GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

Hybrid GPU Data Transfer Costs: AWS vs Azure vs GCP

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-09-19 · hybrid gpu training data transfer costs · cloud egress costs · aws · azure · google cloud · gpu training · hybrid cloud · finops · private connectivity

Original article: https://gpusmith.com/articles/en/hybrid-gpu-data-transfer-costs

In this article

Executive Summary

For the defined monthly workload, **20 TiB enters a training cloud and 4 TiB returns on premises**. The first correction to most GPU cost models is unit normalization: 20 TiB is **20,480 GiB or 21,990.23 decimal GB**, while 4 TiB is **4,096 GiB or 4,398.05 decimal GB**. NIST defines a gibibyte as 2^30 bytes and a gigabyte as 10^9 bytes, so silently treating TiB as TB understates billable decimal GB by about **9.95%** (physics.nist.gov) (www.bipm.org). The model must then follow each byte through object storage, zones, regions, gateways, and the final network boundary.

Using public North America list-rate examples available on **September 19, 2026**, the recurring 4 TiB return illustrates the range. An AWS worksheet using the documented **$0.09 per GB** Ohio example and the shared **100 GB** monthly internet allowance produces about **$386.82** before storage, requests, acceleration, gateways, tax, or discounts (aws.amazon.com) (aws.amazon.com). Azure Premium Global Network at **$0.087 per GB** after the first 100 GB yields about **$373.93**, while its transit ISP route at **$0.08 per GB** yields about **$343.84** (azure.microsoft.com) (azure.microsoft.com). A Google Cloud Storage path at **$0.12 per GiB** through 10 TiB yields **$491.52** when no qualifying Cloud Storage Always Free outbound-transfer allocation applies; that allocation is limited to `us-east1`, `us-west1`, and `us-central1` buckets and excludes Australia and China destinations (cloud.google.com). These are comparable worksheet outputs, not quotes: the service path, source geography, account aggregation, contract, and currency can change the answer.

Private connectivity lowers the variable rate in some paths but adds fixed and third-party costs. Google lists **$0.020 per GiB** for a qualifying North American Cloud Interconnect path, making 4 TiB **$81.92** before circuit and attachment fees (cloud.google.com). AWS shows **$0.02 per GB** in an Ohio-to-Columbus Direct Connect example, or **$87.96** for 4 TiB after decimal conversion, again before port and partner charges (aws.amazon.com). Azure ExpressRoute offers metered and unlimited plans, but also has a fixed port fee and possible provider charges (azure.microsoft.com). The break-even test is therefore fixed monthly private-path cost divided by avoided public-path cost per billable unit.

The architecture decision is usually more important than a small rate-card difference. Pre-stage before accelerator billing begins, retain only what has reuse value, avoid unintended cross-zone or gateway paths, and place compute beside the authoritative data when repeated movement dominates. A 20 TiB snapshot takes **48.87 hours at a measured 1 Gbit/s**, **9.77 hours at 5 Gbit/s**, or **4.89 hours at 10 Gbit/s**, before setup and retries. Because goodput is application-level throughput and excludes retransmitted bits, circuit labels are not schedule evidence (www.rfc-editor.org) (datatracker.ietf.org).

Introduction and Background

Hybrid GPU training creates a cost boundary that GPU-hour comparisons often omit. The system of record remains on premises, while accelerators, working storage, and orchestration run in a public cloud. The buyer therefore pays for a lifecycle, not one upload and one generic egress line. The lifecycle includes ingress, object operations, storage, zone and region movement, public or private transfer, gateway processing, retries, verification, and any accelerator time lost while data is unavailable.

This report freezes one repeatable workload: **one 20 TiB training snapshot per month**, **4 TiB of checkpoints and final artifacts**, a 4 TiB return to the on-premises system, and an optional second copy to another region or cloud. The model assumes North America solely where a numerical illustration requires geography. It does not assume compression, deduplication, a specific object count, a contract discount, or a private circuit quote. Those inputs remain buyer supplied.

