©2023 NVIDIA Corporation & affiliates. All Rights Reserved. This manual or parts thereof may not be reproduced in any form unless permitted by contract or by written permission of NVIDIA Corporation.

**Trademarks**

Linux is a registered trademark of Linus Torvalds. PathScale is a registered trademark of Cray, Inc. Red Hat and all Red Hat-based trademarks are trademarks or registered trademarks of Red Hat, Inc. SUSE is a registered trademark of SUSE LLC. NVIDIA, CUDA, GPUDirect, HPC SDK, NVIDIA DGX, NVIDIA Nsight, and NVLink are registered trademarks of NVIDIA Corporation. FLEXlm is a registered trademark of Flexera Software, Inc. PBS Professional, and Green Provisioning are trademarks of Altair Engineering, Inc. All other trademarks are the property of their respective owners.

**Rights and Restrictions**

All statements, specifications, recommendations, and technical information contained herein are current or planned as of the date of publication of this document. They are reliable as of the time of this writing and are presented without warranty of any kind, expressed or implied. NVIDIA Corporation shall not be liable for technical or editorial errors or omissions which may occur in this document. NVIDIA Corporation shall not be liable for any damages resulting from the use of this document.

**Limitation of Liability and Damages Pertaining to NVIDIA Corporation**

The NVIDIA Base Command Manager product principally consists of free software that is licensed by the Linux authors free of charge. NVIDIA Corporation shall have no liability nor will NVIDIA Corporation provide any warranty for the NVIDIA Base Command Manager to the extent that is permitted by law. Unless confirmed in writing, the Linux authors and/or third parties provide the program as is without any warranty, either expressed or implied, including, but not limited to, marketability or suitability for a specific purpose. The user of the NVIDIA Base Command Manager product shall accept the full risk for the quality or performance of the product. Should the product malfunction, the costs for repair, service, or correction will be borne by the user of the NVIDIA Base Command Manager product. No copyright owner or third party who has modified or distributed the program as permitted in this license shall be held liable for damages, including general or specific damages, damages caused by side effects or consequential damages, resulting from the use of the program or the un-usability of the program (including, but not limited to, loss of data, incorrect processing of data, losses that must be borne by you or others, or the inability of the program to work together with any other program), even if a copyright owner or third party had been advised about the possibility of such damages unless such copyright owner or third party has signed a writing to the contrary.
# Table of Contents

Table of Contents ................................................................. i
0.1 About This Manual ........................................................... v
0.2 About The Manuals In General .......................................... v
0.3 Getting Administrator-Level Support ................................ vi
0.4 Getting Professional Services ........................................ vi

1 Introduction To Containerization On NVIDIA Base Command Manager 3

2 Docker Engine 5
   2.1 Docker Setup .......................................................... 5
   2.2 Integration With Workload Managers ................................ 7
   2.3 DockerHost Role ...................................................... 7
   2.4 Iptables ............................................................... 10
   2.5 Storage Backends .................................................... 10
      2.5.1 Device Mapper Driver Settings Support ..................... 12
   2.6 Docker Monitoring ................................................. 13
   2.7 Docker Setup For NVIDIA ......................................... 14

3 Docker Registries 17
   3.1 Docker And Harbor Registries: Introduction .................... 17
      3.1.1 Docker Hub, A Remote Registry .............................. 17
      3.1.2 Local Image Registry Options: Classic Docker Registry Vs Harbor ................................. 17
   3.2 Docker And Harbor Registries: Setup And Configuration .... 17
      3.2.1 Docker Registry Daemon Configuration Using The Docker Registry Role ...................... 18
      3.2.2 Harbor Daemon Configuration Using The Harbor Role ................................................. 19

4 Kubernetes 21
   4.1 Reference Architecture ............................................. 21
      4.1.1 Kubernetes HA .................................................. 22
   4.2 Kubernetes Setup ................................................... 22
      4.2.1 Kubernetes Networking ........................................ 23
      4.2.2 Kubernetes Core Add-ons .................................... 24
      4.2.3 Kubernetes Optional Add-ons ................................. 25
      4.2.4 Kubernetes Setup From The Command Line ................ 26
      4.2.5 Kubernetes Setup From A TUI Session ...................... 29
      4.2.6 Testing Kubernetes ........................................... 31
   4.3 Using GPUs With Kubernetes: NVIDIA GPUs .................... 32
      4.3.1 Prerequisites .................................................... 32
      4.3.2 Existing Contained Deployment ............................... 33
      4.3.3 Existing Docker Deployment ................................... 33
      4.3.4 New Kubernetes Installation ................................. 34
      4.3.5 Existing Kubernetes Installation ............................ 36
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4 Using GPUs With Kubernetes: AMD GPUs</td>
<td>37</td>
</tr>
<tr>
<td>4.4.1 Prerequisites</td>
<td>37</td>
</tr>
<tr>
<td>4.4.2 Managing The YAML File Through CMDaemon</td>
<td>37</td>
</tr>
<tr>
<td>4.4.3 Including Head Nodes as part of the DaemonSet</td>
<td>38</td>
</tr>
<tr>
<td>4.4.4 Running The DaemonSet Only On Specific Nodes</td>
<td>39</td>
</tr>
<tr>
<td>4.4.5 Running An Example Workload</td>
<td>40</td>
</tr>
<tr>
<td>4.5 Kubernetes Configuration Overlays</td>
<td>41</td>
</tr>
<tr>
<td>4.6 Removing A Kubernetes Cluster</td>
<td>42</td>
</tr>
<tr>
<td>4.7 Kubernetes Cluster Configuration Options</td>
<td>43</td>
</tr>
<tr>
<td>4.8 EtcdCluster</td>
<td>45</td>
</tr>
<tr>
<td>4.9 Kubernetes Roles</td>
<td>46</td>
</tr>
<tr>
<td>4.9.1 EtcHost Role</td>
<td>47</td>
</tr>
<tr>
<td>4.9.2 The KubernetesAPIServerProxy Role</td>
<td>47</td>
</tr>
<tr>
<td>4.9.3 The KubernetesApiServer Role</td>
<td>48</td>
</tr>
<tr>
<td>4.9.4 KubernetesController Role</td>
<td>50</td>
</tr>
<tr>
<td>4.9.5 KubernetesScheduler Role</td>
<td>52</td>
</tr>
<tr>
<td>4.9.6 KubernetesProxy Role</td>
<td>53</td>
</tr>
<tr>
<td>4.9.7 KubernetesNode Role</td>
<td>54</td>
</tr>
<tr>
<td>4.10 Security Model</td>
<td>57</td>
</tr>
<tr>
<td>4.10.1 Kyverno</td>
<td>57</td>
</tr>
<tr>
<td>4.10.2 PodSecurityPolicy</td>
<td>58</td>
</tr>
<tr>
<td>4.11 Addition Of New Kubernetes Users</td>
<td>59</td>
</tr>
<tr>
<td>4.11.1 Adding Users Non-Interactively With cm-kubernetes-set-up</td>
<td>59</td>
</tr>
<tr>
<td>4.12 Getting Information And Modifying Existing Kubernetes Users</td>
<td>61</td>
</tr>
<tr>
<td>4.13 List Of Resources Defined For Users</td>
<td>61</td>
</tr>
<tr>
<td>4.14 Pod Security Policies</td>
<td>63</td>
</tr>
<tr>
<td>4.14.2 Disabling Pod Security Policies For Kubernetes</td>
<td>63</td>
</tr>
<tr>
<td>4.14.3 Enabling Manually Via cmsh Instead</td>
<td>63</td>
</tr>
<tr>
<td>4.14.4 The psp Application</td>
<td>64</td>
</tr>
<tr>
<td>4.14.5 The psp_system Application</td>
<td>65</td>
</tr>
<tr>
<td>4.14.6 Users And PodSecurityPolicies</td>
<td>67</td>
</tr>
<tr>
<td>4.15 Kyverno</td>
<td>67</td>
</tr>
<tr>
<td>4.15.1 Kyverno Installation</td>
<td>67</td>
</tr>
<tr>
<td>4.15.2 Kyverno Policies</td>
<td>69</td>
</tr>
<tr>
<td>4.16 Kubernetes Permission Manager</td>
<td>70</td>
</tr>
<tr>
<td>4.17 Providing Access To External Users</td>
<td>72</td>
</tr>
<tr>
<td>4.18 Networking Model</td>
<td>74</td>
</tr>
<tr>
<td>4.19 Kubernetes Monitoring</td>
<td>75</td>
</tr>
<tr>
<td>4.20 Local Path Storage Class</td>
<td>75</td>
</tr>
<tr>
<td>4.21 Setup Of A Storage Class For Ceph</td>
<td>75</td>
</tr>
<tr>
<td>4.22 Integration With Harbor</td>
<td>79</td>
</tr>
<tr>
<td>4.23 Kubernetes Upgrades</td>
<td>79</td>
</tr>
<tr>
<td>4.23.1 Upgrade Prerequisites</td>
<td>79</td>
</tr>
<tr>
<td>4.23.2 Example RHEL8 Cluster</td>
<td>80</td>
</tr>
<tr>
<td>4.23.3 Updating The First Control Plane Node</td>
<td>81</td>
</tr>
</tbody>
</table>
## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.23.4 Updating Subsequent Control Plane Nodes</td>
<td>82</td>
</tr>
<tr>
<td>4.23.5 Updating The Worker Nodes</td>
<td>83</td>
</tr>
<tr>
<td>4.23.6 Updating The Status In NVIDIA Base Command Manager</td>
<td>83</td>
</tr>
<tr>
<td>4.23.7 Notes For Ubuntu</td>
<td>84</td>
</tr>
<tr>
<td>4.23.8 Notes For SLES</td>
<td>84</td>
</tr>
<tr>
<td>4.23.9 Other Approaches</td>
<td>84</td>
</tr>
<tr>
<td>5 Kubernetes Apps</td>
<td>87</td>
</tr>
<tr>
<td>5.1 Providing Custom Docker Images</td>
<td>89</td>
</tr>
<tr>
<td>6 Kubernetes Operators</td>
<td>91</td>
</tr>
<tr>
<td>6.1 Helm Charts For The BCM Operators</td>
<td>91</td>
</tr>
<tr>
<td>6.2 The Jupyter Kernel Operator</td>
<td>93</td>
</tr>
<tr>
<td>6.2.1 Installing The Jupyter Kernel Operator</td>
<td>93</td>
</tr>
<tr>
<td>6.2.2 Architecture Overview</td>
<td>93</td>
</tr>
<tr>
<td>6.2.3 Running Jupyter Kernel Using The Operator</td>
<td>95</td>
</tr>
<tr>
<td>6.2.4 Jupyter Kernel Operator Tunables</td>
<td>96</td>
</tr>
<tr>
<td>6.2.5 Sidecar Arguments And Environment Variables</td>
<td>97</td>
</tr>
<tr>
<td>6.2.6 Running Spark-based Kernels In Jupyter Kernel Operator</td>
<td>98</td>
</tr>
<tr>
<td>6.2.7 Example: Creating An R Kernel From The Kernel Template</td>
<td>98</td>
</tr>
<tr>
<td>6.2.8 Example: Letting Kubernetes Access Private Registries From The Kernel Template</td>
<td>103</td>
</tr>
<tr>
<td>6.2.9 Example: Adding The PVC Parameter To The Kernel Template</td>
<td>105</td>
</tr>
<tr>
<td>6.3 The NVIDIA GPU Operator</td>
<td>107</td>
</tr>
<tr>
<td>6.3.1 Installing The NVIDIA GPU Operator</td>
<td>107</td>
</tr>
<tr>
<td>6.3.2 Installing The NVIDIA GPU Operator On An Existing Cluster</td>
<td>108</td>
</tr>
<tr>
<td>6.3.3 Removing The NVIDIA GPU Operator</td>
<td>109</td>
</tr>
<tr>
<td>6.3.4 Validating The NVIDIA GPU Operator</td>
<td>109</td>
</tr>
<tr>
<td>6.3.5 Validating The NVIDIA GPU Operator In Detail</td>
<td>109</td>
</tr>
<tr>
<td>6.3.6 Running A GPU Workload</td>
<td>113</td>
</tr>
<tr>
<td>6.4 The Run:ai Operator</td>
<td>114</td>
</tr>
<tr>
<td>6.4.1 Installing The Run:ai Operator</td>
<td>114</td>
</tr>
<tr>
<td>6.4.2 Removing The Run:ai Operator</td>
<td>115</td>
</tr>
<tr>
<td>6.4.3 Completing The Run:ai Installation</td>
<td>115</td>
</tr>
<tr>
<td>6.4.4 Post-installation</td>
<td>118</td>
</tr>
<tr>
<td>6.5 Kubernetes Spark Operator</td>
<td>119</td>
</tr>
<tr>
<td>6.5.1 Installing The Kubernetes Spark Operator</td>
<td>119</td>
</tr>
<tr>
<td>6.5.2 Example Spark Operator Run: Calculating Pi</td>
<td>120</td>
</tr>
<tr>
<td>7 Kubernetes On Edge</td>
<td>123</td>
</tr>
<tr>
<td>7.1 Flags For Edge Installation</td>
<td>123</td>
</tr>
<tr>
<td>7.1.1 Speeding Up Kubernetes Installation To Edge Nodes With The --skip-* Flags: Use Cases</td>
<td>124</td>
</tr>
<tr>
<td>8 Kubernetes Cluster API</td>
<td>125</td>
</tr>
<tr>
<td>8.1 Kubernetes Cluster API Components</td>
<td>125</td>
</tr>
<tr>
<td>8.1.1 Kubernetes Management Cluster</td>
<td>125</td>
</tr>
<tr>
<td>8.1.2 Kubernetes CAPI Cluster</td>
<td>126</td>
</tr>
</tbody>
</table>
Preface

Welcome to the Containerization Manual for NVIDIA Base Command Manager 10.

0.1 About This Manual

This manual is aimed at helping cluster administrators install, understand, configure, and manage the containerization integration capabilities of NVIDIA Base Command Manager. The administrator is expected to be reasonably familiar with the Administrator Manual.

0.2 About The Manuals In General

Name Changes From Version 9.2 To 10

The cluster manager software was originally developed by Bright Computing and the name “Bright” featured previously in the product, repositories, websites, and manuals.

Bright Computing was acquired by NVIDIA in 2022. The corresponding name changes, to be consistent with NVIDIA branding and products, are a work in progress. There is some catching up to do in places. For now, some parts of the manual still refer to Bright Computing and Bright Cluster Manager. These remnants will eventually disappear during updates.

BCM in particular is a convenient abbreviation that happens to have the same letters as the former Bright Cluster Manager. With the branding change in version 10, Base Command Manager is the official full name for the product formerly known as Bright Cluster Manager, and BCM is the official abbreviation for Base Command Manager.

Regularly updated versions of the NVIDIA Base Command Manager 10 manuals are available on updated clusters by default at /cm/shared/docs/cm. The latest updates are always online at https://docs.nvidia.com/base-command-manager.

- The Installation Manual describes installation procedures for the basic cluster.
- The Administrator Manual describes the general management of the cluster.
- The User Manual describes the user environment and how to submit jobs for the end user.
- The Cloudbursting Manual describes how to deploy the cloud capabilities of the cluster.
- The Developer Manual has useful information for developers who would like to program with BCM.
- The Machine Learning Manual describes how to install and configure machine learning capabilities with BCM.
- The Edge Manual explains how BCM can be used with edge sites.

If the manuals are downloaded and kept in one local directory, then in most pdf viewers, clicking on a cross-reference in one manual that refers to a section in another manual opens and displays that section in the second manual. Navigating back and forth between documents is usually possible with keystrokes or mouse clicks.

For example: <Alt>-<Backarrow> in Acrobat Reader, or clicking on the bottom leftmost navigation button of xpdf, both navigate back to the previous document.

The manuals constantly evolve to keep up with the development of the BCM environment and the addition of new hardware and/or applications. The manuals also regularly incorporate feedback from
administrators and users, and any comments, suggestions or corrections will be very gratefully accepted at manuals@brightcomputing.com.

There is also a feedback form available via Base View, via the menu icon, following the clickpath: ☰ → Help → Feedback

0.3 Getting Administrator-Level Support

If the reseller from whom BCM was bought offers direct support, then the reseller should be contacted. Otherwise the primary means of support is via the website https://www.nvidia.com/en-us/data-center/bright-cluster-manager/support/. This allows the administrator to submit a support request via a web form, and opens up a trouble ticket. It is a good idea to try to use a clear subject header, since that is used as part of a reference tag as the ticket progresses. Also helpful is a good description of the issue. The followup communication for this ticket goes via standard e-mail. Section 16.2 of the Administrator Manual has more details on working with support.

0.4 Getting Professional Services

The BCM support team normally differentiates between

- regular support (customer has a question or problem that requires an answer or resolution), and
- professional services (customer asks for the team to do something or asks the team to provide some service).

Professional services can be provided via the NVIDIA Enterprise Services page at: https://www.nvidia.com/en-us/support/enterprise/services/
Trademarks
Linux is a registered trademark of Linus Torvalds. PathScale is a registered trademark of Cray, Inc. Red Hat and all Red Hat-based trademarks are trademarks or registered trademarks of Red Hat, Inc. SUSE is a registered trademark of SUSE LLC. NVIDIA, CUDA, GPUDirect, HPC SDK, NVIDIA DGX, NVIDIA Nsight, and NVLink are registered trademarks of NVIDIA Corporation. FLEXlm is a registered trademark of Flexera Software, Inc. PBS Professional, and Green Provisioning are trademarks of Altair Engineering, Inc. All other trademarks are the property of their respective owners.

Rights and Restrictions
All statements, specifications, recommendations, and technical information contained herein are current or planned as of the date of publication of this document. They are reliable as of the time of this writing and are presented without warranty of any kind, expressed or implied. NVIDIA Corporation shall not be liable for technical or editorial errors or omissions which may occur in this document. NVIDIA Corporation shall not be liable for any damages resulting from the use of this document.

Limitation of Liability and Damages Pertaining to NVIDIA Corporation
The NVIDIA Base Command Manager product principally consists of free software that is licensed by the Linux authors free of charge. NVIDIA Corporation shall have no liability nor will NVIDIA Corporation provide any warranty for the NVIDIA Base Command Manager to the extent that is permitted by law. Unless confirmed in writing, the Linux authors and/or third parties provide the program as is without any warranty, either expressed or implied, including, but not limited to, marketability or suitability for a specific purpose. The user of the NVIDIA Base Command Manager product shall accept the full risk for the quality or performance of the product. Should the product malfunction, the costs for repair, service, or correction will be borne by the user of the NVIDIA Base Command Manager product. No copyright owner or third party who has modified or distributed the program as permitted in this license shall be held liable for damages, including general or specific damages, damages caused by side effects or consequential damages, resulting from the use of the program or the un-usability of the program (including, but not limited to, loss of data, incorrect processing of data, losses that must be borne by you or others, or the inability of the program to work together with any other program), even if a copyright owner or third party had been advised about the possibility of such damages unless such copyright owner or third party has signed a writing to the contrary.
Introduction To Containerization
On NVIDIA Base Command Manager

Containerization is a technology that allows processes to be isolated by combining cgroups, Linux namespaces, and (container) images.

- Cgroups are introduced in section 7.10 on workload management of the Administrator Manual.

- Linux namespaces represent independent spaces for different operating system facilities: process IDs, network interfaces, mount points, inter-process communication resources and others. Cgroups and namespaces allow processes to be isolated from each other by separating the available resources as much as possible.

- A container image is a component of a container, and is a file that contains one or several layers. The layers cannot be altered as far the container is concerned, and a snapshot of the image can be used for other containers. A union file system is used to combine these layers into a single image. Union file systems allow files and directories of separate file systems to be transparently overlaid, forming a single coherent file system.

Cgroups, namespaces and image are the basis of a container. When the container is created, then a new process can be started within the container. Containerized processes running on a single machine all share the same operating system kernel, so they start immediately, without the delay of requiring a kernel to first boot up. No process is allowed to change the layers of the image. All changes are applied on a temporary layer created on top of the image, and these changes are destroyed when the container is removed.

There are several ways to manage the containers, but the most powerful approaches use Docker, also known as Docker Engine, and Kubernetes.

Docker manages containers on individual hosts, while Kubernetes manages containers across a cluster. BCM integrates both of these solutions, so that setup, configuration and monitoring of containers becomes an easily-managed part of BCM.

Chapter 2 describes how Docker integration with BCM works.
Chapter 3 covers how Docker registries are integrated.
Chapter 4 covers Kubernetes integration.
Chapter 5 covers Kubernetes application configuration and groups of Kubernetes applications.
Chapter 6 covers Kubernetes operators, which are a way to manage Kubernetes cluster applications.
Chapter 7 covers Kubernetes deployment on edge sites.
Chapter 8 describes the installation and usage of the NVIDIA Base Command Manager CAPI extension called BCM Kubernetes CAPI Infrastructure Provider. Kubernetes Cluster API (CAPI), is an API for managing Kubernetes clusters.

Chapter 9 describes the use of Singularity, which is an application containerization tool. Singularity is designed to execute containers as if they are just native applications on a host computer, and to work with HPC.
Docker integration with NVIDIA Base Command Manager 10 for Docker version 24.0.5 is available at the time of writing of this section (September 2023) on the x86_64 architecture for all the BCM-supported Linux distributions. For a more up-to-date status, the features matrix at https://support.brightcomputing.com/feature-matrix/ can be checked.

Docker integration with NVIDIA Base Command Manager 10 is part of BCME (Appendix A), which means that it is certified for NVIDIA AI Enterprise.

Docker Engine (or just Docker) is a tool for container management. Docker allows containers and their images to be created, controlled, and monitored on a host using Docker command line tools or the Docker API.

Swarm mode, which allows containers to spawn on several hosts, is not formally supported by NVIDIA Base Command Manager 10. This is because NVIDIA Base Command Manager 10 provides Kubernetes for this purpose instead.

Docker provides a utility called `docker`, and two daemons called `containerd` (the default provided by BCM), and `dockerd`. Additional functionality includes pulling the container image from a specific image registry (Chapter 3), configuring the container network, setting `systemd` limits, and attaching volumes.

### 2.1 Docker Setup

BCM provides the `cm-docker` package. The package includes the following components:

- Docker itself, that provides an API and delegates the container management to Containerd;
- Containerd runtime, that manages OCI images and OCI containers (via runC);
- runC, a CLI tool for spawning and running containers according to the OCI specification runtime;
- docker-py, a Python library for the Docker API.

Typically, however, the administrator is expected to simply run the `cm-docker-setup` utility, which is provided by BCM’s `cm-setup` package. Running `cm-docker-setup` takes care of the installation of the `cm-docker` package and also takes care of Docker setup. If run without options then the utility starts up a TUI dialog (figure 2.1).
The cm-docker-setup utility asks several questions, such as which Docker registries are to be used, what nodes Docker is to be installed on, whether the NVIDIA container runtime should be installed, and so on. If cm-docker-setup is used with the -c option, and given a YAML configuration file <YAMLfile>, then a runtime configuration is loaded from that file. The YAML file is typically generated and saved from an earlier run.

When the questions in the TUI dialog have been answered and the deployment is carried out, the utility:

- installs the cm-docker package, if it has not been installed yet
- then assigns the DockerHost role to the node categories or head nodes that were specified
- adds health checks to the BCM monitoring configuration
- performs the initial configuration of Docker.

The regular nodes on which Docker is to run, are restarted by the utility, if needed. The restart operation provisions the updated images from the image directory onto the nodes.

The cm-docker package also includes a modules environment file, which must be loaded in order to use the docker command. The modules environment and modules are introduced in section 2.2 of the Administrator Manual.

By default only the administrator can run the docker commands after setup (some output ellipsized):

Example

```
[root@basecm10 ~]# ssh node001
[root@node001 ~]# module load docker
[root@node001 ~]# docker info
Containers: 0
Images: 0
...  
Docker Root Dir: /var/lib/docker
Debug Mode: false
Registry: https://index.docker.io/v1/
Labels: 
Experimental: false
Insecure Registries: 
127.0.0.0/8
Registry Mirrors: 
https://harbor-proxy.brightcomputing.com/
Live Restore Enabled: false
[root@node001 ~]# 
```

and the hello-world image can be run as usual with:
Example

[root@node001 ~]# docker run hello-world
Unable to find image 'hello-world:latest' locally
latest: Pulling from library/hello-world
2db29710123e: Pull complete
Digest: sha256:cc15c5b292d8525effc0f5989f29f1804f3a725c82d05e156653a563f15e4f695
Status: Downloaded newer image for hello-world:latest

Hello from Docker!
This message shows that your installation appears to be working correctly.
...

Or, for example, importing and running Apache containers with Docker may result in the following output:

Example

[root@node001 ~]# module load docker
[root@node001 ~]# docker run httpd & docker run httpd &
... runs a couple of Apache containers...
[root@node001 ~]# docker container ls
CONTAINER ID IMAGE COMMAND CREATED STATUS PORTS NAMES
acdbe2f9367b httpd "httpd-foreground" 13 seconds ago Up 11 seconds 80/tcp quizzical_hbabha
64787a86524dd httpd "httpd-foreground" 13 seconds ago Up 11 seconds 80/tcp funny_hypatia
...
[root@node001 ~]#

Using Docker directly means being root on the host. It is rarely sensible to carry out regular user actions as the root user at all times.

So, to make Docker available to regular users, Kubernetes provides a user management layer and restrictions.

After Docker has been installed, Kubernetes can be set up to allow regular user access to the Docker containers as covered in Chapter 4. It is a best practice for regular users to use Kubernetes instead of Docker commands directly.

2.2 Integration With Workload Managers

BCM does not provide integration of Docker with workload managers. The administrator can however tune the workload managers in some cases to enable Docker support.

- LSF – An open beta version of LSF with Docker support is available from the IBM web site. This LSF version allows jobs to run in Docker containers, and monitors the container resources per job.

- PBS Professional – Altair provides a hook script that allows jobs to start in Docker containers. Altair should be contacted to obtain the script and instructions.

2.3 DockerHost Role

When cm-docker-setup is executed, the DockerHost role is assigned to nodes or categories. The DockerHost role is responsible for Docker service management and configuration.

From cmsh, the configuration parameters can be managed from the Docker::Host role:

Example
The Docker host parameters that CMDaemon can configure in the DockerHost role, along with a description, are shown in table 2.1:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add services*</td>
<td>Add services to nodes belonging to this node. Care must be taken if setting this to no. (default: yes)</td>
</tr>
<tr>
<td>Spool</td>
<td>Root of the Docker runtime (default: /var/lib/docker)</td>
</tr>
<tr>
<td>Tmp dir</td>
<td>Location for temporary files. Default: $$&lt;spool&gt;/tmp$$, where $$&lt;spool&gt;$$ is replaced by the path to the Docker runtime root directory</td>
</tr>
</tbody>
</table>

...continues
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enable SELinux</td>
<td>Enable selinux support in Docker daemon (default: yes)</td>
</tr>
<tr>
<td>Default Ulimits</td>
<td>Set the default ulimit options for all containers</td>
</tr>
<tr>
<td>Debug</td>
<td>Enable debug mode (default: no)</td>
</tr>
<tr>
<td>Log Level</td>
<td>Set the daemon logging level. In order of increasing verbosity: fatal, error, warn, info, debug. (default: info)</td>
</tr>
<tr>
<td>Bridge IP</td>
<td>Network bridge IP (not defined by default)</td>
</tr>
<tr>
<td>Bridge</td>
<td>Attach containers to a network bridge (not defined by default)</td>
</tr>
<tr>
<td>MTU</td>
<td>Set the containers network MTU, in bytes (default: 0, which does not set the MTU at all)</td>
</tr>
<tr>
<td>API Sockets</td>
<td>Daemon socket(s) to connect to (default: unix:///var/run/docker.sock)</td>
</tr>
<tr>
<td>Iptables</td>
<td>Enable iptables rules (default: yes)</td>
</tr>
<tr>
<td>User Namespace Remap</td>
<td>User/Group setting for user namespaces (not defined by default). It can be set to any of &lt;UID&gt;, <a href="">UID:GID</a>, &lt;username&gt;, <a href="">username:groupname</a>. If it is used, then user_namespace.enable=1 must be set in the kernel options for the relevant nodes, and those nodes must be rebooted to pick up the new option.</td>
</tr>
<tr>
<td>Insecure Registries</td>
<td>If registry access uses HTTPS but does not have proper certificates distributed, then the administrator can make Docker accept this situation by adding the registry to this list (empty by default)</td>
</tr>
<tr>
<td>Enable TLS</td>
<td>Use TLS (default: no)</td>
</tr>
<tr>
<td>Verify TLS</td>
<td>Use TLS and verify the remote (default: no)</td>
</tr>
<tr>
<td>TLS CA</td>
<td>Trust only certificates that are signed by this CA (not defined by default)</td>
</tr>
</tbody>
</table>
Parameter | Description
--- | ---
TLS Certificate | Path to TLS certificate file (not defined by default)
TLS Key | Path to TLS key file (not defined by default)
Certificates Path | Path to Docker certificates (default: /etc/docker)
Storage Backends | Docker storage back ends. Storage types can be created and managed, in a submode under this mode. The available types are described in table 2.2. Each of these storage types has options that can be set from within the submode.
Containerd Socket | Path to the containerd socket (default: not used)
Runtime | Docker runtime
Options | Additional parameters for docker daemon

* Boolean (takes yes or no as a value)

Table 2.1: Docker::Host role options

2.4 Iptables

By default iptables rules have been added to nodes that function as a Docker host, to let network traffic go from the containers to outside the pods network. If this conflicts with other software that uses iptables, then this option can be disabled. For example, if the `docker::host` role has already been assigned to the nodes via the default category, then the iptables rules that are set can be disabled by setting the `iptables` parameter in the `Docker::Host` role to `no`:

Example

```
[root@basecm10 ~]# cmsh
[basecm10]# category use default
[basecm10->category[default]]% roles
[basecm10->category[default]->roles]% use docker::host
[basecm10->category[default]->roles[Docker::Host]]% set iptables no
[basecm10->category[default]->roles[Docker::Host]]% commit
```

2.5 Storage Backends

A core part of the Docker model is the efficient use of containers based on layered images. To implement this, Docker provides different storage back ends, also called storage drivers. These storage back ends rely heavily on various filesystem features in the kernel or volume manager. Some storage back ends perform better than others, depending on the circumstances.

The default storage back end configured by `cm-docker-setup` is overlay2. Storage back ends supported by Docker are listed in table 2.2:
### Table 2.2: Docker storage back ends

<table>
<thead>
<tr>
<th>Technology</th>
<th>Description</th>
<th>Backend Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>OverlayFS</td>
<td>This is a modern union filesystem. It is the preferred storage driver for recent Docker versions. It has been in the mainline Linux kernel since version 3.18, with additional improvements for Docker in version 4.0. All of the distributions that BCM supports have backported the kernel changes needed for this to work.</td>
<td>overlay2</td>
</tr>
<tr>
<td>Device Mapper</td>
<td>Deprecated since Docker Engine 18.09. It is a kernel-based framework that has been included in the mainline Linux kernel since version 2.6.9. It underpins many advanced volume management technologies on Linux. The driver stores every image and snapshot on its own virtual device, and works at the block level rather than the file level.</td>
<td>devicemapper</td>
</tr>
<tr>
<td>• A loopback mechanism can be implemented using <code>loop-lvm</code> mode. This allows files on a local disk to be managed as if they are on a physical disk or block device. This is simpler than the thin pool mode, but is strongly discouraged for production use. In BCM this mode is implemented by selecting the option <code>loopback</code> (testing only). This is selected in the storage back end selection screen of the <code>cm-docker-setup</code> installation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• A thin pool mode can be implemented using <code>direct-lvm</code> mode. This uses a logical volume as a thin pool to use as backing for the storage pool, and uses a spare block device. Configuring this is normally more involved. In BCM this mode is implemented by selecting the option <code>block</code> (production ready). This is selected in the storage back end selection screen of the <code>cm-docker-setup</code> installation session.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUFS</td>
<td>This was the first storage back end that Docker used. AUFS is not included in the mainline Linux kernel. Out of the distributions that NVIDIA Base Command Manager 8.2 supports, it is only Ubuntu that supports it.</td>
<td>aufs</td>
</tr>
</tbody>
</table>

The `docker info` command, amongst many other items, shows the storage driver and related settings that are being used in Docker:

**Example**

```
[root@basecm10 ~]# module load docker
[root@basecm10 ~]# docker info
```

Client:
- Context: default
- Debug Mode: false
Server:
Containers: 18
Running: 8
Paused: 6
Stopped: 4
Images: 1
Server Version: 20.10.9
Storage Driver: overlay2
  Backing Filesystem: xfs
  Supports d_type: true
  Native Overlay Diff: true
  userxattr: false
Logging Driver: json-file
Cgroup Driver: cgroupfs
Cgroup Version: 1
Plugins:
  Volume: local
  Network: bridge host ipvlan macvlan null overlay
  Log: awslogs fluentd gcplogs gelf journald json-file local logentries splunk syslog
...

Docker data volumes are not controlled by the storage driver. Reads and writes to data volumes bypass the storage driver. It is possible to mount any number of data volumes into a container. Multiple containers can also share one or more data volumes.

More information about Docker storage back ends is available at https://docs.docker.com/engine/userguide/storagedriver.

### 2.5.1 Device Mapper Driver Settings Support

BCM supports device mapper driver settings more explicitly than the other driver back end settings.

