Simplifying GPU monitoring in OCI Kubernetes Engine (OKE) with Node Manager

Augest 8, 2024

We are happy to announce the release of the Node Manager, a tool to enhance GPU, networking, and other infrastructure level monitoring in Kubernetes clusters created through Oracle Cloud Infrastructure Kubernetes Engine (OKE). This tool gathers relevant information from the nodes within your cluster and surfaces it for you to consume in a Kubernetes native way. This tool is actively in development and over time we will continue to add useful data, such as from the NVIDIA Data Center GPU Manager (DCGM) daemon, and other sources relevant to managing the health of instances running in your clusters.

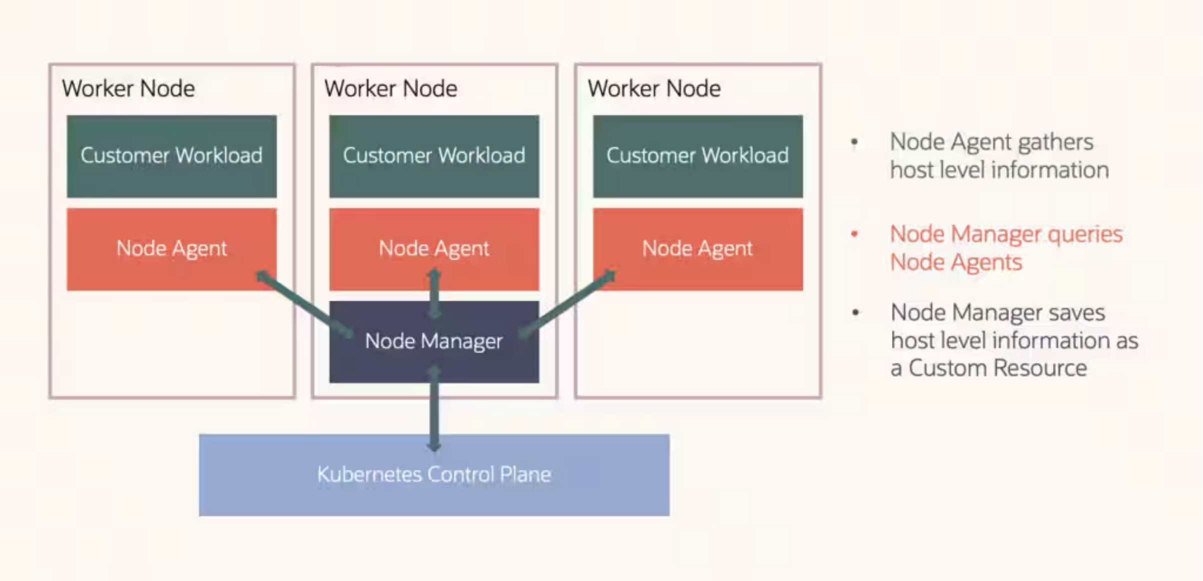
Kubernetes, an open source container orchestrator, has become the de facto standard for the cloud native development practices that characterize modern IT. It is used to deploy, manage, and scale containerized workloads and handles common operational requirements, such as node management, load balancing, self-healing, and storage mounting, all of which are relevant to customers looking to run artificial intelligence (AI) and machine learning (ML) workloads. More and more of our customers are running AI and ML workloads, such as model training and inferencing, on OKE clusters. These clusters consist of bare metal and virtual machine compute shapes that include Intel or

AMD CPUs and NVIDIA graphics processors designed for hardware-accelerated workloads. Customers with greater performance requirements leverage Cluster Networks with Instance Pools, groups of identical high performance computing (HPC), graphical processing unit (GPU), or optimized instances connected through a high-bandwidth, ultra low-latency network. Each node in the cluster is a bare metal machine located in close physical proximity to the other nodes. A remote direct memory access (RDMA) network between nodes provides latency as low as single-digit microseconds, comparable to on-premises HPC clusters.