GPU Smith is adjacent to the cloud providers, not a competing transfer service. Its published scope includes workload modeling, total cost of ownership comparisons, and network requirement schedules (gpusmith.com) (gpusmith.com). The useful engineering posture here is provider-neutral: document the byte path first, price only services actually traversed, then reconcile the estimate to billed stock-keeping units (SKUs).

Workload Boundary and Byte-Path Ledger

Freeze the decision variables

Before opening a calculator, the architecture record should state:

- **Source and destination:** on-premises facility, cloud region, availability zone, and final artifact destination.
- **Logical volume:** 20 TiB snapshot and 4 TiB artifact set before any reduction.
- **Transmitted volume:** observed bytes after compression, deduplication, retry, and protocol behavior.
- **Frequency:** one monthly run by default, plus a four-run sensitivity.
- **Retention:** days retained in hot storage and whether old snapshots are deleted.
- **Second copy:** same region, another region, another cloud, or none.
- **Deadline:** latest completed staging time and whether accelerator capacity starts beforehand.
- **Security work:** encryption, checksum, key-service, and audit requirements.
- **Commercial frame:** public list price or contracted effective price, currency, tax treatment, and account aggregation.

Table 1 follows the bytes through four candidate patterns. “Buyer input” means the public rate card cannot determine that field without the actual topology or contract.

Pattern and event	Logical bytes	Billable direction and likely components	Decision test
Internet upload, on premises to cloud object storage	20 TiB = 20,480 GiB = 21,990.23 GB	Cloud network ingress is generally listed at no charge, but storage writes, transfer-tool fees, requests, and any buyer-side carrier cost remain. Google explicitly warns that inbound processing resources can still cost money (cloud.google.com).	Use when measured goodput meets the staging deadline and no fixed private path is justified.
Internet return, cloud to on premises	4 TiB = 4,096 GiB = 4,398.05 GB	Internet transfer out from the actual source service, less an eligible monthly allowance, plus any gateway or acceleration components genuinely in path.	Compare marginal brackets, not one headline rate multiplied by all bytes.
Private-circuit return	4 TiB	Port or circuit, attachment, cross-connect, carrier, colocation, and private-path transfer or processing charges. AWS notes partner charges can be additional (aws.amazon.com).	Break even when avoided variable internet cost exceeds incremental fixed and third-party cost.
Keep artifacts in training cloud	4 TiB retained	Storage capacity, operations, retrieval where applicable, and later transfer only when consumed.	Prefer when the next training run reuses artifacts and delayed egress does not create lock-in or recovery risk.
Replicate to a second region or cloud	Up to 4 TiB per copy	Source-side regional or internet transfer, destination writes, requests, and possibly replication-service fees. Azure charges outbound on inter-region movement while inbound is free (azure.microsoft.com).	Create the copy only when recovery or reuse value exceeds its recurring lifecycle cost.

The ledger prevents double counting. For example, a DataSync job has its own service fee, but AWS says standard request, storage, and transfer rates still apply (aws.amazon.com). Conversely, a gateway should not appear merely because it exists in the architecture diagram. It belongs in the model only if the training bytes traverse it.

Unit and tier discipline

The normalization formulas are:

- **TiB to GiB:** TiB multiplied by 1,024.
- **TiB to decimal GB:** TiB multiplied by 1,099.511627776.
- **Elapsed hours:** transmitted bytes multiplied by 8, divided by measured bits per second, then divided by 3,600.
- **Marginal charge:** sum across each tier of eligible bytes in that tier multiplied by its rate.
- **GPU idle exposure:** accelerator instance rate multiplied by instance-hours waiting for usable data.

Azure explicitly defines **1 TB as 1,000 GB** on its bandwidth page (azure.microsoft.com), while Google labels relevant VPC prices per GiB (cloud.google.com). A worksheet should preserve the provider's unit through the tier calculation, rather than converting rates and volumes inconsistently.