By default the device mapper storage back end is added automatically, and can be configured in the storagebackends submode of the DockerHost role:

**Example**

```bash
[basecm10->device[basecm10]->roles[docker::host]]% storagebackends
[basecm10->device[basecm10]->roles[docker::host]->storagebackends]% use devicemapper
[basecm10->device[basecm10]->roles[docker::host]->storagebackends[devicemapper]]% show
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>devicemapper</td>
</tr>
<tr>
<td>Revision</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>devicemapper</td>
</tr>
<tr>
<td>Loop Data Size</td>
<td></td>
</tr>
<tr>
<td>Loop Metadata Size</td>
<td></td>
</tr>
<tr>
<td>Loop Device Size</td>
<td></td>
</tr>
<tr>
<td>Pool Device</td>
<td></td>
</tr>
<tr>
<td>Filesystem</td>
<td>xfs</td>
</tr>
<tr>
<td>Block Size</td>
<td>64K</td>
</tr>
<tr>
<td>Blk Discard</td>
<td>yes</td>
</tr>
<tr>
<td>Mkfs Arguments</td>
<td></td>
</tr>
<tr>
<td>Mount Options</td>
<td></td>
</tr>
</tbody>
</table>

---
2.6 Docker Monitoring

The parameters that are used in the Docker device mapper back end are described in table 2.3:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to Docker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blk Discard*</td>
<td>Enables or disables the use of blkdiscard when removing device mapper devices (default: yes)</td>
<td>dm.blkdiscard</td>
</tr>
<tr>
<td>Block Size</td>
<td>Custom blocksize to use for the thin pool (default: 64kB)</td>
<td>dm.blocksize</td>
</tr>
<tr>
<td>Filesystem</td>
<td>Filesystem type to use for the base device (default: xfs)</td>
<td>dm.fs</td>
</tr>
<tr>
<td>Loop Data Size</td>
<td>Size to use when creating the loopback file for the data virtual device which is used for the thin pool (default: 100GB)</td>
<td>dm.loopdatasize</td>
</tr>
<tr>
<td>Loop Device Size</td>
<td>Size to use when creating the base device, which limits the size of images and container (not set by default)</td>
<td>dm.basesize</td>
</tr>
<tr>
<td>Loop Metadata Size</td>
<td>Size to use when creating the loopback file for the metadata device which is used for the thin pool (default: 2GB)</td>
<td>dm.loopmetadatasize</td>
</tr>
<tr>
<td>Mkfs Arguments</td>
<td>Extra mkfs arguments to be used when creating the base device</td>
<td>dm.mkfsarg</td>
</tr>
<tr>
<td>Mount Options</td>
<td>Extra mount options used when mounting the thin devices</td>
<td>dm.mountopt</td>
</tr>
<tr>
<td>Pool Device</td>
<td>Custom block storage device to use for the thin pool (not set by default)</td>
<td>dm.thinpooldev</td>
</tr>
</tbody>
</table>

* Boolean (takes yes or no as a value)

Table 2.3: Device mapper back end Docker options

For back end driver storage settings other than device mapper, such as AUFS or OverlayFS, settings can be added as options if needed. In cmsh this can be done by setting the options parameter in the storagebackend submode under the docker::host role.

2.6 Docker Monitoring

When cm-docker-setup runs, it configures and runs the following Docker health checks:

1. makes a test API call to the endpoint of the Docker daemon
2. checks containers to see that none is in a dead state

The Docker daemon can be monitored outside of BCM with the usual commands directly. BCM ways to manage or check on Docker include the following:

In CMDaemon, the docker service can be checked:

Example

```
[basescm10->device[node001]->services]% list
Service (key) Monitored Autostart
------------------------ ---------- ----------
docker                   yes        yes
nslcd                    yes        yes
[basescm10->device[node001]->services]% show docker
Parameter        Value
```
### 2.7 Docker Setup For NVIDIA

NVIDIA GPU Cloud (NGC) is a cloud platform that runs on NVIDIA GPUs. NGC containers are lightweight containers that run applications on NVIDIA GPUs. The applications are typically HPC, machine learning, or deep learning applications.

An NGC can be set up to run NGC containers from the registry `http://ngc.nvidia.com`. Docker can be configured as an NGC running NGC containers by using the NVIDIA Container Toolkit.

The BCM package provided for this is `cm-nvidia-container-toolkit`.

One way to install and deploy this package is as part of the Docker installation, when running `cm-docker-setup` (section 2.1), where the cluster administrator selects `yes` as the answer to the request: "Do you want to install the NVIDIA Runtime for Docker?".

Alternatively, if Docker has already been installed via `cm-docker-setup`, and if the package has not been installed, then it can be installed via the package manager, `yum` or `apt`. The toolkit has to be running on the compute nodes that have GPUs, which means that the installation must go to the appropriate node image (section 11.4 of the Administrator Manual). For example, if the appropriate image is `gpu-image`, then the package manager command for RHEL-based distributions would be:

**Example**

```bash
[basecm10->device[node001]]% nvidia-docker run --runtime=nvidia d2012 
```
# yum install --installroot=/cm/images/gpu-image cm-nvidia-container-toolkit

The nodes that use that GPU image can then be rebooted to pick up the new package.
The GPU status can then be printed with the NVIDIA system management interface command. For example, if the image has been picked up by node001:

Example

[root@basecm10 ~]# ssh node001
Last login: Thu Dec  2 09:24:03 2021 from basecm10.cm.cluster
[root@node001 ~]# module load docker
[root@node001 ~]# docker run --runtime=nvidia --rm nvidia/cuda:11.4-base nvidia-smi
Unable to find image 'nvidia/cuda:11.4.0-base' locally
11.4.0-base: Pulling from nvidia/cuda...

Example

```
Example

[root@basecm10 ~]# ssh node001
Last login: Thu Dec  2 09:24:03 2021 from basecm10.cm.cluster
[root@node001 ~]# module load docker
[root@node001 ~]# docker run --runtime=nvidia --rm nvidia/cuda:11.4-base nvidia-smi
Unable to find image 'nvidia/cuda:11.4.0-base' locally
11.4.0-base: Pulling from nvidia/cuda...
```
3

Docker Registries

When a user creates a new container, an image specified by the user should be used. The images are kept either locally on a host, or in a registry. The image registry can be provided by a cloud provider or locally.

3.1 Docker And Harbor Registries: Introduction

3.1.1 Docker Hub, A Remote Registry

By default, Docker searches for images in Docker Hub, which is a cloud-hosted public and private image registry. Docker Hub serves a huge collection of existing images that users can make use of. Every user is allowed to create a new account, and to upload and share images with other users. Using the Docker client, a user can search for already-published images, and then pull them down to a host in order to build containers from them.

When an image is found in the registry, the Docker client verifies if the latest version of the image has already been downloaded. If not, then it downloads the images, and stores them locally. It also tries to synchronize them when a new container is created. When the latest image is downloaded, Docker creates a container from the image layers that are formatted to be used by a union file system. Docker can make use of several union file system variants, including AUFS, btrfs, vfs, and DeviceMapper.

3.1.2 Local Image Registry Options: Classic Docker Registry Vs Harbor

Besides using Docker Hub for the image registry, the administrator can also install a local image registry on the cluster. BCM provides two ways to integrate such a local registry with the cluster, based on existing open source projects:

- The first one is the classic docker registry provided by Docker Inc, and can be useful if the registry is used by trusted users.

- The second registry, Harbor, developed by VMware and introduced in NVIDIA Base Command Manager version 8.1-5, provides additional features such as security and identity management, and is aimed at the enterprise.

Both options can be installed with the cm-container-registry-setup utility, which comes with BCM’s cm-setup package.

3.2 Docker And Harbor Registries: Setup And Configuration

Docker Registry and Harbor can be installed via the cm-container-registry-setup command-line utility. They can also be installed via Base View in NVIDIA Base Command Manager for versions beyond 8.1-6 as follows:

- The Docker Registry Deployment Wizard is launched via the clickpath: Containers → Docker → Docker Registry Wizard
• Either Docker Registry, or Harbor, should be chosen as a registry.

• A single node is ticked for the deployment. The address, port, and the root directory for storing the container images are configured. If the head node is selected for Harbor, then the setup later on asks to open the related port on the head node. This is to make Harbor and the Harbor UI externally accessible.

• In the summary page, if the Ready for deployment box is ticked, then the administrator can go ahead with deploying the registry.

• When the deployment is complete, the Docker Registry becomes ready for use. Harbor can take a few additional minutes to be ready for use.

Similar to the case of etcd nodes (section 4.2), nodes that run Harbor or Docker Registry have the datanode option (page 255 of the Administrator Manual) when installed by BCM utilities. The option helps prevent the registry being wiped out by accident, and is added when the cm-container-registry-setup utility is used to install Harbor or Docker Registry. This extra protection is put into place because if a registry is wiped out, then the state of images in the container becomes incoherent and cannot be restored.

**Harbor UI**

If the head node is where Harbor is to be installed, and is to be made externally accessible, then the Harbor UI can be accessed at https://<head node hostname>:9443.

If a different node is used for Harbor to be installed, then the related port must be forwarded locally. Harbor can be logged into by default with the admin user and the default Harbor12345 password.
It is recommended to change that password after the first login.

**Clair**

Harbor comes with Clair, a tool for vulnerability static analysis of container images.

More information on using Clair with Harbor can be found at https://github.com/vmware/harbor/blob/master/docs/user_guide.md#vulnerability-scanning-via-clair.

**Dealing With A Pre-existing Kubernetes Or Harbor Installation**

Since Harbor uses Docker internally, and because Kubernetes customizes Docker networking, it means that nodes where Kubernetes is running cannot be reused for Harbor, and that nodes where Harbor is running cannot be reused for Kubernetes.

### 3.2.1 Docker Registry Daemon Configuration Using The Docker Registry Role

The Docker Registry role is used to configure and manage the docker-registry daemon, and its parameters are described in table 3.1:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Main domain name (default: hostname of the node)</td>
</tr>
<tr>
<td>Alt Domains</td>
<td>Alternative domain names (default: FQDN of the node)</td>
</tr>
<tr>
<td>Port</td>
<td>Port (default: 5000)</td>
</tr>
</tbody>
</table>

...continues
3.2 Docker And Harbor Registries: Setup And Configuration

...continued

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spool Dir</td>
<td>Spool directory (default: /var/lib/docker-registry)</td>
</tr>
</tbody>
</table>

* Boolean (takes yes or no as a value)

**Table 3.1: Docker Registry role parameters**

The values stored in the Docker Registry role are not supposed to be changed, but they are useful for the uninstall procedure, and also as a record of the settings for the administrator.

```
[basecm10->device[basecm10]->roles[generic::docker_registry]]% environments
[basecm10->device[basecm10]->roles[generic::docker_registry]->environments]% list
```

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Node Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>alt_domains</td>
<td>node001.cm.cluster</td>
<td>no</td>
</tr>
<tr>
<td>domain</td>
<td>node001</td>
<td>no</td>
</tr>
<tr>
<td>port</td>
<td>5000</td>
<td>no</td>
</tr>
<tr>
<td>spool_dir</td>
<td>/var/lib/docker-registry</td>
<td>no</td>
</tr>
</tbody>
</table>

Further details on the `docker-registry` daemon can be found at https://github.com/docker/distribution.

3.2.2 Harbor Daemon Configuration Using The Harbor Role

The Harbor role is used to configure and manage the `harbor` daemon. The parameters of the role are described in table 3.2:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Main domain name (default: hostname of the node)</td>
</tr>
<tr>
<td>Alt Domains</td>
<td>Alternative domain names (default: FQDN of the node)</td>
</tr>
<tr>
<td>Port</td>
<td>Port (default: 9443)</td>
</tr>
<tr>
<td>Spool Dir</td>
<td>Spool directory (default: /var/lib/harbor)</td>
</tr>
<tr>
<td>Default Password</td>
<td>Default password of the Harbor admin user (default: Harbor12345)</td>
</tr>
<tr>
<td>DB Password</td>
<td>Password of the Harbor database (default: randomly generated)</td>
</tr>
<tr>
<td>Clair DB Password</td>
<td>Password of the Clair database (default: randomly generated)</td>
</tr>
</tbody>
</table>

...continues
...continued

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Boolean</td>
<td>(takes yes or no as a value)</td>
</tr>
</tbody>
</table>

**Table 3.2: Harbor role parameters**

The values stored in the Harbor role are not supposed to be changed, but they are useful for the uninstall procedure, and also as reminder of the settings for the administrator.

```
[basecm10->device[basecm10]->roles[generic::harbor]]% environments
[basecm10->device[basecm10]->roles[generic::harbor]->environments]% list
```

<table>
<thead>
<tr>
<th>Name (key)</th>
<th>Value</th>
<th>Node Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>alt_domains</td>
<td>harbor,node001.cm.cluster</td>
<td>no</td>
</tr>
<tr>
<td>db_password</td>
<td>&lt;generated password&gt;</td>
<td>no</td>
</tr>
<tr>
<td>domain</td>
<td>node001</td>
<td>no</td>
</tr>
<tr>
<td>external_network</td>
<td>True</td>
<td>no</td>
</tr>
<tr>
<td>port</td>
<td>9443</td>
<td>no</td>
</tr>
<tr>
<td>redirect_port</td>
<td>65535</td>
<td>no</td>
</tr>
<tr>
<td>spool_dir</td>
<td>/var/lib/harbor</td>
<td>no</td>
</tr>
</tbody>
</table>

Further details on Harbor can be found at https://vmware.github.io/harbor.
Kubernetes

Kubernetes is an open-source platform for automating deployment, scaling, and operations of application containers across clusters of hosts. With Kubernetes, it is possible to:

- scale applications on the fly
- seamlessly update running services
- optimize hardware usage by using only the resources that are needed

BCM provides the administrator with the required packages, allows Kubernetes to be set up on a cluster, and manages and monitors Kubernetes. More information about the design of Kubernetes, its command line interfaces, and other Kubernetes-specific details, can be found at the official online documentation at https://kubernetes.io/docs/.

Kubernetes integration with NVIDIA Base Command Manager 10 is available for Kubernetes v1.24, v1.25, v1.26, v1.27, and v1.28 at the time of writing of this paragraph (January 2024). Kubernetes runs on the x86_64 architecture for all the BCM-supported Linux distributions. For a more up-to-date status, the features matrix at https://support.brightcomputing.com/feature-matrix/ can be checked.

4.1 Reference Architecture

A reference architecture for Kubernetes in BCM comprises:

- **etcd nodes**: An etcd cluster—the Kubernetes distributed key-value storage—runs on regular nodes only, and uses an odd number (1, 3, 5 ... ) of nodes.

- **master nodes**: Kubernetes master units run on head or compute nodes. 2 (or 3) are recommended for High Availability (HA). In a BCM HA configuration, both or none of the head nodes should be selected. That is, it must not run on only one head node of an HA configuration.

- **worker nodes**: Kubernetes worker units run on regular nodes only.

Since NVIDIA Base Command Manager version 8.2, multiple clusters of Kubernetes can be deployed. In such a configuration the same node cannot be shared across different Kubernetes clusters.

A Kubernetes API server proxy based on NGINX runs on every node, except for on nodes that run etcd. The proxy also runs on the head node(s).

Because of the server proxy, a port is reserved on the head node(s) for every Kubernetes cluster. This is required for Kubernetes HA (section 4.1.1), and it also allows `kubectl` and other tools such as Helm to be used from the head node, to access each Kubernetes cluster.
4.1.1 Kubernetes HA
For a Kubernetes HA setup the minimum node requirements are:

- at least 3 nodes for Etcd
- at least 2 nodes for Kubernetes Master

In an average BCM cluster, 3 nodes for the etcd cluster, and 3 nodes for the Kubernetes master are recommended.

Even without a Kubernetes master in an HA configuration, there is no downtime for existing pods running on the worker nodes. The worker nodes will still continue to work.

However, Kubernetes HA is needed to be able to schedule tasks, spawn new pods, and in general keep the cluster in the desired state.

4.2 Kubernetes Setup
BCM allows deployment of several Kubernetes versions:

![Figure 4.1: Kubernetes setup TUI session: version selection screen](image1)

Some Kubernetes versions allow selection of the container runtime interface:

![Figure 4.2: Kubernetes setup TUI session: container runtime selection screen](image2)

The container runtime options at the time of writing of this section (December 2022) are shown in table 4.1.
4.2 Kubernetes Setup

Kubernetes Package Container runtime options

<table>
<thead>
<tr>
<th>Kubernetes version</th>
<th>Package</th>
<th>Container runtime options</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.21.4</td>
<td>cm-kubernetes121</td>
<td>containerd, cri-dockerd</td>
</tr>
<tr>
<td>1.22.15</td>
<td>cm-kubernetes122</td>
<td>containerd, cri-dockerd, BCM Docker (deprecated), third party docker (deprecated)</td>
</tr>
<tr>
<td>1.24.0</td>
<td>cm-kubernetes124</td>
<td>containerd, cri-dockerd, BCM docker (deprecated), third party docker (deprecated)</td>
</tr>
</tbody>
</table>

**Table 4.1: Kubernetes versions**

BCM provides the following Kubernetes-related packages:

- The **conntrack** and **nginx** packages: These are always installed on the head node(s) and on the master and worker node(s).

- The **cm-etcd** package is installed on the nodes selected for etcd. In a similar way to the case of Harbor or Docker Registry (section 3.2), the nodes that run etcd are protected by BCM with the **datanode** option (page 255 of the *Administrator Manual*). For etcd nodes, the option is added during the **cm-kubernetes-setup** installation. As in the case for the registries, the datanode option is set in order to help prevent the administrator from wiping out the existing state of etcd nodes. Wiping out the state of etcd nodes means that the Kubernetes cluster becomes incoherent and that it cannot be restored to where it was just before the etcd nodes were wiped. The etcd version installed by the BCM package is 3.5.8.

### 4.2.1 Kubernetes Networking

Early on during the wizard (figure 4.3), a name for the cluster is requested. The wizard pre-fills it with **default**, but this should not be confused with the Kubernetes **default** namespace. Here, the name is used instead, inside BCM, to identify the cluster, configuration files, and other resources such as module files.

![Insert basic values of the new Kubernetes cluster](image)

**Figure 4.3: Kubernetes setup TUI session: networking selection screen**

This screen also allows the following important choices:

- **Kubernetes external FQDN**: This is the FQDN that is placed as one of the subjects in the public-facing certificates generated for this Kubernetes cluster.

- **Service network base address and Service network netmask bits**: These define the CIDR for the service network. The wizard pre-fills the fields. It also tries to avoid pre-filling them with overlapping network ranges, by taking any existing network known to BCM into account.
• Pod network base address and Pod network netmask bits: These define the CIDR for the pod network. The wizard pre-fills these. It also tries to avoid pre-filling them with overlapping network ranges, by taking any existing network known to BCM into account. By default, entire /24 network ranges are assigned to individual Kubernetes nodes from the pod CIDR.

The packages are installed automatically from the repository when the administrator runs `cm-kubernetes-setup` from the command line.

The log file produced by the setup can be found in `/var/log/cm-kubernetes-setup.log`.

### 4.2.2 Kubernetes Core Add-ons

During setup, some critical add-on components such as CoreDNS and Calico are automatically deployed in the `kube-system` namespace. In BCM all add-ons are treated as Kubernetes applications (Chapter 5), and belong to the default app group `system`.

A `cmsh` treeview illustrating the hierarchy to access these applications is:

```
[cmsh]
|-- ...
|-- kubernetes[default]
   ||-- appgroups[system]
   |   ||-- applications
   |-- ...
```

**CoreDNS**

CoreDNS is the DNS server add-on for internal service discovery. It reads the IP addresses of services and pods from Etcd, and resolves domain names for them. If a domain name is not found because the domain is external to the Kubernetes cluster, then CoreDNS forwards the request to the main DNS server. BCM uses CoreDNS version 1.10.1 with Kubernetes version 1.28.

**Calico**

Calico is a modern SDN (Software-Defined Networking) add-on to manage a cluster-wide network for pods. Calico uses an agent called Felix to run on each node as a pod. BCM uses Calico version 3.26.

Further details on Calico can be found at [https://docs.projectcalico.org/](https://docs.projectcalico.org/).

If the Kubernetes cluster is composed of more than 50 nodes, then the Calico component Typha is also automatically deployed for better scalability. The number of Typha replicas is calculated by allocating one Typha replica per 150 nodes, with a minimum of 3 (above 50 nodes) and a maximum of 20.

If an initial deployment of the Kubernetes cluster has fewer than 50 nodes, but nodes are then added to the Kubernetes cluster so that the 50 node threshold is exceeded, then Typha is not automatically enabled. In this case, Typha can be enabled manually via `cmsh` as follows:

**Example**

```
[basecm10->kubernetes[default]->appgroups[system]->applications[calico]]% environment
[basecm10->kubernetes[default]->appgroups[system]->applications[calico]->environment]% list
Name (key)   Value          Nodes
---------------------- ------------------
calico_typha_replicas 0     no
head_node_internal_ip 10.141.255.254 no
[basecm10->kubernetes[default]->appgroups[system]->applications[calico]->environment]% set calico_typha_service value calico-typha
[basecm10->kubernetes[default]->appgroups[system]->applications[calico]->environment]% set calico_typha_replicas value 3
[basecm10->kubernetes[default]->appgroups[system]->applications[calico]->environment]% commit
```
4.2 Kubernetes Setup

4.2.3 Kubernetes Optional Add-ons

The following add-ons are installed by default unless otherwise noted. However, the user can choose to skip some or all of them during the setup.

Kubernetes Dashboard

Kubernetes Dashboard is the web user interface add-on for GUI cluster administration and metrics visualization. BCM uses Kubernetes Dashboard version 2.7.0.

There are two ways to access the dashboard:


- Users on an external network can log in to `kubectl` or Kubernetes Dashboard by following the procedures described in section 4.17.

Kubernetes Metrics Server

The Kubernetes Metrics Server is an add-on that is a replacement for Heapster. It aggregates metrics, and provides container monitoring and performance analysis. It exposes metrics via an API. BCM runs Metrics server version 0.6.3.

Helm

Helm is an add-on that manages `charts`, which are packages of pre-configured Kubernetes resources. The Helm component is installed and properly configured with BCM’s Kubernetes installation by default. It is initialized and ready for use for every Kubernetes user when the Kubernetes module is loaded. BCM uses Helm version 3.

NGINX Ingress Controller

The official Kubernetes ingress controller add-on is built around the Kubernetes Ingress resource, using a ConfigMap to store the NGINX configuration. Ingress provides HTTP and HTTPS routes from outside a Kubernetes cluster to services within the cluster. Traffic routing is controlled by rules defined in the Ingress resource.

By default, BCM provides an ingress for Kubernetes Dashboard during the `cm-kubernetes-setup` run, so that the Dashboard works. Port 30080 is the default that is set for the HTTP, and port 30443 is the default that is set for HTTPS.

These 2 ports are exposed on every Kubernetes node, both masters and workers.

The Ingress Controller is deployed as a `NodePort` which means it comes with a default range of possible port values of 30000-32767.

NVIDIA Device Plugin For Kubernetes

An older alternative to the Kubernetes NVIDIA GPU Operator (section 6.3) is the NVIDIA device plugin for Kubernetes, which is an add-on option in the `cm-kubernetes-setup` run. By default it is not selected. To be used, in addition to being selected, it requires that the NVIDIA GPU drivers are first installed (section 9 of the Installation Manual) on the head node and regular node container hosts.

This means that, as described in that section, the regular nodes on which the GPUs are located must have the `cuda-driver` and `cuda-dcgm` packages installed, inside the software image.

After booting up the container host regular nodes, the functioning of the DCGM (Data Center GPU Manager) tools should be checked by running `module load cuda-dcgm; dcmi config --get`.

The plugin add-on can then be installed as a selection option in the `cm-kubernetes-setup` session that is run by the system administrator after the NVIDIA GPU drivers have been installed. The plugin is somewhat beta at the time of writing (March 2019).

If `cuda-dcgm` and `cuda-driver` are installed for GPUs on the head node, then to ensure that the head node GPUs are detected by CMDaemon, it is recommended to restart CMDaemon on the head node.
with `service cmdaemon restart`.

The plugin then allows the GPUs to be consumed from the containers. Overcommitting GPUs (sharing) is not possible from containers or pods. Multiple GPUs can be requested by the container. Some documentation on its use can be found at https://github.com/NVIDIA/k8s-device-plugin.

BCM provides NVIDIA device plugin version 1.11

### 4.2.4 Kubernetes Setup From The Command Line

The `cm-kubernetes-setup` command line utility has the following usage synopsis:

```
[root@basecm10 ~]# cm-kubernetes-setup -h
usage: Kubernetes Setup cm-kubernetes-setup [-c <config_file>]
    [--cluster CLUSTER_NAME]
    [--skip-docker] [--skip-reboot]
    [--skip-image-update]
    [--add-user USERNAME_ADD]
    [--list-users]
    [--get-user GET_USER]
    [--modify-user USERNAME_MODIFY]
    [--remove-user USERNAME_REMOVE]
    [--namespace NAMESPACE]
    [--add-to-namespace]
    [--remove-from-namespace]
    [--role edit,admin,view,cluster-admin]
    [--runas-uid RUNAS_UID]
    [--runas-gids RUNAS_GIDS]
    [--user-paths USER_PATHS]
    [--allow-all-uids]
    [--operators OPERATORS]
    [--backup-permissions FILE]
    [--restore-permissions FILE]
    [--list-operators] [--psp]
    [--update-addons] [--remove]
    [--yes-i-really-mean-it]
    [--remove-ceph-storage] [--pull]
    [--images IMAGES] [--nodes NODES]
    [--node-selector NODE_SELECTOR]
    [--pull-registry-server PULL_REGISTRY_SERVER]
    [--pull-registry-username PULL_REGISTRY_USERNAME]
    [--pull-registry-email PULL_REGISTRY_EMAIL]
    [--pull-registry-password PULL_REGISTRY_PASSWORD]
    [--allow-device-mapper] [-v]
    [--store-name-aliases]
    [--no-distro-checks] [--json]
    [--output-remote-execution-runner]
    [--on-error-action debug,remotedebug,undo,abort]
    [--skip-packages]
    [--min-reboot-timeout <reboot_timeout_seconds>]
    [--allow-running-from-secondary]
    [--dev] [-h]
```

Optional arguments:

- `--cluster CLUSTER_NAME`
  Name of the referred Kubernetes cluster

- `-h, --help`
  Print this screen
4.2 Kubernetes Setup

common:
  Common arguments
  
  -c <config_file>  Load runtime configuration for plugins from a YAML config file

installing Kubernetes clusters:
  Flags for installing or managing Kubernetes clusters
  
  --skip-docker  Skip the Docker installation steps.
  --skip-reboot  Skip the reboot steps.
  --skip-image-update  Skip the image update steps.

user management:
  Flags for adding a new user in a Kubernetes cluster
  
  --add-user USERNAME_ADD  Create a new user in a Kubernetes cluster
  --list-users  Get information about configured Kubernetes users
  --get-user GET_USER  Get information about configured Kubernetes users
  --modify-user USERNAME_MODIFY  Modify user in a Kubernetes cluster
  --remove-user USERNAME_REMOVE  Remove existing user from a Kubernetes cluster
  --namespace NAMESPACE  Specify namespace for user (--get-user, --modify-user) role binding
  --add-to-namespace  Indicate if permissions to manage namespace needs to be granted for a given user
                       (--modify-user)
  --remove-from-namespace  Indicate if permissions to manage namespace needs to be revoked for a given user
                           (--modify-user)
  --role edit,admin,view,cluster-admin  Specify role for the new (--add-user) and existing (--modify-user) role binding
                                      (Default: edit). For 'cluster-admin' namespace flag is ignored
  --runas-uid RUNAS_UID  UID is allowed to be used in unprivileged pods (--add-user, --modify-user)
  --runas-gids RUNAS_GIDS  Comma-separated list of GIDs allowed to be used in unprivileged pods
                           (--add-user, --modify-user)
  --user-paths USER_PATHS  Comma-separated list of paths user is able to mount in pods
                           (--add-user, --modify-user)
  --allow-all-uids  Allow user to run processes in pods as any user (--add-user, --modify-user)
                   hostPath volumes will be disabled for such pods
  --operators OPERATORS  Comma-separated list of operators user has access to (--add-user, --modify-user)

backup or restore Permission Manager user configurations:
  Flag for managing permission manager user configuration
  
  --backup-permissions FILE  Save permissions to file
  --restore-permissions FILE  Restore permissions from file. Workload which is already run by users in their
                               namespaces will be affected
list available operators:
  Flag to list available Kubernetes operators

  --list-operators List available Kubernetes operators

update kubernetes addons:
  Flags for updating Kubernetes addons

  --update-addons Update Addons

removing Kubernetes clusters:
  Flags for removing a Kubernetes cluster

  --remove Remove a Kubernetes cluster
  --yes-i-really-mean-it Required for additional safety
  --remove-ceph-storage Remove Kubernetes osd pool from Ceph cluster

pulling images to the nodes:
  Flags for pulling images to the nodes

  --pull Pull images to the nodes
  --images IMAGES Comma-separated list of images to pull (--pull)
  --nodes NODES Comma-separated list of nodes to pull images to (--pull)
  --node-selector NODE_SELECTOR
    Selector (label query) to filter on. supports '=', '==', and '!='. (e.g. key1=value1,key2=value2) (--pull)
  --pull-registry-server PULL_REGISTRY_SERVER
    Registry server to authenticate
  --pull-registry-username PULL_REGISTRY_USERNAME
    Registry username
  --pull-registry-email PULL_REGISTRY_EMAIL
    Email to authenticate
  --pull-registry-password PULL_REGISTRY_PASSWORD
    Registry password

Docker storage backend specific:

  --allow-device-mapper
    Allow to select DeviceMapper (DEPRECATED) storage in wizard

advanced:
  Various *advanced* configuration options flags.

  -v, --verbose Verbose output
  --store-name-aliases Store hostname aliases for head nodes (active and passive) and default category
  --no-distro-checks Disable distribution checks based on ds.json
  --json Use json formatting for log lines printed to stdout
  --output-remote-execution-runner
    Format output for CMDaemon
  --on-error-action debug,remotedebug,undo,abort
    Upon encountering a critical error, instead of asking the user for choice, setup will do selected action.
  --skip-packages Skip the any stages which install packages. Requires packages to be already installed.
4.2 Kubernetes Setup

--min-reboot-timeout <reboot_timeout_seconds>
    How long to wait for nodes to finish reboot (default and minimum allowed: 300 seconds).
--allow-running-from-secondary
    Allow to run the wizard from the secondary when it is the active head node.
--dev
    Enables additional command line arguments

The cm-kubernetes-setup utility should be executed on the console.

Dealing With A Pre-existing Docker Installation
Docker is required for Kubernetes configured by NVIDIA Base Command Manager. The setup wizard checks if Docker has been installed (page 5), and automatically installs Docker, if needed. However, if Docker has already been configured on the same category of nodes on which Kubernetes is to be installed, then the installation stops, because overriding the existing Docker configuration may not be what is wanted. To override the existing Docker configuration, Docker for that category should first be removed with the cm-docker-setup --remove command.

Dealing With A Pre-existing Etcd Cluster
Etcd is required by Kubernetes to store all the key-value states of the Kubernetes cluster. If no Etcd cluster is found, then the setup wizard prompts to deploy an Etcd cluster. If Etcd is already installed, or present from a previous Kubernetes cluster, then the setup wizard prompts on whether to use the existing Etcd cluster.

Dealing With A Pre-existing Harbor Installation
Since Harbor uses Docker internally, and because Kubernetes customizes Docker networking, it means that nodes where Harbor is running cannot be reused for Kubernetes.

4.2.5 Kubernetes Setup From A TUI Session
When the Kubernetes installation is carried out using cm-kubernetes-setup without any options, a TUI wizard starts up. The administrator can answer several questions in the wizard. Questions that are asked include questions about which of the node categories or which of the individual nodes should be configured to run the Kubernetes services. There are also questions about the service and pod networks parameters, the port numbers that will be configured for the daemons, whether to install specific add-ons, and so on. After the wizard has been completed, a configuration file can be saved that can be used to set up Kubernetes.

The configuration file can be deployed immediately from the wizard, or it can be deployed later by specifying it as an option to cm-kubernetes-setup, in the form -c <file>.

If no deployment has been carried out earlier, then the main operations screen of the wizard shows just two options, Deploy and Exit.

If deployment has already been carried out, then the further options that are available are also displayed (figure 4.4):

Example

[root@basecm10 ~]# cm-kubernetes-setup

TUI session starts up:
The deployment via CLI or via TUI assigns the appropriate roles, and adds the new Kubernetes cluster. The deployment adds health checks to the monitoring configuration, and it generates certificates for the Kubernetes daemons.

It installs a container runtime:

- The container runtime deployed by default is Containerd (figure 4.5).
- Alternatively, the following Docker runtimes can be deployed:
  - The NVIDIA Base Command Manager Docker runtime
  - A third party Docker package can also be used, for example: an existing (non-BCM) Docker already on the cluster.

However, Docker is deprecated since Kubernetes version 1.21, so neither of the Docker runtime options is recommended.

Calico is set as the Kubernetes network plugin by default. Flannel is an option.

The master, worker, and etcd nodes can be assigned to specific nodes or categories.

The network configuration settings for the Kubernetes cluster can be specified. Ports have default assignments, but can be re-assigned as needed. The etcd spool file path can be set.

The following options are also possible:

- a registry mirror from DockerHub can be specified
- the Kubernetes API server can be exposed to the external network
4.2 Kubernetes Setup

- the internal network used by Kubernetes nodes can be selected

Add-ons that are available are:

- Ingress Controller (Nginx)
- Kubernetes Dashboard
- Kubernetes Metrics Server
- Kubernetes State Metrics
- NVIDIA device plugin for Kubernetes

Operator packages are application managers, and are described further in Chapter 6. Operators that can be installed are:

- cm-jupyter-kernel-operator
- cm-kubernetes-postgresql-operator
- cm-kubernetes-spark-operator

The permission manager—an application for role-based access control—can also be configured.

4.2.6 Testing Kubernetes

To test that Kubernetes works, the `cluster-info` command can be run. For example, on the head node, `basecm10`:

Example

```
[root@basecm10 ~]# module load kubernetes
[root@basecm10 ~]# kubectl cluster-info
Kubernetes control plane is running at https://localhost:10443
CoreDNS is running at https://localhost:10443/api/v1/namespaces/kube-system/services/kube-dns:dns/proxy
```

To further debug and diagnose cluster problems, use `kubectl cluster-info dump`.

After `cm-kubernetes-setup` finishes, and the regular nodes have been rebooted, the state of the nodes can be found by running the `get nodes` command:

Example

```
[root@basecm10 ~]# kubectl get nodes
NAME STATUS ROLES AGE VERSION
node001 Ready <none> 25m v1.28
node002 Ready <none> 25m v1.28
node003 Ready <none> 25m v1.28
node004 Ready <none> 25m v1.28
node005 Ready <none> 25m v1.28
node006 Ready <none> 25m v1.28
```

A six node cluster should show the following Kubernetes installed add-ons, when using `kubectl` with the `get all -n kube-system` option (some lines truncated):

Example
The administrator can now configure the cluster to suit the particular site requirements.

4.3 Using GPUs With Kubernetes: NVIDIA GPUs

4.3.1 Prerequisites

The GPUs have to be recognized by the nodes, and have the appropriate drivers (such as cuda-driver) installed. Details on how to do this are given in section 9 of the Installation Manual.

To verify the GPUs are recognized and have drivers in place, the nvidia-smi command can be run. The response displayed for a GPU should look similar to the following:

Example

root@node001:~# nvidia-smi
Tue Dec 7 11:25:21 2021
+-----------------------------------------------------------------------------+
4.3 Using GPUs With Kubernetes: NVIDIA GPUs

Docker as a container runtime is deprecated since Kubernetes v1.21 (page 30). It is still possible to use Docker with Kubernetes but it is recommended to deploy Containerd for new Kubernetes setups. The setups for NVIDIA GPUs are described next for both Docker and Containerd:

4.3.2 Existing Containerd Deployment
If a non-BCM Containerd has already been deployed before Kubernetes is deployed, then cm-kubernetes-setup may replace an existing Containerd configuration file in order to enable NVIDIA GPU integration via a Kubernetes CNI plugin. This is because Containerd is configured by cm-kubernetes-setup, overwriting any previous configuration.

4.3.3 Existing Docker Deployment
If a non-BCM Docker has already been deployed before Kubernetes is deployed, then the instructions from section 2.7 must be followed first. These are the instructions on making sure that GPUs can be used inside Docker. The NVIDIA container toolkit (cm-nvidia-container-toolkit) has to be present on the nodes.

To verify Docker is working with GPUs, nvidia-smi can be run from inside a container:

Example

root@node001:~# module load docker
root@node001:~# docker run --runtime=nvidia --rm nvidia/cuda nvidia-smi

Mon Sep 28 13:13:39 2020

+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.51.06 Driver Version: 450.51.06 CUDA Version: 11.0 |
|-----------------------------------------------------------------------------+
| GPU Name Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC |
| Fan Temp Perf Pwr:Usage/Cap| Memory-Usage | GPU-Util Compute M. |
| | | MIG M. |
|================================================================-------------|
| 0 Tesla K40c On | 00000000:00:06.0 Off | Off |
| 23% 28C P8 21W / 235W | 0MiB / 12206MiB | 0% Default |
| | | N/A |
+-----------------------------------------------------------------------------+
4.3.4 New Kubernetes Installation

If Containerd is selected as container runtime, then cm-kubernetes-setup assigns a new role to the Kubernetes worker nodes: generic::containerd.

The role has a Configurations submode, in which the containerd-cri object can be configured. The entry for Filename specifies the path to the cri.toml file, which contains content used by the container runtime interface on the Kubernetes worker nodes that have been assigned the role.

Example

```
[basecm10->configurationoverlay[kube-default-worker]->roles]% use generic::containerd
[...]->roles[generic::containerd]]% show
Parameter       Value
----------------- -----------------------------------------------
Name             generic::containerd
Type             GenericRole
Add services     yes
Provisioning associations <0 internally used>
Services         containerd
Configurations   <2 in submode>
Environments     <1 in submode>
Exclude list snippets <1 in submode>
Data node        no
[...]->roles[generic::containerd]]% configurations
[...]->roles[generic::containerd]->configurations% use containerd-cri
[...]->roles[generic::containerd]->configurations[containerd-cri]% show
Parameter       Value
----------------- -----------------------------------------------
Name             containerd-cri
Type             static
Create directory yes
Filename         /cm/local/apps/containerd/var/etc/conf.d/cri.toml
Filemask directory 0644
User name
Group name
Disabled         no
Service action on write RESTART
Service stop on failure yes
Content          <645B>
Filemask         0644
```

The file with the CRI (Container Runtime Interface) configuration is created in directory:

```
/cm/local/apps/containerd/var/etc/conf.d
```

and included into the main Containerd configuration file:

```
/cm/local/apps/containerd/var/etc/config.toml
```

with the imports statement:

```
imports = ["/cm/local/apps/containerd/var/etc/conf.d/*.toml"]
```
Whatever the container runtime that is selected, if NVIDIA GPU integration is required then the NVIDIA container toolkit is taken care of by the installer.

There are two ways to integrate NVIDIA GPUs into Kubernetes:

1. With the NVIDIA GPU operator for Kubernetes. This is discussed further in section 6.3.

2. With an add-on called the NVIDIA device plugin for Kubernetes. This is the way discussed in this section.

During the installation of Kubernetes via cm-kubernetes-setup (section 4.2.5) there is a step where custom add-ons can be selected (figure 4.6):

Figure 4.6: Kubernetes setup TUI session: deployment of add-ons

The add-on NVIDIA device plugin for Kubernetes should be enabled in that step. This ensures that cm-nvidia-container-toolkit is installed in the software image.

This is the default approach. It can be verified to work by deploying a pod that requests a GPU, as requested using the nvidia.com/gpu: 1 line inside the yaml file. Alternatively, the GPU operator can be deployed via Helm (page 107).

If more GPUs are available on a single host, then only one GPU should be made visible, and recognized inside the pod, when requesting a single GPU, as in the example:

Example

```yaml
root@cluster:~> cat gpu.yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-pod
spec:
  restartPolicy: Never
containers:
- name: cuda-container
  image: nvidia/cuda:9.2-runtime
  command: ["nvidia-smi"]
  resources:
    limits:
      nvidia.com/gpu: 1
root@cluster:~> kubectl apply -f gpu.yaml
pod/gpu-pod configured
root@cluster:~> kubectl logs gpu-pod
Mon Sep 28 12:12:46 2020
+-----------------------------------------------------------------------------+
<table>
<thead>
<tr>
<th>NVIDIA-SMI 450.51.06 Driver Version: 450.51.06 CUDA Version: 11.0</th>
</tr>
</thead>
</table>
```

4.3.5 Existing Kubernetes Installation

The NVIDIA device plugin can be deployed either as an add-on (this chapter) or as an operator (section 6.3). If during the installation the add-on was not selected (section 4.3.4 shows a screenshot), then it can be enabled in cmsh or Base View afterwards.

