Shelby + DoubleZero

Cost savings analysis for Northern Data /
Rumble / Tether GPU infrastructure

TOTAL CAPITAL DEPLOYED >$2.5B

GPU FLEET 22,400 H100/H200 | 9 EU DATA CENTERS

TETHER GPU LEASE $150M / 2 YEARS ($75M/YR)

FULL FLEET CLOUD REVENUE ~$135M/YR ANNUALIZED

UTILIZATION 66% JAN 2026, TARGET 85% Q1 2026

CONSERVATIVE: $24M/YR | OPTIMISTIC: $53M/YR

WRITTEN BY BUZZ BOULANGER, MAX MOHAMMADI, PRANAV RAVAL

APRIL 2026 | CONFIDENTIAL

1. Capital structure

The $75M/year Tether GPU lease is just the anchor tenant contract. Shelby and DoubleZero efficiency improvements apply to the entire 22,400 GPU fleet serving all customers, not just Tether's portion. The total capital at play is substantially larger.

CAPITAL FLOW	AMOUNT	TYPE
Tether investment in Rumble	$775M	Equity
Tether loan to Northern Data	EUR 610M (~$665M)	Debt (restructuring)
Rumble acquisition of Northern Data	~$767M	All-share deal
Tether GPU services lease	$150M / 2 years	Revenue contract
Tether advertising commitment	$100M / 2 years	Revenue contract
Northern Data net debt forecast	EUR 750M+	Balance sheet
Total capital deployed	>$2.5B	

Northern Data cloud revenue trajectory

PERIOD	CLOUD REVENUE	GPU UTILIZATION
Q3 2025	EUR 8M	~11% (Aug) rising
Q4 2025	EUR 31M	62% (Dec)
Jan 2026	Trending higher	66%
Q1 2026 target	EUR 31M+ run rate	85% allocation
FY outlook	EUR 240-320M group	EUR 80-130M adj. EBITDA

2. GPU fleet utilization breakdown

Northern Data's 22,400 GPUs split between training (QVAC model development, Genesis dataset generation, BitNet compression), platform inference (Rumble auto-captioning, classification, live translation), and unsold capacity.

SEGMENT	GPUS	% OF FLEET	TRUE COMPUTE UTIL.
Training (QVAC, Genesis, BitNet)	~10,200	~46%	60-65%
Platform inference (Rumble AI features)	~4,360	~19%	70-75%
Unsold capacity	~7,840	~35%	N/A (sales problem)
Total fleet	22,400	100%	

The opportunity cost gap

The delta between commercially sold (65%) and true compute utilization (39-49%) represents GPUs that customers pay for but produce zero useful compute.

METRIC	VALUE
Commercially sold	65% of fleet = 14,560 GPUs
True compute utilization	39-49% of fleet = 8,736-10,976 GPUs
The gap (our opportunity)	16-26% = 3,584-5,824 GPUs
Dollar value of gap	$12M - $19.5M / year
As % of Tether $75M/yr lease	16% - 26%
Each 1% recovered	~$750K / year back to Tether

3. The I/O bottleneck

Industry data on GPU idle time from data loading:

- **Run:ai (2025):** ~40% of enterprise GPU idle time is attributed to I/O wait states.
- **Microsoft:** Up to 70% of model training time consumed by I/O operations across millions of analyzed workloads.
- Poorly optimized pipelines reduce GPU utilization to 40-60%. Optimized pipelines achieve 90%+.
- Only 7% of organizations report GPU utilization exceeding 85% during peak training.

How this breaks at Northern Data

- **Cross-site fetch over public internet.** QVAC training data in Frankfurt, GPU cluster in Helsinki. 20-40ms baseline latency with spikes to 180ms+ on congested hops. GPU sits idle during every spike.
- **Full dataset replication before training starts.** 200TB dataset x 9 sites = 1.8PB of copies. Sync takes hours to days. Fresh Rumble content means stale copies.
- **Slowest node drags entire sync group.** Data-parallel all-reduce means one jitter spike stalls every GPU in the distributed job. At 10,000+ GPUs across 9 sites, probability approaches 100%.

Before vs after: Shelby + DoubleZero

METRIC	BEFORE (CURRENT)	AFTER (SHELBY + DZ)
Data fetch latency	20-180ms (variable)	Sub-ms (cache) / 2-5ms (DZ)
Jitter	Unpredictable spikes	Zero (deterministic fiber)
Dataset availability	Wait for replication	Instant (global namespace)
Cache hit rate	0% (no shared cache)	85-90% by epoch 2
Checkpoint restore	Minutes (cold S3 fetch)	Seconds (Shelby edge cache)
GPU compute utilization	60-65%	85-90%
Annual I/O waste	$10-12M	$1.6-3.6M

4. Checkpoint offload: the largest cost lever

During training, thousands of GPUs periodically dump checkpoint data to parallel file systems like VAST (NVMe all-flash). This storage is expensive and data is locked within a rack/cluster. Shelby is an HDD-based object store deployed across all Northern Data data centers, connected through dedicated 100 Gbps fiber. At approximately 20x cheaper per TB than VAST, checkpoint offload fundamentally changes the economics.

