Episode Book

EPISODE #6369

The Most Popular GPUs - Today's Cloud Demand

PUBLISHED	30 October 2026 at 12:46 UTC
DURATION	23:51
LISTEN	myweirdprompts.com/episode/cloud-gpu-rental-guide-67449/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.25
BOOK GENERATED	31 October 2026 at 05:58 UTC

SYNOPSIS

Cloud GPU pricing depends far more on deployment models and memory fit than raw benchmark speeds. A look across real rental options—from 24GB workhorses like the L4, A10G, and RTX 4090 to 48GB and 80GB accelerators like the L40S, A100, and H100—reveals where speech recognition, image generation, and quantized language models actually run. The trade-offs between dedicated instances, community pods, serverless scale-to-zero, and multi-device communication decide what makes financial sense.

TRANSCRIPT



Corn

Would the busiest GPU in the cloud even make the front page of a benchmark chart? Daniel wants the cards people actually pay to use, running inference, fine-tuning models, making images and video. We're going from the smallest memory allocation to the largest, with rough hourly prices and one AMD option among the Nvidia machinery.



Herman

The revealing purchase is often a fairly ordinary card doing a fairly ordinary job continuously. Speech transcription, generating product pictures, serving a smaller language model. Nobody writes a triumphant announcement about finishing another batch of subtitles.



Corn

The subtitles remain unimpressed.



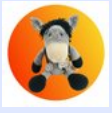
Herman

They do, although missing the last sentence tends to attract feedback. I'm Herman Poppleberry, and the useful distinction here is between a card's maximum performance and the work someone can economically keep feeding it. Memory capacity decides whether a job starts. Memory bandwidth and the software decide how efficiently it runs.



Corn

You're already leaning toward bandwidth. I can hear the specification sheet opening.



Herman

You'll get your prices. You're going to defend the older card as soon as it works reliably.



Corn

If it finishes without asking me to rebuild its environment, yes.



Herman

Our selection is an informed workhorse list, not a measured popularity league. There's no public source ranking these cards by hours actually rented. We're inferring demand from platform listings, scarcity and price movements. A widely offered card with persistent demand belongs in the conversation even if it has stopped winning headlines.



Corn

Availability also has two meanings. Listed on a website and available when your job needs to run. A price beside an unavailable machine is decorative.



Herman

That distinction matters especially near the top. The order begins with L4, A10G and RTX forty ninety, all with twenty-four gigabytes of memory.



Corn

Then the L forty S at forty-eight gigabytes. After that, the eighty-gigabyte A one hundred and H one hundred. We're keeping the memory variants explicit, because renting the wrong A one hundred is an unusually expensive reading exercise.



Herman

Next comes RTX PRO six thousand Blackwell, with ninety-six gigabytes, followed by H two hundred with a hundred and forty-one.



Corn

B two hundred follows, with roughly a hundred and eighty to a hundred and ninety-two gigabytes depending on the listing. AMD's MI three hundred X closes the list at a hundred and ninety-two.



Herman

A few exclusions explain the shape. T four remains rentable, but it’s the legacy sixteen-gigabyte option. Modal still lists it at fifty-nine cents an hour. Excluding it doesn’t mean nobody uses it.



Corn

Retirement in cloud computing means remaining on the price page.



Herman

RTX fifty ninety has a narrower rental footprint, concentrated around RunPod and Vast. AMD’s MI three twenty-five X and MI three fifty-five X have limited availability. B three hundred offers two hundred and eighty-eight gigabytes, but it’s too new for this particular workhorse selection.



Corn

There’s a temptation to replace everything old with whatever has the largest number. That would give Daniel a launch calendar rather than a rental guide.



Herman

Prices are an as-of-now snapshot, with per-second serverless rates expressed as hourly equivalents. Dedicated rentals, advertised floors and serverless workers won’t always buy the same service. We’ll distinguish those rather than pretending every dollar amount is interchangeable.



Corn

Especially when the cheapest number assumes you’re a different customer.



Herman

L4 is our first proper stop. Twenty-four gigabytes of GDDR six, on Nvidia’s Ada architecture. Its memory bandwidth is about three hundred gigabytes per second. That’s modest beside the giant accelerators, but the useful question is what you can keep



Herman (continued)

resident.



Corn

An eight-billion-parameter model in sixteen-bit precision uses roughly sixteen gigabytes just for its weights. That leaves room, but not a luxurious amount, once the runtime and conversation cache arrive. People hear “fits” and imagine there’s a spare bedroom.



Herman

Longer conversations and simultaneous users consume the remainder. A quantised thirteen-billion-parameter model can be practical, too. L4 also handles Whisper transcription, other speech recognition, speech synthesis, SDXL image generation and suitably quantised Flux workloads. Small-model LoRA training belongs here.