```
[cluster->kubernetes[default]->appgroups[system]->applications[nvidia]]% set enabled yes
[cluster->kubernetes[default]->appgroups[system]->applications[nvidia]]% commit
```

This add-on deploys a DaemonSet that runs the NVIDIA device plugin for Kubernetes on nodes with the `brightcomputing.com/gpu-accelerator` label.

NVIDIA Base Command Manager is responsible for setting this GPU label on the appropriate Kubernetes nodes, with parameter values specified for the labelset:

```
[cluster->kubernetes[default]->labelsets[nvidia]]% show
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>nvidia</td>
</tr>
<tr>
<td>Revision</td>
<td></td>
</tr>
<tr>
<td>Labels</td>
<td><code>brightcomputing.com/gpu-accelerator=</code></td>
</tr>
<tr>
<td>Nodes</td>
<td>node001</td>
</tr>
<tr>
<td>Categories</td>
<td></td>
</tr>
<tr>
<td>Overlays</td>
<td></td>
</tr>
</tbody>
</table>

These labels should appear in Kubernetes Node resources as a result, and the device plugin running on these particular nodes (some columns of output have been removed for clarity):

```
[root@head ~] kubectl get nodes --show-labels
```

<table>
<thead>
<tr>
<th>NAME</th>
<th>STATUS</th>
<th>ROLES</th>
<th>AGE</th>
<th>VERSION</th>
<th>LABELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>node001</td>
<td>Ready</td>
<td>&lt;none&gt;</td>
<td>16h</td>
<td>v1.21.4</td>
<td>beta.kube...com/gpu-accelerator=...hostname=node001...</td>
</tr>
<tr>
<td>node002</td>
<td>Ready</td>
<td>&lt;none&gt;</td>
<td>16h</td>
<td>v1.21.4</td>
<td>beta.kube...com/node-category=de...hostname=node002...</td>
</tr>
<tr>
<td>node003</td>
<td>Ready</td>
<td>&lt;none&gt;</td>
<td>16h</td>
<td>v1.21.4</td>
<td>beta.kube...com/node-category=de...hostname=node003...</td>
</tr>
<tr>
<td>node004</td>
<td>Ready</td>
<td>&lt;none&gt;</td>
<td>16h</td>
<td>v1.21.4</td>
<td>beta.kube...com/node-category=de...hostname=node004...</td>
</tr>
<tr>
<td>node005</td>
<td>Ready</td>
<td>&lt;none&gt;</td>
<td>16h</td>
<td>v1.21.4</td>
<td>beta.kube...com/node-category=de...hostname=node005...</td>
</tr>
<tr>
<td>node006</td>
<td>Ready</td>
<td>&lt;none&gt;</td>
<td>16h</td>
<td>v1.21.4</td>
<td>beta.kube...com/node-category=de...hostname=node006...</td>
</tr>
<tr>
<td>node007</td>
<td>Ready</td>
<td>&lt;none&gt;</td>
<td>16h</td>
<td>v1.21.4</td>
<td>beta.kube...com/node-category=de...hostname=node007...</td>
</tr>
</tbody>
</table>

4.4 Using GPUs With Kubernetes: AMD GPUs

4.4.1 Prerequisites

The GPUs have to be recognized by the node. One way to check this from within BCM is to run `sysinfo` for the node:

Example

```
% sysinfo | grep GPU
Number of GPUs 1
GPU Driver Version 4.18.0-193.el8.x86_64
GPU0 Name Radeon Instinct MI25
```

In order to make Kubernetes aware of nodes that have AMD GPUs, the AMD GPU device plugin has to be deployed as a DaemonSet inside Kubernetes. The official GitHub repository that hosts this plugin can be found at:

https://github.com/RadeonOpenCompute/k8s-device-plugin

The device plugin requires Kubernetes v1.16+, which has been around since NVIDIA Base Command Manager version 9.0. With some extra instructions, the plugin can also be made a part of NVIDIA Base Command Manager version 8.2.

The DaemonSet YAML file can be deployed with:

Example

```
kubectl create -f https://raw.githubusercontent.com/RadeonOpenCompute/k8s-device-plugin/v1.16/k8s-ds-amdgpu-dp.yaml
```

4.4.2 Managing The YAML File Through CMDaemon

The plugin can be added by the user via the Kubernetes appgroups as an application. In the session that follows, it is given the arbitrary name `device-plugin`:

Example

```
% wget https://raw.githubusercontent.com/RadeonOpenCompute/k8s-device-plugin/v1.16/k8s-ds-amdgpu-dp.yaml -O /tmp/k8s-ds-amdgpu-dp.yaml
% cmsh
% kubernetes
% appgroups
% add amd
% applications
% add device-plugin
```

The configuration of the plugin can be set to the YAML file, by setting the `config` parameter to take the value of the YAML file path.

```
% set /tmp/k8s-ds-amdgpu-dp.yaml/config
```
Example

```bash
[basecm10->...[amd*]->applications*[device-plugin*]]% set config /tmp/k8s-ds-amdgpu-dp.yaml
[basecm10->kubernetes*[default*]->appgroups*[amd*]->applications*[device-plugin*]]% show
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>device-plugin</td>
</tr>
<tr>
<td>Revision</td>
<td></td>
</tr>
<tr>
<td>Enabled</td>
<td>yes</td>
</tr>
<tr>
<td>Config</td>
<td>&lt;914B&gt;</td>
</tr>
<tr>
<td>Environment</td>
<td>&lt;0 in submode</td>
</tr>
<tr>
<td>Exclude list snippets</td>
<td>&lt;0 in submode</td>
</tr>
</tbody>
</table>
```

The YAML file can also be edited within `cmsh` after it has been set, by running `set config` without a value.

There are older releases available, starting from Kubernetes v1.10, if needed. Saving this device-plugin YAML should result in pods being scheduled on all the non-tainted nodes, as seen by listing the pods (some columns elided):

```bash
[root@basecm10 ~]# module load kubernetes/default/
[root@basecm10 ~]# kubectl get pod -n kube-system -l name=amdgpu-dp-ds -o wide
| NAME                              | READY | STATUS  | ... | IP       | NODE ...
|-----------------------------------|-------|---------|-----|----------|---------|
| amdgpu-device-plugin-daemonset-66jl7 | 1/1   | Running | ... | 172.29.112.135 | gpu001 ...
| amdgpu-device-plugin-daemonset-8mh9w | 1/1   | Running | ... | 172.29.152.130 | gpu002 ...
```

4.4.3 Including Head Nodes as part of the DaemonSet:

BCM taints head nodes, so that they do not run non-critical pods. The taint can be removed with the "-" operator to allow non-critical pods to run:

Example

```bash
kubectl taint nodes basecm10 node-role.kubernetes.io/master-
```

However, a more specific exception can be configured in the DaemonSet itself.

Within the YAML file, the following existing tolerations definition has to be modified, from:

```yaml
tolerations:
  - key: CriticalAddonsOnly
    operator: Exists
```

to:

```yaml
tolerations:
  - key: node-role.kubernetes.io/master
    effect: NoSchedule
    operator: Exists
```

The modified toleration tolerates this taint, and therefore has the device plugin run on such tainted nodes.
Verifying That AMD GPUs Are Recognized By Kubernetes

If Kubernetes is aware of the AMD GPUs for a node then several mentions of `amd.com/gpu` are displayed when running the `kubectl describe node` command for the node. The following session shows output for a node `gpu01`, ellipsized for clarity:

**Example**

```
[root@basecm10 ~]# kubectl describe node gpu01
Name: gpu01
...
Capacity:
  amd.com/gpu: 3
  cpu: 64
  ephemeral-storage: 1813510Mi
  hugepages-1Gi: 0
  hugepages-2Mi: 0
  memory: 527954676Ki
  pods: 50
...
```

### 4.4.4 Running The DaemonSet Only On Specific Nodes

The AMD GPU device plugin, unlike the NVIDIA GPU device plugin Daemonset, is scheduled to run on each Kubernetes host. This means that it runs even if the host has no GPU.

This can be prevented with the following steps:

A LabelSet can be created via `cmsh`, and the nodes or categories that have GPUs are assigned within the `labelsets` mode:

**Example**

```
[root@basecm10 ~]# cmsh
[basecm10]% kubernetes
[basecm10->kubernetes[default]]% labelsets
[basecm10->kubernetes[default]->labelsets]% use nvidia
[basecm10->kubernetes[default]->labelsets[nvidia]]% .. #but, we're using AMD GPUs, so let's go back up:
[basecm10->kubernetes[default]->labelsets]% add amd
[basecm10->kubernetes[default]->labelsets[amd]]% set labels brightcomputing.com/amd-gpu-accelerator=
[basecm10->kubernetes[default]->labelsets[amd]]% append categories gpu-nodes
[basecm10->kubernetes[default]->labelsets[amd]]% commit
```

This assigns the labels to the nodes with GPUs. This can be verified with:

**Example**

```
kubectl get nodes -l brightcomputing.com/amd-gpu-accelerator=
NAME      STATUS   ROLES   AGE   VERSION
gpu001     Ready    master 66m   v1.18.8
gpu002     Ready    master 66m   v1.18.8
...
```

The DaemonSet YAML can now be adjusted to only run the device plugin on nodes with this new label. This can be done by adding an affinity block after the tolerations block:

**Example**
tolerations:
- key: CriticalAddonsOnly # toleration may be different, if changes were made to it
  operator: Exists
affinity:
  nodeAffinity:
    requiredDuringSchedulingIgnoredDuringExecution:
      nodeSelectorTerms:
        - matchExpressions:
          - key: 'brightcomputing.com/amd-gpu-accelerator'
            operator: Exists

This results in the device plugin pods being removed immediately from all nodes that do not have the label.

4.4.5 Running An Example Workload
An example workload can be run as described in the official AMD GPU Kubernetes device plugin documentation at:

https://github.com/RadeonOpenCompute/k8s-device-plugin/tree/v1.16#example-workload

Thus it should now be possible to run:

[root@basecm10 ~]# kubectl create -f https://raw.githubusercontent.com/RadeonOpenCompute/k8s-device-plugin/v1.16/example/pod/alexnet-gpu.yaml

The YAML requests only one GPU at the bottom of the YAML file:

apiVersion: v1
kind: Pod
metadata:
  name: alexnet-tf-gpu-pod
  labels:
    purpose: demo-tf-amdgpu
spec:
  containers:
    - name: alexnet-tf-gpu-container
      image: rocm/tensorflow:latest
      workingDir: /root
      env:
      - name: HIP_VISIBLE_DEVICES
        value: "0" # # 0,1,2,...,n for running on GPU and select the GPUs, -1 for running on CPU
      command: ["/bin/bash", ":c", "--"]
      args: ["python3 benchmarks/scripts/tf_cnn_benchmarks/tf_cnn_benchmarks.py --model=alexnet;\ntrap : TERM INT; sleep infinity & wait"]
      resources:
      limits:
        amd.com/gpu: 1 # requesting a GPU

Container creation might take a while due to the image size. Once scheduled, it prints out that it found exactly one GPU, and proceeds to run a TensorFlow workload.

Example

[root@basecm10 ~]# kubectl logs -f alexnet-tf-gpu-pod
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/tensorflow/python/compat/v2_compat.py:96:
4.5 Kubernetes Configuration Overlays

disable_resource_variables (from tensorflow.python.ops.variable_scope) is deprecated and will be removed in a future version.

Instructions for updating:
- non-resource variables are not supported in the long term

Had more GPUs been requested, more would have been made available to the container.

For comparison, a CPU version of the container is also available. The official instructions can be referred to for these, too.

4.5 Kubernetes Configuration Overlays

A list of configuration overlays can be seen from within configurationoverlay mode:

Example

```
[basecm10->configurationoverlay]% list
Name (key)     Priority Nodes Categories Roles
-------------- ---------- ----------------- ------------- ------------------...
kube-default-etcd 500 node001..node003 Etdc::Host
kube-default-master 510 node001..node003 Docker::Host. Kube...
kube-default-worker 500 node004..node006 default Docker::Host. Kube...
```

Configuration overlays can be used to manage the Kubernetes services used with a particular configuration. For example, when managing the Kubernetes services used for a Kubernetes engine within an Auto Scale tracker (section 8.4.9 of the Administrator Manual).
A Kubernetes cluster can be removed using `cm-kubernetes-setup` with the `--remove` and `--yes-i-really-mean-it` options. Also, if there more than one cluster present, then the cluster name must be specified using the `--cluster` parameter.

A removal run looks as follows (some output ellipsized):

Example

```
[root@basecm10 ~]# cm-kubernetes-setup --remove --cluster default --yes-i-really-mean-it
```

Connecting to CMDaemon
Executing 20 stages

```
# Starting execution for 'Kubernetes Setup'

- kubernetes
- docker
## Progress: 0
### stage: kubernetes: Get Kube Cluster
## Progress: 5
### stage: kubernetes: Check Kube Cluster Exists
## Progress: 10
### stage: kubernetes: Find Installed Components
## Progress: 15
### stage: kubernetes: Find Files On Headnodes
## Progress: 20
### stage: kubernetes: Firewall Zone Close
## Progress: 25
### stage: kubernetes: Firewall Interface Close
## Progress: 30
### stage: kubernetes: Firewall Policy Close
## Progress: 35
### stage: kubernetes: Nginx Reverse Proxy Close
## Progress: 40
### stage: kubernetes: IP Ports Close
## Progress: 60
### stage: kubernetes: Remove Installed Components
## Progress: 65
### stage: kubernetes: Remove Files On Headnodes
## Progress: 70
### stage: kubernetes: Remove Etcd Spool
## Progress: 80
### stage: kubernetes: Set Reboot Required
You need to reboot 2 nodes to cleanup the network configuration
## Progress: 85
### stage: kubernetes: Collection Update Provisioners
## Progress: 100

Took: 00:08 min.
Progress: 100/100
```

Finished execution for 'Kubernetes Setup'. status: completed

Kubernetes Setup finished!

Using the `--remove` option removes the Kubernetes cluster configuration from BCM, unassigns Kubernetes-related roles—including the EtcdHost role—and removes Kubernetes health checks. The command does not remove packages that were installed with a `cm-kubernetes-setup` command before that.
After the disabling procedure has finished, the cluster has no Kubernetes configured and running.

4.7 Kubernetes Cluster Configuration Options

Kubernetes allows many Kubernetes clusters to be configured. These are separated sets of hosts with different certificates, users and other global settings.

When carrying out the Kubernetes setup run, a Kubernetes cluster name will be asked, and a new object with the cluster settings is then added into the CMDaemon configuration. The administrator can change the settings of the cluster from within the Kubernetes mode of cmsh or within the Kubernetes Clusters options window of Base View, accessible via the clickpath Containers→Kubernetes Clusters.

The cmsh equivalent looks like:

