HARDWARE-TEMPLATE

On-Prem AI Workstation — {{CLIENT_COMPANY}} Office Deployment

Prepared for: {{CLIENT_NAME}} · {{CLIENT_TITLE}}, {{CLIENT_COMPANY}}
Prepared by: Kevan Burns · Founder & CEO, UnyKorn LLC
Contact: kevan@ldrcllc.com
Date: {{DATE}}

Why on-prem, alongside the cloud

Your sovereign AI runs primarily on hardened cloud infrastructure I operate under UnyKorn. That's the right default for reliability and geographic redundancy. **But the highest-value work — confidential deal analysis, LP correspondence, tenant financials, board-level strategy — should also be able to run on hardware that physically sits in your office, air-gapped from the internet whenever you want.** That is what this workstation buys you.

Three things this on-prem machine gives you that pure cloud cannot:

1. **Sovereignty.** Sensitive analysis never leaves the building unless you say so. If we ever unplug the internet cable, this machine keeps working.
2. **Speed on private data.** RAG over your entire document archive (100k+ documents, closing binders, financials, correspondence) runs locally at millisecond latency. No egress costs.
3. **Cost floor.** Once purchased, the incremental cost of a query is electricity. The cloud stays for burst capacity, model access, and geographic failover.

We run a **hybrid design**: routine queries and public-web work → cloud. Confidential work, RAG over private archives, model fine-tuning → on-prem. You choose the routing rule per project.

Recommended build — Lambda Scalar or Bizon G3000-class

Target total: ~$46,200 delivered & configured

Component	Spec	Vendor / SKU	Est. Price
GPU × 2	NVIDIA **RTX PRO 6000 Blackwell Workstation Edition**, 96 GB GDDR7 each	nvidia.com/rtx-pro-6000 · via PNY, Leadtek, ASUS, Lambda, Bizon	$17,000 (2 × $8,500)
CPU	AMD **Threadripper PRO 7995WX** (96-core / 192-thread, 5.1 GHz boost)	amd.com	$9,900
Motherboard	ASUS **Pro WS WRX90E-SAGE SE** (WRX90, 7× PCIe 5.0 x16, 8× DDR5 ECC)	asus.com	$1,400
RAM	**512 GB DDR5-5600 ECC RDIMM** (8 × 64 GB Micron/Samsung/Kingston)	kingston.com KSM56R46BD4PMI-64HAI	$3,400
NVMe (fast tier)	4 × **Samsung 990 Pro 8 TB** NVMe Gen4 (RAID 10 → 16 TB)	samsung.com	$2,400
HDD (bulk tier)	4 × **Seagate Exos X24 24 TB** SATA (RAID 6 → ~48 TB usable)	seagate.com Exos	$1,800
PSU	**Corsair AX1600i** 1600 W Titanium	corsair.com AX1600i	$500
Chassis	Fractal Design **Define 7 XL** (silent, 18 drive bays)	fractal-design.com	$250
CPU cooler	Noctua **NH-U14S TR5-SP6** (sTR5 socket)	noctua.at NH-U14S TR5-SP6	$150
Networking	Onboard 10 GbE + **Mellanox ConnectX-6 Dx** dual-port 25 GbE add-in	nvidia.com ConnectX-6 Dx	$600
UPS	**APC Smart-UPS SRT 3000 VA** (rack/tower, sinewave, SNMP)	apc.com SRT3000RMXLA	$1,800
KVM & display	Small 1080p touchscreen + wireless kbd/mouse for local console	Amazon Business	$300

Component	Spec	Vendor / SKU	Est. Price
Assembly + burn-in + 3-yr warranty	**Lambda** (lambdalabs.com) or **Bizon** (bizon-tech.com) — quote both	—	$6,700
Delivered total			**~$46,200**

Alternative — turnkey NVIDIA reference platform:

- **NVIDIA DGX Station** (Blackwell refresh) — ~$70–100K. Whisper-quiet, 784 GB unified memory, NVLink, InfiniBand out-of-box, NVIDIA AI Enterprise included. Buy via nvidia.com/dgx-station or partners (SHI, CDW).
- **NVIDIA DGX Spark cluster** (2–4 units, ~$4K each = $8–16K) — small, silent, 128 GB unified per node. Great as edge nodes at branch offices later. nvidia.com/dgx-spark.