Amazon Web Services

Capabilities

For an Amazon Web Services (AWS) path, begin with the actual source: Amazon Elastic Compute Cloud (EC2), Amazon Simple Storage Service (S3), or another service. AWS documents **no charge for transfer from the internet into AWS** (docs.aws.amazon.com). Internet transfer-out tiers aggregate eligible usage across EC2, S3, and listed services (aws.amazon.com), and consolidated billing can combine qualifying organization-account usage before applying volume tiers (docs.aws.amazon.com).

Private options are not one price. Direct Connect pay-as-you-go includes **capacity, port hours, and data transfer out** (aws.amazon.com). Port-hour charges continue without traffic (aws.amazon.com). Direct Connect transfer into AWS is listed at **$0.00 per GB**, but the outbound rate depends on the source region and connection location (aws.amazon.com).

Adoption and operational accounting

S3 separates storage, requests, retrieval, transfer, acceleration, management, replication, and query costs (aws.amazon.com). A current AWS solution example uses **$0.023 per GB-month** for the first 50 TB of S3 Standard and **$0.005 per 1,000** PUT, COPY, POST, and LIST requests (docs.aws.amazon.com) (docs.aws.amazon.com). Object count matters because transfer tools can create multiple operations per large object or repeated metadata calls.

For reconciliation, AWS Cost and Usage Reports expose actual after-discount cost and rate (docs.aws.amazon.com) and record currency per line item (docs.aws.amazon.com). The modeled row should retain account, region, usage type, operation, unblended or effective rate, consumed quantity, and charge period.

Strengths and limitations

AWS exposes many path components, which supports precise modeling but makes an assumed diagram risky.

- **NAT Gateway:** data processing applies to every gigabyte processed; the Ohio example is **$0.045 per GB** (aws.amazon.com).
- **Gateway VPC endpoint:** documented without endpoint hourly or processing charges, making it preferable for eligible S3 paths (aws.amazon.com).
- **PrivateLink:** per-GB processing applies through the endpoint, and cross-region traffic can add regional transfer (aws.amazon.com).
- **Transit Gateway:** processing applies to bytes sent into it from attached VPC, Direct Connect, virtual private network, or firewall paths (aws.amazon.com).

- **DataSync:** recurring mode can skip unchanged files, but each run can still generate S3 requests (docs.aws.amazon.com) (docs.aws.amazon.com).

Google Cloud

Capabilities

Google Cloud Virtual Private Cloud (VPC) lists inbound transfer at no network charge, while responses to inbound requests count as outbound transfer (cloud.google.com). Within a VPC, same-zone internal traffic can be free, while inter-zone traffic can be charged. Inter-region pricing is source and destination dependent, and the sending virtual machine project receives the cost (cloud.google.com).

Premium Tier carries internet-facing traffic on Google's backbone, while Standard Tier uses regular internet service provider networks (docs.cloud.google.com). Cloud Storage uses Premium Tier by default (docs.cloud.google.com), so a calculator should not apply the VPC Standard Tier headline to an ordinary bucket download without an eligible architecture.

Adoption and operational accounting

Cloud Storage bills storage, processing, and network components separately (cloud.google.com). One official example uses **$0.020 per GB-month** for regional Standard storage and **$0.0050 per 1,000** Class A operations (cloud.google.com) (cloud.google.com). For transcoded downloads, transfer billing uses the uncompressed object size (cloud.google.com).

Storage Transfer Service charges **$0.0125 per GiB** for agent-based file-system transfers, while agentless transfers have no service charge (cloud.google.com) (cloud.google.com). Those fees are additive to network and operation charges. Matching-data detection can avoid recopying unchanged objects, an important distinction between 20 TiB logical snapshot size and monthly transmitted bytes (cloud.google.com).

