SUSE AI requirements

WHAT?

Hardware, software and networking requirements for successful deployment and operation of SUSE AI.

WHY?

To ensure efficient operation of SUSE AI.

EFFORT

Less than 30 minutes of reading and a basic knowledge of hardware and RKE2 environment.

This page describes the software, hardware and networking requirements for the cluster nodes where you plan to install SUSE AI.

Publication Date: 12 Nov 2024

Contents

- 1 Hardware requirements 2
- 2 GPU hardware 3
- 3 Application-specific requirements 4
- 4 Legal Notice 9

Glossary 10

A GNU Free Documentation License 10

*

1 Hardware requirements

For successful deployment and operation, SUSE AI has the same hardware prerequisites as RKE2 cluster.

1.1 Recommended hardware (basic functionality)

RAM

At least 32 GB of RAM per node. This is the minimum recommendation for the control plane node. Additional resources may be needed for the worker nodes based on workload.

CPU

A multi-core processor with the minimum of 4 cores. 8 cores or more may be necessary depending on the cluster scale and application demands.

Disk space

- 50 GB or more is recommended for control plane nodes.
- Additional space for data storage, such as application data or log files, are required depending on the deployment scale and the workloads running on the cluster.
- SSDs or high-speed storage is preferred for faster data access and efficient operation of containerized workloads.

Networking

- A reliable and stable network connection between all nodes in the cluster.
- Cluster nodes need to be able to communicate with each other and access external resources, such as container images or software updates.
- Ensure that all nodes have public IPs or are accessible via VPN or other private network if deploying across multiple data centers.

1.2 Recommended hardware (for High Availability)



Important

While 32 GB of RAM is the minimum for basic functionality, a production-grade deployment with high availability, multi-node clusters or running resource-intensive applications like AI/ML workloads might require more.

RAM

64 GB or more per node is recommended for larger clusters or to run applications with high resource demands.

CPU

At least 8 cores, ideally 16 or more cores depending on the expected load.

Disk space

- For larger-scale clusters or persistent storage applications, 100 GB or more of disk space per node might be required.
- Using high-performance SSDs is recommended, especially for workloads with high I/O requirements such as databases or AI/ML model training.

Networking

Ensure a low-latency, high-throughput network for efficient communication between nodes, especially if deploying in multi-region or multi-cloud environments.



Тір

For more detailed and specific hardware recommendations, refer to the official RKE2 installation requirements documentation at https://docserv.suse.de:8085/cloudnative/rke2/ latest/en/install/requirements.html

2 GPU hardware

To run AI/ML workloads such as training machine learning models or running inference workloads, deploy cluster nodes with compatible NVIDIA GPUs to gain acceleration.

2.1 Using the NVIDIA GPU Operator

Configuring and managing nodes with hardware resources can require multiple configurations for software components. These include drivers, container runtimes and libraries. To use NVIDIA GPUs in a Kubernetes cluster, you need to configure the NVIDIA GPU Operator. Because GPU is a special resource in the cluster, you need to install the following components to enable deployment of workloads for processing on the GPU.

- NVIDIA drivers (to enable CUDA)
- Kubernetes device plug-in
- Container runtime
- Other tools to provide capabilities such as monitoring or automatic node labelling.

To ensure that the NVIDIA GPU Operator is installed correctly, the Kubernetes cluster must meet the following prerequisites:

- All worker nodes must run the same operating system version to use the NVIDIA GPU Driver container.
- Nodes must be configured with a container engine such as Docker (CE/EE), containerd, or Podman.
- Nodes should be equipped with NVIDIA GPUs.
- Nodes should have NVIDIA drivers installed.

2.2 Supported GPUs

The NVIDIA GPU Operator is compatible with a range of NVIDIA GPUs. For a full list of supported GPUs, refer to NVIDIA GPU Operator Platform Support (https://docs.nvidia.com/datacenter/cloud-native/gpu-operator/latest/platform-support.html) **?**.

3 Application-specific requirements

The SUSE AI stack consists of multiple applications. We recommend running each application on nodes that meet or exceed the corresponding hardware requirements.

3.1 Milvus hardware and software requirements

This topic describes requirements for the Milvus application.

3.1.1 Hardware requirements

3.1.1.1 Minimum requirements

The following requirements are for basic Milvus deployment on a single node or a small scale.

RAM

A minimum of 32 GB of RAM.

CPU

At least 8 CPU cores.

Disk space

At least 100 GB of storage, preferably SSD.

Networking

A stable connection with 1 Gbps network bandwidth.

3.1.1.2 Recommended hardware for large-scale workloads

The following requirements are for multi-node Milvus clusters or heavy workloads, such as large vector databases.

RAM

A minimum of 64 GB of RAM per node.

CPU

8-16 CPU cores per node or more.

Disk space

500 GB or more of high-speed storage, ideally SSD or NVMe SSD.

Networking

10 Gbps Ethernet or faster for high-performance clusters.

*

3.1.1.3 CPU instruction set requirements

The following CPU instruction sets are required for Milvus:

- SSE4.2
- AVX
- AVX2
- AVX-512

3.1.2 Software requirements

Running Milvus requires specific versions of the following software:

Kubernetes

SUSE supported versions of RKE2 that use Kubernetes 1.18 or higher.

Helm

The recommended version is 3.5.0 or later.