Kevan's recommendation: the custom Lambda/Bizon build. 192 GB VRAM (more than a single DGX Station) at 40–60% of the cost, serviceable by anyone. If you want NVIDIA's turnkey warranty and support contract, DGX Station is the answer at ~2× the price.

What this machine actually runs

Simultaneously, in production:

- **Llama 3.3 70B** (or Llama 4 when available) in FP8 — the general-purpose brain, ~40 GB VRAM.
- **DeepSeek V3 / Qwen 3** in FP4 — for coding, research, analytical work.
- **Llama 405B** in FP4 (single instance) — heaviest reasoning on private data.
- **Local voice pipeline** — Whisper (STT) + XTTS/Piper (TTS) — assistant works even if internet drops.
- **Embedding + reranking models** — for the RAG index over your document archive.
- **Image + video gen** (ComfyUI + Wan alternatives) — pitch decks, one-pagers.
- **Fine-tuning jobs overnight** — LoRA adapters trained on your own writing style, memos, letters.

Storage capacity:

- 16 TB NVMe fast tier — active models, hot RAG index, working data.
- 48 TB HDD bulk tier — historical documents, archives, recorded meetings.
- Nightly encrypted backup → GCS (UnyKorn GCP billing).

Power / thermal:

- Idle: ~450 W. Peak load (both GPUs): ~1,200 W.

- Draws from a dedicated 20A 120V circuit (or one 15A 220V) — most offices already have this near the server closet.
- Heat output: ~4,100 BTU/hr peak. Fine for any office with normal HVAC.
- Noise: ~30 dB idle, ~42 dB full load with the Define 7 XL chassis.

Software stack (pre-configured before delivery)

Layer	Software
OS	Ubuntu 24.04 LTS Server (unattended-upgrades, full-disk LUKS encryption)
Container runtime	Docker CE + NVIDIA Container Toolkit
GPU orchestration	NVIDIA Fabric Manager + DCGM monitoring
Model serving	vLLM, Text Generation Inference (TGI), Ollama
Chat UI	Open WebUI (private, YubiKey-gated login)
RAG stack	LlamaIndex + Qdrant (local vector DB) + BGE-M3 embeddings
Voice	Whisper.cpp + Piper + Coqui XTTS
Media gen	ComfyUI, Fooocus, Wan-2.1
Observability	Grafana + Prometheus + Loki (local dashboards)
Remote access	Cloudflare Tunnel + Zero Trust (only your phone + Kevan's laptop, YubiKey-gated)

NVIDIA AI Enterprise license — $4,500/GPU/yr × 2 = **$9,000/yr, optional.** Adds enterprise NIM microservices, direct NVIDIA support, tuned model catalog. Recommend Year 2 once we know the workload.

Physical placement & security

- **Location:** locked server closet, or a locked cabinet in a low-traffic area. Not on an open desk.
- **Power:** dedicated 20A 120V circuit, UPS between wall and machine.

- **Network:** CAT6a to the office LAN + separate CAT6a direct to a business-class internet drop. VLAN-isolated from the general office network.
- **Physical security:** Kensington cable lock + case intrusion sensor + PIN-locked BIOS. YubiKey required to boot LUKS. Camera on the closet door.
- **Remote hands:** during the first 90 days, Kevan is on-site or on-call for any hardware event. After that, break-fix contract with a local IT partner.