Example

```bash
[root@basecm10 ~]# cmsh
[root@basecm10 %] kubernetes list
Default
[root@basecm10 %] kubernetes use default
[root@basecm10->kubernetes[default]]% show
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kubernetes role bindings</td>
<td>&lt;0 in submode&gt;</td>
</tr>
<tr>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>default</td>
</tr>
<tr>
<td>Authorization Mode</td>
<td>Node,RBAC</td>
</tr>
<tr>
<td>Kube Config</td>
<td>/cm/local/apps/kubernetes/var/etc/node.kubeconfig</td>
</tr>
<tr>
<td>Kube Client Config</td>
<td>/cm/local/apps/kubernetes/var/etc/kubelet.kubeconfig</td>
</tr>
<tr>
<td>Kube Config Template</td>
<td>&lt;409 bytes&gt;</td>
</tr>
<tr>
<td>CA</td>
<td>/cm/local/apps/kubernetes/var/etc/kubeca-default.pem</td>
</tr>
<tr>
<td>CA Key</td>
<td>/cm/local/apps/kubernetes/var/etc/kubeca-default.key</td>
</tr>
<tr>
<td>Kubernetes Certificate</td>
<td>/cm/local/apps/kubernetes/var/etc/node.pem</td>
</tr>
<tr>
<td>Kubernetes Key</td>
<td>/cm/local/apps/kubernetes/var/etc/node.key</td>
</tr>
<tr>
<td>Kubernetes Client Certificate</td>
<td>/cm/local/apps/kubernetes/var/etc/kubelet.pem</td>
</tr>
<tr>
<td>Kubernetes Client Key</td>
<td>/cm/local/apps/kubernetes/var/etc/kubelet.key</td>
</tr>
<tr>
<td>Service Accounts Certificate</td>
<td>/cm/local/apps/kubernetes/var/etc/sa-default.pem</td>
</tr>
<tr>
<td>Service Accounts Certificate Key</td>
<td>/cm/local/apps/kubernetes/var/etc/sa-default.key</td>
</tr>
<tr>
<td>Kubernetes API Aggregator Certificate</td>
<td>/cm/local/apps/kubernetes/var/etc/apiaggregator.pem</td>
</tr>
<tr>
<td>Kubernetes API Aggregator Certificate Key</td>
<td>/cm/local/apps/kubernetes/var/etc/apiaggregator.key</td>
</tr>
<tr>
<td>Cluster Domain</td>
<td>cluster.local</td>
</tr>
<tr>
<td>Etcd Cluster</td>
<td>kube-default</td>
</tr>
<tr>
<td>Etcd Prefix</td>
<td>/kube-apiserver</td>
</tr>
<tr>
<td>Etcd Servers</td>
<td></td>
</tr>
<tr>
<td>Service Network</td>
<td>kube-default-service</td>
</tr>
<tr>
<td>Trusted domains</td>
<td>basecm10.local,kubernetes,kubernetes.default,kubernetes.default.svc,master,localhost</td>
</tr>
<tr>
<td>Pod Network</td>
<td>kubectl-pod</td>
</tr>
<tr>
<td>Pod Network Node Mask</td>
<td></td>
</tr>
<tr>
<td>Internal Network</td>
<td>internalenet</td>
</tr>
<tr>
<td>KubeDNS IP</td>
<td>10.150.255.254</td>
</tr>
<tr>
<td>Kubernetes API server</td>
<td></td>
</tr>
<tr>
<td>Kubernetes API server proxy port</td>
<td>10444</td>
</tr>
<tr>
<td>App Groups</td>
<td>&lt;1 in submode&gt;</td>
</tr>
</tbody>
</table>
The preceding `kubernetes` mode parameters are described in table 4.2:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorization Mode</td>
<td>Selects how to authorize on the secure port (default: RBAC (Role-Based Access Control) and Node Authorization modes)</td>
</tr>
<tr>
<td>Kube Config</td>
<td>Path to a kubeconfig file, specifying how nodes authenticate to the API server</td>
</tr>
<tr>
<td>Kube Client Config</td>
<td>Path to a kubeconfig file, specifying how kubelets authenticate to the API server</td>
</tr>
<tr>
<td>Kube Config Template</td>
<td>Template BCM uses to generate kubeconfig files for services and users.</td>
</tr>
<tr>
<td>CA</td>
<td>Path to PEM-encoded RSA or ECDSA certificate used for the CA</td>
</tr>
<tr>
<td>CA Key</td>
<td>Path to PEM-encoded RSA or ECDSA private key used for the CA</td>
</tr>
<tr>
<td>Kubernetes Certificate</td>
<td>File containing x509 certificate used by the kubelets</td>
</tr>
<tr>
<td>Kubernetes Key</td>
<td>File containing x509 private key used by the Kubernetes kubelets</td>
</tr>
<tr>
<td>Kubernetes Client Certificate</td>
<td>File containing x509 certificate used for the kubelets</td>
</tr>
<tr>
<td>Kubernetes Client Key</td>
<td>File containing x509 private key used for the Kubernetes kubelets</td>
</tr>
<tr>
<td>Service Accounts Certificate</td>
<td>File containing x509 certificate used for Kubernetes service accounts.</td>
</tr>
<tr>
<td>Service Accounts Key</td>
<td>File containing x509 private key used for Kubernetes Service Accounts.</td>
</tr>
<tr>
<td>Cluster Domain</td>
<td>Domain for this cluster</td>
</tr>
<tr>
<td>Etcd Cluster</td>
<td>The Etcd cluster instance.</td>
</tr>
<tr>
<td>Etcd Prefix</td>
<td>The prefix for all resource paths in <code>etcd</code></td>
</tr>
</tbody>
</table>

...continues
### 4.8 EtcdCluster

The EtcdCluster mode sets the global Etcd cluster settings. It can be accessed via the top level `etcd` mode of `cmsh`.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to <code>etcd</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Etcd Cluster Name.</td>
<td><code>--initial-cluster-token</code></td>
</tr>
<tr>
<td>Election Timeout</td>
<td>Election Timeout, in milliseconds.</td>
<td><code>--election-timeout</code></td>
</tr>
<tr>
<td>Heart Beat Interval</td>
<td>Heart Beat Interval, in milliseconds.</td>
<td><code>--heartbeat-interval</code></td>
</tr>
<tr>
<td>CA</td>
<td>The Certificate Authority (CA) Certificate path for Etcd, used to generate certificates for Etcd.</td>
<td><code>--peer-trusted-ca-file</code></td>
</tr>
</tbody>
</table>

---

Table 4.2: `kubernetes` mode parameters

---

...continued
...continued

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to <code>etcd</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>CA Key</td>
<td>The CA Key path for Etcd, used to generate certificates for Etcd.</td>
<td></td>
</tr>
<tr>
<td>Member Certificate</td>
<td>The Certificate path to use for Etcd cluster members, signed with the Etcd CA. The EtcdHost Role can specify a Member CA as well, and in that case it overwrites any value set here.</td>
<td>--peer-cert-file</td>
</tr>
<tr>
<td>Member Certificate Key</td>
<td>The Key path to use for Etcd cluster members, signed with the Etcd CA. The EtcdHost Role can specify a Member CA as well, and in that case it overwrites any value set here.</td>
<td>--peer-key-file</td>
</tr>
<tr>
<td>Client CA</td>
<td>The CA used for client certificates. When set it is assumed client certificate and key will be generated and signed with this CA by another party. Etcd still expects the path to be correct for the Client Certificate and Key.</td>
<td>--trusted-ca-file</td>
</tr>
<tr>
<td>Client Certificate</td>
<td>The Client Certificate, used by Etcdctl for example.</td>
<td>--cert-file</td>
</tr>
<tr>
<td>Client Certificate Key</td>
<td>The Client Certificate Key, used by Etcdctl for example.</td>
<td>--key-file</td>
</tr>
</tbody>
</table>

* Boolean (takes yes or no as a value)

Table 4.3: EtcdCluster role parameters and `etcd` options

4.9 Kubernetes Roles

Kubernetes roles include the following roles:

- EtcdHost (page 47)
- KubernetesApiServerProxy (page 47)
- KubernetesApiServer (page 48)
- KubernetesController (page 50)
- KubernetesScheduler (page 52)
- KubernetesProxy (page 53)
- KubernetesNode (page 54)

When nodes are configured using Kubernetes roles, then settings in these roles may sometimes use the same pointer variables—for example the Kubernetes or Etcd cluster instance. Pointer variables such as these have definitions that are shared across the roles, as indicated by the parameter description tables for the roles, and which are described in the following pages.

In cmsh, the roles can be assigned:

- for individual nodes via the `roles` submode of `device` mode
- for a category via the `roles` submode of a category
- for a configuration overlay via the `roles` submode of `configurationoverlay` mode
4.9 Kubernetes Roles

4.9.1 EtcdHost Role

The EtcdHost role is used to configure and manage the etcd service for a node.

The etcd service manages the etcd database, which is a hierarchical distributed key-value database. The database is used by Kubernetes to store its configurations. The EtcdHost role parameters are described in table 4.4:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to etcd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Member Name</td>
<td>The human-readable name for this etcd member ($hostname will be replaced by the node hostname)</td>
<td>--name</td>
</tr>
<tr>
<td>Spool</td>
<td>Path to the data directory (default: /var/lib/etcd)</td>
<td>--data-dir</td>
</tr>
<tr>
<td>Advertise Client URLs</td>
<td>List of client URLs for this member to advertise publicly (default: http://$hostname:5001)</td>
<td>--advertise-client-urls</td>
</tr>
<tr>
<td>Advertise Peers URLs</td>
<td>List of peer URLs for this member to advertise to the rest of the cluster (default: http://$hostname:5002)</td>
<td>--initial-advertise-peer-urls</td>
</tr>
<tr>
<td>Listen Peer URLs</td>
<td>List of URLs to listen on for peer traffic (default: http://$hostname:5002)</td>
<td>--listen-peer-urls</td>
</tr>
<tr>
<td>Snapshot Count</td>
<td>Number of committed transactions that trigger a snapshot to disk (default: 5000)</td>
<td>--snapshot-count</td>
</tr>
<tr>
<td>Debug*</td>
<td>Drop the default log level to DEBUG for all subpackages? (default: no)</td>
<td>--debug</td>
</tr>
<tr>
<td>Member Certificate</td>
<td>Etcd member certificate, signed with CA specified in the Etcd Cluster. When set it will overrule the value from the EtcdCluster object. Default empty.</td>
<td>--peer-cert-file</td>
</tr>
<tr>
<td>Member Certificate Key</td>
<td>Etcd member certificate key, signed with CA specified in the Etcd Cluster. When set it will overrule the value from the EtcdCluster object. Default empty.</td>
<td>--peer-keyt-file</td>
</tr>
<tr>
<td>Options</td>
<td>Additional parameters for the etcd daemon (empty by default)</td>
<td></td>
</tr>
</tbody>
</table>

* Boolean (takes yes or no as a value)

Table 4.4: EtcdHost role parameters and etcd options

The etcd settings are updated by BCM in /cm/local/apps/etcd/current/etc/cm-etcd.conf.

4.9.2 The KubernetesApiServerProxy Role

The KubernetesApiServerProxy role sets up a proxy that provides the entry point for one or more instances of the Kubernetes API server. The proxy runs on every node of a Kubernetes cluster instance, including the head node.
If multiple Kubernetes master nodes are present, then it enables HA for the Kubernetes master components, as described in section 4.1.

4.9.3 The KubernetesApiServer Role
The KubernetesApiServer role is used to configure and manage the `kube-apiserver` daemon. The `kube-apiserver` daemon is a Kubernetes API server that validates and configures data for the Kubernetes API objects. The API objects include pods, services, and replication controllers. The API Server processes REST operations, and provides a front end to the shared state of the cluster through which all the other components interact.

The KubernetesApiServer role parameters are described in table 4.5:
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to kube-apiserver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kubernetes Cluster</td>
<td>The Kubernetes cluster instance (pointer)</td>
<td></td>
</tr>
<tr>
<td>Insecure API Port</td>
<td>The port on which to serve unsecured, unauthenticated access (disabled by default)</td>
<td>--insecure-port</td>
</tr>
<tr>
<td>Secure API Port</td>
<td>The port on which to serve HTTPS with authentication and authorization. If 0, then HTTPS will not be served at all. (default: 6443)</td>
<td>--secure-port</td>
</tr>
<tr>
<td>Advertise Address</td>
<td>The IP address on which to advertise the API server to members of the cluster with --advertise-address. If set to 0.0.0.0, then the IP address of the management network of the head node is used. (default: 0.0.0.0)</td>
<td></td>
</tr>
<tr>
<td>Insecure Bind Address</td>
<td>IP address to serve on (default: 127.0.0.1)</td>
<td>--insecure-bind-address</td>
</tr>
<tr>
<td>Secure Bind Address</td>
<td>The IP address on which to serve the read- and secure ports (default: 0.0.0.0)</td>
<td>--bind-address</td>
</tr>
<tr>
<td>Admission Control</td>
<td>Ordered list of plug-ins to control the admission of resources into the cluster) (default: NamespaceLifecycle,LimitRanger, ServiceAccount,PersistentVolumeLabel, DefaultStorageClass,ValidatingAdmissionWebhook, ResourceQuota,DefaultTolerationSeconds, MutatingAdmissionWebhook)</td>
<td>--admission-control</td>
</tr>
<tr>
<td>Allowed Privileged*</td>
<td>If yes, then allow privileged containers. (default: no)</td>
<td>--allow-privileged</td>
</tr>
<tr>
<td>Event TTL</td>
<td>Time period that events are retained. Empty by default. A format example: 1h0m0s</td>
<td>--event-ttl</td>
</tr>
<tr>
<td>Kubelet Timeout</td>
<td>Kubelet port timeout (default: 5s)</td>
<td>--kubelet-timeout</td>
</tr>
<tr>
<td>Log Level</td>
<td>Log level (default: 0)</td>
<td>--v</td>
</tr>
<tr>
<td>Log To StdErr*</td>
<td>Logging to stderr means it goes into the systemd journal (default: yes)</td>
<td>--logtostderr</td>
</tr>
</tbody>
</table>

...continues
Further details on the Kubernetes API Server can be found at https://kubernetes.io/docs/admin/kube-apiserver/.

### 4.9.4 KubernetesController Role

The Kubernetes Controller role is used to configure and manage the `kube-controller-manager` daemon that embeds the core control loops shipped with Kubernetes. In Kubernetes, a controller is a control loop that watches the shared state of the cluster through the API server, and it makes changes in order to try to move the current state towards the desired state. Examples of controllers that ship with Kubernetes at the time of writing (January 2018) are:

- the replication controller
- the endpoints controller,
- the namespace controller,
- the serviceaccounts controller

The KubernetesController role parameters are described in table 4.6:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to kube-controller-manager</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kubernetes Cluster</strong></td>
<td>The Kubernetes cluster instance (pointer)</td>
<td></td>
</tr>
<tr>
<td><strong>Address</strong></td>
<td>IP address to serve on (default: 0.0.0.0)</td>
<td>--address</td>
</tr>
<tr>
<td><strong>Port</strong></td>
<td>Port to serve on (default: 10252)</td>
<td>--port</td>
</tr>
<tr>
<td><strong>Concurrent Endpoint Syncs</strong></td>
<td>Number of endpoint syncing operations that will be done concurrently. (default: 5)</td>
<td>--concurrent-endpoint-syncs</td>
</tr>
<tr>
<td><strong>Concurrent Rc Syncs</strong></td>
<td>The number of replication controllers that are allowed to sync concurrently. 5</td>
<td>--concurrent-rc-syncs</td>
</tr>
<tr>
<td><strong>Namespace Sync Period</strong></td>
<td>Period for syncing namespace life-cycle updates</td>
<td>--namespace-sync-period</td>
</tr>
<tr>
<td><strong>Node Monitor Grace Period</strong></td>
<td>Period for syncing NodeStatus in NodeController</td>
<td>--node-monitor-grace-period</td>
</tr>
<tr>
<td>Parameter</td>
<td>Description</td>
<td>Option to</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Node Monitor Period</td>
<td>Period the running Node is allowed to be unresponsive before marking it unhealthy</td>
<td>--node-monitor-period</td>
</tr>
<tr>
<td>Node Startup Grace Period</td>
<td>Period the starting Node is allowed to be unresponsive before marking it unhealthy</td>
<td>--node-startup-grace-period</td>
</tr>
<tr>
<td>Node Sync Period</td>
<td>Period for syncing nodes from cloud-provider</td>
<td>--node-sync-period</td>
</tr>
<tr>
<td>Pod Eviction Timeout</td>
<td>Grace period for deleting pods on failed nodes</td>
<td>--pod-eviction-timeout</td>
</tr>
<tr>
<td>Pv Claim Binder Sync Period</td>
<td>Period for syncing persistent volumes and persistent volume claims</td>
<td>--pvclaimbinder-sync-period</td>
</tr>
<tr>
<td>Register Retry Count</td>
<td>Number of retries for initial node registration</td>
<td>--register-retry-count</td>
</tr>
<tr>
<td>Resource Quota Sync Period</td>
<td>Period for syncing quota usage status in the system</td>
<td>--resource-quota-sync-period</td>
</tr>
<tr>
<td>Log Level</td>
<td>Log level (default: 0)</td>
<td>--v</td>
</tr>
<tr>
<td>Log To StdErr</td>
<td>Logging to stderr means getting it into the systemctl journal (default: yes)</td>
<td>--logtostderr</td>
</tr>
<tr>
<td>Options</td>
<td>Additional parameters for the kube-controller-manager daemon</td>
<td></td>
</tr>
<tr>
<td>Cluster signing Cert file</td>
<td>Filename containing a PEM-encoded X509 CA certificate used to issue cluster-scoped certificates. (leave empty to use the value of CA defined in the Kubernetes Cluster instance).</td>
<td>--cluster-signing-cert-file</td>
</tr>
<tr>
<td>Cluster signing Cert Key file</td>
<td>Filename containing a PEM-encoded RSA or ECDSA private key used to sign cluster-scoped certificates. (leave empty to use the value of CA Key defined in the Kubernetes Cluster instance).</td>
<td>--cluster-signing-key-file</td>
</tr>
<tr>
<td>Use Service Account Credentials</td>
<td>Flag to enable or disable use of Service Account Credentials.</td>
<td>--use-service-account-credentials</td>
</tr>
<tr>
<td>Allocate Node Cidrs</td>
<td>Allocate node CIDR in cluster using Pod Network and Pod Network Node Mask size defined in the Kubernetes cluster object.</td>
<td>--allocate-node-cidrs</td>
</tr>
<tr>
<td>Kube Config</td>
<td>Path to a kubeconfig file, specifying how to authenticate to API server.</td>
<td>--kubeconfig</td>
</tr>
<tr>
<td>Kubernetes Certificate</td>
<td>File containing x509 Certificate used by Kubernetes Controller Manager. This will be embedded in the Kube Config file.</td>
<td>(1)</td>
</tr>
<tr>
<td>Kubernetes Key</td>
<td>File containing x509 private key used by Kubernetes Controller Manager. This will be embedded in the Kube Config file.</td>
<td>(2)</td>
</tr>
</tbody>
</table>
Further details on the Kubernetes controller manager can be found at https://kubernetes.io/docs/admin/kube-controller-manager/.

### 4.9.5 KubernetesScheduler Role

The KubernetesScheduler role is used to configure and manage the kube-scheduler daemon. The Kubernetes scheduler defines pod placement, taking into account the individual and collective resource requirements, quality of service requirements, hardware/software/policy constraints, affinity and anti-affinity specifications, data locality, inter-workload interference, deadlines, and so on.

The KubernetesScheduler role parameters are described in table 4.7:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to kube-scheduler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kubernetes Cluster</td>
<td>The Kubernetes cluster instance (pointer)</td>
<td></td>
</tr>
<tr>
<td>Address</td>
<td>IP address to serve on (default: 0.0.0.0)</td>
<td>--address</td>
</tr>
<tr>
<td>Scheduler Port</td>
<td>Port to serve on (default: 10253)</td>
<td>--port</td>
</tr>
<tr>
<td>Algorithm Provider</td>
<td>The scheduling algorithm provider to use</td>
<td>--algorithm-provider</td>
</tr>
<tr>
<td>Policy Config</td>
<td>File with scheduler policy configuration</td>
<td>--policy-config-file</td>
</tr>
<tr>
<td></td>
<td>(default: /cm/local/apps/kubernetes/var/etc/scheduler-policy.json)</td>
<td></td>
</tr>
<tr>
<td>Log Level</td>
<td>Log level (default: 0)</td>
<td>--v</td>
</tr>
<tr>
<td>Log To StdErr*</td>
<td>Logging to STDERR means getting it into the systemd journal (default: yes)</td>
<td>--logtostderr</td>
</tr>
<tr>
<td>Options</td>
<td>Additional parameters for the kube-scheduler daemon</td>
<td></td>
</tr>
<tr>
<td>Kube Config</td>
<td>Path to a kubeconfig file, specifying how to authenticate to API server.</td>
<td>--kubeconfig</td>
</tr>
<tr>
<td>Kubernetes Certificate</td>
<td>File containing x509 Certificate used by Kubernetes Scheduler. This certificate will be embedded in the Kube Config file.</td>
<td></td>
</tr>
<tr>
<td>Kubernetes Key</td>
<td>File containing x509 private key used by Kubernetes Scheduler. This certificate key will be embedded in the Kube Config file.</td>
<td></td>
</tr>
</tbody>
</table>

* Boolean (takes yes or no as a value)

Further details on the Kubernetes scheduler can be found at https://kubernetes.io/docs/admin/kube-scheduler/.
4.9 Kubernetes Roles

4.9.6 KubernetesProxy Role

The KubernetesProxy role is used to configure and manage `kube-proxy` daemon. The `kube-proxy` daemon runs on each node, and reflects services as defined in the Kubernetes API. It can do simple TCP and UDP stream-forwarding or round-robin TCP and UDP forwarding across a set of back ends.

The KubernetesProxy role parameters are described in table 4.8:
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Configuration Parameter Passed To kubectl-proxy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kubernetes Cluster</td>
<td>The Kubernetes cluster instance (pointer)</td>
<td>--address</td>
</tr>
<tr>
<td>Address</td>
<td>IP address to serve on (default: 0.0.0.0)</td>
<td>--proxy-port-range</td>
</tr>
<tr>
<td>Proxy Port Range Start</td>
<td>Bottom of range of host ports that may be consumed in order to proxy service traffic (not set by default)</td>
<td>--proxy-port-range</td>
</tr>
<tr>
<td>Proxy Port Range End</td>
<td>Top of range of host ports that may be consumed in order to proxy service traffic (not set by default)</td>
<td>--proxy-port-range</td>
</tr>
<tr>
<td>Masquerade All</td>
<td>If using the pure iptables proxy, SNAT everything (default: yes)</td>
<td>--masquerade-all</td>
</tr>
<tr>
<td>Health Check Address</td>
<td>IP address for the health check server to serve on</td>
<td>--healthz-port</td>
</tr>
<tr>
<td>Health Check Port</td>
<td>Port to bind the health check server to serve on (default: 10251, use 0 to disable)</td>
<td>--healthz-port</td>
</tr>
<tr>
<td>Oom Score Adj</td>
<td>The oom_score_adj value for the kube-proxy process, in range [-999, 999] (default: -999)</td>
<td>--oom-score-adj</td>
</tr>
<tr>
<td>Kube Config</td>
<td>Path to a kubeconfig file, specifying how to authenticate to the API server.</td>
<td>--kubeconfig</td>
</tr>
<tr>
<td>Kubernetes Certificate</td>
<td>File containing x509 Certificate used by kube-proxy. This certificate is embedded in the Kube Config.</td>
<td></td>
</tr>
<tr>
<td>Kubernetes Key</td>
<td>File containing x509 private key used by Kubernetes API server. This certificate key is embedded in the Kube Config.</td>
<td></td>
</tr>
<tr>
<td>Log Level</td>
<td>Log level (default: 0)</td>
<td>--v</td>
</tr>
<tr>
<td>Log To StdErr*</td>
<td>Logging to STDERR means it goes in the systemd journal (default: yes)</td>
<td>--logtostderr</td>
</tr>
</tbody>
</table>

* Boolean (takes yes or no as a value)

Table 4.8: KubernetesProxy role parameters and kubectl-proxy options

Further details on the Kubernetes network proxy can be found at https://kubernetes.io/docs/admin/kube-proxy/.

4.9.7 KubernetesNode Role

The KubernetesNode role is used to configure and manage the kubelet daemon, which is the primary node agent that runs on each node. The kubelet daemon takes a set of pod specifications, called PodSpecs, and ensures that the containers described in the PodSpecs are running and healthy.

The KubernetesNode role parameters are described in table 4.9:
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to kubelet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kubernetes Cluster</td>
<td>The Kubernetes cluster instance (pointer)</td>
<td></td>
</tr>
<tr>
<td><strong>Address</strong></td>
<td>IP address to serve on (default: 0.0.0.0)</td>
<td><code>--address</code></td>
</tr>
<tr>
<td><strong>Kubelet Port</strong></td>
<td>Port that the HTTP service of the node runs on (default: 10250)</td>
<td><code>--port</code></td>
</tr>
<tr>
<td><strong>CNI plugin binaries path</strong></td>
<td>The full path of the directory in which to search for CNI plugin binaries. (default: /cm/local/apps/kubernetes/current/bin/cni)</td>
<td><code>--cni-bin-dir</code></td>
</tr>
<tr>
<td><strong>Enable Server</strong></td>
<td>Enable server mode of Kubelet</td>
<td><code>--enable-server</code></td>
</tr>
<tr>
<td><strong>Host Network Sources</strong></td>
<td>List of sources from which Kubelet allows pods use of the host network (default: file)</td>
<td><code>--host-network-sources</code></td>
</tr>
<tr>
<td><strong>Hostname Override</strong></td>
<td>If non-empty, use this string as identification instead of the actual hostname (not set by default)</td>
<td><code>--hostname-override</code></td>
</tr>
<tr>
<td><strong>Manifests Path</strong></td>
<td>Path to the config file or directory of files (default: /cm/local/apps/kubernetes/var/etc/manifests)</td>
<td><code>--pod-manifest-path</code></td>
</tr>
<tr>
<td><strong>Network plugin</strong></td>
<td>The name of the network plugin to be invoked for various events in kubelet/pod lifecycle. (default: cni)</td>
<td><code>--network-plugin</code></td>
</tr>
<tr>
<td><strong>Spool</strong></td>
<td>Directory path for managing Kubelet files (default: /cm/local/apps/kubernetes/var/kubelet)</td>
<td><code>--root-dir</code></td>
</tr>
<tr>
<td><strong>Cgroup Root</strong></td>
<td>Optional root cgroup to use for pods</td>
<td><code>--cgroup-root</code></td>
</tr>
<tr>
<td><strong>Docker Endpoint</strong></td>
<td>Docker endpoint address to connect to (default: unix:///var/run/docker.sock)</td>
<td><code>--docker-endpoint</code></td>
</tr>
<tr>
<td><strong>Docker Spool</strong></td>
<td>Absolute path to the Docker state root directory (default: /var/lib/docker)</td>
<td><code>--docker-root</code></td>
</tr>
<tr>
<td><strong>Resource Container</strong></td>
<td>Absolute name of the resource-only container to create and run the Kubelet in (default: /kubelet)</td>
<td><code>--resource-container</code></td>
</tr>
<tr>
<td><strong>Allowed Privileged</strong></td>
<td>If true, allow privileged containers. (default: no)</td>
<td><code>--allow-privileged</code></td>
</tr>
<tr>
<td><strong>Labels</strong></td>
<td>List of node labels</td>
<td></td>
</tr>
<tr>
<td><strong>Register On Start</strong></td>
<td>Register the node with the API server (default: yes)</td>
<td><code>--register-node</code></td>
</tr>
<tr>
<td><strong>Eviction minimum reclaim</strong></td>
<td>Minimum amount of resources reclaimed in an eviction (default: imagefs.available=1Gi)</td>
<td><code>--eviction-minimum-reclaim</code></td>
</tr>
</tbody>
</table>

...continues
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Option to <code>kubelet</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard eviction</td>
<td>Hard eviction constraints (default: <code>imagefs.available&lt;1%</code>)</td>
<td><code>--eviction-hard</code></td>
</tr>
<tr>
<td>Max Pods</td>
<td>Number of pods that can run on this node</td>
<td><code>--max-pods</code></td>
</tr>
<tr>
<td>Max pod eviction grace period</td>
<td>Maximum allowed grace period (in seconds) allowed to terminated pods (default: 60)</td>
<td><code>--eviction-max-pod-grace-period</code></td>
</tr>
<tr>
<td>Soft eviction</td>
<td>Soft eviction constraints (default: <code>imagefs.available&lt;5%</code>)</td>
<td><code>--eviction-soft</code></td>
</tr>
<tr>
<td>Soft eviction grace period</td>
<td>Soft eviction grace period (default: <code>imagefs.available=1m30s</code>)</td>
<td><code>--eviction-soft-grace-period</code></td>
</tr>
<tr>
<td>File Check Frequency</td>
<td>Duration between checking configuration files for new data (default: 20s)</td>
<td><code>--file-check-frequency</code></td>
</tr>
<tr>
<td>HTTP Flush Frequency</td>
<td>Duration between checking HTTP for new data (default: 20s)</td>
<td><code>--http-check-frequency</code></td>
</tr>
<tr>
<td>Node Status Update Frequency</td>
<td>The absolute free disk space, in MB, to maintain (default: 10s)</td>
<td><code>--node-status-update-frequency</code></td>
</tr>
<tr>
<td>Run Once</td>
<td>If true, exit after spawning pods from local manifests or remote URLs (default: no)</td>
<td><code>--runonce</code></td>
</tr>
<tr>
<td>Streaming Connection Idle Timeout</td>
<td>Maximum time a streaming connection can be idle before the connection is automatically closed (default: 1h)</td>
<td><code>--streaming-connection-idle-timeout</code></td>
</tr>
<tr>
<td>Sync Frequency</td>
<td>Maximum period between synchronizing running containers and config (default: 1s)</td>
<td><code>--sync-frequency</code></td>
</tr>
<tr>
<td>Image GC High Threshold</td>
<td>Percent of disk usage after which image garbage collection is always run (default: 90)</td>
<td><code>--image-gc-high-threshold</code></td>
</tr>
<tr>
<td>Image GC Low Threshold</td>
<td>Percent of disk usage before which image garbage collection is never run (default: 80)</td>
<td><code>--image-gc-low-threshold</code></td>
</tr>
<tr>
<td>Oom Score Adjust</td>
<td>The <code>oom_score_adj</code> value for the kube-proxy process, in range [-999, 999] (default: -999)</td>
<td><code>--oom-score-adj</code></td>
</tr>
<tr>
<td>Log Level</td>
<td>Log level (default: 0)</td>
<td><code>--v</code></td>
</tr>
<tr>
<td>Log To StdErr</td>
<td>Logging to STDERR means it gets into the systemd journal (default: yes)</td>
<td><code>--logtostderr</code></td>
</tr>
<tr>
<td>Options</td>
<td>Additional parameters for the kube-scheduler daemon</td>
<td></td>
</tr>
</tbody>
</table>
Further details on the kubelet daemon can be found at https://kubernetes.io/docs/admin/kubelet/.

4.10 Security Model

The Kubernetes security model allows authentication using a certificate authority (CA), with the user and daemon certificates signed by a Kubernetes CA. The Kubernetes CA should not be confused with the BCM CA.

BCM will create a CA specifically for issuing all Kubernetes-related certificates. The certificates are put into /cm/local/apps/kubernetes/var/etc/ by default, and /etc/kubernetes/ is made a link to this directory.

In Kubernetes terminology a user is a unique identity accessing the Kubernetes API server. The user may be a human or an automated process. For example an admin or a developer are human users, but kubelet represents an infrastructure user. Both types of users are authorized and authenticated in the same way against the API server.

Kubernetes uses client certificates, tokens, or HTTP basic authentication methods to authenticate users for API calls. BCM configures client certificate usage by default. The authentication is performed by the API server which validates the user certificate using the common name part of the certificate subject.

In Kubernetes, authorization happens as a separate step from authentication. Authorization applies to all HTTP accesses on the main (secure) API server port. BCM by default enables RBAC (Role-Based Access Control) combined with Node Authorization. The authorization check for any request thus takes the common name and/or organization part of the certificate subject to determine which roles the user or service has associated. Roles carry a certain set of privileges for resources within Kubernetes.

4.10.1 Kyverno

BCM has support for the Kyverno policy engine (https://kyverno.io/). If Kyverno is installed, then Kubernetes Permissions Manager (section 4.16) creates policy manifests packed as a Helm chart for every user added to Kubernetes via cm-kubernetes-setup. In addition, a kyverno-policy chart is installed in enforce mode to implement Pod Security Standards (https://kyverno.io/policies/). During installation, some exclusions are added to the policies automatically to make chosen features of Kubernetes cluster work.

For every created user the following defaults are applied:

- The user has an associated service account with the same name
- A <username>-restricted namespace is created. So, for a user john the namespace is john-restricted.
- An edit cluster role is bound to the service account in <username>-restricted namespace. The user is allowed to create pods, services, configmaps, etc. in the namespace
- The user is allowed to list nodes in the cluster
- Kyverno policies are applied to the resources in <username>-restricted namespace or to pod created or updated by the associated user
  - If hostPath is not the home directory of the user (of the format /home/<username>) then the creation of the resource is denied
The UID and GID of the running process are set to the same value as the UID and GID of the PAM user.

Modifications from the defaults are:

- If the **Allow any UID process in pods** checkbox is ticked, or if the `--allow-all-uids` argument is specified, then the UID and GID of the running process becomes the user’s UID and GID only if the hostPath volume is specified. Otherwise it can be set to any UID and GID.

- Cluster roles can be set not only to
  - `edit`
  but also to
  - `view`
  - `admin`
  - `cluster-admin`

More details on these roles can be found at: https://kubernetes.io/docs/reference/access-authn-authz/rbac/#user-facing-roles.

- In addition, the user can be given access to custom CRDs, such as Zalando Postgres Operator, Jupyter Operator or Google Spark Operator.

### 4.10.2 PodSecurityPolicy

BCM also has support for PodSecurityPolicy (PSP) in the Kubernetes API Server. PSP can be explicitly enabled or disabled (section 4.14). However, it should be noted that PodSecurityPolicy was deprecated in Kubernetes v1.21, and removed from Kubernetes in v1.25.

For each user, a PSP is generated and assigned using Helm charts generated by Kubernetes Permissions Manager. (section 4.16).

The following defaults are applied:

- Users can only mount their own home directory. They cannot mount other paths such as `/etc`

- Users cannot run privileged pods.

- Users can only bind on ports higher than 1024.

- Users can run with their own `uid` and `gid`.

- It is possible to allow users to run as `root`, *without* using hostPath volumes.

There are also

- default role bindings that grant access to the resources covered by the `kubectl get all` command.

A higher-level explanation about each of the resources is given in section 4.13.

These policies only do something if PodSecurityPolicies are enabled, or enforced by Kubernetes, which is not the default. Section 4.14 can be referred to for enabling and disabling PodSecurityPolicies.
4.11 Addition Of New Kubernetes Users

BCM users can use Kubernetes by making them Kubernetes users. This means having Kubernetes configuration and access set up for them. This can be carried out via the `cm-kubernetes-setup` TUI utility, and choosing the `Add user` option (figure 4.4). The utility then prompts for

- a Kubernetes cluster
- a user name
- a namespace that the privileges are to be assigned to
- a role for the user, with choices provided from:
  - `cluster-admin`: cluster-wide administrator
  - `admin`: administrator
  - `edit`: regular user
  - `view`: read-only user
- a switch if the user is allowed to run as any user, including root, inside pods
- a comma-separated list of paths that the user is able to mount to pods
- the UIDs and GIDs for user processes in pods
- a list of the Kubernetes operators that a user can use

Based on the input, a YAML for the Kubernetes Permission Manager is generated. This in turn, creates a Helm chart with all the required roles, role bindings, and PSP or Kyverno rules.

Creation of the user also triggers CMDaemon to create certificate and configuration files in the `~/.kube` directory.

4.11.1 Adding Users Non-Interactively With `cm-kubernetes-setup`

The `cm-kubernetes-setup` CLI wizard provides the following options:

```
usage: Kubernetes Setup cm-kubernetes-setup
 [-c <config_file>]
 [--cluster CLUSTER_NAME]
 [--skip-docker] [--skip-reboot]
 [--skip-image-update]
 [--add-user USERNAME_ADD] [--list-users] [--get-user GET_USER]
 [--modify-user USERNAME_MODIFY] [--remove-user USERNAME_REMOVE]
 [--namespace NAMESPACE] [--add-to-namespace] [--remove-from-namespace]
 [--role edit,admin,view,cluster-admin]
 [--runas-uid RUNAS_UID] [--runas-gids RUNAS_GIDS]
 [--user-paths USER_PATHS]
 [--allow-all-uids]
 [--operators OPERATORS]
 [--backup-permissions FILE] [--restore-permissions FILE]
 [--list-operators]
 [--update-addons] [--remove]
 [--yes-i-really-mean-it]
 [--remove-ceph-storage] [--pull]
 [--images IMAGES] [--nodes NODES]
 [--node-selector NODE_SELECTOR]
 [--pull-registry-server PULL_REGISTRY_SERVER]
```
The user has to be a user that exists on the cluster already and available via PAM.

If `--add-to-namespace` is specified, then the namespace has to exist on the Kubernetes cluster already.

Example

```
cm-kubernetes-setup --add-user john
```

The preceding example creates a user `john` for the default `john-restricted` namespace. It also assigns the `edit` role, and gives permission to run processes in the pod with the current UID/GIDs of the user. The ability to mount `~/.john` as a hostPath is also provided.

A way to assign any of the default Kubernetes user-facing roles is also provided by using `--role` key, as documented at https://kubernetes.io/docs/reference/access-authn-authz/rbac/

The possible roles are: `view`, `edit`, `admin`, and `cluster-admin`.

Example

```
cm-kubernetes-setup --add-user john --role view
```

The preceding example creates a user `john` with `view` privileges only, for the default `john-restricted` namespace.

Example

```
cm-kubernetes-setup --add-user john --user-paths /home/john,/scratch --allow-all-uids \
  --operators cm-jupyter-kernel-operator
```

The preceding example creates a user `john` with the following privileges:

- **edit** privileges
- able to mount `/home/john` and `/scratch` as hostPath volume, when the process runs with UID/GIDs taken from the PAM subsystem on the moment of creation
- able to run as any user, including root (attempts to mount any hostPath volume will be rejected)
- access to the Jupyter Kernel Operator, i.e. with access to the resource kind: `CmKubernetesOperatorPermissionsJupyterKernel`
4.12 Getting Information And Modifying Existing Kubernetes Users

It is possible to edit user properties and permissions. \texttt{cm-kubernetes-setup} provides 2 ways of doing it: interactively or via CLI options.

Modifying users can be done interactively by choosing Modify User in the \texttt{cm-kubernetes-setup} main menu. Guidance is then given on choosing the cluster, users, and on modifying permissions.

Modifying users can also be done via CLI options, by specifying the \texttt{--modify-user} argument:

\texttt{Example}

\begin{verbatim}
  cm-kubernetes-setup --add-user john --user-paths /home/john \ 
    --allow-all-uids --operators cm-jupyter-kernel-operator
\end{verbatim}

In addition to what is specified on creation with the \texttt{--add-user} argument in the preceding example, the following example adds permission to mount the \textasciitilde{scratch} hostPath into pods, and gives access to the Zalando PostgreSQL Operator:

\begin{verbatim}
  cm-kubernetes-setup --modify-user john --user-paths /home/john,/scratch \ 
    --allow-all-uids --operators cm-jupyter-kernel-operator.cm-kubernetes-postgresql-operator
\end{verbatim}

Information about existing users can be found with:

\texttt{Example}

\begin{verbatim}
  cm-kubernetes-setup --list-users
\end{verbatim}

Permission for user john to operate in the dev namespace can be added with:

\texttt{Example}

\begin{verbatim}
  kubectl create namespace dev
  cm-kubernetes-setup --modify-user john --namespace dev --add-to-namespace
\end{verbatim}

Permission for the user john to operate in the dev namespace can be revoked with:

\texttt{Example}

\begin{verbatim}
  kubectl create namespace dev
  cm-kubernetes-setup --modify-user john --namespace dev --remove-from-namespace
\end{verbatim}

4.13 List Of Resources Defined For Users

These resources are rendered by the Permission Manager Operator, and can therefore be found inside Kubernetes.

\textbf{The Role Bindings Deployed For Every User By Default}

By default, the role bindings deployed for the user john created in the preceding section are:

- ClusterRole/john-nodes (in namespace john-restricted)
- ClusterRoleBinding/john-nodes (in namespace john-restricted)

User john is given read-only rights for the Nodes resource (for \texttt{kubectl get nodes}).
The Secure Namespace Related Resources

The secure namespace for user john is:

- Namespace/john-restricted

The service account used by john:

- ServiceAccount/john (in namespace john-restricted)

This is found referenced, for example, in john’s $HOME/.kube/config.

The PodSecurityPolicy that defines the user can run non-privileged pods, and use only ports above 1024, and so on:

- PodSecurityPolicy/john-restricted (in namespace john-restricted)

More details on this can be found in section 4.10, page 58. This policy will only do something as soon as the PodSecurityPolicy Admission Controller is enabled in the API server.

A PodSecurityPolicy that defines the user can run as root as well, but without hostPath volumes:

- PodSecurityPolicy/john-restricted-root (in namespace john-restricted)

To give the aforementioned privileges to john’s secure namespace, so that john can run workloads, execute kubectl get all, and more:

- Role/john-restricted (in namespace john-restricted)
- RoleBinding/john-restricted (in namespace john-restricted)

The RoleBinding assigns it to the user john and ServiceAccount account for john. The upstream documentation at https://kubernetes.io/docs/reference/access-authn-authz/service-accounts-admin has more details on this.

The user john can be given the ability to use the PodSecurityPolicy defined earlier in his secure namespace, but also in other namespaces:

- ClusterRole/john-psp (in namespace john-restricted)
- ClusterRoleBinding/john-psp (in namespace john-restricted)

The same ability can be given for the second root but no hostPath PodSecurityPolicy:

- ClusterRole/john-psp-root (in namespace john-restricted)
- ClusterRoleBinding/john-psp-root (in namespace john-restricted)

If the Kyverno engine is installed then several policies are added:

- clusterpolicies.kyverno.io/john-*.drop-privs-w-hostpath*: Policy to modify pod manifests to run process with specified UID/GID
- clusterpolicies.kyverno.io/john-*.limit-hostpath-vols: Policy to deny pods if hostPath volumes does not match specified paths

The full content of all the documents created for the user can be viewed by checking the generated Helm manifest:

Example

helm get manifest -n cm-permissions john-XXXXXX
4.14 Pod Security Policies

Without additional policy engines, Kubernetes has very few restrictions on users. For a more fine-grained authorization it is possible to enable Pod Security Policies (PSPs).

This can be done via the `cm-kubernetes-setup` TUI wizard and choosing “Enable PSP”. Alternatively, it can be done non-interactively using `cm-kubernetes-setup --psp`.

Optionally, AppArmor can be enabled as well, using non-interactive command option `-apparmor`. Enabling PSP creates two new Applications within the Kubernetes AppGroup `system`:

- `psp`: this defines the policy and roles for privileged services. The cluster administrator needs to bind these to services typically running in the `kube-system` namespace.
- `psp_system`: this is auto-generated by the wizard. It binds the previously-mentioned privileges to service accounts, for services defined in the system namespaces. This way Calico, CoreDNS, Ingress, and so on, can still function.

The wizard also removes access to the default namespace for existing users, and it restarts the Kubernetes API Server with the `PodSecurityPolicy` feature enabled to enforce all privileges.

Each user should already have their own `<user>-restricted` namespace and privileges to work within this namespace. After the Kubernetes API Server is enabled with the `PodSecurityPolicy` feature, these policies are enforced after the API server has restarted.

A useful command to check exactly what a user—for example, the user `test`—can do is the following:

```
kubectl --kubeconfig=/home/test/.kube/config auth can-i --list --namespace=test-restricted
```

4.14.2 Disabling Pod Security Policies For Kubernetes
Disabling PSP can be done via the `cm-kubernetes-setup` TUI wizard, and choosing `Disable PSP`. It can alternatively be carried out non-interactively using `cm-kubernetes-setup --disable-psp`.

When disabling, it should be noted that:

- Existing users are not automatically re-added to the `default` namespace.
- Policies are still defined as resources, but are no longer enforced. This may result in more privileges for users then they had before. That is, they may be able to run as `root` in containers again.

4.14.3 Enabling Manually Via `cmsh` Instead
The `PodSecurityPolicy` feature is an admission control setting that can be added via `cmsh`:

```
[cluster->configurationoverlay[kube-default-master]->roles[Kubernetes::ApiServer]]% get admissioncontrol
NamespaceLifecycle LimitRanger ServiceAccount DefaultStorageClass DefaultTolerationSeconds MutatingAdmissionWebhook ValidatingAdmissionWebhook ResourceQuota

PodSecurityPolicy
```

Adding or removing `PodSecurityPolicy` from here in `cmsh` triggers CMDaemon to restart the `kube-apiserver` services.

It could be that running pods are not affected. However, if the cluster administrator re-creates them, then it may be that new pods are not created by ReplicaSets, DaemonSets and similar. Errors may show up as follows:
Warning FailedCreate 5m22s (x20 over 5m33s) replicaset-controller Error creating:
pods "coredns-b5cdc886c-" is forbidden: unable to validate against any pod security policy:
	[spec.containers[0].securityContext.capabilities.add: Invalid value: "NET_BIND_SERVICE":
capability may not be added spec.containers[0].securityContext.capabilities.add: Invalid
value: "NET_BIND_SERVICE": capability may not be added spec.containers[0].securityContext.capabilities.add:
Invalid value: "NET_BIND_SERVICE": capability may not be added]

Also, the kubelet services themselves will not have the proper privileges to manage their pods. If the
PodSecurityPolicy value is enabled in the admissioncontrol settings, then the cluster administrator
must be explicit and define the Pod Security Policy (PSP).

4.14.4 The psp Application

The purpose of this application is to define a PSP for these more privileged components. It starts by
assigning it to the kubelet component. Other system components such as CoreDNS or Calico are not
yet assigned to this PSP.

In cmsh, parts of the configuration setting can be seen with:

```
[cluster->kubernetes[default]->appgroups[system]->applications[psp]]% get config | grep -E "^kind|^ name"
|grepl kind -A 2
kind: PodSecurityPolicy
  name: privileged
kind: ClusterRole
  name: privileged-psp
kind: RoleBinding
  name: privileged-psp-nodes
  namespace: kube-system
```

In the preceding configuration, resources define PodSecurityPolicy/privileged, which the cluster
administrator binds to Group/system:nodes with a RoleBinding/privileged-psp-nodes, and using a
role ClusterRole/privileged-psp. The Group/system:nodes allows access to resources required by
the kubelet component, including read access to secrets, and write access to pods.

The psp Application Configuration

The full YAML configuration for the psp application follows:

```yaml
# privileged psp to be used for kube system services only
apiVersion: policy/v1beta1
kind: PodSecurityPolicy
metadata:
  name: privileged
annotations:
  seccomp.security.alpha.kubernetes.io/allowedProfileNames: '*/'
spec:
  privileged: true
  allowPrivilegeEscalation: false
  allowedCapabilities: ['*']
  volumes: ['*']
  hostNetwork: true
  hostPorts:
    - min: 0
      max: 65535
  hostIPC: true
  hostPID: true
  runAsUser:
    rule: 'RunAsAny'
```
4.14 Pod Security Policies

seLinux:
  rule: 'RunAsAny'
supplementalGroups:
  rule: 'RunAsAny'
fsGroup:
  rule: 'RunAsAny'

---

# cluster role privileged psp
kind: ClusterRole
apiVersion: rbac.authorization.k8s.io/v1
metadata:
  name: privileged-psp
rules:
  - apiGroups: ["policy"]
    resources: ["podsecuritypolicies"]
    verbs: ["use"]
    resourceNames: ["privileged"]
  - apiGroups: ["extensions"]
    resources: ["podsecuritypolicies"]
    verbs: ["use"]
    resourceNames: ["privileged"]

---

# role binding for privileged psp to system:nodes
kind: RoleBinding
apiVersion: rbac.authorization.k8s.io/v1
metadata:
  name: privileged-psp-nodes
  namespace: kube-system
roleRef:
  kind: ClusterRole
  name: privileged-psp
  apiGroup: rbac.authorization.k8s.io
subjects:
  - kind: Group
    name: system:nodes
    apiGroup: rbac.authorization.k8s.io
    namespace: kube-system

4.14.5 The psp_system Application:
The purpose of this application is to also assign the more privileged PodSecurityPolicy created in the psp application to all system applications. To be more specific, this is done by binding it to all the ServiceAccounts and appropriate Namespaces for those services.

In cmsh the binding configuration can be viewed with:

```
[cluster->kubernetes[default]->appgroups[system]->applications[psp_system]] \% get config | grep -E "^kind|" name" \| grep kind -A 2

kind: RoleBinding
  name: privileged-psp-calico-kube-controllers
  namespace: kube-system

--

kind: RoleBinding
  name: privileged-psp-nginx-ingress-serviceaccount
  namespace: ingress-nginx

--

kind: RoleBinding
  name: privileged-psp-metrics-server
```
namespace: kube-system

--

kind: RoleBinding
name: privileged-psp-coredns
namespace: kube-system

kind: RoleBinding
name: privileged-psp-kubernetes-dashboard
namespace: kubernetes-dashboard

kind: RoleBinding
name: privileged-psp-calico-node
namespace: kube-system

This `psp_system` is generated by reading all the existing system add-ons from cmsh, and binding the ClusterRole/privileged-psp to all ServiceAccounts used by those services with a RoleBinding/privileged-psp-<serviceaccount> for each of them.

The `psp_system` Application Configuration

The YAML configuration for the `psp_system` application can be seen with:

```yaml
[cluster->kubernetes[default]->appgroups[system]->applications[psp_system]]% get config
kind: RoleBinding
apiVersion: rbac.authorization.k8s.io/v1
metadata:
  name: privileged-psp-calico-kube-controllers
  namespace: kube-system
roleRef:
  kind: ClusterRole
  name: privileged-psp
  apiGroup: rbac.authorization.k8s.io
subjects:
- kind: ServiceAccount
  name: calico-kube-controllers
  namespace: kube-system
```

The display in the preceding session is truncated, but it is followed by a long list of similar blocks for each of the service accounts used by Calico, DNS, metrics server, device plugins, and so on.

The role has been assigned specifically to ServiceAccounts. One alternative way could be to assign once to the system:serviceaccounts group:

```yaml
kind: RoleBinding
apiVersion: rbac.authorization.k8s.io/v1
metadata:
  name: privileged-psp-system-serviceaccounts
  namespace: kube-system
roleRef:
  kind: ClusterRole
  name: privileged-psp
  apiGroup: rbac.authorization.k8s.io
subjects:
- apiGroup: rbac.authorization.k8s.io
  kind: Group
  name: system:serviceaccounts
```
4.15 Kyverno

Kyverno (https://kyverno.io/) is a policy engine designed for Kubernetes. With Kyverno, policies are managed as Kubernetes resources, and no new language is required to write policies. This allows the use of familiar tools such as kubectl, git, and kustomize to manage policies. Kyverno policies can validate, mutate, and generate Kubernetes resources, as well as ensure OCI image supply chain security.

4.15.1 Kyverno Installation

Kyverno engine and Kyverno policy Helm charts can be installed as a part of cm-kubernetes-setup:

Figure 4.7: Choosing Kyverno installation
The installation adds 2 Helm charts in the namespace 'kyverno':

```
[root@basecm10 ~]# module load kubernetes/
[root@basecm10 ~]# helm list -n kyverno
NAME      NAMESPACE   STATUS  ...  CHART            APP VERSION       
kgverno    kyverno     deployed  ...  kyverno-v2.5.2  v1.7.2             
kyverno-policies kyverno     deployed  ...  kyverno-policies-v2.5.2  v1.7.2
```

If the HA option is chosen, then the replica count value is set to 3.

