Project Eureka

On-Premises Inference Platform

White Paper, July 2026

David Beauchamp, Eureka Endeavors

256 GB of on-prem VRAM in a purpose-built 4U GPU server; approximately $45K using public-reseller CapEx pricing.

Executive Summary

Small and mid-size ventures running AI workloads all hit the same wall: API spend scales with usage, rate limits cap throughput, and every request ships your data (and your customers' data) to someone else's datacenter. The standard fix is NVIDIA hardware, which means either datacenter cards at datacenter prices or a consumer-card build that pulls 3,400W and needs its own electrical work.

This paper lays out a third option: 8× Intel Arc Pro B70 GPUs in an ASUS ESC8000A-E13P 4U GPU server barebone on a dual AMD EPYC 9004/9005 platform. 256 GB of VRAM for roughly $45K all-in using public-reseller CapEx pricing. It runs frontier open-weight models (GLM-5.2, DeepSeek-V3, 70B dense architectures) at production quality, draws about 2,200–2,600W sustained in the reference B70 case, and keeps every byte of data in the building. At a $2,000/month API spend, the hardware pays for itself in roughly 23 months.

1 The Build

1.1 Core Platform

Component	Selection	Rationale
Server barebone	ASUS ESC8000A-E13P-32WP 4U GPU server	Purpose-built 8× dual-slot GPU chassis, dual SP5 sockets, 24 DDR5 RDIMM slots, 8× Gen5 x16 rear GPU slots, 8× 2.5 in NVMe bays, onboard 10GbE, redundant Titanium server PSUs
CPUs	2× AMD EPYC 9004/9005, baseline EPYC 9124	PCIe lanes, not cores. The platform exposes the GPU slots and memory channels needed for inference; low-core Genoa keeps cost and power down.
GPUs	8× Intel Arc Pro B70 (32 GB GDDR6)	256 GB total VRAM, 230W reference/TBP varies by board SKU, dual-slot, PCIe 5.0 x16
Memory	384 GB DDR5 ECC RDIMM baseline	Enough host memory for staging and page cache while models and KV cache remain in VRAM
Storage	2× model NVMe + 1× boot NVMe	Uses onboard M.2 / NVMe capacity rather than an external storage shelf
Power	3+1 redundant 3,200W 80 PLUS Titanium server PSU set	Large margin and serviceability, but requires 220–240V server power and final connector validation for the selected B70 SKU
Cooling	Integrated 4U GPU-server airflow	Avoids a custom fan wall / MCIO-riser / ATX multi-PSU build

Total planning cost: approximately $45K all-in using public-reseller CapEx pricing. Procurement can reduce that number through approved surplus, refurb, RFQ, or volume channels, but those savings are intentionally not baked into the official estimate.

1.2 Bill of Materials

Street pricing, USD, July 2026. The SKUs are the reference configuration, not requirements; substitute where stock and pricing dictate. The hard requirements are the lanes, the VRAM, and the wattage.

This BoM uses public-reseller CapEx pricing. It intentionally does not rely on surplus, refurbished, private RFQ, or lucky-stock pricing. Engineers and procurement teams can still reduce individual line items where approved, but the whitepaper budget should stand on its own without assuming those savings.

Component	Part	Qty	Unit	Ext.
GPU	Intel Arc Pro B70 32 GB, ASRock Creator preferred; MAXSUN SKU requires power/TBP verification	8	$1,100	$8,800
CPU	AMD EPYC 9124, Genoa, 16C, 200W, SP5	2	$1,260	$2,520
Server barebone	ASUS ESC8000A-E13P-32WP 4U GPU server barebone: dual SP5, 24× DDR5 RDIMM, 8× rear PCIe Gen5 x16 dual-slot GPU positions, 8× 2.5 in NVMe bays, 2× 10GbE, 3+1 redundant 3200W Titanium PSUs	1	$11,000	$11,000
Memory	32 GB DDR5-4800/5600 ECC RDIMM, 384 GB total	12	$929	$11,148
Model storage	4 TB NVMe Gen4/Gen5 SSD	2	$600	$1,200
Boot drive	1 TB NVMe SSD	1	$130	$130
CPU heatsinks / thermal kit	ASUS-compatible SP5 heatsinks; verify whether included with barebone	1 set	$320	$320
GPU power adapters	Server GPU power harness/adapters for selected B70 SKU, connector-dependent	1 set	$400	$400
Rail kit / cable management	ASUS rail kit and cable arm if not included	1 set	$600	$600
Hardware misc	Standoffs, thermal paste, ties, spare screws	1 set	$200	$200
Parts subtotal				**$36,318**
Tax and shipping, estimated 8%				∼**$2,905**
Spares kit, PSU/fans/cables as available				∼**$2,500**
Contingency, estimated 10%				∼**$3,632**
Planning total				∼**$45,400**

