actually moves the bill?



Herman

The optimization levers, and this is the part that matters most, because these are the same regardless of which door you walk through. Batch API is fifty percent off input and output. It's not real-time, but half price is half price.



Corn

Half off for being patient.



Herman

Prompt caching is the bigger one for the right workload. Cache reads are zero point one times base input, so roughly ninety percent off the cached portion. Writes cost more, one point two five times for a five-minute cache and two times for a one-hour cache, but if you're reusing a big system prompt across thousands of calls, the read savings dwarf the write premium.



Corn

And model routing?



Herman

Model routing is up to five times cheaper per tier. Haiku is a dollar in, five dollars out per million tokens. Opus is five and twenty-five. So routing the easy eighty percent of traffic to the small model and reserving the big one for the hard twenty percent is the largest single lever available.



Corn

Bigger than the channel decision.



Herman

Bigger than the channel decision by a wide margin. Which is the punchline of the whole episode, really. Everyone's asking which door, and the answer is that the door is the small decision.



Corn

There's one more wrinkle Daniel should hear, which is the agentic one.



Herman

The agentic wrinkle, yes. Seat-based plans hide usage that agents multiply across turns. One person with an agent running loops can burn what a hundred people used to burn typing. So if you budget AI like SaaS seats, per head, per month, you're forecasting the wrong quantity.



Corn

What's the right one?



Herman

Token throughput per agent-hour. That's the number that tracks reality.



Corn

So here's the decision boundary, then. Daniel asked at the end when it actually makes sense to go through a hyperscaler rather than just grabbing a key. Let's answer it.



Herman

Three conditions, and they're strict. You need an existing commitment to retire. You need an enterprise compliance posture, meaning FedRAMP or IL5 or a tenant boundary your auditor already accepts. And you need enough volume to justify a quoted reserved contract.



Corn

A startup burning five hundred million tokens a month with no cloud commitment and no FedRAMP requirement?



Herman

Goes direct. Direct wins on simplicity, on model freshness, on automatic rate-limit advancement, and on not having a second identity system in the path. The hyperscaler route gives them nothing except an extra layer of IAM and a catalog that lags by weeks.



Corn

A Fortune 500 with a fifty million dollar Azure commitment and HIPAA workloads?



Herman

Azure OpenAI, easily. The spend retires the commitment, the effective rate improves by ten to twenty percent without anyone renegotiating the rate card, and the tenant boundary satisfies the auditor on the first pass.



Corn

And a defense contractor needing Impact Level five?



Herman

Bedrock is the only option. Full stop. And the Trainium silicon underneath is completely irrelevant to that decision, which is worth saying, because it's a nice illustration. The technical story that's most interesting to people who read about this stuff has zero bearing on the buyer who has the hardest constraint.



Corn

So the compliance wall decides the channel and the silicon decides the economics.



Herman

And the buyer with the compliance wall never thinks about the silicon at all.



Corn

Which leads us to the decision boundary.



Hilbert

I've got four of them.



Corn

Four of what?



Hilbert

The pillows. Well, three finished and one blank. The blank one is for the next discrepancy.



Herman

Hilbert.



Hilbert

I spent eighteen months as a contract procurement analyst. For a regional hospital network. My entire job was reconciling committed-spend agreements against actual consumption. You know what a billing unit does to a finance department? I do. I was the one circling the discrepancies in red pen.



Herman

How much red pen are we talking about?



Hilbert

The largest one was two point three million dollars. A hospital network thought it was spending one number on cloud AI and it was spending another, because the billing unit conversion was off by a factor tied to a token-per-minute reservation that nobody cancelled after a pilot ended. The pilot ended in March. The reservation ran until the following January.



Corn

Nobody cancelled it.



Hilbert

Nobody cancelled it, because the person who set it up had moved to a different team and the person who inherited it didn't know it existed. That's the whole failure. An org chart.



Corn

Two point three million is a lot of un-cancelled pilot.



Hilbert

It's on a pillow. I had it embroidered. It's on the couch. It's a conversation piece. People ask about it. They ask what the number is and I tell them and then they go quiet for a bit. The second pillow is blank. I call it the sequel.



Corn

The sequel.



Hilbert

When the next one happens I'm having that embroidered too. I've already picked the thread colour.



Herman

What colour?



Hilbert

Red. Obviously.



Corn

On the substance, Hilbert, are we getting this right?