```
[root@basecm10 ~]# helm get values -n kyverno kyverno
USER-SUPPLIED VALUES:
replicaCount: 3
```

This means that at any given time Kubernetes scheduler tries to run 3 pods at the same time:
<table>
<thead>
<tr>
<th>NAME</th>
<th>READY</th>
<th>STATUS</th>
<th>RESTARTS</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>kyverno-5bfb99b9c9-dmmw</td>
<td>1/1</td>
<td>Running</td>
<td>0</td>
<td>1h</td>
</tr>
<tr>
<td>kyverno-5bfb99b9c9-hgfsc</td>
<td>1/1</td>
<td>Running</td>
<td>0</td>
<td>1h</td>
</tr>
<tr>
<td>kyverno-5bfb99b9c9-n67rv</td>
<td>1/1</td>
<td>Running</td>
<td>0</td>
<td>1h</td>
</tr>
</tbody>
</table>

### 4.15.2 Kyverno Policies

It is also recommended to install Kyverno policies in order to enforce Pod Security Standards [https://kyverno.io/policies/](https://kyverno.io/policies/). NVIDIA Base Command Manager configures Kyverno policies in 'enforce' mode, adding service namespaces as exclusions. The list of namespaces to be excluded from particular policies depend on the selected features during install:

```
[root@basecm10 ~]# helm get values -n kyverno kyverno-policies
USER-SUPPLIED VALUES:
validationFailureAction: enforce
policyExclude:
  disallow-host-namespaces:
    any:
    - resources:
      kinds:
      - Pod
      namespaces:
      - default
      - prometheus
  disallow-host-path:
    any:
    - resources:
      kinds:
      - Pod
      namespaces:
      - default
      - local-path-storage
      - '*-restricted'
      - prometheus
      - kube-system
      - gpu-operator
  disallow-host-ports:
    any:
    - resources:
      kinds:
      - Pod
      namespaces:
      - default
      - prometheus
```

In the preceding output, all namespaces that match the wildcard ‘*-restricted’ are excluded from the policy named ‘disallow-host-path’ ([https://kyverno.io/policies/pod-security/baseline/disallow-host-path/disallow-host-path/](https://kyverno.io/policies/pod-security/baseline/disallow-host-path/disallow-host-path/)). This means that, without additional restrictions, all pods in the user namespaces can mount any host path from an underlying node.

To prevent that Kubernetes Permission Manager creates a Kyverno Cluster Policy for every newly-created user, and restricts the hostPath to only the home directory of the user:

```
[root@basecm10 ~]# kubectl get clusterpolicies.kyverno.io | grep john
john-n730sr0-drop-privs-w-hostpath false enforce true
john-n730sr0-drop-privs-w-hostpath-containers false enforce true
```
The Kubernetes permission manager is a custom operator based on Helm. It helps to manage user and system account permissions, roles, role bindings and pod security policies. The operator itself is packed and distributed as a Helm chart, so it can be installed during Kubernetes cluster creation via the cm-kubernetes-setup TUI. The Helm chart for the operator is located in /cm/shared/apps/kubernetes-permissions-manager/current/helm. The output to the following command shows if it is installed:

```
[root@basecm10 ~]# module load kubernetes/
[root@basecm10 ~]# helm list -n cm
NAME        NAMESPACE  STATUS     ... CHART APP VERSION
local-path-provisioner cm    deployed ... cm-kubernetes-local-path-provisioner-0.0.20 0.0.20
permissions-manager cm    deployed ... cm-kubernetes-permissions-manager-0.0.1 0.0.1
```

The Helm chart of the operator includes custom resource definitions (CRD), and makes it possible for the administrator to manage resources using the kubectl tool:

**Example**

```
[root@basecm10 ~]# cat > permissions.yaml<<EOF
apiversion: charts.brightcomputing.com/v1alpha1
kind: CmKubernetesPermissionUser
metadata:
  labels:
    namespace: cmsupport-restricted
  namespace: cm-permissions
spec:
  allow_all_uids: false
  allowPrivilegeEscalation: false
  allowPrivileged: false
  create_namespace: true
  create_service_account: true
  gids:
    - 1000
  namespace: cmsupport-restricted
  psp_spec_override:
    role: edit
    uid: 1000
  user_paths:
    - /home/cmsupport
  username: cmsupport
EOF
[root@basecm10 ~]# kubectl apply -f permissions.yaml
cmkubernetespermissionuser.charts.brightcomputing.com/cmsupport-c7tk7ft created
[root@basecm10 ~]# kubectl get cmkubernetespermissionusers -A
NAMESPACE  NAME     AGE
cm-permissions cmsupport-c7tk7ft 22s
```

At the time of writing of this section (December 2021), the permission manager handles these 4 CRDs:

1. cmkubernetespermissionusers to manage user access to generic resources of the cluster, such as pods, services, secrets, configmaps, etc.
2. `cmkubernetesoperatorpermissionsjupyterkernels` to manage access to the Jupyter Kernels.

3. `cmkubernetesoperatorpermissionspostgresqls` to manage access to the Zalando PostreSQL operator (https://github.com/zalando/postgres-operator).


Providing access to third party operators is necessary if pod security policy is enabled. This is because, by default, not only does the user have no access to CRDs, but also the service accounts of the third party operators have no access to the user namespace.

The following example is a YAML document that provides access to the Jupyter Kernel Operator:

```yaml
apiVersion: charts.brightcomputing.com/v1alpha1
kind: CmKubernetesOperatorPermissionsJupyterKernel
metadata:
  labels:
    namespace: cmsupport-restricted
    username: cmsupport
  name: cmsupport-unz4wlf
  namespace: cm-permissions
spec:
  namespace: cmsupport-restricted
  username: cmsupport
  namespace: cm-permissions
  username: cmsupport
```

Every installed CRD document triggers the Kubernetes permission operator to create a corresponding Helm chart:

```
# helm get values -n cm-permissions cmsupport-unz4wlf
USER-SUPPLIED VALUES:
  namespace: cmsupport-restricted
  username: cmsupport

# helm get manifest -n cm-permissions cmsupport-unz4wlf
---
# Source: cm-kubernetes-operator-permissions-jupyter-kernel/templates/user-permissions.yaml
# Bind policy to service user
apiVersion: rbac.authorization.k8s.io/v1
kind: ClusterRoleBinding
metadata:
  name: cmsupport-unz4wlf-cmsupport-psp
  labels:
  - helm.sh/chart: cmsupport-unz4wlf
    app.kubernetes.io/name: cm-kubernetes-operator-permissions-jupyter-kernel
    app.kubernetes.io/instance: cmsupport-unz4wlf
    app.kubernetes.io/version: "0.0.1"
    app.kubernetes.io/managed-by: Helm
  subjects:
  - kind: ServiceAccount
    name: default
    namespace: cmsupport-restricted
    roleRef:
      kind: ClusterRole
      name: cmsupport-psp
      apiGroup: rbac.authorization.k8s.io
```

...
It is also possible to customize the resulting Helm chart by specifying additional values to specify a section of the CRD. Available values for the Jupyter kernel can be checked using the following command:

```bash
kubectl exec -it -n cmkpm-system $(kubectl get pods -n cmkpm-system -l control-plane=controller-manager -o name) -c manager -- cat /opt/helm/helm-charts/cm-kubernetes-operator-permissions-jupyter-kernel/values.yaml
```

Similarly, tunables for the generic user permissions of the user are available via:

```bash
kubectl exec -it -n cmkpm-system $(kubectl get pods -n cmkpm-system -l control-plane=controller-manager -o name) -c manager -- cat /opt/helm/helm-charts/cm-kubernetes-permission-user/values.yaml
```

The output should be similar to:

```yaml
username: "" # name of the user
create_service_account: true # whether to create kubernetes serviceaccount for the user
role: edit # user role
user_paths: [] # hostPath user able to mount to pods
uid: -1 # UID to run process inside pods
gids: [-1] # list of the GIDs for process inside pods
namespace: "" # namespace to give user permissions to
create_namespace: true # create or not the namespace
allowPrivilegeEscalation: false
allowPrivileged: false
allow_all_uids: true # allow or not to run process as any user including root
    # if 'true' and process is run not with user's UID, then
    # all hostPath volumes are denied
psp_spec_override: # custom PSP definition for the user
```

## 4.17 Providing Access To External Users

To provide access to users on an external network, the requirements are:

- for `kubectl`, an entry in the company/internal DNS server should resolve the external FQDN to the head node or to one of the nodes where Kubernetes is running;

- for the Kubernetes Dashboard, `dashboard` is a subdomain that must be included as a DNS entry under the external FQDN.

The external FQDN, which is set during the Kubernetes cluster setup, is the first item in the list of trusted domains. This can be retrieved from the Kubernetes cluster entity with `cmsh` as follows:

```
Example

[basecm10->kubernetes[default]]% get trusteddomains
basecm10.example.com
kubernetes
kubernetes.default
kubernetes.default.svc
master
localhost
```

In the preceding example, the FQDN of the cluster is `basecm10.example.com`. The cluster administrator managing their own cluster will have another FQDN, and not this FQDN.

For `kubectl`, the Kubernetes API server proxy port should be open to the external network. The proxy port can be retrieved from the Kubernetes cluster entity as follows:
For the Kubernetes Dashboard, the Ingress Controller HTTPS port should be open to the external network. This port, by default with a value of 30443, can be retrieved from the ingress_controller add-on environment:

Example

```bash
[basecm10->kubernetes[default]]% appgroups
[basecm10->kubernetes[default]->appgroups]% applications system
[basecm10->kubernetes[default]->appgroups[system]->applications]% environment ingress_controller
[basecm10->...applications[ingress_controller]->environment]% list
Name (key) Value Nodes environment
------------------------------- ----------------- ------------------
CM_KUBE_EXTERNAL_FQDN basecm10.example.com yes
CM_KUBE_INGRESS_HTTPS_PORT 30443 yes
CM_KUBE_INGRESS_HTTP_PORT 30080 yes
ingress_controller_label brightcomputing.com/ingress-controller no
replicas 1 no
```

If exposing the Kubernetes API server to the external network is selected during setup with `cm-kubernetes-setup`, then the HTTPS and HTTP ports in the preceding example are opened on the Shorewall service that runs on the head node. Exposure to the external network is enabled by default.

Convention Of Using A Domain Name As A Prefix Label

In the preceding example, the brightcomputing.com prefix that is part of the value for ingress_controller_label is just a label rather than a domain. The reason that prefix is used is that it simply follows the convention of using domain names as labels, such as is done by the Kubernetes community (domain: kubernetes.io) and RHEL OpenShift (domain: openshift.io). The prefix brightcomputing.com could equally well have been the prefix brightaccess instead. However it is probably less confusing now to follow the established convention. So that is what is done here for the label.

Users Can Access The Kubernetes Dashboard

Users can access the Kubernetes Dashboard using `dashboard`. By default, the URL takes the FQDN and the port value along with the `dashboard` subdomain, and has the form:

```
https://dashboard.<CM_KUBE_EXTERNAL_FQDN>:<CM_KUBE_INGRESS_HTTPS_PORT>
```

So, for example, it could be something like:

Example

```
https://dashboard.basecm10.example.com:30443
```

Ingress Configuration For Dashboard In cmsh

The default Ingress rule described earlier can be found as an object within cmsh:

```bash
[basecm10->kubernetes[default]->appgroups[system]->applications[dashboard_ingress]]% get config
apiVersion: networking.k8s.io/v1beta1
kind: Ingress
metadata:
  name: kubernetes-dashboard
namespace: kubernetes-dashboard
```
annnotations:
  kubernetes.io/ingress.class: "nginx"
  nginx.ingress.kubernetes.io/secure-backends: "true"
  nginx.ingress.kubernetes.io/ssl-passthrough: "true"
  nginx.ingress.kubernetes.io/backend-protocol: "HTTPS"

spec:
  rules:
  - host: "dashboard.$CM_KUBE_EXTERNAL_FQDN"
    http:
      paths:
      - path: /n
        backend:
          serviceName: kubernetes-dashboard
          servicePort: 443

Using `kubectl`, the Ingress resource can be found with:

```
bash$ kubectl get ingress -n kubernetes-dashboard
```

<table>
<thead>
<tr>
<th>NAME</th>
<th>HOSTS</th>
<th>ADDRESS</th>
<th>PORTS</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>kubernetes-dashboard</td>
<td>dashboard.cluster1.local</td>
<td>10.150.153.251</td>
<td>80</td>
<td>45h</td>
</tr>
</tbody>
</table>

The official documentation for Ingress, at https://v1-16.docs.kubernetes.io/docs/concepts/services-networking/ingress/, explains it well. Path rewrites without domain names can also be used to set up Ingress with multiple backends (serviceName and servicePort pairs), without having to deal with setting up a DNS.

### Ingress Controller Running On Compute Nodes

For scenarios where the head node is not involved in a Kubernetes setup, BCM does not currently set up any forwarding for the Ingress Controller. BCM does set up an NGINX proxy to expose the Kubernetes API Server in such cases, and accessing the Dashboard can then be done with the `kubectl` proxy approach.

For now a workaround to forward Ingress to a compute node can be achieved with port-forwarding, for example by adding the following line to `/etc/shorewall/rules` in Shorewall (section 7.2 of the *Installation Manual*):

```
Example
DNAT net nat:10.141.0.1:30443 tcp 30443
```

### Using One Ingress Controller For Multiple Kubernetes Clusters

BCM does not offer an out-of-the-box solution for one Ingress Controller with multiple Kubernetes clusters. This configuration can be achieved by configuring software such as NGINX to proxy, based on the domain name to the appropriate backend(s).

### 4.18 Networking Model

Kubernetes expects all pods to have unique IP addresses, which are reachable from within the cluster. This can be implemented in several ways, including adding pod network interfaces to a network bridge created on each host, or by using 3rd party tools to manage pod virtual networks.

Since NVIDIA Base Command Manager version 9.2, the pod network provider is Calico (https://www.projectcalico.org/). Calico uses the Border Gateway Protocol (BGP) to distribute routes for every Kubernetes pod. This allows the Kubernetes cluster to be integrated without the need for overlays (IP-in-IP). Calico is particularly suitable for large Kubernetes deployments on bare metal, or in private clouds. This is because for larger deployments the performance and complexity costs of overlay networks can become significant.
4.19 Kubernetes Monitoring

When cm-kubernetes-setup is run, it configures the following Kubernetes-related health checks:

1. KubernetesChildNode: checks if all the expected agents and services are up and running for active nodes
2. KubernetesComponentsStatus: checks if all the daemons running on a node are healthy
3. KubernetesNodesStatus: checks if Kubernetes nodes have a status of Ready
4. KubernetesPodsStatus: checks if all the pods are in one of these states: Running, Succeeded, or Pending

4.20 Local Path Storage Class

For storage, instead of creating Kubernetes PersistentVolumes every time, a modern and practical way is to use the StorageClass feature.

Further documentation on StorageClass is available at:

- https://kubernetes.io/docs/concepts/storage/persistent-volumes/#storageclasses

As a part of initial installation it is possible to choose a Local Path Storage class to utilize the shared storage mounted on every node of the Kubernetes cluster. Possible options include any POSIX shared filesystems, such as NFS, BeeGFS, LustreFS, etc.

![Figure 4.10: Kubernetes setup TUI session: local storage configuration](image)

During setup, the installation wizard asks for a path for where Kubernetes physical volumes (PV) will be physically located. This path should be located on a shared filesystem accessible from all nodes.

After installation, the storage class can be seen to be available with:

```
# kubectl get storageclasses.storage.k8s.io
```

<table>
<thead>
<tr>
<th>NAME</th>
<th>PROVISIONER</th>
<th>RECLAIMPOLICY</th>
<th>VOLUMEBINDINGMODE</th>
<th>ALLOWVOLUMEEXPANSION</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>local-path (default)</td>
<td>rancher.io/local-path</td>
<td>Delete</td>
<td>Immediate</td>
<td>false</td>
<td>1h</td>
</tr>
</tbody>
</table>

Users of the cluster can then freely create persistent volume claims (PVC) resources and use them in running pods.

4.21 Setup Of A Storage Class For Ceph

Pods running on Kubernetes can use Ceph as a distributed storage system to store data in a persistent way.

This section assumes a working Ceph cluster. Ceph installation for BCM is covered in Chapter 9 of the Administrator Manual.

A new pool kube can be created with a replication factor of 3:

Example
The pods of a given namespace have to have access to the Ceph RBD images created to back the volumes. A kube client can be created with:

Example

```
[root@basecm10 ~]# ceph auth get-or-create client.kube mon 'allow r' osd 'allow rwx pool=kube'
[client.kube]
    key = AQCnOvdZpYewBBAAWv1d7c7/XbEvj7QQ7N0THg==
```

A list of the current users, and their access control can be viewed with (some output elided):

Example

```
[root@basecm10 ~]# ceph auth list
installed auth entries:

    osd.0
        key: AQD9M/dZw6HPNRAAT+X8mG5gRUkjLnQo38j4EA==
        caps: [mon] allow rwx
        caps: [osd] allow *
    osd.1
    ...
    client.admin
        key: AQCnM/dZv6OP2x4xwqY9ADB.VV+6i2Uq/ZNqP6iAA==
        caps: [mds] allow *
        caps: [mgr] allow *
        caps: [mon] allow *
        caps: [osd] allow *
    ...
    client.kube
        key: AQCnOvdZpYewBBAAWv1d7c7/XbEvj7QQ7N0THg==
        caps: [mon] allow r
        caps: [osd] allow rwx pool=kube
```

The kube user must be able to map images. The kube configuration must therefore look similar to the section for client.kube in the preceding example.

A Kubernetes secret must be created in the default namespace, and in every Kubernetes namespace that needs storage, using the Ceph user key:

Example

```
```
4.21 Setup Of A Storage Class For Ceph

[root@basecm10 ~]# kubectl create secret generic ceph-secret-user
  --from-literal=userKey=$(ceph auth get-key client.kube) --from-literal=userID=kube
  --namespace=default
secret "ceph-secret-user" created

The Ceph cluster ID can be retrieved by running Ceph's fsid command. The fsid command was given its name because it originally retrieved a file system ID. However, Ceph has evolved since that time, and a file system is not required for the ID to be retrieved for Ceph storage. The ID is now a general ID for the Ceph storage system:

Example

[root@basecm10 ~]# ceph fsid
fe513405-53f6-40a0-8ccb-651e1935c5e4

Ceph monitor <IP address>:<port> values can be found by running ceph mon stat:

Example

[root@basecm10 ~]# ceph mon stat
e1: 3 mons at {node001=10.141.0.1:6789/0,node002=10.141.0.2:6789/0,node003=10.141.0.3:6789/0},
  election epoch 38, quorum 0,1,2 node001,node002,node003

The Helm repository ceph-csi has the container storage interface drivers for Ceph. It can be added with:

Example

[root@basecm10 ~]# helm repo add ceph-csi https://ceph.github.io/csi-charts
"ceph-csi" has been added to your repositories

A request can be made to add a chart ceph-csi-rbd for a RADOS block device image that provides default information for the pods to come up:

Example

[root@basecm10 ~]# helm install --namespace "ceph-csi-rbd" "ceph-csi-rbd" ceph-csi/ceph-csi-rbd
  --create-namespace --set nodeplugin.httpMetrics.containerPort=8082
  --set provisioner.httpMetrics.containerPort=8082

The configmap has to be adjusted to add the Ceph cluster that is used for the storageclass in config.json. The cluster ID, as well as the Ceph monitor IP addresses and ports match the Ceph information from earlier on.

Example

# kubectl describe configmap -n ceph-csi-rbd ceph-csi-config
Name: ceph-csi-config
Namespace: ceph-csi-rbd
Labels:
  app=ceph-csi-rbd
  app.kubernetes.io/managed-by=Helm
  chart=ceph-csi-rbd-3.9.0
  component=nodeplugin
  heritage=Helm
  release=ceph-csi-rbd
Annotations:
  meta.helm.sh/release-name: ceph-csi-rbd
  meta.helm.sh/release-namespace: ceph-csi-rbd
At this point, a storage-class.yml file can then be created, similar to:

Example

```yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: fast
parameters:
  clusterID: fe513405-53f6-40a0-8ccb-651e1935c5e4
  imageFeatures: layering
  imageFormat: "2"
  pool: kube
  csi.storage.k8s.io/provisioner-secret-name: ceph-secret-user
  csi.storage.k8s.io/provisioner-secret-namespace: default
  csi.storage.k8s.io/provisioner: rbd.csi.ceph.com
  reclaimPolicy: Delete
  volumeBindingMode: Immediate
```

In the preceding YAML file, `clusterID` and `pool` must be set to the appropriate values.

Details about the `StorageClass` parameters can be found at: https://github.com/ceph/ceph-csi/blob/devel/examples/rbd/storageclass.yaml

The Kubernetes storage class for Ceph RBD can now be created:

Example

```
[root@basecm10 ~]# kubectl apply -f storage-class.yml
storageclass.storage.k8s.io/fast created
```

To verify that it has been created, the new `StorageClass` can be listed with:

Example

```
[root@basecm10 ~]# kubectl get sc
```

<table>
<thead>
<tr>
<th>NAME</th>
<th>PROVISIONER</th>
<th>RECLAIMPOLICY</th>
<th>VOLUMEBINDINGMODE</th>
<th>ALLOWVOLUMEEXPANSION</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>fast</td>
<td>rbd.csi.ceph.com</td>
<td>Delete</td>
<td>Immediate</td>
<td>false</td>
<td>33s</td>
</tr>
</tbody>
</table>
4.22 Integration With Harbor

In order to spawn pods that use images from the Harbor registry, a secret must first be created with the credentials:

```
[root@basecm10 ~]# kubectl create secret docker-registry myregistrykey \
--docker-server=node001:9443 --docker-username=admin --docker-password=Harbor12345
```

The secret must then be referenced from the pod:

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: foo
spec:
  containers:
    - name: foo
      image: node001:9443/library/nginx
      imagePullSecrets:
        - name: myregistrykey
```

Further information on this is available at https://kubernetes.io/docs/concepts/containers/images/#specifying-imagepullsecrets-on-a-pod.

4.23 Kubernetes Upgrades

Upgrades to Kubernetes can be done by following the official upstream documentation at https://kubernetes.io/docs/tasks/administer-cluster/kubeadm/kubeadm-upgrade/.

There are a few things to keep in mind when upgrading a BCM Kubernetes cluster:

- Nodes that are part of the cluster may be provisioned through their software image.
- The Kubernetes version in NVIDIA Base Command Manager must be updated to match the version after the update.

A list of supported available versions can be retrieved with:

```
Example
```

```
cm-kubernetes-setup --list-versions
1.28
1.27 (NVIDIA AI Enterprise certified)
1.26 (NVIDIA AI Enterprise certified)
1.25 (NVIDIA AI Enterprise certified)
1.24 (NVIDIA AI Enterprise certified)
```

In the example that follows, a BCM cluster is upgraded from Kubernetes 1.25.11 to Kubernetes 1.26.6 to illustrate the process.

The URL for the upstream documentation redirects the readers to the upgrade instructions for 1.25 to 1.26. At the time of writing (July 2023) these are at https://v1-26.docs.kubernetes.io/docs/tasks/administer-cluster/kubeadm/kubeadm-upgrade/.

4.23.1 Upgrade Prerequisites

Kubernetes is not a single entity for upgrade purposes. There is a single driver, the `kubeadm` program, which plans and executes the upgrade only for itself and a handful of components central to Kubernetes’ functionality. Kubernetes does not have ways to manage its third-party components with `kubeadm`’s upgrade planning and execution. This means that multiple Kubernetes components that are essential
to Kubernetes are not involved in the upgrade process. If these components are neglected during an upgrade, then they can either prevent successful completion of the upgrade, or can themselves fail as a result of the upgrade.

Some noteworthy components that are excluded by upgrades are the Etcd database, CNIs, CSIs, container runtimes, as well as individual pods not managed by higher-level entities such as Deployment, ReplicaSet and so on. The Kubernetes documentation at

https://kubernetes.io/docs/concepts/workloads/pods/disruptions/#pod-disruption-budgets

describes the workings of the eviction process in greater detail.

Kubernetes upgrades assume that various entities, in particular, pods can migrate away from the node being upgraded. This assumption may not be true because of various constraints on pod scheduling. The constraints need to be satisfied or worked around in order for migration to work. Particular care has to be taken when dealing with persistent volumes or physical resources such as GPUs that cannot migrate, but are needed for a pod to be deployed.

**Hot And Cold Upgrades**

- A hot upgrade is one that allows the system to stay online at all times.
- A cold upgrade is one that requires at least a brief service interruption.

Since Kubernetes is aimed at providing 100% system uptime, the upgrade process is designed to be hot. The cluster administrator carrying out the hot upgrade does however need to make an effort for the instances where Kubernetes cannot ensure service uptime.

**Upgrading components that cannot be stopped in a regular manner:** One common problem with hot upgrades is the need to migrate or to stop DaemonSet pods. These are intended to run on all Kubernetes nodes, even those being drained. Kubernetes can be told to ignore DaemonSet pods by using --ignore-daemonsets option. The problem with doing that is that it avoids upgrading those pods, and also typically avoids upgrading the third-party component that created them. To ensure upgrade completion, the administrator has to carry out a separate upgrade action, appropriate for each component ignored in this way.

**Upgrading gradually:** Kubernetes upgrades come without any tools to perform upgrades at scale. The available tools can only upgrade a single node at a time. This means that, especially for larger clusters, the administrator needs to carry out the upgrade gradually, and has to plan and allocate extra capacity to accommodate pods migrated from nodes being upgraded. Special precautions need to be taken when dealing with deployments which require larger replica counts. Administrators may need to perform special procedures. Third-party components in particular may require the simultaneous presence of multiple entities across multiple nodes, in order to be able to upgrade those components without a service interruption.

Finally, it is possible that some components are simply not designed for hot upgrades. In that case, administrators have to remove those components from the system, and install the newer version, with the downtime depending on each individual case.

**4.23.2 Example RHEL8 Cluster**

A cluster with 3 nodes that run the control plane, based on RHEL8, is considered as a reference example. One control plane node is the single head node, and the other two control plane nodes are two regular (compute) nodes. There are four additional worker-only nodes, and all the non-head nodes share the same category and software image.

[root@basecm10 ~]# module load kubernetes/default/1.25.11-0
4.23 Kubernetes Upgrades

The upgrade procedure begins by starting with one of the control plane nodes. Instead of going into the details—the official documentation does a good job—the commands for this particular example cluster are listed.

For the example, the cluster administrator starts with updating the control plane being run by the head node. However, for completeness, the two subsections that follow describe updating the control plane if starting with

- the head node, or
- a compute node

The commands are for RHEL. Instructions for Ubuntu may differ slightly. Both are documented in the upstream documentation.

### Upgrading A Control Plane Node Starting With The Head Node

To upgrade the control plane when starting with the head node, the list of commands is:

```
[basecm10%]# kubectl get nodes
NAME   STATUS   ROLES              AGE       VERSION
node001 Ready  control-plane,master,worker 8m25s  v1.25.11
node002 Ready  control-plane,master,worker 8m25s  v1.25.11
node003 Ready  worker              8m43s  v1.25.11
node004 Ready  worker              8m41s  v1.25.11
node005 Ready  worker              8m42s  v1.25.11
node006 Ready  worker              8m41s  v1.25.11
basecm10 Ready control-plane,master 9m22s  v1.25.11
```

#### Upgrading A Control Plane Node Starting With The Head Node

To upgrade the control plane when starting with the head node, the list of commands is:

```
yum list --showduplicates kubeadm --disableexcludes=kubernetes # we find 1.26.6-0
yum install -y kubeadm-1.26.6-0 --disableexcludes=kubernetes
kubeadm version
kubeadm upgrade plan
kubeadm upgrade apply v1.26.6
kubectl drain $(hostname) --ignore-daemonsets
yum install -y kubelet-1.26.6-0 kubectl-1.26.6-0 --disableexcludes=kubernetes
systemctl daemon-reload
systemctl restart kubelet
kubectl uncordon $(hostname)
```

### Upgrading The Control Plane Node Starting With A Compute Node

This section describes the list of commands for upgrading the control plane when starting with a regular node.

The appropriate software image and related information for the node must first be found. For node001 this can be done with:

```
[root@basecm10 ~]# cmsh
[basecm10]% device use node001
[basecm10->device[node001]]% get softwareimage
default-image (category:default)
[basecm10->device[node001]]% softwareimage
default-image
[basecm10->softwareimage]% use default-image
[basecm10->softwareimage[default-image]]% get path
/cm/images/default-image
```

The software image page is then entered. The kubeadm, kubelet and kubectl packages are upgraded in advance:
The following set of commands then carries out the update:

```bash
export host=node001
cmsh -c "device use $host; imageupdate -w --wait"
ssh $host kubeadm version
ssh $host kubeadm upgrade plan
ssh $host kubeadm upgrade apply v1.26.6
kubectl drain $host --ignore-daemonsets
ssh $host sudo systemctl daemon-reload
ssh $host sudo systemctl restart kubelet
kubectl uncordon $host
```

### 4.23.4 Updating Subsequent Control Plane Nodes

If there is only one control plane node, then this section can be skipped.

However, assuming the control plane managed by the head node has been updated, then the nodes seen by `kubectl` show up with their roles as:

**Example**

```bash
[root@basecm10 ~]# kubectl get nodes
```

<table>
<thead>
<tr>
<th>NAME</th>
<th>STATUS</th>
<th>ROLES</th>
<th>AGE</th>
<th>VERSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>node001</td>
<td>Ready</td>
<td>control-plane,master,worker</td>
<td>26m</td>
<td>v1.25.11</td>
</tr>
<tr>
<td>node002</td>
<td>Ready</td>
<td>control-plane,master,worker</td>
<td>26m</td>
<td>v1.25.11</td>
</tr>
<tr>
<td>node003</td>
<td>Ready</td>
<td>worker</td>
<td>26m</td>
<td>v1.25.11</td>
</tr>
<tr>
<td>node004</td>
<td>Ready</td>
<td>worker</td>
<td>26m</td>
<td>v1.25.11</td>
</tr>
<tr>
<td>node005</td>
<td>Ready</td>
<td>worker</td>
<td>26m</td>
<td>v1.25.11</td>
</tr>
<tr>
<td>node006</td>
<td>Ready</td>
<td>worker</td>
<td>26m</td>
<td>v1.25.11</td>
</tr>
<tr>
<td>basecm10</td>
<td>Ready</td>
<td>control-plane,master</td>
<td>27m</td>
<td>v1.26.6</td>
</tr>
</tbody>
</table>

The other control plane nodes can now be updated. In the reference example case, this is the control plane as run on `node001` and `node002`. The commands in the section on **Upgrading A Control Plane Node Starting With A Compute Node** (page 81) should be followed, but with one small difference. Instead of running:

```
kubeadm upgrade apply v1.26.6
```

the command

```
kubeadm upgrade node
```

is run.

The software image should also be prepared with the new packages. The subsequent control plane nodes can then be updated with the following set of commands:

**Example**

```bash
export host=node001
cmsh -c "device use $host; imageupdate -w --wait"
ssh $host kubeadm version
ssh $host kubeadm upgrade plan
ssh $host kubeadm upgrade node
```
4.23 Kubernetes Upgrades

kubectl drain $host --ignore-daemonsets
ssh $host sudo systemctl daemon-reload
ssh $host sudo systemctl restart kubelet
kubectl uncordon $host

Output after this should now look similar to:

```
[root@basecm10 ~]# kubectl get nodes
NAME       STATUS   ROLES                  AGE    VERSION
node001     Ready    control-plane,master,worker 26m     v1.26.6
node002     Ready    control-plane,master,worker 26m     v1.25.11
node003     Ready    worker                  26m     v1.25.11
node004     Ready    worker                  26m     v1.25.11
node005     Ready    worker                  26m     v1.25.11
node006     Ready    worker                  26m     v1.25.11
basecm10    Ready    control-plane,master    27m     v1.26.6
```

The procedure can be repeated for node002 in the reference example system. After that, the control plane nodes are all in an updated state.