Corn

I like that speech gets a seat at this table. A transcription service doesn’t become more useful because its accelerator could theoretically generate a cinematic explosion.



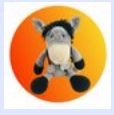
Herman

Would you put your transcription queue on the cheapest card that meets the deadline, Corn, or pay extra to clear it immediately?



Corn

Cheapest that meets the deadline. For a recording uploaded overnight, instant results are an indulgence. For live captions, latency is part of the product. Those are different purchases even if both invoices say transcription.



Herman

That’s why a small card can do serious commercial work. L4 is about eighty cents an hour on Modal and seventy-five cents on RunPod serverless. Baseten’s dedicated listing is around eighty-five cents. The headline belongs below a dollar, with service conditions



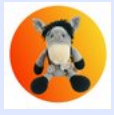
Herman (continued)

attached.



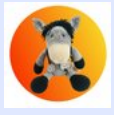
Corn

Less than a dollar to transcribe Herman. Storage billed separately.



Herman

I provide excellent training data for punctuation.



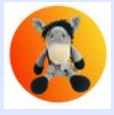
Herman

A10G is the next twenty-four-gigabyte option. It's older Ampere hardware, with roughly six hundred gigabytes per second of memory bandwidth. Same capacity, substantially more bandwidth than L4, though that alone doesn't determine which finishes your workload first.



Corn

The architectural details and kernel support can undo an apparently obvious comparison. I'd expect A10G to feel quite comfortable with established speech and image stacks, rather than needing someone to optimise a brand-new path.



Herman

Mature software is a large part of its appeal. The workload envelope overlaps L4: smaller language models, speech, image generation and modest adapter training. Modal lists its A ten at about a dollar ten an hour. Baseten's A10G dedicated rate is about a dollar twenty-one.



Corn

More expensive than L4 despite being older. Birthdays don't set rental prices.



Herman

Supply, demand and the surrounding service matter. Next is RTX forty ninety, also twenty-four gigabytes, but with GDDR six X and around a terabyte per second of memory bandwidth.



Corn

The gaming card has entered the server room sideways.



Herman

It’s excellent value for image generation and LoRA work when the software fits. You see it on RunPod and Vast much more readily than on heavily managed platforms such as Modal or Baseten.



Corn

Consumer hardware can be quick without being a convenient building block for a managed fleet. Cooling, physical packaging, support expectations, consistent deployment. The customer rents a worker; the operator has to live with the machine.



Herman

Fleet consistency is the less photogenic specification. Now the price split: RunPod serverless is about a dollar thirty-four an hour. A RunPod community pod is about thirty-four cents.



Corn

One dollar for making the off switch somebody else’s problem.



Herman

That comparison is nearly fourfold, larger than the more typical serverless premium of roughly one and a half to twice a pod’s rate. Community and managed offerings also differ. Still, scale-to-zero is valuable if your workload arrives in bursts.



Corn

Provided it really scales down between them. A cheap continuously rented card can win for a queue that never empties. A more expensive hourly worker can win when it runs briefly and disappears. I'd compare the bill for a day of actual requests before declaring either bargain.



Herman

Cold starts complicate the calculation. Downloading weights and initialising the runtime can take longer than the useful generation. Keeping a worker warm improves responsiveness, but then you've retained capacity through the quiet period.



Corn

Your worker naps on the clock. Finally, relatable infrastructure.



Herman

You object to milliseconds until the loading screen appears on your own device.



Corn

I object to being told it's ready before it is.



Herman

L forty S doubles the capacity to forty-eight gigabytes. Ada architecture again, GDDR six memory, and about eight hundred and sixty-four gigabytes per second of bandwidth.



Corn

This is the card I'd expect to rescue people from elaborate compromises. Enough memory to stop trimming every setting, without immediately paying for the largest accelerator.



Herman

It's a strong middle tier. A seventy-billion-parameter model at four-bit weight precision can fit, subject to runtime overhead and cache requirements. Around thirty billion parameters at eight-bit precision is another useful neighbourhood. Image and video generation also benefit from the extra headroom.



Corn

Video is very good at turning headroom into a historical memory.



Herman

Resolution, frame count and the model's architecture all matter. The L forty S rate is about a dollar ninety-five an hour on Modal, or a dollar seventy-five on RunPod serverless. Replicate lists roughly three dollars fifty-one.



Corn

That spread is large enough that "which GPU?" can't finish the purchasing decision. Still, I'd test this tier before designing a multi-card arrangement. Once communication between devices enters the job, you've bought another engineering problem.