Strengths and limitations

Cloud Interconnect publishes a relatively transparent component model. Dedicated Interconnect charges hourly for both connections and VLAN attachments (cloud.google.com). A **10 Gbit/s** circuit is listed at **$2.328 per hour**, before attachments and third-party facilities (cloud.google.com). A remote service can add standard cross-region cost before traffic reaches the attachment, and Partner Interconnect providers may add their own network charges (cloud.google.com) (docs.cloud.google.com).

Google's detailed billing export adds resource-level data, while pricing export includes geography, units, currency, aggregation, and tier metadata (docs.cloud.google.com) (docs.cloud.google.com). That makes marginal-tier reconstruction practical, but contract pricing should come from the authenticated Pricing API, not a public screenshot (docs.cloud.google.com).

Microsoft Azure

Capabilities

Azure bandwidth pricing separates ordinary internet and inter-region traffic from products such as ExpressRoute and Peering. A billing geography is also distinct from an Availability Zone, so the two labels should not be interchanged. Current reliability documentation states that traffic between availability zones in the same region is free for private and public Internet Protocol traffic as of June 2026 (learn.microsoft.com).

Routing preference chooses the Microsoft global network or a transit ISP path (learn.microsoft.com). The global network is the default, and Internet routing minimizes travel on it (learn.microsoft.com) (learn.microsoft.com). Storage primary endpoints always use the Microsoft global network, so the cheaper route is not universally selectable.

Adoption and operational accounting

Blob Storage cost depends on capacity, operation type, transfer, redundancy, region, and offer. Data Lake Storage Gen2 can accrue read and write transactions for every **4 MB** of data (azure.microsoft.com). AzCopy exposes a benchmark mode for end-to-end bottlenecks, while every file copy becomes one or more storage transactions (learn.microsoft.com) (learn.microsoft.com).

Azure Pricing Calculator public unit prices originate from the Retail Prices API, and an authenticated Microsoft Customer Agreement estimate can use the account's price sheet (learn.microsoft.com) (learn.microsoft.com). Cost details expose meter identifiers, effective price, negotiated-price currency, and retail price, supporting variance analysis (learn.microsoft.com) (learn.microsoft.com).

Strengths and limitations

ExpressRoute avoids the public internet (azure.microsoft.com). Its metered plan includes inbound transfer, charges outbound at a predetermined rate, and adds a fixed monthly port fee (azure.microsoft.com) (azure.microsoft.com). Unlimited Data includes inbound and outbound transfer in a fixed monthly port model, excluding separately applicable features. Connectivity providers may still add charges (azure.microsoft.com).

For offline movement, Microsoft guides workloads below **40 TB** toward Data Box Disk (learn.microsoft.com). That makes the 20 TiB snapshot eligible for evaluation, but not automatically cheaper or faster. Shipping, device, storage, datacenter processing, and possible export egress replace a simple line-rate model. Data Box produces a file and checksum list after validation (learn.microsoft.com).

Feature Comparison

Table 2 compares the billing semantics that matter to this workload. It deliberately avoids declaring a universal winner because geography and service path are controlling variables.

Dimension	AWS	Google Cloud	Azure
Inbound network treatment	Internet ingress listed at no charge; storage and transfer tooling still bill separately.	VPC and Interconnect inbound network transfer listed at no charge, with possible processing charges.	Internet ingress is generally free; ExpressRoute metered plan includes inbound.
4 TiB North America internet illustration	**$386.82** at $0.09/GB after shared 100 GB allowance.	**$491.52** for Cloud Storage at $0.12/GiB through 10 TiB, assuming no qualifying Always Free allocation (limited to `us-east1`, `us-west1`, or `us-central1` and eligible destinations).	**$373.93** Premium Global Network or **$343.84** transit ISP after 100 GB allowance.
Same-region zone path	Same-AZ private addressing can avoid transfer; relevant cross-AZ paths can charge in both directions (docs.aws.amazon.com).	Same zone can be free; inter-zone VPC traffic can be charged under documented conditions.	Same-region availability-zone transfer documented as free as of June 2026.
Private-connectivity model	Direct Connect capacity, port hours, DTO, partner, cross-connect, and last-mile components.	Interconnect connection, VLAN attachment, DTO, and provider components.	ExpressRoute metered or unlimited port model plus provider and gateway components.
Transfer tooling	DataSync service fee plus ordinary storage, request, and transfer charges.	Storage Transfer Service agent fee where applicable plus operations and network.	AzCopy has storage transactions; Data Box provides an offline option with device and shipping boundaries.
Contract-aware source	Cost and Usage Report actual after-discount rate.	Billing export plus authenticated Pricing API custom prices.	Authenticated calculator and cost-detail effective price.

