On-Premises Hardware Sizing

Infrastructure Specifications for On-Premise Deployments

Overview

This guide provides hardware sizing specifications required to support different numbers of users and data volumes for on-premises deployments.

The primary parameter to consider is the number of concurrent users. Any GPU-based sizing can support a large number of users; however, if many users submit requests simultaneously, some may experience delays while the GPU processes previous prompts.

Single Unit Deployments

Single unit deployments require two virtual or physical servers — one Windows server and one Linux server to host Docker containers. The Docker containers may be collocated on an existing Linux Docker host.

Component	Small	Medium	Large	Large Plus Storage Optimized	X-Large
Licensed Users	150 (Data Analysis not supported)	600	2,000	2,000	3,000
Concurrent Users	5	20	40	40	100
Data Supported	500 GB	700 GB	1.4 TB	30 TB	30 TB
CPU	12 Cores i7 / Xeon	16 Cores i7 / Xeon	32 Cores i7 / Xeon	64 Cores i7 / Xeon	64 Cores i7 / Xeon
Memory	48 GB DDR4+	64 GB DDR4+	128 GB DDR4+	256 GB DDR4+	256 GB DDR4+
NVMe SSD	256 GB	1 TB	4 TB	8 TB (Source Files not stored) 48 TB (Files Stored)	8 TB (Source Files not stored) 48 TB (Files Stored)
Consumer GPU	1 x RTX 4090 24GB 1 x RTX 4070 Ti 12GB	4 x RTX 4090 24GB 1 x RTX 4070 Ti	8 x RTX 4090 24GB 2 x RTX 4070 Ti	10 x RTX 4090 24GB	—
Data Center GPU	2 x L4 24 GB	1 x H100NVL GPU	2 x H100NVL GPU	2 x H100NVL GPU	3 x H100NVL GPU
LLM Options	1x GPT OSS 20B 1 Embedding model	1x GPT OSS 120B 1 Embedding model	2x Gemma 27B OR 2x GPT OSS 120B Embedding model supported	2x Gemma 27B OR 2x GPT OSS 120B Embedding model supported	3x Gemma 27B OR 3x GPT OSS 120B Embedding model supported

Enterprise Deployments

Component	Enterprise 1	Enterprise 2
Licensed Users	6,000	10,000
Concurrent Users	500	1,000
Data Supported	60 TB	100 TB
Windows Servers	2x Active-Active (may be VMs) Per server: 32 GB RAM | 16 Cores | 200 GB SSD/NVMe	2x Active-Active (may be VMs) Per server: 64 GB RAM | 32 Cores | 200 GB SSD/NVMe
Linux Servers	2x Active-Active (may be on one host with multiple instances or separate hosts) Total resources required: 512 GB RAM | 128 Cores | 16 TB SSD/NVMe 4 x H100 / H200	2x Active-Active (may be on one host with multiple instances or separate hosts) Total resources required: 768 GB – 1 TB DDR4+ RAM | 192 Cores | 24 TB SSD/NVMe 6 x H100 / H200

LLM GPU Compatibility

Model	Min VRAM	Recommended Single-GPU	Multi-GPU (Tensor Parallel)	Notes / Requirements
OSS 20B	16–24 GB	RTX 3090 / 4090 (24 GB) RTX A5000 (24 GB) A10 (24 GB), A40 (48 GB) A100 (40/80 GB), H100	2 × 16–24 GB (TP=2) 4 × 16 GB (TP=4) 2 × 24 GB (TP=2)	CUDA 12.x, Tensor Cores, Compute Capability ≥ 7.0. NVLink recommended for TP.
OSS 120B	80 GB+	A100 80 GB, H100 80 GB H200, MI300X	2 × 80 GB (TP=2) 4 × 48 GB (TP=4) 4 × 80 GB (TP=4, ideal)	Designed for 80 GB-class GPUs. PCIe-only not supported. Requires CUDA 12.x & CC ≥ 7.0.
Gemma 3 27B	48 GB	RTX 6000 Ada (48 GB) A40 (48 GB) A100 80 GB, H100	2 × 24 GB (TP=2) 2 × 48 GB (TP=2) 4 × 24 GB (TP=4)	~54 GB in FP16; ≥ 48 GB VRAM recommended for full precision. CUDA 12.x & CC ≥ 7.0.

General GPU Requirements

- Compute Capability: ≥ 7.0 (Volta or newer)
- Tensor Cores: Required (Volta / Turing / Ampere / Ada / Hopper)
- CUDA Version: CUDA 12.x required for vLLM / SGLang / Triton / FlashAttention
- VRAM Pooling: VRAM does not combine across GPUs; multi-GPU requires Tensor Parallelism (TP)
- Interconnect: NVLink strongly recommended for multi-GPU TP on 20B+ models
- Not Supported: Maxwell (M10 / M60 / M40), Pascal (P100 / P40 / GTX 10xx), GPUs with < 12 GB VRAM

High Availability Considerations

The following components must be considered when designing a highly available solution:

- Dashboard (User Interface) + Background Service — Two Windows servers required. May be load balanced with Active-Active or Active-Passive configurations.
- Gateway — Requires two Linux hosts. Paired with one Dashboard / BG Service each.
- LLMs — At least 2 embedding GPUs and 2 LLM GPUs. Active-Active configuration recommended to maximise routine GPU utilisation for speed.
- Databases — Postgres and MS SQL databases will need to be configured to run HA replicas.

For technical enquiries and deployment support, visit www.defenix.ai