

# System information:

RHOAI 2.25

OCP 4.18.11

Spyre Operator v0.2.1

# Setup Instructions

1. Install Spyre operator and OpenShift AI from OpenShift operator Hub

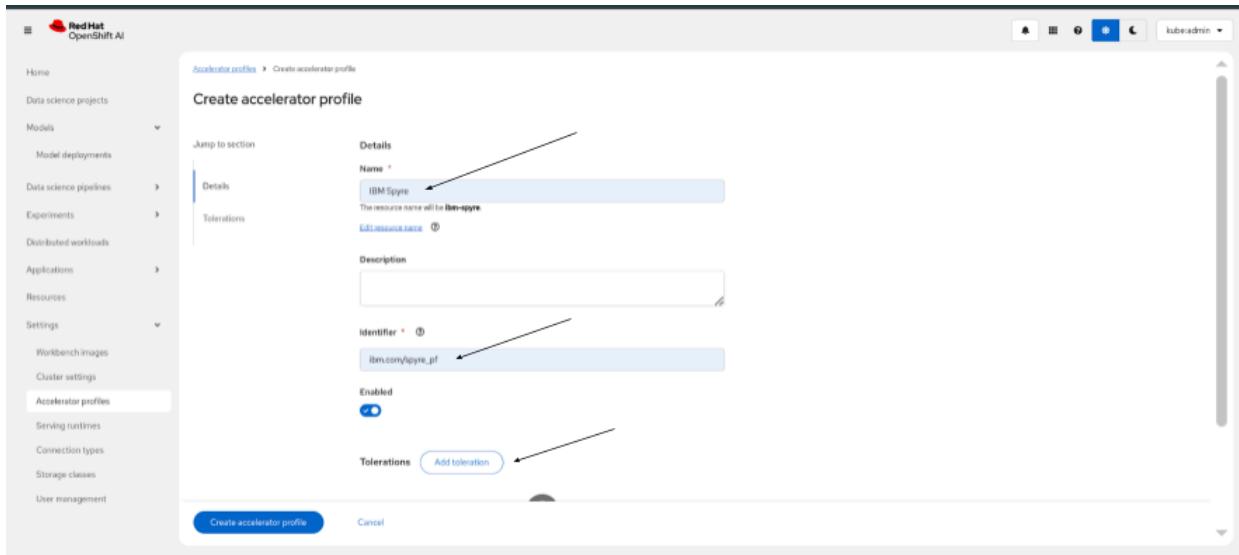
[https://catalog.redhat.com/en/software/containers\(ibm-aiu/spyre-operator/688a1121575e62c686a471d4](https://catalog.redhat.com/en/software/containers(ibm-aiu/spyre-operator/688a1121575e62c686a471d4)

<https://catalog.redhat.com/software/container-stacks/detail/63b85b573112fe5a95ee9a3a>

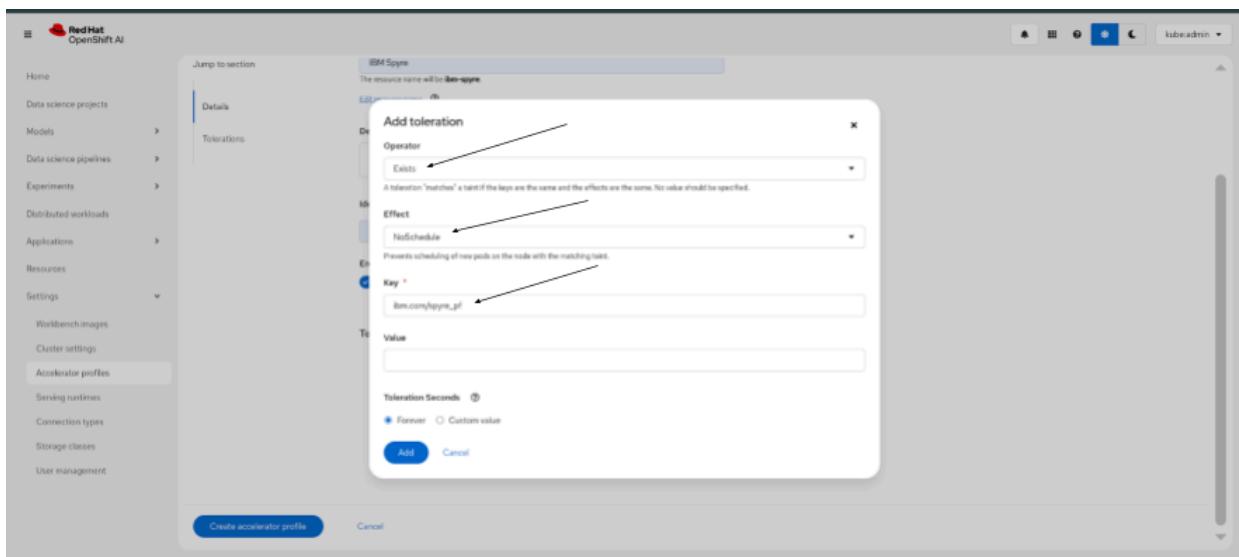
2. Create AcceleratorProfile in OpenShift AI Dashboard settings.