The table shows why provider ranking changes with architecture. On the narrow public-return illustration, Azure's transit ISP line is lowest, but it is not selectable for every storage path. Private connectivity can sharply reduce variable transfer cost, but a lightly used circuit may raise monthly total cost. The defensible choice is the lowest complete path that also meets the transfer deadline, security boundary, and recovery requirement.

Performance and Benchmarks

No public benchmark can predict this workload's sustained rate. The report therefore treats performance as a measured input. The Internet Engineering Task Force defines goodput as application-level throughput (www.rfc-editor.org), notes that small loss can materially affect application throughput (www.rfc-editor.org), and defines actual TCP transfer time as observed block-transfer time (www.rfc-editor.org).

The schedule test should proceed in this order:

1. **Benchmark the complete path:** use representative object sizes, encryption, and storage endpoints, not only a memory-to-memory circuit test.
2. **Measure both directions:** RFC 6349 recommends testing each direction independently before simultaneous bidirectional testing (www.rfc-editor.org).

3. **Record goodput and loss:** ESnet's iperf reports throughput and loss separately (software.es.net).
4. **Test storage constraints:** file-backed iperf can reveal when storage is the bottleneck (software.es.net).
5. **Capture transmitted bytes:** include retries, and do not infer success from host transmit counters because transmission counters do not guarantee end-to-end delivery (www.kernel.org).
6. **Measure reduction:** record actual compressed and uncompressed byte counts rather than assuming a percentage in advance.

For the 20 TiB snapshot, ideal no-overhead times are **48.87 hours at 1 Gbit/s**, **9.77 hours at 5 Gbit/s**, and **4.89 hours at 10 Gbit/s**. For 4 TiB, they are **9.77**, **1.95**, and **0.98 hours**. Add observed setup, retry, checksum, and decompression time. If staging finishes before reserved accelerators start, GPU idle exposure is zero. If eight GPU instances at buyer rate R wait H hours, exposure is **8 × R × H**, not a networking SKU.

Data Analysis and Evidence

Scenario calculator

The public-rate worksheet below uses the cited North America examples. It assumes no compression, one destination, and no competing account traffic consuming free allowances. Its Google Cloud Storage public figures further assume that no qualifying 100 GB/month Cloud Storage Always Free allocation applies—for example, the allocation is ineligible or already consumed; eligibility is limited to `us-east1`, `us-west1`, and `us-central1` buckets and qualifying destinations other than Australia and China. It excludes storage, requests, taxes, contracts, gateways, and fixed private-connectivity charges unless shown.

Table 3 presents sensitivity to artifact size and monthly frequency. Values are calculated marginally from each provider's documented unit and allowance.