4.23.5 Updating The Worker Nodes

The commands for the worker nodes follow a similar pattern. They are updated one by one, and the software images must also be updated, if needed. Since nodes are drained as part of their update procedure, it is best to do them one by one, or at most in limited batches to avoid doing too many at once.

The set of commands to update a single worker is:

```
export host=node003
cmsh -c "device use $host; imageupdate -w --wait"
ssh $host kubeadm upgrade node
kubectl drain $host --ignore-daemonsets
ssh $host sudo systemctl daemon-reload
ssh $host sudo systemctl restart kubelet
kubectl uncordon $host
```

The `kubectl drain` command might complain for other reasons. The administrator can decide on proceeding further by adding additional flags.

For example:

```
kubectl drain $host -ignore-daemonsets --delete-emptydir-data
```

forces the drain even if some pods have stateful data in `emptyDir` volumes.

Multiple nodes can be updated as follows:

**Example**

```
export hosts=['node00[4-6]'
cmsh -c "device; imageupdate -N $hosts -w --wait"
pdsh -W $hosts kubeadm upgrade node
dehost -w $hosts hostname | xargs -n 1 -I{} kubectl drain {} --ignore-daemonsets --delete-emptydir-data
dehost -w $hosts sudo systemctl daemon-reload
dehost -w $hosts sudo systemctl restart kubelet
dehost -N -w $hosts hostname | xargs -n 1 -I{} kubectl uncordon {}
```

4.23.6 Updating The Status In NVIDIA Base Command Manager

At the time of writing of this section (July 2023), the following change is not yet carried out automatically by BCM. It is however necessary in order to get the correct version of Kubernetes reflected in the module file. The procedure to update the old version can be as follows:
4.23.7 Notes For Ubuntu

The upstream documentation covers Ubuntu upgrades when it diverges from RHEL procedures. As an aid, the following commands for RHEL are followed by the equivalent commands for Ubuntu are listed next:

# RHEL
yum list --showduplicates kubeadm --disableexcludes=kubernetes

sudo yum install -y kubeadm-1.26.6-0 kubelet-1.26.6-0 kubectl-1.26.6-0 --disableexcludes=kubernetes

# Ubuntu
apt-cache madison kubeadm

apt-mark unhold kubeadm kubelet kubectl && \
apt-get update && apt-get install -y kubeadm=1.26.6-00 kubelet=1.26.6-00 kubectl=1.26.6-00 && \
apt-mark hold kubeadm kubelet kubectl

4.23.8 Notes For SLES

The upstream documentation does not provide explicit instructions for distributions other than RHEL and Ubuntu.

However, for SLES, BCM uses the “tarball” approach. The installation of such non-package-manager packages is documented at https://kubernetes.io/docs/setup/production-environment/tools/kubeadm/install-kubeadm/, in the “Without a package manager” section. The commands in that section can be used as a guide to find alternatives for the yum and apt commands used so far.

For example, the following command from the installation instructions documentation can be used to get the binaries for the same Kubernetes version used so far in the reference example: This is the “Without a package manager” equivalent of the previous yum install and apt install commands:

RELEASE="v1.26.6"
ARCH="amd64"
pushd /usr/bin/
  sudo curl -L --remote-name-all \
  https://dl.k8s.io/release/$RELEASE\/bin/linux/$ARCH/{kubeadm,kubelet,kubectl}
  sudo chmod +x {kubeadm,kubelet,kubectl}
popd

4.23.9 Other Approaches

When carrying out the steps described in this chapter, some things that may in practice end up being done differently from what is suggested.

For example: control plane nodes often have different categories and software-images compared to the workers. Perhaps the master nodes do not have GPUs, and need different packages, etc. In that case, multiple software images have to be prepared with new packages.
Earlier on, a process to update nodes one at a time was described. However, when a software image is updated, and multiple nodes are tied to that software image, then those nodes can all be provisioned at the same time. The binaries in the updated software image can therefore also be provisioned to all the nodes using that image.

The problem with this is that it could result in nodes getting the new binaries prematurely if they happen to reboot during the update. If this is an unwanted risk, then it can be avoided in several ways, described next.

**Additional Software Images**

A separate software image can be introduced. Nodes can be moved to the new software image one at a time. A software image can also be overruled at the level of a node.

**Example**

```
[root@basecm10 ~]# cmsh
[basecm10]% device use node001
[basecm10->device[node001]]% get softwareimage
  default-image (category:default)
[basecm10->device[node001]]% set softwareimage my-new-image
[basecm10->device*[node001*]]% commit
[basecm10->device[node001]]% get softwareimage
  my-new-image
[basecm10->device[node001]]% quit
```

The cluster administrator should remember to undo the preceding settings, or should move the new software image to the appropriate category, and then clear the node-level override again. This is so that, after the upgrade, the system is organized as before.

**Using NOSYNC**

The other option is to configure the nodes with NOSYNC for their `Next install mode`. This prevents them from syncing with their software image when rebooting (in the case that they still reboot):

**Example**

```
[root@basecm10 ~]# cmsh
[basecm10]% device use node001
[basecm10->device[node001]]% set nextinstallmode nosync
[basecm10->device*[node001*]]% commit
```

Both these approaches make updating slightly more tedious, but also more straightforward.

During testing by BCM developers, nodes getting their binaries updated prematurely due to an unexpected reboot was not seen to be a significant issue. This is presumably because as long as the first control plane node is updated successfully, and the reboot of the extra node is by accident, there is an interruption anyway. The kubelet simply comes back up with the new version. However this is not the official recommended approach.
Kubernetes add-ons were introduced in NVIDIA Base Command Manager version 8.1, and could be managed in that version as part of the addons submode of the kubernetes mode in cmsh. In NVIDIA Base Command Manager version 8.2 this feature was expanded into the Kubernetes Applications & Groups feature. Kubernetes Applications & Groups, less formally called app groups, can be accessed via the appgroups submode of cmsh:

**Example**

```
root@basecm10 ~# cmsh
[basecm10]% kubernetes
[basecm10->kubernetes[default]]% appgroups
[basecm10->kubernetes[default]->appgroups]% list
Name (key)    Applications
            13 in submode>
[basecm10->kubernetes[default]->appgroups]%
```

The version 8.1, addons mode parameters are now accessed from version 8.2 onward via a default system app group instance. The system instance is accessed in the appgroups submode.

**Example**

```
[basecm10->kubernetes[default]->appgroups]% use system
[basecm10->kubernetes[default]->appgroups[system]]% show
Parameter    Value
-------------------------------- ------------------------------------------------
Name         system
Revision
Enabled      yes
applications 13 in submode>
[basecm10->kubernetes[default]->appgroups[system]]% applications
[basecm10->kubernetes[default]->appgroups[system]->applications]% list
Name (key)    Format   Enabled
------------------- ------ -------
bootstrap     Yaml     yes
calico        Yaml     yes
dashboard     Yaml     yes
dashboard_ingress Yaml     yes
dns           Yaml     yes
flannel       Yaml     no
ingress_controller Yaml     yes
```
A Kubernetes application can span multiple namespaces. A name in appgroups therefore only exists to group logically-related applications. Each application contains a YAML configuration file, which BCM synchronizes to the Kubernetes API.

The default system app group is pre-defined. Other app groups can be created as needed. For example, an app group called monitoring could be created to group applications for running Prometheus, node exporters, and anything else related to exposing or viewing Prometheus metrics.

Toggling the Enable parameter of an app group enables or disables all of its application components in Kubernetes. Finer-grained control is possible within the applications mode level, by toggling the enabled parameter per application component instance. For example, within the calico application component instance:

**Example**

```
[basecm10->kubernetes[default]->appgroups[system]->applications]% use calico
[basecm10->kubernetes[default]->appgroups[system]->applications[calico]]% show
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>calico</td>
</tr>
<tr>
<td>Revision</td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td>Yaml</td>
</tr>
<tr>
<td>Enabled</td>
<td>yes</td>
</tr>
<tr>
<td>Config</td>
<td>&lt;244KiB&gt;</td>
</tr>
<tr>
<td>Environment</td>
<td>&lt;3 in submode&gt;</td>
</tr>
<tr>
<td>Exclude list snippets</td>
<td>&lt;2 in submode&gt;</td>
</tr>
</tbody>
</table>

A large YAML configuration file for each application component instance can be configured via the Config parameter property, using the set option of cmsh. This opens up a text editor and allows the environment variables in the YAML configuration file to be managed.

Exclude list snippets are short exclude lists that can be set up for Kubernetes apps computing within the excludelistsnippets submode. They are used to prevent BCM software image updates from overwriting the provisioned files or directories of the container image that are important to the associated Kubernetes application.

Using exclude list snippets within an excludelistsnippets submode is discussed in detail in section 4.4.1 of the Cloudbursting Manual. Similar to the case of Kubernetes apps images, in cloud computing exclude list snippets are used to prevent overwriting of the provisioned files and directories of cloud images.

Environment entries can be set via the Environment submode. Environment entries are similar to environment variables, and are used to replace variables inside the YAML configuration file. Environment entries can be added to the environment as well, if the Nodes environment value inside the Environment submode is set to yes.

**Example**

```
[basecm10->kubernetes[default]->appgroups[system]->applications[calico]]% environment
[basecm10->kubernetes[default]->appgroups[system]->applications[calico]->environment]% list
```

<table>
<thead>
<tr>
<th>Name (key)</th>
<th>Value</th>
<th>Nodes environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>calico_typha_replicas</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>calico_typha_service</td>
<td>none</td>
<td>no</td>
</tr>
<tr>
<td>head_node_internal_ip</td>
<td>10.141.255.254</td>
<td>no</td>
</tr>
</tbody>
</table>
5.1 Providing Custom Docker Images

Allowing users to work with custom Docker images on the cluster requires adding the user to the `docker` group. This can be carried out with `usermod -aG docker <user>`.
Kubernetes Operators

Kubernetes operators are the modern way to manage Kubernetes cluster applications (https://kubernetes.io/docs/concepts/extend-kubernetes/operator/). It is usually recommended that Kubernetes operators are used instead of the legacy applications.

At the time of writing of this section (July 2023), NVIDIA Base Command Manager provides and packages several operators which are validated to perform basic functionalities on a Kubernetes BCM setup.

- the NVIDIA GPU Operator (section 6.3)
- the Prometheus Stack Operator
- the Prometheus Adapter Operator
- the Run:ai Operator (section 6.4)
- the Jupyter Kernel Operator (section 6.2)
- the Spark Operator (section 6.5)
- the Zalando PostreSQL Operator (https://github.com/zalando/postgres-operator)
- the MPI Operator

6.1 Helm Charts For The BCM Operators

During the initial setup, the installation wizard displays a menu to select which operators are to be installed (figure 6.1).

In the case of BCM operators this results in .deb or .rpm packages being deployed. In the case of cloud operators, this already results in a Helm chart being deployed:
Then, based on the selection, for the BCM operators, the wizard asks which of the installed BCM operators to install (roll out) (figure 6.2):

```
[root@basecm10 ~]# yum install cm-jupyter-kernel-operator -y
[root@basecm10 ~]# helm install cm-jupyter-kernel-operator -n cm --wait \
/cm/shared/apps/jupyter-kernel-operator/current/helm/*.tgz
```

If additional tuning is required then tunable values can be set with a command line similar to the following:

```
[root@basecm10 ~]# helm install cm-jupyter-kernel-operator -n cm --wait \
--values tunables.yaml \
/cm/shared/apps/jupyter-kernel-operator/current/helm/*.tgz
```

Possible values can be displayed as follows:

```
[root@basecm10 ~]# helm show values /cm/shared/apps/jupyter-kernel-operator/current/helm/*.tgz
```

Installed operators can be listed by using the CLI option `--list-operators`:

```
[root@basecm10 ~]# cm-kubernetes-setup --list-operators
...
### stage: kubernetes: Display Available Operators
OPERATOR________________________________: api_available___________________________
cm-jupyter-kernel-operator : 1
postgresql-operator : 1
```
6.2 The Jupyter Kernel Operator

6.2.1 Installing The Jupyter Kernel Operator

The Kubernetes Jupyter Kernel Operator can be installed as a part of the cm-kubernetes-setup procedure (section 4.2.5), which eventually leads to the selection screen displayed in figure 6.1.

The Kubernetes Jupyter Kernel Operator can alternatively be installed later on using the OS package manager and Helm:

```
[root@basecm10 ~]# yum install cm-jupyter-kernel-operator -y
[root@basecm10 ~]# helm install cm-jupyter-kernel-operator /
/cm/shared/apps/jupyter-kernel-operator/current/helm/cm-jupyter-kernel-operator-*.tgz
```

The Jupyter Kernel Operator can be removed with:

```
Example
[root@basecm10 ~]# helm uninstall cm-jupyter-kernel-operator
```

It is recommended to enable the PodSecurityPolicy (PSP, section 4.10.2), for the cluster before allowing a user, for example alice, to create resources in the Kubernetes cluster.

```
Example
[root@basecm10 ~]# cm-kubernetes-setup --psp
[root@basecm10 ~]# cm-kubernetes-setup --add-user alice --operators cm-jupyter-kernel-operator
```

The Kubernetes Jupyter Kernel Operator Helm chart creates a CRD that can be used in the Kubernetes API. To check the availability of the CRD, the following command can be run:

```
Example
[root@basecm10 ~]# module load kubernetes
[root@basecm10 ~]# kubectl get crd | grep jupyterkernels
```

6.2.2 Architecture Overview

The Kubernetes Jupyter Kernel Operator has two main components:

- the operator itself
- the sidecar. This is attached to every user-defined kernel pod, and communicates with Jupyter Enterprise Gateway, acting as a proxy for the kernel process.

The following is an overview of the kernel setup and pod lifecycle when the user runs the Kubernetes Jupyter Kernel Operator:

1. User initiates creating kernel in JupyterLab.
2. JupyterLab delegates this task to Jupyter Enterprise Gateway (JEG).
3. JEG opens a service TCP/IP socket and creates a CRD in Kubernetes specifying this port.
4. KubeApi notifies Jupyter Kernel Operator about the newly created CRD.
5. Jupyter Kernel Operator creates services, configmaps, secrets.

6. Jupyter Kernel Operator creates pod to run Jupyter kernel based on the specification. The sidecar is added to the kernel pod during this step.

7. The sidecar waits for the connection file created by the kernel. Alternatively, it relies on the connection file created by the operator (if requested), as not all kernels create a connection file.

8. The sidecar runs a proxy to forward kernel communications to JEG (stdin, shell, iopub, etc).

9. The sidecar notifies JEG about connection parameters and handles kernel communications.

10. If JEG disappears, or if communication drops, then the sidecar stops. This causes the kernel operator to get a notification via the KubeApi service.

11. The Kubernetes Jupyter Kernel Operator removes the unneeded pod, service, configmap and secrets. It also tries to gather stdout and std error of the kernel pod for debug purposes.

The pod created in step 6 is heavily customized by the kernel operator. For security reasons, running a process inside the pod must be carried out as an unprivileged user.

For the convenience of the Jupyter user, the UID/GID of the process inside the pod should match the UID/GID of the Jupyter user. If that is not the case, then the files created in the container are inaccessible for the Jupyter user.

To achieve matching UID/GIDs, the operator dynamically creates `/etc/passwd` and `/etc/group` files inside the pod and populates them with the data from corresponding templates. At the same time the operator can create a kernel communication file, if requested—some kernels rely on that.
6.2.3 Running Jupyter Kernel Using The Operator

An example of a basic YAML definition for the CMJupyterKernel is:

```
$ cat cmjk.yaml
---
apiVersion: apps.brightcomputing.com/v1
kind: CMJupyterKernel
metadata:
  name: cmjk-test
  namespace: alice-restricted
spec:
  username: alice
  uid: 1001
  gid: 1001
  kernel_id: testtesttest
  homedir: /home/alice
  pod:
    volumes:
      - name: homedir
        hostPath:
          path: /home/alice
          type: DirectoryOrCreate
    containers:
      - name: kernel
        image: jupyter/datascience-notebook
        command:
          - "python"
          - "-m"
          - "ipykernel_launcher"
          - "-f"
          - "'/var/tmp/kernel-parm.json"
        workingDir: /home/alice
        securityContext:
          allowPrivilegeEscalation: false
          privileged: false
          runAsNonRoot: true
          runAsUser: 1001
          runAsGroup: 1001
        volumeMounts:
          - name: homedir
            mountPath: /home/alice
```

This can be submitted, but the operator removes it in approximately 1 minute:

```
$ module load kubernetes
$ kubectl apply -f cmjk.yaml
```

The logs of the operator can be checked for debug purposes:

```
$ module load kubernetes
$ kubectl logs
```
This indicates that the sidecar was stopped because there was no connection from Jupyter Enterprise Gateway to the kernel. This is expected, since the kernel has been run manually, and not using Jupyter. After the sidecar shutdown, the kube-api server notifies the operator, which, in turn, removes objects such as CMJupyterKernel, pods, and services.

### 6.2.4 Jupyter Kernel Operator Tunables

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>kernel_id</td>
<td>Kernel identifier (random UUID) given by Jupyter server</td>
</tr>
<tr>
<td>username</td>
<td>Name of the user</td>
</tr>
<tr>
<td>uid, gid, homedir, usershell</td>
<td>UID, GID, home directory and default shell of the user</td>
</tr>
<tr>
<td>image_os_flavor</td>
<td>Defines template of /etc/passwd and /etc/group files, where uid, gid, homedir, and usershell will be added. Could be one of ubuntu1804, rhel7, rhel8, ubuntu2004, sles12, sles15.</td>
</tr>
<tr>
<td>etc_passwd, etc_group</td>
<td>Custom content of the /etc/passwd or /etc/group, if necessary.</td>
</tr>
<tr>
<td>sidecar_command, sidecar_args</td>
<td>Commands and arguments to run the sidecar. By default empty. Most of the arguments for the sidecar are passed via environment variables (section 6.2.5).</td>
</tr>
<tr>
<td>kernel_connection_file_path</td>
<td>Where to expect to find kernel connection file. Default: /var/tmp/kernel-parm.json</td>
</tr>
</tbody>
</table>

...continues
### 6.2.5 Sidecar Arguments And Environment Variables

**Sidecar Arguments**

A timeout can be set as an argument for the sidecar.

- `--timeout`: Defines how long, in seconds, that the sidecar waits for the Jupyter Enterprise Gateway proxy to connect before shutdown. Default: 60

**Environment Variables**

The following environment variables can be used by the sidecar:

<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMJK_CONNECTION_FILE</td>
<td>Path to find a connection file. The sidecar uses the file to establish a connection to the kernel and to pass data between Jupyter Enterprise Gateway and the kernel. Default: /var/tmp/kernel-param.json.</td>
</tr>
<tr>
<td>CMJK_KERNEL_ID</td>
<td>Unique identifier of the kernel. Usually the UUID in table 6.1.</td>
</tr>
<tr>
<td>CMJK SHELL_PORT</td>
<td>Proxy port to open to forward shell communication. Default: 5001.</td>
</tr>
</tbody>
</table>

...continues
<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMJK_IOPUB_PORT</td>
<td>Proxy port to open to forward iopub communication. Default: 5002.</td>
</tr>
<tr>
<td>CMJK_STDIN_PORT</td>
<td>Proxy port to open to forward stdin communication. Default: 5003.</td>
</tr>
<tr>
<td>CMJK_CONTROL_PORT</td>
<td>Proxy port to open to forward control communication. Default: 5004.</td>
</tr>
<tr>
<td>CMJK_HB_PORT</td>
<td>Proxy port to open to forward heartbeat communication. Default: 5005.</td>
</tr>
<tr>
<td>CMJK_COMM_PORT</td>
<td>Proxy port to open to forward comm communication. Default: 5006.</td>
</tr>
</tbody>
</table>

### 6.2.6 Running Spark-based Kernels In Jupyter Kernel Operator

Jupyter integration for BCM provides a kernel template (`jupyter-eg-kernel-k8s-cmjkop-py-spark`) and a sample container image (`brightcomputing/jupyter-kernel-sample:k8s-spark-py39-2.0.0`) to run Jupyter kernels in a Spark environment. The image can be altered or created from scratch based on the scripts provided in `/cm/shared/examples/jupyter/kubernetes-kernel-image-spark-py39/`.

The Jupyter kernel is not run directly. Instead, the kernel process is run and controlled by the `spark-submit` executable inside the container.

Jupyter Kernel Operator alters the provided image based on the CRD definition.

Spark-specific tunables are `spark_pod_template_path` and `spark_pod_template`. The operator creates a file inside of the Spark driver pod and puts the content of `spark_pod_template` in it. After that, `spark-submit` uses this file, via the `--spark.kubernetes.executor.podTemplateFile` configuration option, to create executor pods.

### 6.2.7 Example: Creating An R Kernel From The Kernel Template

The Jupyter Kernel Operator can be used out-of-the-box to support more kernels.

For example, an R kernel can be added.


The default entry point cannot be used as that would start Jupyter notebook, while the aim for this section is to use the kernel only.

Some exploratory investigation should reveal the command to start the kernel:

The pod can be run interactively in a Jupyter notebook terminal by a user:

```bash
cat /opt/conda/share/jupyter/kernels/ir/kernel.json
"argv": ["R", "--slave", "-e", "IRkernel::main()", "--args", "connection_file"],
"display_name": "R",
```

The kernel specifications can then be investigated:

```bash
jupyter-kernelspec list
```
Based on this information the Jupyter Kernel Operator CRD can be created for the user:

```yaml
cat cmjk-ir.yaml
---
apiVersion: apps.brightcomputing.com/v1
kind: CMJupyterKernel
metadata:
  name: cmjk-test
  namespace: alice-restricted
spec:
  username: alice
  uid: 1001
  gid: 1001
  kernel_id: testtesttest
  homedir: /home/alice
  create_connection_file: true  # R kernel expects connection file be created
pod:
  volumes:
    - name: homedir
      hostPath:
        path: /home/alice
        type: DirectoryOrCreate
  containers:
    - name: kernel
      image: jupyter/r-notebook  # image
      command:
        - "R"
      args:
        - "--slave"
        - "-e"
        - "IRkernel::main()"
        - "--args"
        - "'/var/tmp/kernel-parm.json'"  # we have static connection file
      workingDir: /home/alice
      securityContext:
        allowPrivilegeEscalation: false
        privileged: false
        runAsNonRoot: true
        runAsUser: 1001
        runAsGroup: 1001
      volumeMounts:
        - name: homedir
          mountPath: /home/alice
```

There are several changes from the previous (section 6.2.3) YAML, and from the IR command line:

- **create_connection_file**: true

If this is not specified then the kernel complains with the following message during startup:

```
kernell: cannot open file '/var/tmp/kernel-parm.json': No such file or directory
```

This means that the kernel expected this file to be created before the start.
Another image needs to be used.

The spec file has a fixed path and name, instead of "connection_file" as in kernel.json earlier.

The resulting cmjk-ir.yaml file can be submitted to Kubernetes, but it will be removed by the operator after one minute, as it is not being started from the Jupyter Enterprise Gateway.

The next step is to create a kernel template. The Python kernel can be used as a reference:

cd /cm/shared/apps/jupyter/current

cd lib/python*/site-packages/cm_jupyter_kernel_creator/kerneltemplates

cp -pr jupyter-eg-kernel-k8s-cmjkop-py jupyter-eg-kernel-k8s-cmjkop-r

The files meta.yaml, kernel.json, and templates/cmjk.yaml.j2 need to be changed in order to able to provide the correct image and command:

vim jupyter-eg-kernel-k8s-cmjkop-r/meta.yaml
vim jupyter-eg-kernel-k8s-cmjkop-r/kernel.json.j2
vim templates/cmjk.yaml.j2

The changes that are applied should look similar to the following:

diff -u jupyter-eg-kernel-k8s-cmjkop-py/kernel.json.j2 jupyter-eg-kernel-k8s-cmjkop-r/kernel.json.j2
+++ jupyter-eg-kernel-k8s-cmjkop-r/kernel.json.j2 2022-02-16 12:22:23.610382929 +0100
@@ -15,8 +15,8 @@
    argv: 
    - "R",
    + "--slave", "-e", "IRkernel::main()",
    + "--args", "./var/tmp/kernel-parm.json"
    ]

diff -u jupyter-eg-kernel-k8s-cmjkop-py/meta.yaml jupyter-eg-kernel-k8s-cmjkop-r/meta.yaml
+++ jupyter-eg-kernel-k8s-cmjkop-r/meta.yaml 2022-02-16 12:20:57.500974886 +0100
@@ -1,5 +1,5 @@
    display_name: "Python on Kubernetes Operator"
    +display_name: "R on Kubernetes Operator"
    features: "k8s-jupyter-operator-enabled"
    +features: "k8s-jupyter-operator-enabled"
    parameters:
    - display_name:
      +display_name:
      definition:
      getter: shell
After instantiating a kernel spec from the template, the R kernel is ready for use:
Figure 6.3: Jupyter Kernel Creator: creating the IR kernel spec
### 6.2.8 Example: Letting Kubernetes Access Private Registries From The Kernel Template

To be able to pull images from private registries, Kubernetes needs to be instructed about the credentials to use.

**Creating The Secret**

This can be achieved by specifying the secret to `spec.imagePullSecrets` of the pod definition.

For Jupyter Kernel Operator this is `spec.pod.imagePullSecrets`:

**Example**

```
pod:
  containers:
    - name: kernel
      image: image
      ...
      imagePullSecrets:
        - name: regcred
          ...
```

Creating secrets can be carried out with `kubectl`:

**Example**

```
kubectl create \\--namespace alice-restricted \\
  --docker-server=<your-registry-server> \\
  --docker-username=<your-name> \\
  --docker-password=<your-pword> \\
  --docker-email=<your-email> \\
```
More details about managing certificates can be found in the Kubernetes documentation at: https://kubernetes.io/docs/tasks/configure-pod-container/pull-image-private-registry/

**Parameterizing The Secret**

The name of the secret can be parameterized, so that users are allowed to select from secrets in their namespaces.

![Figure 6.5: Jupyter Kernel Creator: secret selection](image)

For parameterization, the `meta.yaml` and `kernel.json.j2` files must also then be modified:

**Example**

```yaml
# cat meta.yaml
...
parameters:
...
  image_pull_secret_name:
    type: str
    definition:
      getter: static
    default: ""
    display_name: "Name of the secret to pull images"

# cat kernel.json.j2
```
6.2 The Jupyter Kernel Operator

```yaml
... "metadata": {
  "process_proxy": {
    "class_name": "cm_jupyter_kernel_creator.eg_processproxies.k8scmjkop.KubernetesCMJupyterKernelOperator",
    "config": {
      "image_pull_policy": "{{ image_pull_policy[0] }}",
      "namespace": "{{ kubernetes_namespace }}",
      "image_pull_secret_name": "{{ image_pull_secret_name }}",
      "gpu_limit": {{ gpu_limit }}
    }
  }
...

# cat templates/cmjk.yaml.j2
...
  pod:
    ...
    {% if image_pull_secret_name %}
      imagePullSecrets:
      - name: {{ image_pull_secret_name }}
    {% endif %}
    containers:
      - name: kernel
        image: {{ image }}
    ...

6.2.9 Example: Adding The PVC Parameter To The Kernel Template

The PVC that is to be mounted can be set in the template:
The settings in `meta.yaml`, `kernel.json.j2` and `templates/cmjk.yaml.j2` for this are:

**Example**

```yaml
# cat meta.yaml
...
parameters:
...
  pvc_name:
    type: list
    definition:
      getter: shell
      exec:
        - source /etc/profile.d/modules.sh
        - module load kubernetes
        - kubectl get pvc -o jsonpath="{range .items[*]}{.metadata.name}{'\n'}{end}"
      display_name: "PVC to mount"
limits:
  max_len: 1
  min_len: 1
```
6.3 The NVIDIA GPU Operator

6.3.1 Installing The NVIDIA GPU Operator

The NVIDIA GPU operator can be installed as a part of the installation session by the `cm-kubernetes-setup` wizard (section 4.2.5). During the setup session, a checkbox can be checkmarked to install and enable the GPU operator (figure 6.1). Nodes that run DGX OS are also supported by the wizard.

The NVIDIA GPU operator can also be deployed on an existing BCM Kubernetes cluster, as described next.
6.3.2 Installing The NVIDIA GPU Operator On An Existing Cluster

The NVIDIA GPU Operator (https://github.com/NVIDIA/gpu-operator) with BCM must always be deployed using Helm.

**Prerequisites:** If the existing cluster uses the NVIDIA device plugin add-on, even if configured by NVIDIA Base Command Manager, then it may be necessary to disable the add-on. This add-on is now deprecated, and will be removed in a future release.

```
[cluster->kubernetes[default]->appgroups[system]->applications[nvidia]]% set enabled no
[cluster->kubernetes[default]->appgroups[system]->applications[nvidia]]% commit
```

One of the prerequisites for the preceding add-on is that it uses labels to identify the nodes to be managed by the add-on. These labels are unnecessary for the GPU operator, and may be removed:

```
[cluster->kubernetes[default]->labelsets]% remove nvidia
[cluster->kubernetes[default]->labelsets]% commit
```

**Installing The NVIDIA GPU Operator:** A knowledge base article that describes how to prepare software images, and how to deploy the NVIDIA GPU Operator using Helm, can be found at:


The article also covers how to deploy the Prometheus Operator Stack, and the Prometheus Adapter for monitoring GPU usage. Deploying these is optional.

Validation methods are described for each step of the deployment.

- For containerd, Helm installation is carried out by the root user with the following options:

  ```
  helm install --wait -n gpu-operator --create-namespace \
  --version v1.10.1 \ 
  --set driver.enabled=false \ 
  --set operator.defaultRuntime=containerd \ 
  --set toolkit.enabled=true \ 
  --set toolkit.env[0].name=CONTAINERD_CONFIG \ 
  --set toolkit.env[0].value=/cm/local/apps/containerd/var/etc/conf.d/nvidia-cri.toml \ 
  gpu-operator nvidia/gpu-operator
  ```

- For docker, Helm installation is carried out by the root user with the following options:

  ```
  helm install --wait -n gpu-operator --create-namespace \
  --version v1.10.1 \ 
  --set driver.enabled=false \ 
  --set operator.defaultRuntime=docker \ 
  --set toolkit.enabled=true \ 
  gpu-operator nvidia/gpu-operator
  ```

**NVIDIA GPU Operator containerd configuration:** The operator provides the toolkit binaries and containerd configuration (nvidia-cri.toml) on each host where a GPU is auto-detected via a host-mount.

The flag that enables this is `--set toolkit.enabled=true`. The path for the configuration file should be set to: `/cm/local/apps/containerd/var/etc/conf.d/nvidia-cri.toml`, which is where BCM’s cm-containerd package expects to find it.

The operator provides a similar configuration functionality for the CUDA drivers. However this is not used in BCM, and it is disabled with the `--set driver.enabled=false` flag. This is because BCM supports CUDA on more Linux distributions and kernel versions than the NVIDIA GPU Operator does. CUDA drivers are therefore expected to already be present on the relevant nodes that have GPUs.
6.3 The NVIDIA GPU Operator

NVIDIA GPU Operator Docker configuration: This is only relevant for older Kubernetes deployments that are deployed on top of Docker or BCM Docker.

Default paths are used, so nothing particularly special has to be done for the operator to deploy properly.

6.3.3 Removing The NVIDIA GPU Operator

The NVIDIA GPU Operator can be found in the `gpu-operator` namespace inside Helm and Kubernetes.

```
root@basecm10 ~# helm list -n gpu-operator
NAME NAMESPACE REVISION ... STATUS CHART APP VERSION
gpu-operator gpu-operator 1 ... deployed gpu-operator-v1.10.1 v1.10.1
```

A `helm uninstall gpu-operator` command can be used to uninstall the operator.

6.3.4 Validating The NVIDIA GPU Operator

A pragmatic way to validate the NVIDIA GPU Operator is to check if the validator pods can be run. A running status for the pods that are to have a GPU on them can be seen with:

Example

```
root@basecm10 ~# kubectl get pod -n gpu-operator -l app=nvidia-operator-validator -o wide
NAME READY STATUS ... IP NODE NOMINATED NODE READINESS GATES
nvidia-operator-validator-2qvz6 1/1 Running ... 172.29.152.172 node001 <none> <none>
nvidia-operator-validator-xkwwv 1/1 Running ... 172.29.112.154 node002 <none> <none>
```

The preceding shows successfully running pods. The log output should show all validations are successful:

Example

```
root@basecm10 ~# kubectl logs -n gpu-operator -l app=nvidia-operator-validator -c nvidia-operator-validator
all validations are successful
all validations are successful
```

6.3.5 Validating The NVIDIA GPU Operator In Detail

The set of pods associated with the NVIDIA GPU Operator can be examined in more detail. The following shows outputs from a GPU operator deployment that is working correctly:

Example

```
root@basecm10 ~# helm list -n gpu-operator
NAME NAMESPACE ... STATUS CHART APP VERSION
gpu-operator gpu-operator ... deployed gpu-operator-v1.10.1 v1.10.1

root@basecm10 ~# kubectl get all -n gpu-operator -o wide
NAME READY STATUS RESTARTS ... NODE
pod/gpu-feature-discovery-gk892 1/1 Running 0 ... node001
pod/gpu-feature-discovery-rmkvj 1/1 Running 0 ... node002
pod/gpu-operator-79c66dc97-1nc1m 1/1 Running 0 ... basecm10
pod/gpu-operator-node-feature-discovery-master-6c65c99969-cjlpq 1/1 Running 0 ... basecm10
pod/gpu-operator-node-feature-discovery-worker-gxz1 1/1 Running 0 ... node002
pod/gpu-operator-node-feature-discovery-worker-ds5mb 1/1 Running 0 ... basecm10
pod/gpu-operator-node-feature-discovery-worker-jf65c 1/1 Running 0 ... node001
pod/nvidia-container-toolkit-daemonset-ffbk7 1/1 Running 1 (46m ago) ... node002
pod/nvidia-container-toolkit-daemonset-lqfkq 1/1 Running 0 ... node001
pod/nvidia-cuda-validator-pxs9b 0/1 Completed 0 ... node001
```
On this particular example cluster, there are two compute nodes with GPUs, and there is one control plane node without a GPU:

```
root@basecm10 ~# kubectl get nodes
NAME STATUS ROLES AGE VERSION
node001 Ready worker 3h2m v1.24.0
node002 Ready worker 3h2m v1.24.0
basecm10 Ready control-plane.master 3h2m v1.24.0
```

**Feature discovery pods:** Node Feature Discovery (NFD, https://intel.github.io/kubernetes-docs/nfd/index.html) is an add-on that is initiated after the operator is installed. A master pod collects discovery information from the worker pods, and schedules more pods in case GPUs have been detected.