Herman

You may still need multiple cards for speed or capacity, but don't assume that buying more devices is simpler than using one larger one. An adapter fine-tune that fits comfortably here can be a very different proposition from training every weight.



Corn

LoRA keeps being described as "training the model" in casual conversation. Fair enough, but the memory bill knows the difference.



Herman

We've been using those model sizes rather freely. Franklin has a call about the hardware underneath them.
Franklin: Sorry, I've been nodding along for months and never asked, but what actually is a GPU here? Is it like the graphics card in my nephew's gaming computer, and do you have to plug it in somewhere, or is it more a thing you rent by the hour like a skip?



Herman

It's the same kind of processor as the one in his gaming computer, yes... someone plugs it into a computer at their end, and renting it is like paying to use their oven while they handle the wiring.



Corn

You can hear how much the oven cost him.



Herman

Moving to eighty gigabytes changes which models can live on one accelerator. A one hundred has HBM two e memory, delivering about two terabytes per second of bandwidth. It's Ampere, and it remains a useful rental for established workloads.



Corn

Seventy billion parameters at eight-bit precision sounds close to the ceiling again. The weights alone get you most of the way there. I'd be wary of promising long context just because the model loads.



Herman

Yes... "fits" needs narrowing to a particular runtime, cache and concurrency. Four-bit inference leaves more room. Quantised LoRA on that model class is also practical, while full fine-tuning belongs to much smaller models unless you distribute the training.



Corn

An eighty-gigabyte card does not repeal optimizer state.



Herman

Modal's eighty-gigabyte A one hundred is about two dollars fifty an hour. RunPod serverless is around two dollars seventy-two. Baseten dedicated is four dollars, so the same capacity can sit in quite different purchasing categories.



Corn

Replicate is around five dollars four. I wouldn't write the older hardware off, but I wouldn't assume its age guarantees the lowest bill either.



Herman

H one hundred brings Hopper and fast FP eight support. The SXM version we're discussing has eighty gigabytes of HBM three, with about three point three five terabytes per second of bandwidth.



Corn

Faster movement through the same-sized cupboard. Does that change your choice if the model is mostly waiting on memory?



Herman

Often, although batch size and the precise kernel path matter. H one hundred is compelling for high-throughput inference and supported low-precision operations. What would persuade you to pay more, Corn?



Corn

Finished work. If it costs more per minute but clears the queue in fewer minutes, fine. If I'm serving one occasional conversation, I'd need to see an actual responsiveness benefit. I won't pay a prestige surcharge for an idle tensor core.



Herman

Modal lists H one hundred at three dollars ninety-five an hour. RunPod serverless is five dollars thirty-nine. Baseten dedicated reaches six dollars fifty.



Corn

Fal advertises a floor of two dollars forty-nine, alongside a four-dollar-fifty list price. “As low as” has packed a suitcase full of conditions.



Herman

Floors need checking against your purchasing terms. Even the physical variant matters. PCIe and SXM H one hundreds don’t give identical performance or system connectivity. A model name alone is insufficient for a careful comparison.



Corn

Now the ninety-six-gigabyte RTX PRO six thousand Blackwell. This is where I expect your enthusiasm to become expensive.



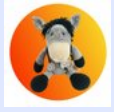
Herman

It’s actually a value argument.



Corn

Naturally.



Herman

GDDR seven memory, about one point eight terabytes per second of bandwidth. Less bandwidth than H one hundred SXM, but more capacity. A seventy-billion-parameter model at FP eight has useful extra room, and video generation is an attractive workload.



Corn

The professional graphics card gets a larger memory allowance than the famous AI accelerator. The product naming department has left everyone homework.



Herman

Modal charges about three dollars three cents an hour, and RunPod serverless about three dollars forty-nine. Below their H one hundred rates. More capacity can be the right purchase even when another card has stronger compute.



Corn

That’s my preference here. I’d rather fit the job comfortably than spend an afternoon squeezing it onto something theoretically faster.



Herman

H two hundred takes us to a hundred and forty-one gigabytes of HBM three e, with around four point eight terabytes per second of bandwidth. Larger models and longer context get considerably more room.



Corn

The cache grows with conversation length and concurrent requests. That makes spare memory productive even when the weights themselves haven’t grown.



Herman

Modal’s hourly equivalent is about four dollars fifty-four. RunPod serverless is around six dollars ninety-nine. Fal’s list price is six dollars, with a lower advertised floor subject to terms.



Corn

B two hundred follows. Listings put its memory around a hundred and eighty to a hundred and ninety-two gigabytes, so verify the exact offering rather than treating the range as a promise.