Timeline & payment

Milestone	Timing
Order placed (Lambda or Bizon)	Day 0
Build & burn-in at vendor	3–4 weeks
Delivery to {{CLIENT_COMPANY}} office	Week 4
On-site installation + software deployment	1 day
RAG indexing of your document archive	Week 5 (overnight jobs)
Full production	Week 6

Payment options:

- **A.** UnyKorn LLC purchases the machine, hosts it at your office under a monthly hardware-as-a-service line item (~$1,500/mo over 36 months, ownership transfers to {{CLIENT_COMPANY}} at month 36). Best if you want to keep this as OpEx with no capex spike.
- **B.** {{CLIENT_COMPANY}} purchases directly, UnyKorn manages. Best for capex allocation and Section 179 depreciation.

Both are fine. Tell Kevan which fits your accounting and paperwork gets structured accordingly.

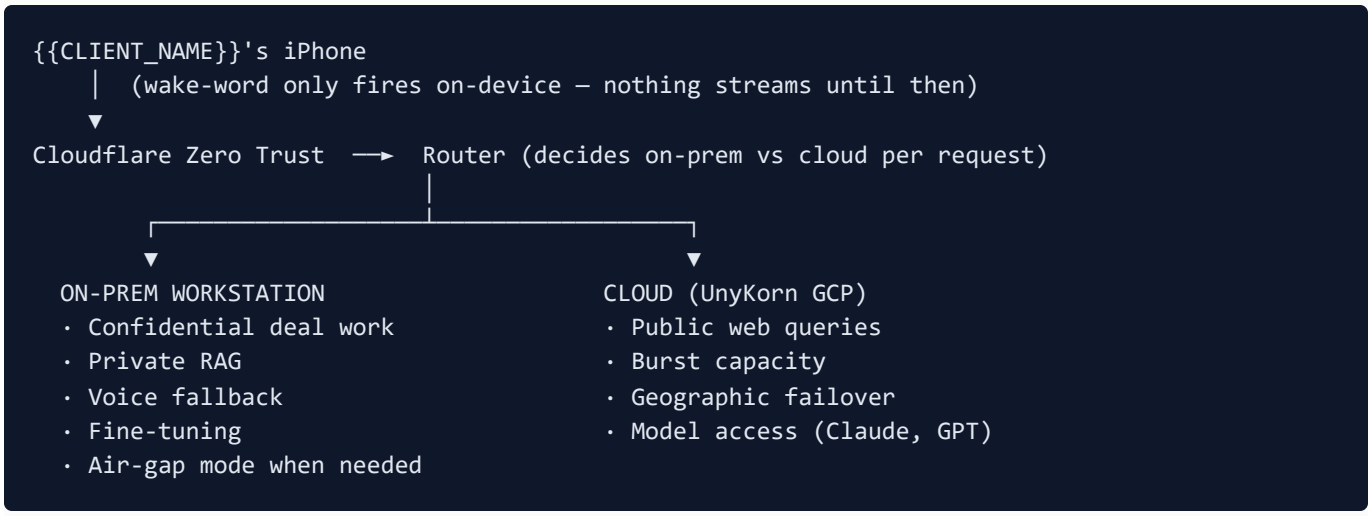
What we're *not* recommending (and why)

- **Cloud-only.** Faster to deploy, but you never get true sovereignty over sensitive work, and inference costs scale with usage forever.
- **A single consumer RTX 5090 machine.** ~$3K instead of $46K, but only 32 GB VRAM. Cannot host frontier models at full precision, no ECC memory, no enterprise warranty. Fine for a hobbyist,

wrong for a principal's production environment.

- **A datacenter H200/B200 build.** ~$300K+, requires liquid cooling and 3-phase power. Overkill for one office — that's the tier you buy when you're renting compute to third parties.
- **Fully on-prem, no cloud.** Loses geographic redundancy and burst capacity. The hybrid design is strictly better.

Hybrid architecture at a glance



Return questions to **kevan@ldrcllc.com**. Happy to walk this through live before you decide on Path A vs Path B.

— Kevan Burns
Founder & CEO, UnyKorn LLC