Name	Identifier	Enable	Last modified
IBM Spyre	ibm.com/aiu_pf	On	...

- a. Add Name and identifier



b. Add tolerations and click Add.



3. Add custom ServingRuntime in OpenShift.

Red Hat OpenShift

Select your new project

Administrator

Home Overview Projects Search API Explorer Events Operators Workloads Pods Deployments DeploymentConfigs StatefulSets Secrets ConfigMaps CronJobs Jobs DaemonSets ReplicaSets

Project: vllm-spyre-scheduler

Create ServingRuntime

Copy-paste serving runtime

```

1 apiVersion: serving.kserve.io/v1alpha1
2 kind: ServingRuntime
3 metadata:
4   name: vllm-spyre-runtime
5 annotations:
6   openShift.io/display-name: vLLM IBM Spyre ServingRuntime for KServe
7   openShift.io/recommended-accelerators: ["ibm.com/spyre_pf"]
8 labels:
9   openShiftHub.io/dashboard: "true"
10 spec:
11   annotations:
12     prometheus.io/port: "8000"
13     prometheus.io/path: "/metrics"
14   multiModel: false
15   supportedModelFormats:
16     - autoSelect: true
17     - name: VLLM
18   containers:
19     - name: kserve-container
20       image: registry.redhat.io/rhaisis/vllm-spyre-mne0-3.2.2
21       command:
22         - /bin/bash
23         - -
24         - source /etc/profile.d/ibm-aiu-setup.sh
25         - exec python3 -m vllm entropypoints openapi_server "80"
26       args: -----
27

```

Create Cancel

4. Add registry pull-secret - Pull secrets are required to fetch images from [registry.redhat.io](#) where vLLM images and pre-built modelcar containers are kept e.g. oci://[registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct:1.5](#) and oci://[registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct:1.5](#) . This example uses a model from OCI registry  
[https://docs.redhat.com/en/documentation/red\\_hat\\_ai\\_inference\\_server/3.0/html/validated\\_models/red\\_hat\\_ai\\_validated\\_models](https://docs.redhat.com/en/documentation/red_hat_ai_inference_server/3.0/html/validated_models/red_hat_ai_validated_models)

Use the pull secrets provided to IBM for this step

- a. Obtain access to [registry.redhat.io](#)
- b. Login using podman. Enter your username and password

None

```
podman login registry.redhat.io
```

When you log into the registry, your credentials are stored in your \${XDG\_RUNTIME\_DIR}/containers/auth.json file. Those credentials are used automatically the next time you pull from that registry. Here is an example of that file:

Shell

```
{ "auths": { "https://registry.redhat.io": { "auth": "c2xmams6c2RmbGtq" } } }
```

- c. Alternatively, you can create the authentication using this command.  
Which is a combination of your username and password.

Shell

```
AUTH_STRING=$(echo -n "YOUR_USERNAME:YOUR_PASSWORD_HERE" | base64)  
echo '{  
  "auths": {  
    "registry.redhat.io": {  
      "auth": "'${AUTH_STRING}'"  
    }  
  }  
'
```