Storage cost comparison

	VAST (NVME ALL-FLASH)	SHELBY (HDD OBJECT STORE)
Media type	QLC NVMe SSD	Enterprise HDD
Cost per TB/year (TCO)	$80-160	$4-8 (20x cheaper)
Latency	Microseconds	Milliseconds (RPC cache: sub-ms)
Best use case	Active training data, GPU feed	Checkpoints, datasets, cross-DC access
Key constraint	Expensive, locked to rack/cluster	Higher base latency for hot reads
Key strength	NVMe speed, GPUDirect	Global namespace, any DC, 100Gbps fiber

Three savings mechanisms

- **Dump checkpoints to Shelby, free VAST immediately.** 30-50% of VAST capacity is checkpoint data at any time. A 70B model checkpoint = 140-280GB. Moving to Shelby HDD at 1/20th cost saves $285K-$760K/yr in storage, plus frees VAST for the active training data that actually needs NVMe speed.
- **Freed VAST capacity enables more concurrent jobs.** VAST capacity is the gating constraint on how many jobs run simultaneously. Freeing 30-50% means 30-50% more concurrent jobs on existing hardware. At $150-200M/yr fleet revenue, each 1% more concurrency = $1.5-2M/yr revenue capacity.
- **Deferred VAST CapEx.** VAST systems cost $1M+ per deployment. Checkpoint offload means Northern Data delays next VAST expansion at 3-4 sites, saving $2-5M in hardware not purchased.

5. Dynamic workload placement + dataset prefetching

Dynamic workload placement

Today, training jobs are locked to whichever site has data replicated locally. If Helsinki has 2,000 idle GPUs but the dataset is on VAST in Frankfurt, the job runs in Frankfurt. With Shelby's global namespace accessible from all 9 data centers over dedicated fiber, a job scheduler can place work wherever GPUs are free.

- Moving from 85% allocation to 90-95% effective utilization by eliminating geographic job stickiness.
- Each 1% improvement = $1.5-2M/year on the full fleet revenue base.
- Inter-job gap (2-6 hours for dataset replication) drops to minutes.

Dataset prefetching

Customer submits a job. Shelby RPC begins prefetching the dataset to local NVMe cache immediately. By the time GPUs are allocated and the job launches, data is already warm. First epoch runs at cache speed, not network speed.

- Without prefetch: first 30-60 min of every job is data loading at reduced GPU utilization.
- Popular public datasets (Common Crawl, LAION, Genesis) pre-staged on RPCs at every site. Any customer job using these starts instantly.
- Competitive advantage vs other GPU cloud providers who require manual data staging.

6. Full annual cost savings summary

CATEGORY	CONSERVATIVE	OPTIMISTIC
CHECKPOINT OFFLOAD (VAST DISPLACEMENT)		
Storage cost delta (VAST $100/TB vs Shelby $5/TB)	$285K	$760K
VAST CapEx deferred (avoid expansion at 3-4 sites)	$2.0M	$5.0M
Freed VAST capacity enables more concurrent jobs	$4.5M	$10.0M
Subtotal	**$6.8M**	**$15.8M**
DYNAMIC WORKLOAD PLACEMENT		
Cross-DC scheduling (utilization 85% to 90-95%)	$3.0M	$7.5M
Reduced inter-job gap (replication wait eliminated)	$1.5M	$3.0M
Subtotal	**$4.5M**	**$10.5M**
DATASET PREFETCHING		
GPU idle time at job start eliminated	$0.8M	$2.0M
Multi-customer dataset staging	$0.5M	$1.5M
Subtotal	**$1.3M**	**$3.5M**
I/O WAIT + INFERENCE + NETWORK (PRIOR ANALYSIS)		
Training I/O wait recovery (DZ fiber + edge cache)	$6.3M	$11.3M
Platform inference model weight loading	$1.5M	$2.8M
Storage replication + egress + transfer	$2.3M	$4.7M
New revenue (DZ fiber contrib + Shelby reads)	$0.8M	$3.5M
Subtotal	**$10.9M**	**$22.3M**
TOTAL ANNUAL IMPACT	**$24.0M**	**$53.0M**
As % of full fleet revenue ($150-200M/yr)	**12-16%**	**27-35%**
Implied valuation lift at 20x EBITDA	**$480M**	**$1.06B**

Conservative: 50% I/O recovery, 3PB checkpoint offload, 5% placement gain. Optimistic: 75% I/O recovery, 8PB offload, 10% placement gain, active DZ monetization.

