Welcome to Onyx

Resourcing

Recommended CPU / RAM / Disk to run Onyx

Running Locally

When running locally through Docker, we recommend making at least 4vCPU cores and 10GB of RAM available to Docker (16GB is preferred). This can be controlled in the **Resources** section of the Docker Desktop settings menu.

Single Cloud Instance

For small-mid scale deployments, we generally recommend setting everything up on a single instance (e.g. an AWS EC2 instance, a Google Compute Engine instance, an Azure VM, etc.) via Docker Compose as it's the simplest way to get started. For a step-by-step guide on how to do this, checkout our **EC2 deployment guide**.

For most use cases a single reasonably sized instance should be more than enough to guarantee excellent performance. A single instance should be able to effectively serve a small-medium sized organization without issue.

If you go with this approach, we recommend:

CPU: >= 4 vCPU cores (we recommend >=8 vCPU cores if possible, this scales with the number of documents you index)

Memory: >= 16 GB of RAM (this also scales with the number of documents you index)

Disk: >= 50 GB + ~2.5x the size of the indexed documents. Disk is generally very cheap, so we would recommend getting extra disk beyond this recommendation to be safe.

onyx

›

Note: Vespa, used by Onyx for document indexing, requires Haswell (2013) or later CPUs. For older CPUs, use the `vespaengine/vespa-generic-intel-x86_64` image in your `docker-compose.yml`. This generic image is slower but ensures compatibility. For details, see **Vespa CPU Support**. To clean up these unused images, run: `docker system prune --all`.

Kubernetes / AWS ECS

If you prefer to give each component it's own dedicated resources for more efficient scaling, we recommend giving each container access to at least the following resources:

`api_server` - 1 CPU, 2Gi Memory

`background` - 2 CPU, 8Gi Memory

`indexing_model_server` / `inference_model_server` - 2 CPU, 4Gi Memory

`postgres` - 2 CPU, 2Gi Memory

`vespa` - >=4 CPU, >= 8Gi Memory. This is the bare minimum for a production deployment, and we would generally recommend higher than this. The resources required here also scales linearly with the number of documents indexed. For reference, with 50GB of documents, we would generally recommend at least 10 CPU, 20Gi Memory + tuning the `VESPA_SEARCHER_THREADS` environment variable. See the **How Resource Requirements Scale** section below for more details.

`nginx` - 250m CPU, 128Mi Memory

All together, this comes out to a total available node size of at least 14 vCPU and 23GB of Memory.

onyx

›

responsible for storing the raw text, embeddings, and handling search requests.

Based on our experience with other large deployments, Vespa needs ~3GB of RAM for each additional 1GB of documents indexed and ~0.5 additional CPU core for each 1 GB of documents indexed. This is on top of the base requirements of 4 CPU cores and 16GB of RAM. Some notes:

> The "1GB of documents" refers to the raw text of the documents. A large PDF may be a few MB in raw size, but this is primarily due the images contained within. The actual text content of the document is likely much less.
>
> This is a rough estimate, and depends on the dimensionality of the embedding model used. There are also techniques that we support like quantization and dimensionality reduction that can reduce the memory requirements.

Example

For a deployment with 10GB of text content, you can expect the following resource requirements for the `index` component:

> **CPU:** 4 + 10 * 0.5 = 9 cores
>
> **Memory:** 4 + 10 * 3 = 34GB

If deploying with a single instance, this would be on top of the base requirements, so we would recommend adding an additional 4 cores and 16GB of RAM to the instance. Overall, that would take us to >= 13 cores and >= 50GB of RAM. To meet these requirements, something like a m7g.4xlarge or a c5.9xlarge instance would be appropriate.

If deploying with Kubernetes or AWS ECS, this would give a per-component resource allocation of:

> **api_server:** 1 CPU, 2Gi Memory
>
> **background:** 2 CPU, 8Gi Memory

onyx

vespa: 10 CPU, 34Gi Memory

which comes out to a total available node size of at least 17 CPU and 66GB of Memory.

Quickstart

Slack Bot Setup