- d. Alternatively apply from the terminal

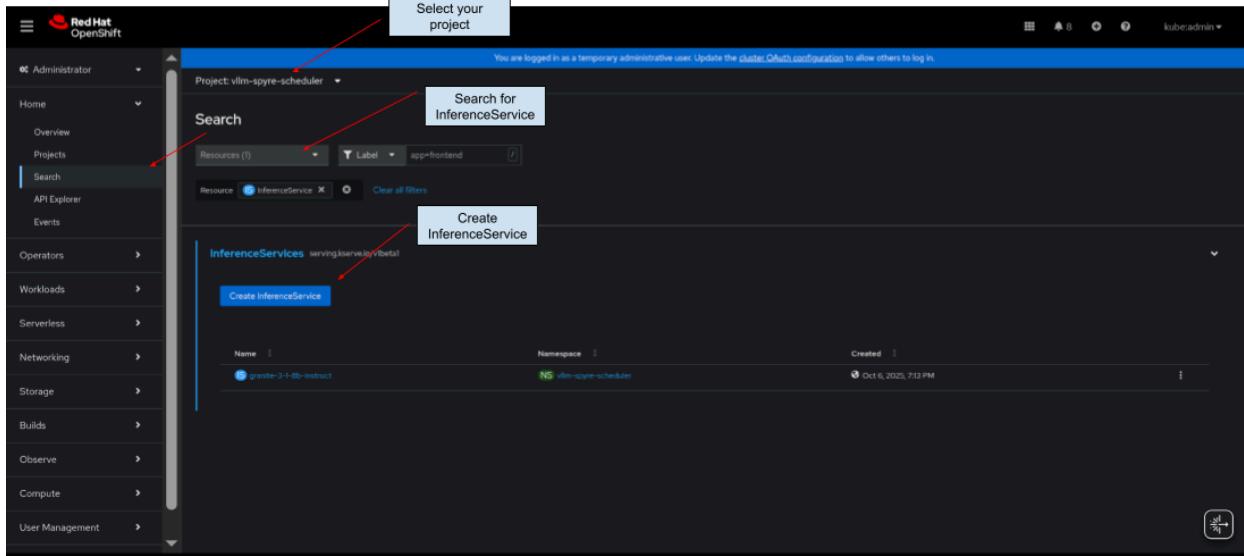
Shell

```
oc create secret docker-registry v11m-registry \  
--docker-server=registry.redhat.io \  
--docker-username=YOUR_USERNAME \  
--docker-password='YOUR_PASSWORD' \  
-n your-namespace
```

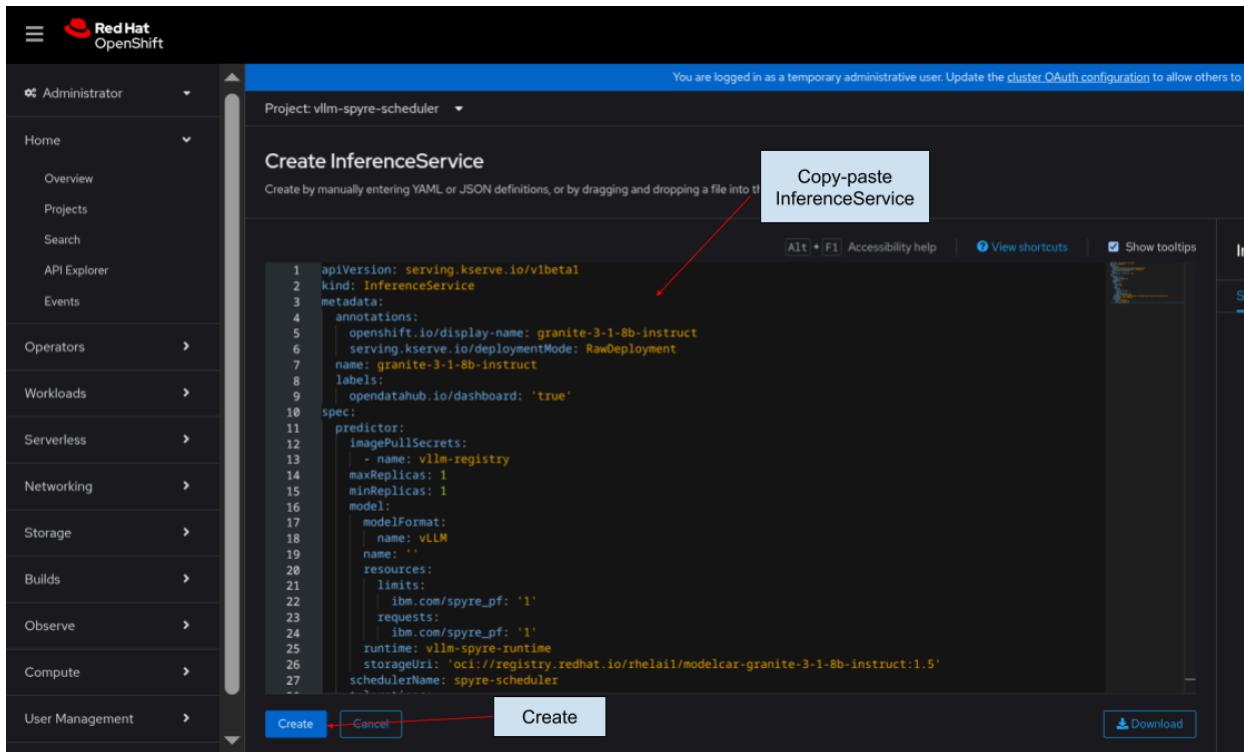
- e. If working on a shared cluster, follow the link below to enable Registry Service Accounts for Shared Environments

