signal65

July 2026
Dell Technologies - Broadcom

Governing AI Economics

Sovereign Infrastructure and the Token Cost Equation on a Dell and Broadcom AI Factory

In the AI-native enterprise, the token has superseded the seat license as the unit of digital cost, and token spend is usage-driven, nonlinear, and volatile. Retail token rates widely reflect subsidized introductory pricing, adding repricing risk on top of volume risk.

THE TOKEN COST EQUATION

AI SPEND
MONTHLY BUDGET EXPOSURE

=

TOKENS CONSUMED
DEMAND — SET BY ARCHITECTURE

×

COST PER TOKEN
PRICE — SET BY VENUE

Govern Demand

TopicLake Insights transforms raw documents into Compute- Ready Documents: enriched, governed artifacts built once at ingestion instead of reprocessing the same source material through inference on every query.

~2 KB per query

STRUCTURED RETRIEVAL FOR CONSISTENT RESULTS

Stabilize Price

Sustained inference moves on-premises with Dell Technologies and Broadcom networking: amortized, owned capacity replaces volatile usage-based pricing, with zero egress fees. The per-token rate becomes internally derived.

UP TO 18x modeled advantage

AT STEADY-STATE UTILIZATION

Governed Demand

Compute-Ready Documents enrich data once at ingestion and retrieve it repeatedly — every subsequent query inherits the savings.

Owned Economics

Sustained inference on Dell and Broadcom infrastructure replaces volatile, subsidized API pricing with predictable amortized cost.

Sovereign by Design

PII/PHI and proprietary context stay behind the firewall; auditable retrieval grounds every answer in a verifiable path of truth.

THE LEVERS COMPOUND

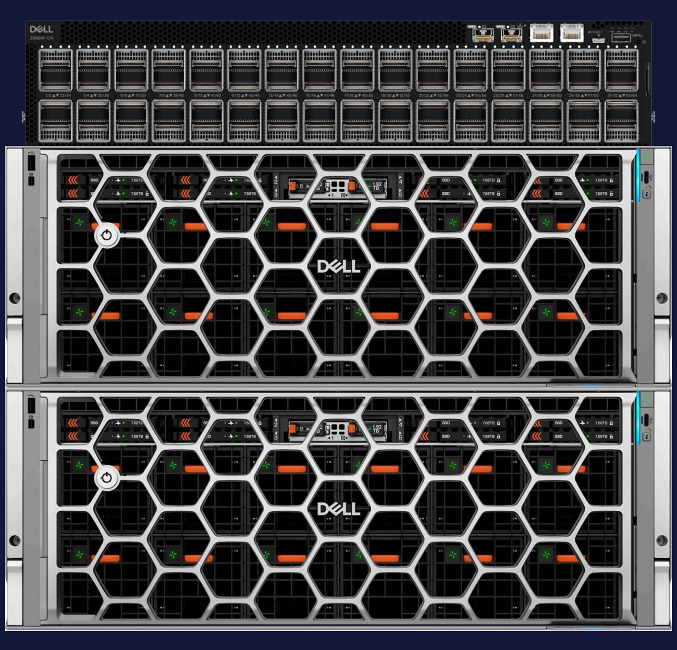
Signal65 modeled an illustrative document-intelligence workload: one million retrieval-augmented queries per month, indexed to a cloud-API baseline of 100. Figures are modeled, not measured.

Scenario	Monthly Token Demand	Cost Index
Raw-text RAG on Cloud APIs (baseline)	~10.8B	100
CRD Retrieval on Cloud APIs (Lever 1)	~1.3B — 8x less	~12
CRD Retrieval on the Dell/Broadcom Factory (Levers 1+2)	~1.3B	~0.7 - 2.4

Modeled scenario; customer-specific results depend on configuration, utilization, pricing, and workload mix. Combined range reflects a conservative 5x venue advantage at the high end and the modeled 18x steady-state ceiling at the low end.

PURPOSE-BUILT FOR PERFORMANCE

The economics of ownership depend on utilization — every idle GPU-hour raises the effective cost of every token produced. The platform is engineered to keep expensive compute saturated.



DELL POWERSWITCH Z9864F-ON + 2× DELL POWEREDGE XE7745

Dell PowerSwitch Z9864F-ON — The Fabric

Built on Broadcom silicon and running enterprise SONiC 4.4.0, the Z9864F-ON interconnects the cluster over 400 GbE RoCEv2, delivering deterministic, low-jitter throughput. In an AI Factory the network is an economic component: congestion and jitter surface as idle accelerator time, and idle accelerators inflate cost per token.

Dell PowerEdge XE7745 — The Compute

A high-memory, multi-GPU platform for large-context inference and high-volume retrieval. Each node in the tested two-server cluster:

- Dual AMD EPYC 9555 · 2.3 TB DDR5 memory
- 8× NVIDIA RTX Pro 6000 Blackwell Server
- 8× Broadcom BCM57608 400GbE RoCEv2 network controllers
- Ubuntu 24.04 LTS · CUDA 13.0 / Driver 580.95.05 · vLLM v0.10.2

BENCHMARKED AT PRODUCTION SCALE

1,385-page FAA technical document set, dense with tables and diagrams — gpt-oss-20b served with vLLM, NVIDIA RTX Pro 6000 Blackwell vs. NVIDIA L40S across the full extraction-and-enrichment pipeline.

1.94x
OCR extraction speedup, Blackwell vs. L40S

11.75
pages per minute across the full document set

0
observed workflow failures in the selfhosted run

6.59kWh
4-GPU efficiency sweet spot — faster than 2-GPU at lower energy

FROM RISK TO RESOLUTION

CFO Risk	How the Sovereign Factory Resolves It
Forecast Volatility	Converts AI spend from an unpredictable variable cost into a manageable capital asset.
Margin Leakage	Reduces uncontrolled escalation of token costs embedded in SaaS and cloud-API contracts.
Capital Timing Risk	Replaces reactive purchasing with a predictable roadmap for GPU/CPU scaling.
Earnings Narrative Risk	Provides an auditable link between infrastructure investment and measurable KPIs.

Govern AI Cost with the Same Rigor Applied to Capital Allocation

Govern tokens with capital-allocation discipline and own the substrate your intelligence is built on to secure a lasting competitive edge.

In partnership with: DELL Technologies | BROADCOM | GADGET SOFTWARE