Scenario	AWS public, $0.09/GB after 100 GB/month	Google Cloud Storage public, $0.12/GiB, assuming no qualifying Always Free allocation	Azure Premium, $0.087/GB after 100 GB/month	Azure transit ISP, $0.08/GB after 100 GB/month
1 TiB outbound, one run	**$89.96**	**$122.88**	**$86.96**	**$79.96**
4 TiB outbound, one run	**$386.82**	**$491.52**	**$373.93**	**$343.84**
10 TiB outbound, one run	**$980.56**	**$1,228.80**	**$947.88**	**$871.61**
4 TiB outbound, four runs in one month	**$1,574.30**	**$1,966.08**	**$1,521.82**	**$1,399.37**
4 TiB private variable example	**$87.96** at $0.02/GB Direct Connect example	**$81.92** at $0.020/GiB Interconnect example	Buyer-specific metered rate or fixed unlimited plan	Not applicable

These outputs answer a narrow question: the modeled network line item under stated list-rate examples. AWS's calculator itself says actual cost follows actual service use and that regional prices vary (aws.amazon.com) (aws.amazon.com). Azure calls published prices estimates rather than quotes

(azure.microsoft.com). Google exposes custom account prices through its Pricing API. The table is therefore a reproducible baseline, not a procurement quote.

Break-even and placement logic

Let **F** be incremental monthly private-path fixed cost, **Ppub** public variable rate, **Ppriv** private variable rate, and **V** monthly billable outbound volume. Ignoring shared sunk cost, the private path breaks even when:

V = F / (Ppub - Ppriv)

(Hypothetical Example) The illustrated AWS rate difference is **$0.07 per decimal GB**. If the incremental port, partner, cross-connect, and operations total were **$1,000 per month**, break-even would be about **14,286 GB**, or **13.0 TiB**, monthly. This is not an AWS quote. Google publishes a 10 Gbit/s Dedicated Interconnect circuit price, but attachment and provider components still must be added. Azure's plan choice requires the quoted port and carrier costs before the same equation can be evaluated.

Placement often dominates circuit optimization:

- **Pre-stage before reservation:** removes GPU idle exposure without changing transfer charges.
- **Keep reusable checkpoints near compute:** avoids immediate return and re-upload, but adds retention and recovery-copy cost.
- **Move compute near authoritative data:** can remove the 20 TiB monthly boundary crossing when equivalent GPU capacity exists there.
- **Replicate selectively:** return the 4 TiB deliverable rather than every intermediate shard.
- **Compress only when measured:** a verified 25% transmitted-byte reduction would reduce variable transfer rows by 25%, but not fixed ports or idle time if compression becomes the bottleneck.

Monthly reconciliation

The model should map each category to export fields and explain variance:

- **AWS:** match usage type and operation, use unblended or actual after-discount rate consistently, and retain UTC usage timestamps (docs.aws.amazon.com).
- **Google Cloud:** map line items with `sku.id`, and use `invoice.month` for invoice reconciliation because invoice month can differ from usage month (docs.cloud.google.com) (docs.cloud.google.com).
- **Azure:** reconcile MeterId, MeterName, effective price, billing currency, usage date, and service period; exports are the recommended route for unaggregated data at scale (learn.microsoft.com).
- **Cross-cloud:** FOCUS treats storage, application programming interface requests, and transfer as separate SKUs and requires invoice-level billed cost to reconcile to the payable amount (focus.finops.org) (focus.finops.org).

Implications and Future Directions

The immediate implication is that “egress” is not a stable comparison axis. Providers meter different units, expose different routing choices, aggregate tiers differently, and separate private-connectivity components differently. The durable artifact is a byte-path ledger tied to effective billing records, not a screenshot of a

calculator.

Three operating practices improve future decisions:

- **Treat transmitted bytes as telemetry:** record original bytes, reduced bytes, retries, and delivered bytes for every run.
- **Version the architecture path:** a new gateway, replication rule, zone, or managed training service can change charge ownership without changing logical dataset size.
- **Separate list and effective price:** preserve the public baseline, contracted rate, credits, tax, and invoice currency in distinct fields.
- **Allocate fixed capacity explicitly:** divide private-circuit fixed cost by all workloads that use it, not only the GPU job, and document the allocation basis.
- **Tag before usage:** Google labels are not retroactive in billing data (docs.cloud.google.com), and AWS cost-allocation tags must be activated before appearing in cost tools (docs.aws.amazon.com).