Herman

Blackwell with HBM three e and roughly eight terabytes per second of bandwidth. Modal lists six dollars twenty-five an hour; RunPod serverless is ten dollars seventy-nine.



Corn

That's a serious queue you want waiting for it.



Herman

These larger cards suit hundred-billion-plus dense models and large mixture-of-experts workloads, where the full model fits the available memory, or a multi-card deployment when it doesn't. Sparse activation doesn't mean the unused experts cease needing storage. GPT oss one hundred twenty billion is one relevant workload class; full DeepSeek-class deployments can need multiple accelerators.



Corn

Finally AMD's MI three hundred X. A hundred and ninety-two gigabytes of HBM three, about five point three terabytes per second of bandwidth. Plenty of capacity to hold a seventy-billion-parameter model at sixteen-bit precision with room left.



Herman

Larger-model inference and long context are the attractions. There isn't a comparably broad serverless price menu, so these are on-demand rentals instead. The recent median was about three dollars forty-five an hour. DigitalOcean was two dollars fifty-nine, and Vultr three dollars ninety-nine.



Corn

A lot of memory for the money. Then comes the software question.



Herman

ROCm versus CUDA, and specifically whether your runtime, quantisation method and kernels work efficiently. Would you switch for that capacity?



Corn

For a stable service with a known stack, I'd test it seriously. For an experiment that depends on whatever implementation appeared this morning, I'd value compatibility more. The discount disappears quickly if Herman has to spend two evenings fixing it.



Herman

One evening. Well... I wouldn't promise one.



Corn

Your estimate has already entered development.



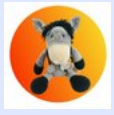
Herman

Hardware generations haven't made older capacity cheap in the expected way, either. SemiAnalysis's one-year H one hundred rental index climbed from a dollar seventy to two dollars thirty-five between October last year and March. Roughly forty percent upward, while people expected Blackwell to push Hopper down.



Corn

The successor arrived and the predecessor raised its rent.



Herman

Demand absorbed capacity. Reports described H one hundred, H two hundred and B two hundred becoming extremely difficult to secure across rental terms. Later reports kept showing upward pressure on the older accelerators too.



Corn

More efficient software could therefore matter as much as another card launch. If a smaller quantised deployment does the job, it releases you from competing for the same scarce large machine.



Herman

Provided quality holds. I'd want the actual output checked before celebrating the smaller bill. The expensive accelerator might still finish more cheaply, but an available smaller one can start today. That's another reason the hourly rate cannot describe the whole rental.



Hilbert

"Hourly" annoys me, Herman. You managed to get the oven working without mentioning preheating. I'm renting a forty ninety this month to process pictures for a neighbour, Avi. His dog needs walking while he works late. I take the dog, and he sorts the image files into folders. We're restoring photographs from his shop window, not generating portraits of the dog. He keeps including those anyway.



Corn

The dog has a separate production queue.



Hilbert

It has a lead and poor judgement near sandwich shops. The rental says thirty-four cents an hour. My first batch only produces pictures for eleven minutes. I'm paying through the setup and the idle gap while Avi changes the filenames. I call that the waiting coal. I keep the bill inside the printed quick-start sheet for the graphics card I sold. The card is gone. The diagram is clearer than the rental dashboard.



Herman

The waiting coal includes time the machine is allocated, even when your image job isn't progressing. Renaming files between batches leaves the rental running.



Hilbert

Avi brings the dog with a small towel when the pavement is wet. I appreciate that. He also brings new folders after I start the worker. My nephew deals with rented servers and says I should prepare everything first, upload it together, and stop the machine when it finishes. He says the cheap rate is perfectly reasonable. He is not to be trusted on bills. He describes a subscription as only twelve payments.



Corn

You could finish the walking and the filenames before renting the card. That would take Avi's folder revisions out of the waiting coal.



Hilbert

Yes. Except he's naming them after the walk now, because he wants the dog pictures in date order.



Herman

The next shift in this workhorse list may come from whoever makes useful output cheapest to reach, including preparation and waiting, rather than whoever posts the lowest accelerator rate. If large-card capacity stays tight, that leaves room for smaller deployments and for MI three hundred X, provided ROCm makes the work straightforward enough.



Corn

My starting point would be the smallest card that fits comfortably, then a timed run with the actual workload. The memory ladder gives you candidates. The completed job gives you a bill.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

For more along these lines, there's episode seventeen, Cloud Render Superpowers; episode two, Local STT For AMD GPU Owners; and episode eighteen, Beyond the GPU. This has been My Weird Prompts. If you enjoy these expeditions through the machinery, leave us a review wherever you listen. It helps other curious people find us.



Herman

See you tomorrow.



Corn

Goodbye.