In the preceding GPU operator output,

- the master pod is running on node001 with the name: `gpu-operator-node-feature-discovery-master-6c65c99969-wtzcx`
- the worker pods run on each node. For example, the worker pod for node002 is: `gpu-operator-node-feature-discovery-worker-z4skv`

The output for the pods is not very verbose by default, but if more pods under the nvidia-name-space are scheduled on a node, besides the `gpu-operator-node-feature-discovery-*` pods, then that means that NFD has detected one or more GPUs.

For example, a GPU discovered on node001 results in a scheduling of the following pods on that node:

- container toolkit
- device plugin
- validator

**Container toolkit pods:** For nodes that have GPUs, the NVIDIA container toolkit installation pods are started. Pod logs show exactly what is being installed.

One of the requirements for the NVIDIA container toolkit installation pods is that the driver has to be in working order, or the init container driver-validation will fail. The following is the log from a successful installation:

**Example**

```
root@basecm10 ~# kubectl logs -f -n gpu-operator nvidia-container-toolkit-daemonset-ffbk7
Defaulted container "nvidia-container-toolkit-ctr" out of: nvidia-container-toolkit-ctr, driver-validation (init)
... 
```

```
time="2022-12-06T14:31:36Z" level=info msg="Installing toolkit"
time="2022-12-06T14:31:36Z" level=info msg="Parsing arguments: [/usr/local/nvidia/toolkit]"
```
Device plugin pods: The device plugin pods are started up next. These have the toolkit as a requirement. If the toolkit is not in working order, then the init container toolkit-validation fails. The following is the log from a successful startup:

Example

```
root@basecm10 ~# kubectl logs -f -n gpu-operator nvidia-device-plugin-daemonset-698hd
Defaulted container "nvidia-device-plugin-ctr" out of: nvidia-device-plugin-ctr, toolkit-validation (init)
2022/12/06 14:32:20 Loading NVML
2022/12/06 14:32:20 Starting FS watcher.
2022/12/06 14:32:20 Starting OS watcher.
2022/12/06 14:32:20 Retreiving plugins.
2022/12/06 14:32:20 No MIG devices found. Falling back to mig.strategy=
2022/12/06 14:32:20 Starting GRPC server for 'nvidia.com/gpu'
2022/12/06 14:32:20 Starting to serve 'nvidia.com/gpu' on /var/lib/kubelet/device-plugins/nvidia-gpu.sock
2022/12/06 14:32:20 Registered device plugin for 'nvidia.com/gpu' with Kubelet
```

The pod log output suggests that the GPU is now registered with the Kubelet as a resource. This can be checked by querying the Node resource:

Example

```
root@basecm10 ~# kubectl describe node node001 | grep nvidia
nvidia.com/cuda.driver.major=520
nvidia.com/cuda.driver.minor=61
nvidia.com/cuda.driver.re=05
nvidia.com/cuda.runtime.major=11
nvidia.com/cuda.runtime.minor=8
nvidia.com/gfd.timestamp=1670337142
nvidia.com/gpu.compute.major=7
nvidia.com/gpu.compute.minor=0
nvidia.com/gpu.count=1
nvidia.com/gpu.deploy.container-toolkit=true
nvidia.com/gpu.deploy.dcm=true
nvidia.com/gpu.deploy.dcm-exporter=true
nvidia.com/gpu.deploy.device-plugin=true
nvidia.com/gpu.deploy.driver=true
nvidia.com/gpu.deploy.gpu-feature-discovery=true
nvidia.com/gpu.deploy.node-status-exporter=true
nvidia.com/gpu.deploy.operator-validator=true
nvidia.com/gpu.family=volta
nvidia.com/gpu.machine=OpenStack-Nova
nvidia.com/gpu.memory=32768
nvidia.com/gpu.present=true
```
nvidia.com/gpu.product=Tesla-V100-SXM3-32GB
nvidia.com/mig.strategy=single
nvidia.com/run.ai-swap.enabled=false
nvidia.com/gpu: 1
nvidia.com/gpu: 1
...

**Validator pods:** If anything goes wrong with either the driver, toolkit, CUDA, or the plugin, then validator pods are a good place to start looking.

If all goes well, the main container outputs all validations are successful:

**Example**

```bash
root@basecm10 ~# kubectl logs -f -n gpu-operator nvidia-operator-validator-2qvz6
Defaulted container "nvidia-operator-validator" out of: nvidia-operator-validator, driver-validation (init), toolkit-validation (init), cuda-validation (init), plugin-validation (init)
all validations are successful
```

It is possible for an init container to fail. The output for the container should then be checked.

The following shows output from successful init containers:

```bash
root@basecm10 ~# kubectl logs -f -n gpu-operator nvidia-operator-validator-2qvz6 -c driver-validation
running command chroot with args [/run/nvidia/driver nvidia-smi]
Tue Dec 6 15:32:14 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 520.61.05 Driver Version: 520.61.05 CUDA Version: 11.8 |
+-----------------------------------------------------------------------------+
| GPU Name Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC |
| Fan Temp Perf Pwr:Usage/Cap| Memory-Usage | GPU-Util Compute M. |
|===============================+======================+======================|
| 0 Tesla V100-SXM3... On | 00000000:00:06.0 Off | 0 |
| N/A 32C P0 46W / 350W | 2MiB / 32768MiB | 0% Default |
+-----------------------------------------------------------------------------+
```

```bash
root@basecm10 ~# kubectl logs -f -n gpu-operator nvidia-operator-validator-2qvz6 -c toolkit-validation
Tue Dec 6 14:32:16 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 520.61.05 Driver Version: 520.61.05 CUDA Version: 11.8 |
+-----------------------------------------------------------------------------+
| GPU Name Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC |
| Fan Temp Perf Pwr:Usage/Cap| Memory-Usage | GPU-Util Compute M. |
|===============================+======================+======================|
| 0 Tesla V100-SXM3... On | 00000000:00:06.0 Off | 0 |
| N/A 32C P0 46W / 350W | 2MiB / 32768MiB | 0% Default |
+-----------------------------------------------------------------------------+
```

```bash
root@basecm10 ~# kubectl logs -f -n gpu-operator nvidia-operator-validator-2qvz6 -c plugin-validation
Tue Dec 6 14:32:16 2022
```
6.3 The NVIDIA GPU Operator

+-------------------------------+----------------------+
| Processes:                   |
| GPU GI CI PID Type Process name GPU Memory |
| ID ID Usage                 |
|=============================================================================|
| No running processes found  |
+-------------------------------+----------------------+

root@basecm10 ~# kubectl logs -f -n gpu-operator nvidia-operator-validator-2qvz6 -c cuda-validation

root@basecm10 ~# kubectl logs -f -n gpu-operator nvidia-operator-validator-2qvz6 -c plugin-validation

This also explains where the pods earlier on came from, the ones marked with status Completed. They are used as part of certain validation steps.

Which init container prints out error messages should indicate where the problem lies—either with the CUDA drivers, or the toolkit, and so on. If the driver or toolkit is not validating correctly, then it may result in a lot of pods stuck in a Pending or an Init stage. Looking at what init container is associated with the stuck pod helps in diagnosing the problem.

DCGM exporter pods: These pods expose metrics endpoints for scraping, and can be considered less critical. They are involved in GPU metrics collection, and can be utilized with, for example, Prometheus Stack Operator, or the Prometheus Adapter, for horizontal pod autoscaling based on GPU metrics.

More information on the Prometheus Stack Operator and the Prometheus Adapter Operator can be found at:

6.3.6 Running A GPU Workload

A GPU workload can be run with the following configuration:

Example

root@basecm10 ~# cat << EOF > gpu.yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-pod
spec:
  restartPolicy: Never
  containers:
  - name: cuda-container
    image: nvidia/cuda:9.2-runtime
    command: ["nvidia-smi"]
    resources:
      limits:
        nvidia.com/gpu: 1
114 Kubernetes Operators

```
root@basecm10 ~# kubectl create -f gpu.yaml
pod/gpu-pod created
```

On a cluster with GPUs available, this pod should get scheduled, and should not stay stuck in the Pending phase.

The preceding example just invokes `nvidia-smi` in the container. The output can be viewed to confirm that it worked:

**Example**

```
root@basecm10 ~# kubectl logs -f gpu-pod
```

```
Tue Dec 6 15:08:03 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 520.61.05 Driver Version: 520.61.05 CUDA Version: 11.8 |
|-------------------------------+----------------------+----------------------+
| GPU Name Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC |
| Fan Temp Perf Pwr:Usage/Cap| Memory-Usage | GPU-Util Compute M. |
| | | MIG M. |
|=============================================================================|
| 0 Tesla V100-SXM3... On | 00000000:00:06.0 Off | 0 |
| N/A 34C P0 47W / 350W | 2MiB / 32768MiB | 0% Default |
| | | N/A |
+-----------------------------------------------------------------------------+

```

Processes:

```
<table>
<thead>
<tr>
<th>GPU</th>
<th>GI</th>
<th>CI</th>
<th>PID</th>
<th>Type</th>
<th>Process name</th>
<th>GPU Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>ID</td>
<td>Usage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

No running processes found

```
+-----------------------------------------------------------------------------+
```

6.4 The Run:ai Operator

The Run:ai operator is a GUI application that can make a cluster installer for Run:ai available via the BCM head node landing page (figure 6.7).

The Run:ai documentation documents the cluster installer bundle for NVIDIA DGX at: https://docs.run.ai/latest/admin/runai-setup/cluster-setup/dgx-bundle/.

![Run:ai installer](image)

Figure 6.7: Run:ai installer available from the BCM landing page
6.4 The Run:ai Operator

6.4.1 Installing The Run:ai Operator

The Run:ai operator can be installed as a part of the cm-kubernetes-setup procedure (section 4.2.5). The Helm status can be checked with, for example:

Example

```
root@basecm10 ~# helm list -n runai
NAME NAMESPACE ... STATUS CHART VERSION
cluster-installer runai ... deployed cluster-installer-2.8.8 0.0.1
```

```
root@basecm10 ~# kubectl get all -n runai
NAME READY STATUS RESTARTS AGE
pod/cluster-installer-deployment-5f4c4cbf4c-82gmx 1/1 Running 0 5m9s
```

```
NAME TYPE CLUSTER-IP EXTERNAL-IP PORT(S) AGE
service/cluster-installer-service ClusterIP 10.150.117.247 <none> 8080/TCP 5m9s
```

```
NAME READY UP-TO-DATE AVAILABLE AGE
deployment.apps/cluster-installer-deployment 1 1 1 5m9s
```

```
NAME DESIRED CURRENT READY AGE
replicaset.apps/cluster-installer-deployment-5f4c4cbf4c 1 1 1 5m9s
```

6.4.2 Removing The Run:ai Operator

The Run:ai operator can be removed via Helm:

Example

```
[root@basecm10 ~]# helm uninstall cluster-installer -n runai
```

Removal of Run:ai access from the BCM head node landing page (figure 6.7) can be carried with a removal of the associated JSON file:

Example

```
root@basecm10 ~# ls -al /var/www/html/kubernetes/runai/
total 4
drwxr-xr-x 4 root root 51 Dec 1 16:24 ..
-rw-r--r-- 1 root root 317 Dec 6 12:21 default.json
root@basecm10 ~# rm -rf /var/www/html/kubernetes/runai/default.json
```

Each Kubernetes cluster has its own JSON file.

Uninstalling the Kubernetes cluster automatically cleans up everything associated with it.

6.4.3 Completing The Run:ai Installation

Right after cm-kubernetes-setup is run, a summary is shown at the end of the installation. This includes URLs that point to Run:ai installer:

Example

```
[root@basecm10 ~]# cm-kubernetes-setup
... 
#### stage: kubernetes: Print Summary
Installation completed. Pods might still be initializing.
```
To add users to the cluster use: refer to `cm-kubernetes-setup --help`

To use kubectl load the module file: kubernetes/default/1.24

Common URLs:
- Kubernetes API server: https://rb-runai.openstacklocal:10443
- Kubernetes dashboard: https://dashboard.rb-runai.openstacklocal:30443/
- Kubernetes dashboard: https://0.0.0.0:30443/dashboard/
- Run:ai installer: http://rb-runai.openstacklocal:30080/runai-installer
- Run:ai installer: https://rb-runai.openstacklocal/#runai

## Progress: 100

Took: 13:18 min.
Progress: 100/100

Finished execution for 'Kubernetes Setup', status: completed

Kubernetes Setup finished!

The Run:ai installer presents a wizard first, which checks for dependencies (figure 6.8).

![Run:ai installer: verification screen](image)

Figure 6.8: Run:ai installer: verification screen

On the next screen, the user is asked for credentials, URLs, and certificates:
6.4 The Run:ai Operator

The Run:ai Operator

Figure 6.9: Run:ai installer: fields to be filled by the user

- The Tenant name and Application secret key are provided by Run:ai.
- The Cluster URL has to be the FQDN of the Kubernetes cluster. For example: `my-kubernetes-cluster-fqdn.com:30443` This is the FQDN that the wizard prompts for in figure 4.3.
- The Private key and Certificate files are the admin key and PEM files from the directory of the user on the cluster. By default:
  - for root these are
    * `/root/.kube/admin-default.key`
    * `/root/.kube/admin-default.pem`
  - for a regular user, such as john, these are
    * `/home/john/.kube/admin-default.key`
    * `/home/john/.kube/admin-default.pem`

The wizard finalizes the Run:ai setup, and then carries out the installation. A progress meter is displayed during installation:
118 Kubernetes Operators

On completion, the wizard redirects the user to the Run:ai dashboard. The NVIDIA Base Command Manager head node landing page is also updated to point to the Run:ai dashboard, with the “+” sign of figure 6.7 now showing a link icon instead, indicating that Run:ai is now installed:

6.4.4 Post-installation

The Run:ai documentation includes post-installation steps in the following URL: https://docs.run.ai/latest/admin/runai-setup/cluster-setup/dgx-bundle/#post-installation.

The most important aspect is configuring Researcher Access Control in BCM: https://docs.run.ai/latest/admin/runai-setup/authentication/researcher-authentication/.

This includes going to cmsh and configuring the Kubernetes API server with additional OIDC pa-
The `runai` binary can be downloaded in various ways, the Runai environment has an option where the binary can be downloaded from the cluster itself. The binary can be copied to `/usr/bin` and made executable by the system administrator.

Figure 6.12: `runai` binary download

### 6.5 Kubernetes Spark Operator

Using the Kubernetes Spark Operator is a simpler alternative to using the `spark-submit` tool for job submission.

#### 6.5.1 Installing The Kubernetes Spark Operator

The Kubernetes Spark Operator can be installed as a part of the `cm-kubernetes-setup` procedure (section 4.2.5), which eventually leads to a display listing the operator packages that may be installed (figure 6.1).

The Kubernetes Spark Operator can alternatively be installed later on using the OS package manager and Helm:

```bash
[root@basecm10 ~]# yum install cm-kubernetes-spark-operator -y
[root@basecm10 ~]# helm install cm-kubernetes-spark-operator /cm/shared/apps/kubernetes-spark-operator/current/helm/spark-operator-*.tgz
```

The Kubernetes Spark Operator can be removed with:

**Example**

```bash
[root@basecm10 ~]# helm uninstall cm-kubernetes-spark-operator
```

The operator installation state can be verified with `--list-operators`:

**Example**

```bash
[root@basecm10 ~]# cm-kubernetes-setup --list-operators
```

```
OPERATOR________________________________: api_available___________________________
cm-jupyter-kernel-operator : 0
cm-kubernetes-postgresql-operator : 0
cm-kubernetes-spark-operator : 1
```

The Helm status can be checked with, for example:

**Example**

```bash
[root@basecm10 ~]# cm-kubernetes-setup
```
The Permission Manager (section 4.16) and PodSecurityPolicy (PSP, section 4.10.2) must both be enabled for the cluster, before allowing a user to create resources in the Kubernetes cluster in their namespace:

Example

```
[root@basecm10 ~]# cm-kubernetes-set up --psp
```

The user alice can be allowed to use the Spark operator, and allowed to run a process as any UID in the pod:

Example

```
[root@basecm10 ~]# cm-kubernetes-set up --add-user alice --operators cm-kubernetes-spark-operator \
--allow-all-uids
```

The Kubernetes Spark operator Helm chart creates a CRD that can be used in the Kubernetes API. For Alice, the CRD is available and can be used with a Spark operator YAML, to build a Spark application carry out a pi run in the restricted namespace.

6.5.2 Example Spark Operator Run: Calculating Pi

Continuing on with the user alice of the preceding section, a YAML file based on the specification at https://github.com/GoogleCloudPlatform/spark-on-k8s-operator/blob/master/examples/spark-py-pi.yaml can be used:

Example

```
[root@basecm10 ~]# su - alice

[alice@basecm10 ~]$ module load kubernetes

[alice@basecm10 ~]$ cat <<EOF > pi-spark.yaml
apiVersion: "sparkoperator.k8s.io/v1beta2"
kinds: SparkApplication
metadata:
  name: pyspark-pi
spec:
  type: Python
  pythonVersion: "3"
  mode: cluster
  image: "gcr.io/spark-operator/spark-py:v3.1.1"
  imagePullPolicy: Always
  mainApplicationFile: local:///opt/spark/examples/src/main/python/pi.py
  sparkVersion: "3.1.1"
  restartPolicy:
    type: OnFailure
    onFailureRetries: 3
EOF
```
onFailureRetryInterval: 10
onSubmissionFailureRetries: 5
onSubmissionFailureRetryInterval: 20
driver:
  cores: 1
coreLimit: "1200m"
memory: "512m"
labels:
  version: 3.1.1
  serviceAccount: spark
executor:
  cores: 1
  instances: 1
  memory: "512m"
  labels:
    version: 3.1.1
EOF

[alice@basecm10 ~]$ kubectl apply -f pi-spark.yaml
sparkapplication.sparkoperator.k8s.io/pyspark-pi created
[alice@basecm10 ~]$ kubectl get pods
NAME           READY STATUS    RESTARTS AGE
pyspark-pi-driver 0/1 ContainerCreating 0 1s
[alice@basecm10 ~]$ kubectl get pods
NAME           READY STATUS    RESTARTS AGE
pyspark-pi-driver 1/1 Running   0 3s
[alice@basecm10 ~]$ kubectl get sparkapplications
NAME    AGE
pyspark-pi 7s
[alice@basecm10 ~]$ kubectl get pods
NAME               READY STATUS      RESTARTS AGE
pythonpi-e768128383a881b3-exec-1 0/1 ContainerCreating 0 0s
[alice@basecm10 ~]$ kubectl get pods
NAME               READY STATUS      RESTARTS AGE
pyspark-pi-driver 0/1 Completed 0 34s
pythonpi-e768128383a881b3-exec-1 0/1 Terminating 0 20s
[alice@basecm10 ~]$ kubectl get pods
NAME               READY STATUS      RESTARTS AGE
pyspark-pi-driver 0/1 Completed 0 36s

Instead of tracking the pod with:
kubectl get pods
as in the preceding session, or with the more convenient:
watch kubectl get pods
the pod could be tracked with the -f|--follow option to stream the driver logs:

Example

[alice@basecm10 ~]$ kubectl logs pyspark-pi-driver -f

To get intended output of the pi run—the calculated value of pi—it is sufficient to grep the log as follows:

Example

[alice@basecm10 ~]$ kubectl logs pyspark-pi-driver | grep 'Pi'
Pi is roughly 3.148800
After the pi run has completed, the resources can be removed from the namespace:

```
[alice@basecm10 ~]$ kubectl delete -f pi-spark.yaml
sparkapplication.sparkoperator.k8s.io "pyspark-pi" deleted

[alice@basecm10 ~]$ kubectl get pods
No resources found in alice-restricted namespace.

[alice@basecm10 ~]$ kubectl get sparkapplications
No resources found in alice-restricted namespace.
```
Kubernetes On Edge

How edge sites can be configured is described in Chapter 2 of the *Edge Manual*. If there are BCM Edge sites configured in the cluster, then the Kubernetes setup prompts the user with edge sites that Kubernetes can be deployed on.

![Figure 7.1: cm-kubernetes-setup prompting for edge sites.](image)

If an edge site is selected, then the rest of the wizard prompts only for nodes available within that edge site; prompts only for the associated network interfaces; and so on.

### 7.1 Flags For Edge Installation

Edge directors often lack high-bandwidth connectivity to the central head node, or they often may benefit from coming up as quickly as possible. It can therefore sometimes be useful to skip stages of the setup.

Running `cm-kubernetes-setup --help` displays some additional flags that allow some setup stages, that bring up a cloud director, to be skipped explicitly:

```
cm-kubernetes-setup --help
...
installing Kubernetes clusters:
  Flags for installing or managing Kubernetes clusters

  --skip-package-install  Skip the package installation steps. Ignores skip_packages flags in the config.
  --skip-reboot          Skip the reboot steps.
  --skip-image-update    Skip the image update steps.
  --skip-disksetup-changes Never change the disk-setup. Use this flag if you manually configure a partition or device for docker thin pool devices for example.
```
7.1.1 Speeding Up Kubernetes Installation To Edge Nodes With The --skip-* Flags: Use Cases

Explanations and use cases for these flags are given in the following table:

<table>
<thead>
<tr>
<th>Flag</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>--skip-package-install</td>
<td>all edge directors share the same software image, and the image is already up to date. So the installer does not need to install packages from that image to the edge director.</td>
</tr>
<tr>
<td>--skip-image-update and --skip-reboot</td>
<td>all edge directors are already provisioned with the up-to-date software image. So the installer does not need to carry out an update from the ISO or head node, and then reboot the edge director.</td>
</tr>
<tr>
<td>--skip-disksetup-changes</td>
<td>all edge directors already have the correct disk layout. This flag can be set if the disk layout was already configured up-front, in order to avoid full provisioning.</td>
</tr>
</tbody>
</table>

These flags can also be configured in the YAML configuration file of the cm-kubernetes-setup wizard.

The flags can be used for scripted installations for quick Kubernetes setups. For a scripted installation of an edge director, preparations can be done beforehand so that all the requirements in the software images that the edge directors use are already installed, the right disk layouts are already configured, and packages are already updated.

All the stages in the flag options can then be skipped for installing onto edge sites. This can make the setup take just a few seconds per Kubernetes deployment.
The Kubernetes Cluster API (CAPI), as explained in the introduction to the online Cluster API Book at https://cluster-api.sigs.k8s.io/, "is a Kubernetes sub-project focused on providing declarative APIs and tooling to simplify provisioning, upgrading, and operating multiple Kubernetes clusters". The Cluster API Book is the official Kubernetes project documentation for CAPI.

This chapter describes the installation and usage of the NVIDIA Base Command Manager CAPI extension called BCM Kubernetes CAPI Infrastructure Provider.

8.1 Kubernetes Cluster API Components

An overview of the CAPI components is shown in figure 8.1. Further details about the components are given in the sections of this chapter that follow.

Figure 8.1: CAPI components

Figure 8.1 shows a standard Kubernetes cluster deployed on an active head node. The cluster has been modified to become a Kubernetes management cluster using the BCM CAPI Infrastructure Provider, and has successfully deployed two additional Kubernetes clusters through CAPI. These additional clusters may be running different versions of Kubernetes.

The system administrator receives module files and kubeconfig files for all three Kubernetes clusters.

8.1.1 Kubernetes Management Cluster

In BCM documentation, the term Kubernetes management cluster is used specifically and precisely to refer to a Kubernetes cluster that operates as the management cluster for CAPI.

The ability to modify an external Kubernetes cluster to operate as a Kubernetes management cluster is under preparation at the time of writing (June 2023) of this section. This includes scenarios such as using an existing Kubernetes cluster, hosted on a public cloud, to serve as the Kubernetes management
cluster for a Kubernetes cluster deployed through BCM. At present, any Kubernetes cluster deployed through BCM (Chapter 4) can be modified into a CAPI management cluster.

8.1.2 Kubernetes CAPI Cluster
BCM documentation uses the term Kubernetes CAPI cluster to refer specifically to a Kubernetes cluster deployed via the Cluster API. Upstream Kubernetes documentation also sometimes refers to a Kubernetes cluster deployed via CAPI as a workload cluster.

8.1.3 BCM CAPI Infrastructure Provider
The BCM CAPI Infrastructure Provider is a derivative of the Bring Your Own Host (BYOH) CAPI provider. The BYOH CAPI provider can also be referred to as the BYOH CAPI Infrastructure Provider, and is available as a GitHub project at:
https://github.com/vmware-tanzu/cluster-api-provider-bringyourownhost

Similar to other CAPI providers, the BCM CAPI Infrastructure Provider is an extension of CAPI itself, and both the CAPI provider and CAPI are required on the Kubernetes management cluster.

The BCM CAPI Operator
The BCM CAPI operator is an operator deployed in the byoh-system namespace on the Kubernetes management cluster. The operator serves as the central point of connection for BCM CAPI host agents (section 8.1.3). It also monitors the state of clusters and machines for the Kubernetes CAPI clusters.

```
root@headnode:~# kubectl get pod -n byoh-system
NAME                   READY STATUS    RESTARTS AGE
byoh-controller-manager-6c98cb44c-7gdbx  2/2   Running 2 (5h55m ago) 23h
```

BCM CAPI Host Agents
BCM CAPI host agents are scheduled to run on selected hosts. These hosts are defined by the cluster administrator when the cm-kubernetes-capi-setup wizard (section 8.2.2) is run. The wizard ensures that a selected host gets the right packages in its software image, and that the host is assigned the right BCM roles.

For CAPI calls to function, the CapiRole is assigned. Assignment can be done either before or after creating a Kubernetes CAPI cluster via the Kubernetes management cluster. The role sets up a capi-agent service on the host for the associated Kubernetes management cluster.

When the role is:
• assigned: BCM initiates and bootstraps the service
• unassigned: BCM halts the service and makes it unavailable for use by CAPI on the associated Kubernetes management cluster

This differs from cloud-based CAPI providers, which do not require managing a limited number of pre-existing nodes, but instead start up new nodes on demand as needed.

If a Kubernetes CAPI cluster receives a cluster request and there are no hosts available, then machines remain in a Pending state until hosts become available. Once hosts become available, the status transitions to Provisioning, and then eventually to Running (section 8.3.1).

BCM CAPI Vs BYOH
Once the actual Kubernetes CAPI clusters are deployed on designated CAPI hosts, BCM uses its Python API to carry out the following operations:

• The requested Kubernetes version for each node is stored in the BCM database.
• The requested Kubernetes version is made available in the corresponding software image for the node.
The node is provisioned with the software image.

The installer is notified that the Kubernetes installation can proceed.

In contrast to the BYOH provider, BCM CAPI has the following extra features:

- Linux distributions can be other than just Ubuntu 20.04.
- Kubernetes versions can be other than only those for which parcels have been created, as the BCM Kubernetes integration does not rely on parcels.

## 8.2 The Kubernetes CAPI Wizard

The `cm-kubernetes-capi-setup` wizard installs the BCM CAPI Infrastructure Provider, and assigns the CAPI role to nodes.

### 8.2.1 The Install CAPI Option

The Install CAPI option of figure 8.2 leads to a screen that prompts the user to select the Kubernetes cluster instance on which to install the BCM CAPI Infrastructure Provider.

If needed:

- `cert-manager` and the Cluster API itself are installed
- the CAPI `clusterctl` tool is run in the back end to carry out initialization

### The Install CAPI Option Actions

Execution of the Install CAPI operation:

- prepares a Kubernetes cluster template for CAPI (section 8.8).
- installs the BCM CAPI Infrastructure Provider on the Kubernetes Cluster with the command: `clusterctl init infrastructure byoh`.
- waits for the BCM CAPI operator to be ready for operation.
• patches the BCM CAPI operator to use the appropriate image. Patching is required until the BCM changes get into the upstream repository.

• makes the capi module available for loading

The **Install CAPI Option Changes To Kubernetes Management Cluster**

After the Install CAPI option has run, the operators installed and running on the Kubernetes management cluster can be seen (some output truncated):

### Example

```
[root@basecm10 ~]# module load kubernetes
[root@basecm10 ~]# kubectl get pod -A | grep -E "byoh-system|capi|cert-manager"
byoh-system byoh-controller-manager-ff7f688bb4-vprrz 2/2 ...
capi-kubeadm-bootstrap-system capi-kubeadm-bootstrap-controller-manager-7945579f8c-4mtb2 1/1 ...
capi-kubeadm-control-plane-system capi-kubeadm-control-plane-controller-manager-66cdefb477b-h5779 1/1 ...
capi-system capi-controller-manager-64cb86f545-8k9zt 1/1 ...
cert-manager cert-manager-5d4c5bc8bc-jshp 1/1 ...
cert-manager cert-manager-cainjector-79b55d9d-qfpbh 1/1 ...
cert-manager cert-manager-webhook-5cf456b6-9qwlx 1/1 ...
```

The **Install CAPI Changes To BCM**

A new Kubernetes cluster template is generated in BCM on running Install CAPI. The template naming convention used follows the form:

```
capi-<mgmt_cluster_name>-template
```

In the present example, where the name of the Kubernetes management cluster is *default*, the resulting template would be called `capi-default-template`. The name *default* is set by default when creating a Kubernetes instance (figure 4.3).

```
[root@basecm10 ~]# cmsh
[basecm10]% kubernetes
[basecm10->kubernetes]% list
Name (key)
----------------------
capi-default-template
default
[basecm10->kubernetes]% show capi-default-template
Parameter Value
------------------------------------
Name capi-default-template
Revision
Etcd Cluster
Pod Network
Pod Network Node Mask
Internal Network
KubeDNS IP 0.0.0.0
Kubernetes API server
Kubernetes API server proxy port 6444
App Groups <0 in submode>
Label Sets <0 in submode>
Notes
Version
Trusted domains kubernetes,kubernetes.default,kubernetes.default.svc,master,localhost
Module file template <690B>
```
This becomes the default template used for generating Kubernetes CAPI clusters associated with this specific Kubernetes management cluster. Users can modify this template, and can even create additional templates as needed. The YAML configuration for a CAPI cluster also lets users specify a template, via annotations within the YAML file (section 8.8).

### 8.2.2 The Assign CAPI Role Option

After the Install CAPI operation has completed its run, the Assign CAPI Role operation can be carried out from the setup screen of figure 8.2. The Assign CAPI Role operation carries out a CAPI role assignment to a group of nodes by creating a configuration overlay (section 4.5). The role can then be assigned either to a category, or to individual nodes.

The software images on the chosen nodes get the packages needed to run the CAPI agents.

When the Assign CAPI Role operation is carried out, the screens in figures 8.4-8.7 may be displayed:

**Figure 8.4: Select a Kubernetes management cluster**

**Figure 8.5: Customize the configuration overlay**

**Figure 8.6: Choose nodes via categories**
Some of the screens may be skipped if they are not needed. For example, the screen for selecting individual nodes (figure 8.7) is not displayed if all the nodes have already been assigned during selection via categories.

**The Assign CAPI Role Actions**

During execution of the Assign CAPI Role operation, the wizard:

- prepares the Kubeadm and Helm repositories in the software images for the nodes.
- writes the IP Forwarding configuration to the software images for the nodes.
- installs necessary packages in the software images for the nodes (for example: cm-capi).
- provisions the nodes from their software images.
- creates a configuration overlay in BCM.
- assigns the selected categories and nodes to the overlay.
- assigns the capi role and containerd roles to the overlay.

**The Assign CAPI Role Changes To BCM**

The configuration overlay created during the execution of the Assign CAPI Role operation is now visible in the configurationoverlay mode (some output elided):

**Example**

```
[root@basecm10 ~]# cmsh
[basecm10]% configurationoverlay
[basecm10->configurationoverlay]% list
Name (key) Priority All head nodes Nodes Categories Roles
------------------- ---------- -------------- ---------------- ----------------- ------------------
capi-default 500 no node001,node002 capi
kube-default-etcd 500 no node003 Etcd::Host
...
[basecm10->configurationoverlay]% show capi-default
Parameter Value
-----------------------------------------------
Name capi-default
Revision
All head nodes no
Priority 500
Nodes node001,node002
Categories
Roles capi
Customizations <0 in submode>
```
The preceding overlay shows that the capi role has been assigned to nodes node001 and node002. The capi role sets the capi-agent service for a node. If it is set, then BCM bootsraps the agent (section 8.4) and initiates the capi-agent systemd service that runs on the node.

The containerd role takes care of the containerd service, so that containerd is also started on the nodes. The containerd service is needed after the agent initializes the provisioning of its node to integrate it into a Kubernetes CAPI cluster. This is because a kubelet service is eventually started, which relies on containerd.

CAPI uses kubeadm for node provisioning and management, which means that containerd must be running in advance. Therefore, as part of its pre-launch checks, kubeadm verifies that the containerd service is operational.

**The Assign CAPI Role Changes To the Kubernetes management cluster**

The BCM CAPI host agents should register with the Kubernetes management cluster, and can then be seen on running `kubectl` with the `byohost` option:

**Example**

```
[root@basecm10 ~]# module load kubernetes/default/1.26.5-0
[root@basecm10 ~]# kubectl get byohost -A
NAMESPACE  NAME       OSNAME          OSIMAGE                       ARCH
default    node001   linux             Rocky Linux 8.7 (Green Obsidian) amd64
default    node002   linux             Rocky Linux 8.7 (Green Obsidian) amd64
```

Registration does not necessarily make these hosts a part of any Kubernetes CAPI cluster. However, if there are clusters that have been created beforehand, and there are machines that are awaiting additional resources, then it might be that some or all of these hosts are immediately provisioned. A simple method to check this is by querying the machine resource.