7. Site architecture

Replicated across all 9 Northern Data European data centers. Each site operates as both a Shelby storage provider and RPC node, connected over DoubleZero dedicated 100 Gbps fiber.

Storage tier

- **VAST DataStore (NVMe flash):** Canonical hot storage. Feeds GPUs via GPUDirect. Active training data only. Write-optimized, internal to the rack.
- **Shelby storage provider (HDD):** Erasure-coded chunks across all sites. Checkpoints, datasets, cross-DC distribution. 20x cheaper than VAST. Global namespace.

Access tier (revenue engine)

- **Shelby RPC node:** Edge cache on local NVMe. Micropayment metering. Data reconstruction from SP chunks. Sits between all consumers and all storage. Every read flows through, generating spread revenue.

Compute consumers

- **GPU cluster:** QVAC training jobs (~10,200 GPUs). Batch fetches from RPC cache. Checkpoint writes to Shelby SP.
- **Inference endpoints:** Translation, transcription, summarization APIs (~4,360 GPUs). Model weights permanently cached on RPC.
- **External buyers:** Government, enterprise, AI companies. Paid API access via Move smart contracts on Aptos.

Network and settlement

- **DoubleZero fiber backbone:** Dedicated 100 Gbps links between all 9 sites plus external DZ network. Northern Data's fiber flips from cost line to revenue-generating DePIN asset. Earns 2Z tokens.
- **Aptos validator:** Storage commitments, audit settlement, micropayment channels, provenance receipts. Earns APT gas fees on your own economic activity.

8. Second and third order effects

Second order effects (direct consequences)

- **Job scheduling becomes location-agnostic.** Any site runs any job. RPC cache warms by epoch 2. Utilization rises across the full 22,400 GPU fleet.
- **Checkpoint portability enables instant failover.** Frankfurt crashes, Helsinki's RPC restores in seconds from Shelby edge cache. MTTR drops from minutes to seconds.
- **Same infrastructure serves training and inference.** No rebuild needed as Northern Data pivots to inference-heavy workloads. Shelby layer is workload-agnostic.
- **Unsold capacity easier to fill.** New customers write data to Shelby once and run on whichever site has available GPUs. Lower onboarding barrier.

Third order effects (compounding consequences)

- **Proprietary dataset moat accumulates.** Every QVAC training run on Rumble's 52M-MAU content library generates derived datasets with cryptographic provenance on Aptos.
- **Government contracts get compliance for free.** Shelby provenance on Aptos satisfies EU data residency and auditable access requirements at the protocol level.
- **Data marketplace emerges without a separate platform.** Shelby RPC + Move smart contracts IS the marketplace. AI companies discover datasets on-chain, pay per-read via micropayment channels.
- **DZ fiber creates compounding network effect.** External DZ participants route through Northern Data's contributed links, generating 2Z revenue. More participants = more revenue.
- **Inference-as-a-service scales without new hardware.** Every QVAC API call generates Shelby micropayment revenue, Aptos gas fees, and DZ traffic. Three token streams compound linearly with user adoption.

9. Multi-layer token economics

By operating Shelby RPC/SP infrastructure and an Aptos validator, revenue is captured at every layer of the stack:

DoubleZero (2Z) - live on Binance, Coinbase

Contribute Northern Data's inter-DC fiber links to the DZ network. Stake 2Z via the DoubleZero Delegation Program Phase II. Earn 2Z from external participants routing traffic through contributed links. SEC no-action letter confirms programmatic transfers are not securities.

Shelby (TBD) - devnet live, mainnet expected 2026

Run RPC nodes across 9 sites to earn micropayment spread on every read served. Run storage providers to earn read-based rewards for data served. Early node operator positioning for retroactive TGE allocation. Mainnet tokenomics (staking, channel operations, gas structure) pending.

Aptos (APT) - live

Run an Aptos validator processing Shelby settlement transactions. Every storage commitment, micropayment channel settlement, audit verification, and provenance receipt is an Aptos transaction. Earn APT gas fees on your own economic activity.

The compounding flywheel

- 1. Facilitate the Tether / Northern Data / Rumble deal with Shelby
- 2. Northern Data contributes fiber to DoubleZero. Earn 2Z.
- 3. Operate Shelby RPC + SP nodes across 9 sites. Earn micropayment spread.
- 4. RPC serves QVAC training data + Rumble video + external buyers.
- 5. Micropayment channels settle on Aptos. Validator earns APT gas.
- 6. Pre-token Shelby activity positions for TGE allocation.
- 7. All three token streams compound on a single deal brokered.

Infrastructure ownership over title accumulation. You are not speculating on tokens. You are manufacturing the utility that underlies their value.