Notes on the line items:

- The ASUS ESC8000A-E13P-32WP becomes the reference platform instead of a hand-assembled ASRock Rack motherboard, generic GPU chassis, MCIO adapter/cable set, ATX PSU stack, sync boards, and fan wall. This is cleaner and more defensible for a whitepaper because the chassis, power, airflow, rear GPU slots, management, and NVMe backplane are designed as one server system.
- The Central Computer listing shows the ESC8000A-E13P-32WP at $10,999.99, special order/out of stock, with dual SP5 sockets, 24 DDR5 RDIMM slots, 8 rear Gen5 x16 dual-slot GPU slots, 8 front 2.5 in NVMe bays, 2× 10GbE, and 3+1 redundant 3200W Titanium power supplies.
- ASUS positions the ESC8000A-E13P as a 4U MGX GPU server for eight dual-slot PCIe accelerators, including 600W-class GPUs. That is more appropriate than treating the build like a commodity 4U case.
- The tradeoff is electrical: this is now a 220–240V server platform with multiple high-current PSU inputs, not a casual single ATX-cord workstation. The actual B70 reference workload may still be roughly 2.2–2.6kW, but the power-infrastructure language should be framed as “standard 240V server/workshop power,” not a normal 120V branch circuit.
- The B70 cost target is now $1,000/card, not $949/card. Intel’s MSRP is still useful context, but public retail listings are closer to $1,000 and can rise with scarcity. The BoM budgets $1,100/card as planning headroom over those listings.
- The 9124 is still a reasonable lowest-cost Genoa choice. Any Genoa-class SP5 CPU keeps the

platform viable; buy lanes, not cores.

- Memory remains the largest pricing swing. The BoM uses public-reseller pricing so the CapEx case is defensible without assuming approved surplus/refurbished sourcing. Procurement may reduce this line materially if refurb or volume channels are acceptable.
- 384 GB of host memory is deliberately light. With the Section 3.3 configuration, models serve from VRAM; host RAM is staging and page cache. Populate more channels only if faster model swaps matter.
- CPU heatsink, rail kit, cable arm, and GPU power harness details must be verified against the exact barebone shipment. Server barebone listings often expose FRU part numbers without making clear which accessories are in the box.

Pricing basis, checked July 2026:

- ASUS ESC8000A-E13P-32WP: Central Computer public listing at $10,999.99, special order/out of stock.
- ASUS ESC8000A-E13P platform specs: dual SP5, 24 DDR5 RDIMM slots, 8 rear PCIe Gen5 x16 GPU slots, 8 front NVMe bays, 3+1 redundant 3200W Titanium PSUs, 220–240V input.
- Intel Arc Pro B70: public ASRock B70 listing around $1,000/card, with Intel MSRP context at $949.
- AMD EPYC 9124: public listings around $1,260 per CPU.
- DDR5 ECC RDIMM: public reseller pricing near $900+ per 32 GB module; approved refurb, surplus, or RFQ channels may reduce this materially but are not included in the budget baseline.
- NVMe storage: 4 TB Gen4 drives are closer to $600 than the earlier $280 assumption.

1.3 GPU Specifications: Intel Arc Pro B70

Spec	Value
GPU Die	BMG-G31 (Battlemage, Xe2-HPG)
Process	TSMC 5 nm
Die Size	368 mm^2
Shading Units / XMX Cores / RT Cores	4096 / 256 / 32
Memory	32 GB GDDR6, 256-bit bus
Memory Bandwidth	608 GB/s (19 Gbps effective)
Compute (FP16 / FP32 / FP64)	45.88 / 22.94 / 2.87 TFLOPS
Base / Boost Clock	2280 / 2800 MHz
TDP / TBP	230W reference; verify board-partner SKU before procurement
Power Connector	SKU-dependent: Intel reference sources indicate 8-pin; ASRock board pages list 12V-2x6 with adapter cabling
Form Factor	Dual-slot, 267×110×39 mm
Bus Interface	PCIe 5.0 x16
MSRP	$949 USD
Launch	March 26, 2026

Raster performance is roughly RTX 4070 class, but the number that matters for inference is the 608 GB/s memory bandwidth: token generation is bandwidth-bound, and 8 cards aggregate to 4,864 GB/s. The 256 XMX cores per card (2,048 across the build) accelerate prompt processing through the SYCL backend.