<https://access.redhat.com/articles/RegistryAuthentication>

5. Create [InferenceService - RawDeployment](#) or [InferenceService - Serverless](#) in Openshift. Note that RHOAI Dashboard is missing the scheduler feature required for Spyre hardware profiles so we're using OpenShift for now.
  - a. Search for InferenceService in your project and Create a new InferenceService



b. Copy-paste [InferenceService Yaml](#) into the editor and click "create".



c. Verify that model is deployed

Project: vllm-spyre-scheduler

Created at: Oct 8, 2025, 11:08 AM

Owner: RS granite-3-1-8b-instruct-predictor-66dd9cf9c5

PodDisruptionBudget: No PodDisruptionBudget

Receiving Traffic: 0

**Init containers**

Name	Image	State	Last State	Restarts	Started	Finished	Exit code
modelcar-init	registry.redhat.io/helix/modelcar-gr...	Terminated		0	Oct 8, 2025, 11:08 AM	Oct 8, 2025, 11:08 AM	0

**Containers**

Name	Image	State	Last State	Restarts	Started	Finished	Exit code
iserve-container	registry.redhat.io/helix/vlm-spyre-rh...	Running		0	Oct 8, 2025, 11:08 AM	-	-
modelcar	registry.redhat.io/helix/modelcar-gr...	Running		0	Oct 8, 2025, 11:08 AM	-	-

**Volumes**

Name	Mount path	SubPath	Type	Permissions	Utilized by
iserve-provision-location	/mnt	No subpath	Read/Write		iserve-container

## 6. Model Inference

### a. Find the URL endpoint for model inference

Project: vllm-spyre-scheduler

You are logged in as a temporary administrative user. Update the cluster OAuth configuration to allow others to log in.

**Routes**

Name	Status	Location	Service
grainite-3-1-8b-instruct	Accepted	https://grainite-3-1-8b-instruct-vlm-spyre-scheduler.apps.spyre-001.nvidia.eng.rdu2.dc.redhat.com	grainite-3-1-8b-instruct-predictor

### b. Send an inference request to the deployed model endpoint using any REST client of your choice.

Shell

```
curl -k
"https://grainite-3-1-8b-instruct-vlm-spyre-scheduler.apps.spyre-001.nvidia.eng.rdu2.dc.redhat.com/v1/completions" \
-H "Content-Type: application/json" \
-d '{"model":"grainite-3-1-8b-instruct","prompt":"Write a short poem.","temperature":0,"max_tokens":128}' \
| jq
```

```
% Total     % Received % Xferd  Average Speed   Time      Time      Time  Current
                                         Dload  Upload   Total  Spent   Left  Speed
100  897  100    798  100     99      38       4  0:00:24  0:00:20  0:00:04    180
{
  "id": "cmpl-4a2c14e12cf2401ca01a12aa03b28ebc",
  "object": "text_completion",
  "created": 1762888026,
  "model": "granite-3-1-8b-instruct",
  "choices": [
    {
      "index": 0,
      "text": "\n\nIn the quiet of the night, under the silver moon's
glow,\nStars twinkle like secrets, in the sky they sow.\nWhispers of the wind,
through the trees they weave,\nA symphony of silence, in the world we
believe.\n\nDreams take flight, on wings of the night,\nIn the canvas of
darkness, they ignite.\nTomorrow's promise, in the stars we see,\nIn the heart
of the night, hope is free.",
      "logprobs": null,
      "finish_reason": "stop",
      "stop_reason": null,
      "prompt_logprobs": null
    }
  ],
  "service_tier": null,
  "system_fingerprint": null,
  "usage": {
    "prompt_tokens": 6,
    "total_tokens": 120,
    "completion_tokens": 114,
    "prompt_tokens_details": null
  },
  "kv_transfer_params": null
}
```