On-Premise System Requirements

Full System Specifications for On-Premises Deployments

Containers

Purpose	Option 1 (GPU)	Option 2 (CPU) POC with LLM running on a CPU
Gateway (Linux) (Vector DB, SQL DB, Gateway API)	16 GB RAM (32 GB for larger databases) 4 cores 150 GB SSD	16 GB RAM 4 cores
Embedder (Linux)	16 GB RAM 16 GB GPU 4 cores 60 GB SSD	16 GB RAM 4 cores
LLM (Linux)	16 GB RAM 24 GB GPU 4 cores 80 GB SSD	16 GB RAM 8 cores
Dashboard / Ingestor (Windows Server — Not a container)	8 GB RAM 4 cores 80 GB SSD	8 GB RAM 4 cores

- Example CPU processors: Intel Xeon Platinum 8000.
- Example GPU for embedding: NVIDIA T4 Tensor Core.
- Example GPU for LLM: NVIDIA L4 Tensor Core.
- Database can be installed on the Windows server or any other convenient location, including an existing instance of Microsoft SQL Server.
- The Linux containers may be deployed on one server or spread over multiple servers.
- Recommended host OS: Ubuntu Linux 22.04 LTS+

Sizing

The above server hardware requirements are estimated to be sufficient for:

- 50 GB of ingested data
- 300 licensed users
- 20 concurrent users asking questions / 75 questions within a 15-minute period

Add more LLM containers with GPUs to support more concurrent users.

AI Model Servers

The main hardware cost of the DefenGPT deployment is the AI servers responsible for answering questions. These utilize GPUs and are also used to create embeddings for processing and preparing data for AI.

Server Specification

Linux server Ubuntu — 2 CPU, 8 GB RAM, HDD/SSD with R/W speed of at least 100 MB/s.

GPU: CUDA 11.8+, minimum 24 GB VRAM.

The disk size should be 30% larger than the original content base file size.

Multiple servers with a load balancer may be used for higher performance and high availability.

Graphics Cards

The following GPU options are available for standard servers. The difference between cards is the number of questions per minute that can be processed. The system supports multiple servers with a load balancer to boost performance.

Card	Answer Speed	Time to Answer 10 Simultaneous Questions	Purchase Cost* (One-Time)
H100 80GB SXM5	1.9 sec	15 sec	$33,000
NVIDIA RTX 4090 24 GB VRAM 4 vCPU	2.1 sec (29 Questions / Min)	17 sec	$2,289
NVIDIA RTX 4070 Ti 12 GB VRAM 8 vCPU	2.5 sec (14 Questions / Min)	25 sec	$790

** Costs were taken from Amazon.com; however, these graphics cards can also be purchased elsewhere.*

*** Approximate times when using Wizard-Vicuna-13B model.*

**** Multiple medium-range cards are preferable to a lower quantity of more expensive cards.*

Dashboard Website

Windows Server 2016 Enterprise (or higher), 4 CPU, 8 GB RAM, 50 GB disk space, IIS installed. HDD/SSD with R/W speed of at least 50 MB/s.

Ingestor Service

May be co-located with the Dashboard server.

Windows Server 2016 Enterprise (or higher), 4 CPU, 8 GB RAM, 50 GB disk space, IIS installed. HDD/SSD with R/W speed of at least 50 MB/s.

Dashboard SQL DB

Configured with SQL file storage.

Windows Server 2016 Enterprise (or higher), 4 CPU, 8 GB RAM, IIS installed. HDD/SSD with R/W speed of at least 50 MB/s.

In general, integrated content is not saved in this database. Instead, the product keeps a link to the source file.

In the case where files are uploaded manually to the Dashboard, they are saved in the Dashboard database as-is (same size).

Gateway Server

Linux server Ubuntu — 4 CPU, 8 GB RAM. HDD/SSD with R/W speed of at least 100 MB/s.
The disk size should be 30% larger than the original content file size.

The server can optionally include an NVIDIA GPU with a minimum of 14 GB VRAM for embedding content.

Offline Installation — File / Container Sizes

Component	Image Size
Gateway Image	7.5 GB
Vector DB Image	500 MB
Microsoft SQL DB Image (Optional)	1.6 GB
LLM Image with Model	17 GB
BG Service + Dashboard Installation Package	250 MB

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