Retail board options include the ASRock Arc Pro B70 Creator (active, 2540 MHz base), ASRock Passive (fanless, 190 mm), MAXSUN Turbo (2600 MHz base), and MAXSUN Fanless. For this build, active-cooled blower or axial variants are preferred; 230W per card in a dense multi-GPU configuration is no place for passive cooling.

1.4 Architectural Rationale

The original concept inherited from a proven 8× RTX 4090D build documented by u/deebuildsthings on r/LocalAIServers. The revised reference build keeps the eight-GPU idea but moves from a hand-cabled MCIO/ATX approach to an ASUS ESC8000A-E13P GPU server barebone. The NVIDIA-to-Intel trade still matters for three reasons:

1. **256 GB vs 192 GB VRAM.** 33% more capacity at the same card count. That is the difference between running a 400B-parameter model at Q4 and squeezing it into IQ3.
2. **1,840W reference GPU draw vs 3,400W-class 4090D GPU draw.** Roughly half the GPU power in the reference case. The ASUS barebone carries far more PSU headroom than the B70s require, which improves serviceability but changes the electrical framing from "ATX build" to "240V server platform."
3. **∼$8–9K vs $20K+ in GPUs.** The GPU savings fund a large share of the dual EPYC server platform. The remaining cost risk is not the GPUs; it is memory pricing, exact board compatibility, and server accessory/adapter details.

The trade is raw compute: 22.94 TFLOPS FP32 per B70 against ∼83 TFLOPS on a 4090D. For inference this shows up mainly at prompt processing time. Token generation is memory-bandwidth-bound, and the B70's 608 GB/s per card is competitive with the GDDR6X cards in its tier.

1.5 PCIe Lane Budget

The ASUS ESC8000A-E13P platform exposes eight rear PCIe Gen5 x16 slots for dual-slot GPU cards, plus additional PCIe slots for NIC/DPU/front expansion and onboard NVMe storage. In the B70 configuration, the key lane consumer is still the same: eight GPUs at Gen5 x16.

Consumer	Lanes
8 GPUs at Gen5 x16	128
Storage, NICs, platform	Platform-provided Gen5 expansion and NVMe paths
Total	**Server platform provides the required eight Gen5 x16 GPU links**

CPU utilization should stay well below full load during sustained inference. The lanes and GPU mechanical/power envelope are the product; the cores are overhead.

2 Power and Thermal Budget

2.1 System Power Draw

Component	Load	Draw
8× Arc Pro B70 (230W reference each)	Sustained inference	1,840W
2× EPYC 9004/9005 low-core CPUs	Moderate	∼200–400W
ASUS server platform, fans, storage, BMC, NICs	Active server baseline	∼150–300W
Total sustained, B70 reference case		**∼2,200–2,600W**
Peak / transient, SKU-dependent	Full compute + CPU/fan burst	**∼2,800–3,200W**

The ASUS ESC8000A-E13P-32WP uses a 3+1 redundant 3200W 80 PLUS Titanium PSU set. That is far more nameplate capacity than the B70 reference workload needs, but it is appropriate for a purpose-built 8-GPU server platform and keeps the build inside a validated thermal/mechanical/power envelope. The final draw must be rechecked if a non-reference B70 board uses a higher TBP.

2.2 Electrical Infrastructure

In the B70 reference-card case, roughly 2.2–2.6kW sustained at 240V is about 9–11A of real load before PSU efficiency and branch-circuit derating. That is still modest compared with high-end eight-GPU NVIDIA builds, but the ASUS server itself is specified for 220–240Vac input with multiple high-current PSU connections. Treat this as a 240V server/workshop-power deployment, not a 120V office-outlet appliance.

Compare the 4090D equivalent: 4,600W at 120V is 38.3A before derating. That means a 50A/240V

circuit, or multiple 20A circuits with careful load balancing.

2.3 Thermal Management

1,840W of reference-case GPU heat in a 4U chassis is non-trivial. The ASUS barebone is designed for eight dual-slot PCIe accelerators and substantially higher GPU power classes than the B70 reference case, so the airflow story is stronger than the earlier generic-chassis plan. The remaining validation item is board-partner cooler style: active-cooled B70 SKUs are preferred, and passive/fanless SKUs should not be used unless ASUS/Intel airflow guidance explicitly supports them in this chassis.