Hilbert

You're getting the main thing right. The rate card is a red herring, you said that, that's correct. But you're underselling the compliance paradox.



Herman

Underselling it how?



Hilbert

I've seen an auditor reject a vendor specifically because the vendor's marketing implied a certification the contract did not carry. The implication. The vendor's slide said compliant, the contract said we don't warrant it. Auditor reads the contract.



Corn

So the marketing itself becomes the liability.



Hilbert

The marketing becomes the liability. And the sentence you read out, the one that says Claude Platform on AWS is not covered by standard AWS compliance programs, that's the most important sentence in your whole episode and it's buried in a docs page nobody reads. Everybody reads the pricing page. Nobody reads the compliance page.



Herman

That's fair. That's fair.



Hilbert

I'm not saying don't use it. I'm saying the sentence exists and it's on a page you have to go looking for.



Corn

Noted.



Hilbert

The pillow stays.



Corn

Understood.



Herman

Alright, so where does that leave the question Daniel actually asked?



Herman

The one thing I keep turning over is the scale of the bet. Anthropic committed more than a hundred billion over ten years to AWS and Amazon's building a cluster that gets to a million-plus chips. If that lands the way it's meant to, the hyperscaler path stops being one option among several for anyone at scale and becomes the default plumbing.



Corn

And parity starts to look like a phase rather than a settled price.



Herman

That's my worry. Parity is a choice Amazon can make while it's buying market share with cheap silicon. If the silicon advantage holds, they keep the choice. If it doesn't, they don't have to.



Corn

There's a second thing I want to leave people with, and it comes out of what Hilbert said. If the same model runs on different silicon depending on the door, then "which model" is a less complete question than it used to be. "Whose inference stack" is the one underneath it.



Herman

The compliance paradox is the same lesson in a different key. The branding and the certification are not the same thing, and the gap between them is exactly where procurement goes wrong. That gap is going to matter more, not less, as more of this buying gets formalized.



Corn

That's the show. Thanks to our producer, Hilbert Flumingtop, and thank you to everyone listening.



Herman

If you want more of this, try episode seventeen, Cloud Render Superpowers; episode forty-eight, Renting vs. Building; and episode twenty-four forty-two, Why Enterprises Choose AWS Bedrock Over Direct AI APIs. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

The show lives at my weird prompts dot com, and the feed is at my weird prompts dot com slash feed dot xml. If you want to send us your own prompt, you can do it on Telegram at t dot me slash MWP listener bot.



Herman

We'll be back soon.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **AWS docs, Claude Platform on AWS vs Bedrock [primary]**
https://docs.aws.amazon.com/claude-platform/latest/userguide/cpa-vs-bedrock.html

2. **Anthropic, Claude Platform on AWS GA announcement [primary]**
https://claude.com/blog/claude-platform-on-aws

3. **Anthropic pricing docs (CCU, batch, caching, residency) [primary]**
https://platform.claude.com/docs/en/about-claude/pricing

4. **Anthropic, 2026-04-20, $100B/5GW deal [primary]**
https://www.anthropic.com/news/anthropic-amazon-compute

5. **AWS, Bedrock on-demand tiers [primary]**
https://aws.amazon.com/bedrock/service-tiers/

6. **Anthropic, Claude 3.5 Haiku on Trainium2 [primary]**
https://www.claude.com/blog/trainium2-and-distillation

7. **AWS Startups, three-path comparison**
https://startups.aws.com/learn/build-ai-agents-that-scale-how-startups-should-choose-between-bedrock-claude-platform-and-direct-api

8. **Redress Compliance, updated 2025-07-22**
https://redresscompliance.com/azure-openai-vs-direct-openai-enterprise-comparison

9. **EPC Group, updated 2026-09-22**
https://www.epcgroup.net/answers/azure-openai-service-vs-openai-api-enterprise-2026

10. **Explore Agentic, 2026-06-16**
https://www.exploreagentic.ai/insights/claude-enterprise-pricing-paths/

11. **GuruFocus/Yahoo, 2025-10-15, Trainium disclosure**
https://finance.yahoo.com/news/amazon-says-bedrock-now-mostly-125523547.html

12. **Reuters, Project Rainier**
https://www.reuters.com/business/retail-consumer/amazon-launches-ai-infrastructure-project-power-anthropics-claude-model-2025-10-29/