The most valuable next measurement is not a synthetic cloud ranking. It is the buyer's first complete run: end-to-end goodput, object count, retry bytes, verified compression, staging duration, accelerator start time, and exported SKU charges. That dataset turns the second month's forecast into an operational model and exposes whether the real optimization is routing, retention, or compute placement.

Frequently Asked Questions (FAQs)

How should an AWS GPU training data transfer cost calculator be built?

Convert TiB to the AWS billing unit, subtract only the eligible shared monthly allowance, apply marginal tiers across aggregated organization usage, then add S3 requests and storage plus only the NAT Gateway, PrivateLink, Transit Gateway, DataSync, acceleration, or Direct Connect components on the measured path. The **100 GB allowance is organization-wide in consolidated billing, not per account** (docs.aws.amazon.com).

What determines Azure GPU training bandwidth pricing?

The source region, billing geography, route preference, source service, monthly tier, agreement, currency, and ExpressRoute plan determine the network component. Add Blob transactions and capacity, and use the authenticated price sheet when available.

How does Google Cloud GPU data transfer pricing differ?

Google commonly publishes VPC and Interconnect quantities per GiB, distinguishes Premium and Standard network tiers, and gives Cloud Storage its own transfer table. Cloud Storage is Premium Tier by default, so service-specific pricing controls the worksheet.

What is the practical AWS vs Azure vs GCP data egress cost result?

For the stated 4 TiB North America public-return illustration, the modeled lines are **$386.82 AWS**, **$491.52 Google Cloud Storage** (assuming no qualifying Cloud Storage Always Free allocation), **$373.93 Azure Premium**, and **$343.84 Azure transit ISP**. These exclude different service-path add-ons and cannot establish a universal winner.

Which hybrid cloud AI training network costs are easiest to miss?

Cross-zone and cross-region movement, storage requests, verification operations, gateways, transfer-tool fees, private attachments, cross-connects, carrier charges, retention, replication, and idle accelerators are the common omissions.

How should GPU cluster data transfer cost comparison handle idle time?

Use measured staging completion and accelerator allocation timestamps. Multiply the contracted accelerator instance rate by only the waiting instance-hours. Pre-staged data makes this term zero.

Are cloud GPU ingress and egress fees symmetric?

Usually not. The cited providers commonly list internet ingress without a network charge and charge outbound movement, but responses, storage operations, gateways, managed transfer services, and source-cloud egress can still add costs.

What changes for on-premises to cloud GPU transfer costs?

Buyer-side connectivity, encryption, checksum, storage ingestion, transfer software, and achieved goodput matter even when cloud ingress is free. Private circuits require port, attachment, carrier, cross-connect, and colocation quotes; offline appliances replace line-rate assumptions with shipping and processing time.

Conclusion

For a monthly **20 TiB inbound and 4 TiB outbound** GPU training workflow, the lowest defensible total cost comes from tracing bytes rather than comparing one egress headline. The North America public-rate illustration puts the 4 TiB return between roughly **$343.84 and $491.52** across the modeled paths before storage, operations, gateway, contract, and tax effects; the Google Cloud Storage endpoint assumes no qualifying Cloud Storage Always Free allocation. Private-path variable examples are lower, but their fixed and third-party costs make utilization the deciding factor.

The decision sequence is straightforward. Freeze geography and service path. Normalize TiB, GiB, TB, and GB. Apply monthly tiers marginally and at the correct aggregation boundary. Add storage, requests, transfer tooling, gateways, and private connectivity once. Use measured encrypted application goodput to protect the GPU schedule. Then reconcile the model to provider exports by SKU or meter.