3.1.3 Additional considerations

Disk and storage

- **Storage type:** SSDs or NVMe SSDs are highly recommended for fast read/write access to large datasets and high-performance vector retrieval.
- **Metadata and data storage:** For large-scale deployments, ensure that metadata and vector data is stored on fast disks (SSD or NVMe).

Network

For high-performance clusters, especially for large-scale deployments, ensure high-bandwidth network connectivity between nodes.

3.1.4 For more information

For more detailed and specific hardware recommendations, please refer to the official Milvus (https://milvus.io/docs/prerequisite-helm.md) and pre-requisite docker (https://milvus.io/docs/v2.1.x/prerequisite-docker.md) documentation.

*

3.2 Ollama hardware and software requirements

The version of Ollama provided with SUSE AI is optimized for NVIDIA GPU hardware. This section guides you through the steps for configuring Ollama on an NVIDIA-enabled system, including necessary configurations for both the hardware and software.

Tip: General recommendations

- **Run Ollama on NVIDIA GPU nodes.** Since Ollama is GPU-optimized, utilizing the power of NVIDIA GPUs is essential for maximum performance. This ensures that the application runs efficiently and makes full use of the hardware capabilities.
- Assign applications to specific nodes. SUSE AI provides a mechanism to assign applications such as Ollama to specific nodes. For more details on how to assign applications to nodes, refer to

3.2.1 Hardware requirements

NVIDIA GPU

- The recommended GPU models include Tesla, A100, V100, RTX 30 series, or other compatible NVIDIA GPUs.
- Ensure that the CUDA Compute Capability of your GPU is compatible with the required version of Ollama.

RAM

At least 16 GB of RAM is recommended. However, higher amounts (32 GB or more) may be necessary for larger models or workloads.

Disk space

At least 50 GB of free disk space is recommended for storing the container images and any data files processed by Ollama.

3.2.2 Software requirements

NVIDIA Docker (nvidia-docker)

• You must install https://docserv.suse.de:8085/suse-ai/1.0/html/ NVIDIA-Operator-installation/index.html https://docserv.suse.de:8085/suse-ai/1.0/html

CUDA Toolkit

You must install the CUDA version supported by your GPU model. For most recent GPUs, CUDA 11.0 or later is required. Refer to CUDA Toolkit installation guide (https://developer.n-vidia.com/cuda-toolkit) a for more details.

NVIDIA driver

Install the NVIDIA driver compatible with your GPU model. Its version must be compatible with the installed CUDA toolkit.



You can check your GP driver version by running the nvidia-smi command.

3.3 Open WebUI hardware and software requirements

While Open WebUI has no specific hardware dependencies beyond those of the underlying platform, consider the following guidelines for optimal performance.

- Because Open WebUI shares most dependencies with Milvus and follow Ollama. the hardware requirements mentioned in Milvus (https://docserv.suse.de:8085/suse-ai/1.0/html/Al-requirements/index.htmrequirements l#ai-requirements-milvus) and Ollama requirements (https://docserv.suse.de:8085/suseai/1.0/html/AI-requirements/index.html#ai-requirements-ollama) 🗗
- Stable network connection is essential, particularly if Open WebUI is integrated with other services or databases. Ensure sufficient bandwidth for Web traffic and API calls.
- For interacting with the Open WebUI interface, use standard Web browsers such as Google Chrome, Mozilla Firefox or Microsoft Edge.

3.4 Apache Pulsar software requirements

Installation of Apache Pulsar and its dependencies may require updating the SELinux policy as follows:

Create a new file with the <u>.te</u> extension, for example, <u>pulsar.te</u>, with the following content:

 Use the <u>checkmodule</u> and <u>semodule_package</u> commands to compile and package the pulsar.te file into a loadable module.

> checkmodule -M -m -o pulsar.mod pulsar.te
semodule_package -o pulsar.pp -m pulsar.mod

3. Load the pulsar.pp module to apply the policy.

> sudo semodule -i pulsar.pp

4 Legal Notice

Copyright© 2006–2024 SUSE LLC and contributors. All rights reserved.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or (at your option) version 1.3; with the Invariant Section being this copyright notice and license. A copy of the license version 1.2 is included in the section entitled "GNU Free Documentation License".

For SUSE trademarks, see https://www.suse.com/company/legal/ \mathbb{Z} . All other third-party trademarks are the property of their respective owners. Trademark symbols (\mathbb{R} , \mathbb{T} etc.) denote trademarks of SUSE and its affiliates. Asterisks (*) denote third-party trademarks.

All information found in this book has been compiled with utmost attention to detail. However, this does not guarantee complete accuracy. Neither SUSE LLC, its affiliates, the authors, nor the translators shall be held liable for possible errors or the consequences thereof.

Glossary

AI

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are designed to learn and solve problems like humans. AI enables computers to understand language, making decisions, and improving from experience.

GenAl

Generative AI (GenAI) is a type of artificial intelligence that can create new content such as text, images or music.

NLG

Natural Language Generation (NLG) is a process of automatically generating human-like text from structured data or other forms of input. NLG systems are designed to convert raw data into coherent and meaningful language easily understood by humans.

NLU

Natural Language Understanding (NLU) is a process the AI uses to analyze and understand the meaning of the input query.

A GNU Free Documentation License

Copyright (C) 2000, 2001, 2002 Free Software Foundation, Inc. 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA. Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or non-commercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent

copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- **B**. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- **G**. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through

arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Ac-knowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See https://www.gnu.org/ copyleft/ .

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

Copyright (c) YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.