3 Software Stack

3.1 Inference Runtime

llama.cpp with **Intel SYCL backend**, compiled from upstream master using Intel's `icx/icpx` compilers via the oneAPI toolkit.

The SYCL backend reaches the hardware through the Level Zero driver and uses:

- **Xe Matrix Extensions (XMX)** for prompt-processing GEMM operations (256 XMX cores per GPU, 2,048 total)
- **Unified Shared Memory (USM)** for optimized host-to-device allocation
- **Immediate command list dispatch** for reduced CPU-to-GPU orchestration latency

The B70 is Battlemage (Xe2-HPG), not Alchemist. This matters: Xe2's matrix engine is significantly more mature than first-gen Arc, particularly for GEMM-heavy SYCL workloads. The lemongravy.me benchmarks were run on the same architecture.

3.2 Containerized Deployment

The runtime environment follows the lemongravy.me Docker blueprint: multi-stage build, pinned llama.cpp commit, SYCL compilation in a oneAPI container, and runtime in a minimal Intel GPU compute layer. This isolates the host from library conflicts and makes GPU driver updates a container-layer concern rather than a system-level one.

One known issue: hybrid models (Qwen 3.6's Gated-DeltaNet layers) trigger a destructive cache flush when the context window fills. The `seq_rm` patch (PR #22534) is applied in the Docker build to preserve cache state. This matters for long-running agent automation workloads.

3.3 Key Configuration

```
llama-server \
  -m <model>.gguf \
  --host 0.0.0.0 --port 8080 \
  -ngl 99 \          # Entire model in VRAM
  -c 65536 \          # Context window
  -fa on \            # Flash attention
```

```
    --cache-ram 0       # No spill to host memory
```

With `-ngl 99` and `--cache-ram 0`, the entire model and KV cache live exclusively in VRAM. The PCIe bus carries only token IDs and logits during inference. For sparse MoE models, where only 3–4B parameters activate per token, host memory bandwidth is effectively irrelevant.

4 Capability Projections

4.1 Benchmarked: Single B70

From published benchmarks on a modest i5-12400F with DDR4 (lemongravy.me, 2026):

Model	Quant	Size	Tok/s (gen)
Qwen 3.6-35B-A3B (MoE)	Q4_K_M	20.8 GB	**63**

Prompt processing: 977 tok/s after SYCL JIT warmup. Generation is memory-bandwidth-bound; prompt processing is compute-bound and benefits from XMX acceleration across the card's 256 cores.

4.2 Projected: 8× B70 (256 GB)

GPU splitting in llama.cpp distributes model weights across cards. Generation throughput scales with aggregate memory bandwidth ($8 \times 608 = 4{,}864$ GB/s). Conservative estimates:

Model	Params (total / active)	Quant	Est. Size	Est. Tok/s	Notes
GLM-5.2	∼395B / ∼32B	Q4_K_M	∼200 GB	∼40–50	Frontier open model, MIT license, 1M context
DeepSeek-V3	671B / 37B	Q3_K_M	∼250 GB	∼15–20	Kept warm for agent automation
Llama 4 70B	70B dense	FP16	∼140 GB	∼50–60	Full precision, no degradation
Qwen 3.6-72B	72B dense	Q4_K_M	∼40 GB	∼100+	Sub-second responses
Gemma 4 26B-A4B	26B MoE	Q5_K_M	∼18 GB	∼200+	Real-time interactive

With 256 GB of VRAM, GLM-5.2 at Q4 leaves ∼56 GB of headroom for KV cache and concurrent workloads. The system can run a production model plus a staging fine-tune simultaneously.

To be clear about what these numbers are: projections from single-card benchmarks, not measurements. The validation plan in Section 7 exists to check them before committing to the full build.

5 Value Proposition

5.1 API Cost Comparison

Metric	API (DeepSeek Pro tier)	On-Prem (This Build)
Monthly cost at 100M tok/day	~$2,000+	$0 (after amortization)
Monthly cost at 300M tok/day	~$6,000+	$0
Rate limits	Provider-enforced	None
Data sovereignty	Every request leaves the building	Data never leaves the box
Fine-tune privacy	Vendor-hosted storage	Local, fully controlled
Model availability	Gated by provider	Open-weight models, no gatekeeping
Internet dependency	Required	None required for inference

At $2,000/month of API spend, the hardware amortizes in roughly 23 months; above that, it generates net savings sooner. The financial case is straightforward. The sovereignty case is the product: customer data that never leaves the box, fine-tunes that sit on your own storage, no provider deciding what you can run.