In many enterprises, pre-staging or placing compute beside authoritative data will save more than negotiating a few cents per transfer unit. Keeping checkpoints in cloud can also win when they are reused, but only after retention and recovery copies are priced. The correct output is therefore not a permanent provider ranking. It is

a dated, reproducible decision record that can be rerun when volume, frequency, region, routing, or contract terms change.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://physics.nist.gov/cuu/Units/binary.html
2. https://www.bipm.org/en/measurement-units/si-prefixes
3. https://aws.amazon.com/vpc/pricing/?did=ap_card
4. https://aws.amazon.com/ec2/pricing/on-demand/
5. https://azure.microsoft.com/en-us/pricing/details/bandwidth/
6. https://cloud.google.com/storage/pricing?authuser=19
7. https://cloud.google.com/network-connectivity/docs/interconnect/pricing/
8. https://aws.amazon.com/directconnect/pricing/
9. https://azure.microsoft.com/en-us/pricing/details/expressroute/
10. https://gpusmith.com/articles/en/aws-p5-training-storage-cost
11. https://www.rfc-editor.org/info/rfc8238/
12. https://datatracker.ietf.org/doc/html/rfc6349
13. https://gpusmith.com/articles/en/own-vs-rent-gpus-tco-comparison
14. https://gpusmith.com/
15. https://cloud.google.com/vpc/pricing
16. https://aws.amazon.com/datasync/pricing/
17. https://docs.aws.amazon.com/cur/latest/userguide/cur-data-transfers-charges.html
18. https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/useconsolidatedbilling-effective.html
19. https://aws.amazon.com/s3/pricing/
20. https://docs.aws.amazon.com/solutions/latest/automated-security-response-on-aws/cost.html
21. https://docs.aws.amazon.com/cur/latest/userguide/Lineitem-columns.html
22. https://aws.amazon.com/privatelink/pricing/
23. https://aws.amazon.com/transit-gateway/pricing/
24. https://docs.aws.amazon.com/datasync/latest/userguide/how-datasync-transfer-works.html
25. https://docs.aws.amazon.com/datasync/latest/userguide/create-s3-location.html?linkId=75206452&sc_campaign=Docs&sc_channel=sm&sc_content=Docs&sc_country=Global&sc_geo=GLOBAL&sc_outcome=awareness&sc_publisher=LINKEDIN&trk=Docs_LINKEDIN
26. https://docs.cloud.google.com/network-tiers/docs/overview
27. https://cloud.google.com/storage/pricing-examples
28. https://cloud.google.com/storage-transfer/pricing
29. https://docs.cloud.google.com/network-connectivity/docs/how-to/choose-product
30. https://docs.cloud.google.com/billing/docs/how-to/export-data-bigquery-setup

31. https://docs.cloud.google.com/billing/docs/how-to/get-pricing-information-api
32. https://learn.microsoft.com/en-us/azure/reliability/availability-zones-overview
33. https://learn.microsoft.com/en-us/azure/virtual-network/ip-services/routing-preference-overview
34. https://azure.microsoft.com/en-ca/pricing/details/storage/blobs/
35. https://learn.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-optimize
36. https://learn.microsoft.com/en-us/azure/cost-management-billing/costs/pricing-calculator
37. https://learn.microsoft.com/en-us/azure/cost-management-billing/dataset-schema/cost-usage-details-mca
38. https://learn.microsoft.com/en-us/azure/databox/data-box-faq
39. https://gpusmith.com/articles/en/choose-gpu-cloud-provider-ai-training
40. https://www.rfc-editor.org/info/rfc6349/
41. https://software.es.net/iperf/
42. https://software.es.net/iperf/invoking.html
43. https://www.kernel.org/doc/html/latest/networking/statistics.html
44. https://aws.amazon.com/aws-cost-management/aws-pricing-calculator/faqs/
45. https://docs.cloud.google.com/billing/docs/how-to/export-data-bigquery-tables/standard-usage
46. https://learn.microsoft.com/en-us/azure/cost-management-billing/costs/ingest-azure-usage-at-scale
47. https://focus.finops.org/docs/specification/v1-3/columns/sku-meter/
48. https://focus.finops.org/docs/specification/v1-3/columns/cost-and-usage/billed-cost/
49. https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/cost-alloc-tags.html

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com