Workloads leveraging GPUs, especially large scale ones, such as those training models, are more sensitive to events impacting underlying hardware and network components. This sensitivity necessitates visibility into underlying hardware issues in order to troubleshoot problems as quickly as possible or ideally to address them before the workload is meaningfully impacted. While OKE offers node health metrics through OCI Monitoring, they are not enough to solve GPU or RDMA specific problems that require low level visibility into the physical health of the underlying hardware. This is where the Node Manager comes in. Node Manager provides a lightweight, real-time hardware level monitoring solution for OKE worker nodes. It exposes detailed physical data, for example GPU counts and RDMA network error codes, both within Kubernetes as Custom Resources for a fully native Kubernetes experience as well as through an OpenMetrics endpoint to allow you to process data through Prometheus or your preferred observability tool.

How does the Node Manager work?

Node Manager relies on an agent running on each worker node to collect hardware level information about the node. These agents expose this information to the Node Manager through an endpoint. The Node Manager queries these endpoints to aggregate cluster-wise node information and saves the information in Kubernetes Custom Resources for users to access.



- Node Agent gathers host level information
- Node Manager queries Node Agents
- Node Manager saves host level information as a Custom Resource

Deploying The Node Manager

The Node Manager is available to deploy from the Oracle Cloud Infrastructure Registry (OCIR) service at this location: ocir.us-ashburn-1.oci.oraclecloud.com/oracle/oke-node-manager:v0.1.0. In addition, you will need the manifest file for the Node Manager, which is available here.

To deploy the Node Manager, run the following kubectl command:

```
$ kubectl apply -f node-manager-manifest.yaml

customresourcedefinition.apiextensions.k8s.io/instances.oci.oraclecloud.com created

customresourcedefinition.apiextensions.k8s.io/interfaces.oci.oraclecloud.com created

customresourcedefinition.apiextensions.k8s.io/networktests.oci.oraclecloud.com created

customresourcedefinition.apiextensions.k8s.io/nodemanagements.oci.oraclecloud.com created

customresourcedefinition.apiextensions.k8s.io/vnicattachments.oci.oraclecloud.com created

serviceaccount/oci-oke-node-agent created

serviceaccount/oci-oke-node-manager created
```

```
role.rbac.authorization.k8s.io/oci-oke-node-leader-election-role created

clusterrole.rbac.authorization.k8s.io/oci-oke-node-agent-role created

clusterrole.rbac.authorization.k8s.io/oci-oke-node-manager-role created

srolebinding.rbac.authorization.k8s.io/oci-oke-node-leader-election-rolebinding created

clusterrolebinding.rbac.authorization.k8s.io/oci-oke-node-agent-rolebinding created

clusterrolebinding.rbac.authorization.k8s.io/oci-oke-node-manager-rolebinding created

configmap/oci-oke-node-agent-config created

configmap/oci-oke-node-custom-node-config created

configmap/oci-oke-node-manager-config created

service/oci-oke-node-manager created

deployment.apps/oci-oke-node-manager created

daemonset.apps/oci-oke-node-agent created
```

Label the nodes you want the Node Agent to collect data from (e.g. GPU nodes) with the following label:

```
oci.oraclecloud.com/node.agent.deploy=true

$ kubectl label node 10.0.10.227

node/10.0.10.227 labeled
```

Note: instance information in custom resources, including shape, status, score, and more may take 10 or more minutes to populate.

Using The Node Manager

The node manager creates a number of custom resource definitions, including:

CRD	Definition
instances.oci.oraclecloud.com	The main resource that holds information about individual
interfaces.oci.oraclecloud.com	Network interfaces.

networktests.oci.oraclecloud.com	Specification for which network tests to run and
nodemanagements.oci.oraclecloud.com	Configuration for the node
vnicattachments.oci.oraclecloud.com	Specification and status of vNIC attachments to the host.