5.2 Capability Ceiling

Model	API Equivalent	Runs Here?
GLM-5.2 (395B MoE)	Competitive with Opus 4.8 / GPT-5.5 on coding, agentic tasks	Yes (Q4)
DeepSeek-V3 (671B MoE)	DS-V4-Pro API tier	Yes (Q3)
Llama 4 70B (70B dense)	Frontier dense	Yes (FP16)
GPT-5.5 / Claude Opus 4.8	Closed frontier	No (closed weights)

This build runs every open-weight model at today's frontier. It does not replace closed-source APIs; it gives you a sovereign alternative for the workloads where open weights are good enough and privacy matters.

6 Risks and Mitigations

Risk	Severity	Mitigation
SYCL stack instability	Medium	Vulkan fallback available; both backends benchmarked on target hardware before commitment
llama.cpp master regressions	Medium	Pinned commit, Docker-versioned builds, rollback path via container tags
8-GPU scaling untested on B70	Low-Medium	Published single-B70 benchmarks validate the software stack; llama.cpp GPU splitting is architecture-agnostic. Multi-GPU validation is the first step after the single-card benchmark.
Barebone accessory / harness mismatch	Medium	Verify rail kit, CPU heatsinks, GPU power harness, and B70 connector compatibility before procurement. The ASUS platform reduces MCIO cabling risk but introduces server-accessory and GPU-harness validation.
Intel Arc driver maturity	Medium	B70 is a workstation card (Pro series) with an enterprise driver track; Battlemage Xe2 is second-gen, substantially more mature than first-gen Alchemist. oneAPI toolkit provides a stable compute runtime. Intel Arc & iGPU driver 101.8861 is current as of July 2026.
Multi-model concurrency	Medium	llama.cpp handles one model at a time; for concurrent serving, deploy multiple instances on separate port ranges or accept context-switching overhead. vLLM/SGLang support for Intel Arc is still experimental; watch upstream.
Dual-slot density	Low	The ASUS ESC8000A-E13P is purpose-built for eight dual-slot PCIe accelerators. 230W-class active B70 SKUs should be easier to cool than 425W-class consumer GPUs, but passive/fanless B70 boards still require chassis-specific airflow validation.

7 Rollout Plan

1. **Procurement:** Source 8× Arc Pro B70 (ASRock Creator preferred until MAXSUN power/TBP is verified), dual EPYC 9004/9005 CPUs, and the ASUS ESC8000A-E13P-32WP barebone.
2. **Single-card validation:** Build a one-B70 test rig. Run the lemongravy.me benchmark suite against your target models. Confirm SYCL performance and stability at 230W.
3. **Multi-GPU test:** 2-card GPU split validation; confirm bandwidth scales linearly. Then 4 cards, then 8.
4. **Platform assembly:** Verify the ASUS barebone accessory set, rail kit, CPU heatsinks, GPU power harnesses, and B70 connector compatibility before power-on. Confirm all 8 cards enumerate at PCIe 5.0 x16.
5. **Software deployment:** Docker-based llama.cpp SYCL build with the `seq_rm` hybrid patch. Benchmark multi-GPU throughput against the single-card baseline.
6. **Production cutover:** Move API workloads to local inference, highest-spend first.

References

- ASUS ESC8000A-E13P-32WP product listing, Central Computer (July 2026). centralcomputer.com
- ASUS ESC8000A-E13P GPU Server specifications, ASUS Servers (2026). servers.asus.com
- Intel Arc Pro B70 Specifications, TechPowerUp GPU Database (2026). techpowerup.com
- u/deebuildsthings, "I built a 8x RTX 4090D with 192 VRAM, here's what I learnt," r/LocalAIServers (June 2026). reddit.com
- Lemongravy, "Taming the Battlemage: 63 Tokens/Sec Sustained on Qwen 3.6 with Intel SYCL" (2026). lemongravy.me
- GLM-5.2 Technical Report, Z.ai (2026). arXiv:2602.15763
- llama.cpp, ggml-org. github.com/ggml-org/llama.cpp
- Intel oneAPI Toolkit. intel.com