```
[root@basecm10 ~]# kubectl get machine -A
No resources found
```

If however there were machines that were pending, the output might look like:

```
[root@basecm10 ~]# kubectl get machines
NAME                  CLUSTER           NODENAME      PROVIDERID           PHASE          AGE     VERSION
byoh-cluster-control-plane-h2d64  byoh-cluster           Provisioning   4s               v1.26.1
byoh-cluster-md-0-56985bf9d6xkhj68-dxvkc  byoh-cluster     Pending             6s               v1.26.1
```

When the machines are provisioned, the Kubernetes CAPI cluster output state might look like:

```
[root@basecm10 ~]# kubectl get machines
NAME                  CLUSTER           NODENAME      PROVIDERID           PHASE          AGE     VERSION
byoh-cluster-control-plane-h2d64  byoh-cluster           node001 byoh://node001/e7fq1b Running 3h3m...
byoh-cluster-md-0-56985bf9d6xkhj68-dxvkc  byoh-cluster           node002 byoh://node002/ngkju0 Running 3h3m...
```

This is discussed further in section 8.3.

### 8.3 Deploying A Kubernetes Cluster Through CAPI

The BCM CAPI Infrastructure Provider configures the Kubernetes management cluster. A Kubernetes cluster can then be created through CAPI, and deployed on the available CAPI nodes.

If resources are not available, then the cluster cannot assign nodes to the newly-created cluster definition for provisioning. This results in a persistent Pending state for various resources.

The capi module makes the command line tool `clusterctl` available, which can be used to generate a cluster manifest. In the following example, the tool is used to define a cluster with Kubernetes version 1.26.0, comprising one control plane node and one worker node.
Example

[root@headnode ~]# module load capi/1.3.0
[root@headnode ~]# CONTROL_PLANE_ENDPOINT_IP=10.141.168.1 clusterctl generate cluster byoh-cluster \
   --infrastructure byoh --kubernetes-version v1.26.0 \
   --control-plane-machine-count 1 --worker-machine-count 1 > cluster.yaml

The environment variable `CONTROL_PLANE_ENDPOINT_IP` must be set by the cluster administrator to a valid unused IP address. The IP address must be within the internal network of the nodes that have been assigned the CAPI role. Here the IP address of 10.141.168.1 is set.

The YAML file that is generated is typically several hundred lines long:

Example

[root@headnode ~]# wc -l cluster.yaml
218 cluster.yaml

The start of it looks like:

Example

[root@headnode ~]# head cluster.yaml
apiVersion: bootstrap.cluster.x-k8s.io/v1beta1
kind: KubeadmConfigTemplate
metadata:
  name: byoh-cluster-md-0
  namespace: default
spec:
  template:
    spec: {}
---
apiVersion: cluster.x-k8s.io/v1beta1

The cluster can now be created with `kubectl`:

Example

[root@headnode ~]# kubectl create -f cluster.yaml
kubeadmconfigtemplate.bootstrap.cluster.x-k8s.io/byoh-cluster-md-0 created
cluster.cluster.x-k8s.io/byoh-cluster created
machinedeployment.cluster.x-k8s.io/byoh-cluster-md-0 created
kubeadmcontrolplane.controlplane.cluster.x-k8s.io/byoh-cluster-control-plane created
byocluster.infrastructure.cluster.x-k8s.io/byoh-cluster created
byomachinetemplate.infrastructure.cluster.x-k8s.io/byoh-cluster-control-plane created
byomachinetemplate.infrastructure.cluster.x-k8s.io/byoh-cluster-md-0 created
k8sinstallerconfigtemplate.infrastructure.cluster.x-k8s.io/byoh-cluster-control-plane created
k8sinstallerconfigtemplate.infrastructure.cluster.x-k8s.io/byoh-cluster-md-0 created

The very first control plane node is assigned this IP address initially, and any other control planes nodes do not try to take it at that time. Only one of the control plane nodes use this IP address at a time.

Load balancing between all three nodes is however possible—it is just not currently configured out-of-the-box at the time of writing (July 2023).

When all prerequisites are met, the BCM CAPI Infrastructure Provider initiates node provisioning, and creates the cluster. BCM generates a module file and kubeconfig for the cluster automatically.
8.3.1 Machine Provisioning

Applying the YAML for the cluster creates the machine resources. These are initially in a Pending state:

Example

```
[root@basecm10 ~]# kubectl get machines
NAME                CLUSTER       PHASE   AGE     VERSION
byoh-cluster-control-plane-jzm22 byoh-cluster  Pending 2s     v1.26.0
byoh-cluster-md-0-5c594b479c9jc29-w56w byoh-cluster  Pending 4s     v1.26.0
```

The operator in the byoh-system namespace then allocates resources. It first selects the machine for the control plane:

```
[root@basecm10 ~]# kubectl logs -n byoh-system -l cluster.x-k8s.io/provider=infrastructure-byoh
I0601 06:16:18.202787 .. "msg"="Attempting host reservation" "cluster"="byoh-cluster" ...
I0601 06:16:18.318478 .. "msg"="Successfully attached Byohost" "byohost"="node001" "cluster"="byoh-cluster" ...
```

The capi-agent service on the node finds an installation script and invokes it, as indicated by the log entry:

Example

```
[root@node001 ~]# journalctl -u capi-agent.service -g executing
Jun 01 06:16:20 .. controller/byohost "msg"="executing install script" "name"="node001" ...
```

At this point the ByoHost resource should be linked to a ByoMachine resource, which is linked to a Machine resource. Until the installation script has completed, it is not easy to go the other way around from a Machine resource to find the related ByoHost resource. Once the installation script has completed, the Machine resource updates the ProviderID column with a value, but it can take some time to show up. It shows up later in this session as the value byoh://node001/f3j7mt in this example session.

The output for the installation script is only printed on completion, and the installation script is automatically removed. Figure 8.9 has more details on the installation process.

The machine transitions to the Provisioning state, and also the ByoHost resource is now tied to the given Machine:

Example

```
[root@basecm10 ~]# kubectl get machines
NAME                CLUSTER       PROVIDERID   PHASE   AGE     VERSION
byoh-cluster-control-plane-jzm22 byoh-cluster  Provisioning 4s     v1.26.0
byoh-cluster-md-0-5c594b479c9jc29-w56w byoh-cluster  Pending 6s     v1.26.0
```

The journal for BCM displays output similar to the following (some output ellipsized):

Example

```
[root@basecm10 ~]# journalctl -u cmd.service | grep -i register
Jun 01 06:16:20 .. controller/byohost "msg"="executing install script" "name"="node001" ...
```

The registration of a node by the wizard is done automatically in the background, by using BCM’s Python API (PythonCM, Chapter 1 of the Developer Manual). The node is registered with the active head node, triggering a cascade of events using:

```
cm-kubernetes-capi-setu --register-node node001
```
Script logs can be found in the log file `/var/log/cm-kubernetes-capi-setup.log`. More on what the script does can be found in section 8.5.

If multiple nodes are being provisioned at the same time, then BCM invokes the script with more nodes as arguments so that the work is parallelized. On completion, the machine transitions to the Running state:

**Example**

```
[root@basecm10 ~]# kubectl get machine
NAME  CLUSTER NODENAME PROVIDERID PHASE     AGE      VERSION
byoh-clus... byoh-cluster node001 byoh://node001/e7fq1b Running 3m4s v1.26.0
byoh-clus... byoh-cluster node002 byoh://node002/i0tj4j Running 3m6s v1.26.0
```

Section 8.5.1 covers the process from a different perspective, which may clarify matters further.

### 8.3.2 Accessing The Cluster

Assuming a Kubernetes CAPI cluster named `byoh-cluster` has been deployed, with three control planes, one worker, with the following machines all running:

**Example**

```
[root@basecm10 ~]# kubectl get machines
NAME  CLUSTER NODENAME PROVIDERID PHASE     AGE      VERSION
byoh-clus... byoh-cluster node006 byoh://node006/ytvpbd Running 26m v1.26.1
byoh-clus... byoh-cluster node004 byoh://node004/7rczjg Running 22m v1.26.1
byoh-clus... byoh-cluster node002 byoh://node002/i0tj4j Running 44m v1.26.1
byoh-clus... byoh-cluster node003 byoh://node003/f18zem Running 44m v1.26.1
```

The `KubeCluster` entity can then be seen in `cmsh`:

**Example**

```
[root@basecm10 ~]# cmsh
[basecm10]% kubernetes
[basecm10->kubernetes]% show byoh-cluster
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td><code>byoh-cluster</code></td>
</tr>
<tr>
<td>Revision</td>
<td></td>
</tr>
<tr>
<td>Etcd Cluster</td>
<td></td>
</tr>
<tr>
<td>Pod Network</td>
<td></td>
</tr>
<tr>
<td>Pod Network Node Mask</td>
<td></td>
</tr>
<tr>
<td>Internal Network</td>
<td></td>
</tr>
<tr>
<td>KubeDNS IP</td>
<td><code>0.0.0.0</code></td>
</tr>
<tr>
<td>Kubernetes API server</td>
<td></td>
</tr>
<tr>
<td>Kubernetes API server proxy port</td>
<td><code>6444</code></td>
</tr>
<tr>
<td>App Groups</td>
<td><code>&lt;0 in submode&gt;</code></td>
</tr>
<tr>
<td>Label Sets</td>
<td><code>&lt;0 in submode&gt;</code></td>
</tr>
<tr>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>Version</td>
<td><code>1.26.1</code></td>
</tr>
</tbody>
</table>
| Trusted domains | `kubernetes,kubernetes.default,kubernetes.default.svc.master=localhost`
| Module file template | `<690B>` |
| Kubeadm init file | `<0B>` |
| Service Network |       |
| Kubeadm CERT Key | `<not set>` |
| Kube CA Cert | `<not set>` |
There should also be a `byoh-cluster` module file available on the head node:

**Example**

```bash
[root@basecm10 ~]# module unload kubernetes/mgmt/1.24.9-00
[root@basecm10 ~]# module load kubernetes/byoh-cluster/1.26.1
[root@basecm10 ~]# kubectl get nodes

NAME     STATUS     ROLES       AGE     VERSION
node002   NotReady   control-plane 29m  v1.26.1
node003   NotReady   <none>      25m  v1.26.1
node004   NotReady   control-plane 23m  v1.26.1
node006   NotReady   control-plane 25m  v1.26.1
```

By default, Kubernetes CAPI clusters do not come with a networking implementation configured. This can be created by the cluster administrator, to see if this improves the state of the cluster (section 4.2.2):

**Example**

```bash
root@rb-capi2:~# kubectl create -f https://raw.githubusercontent.com/projectcalico/calico/v3.24.5 ... /manifests/calico-typha.yaml
poddisruptionbudget.policy/calico-kube-controllers created
poddisruptionbudget.policy/calico-typha created
serviceaccount/calico-kube-controllers created
serviceaccount/calico-node created
configmap/calico-config created
customresourcedefinition.apiextensions.k8s.io/bgpconfigurations.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/bgppeers.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/blockaffinities.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/caliconodestatuses.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/clusterallocations.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/felixconfigurations.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/globalnetworkpolicies.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/globalnetworksets.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/hostendpoints.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/ipamblocks.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/ipamconfigurations.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/ipamnetworkpolicies.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/ipamnetworksets.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/ipamnetworksets.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/kubecontrollersconfigurations.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/networkpolicies.crd.projectcalico.org created
customresourcedefinition.apiextensions.k8s.io/networkpolicies.crd.projectcalico.org created
clusterrole.rbac.authorization.k8s.io/calico-kube-controllers created
clusterrole.rbac.authorization.k8s.io/calico-node created
clusterrolebinding.rbac.authorization.k8s.io/calico-kube-controllers created
clusterrolebinding.rbac.authorization.k8s.io/calico-node created
service/calico-typha created
```
The preceding session ends up with the nodes then ends up configured with Calico networking:

Example

```
root@rb-capi2:~# kubectl get nodes
NAME     STATUS       ROLES       AGE     VERSION
node002   Ready        control-plane 31m     v1.26.1
node003   Ready        <none>      27m     v1.26.1
node004   Ready        control-plane 25m     v1.26.1
node006   Ready        control-plane 28m     v1.26.1
```

8.3.3 Scaling Control Planes Or Workers

Control planes can be scaled at the level of the KubeAdmControlPlane resource (some output ellipsized):

Example

```
[root@basecm10 ~]# module load kubernetes/default
[root@basecm10 ~]# kubectl get kubeadmcontrolplane
NAME     CLUSTER INITIALIZED ... REPLICAS ... UPDATED UNAVAILABLE ...
byoh-cluster-control-plane byoh-cluster true ... 1 ... 1 1 ...
```

```
[root@basecm10 ~]# kubectl patch kubeadmcontrolplane byoh-cluster-control-plane --patch='{"spec": {"replicas": 3}}' --type=merge
kubeadmcontrolplane.controlplane.cluster.x-k8s.io/byoh-cluster-control-plane patched
```

Workers can be scaled at the level of the MachineDeployment resource:

Example

```
[root@basecm10 ~]# kubectl get machinedeployment
NAME     CLUSTER        REPLICAS ... UPDATED UNAVAILABLE PHASE     AGE     VERSION
byoh-cluster-md-0 byoh-cluster 1 ... 1 1 ScalingUp 14h v1.26.1
```

```
[root@basecm10 ~]# kubectl scale --replicas=2 machinedeployment/byoh-cluster-md-0
machinedeployment.cluster.x-k8s.io/byoh-cluster-md-0 scaled
```

8.3.4 Upgrading Control Planes Or Workers

For this section, some background understanding of how Kubernetes upgrades work is recommended. A good introduction can be found at:

https://cluster-api.sigs.k8s.io/tasks/upgrading-clusters.html

Further hints and suggestions on upgrading can be found at:

- https://github.com/kubernetes/sig-release/blob/master/release-engineering/versioning.md#kubernetes-release-versioning: discusses release versioning, and is a recommended first read

- https://kubernetes.io/releases/version-skew-policy/: discusses version skew policy, but is also a document that provides more detailed information

One suggestion from these resources that should be followed, is first to upgrade to the latest patch version of the current minor version, and then to upgrade to the next minor version. Typically all the control planes are upgraded first, and then the workers.
Rolling Upgrades
The default upgrade method is with rolling upgrades. More information on rolling upgrades can be found at:

https://cluster-api.sigs.k8s.io/tasks/upgrading-clusters.html#
upgrading-machines-managed-by-a-machinedeployment

The rolling upgrades method requires that at least one spare ByoHost is available, since machines are replaced one by one, and both the old and new machine need to run at the same time during a part of the procedure. If assigning an extra CAPI role is a problem, then the upgrade strategy based on OnDelete can be followed instead, which is also described at that URL.

Another option, depending on the cluster, could be to temporarily scale down the workers by one during the rolling upgrade, via the MachineDeployment resource.

Deprecated APIs
- The deprecation guide at:


  should be read before carrying out upgrades. The cluster administrator should check the changelogs and upstream documentation for obsolete APIs in the target version.
- The pluto utility (https://github.com/FairwindsOps/pluto) can check for deprecated API usage.
- The kubent utility (https://github.com/doitintl/kube-no-trouble) can also be used, but seems a few Kubernetes versions behind at the time of writing (June 2023).

Upgrading The Control Plane
The following example session upgrades the control plane from version v1.26.1 to v1.26.2:

Example

```
[root@basecm10 ~]# kubectl get kubeadmcontrolplane
NAME                  CLUSTER INITIALIZED .. REPLICAS .. UPDATED UNAVAILABLE AGE VERSION
byoh-cluster-control-plane byoh-cluster true 1 1 1 16h v1.26.1

[root@basecm10 ~]# kubectl patch kubeadmcontrolplane byoh-cluster-control-plane
  --type=merge
  -p '{"spec": {"version": "v1.26.2"}}'

kubeadmcontrolplane.controlplane.cluster.x-k8s.io/byoh-cluster-control-plane patched
```

The resource immediately shows the new version. However, the upgrade is not performed immediately:

```
[root@basecm10 ~]# kubectl get kubeadmcontrolplane
NAME                  CLUSTER INITIALIZED .. REPLICAS .. UPDATED UNAVAILABLE AGE VERSION
byoh-cluster-control-plane byoh-cluster true 1 0 1 16h v1.26.2
```

First a new control plane is provisioned with the new version. Only when it is fully up, is the old control plane deleted:

```
[root@basecm10 ~]# kubectl get machine
NAME   CLUSTER NODENAME PROVIDERID PHASE AGE VERSION
byoh-clu.. byoh-cluster node001 byoh://node005/gije0o Running 125m v1.26.1
byoh-clu.. byoh-cluster node002 byoh://node004/98s8gu Running 98m v1.26.1
```

The deletion is not carried out immediately. There is a grace period in which both control planes run. The old control plane is deleted shortly afterwards.

```
[root@basecm10 ~]# kubectl get machine
NAME    CLUSTER   NODENAME   PROVIDERID   PHASE   AGE  VERSION
byoh-clus... byoh-cluster node001 byoh://node005/gije0o Running 127m v1.26.1
byoh-clus... byoh-cluster node003 byoh://node003/mjvvvg Running 2m30s v1.26.2
byoh-clus... byoh-cluster node004 byoh://node004/98s8gu Running 101m v1.26.1
```

If there is than one control plane, then a rolling upgrade takes place, one control plane at a time.

### Upgrading The Workers

The following session shows the workers being upgraded from version v1.26.1 to v1.26.2.

**Example**

```
[root@basecm10 ~]# kubectl get machinedeployment
NAME    CLUSTER   REPLICAS .. UPDATED UNAVAILABLE PHASE   AGE  VERSION
byoh-cluster-md-0 byoh-cluster 3 .. 3 3  ScalingUp 17h v1.26.1
```

In this case there are three workers. The version is patched with:

**Example**

```
[root@basecm10 ~]# kubectl patch machinedeployment byoh-cluster-md-0 \
  --type=merge \
  -p '{"spec": {"template": {"spec": {"version": "v1.26.2"}}}}'
```

The resource immediately shows the new desired version. However, the upgrade is not performed immediately:

**Example**

```
[root@basecm10 ~]# kubectl get machinedeployment
NAME    CLUSTER   REPLICAS .. UPDATED UNAVAILABLE PHASE   AGE  VERSION
byoh-cluster-md-0 byoh-cluster 4 .. 1 4  ScalingUp 17h v1.26.2
```

First a new worker is provisioned with the new version. Only when it is fully up, is the old worker deleted:

```
[root@basecm10 ~]# kubectl get machine
NAME    CLUSTER   NODENAME   PROVIDERID   PHASE   AGE  VERSION
byoh-clus... byoh-cluster node003 byoh://node003/mjvvvg Running 5m45s v1.26.2
byoh-clus... byoh-cluster node004 byoh://node004/98s8gu Running 104m v1.26.1
byoh-clus... byoh-cluster node002 byoh://node002/f3j7mt Running 107m v1.26.1
byoh-clus... byoh-cluster node006 byoh://node006/paityf Running 139m v1.26.1
byoh-clus... byoh-cluster Provisioning 7s v1.26.2
```

The process repeats itself until all the workers are upgraded.

### 8.4 BCM Host Agent Registration

The registration process in figure 8.8 is sourced from the documentation at the upstream BYOH project:

The roles of the platform operator and site operator in the illustration are automated by BCM.

The process initiated by BCM CAPI Infrastructure Provider starts with BCM.

When initiating, the capi-agent service verifies the existence of a configuration before launching. If a configuration does not exist, then the BCM API is used to request one.

BCM then automates all the steps necessary to generate this configuration, and prepares it for the capi-agent so it can start its process.

To clarify figure 8.8 further, before step 9 (create a new kubectl file at ~/.kube/config), only the bootstrap configuration exists on the node. However, after this step, two configuration files exist: the bootstrap and the definitive configuration files. The presence of these files enables BCM to determine the current phase of the bootstrap process that the capi-agent is in.

Manual generation of bootstrap configurations is also possible. Details on this are given in the documentation at https://github.com/vmware-tanzu/cluster-api-provider-bringyourownhost/blob/main/docs/getting_started.md#register-byoh-host-to-management-cluster

The host agent registration process described in the preceding is part of what is carried out for nodes when running the Assign Capi Role option in section 8.2.2.

```
[root@basecm10 ~]# kubectl get byohost -A
NAMESPACE   NAME   OSNAME   OSIMAGE                  ARCH
default     node001 linux Rocky Linux 8.7 (Green Obsidian)  amd64
default     node002 linux Rocky Linux 8.7 (Green Obsidian)  amd64
default     node003 linux Rocky Linux 8.7 (Green Obsidian)  amd64
default     node004 linux Rocky Linux 8.7 (Green Obsidian)  amd64
default     node005 linux Rocky Linux 8.7 (Green Obsidian)  amd64
default     node006 linux Rocky Linux 8.7 (Green Obsidian)  amd64
```

### 8.5 Install Process BCM CAPI

The preceding section (section 8.4) focused on BCM CAPI host agent registration. The current section (section 8.5) discusses the deployment of Kubernetes clusters on these CAPI Hosts.

The deployment uses installation scripts transmitted to the designated nodes through “installation secrets”.

Node registration with BCM is also carried out as part of the deployment of Kubernetes on the CAPI hosts, and is distinct from the host agent registration of section 8.4. The deployment process flow is illustrated in figure 8.9:
8.5 Install Process BCM CAPI

Figure 8.9: BYOH installer flow
8.5.1 Registration Process Of The Node With BCM

With BCM CAPI Infrastructure Provider, the installation script is a python cm script. It establishes communication with BCM on the active Head Node. This is referred to as “registering” the node with BCM.

The registration process involves essential bookkeeping tasks, such as tracking which nodes belong to specific Kubernetes CAPI clusters, and preparing software images with the appropriate Kubernetes versions.

The example that follows illustrates various components involved in creating a Kubernetes cluster using CAPI. In the example, a single head node operates as the Kubernetes management cluster, and multiple nodes are assigned the CAPI role. A cluster definition is generated for a control plane node and a worker node using the procedure described in section 8.3. An overview of the components involved when the cluster definition is applied is seen in figure 8.10:
Figure 8.10: CAPI summary
In figure 8.10, the Kubernetes CAPI cluster has the control plane deployed on node001, followed by the deployment of the worker node software image on node002. The diagram of the is expanded from the node001 CAPI agent block illustrates the steps executed by the node001 capi-agent.

Upon registering with the BCM API, the more important actions that BCM carries out are:

- **Step 1**: Preparation of Kubernetes version 1.26.0 within the software image for the registering node.

- **Step 2**: Provisioning of the node through an image update, for example using the `imageupdate` command of cmsh (section 5.6.2 of the Administrator Manual), to ensure that it stays synchronized with its software image.

- **Step 3**: Handing control back to the capi-agent, which proceeds with the subsequent step in the CAPI cluster creation process.

The steps also include processes beside the ones illustrated in figure 8.10. The more complete sequence is:

- **Step 1**: The installation script is invoked with contextual information, such as the desired Kubernetes version (for example: 1.26.0) and the cluster name.

- **Step 2**: The installation script, implemented as a PythonCM script, establishes communication with BCM. It registers itself with the active head node and remains in a waiting state until notified to terminate.

- **Step 3**: BCM creates a mapping that specifies which node should be provisioned with the corresponding Kubernetes version.

- **Step 4**: BCM ensures that the software image for the node contains the requested Kubernetes version.

- **Step 5**: BCM ensures that the node is provisioned with its designated software image.

- **Step 6**: BCM signals to the node that the installation script has completed.

From this point onward, the default logic for the BCM CAPI operator takes over, and kubeadm initializes the node accordingly.

**8.5.2 Creating A Kubernetes Cluster Via CAPI**

The steps in section 8.5.1, are about the node registration process and node provisioning during Kubernetes cluster creation with CAPI. The following steps zoom out further, and describe the complete process of creating a Kubernetes cluster through CAPI.

- **Step 1**: The system administrator defines a cluster to be deployed by the BCM CAPI operator, and feeds it to the Kubernetes API server using kubectl.

- **Step 2**: The definition results in a number of Kubernetes resources being created, such as a Cluster, ByoCluster, MachineDeployment.

- **Step 3**: The BCM CAPI operator responds to these newly-created resources.

- **Step 4**: CAPI starts assigning CAPI agents to reconcile specific Machines, beginning with the first control plane. This is indicated by the red provisioning line, numbered with a 1.

- **Step 5**: BCM prepares a module file and kubecfg for the new Kubernetes CAPI cluster, and writes these to disk for the system administrator.

- **Step 6**: BCM CAPI operator updates its records using the newly-created control plane and proceeds with provisioning the additional control planes or workers, such as the additional worker. This is indicated by the red line numbered with a 2.
8.6 Configuring CAPI Versions In Software Images

BCM takes care of configuring CAPI versions in software images automatically. This is part of its BCM CAPI node registration process, as mentioned earlier in section 8.5.

An alternative is to manually pre-install a specific version of Kubernetes in a software image. This can be done from within the softwareimage mode of cmsh:

```
[root@basecm10 ~]# cmsh
[basecm10]# softwareimage
[basecm10->softwareimage]# use default-image
[basecm10->softwareimage[default-image]]# help capi
Name:
capi - Manage Kubernetes CAPI versions on the image

Usage:
capi [OPTIONS] list
    List CAPI versions on all / current image

capi [OPTIONS] add <version> [ <version> ... ]
    Add specified versions on all / current image

capi [OPTIONS] remove <version> [ <version> ... ]
    Remove all CAPI versions on all / current image

capi [OPTIONS] clear
    Remove all CAPI clusters

Options:
-v, --verbose
    Be more verbose

-d, --delimiter <string>
    Use <string> as delimiter between columns. Use {} for JSON, and {<digit>} for JSON with a specific indentation.

--image, -i <list of images>
    Perform action on comma separated list of images

--repo-refresh, -r
    Refresh the repository cache before adding new versions

--debug
    Run script with debug on

Examples:
capi list
    List CAPI versions on all / current image

capi clear
    Remove all CAPI versions on all / current image

capi add 1.26.0 1.27.*
    Add specified versions on all / current image

Example
```

```
[basecm10->softwareimage[default-image]]# capi list

<table>
<thead>
<tr>
<th>Node</th>
<th>image</th>
<th>versions</th>
<th>Result</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>basecm10</td>
<td>/cm/images/default-image 1.24.9, 1.23.0, 1.26.1, 1.26.2</td>
<td>good</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[basecm10->softwareimage[default-image]]# capi add 1.27.*

<table>
<thead>
<tr>
<th>Node</th>
<th>image</th>
<th>versions</th>
<th>Result</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>basecm10</td>
<td>/cm/images/default-image 1.24.9, 1.23.0, 1.26.1, 1.26.2, 1.27.0, 1.27.1, 1.27.2</td>
<td>good</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8.7 Removing Kubernetes CAPI clusters

The following steps remove the CAPI clusters:

• **Step 1:** The removal of the CAPI clusters is started with the `kubectl delete` command:
Step 2: If all CAPI hosts are not part of any significant clusters, then the configuration overlay configuration overlays are removed:

```
[...]
* Step 2: If all CAPI hosts are not part of any significant clusters, then the configuration overlay configuration overlays are removed: *
```

The removal of the configuration overlays causes containerd and the CAPI agents to stop on the hosts.

Step 3: For additional cleanliness, the bootstrap configurations for each of the hosts should be removed:

```
[...]
* Step 3: For additional cleanliness, the bootstrap configurations for each of the hosts should be removed: *
```

Step 4: The infrastructure provider is then removed:

```
[...]
* Step 4: The infrastructure provider is then removed: *
```

Step 5: The removal of all the ByoHost resources can be checked:

```
[...]
* Step 5: The removal of all the ByoHost resources can be checked: *
```

Step 6: Finally, the Kubernetes CAPI clusters is eliminated from BCM itself:

```
[...]
* Step 6: Finally, the Kubernetes CAPI clusters is eliminated from BCM itself: *
```
8.8 Kubernetes CAPI Templates

In BCM, each Kubernetes cluster is represented as a KubeCluster entity stored in the BCM database. These entities can be viewed in the kubernetes submode of cmsh:

Example

[root@basecm10 ~]# cmsh
[root@basecm10]$ kubernetes
[root@basecm10->kubernetes]$ list
Name (key)
-------------------
default

- For every Kubernetes management cluster, a default KubeCluster CAPI template is generated once the CAPI role node is assigned using the wizard, as discussed in section 8.2.2. This template serves as a base for all subsequent Kubernetes clusters created via CAPI.

- For example, if an administrator creates a cluster my-capi-cluster via CAPI, for the Kubernetes management cluster mgmt, then BCM clones the capi-mgmt-template KubeCluster entity to create a new KubeCluster my-capi-cluster.

- At the time of CAPI cluster creation, the KubeCluster template can also be customized using an annotation in the Cluster resource definition:

Example

apiVersion: cluster.x-k8s.io/v1beta1
class: Cluster
metadata:
  labels:
    cni: byoh-cluster-crs-0
crs: "true"
infra1.mvidia.x-k8s.io/capi-kube-template: "my-capi-cluster-template"
name: byoh-cluster
namespace: default
spec:
  clusterNetwork:
    pods:
      cidrBlocks:
      - 192.168.0.0/16
    serviceDomain: cluster.local
  services:
    cidrBlocks:
      - 10.128.0.0/12
  controlPlaneRef:
    apiVersion: controlplane.cluster.x-k8s.io/v1beta1
class: KubeadmControlPlane
name: byoh-cluster-control-plane
  infrastructureRef:
    apiVersion: infrastructure.cluster.x-k8s.io/v1beta1
class: ByoCluster
name: byoh-cluster

The provided YAML illustrates a cluster byoh-cluster definition, with the specified capi-template set to my-capi-cluster-template. If present, then this is used instead of the default.
• The fallback sequence for the CAPI template is: The specified label (for example: my-capi-cluster-template) is used if it exists; otherwise the default capi-mgmt-template is used. If neither is available, then capi-template is used.
BCM provides an application containerization tool called Singularity. Singularity is designed to execute containers as if they are just native applications on a host computer, and to work with HPC. Singularity users can therefore run Singularity containers just as they run any other program on an HPC cluster. BCM provides Singularity version 3.8.4.

9.1 Use Cases

Adding Singularity to BCM brings a stronger integration of containerization with HPC. While Docker and Kubernetes can work within HPC, some drawbacks still prevent the use of HPC resources in the way that HPC users and administrators are used to.

Besides the use of Singularity containers in HPC jobs, Singularity users can create portable images with their applications. Singularity images are files that represent the container filesystem. These images can be copied from one environment (cluster) to another and executed without modification. Thus, when a user creates a container image file, it is up to the user what files, or which RPMs, to install in the image. For example, the user can create an image file that bundles Open MPI with the user’s application. This guarantees that the application will be able to run if it requires that MPI implementation, even if no MPI libraries are installed on the execution host or if there is some version incompatibility.

There is no need for a special configuration inside workload managers in order to use Singularity. This is because the containers are designed to be run like any application on the system. Users need just to use the image file as the usual script or binary to be executed in their jobscripts or in a shell. The singularity command can also be used to apply special options to the container, when executing the image file in the jobscript or shell.

9.2 Package cm-singularity

Singularity is packaged for SLES12, SLES15, Ubuntu 18.04, Ubuntu 20.04, RHEL7/CentOS7, and RHEL8/Rocky Linux 8. It is available from the YUM or Zypper repositories from version 7.3 of NVIDIA Base Command Manager onward, and is distributed as a package called cm-singularity. The package should be installed in the software image for each node. The user is able to run a Singularity image only if the Singularity package is installed on the node. In order to allow users to build an image, it makes sense to install the package on the head and login nodes as well. The tool does not provide services that run in the background, so a simple installation of the package is enough to start using it.

Singularity contexts are always run as the user running them. This means that there is no risk in allowing the containers to have access to, and interact with, the file system of the host.

This means that, if an image is created by the root user on a machine, then the files that require root access inside the image, still need to be allowed root permissions on any other machine. Thus, if a user creates an image on a laptop, and adds a file that can be read only by the root user, then when the container is started on another machine by a regular user, that regular user has no access to the root-only readable file inside the container.
While there is no daemon running as root, nor any persistent processes that an attacker may use to escalate privileges, there is a need to run some system calls as root so that the container is encapsulated. For this part of the run flow, there is a single SUID binary called Sexec (Singularity Exec). This is a simple binary that is as small as possible, and which the Singularity developers claim has been audited by multiple security experts.

9.3 MPI Integration

Because of the nature of Singularity, all MPI implementations should work fine inside a Singularity container. The developers of the tool have spent a lot of effort in making Singularity aware of Open MPI, as well as adding a Singularity module into Open MPI so that running at extreme scale is as efficient as possible. However, in some cases, starting an MPI process may not be as optimal as execution outside the container. So, specifically for Open MPI, Singularity provides a special mechanism to handle the execution of MPI processes. It adds all the MPI processes of the same MPI application to the same container on a host. This also reduces the application startup time. The Open MPI daemon orted in this case is not added to the running container, which means the overhead of starting up daemons is reduced.

When an Open MPI application that has been packaged to an image is started, the following steps take place:

1. mpirun is called;
2. mpirun forks and executes orted;
3. orted initializes the PMI (process management interface);
4. orted forks as many times as the number of processes per node requested;
5. the container image is started in each fork (because it is the original command specified in mpirun arguments);
6. each container process executes the command (that is, the MPI application) passed inside the given container;
7. each of the MPI process links to the dynamic Open MPI library, which loads shared libraries with dlopen system call;
8. Open MPI libraries connect back to the original orted process via PMI;
9. all non-shared memory communication then occurs through the PMI, and then passes on to local network interfaces.

Additional information about Singularity usage can be found in Chapter 11 of the User Manual. The official web site of the tool is https://www.sylabs.io/singularity.
Base Command Manager Essentials And NVIDIA AI Enterprise

Base Command Manager Essentials (BCME) is the NVIDIA AI Enterprise (https://docs.nvidia.com/ai-enterprise/index.html) edition of Base Command Manager.

A.1 Scope Of BCME

BCME:

- provisions clusters. This includes:
  - operating system installation
  - networking setup
  - security configuration
  - DNS configuration
  while ensuring cluster integrity
- automates server management and updates, preventing server drift
- manages AI workloads with:
  - Kubernetes
  - automated scaling
  - a tightly integrated Run:ai
- can install and manage Slurm workload manager
- enables a streamlined Jupyter setup with NGC containers
- provides comprehensive management for cluster control and job monitoring. This includes managing and monitoring for
  - GPU metrics
  - resource allocation
  - access control
  - chargeback options
A.2 BCME And Support For NVIDIA AI Enterprise

A.2.1 Certified Features Of BCME For NVIDIA AI Enterprise
Some features of BCME are certified for NVIDIA AI Enterprise.
The BCM Feature Matrix at:
https://support.brightcomputing.com/feature-matrix/
has a complete list of the features of BCME that are certified for NVIDIA AI Enterprise.

A.2.2 NVIDIA AI Enterprise Compatible Servers
BCME must be deployed on NVIDIA AI Enterprise compatible servers.
The NVIDIA Qualified System Catalog at:
displays a complete list of NVIDIA AI Enterprise compatible servers, if the options
• AI Enterprise Bare Metal and
• AI Enterprise vSphere

are ticked in the NVIDIA Cert. Type filter menu dropdown.

A.2.3 NVIDIA Software Versions Supported
NVIDIA AI Enterprise supports specific versions of NVIDIA software, including
• NVIDIA drivers
• NVIDIA containers
• the NVIDIA Container Toolkit
• the NVIDIA GPU Operator
• the NVIDIA Network Operator

The NVIDIA AI Enterprise Catalog On NGC at:
https://catalog.ngc.nvidia.com/enterprise
lists the specific versions of software included in a release.

A.2.4 NVIDIA AI Enterprise Product Support Matrix
The NVIDIA AI Enterprise Product Support Matrix at:
lists the platforms that are supported.