Node Manager surfaces information collected on nodes through custom resources associated with the definitions shared above. Describe the custom resource, for example instances.oci.oraclecloud.com to see information supplied by the node manager:

```
$ kubectl get instances.oci.oraclecloud.com

NAME        SERIAL   IP          ENABLED   DETAIL   SHAPE
STATUS   SCORE   RDMA   MODIFIED   AGE

10.0.10.227         10.0.10.227   true      Node     VM.Standard.E3.Flex   OK
100          114s       19m

10.0.10.46          10.0.10.46    true      Node     VM.Standard.E3.Flex   OK
100          85s        19m

10.0.10.9           10.0.10.9     true      Node     VM.Standard.E3.Flex   OK
100          92s        19m

10.0.10.98          10.0.10.98    true      Node     VM.Standard.E3.Flex   OK
100          65s        19m
```

Additional information is available from running the command with the -o wide option:

```
$ kubectl get instances.oci.oraclecloud.com -o wide

NAME                       SERIAL      IP            ENABLED   DETAIL   SHAPE
STATUS        SCORE   RDMA   OCA        GPU          MLXFW        ERRORS
MODIFIED   AGE

abc-gpu-111-333-aabbccddee   aabbccddee   10.114.238.6     true      Node
BM.GPU.H100.8        OK           100      16      1.39.0-9   535.161.08
28.39.2500                              25s       3d19h

abc-gpu-111-444-aaaaccddee   aaaaccddee   10.114.189.118   true      Node
BM.GPU.H100.8        OK           100      16      1.39.0-9   535.161.08
28.39.2500                              78s       3d19h
```

```
abc-gpu-222-555-aaaaaaddee   aaaaaaddee   10.114.137.227   true      Node
BM.GPU.H100.8        Fault        82       16     1.39.0-9   535.161.08   28.39.2500
["rdma15 physical errors"]       97s        3d19h
```

Node Manager Endpoints and Data

The Node Manager exposes the following endpoints for retrieving data about your Kubernetes cluster:

Endpoints	Description	Data Provided
$NODE_AGENT_IP:8086/instance	Provides instance information	• Version information for the GPU Driver, MLX OFED, Oracle Cloud Agent (OCA), OKE, Kubelet, CRI-O, MLX Firmware and MLX Firmware Tool (MFT) • Instance information such as instance OCID, instance configuration OCID, image OCID, node pool/instance pool OCID • Number of RDMA devices, compute shape information, vNIC information • Tunable rules information such as tunable rule specifications, failures, values identified for a rule
$NODE_MANAGER_IP:8087/metrics	Provides performance and runtime metrics for Node Manager	• Performance and runtime information regarding time taken by Controller loop runs • Performance and runtime information regarding time taken to query individual Agents • Mellanox Effective Physical Errors • Mellanox Troubleshooting Codes
$NODE_AGENT_IP:8088/metrics	Provides performance and runtime metrics for Node Agent	• Performance and runtime information regarding time taken by the Agent to gather data

Conclusion

Node Manager simplifies monitoring hardware level information, including from GPUs, networking, and other infrastructure components, from worker nodes in your OCI Kubernetes Engine (OKE) clusters. It captures information regarding physical errors impacting hosts and host networks in addition to instance information and surfaces the information in a Kubernetes-native way, making it easier for you to make use of it to operate your clusters. We designed this tool to be extensible and plan to add checks and surface additional metrics over time.

For more information, see the following resources:

- OCI Kubernetes Engine Documentation
- Running Applications on GPU-based Nodes
- Kubernetes at scale just got easier
- OKE best practices
- Start your free trial of Oracle Cloud Infrastructure

About Cloudsway

Cloudsway is a subsidiary of Wangsu Science and Technology (stock code: 300017), established in March 2023. Wangsu Science and Technology is a global leading provider of information infrastructure platform services, with business spread across more than 70 countries and regions worldwide.

Cloudsway is one of the three innovation engines in Wangsu's “2+3” strategy, providing enterprises with integrated products and solutions, such as cloud strategy consulting, modernized application construction, generative AI, and enterprise-grade cloud hosting services. solutions based on AWS.

Cloudsway is committed to become a leading provider of hybrid cloud solutions,offering secure, efficient, and convenient cloud services to enterprises, helping them with digital and intelligent transformation, and boosting their operational efficiency.