GPU as a Service on Cisco UCS using Cisco Container Platform

© 2020 Cisco and/or its affiliates. All rights reserved.

Contents

© 2020 Cisco and/or its affiliates. All rights reserved.

Overview

Graphics processing units (GPUs) are commonly used for deep learning today. IT administrators want to offer GPU-as-a-service capabilities for their data scientist teams. IT teams don't necessarily want to assign dedicated GPU servers for each data science team because data science workloads have varying GPU requirements and don't always require access to the full set of GPUs in a single server. IT teams need deep-learning servers to be sharable to maximize utilization of this crucial resource.

In addition to these challenges, the deep-learning stack is difficult to install. Data scientists don't want to deal with installation, compatibility, and update problems.

The Cisco Unified Computing System™ (Cisco UCS®) and Cisco® Container Platform solve these challenges by setting a strong foundation for artificial intelligence (AI) projects that flexibly adapts to workload demands as they change over time.

Use case: Provision GPUs using multi-GPU-as-a-service capability

Challenge: Efficient GPU utilization

IT teams want deep-learning infrastructure to be used more efficiently. IT teams want to manage the infrastructure in a manner similar to the way that they manage other enterprise applications. IT teams need a solution that enables GPU servers to be shared between data scientists, but that also has the isolation, security, and reporting benefits of separated environments. IT teams have cooling and space requirements and want to reduce power consumption for their high-performance deep-learning servers.

For example, The Cisco UCS C480ML M5 Rack Server comes with eight NVIDIA Tesla V100 GPUs. There are many use cases in which customers don't need the entire C480ML server for a single deep learning job. A data scientist working with a small data set for model development purposes may want to iterate quickly by using only a single GPU.

Optimally, all GPUs across the servers are available as a pool that can be shared across various lines of business (LOBs).

Solution: Cisco Container Platform is a fully curated, lightweight container management platform for production grade environments and is delivered with Cisco enterprise-class support. It reduces the complexity of configuring, deploying, securing, scaling, and managing containers through automation. Based on upstream Kubernetes, Cisco Container Platform presents a user interface for self-service deployment and management of container clusters.

When users create container clusters in Cisco Container Platform, they can attach GPUs to the worker nodes (Figure 1). With Cisco Container Platform, heterogeneous GPUs can be pooled and shared between data science teams, leading to better GPU utilization and better return on machine learning infrastructure. Users can effectively put unused GPUs back into the pool so that the unused GPUs can be used by other data science teams who need them.

© 2020 Cisco and/or its affiliates. All rights reserved.

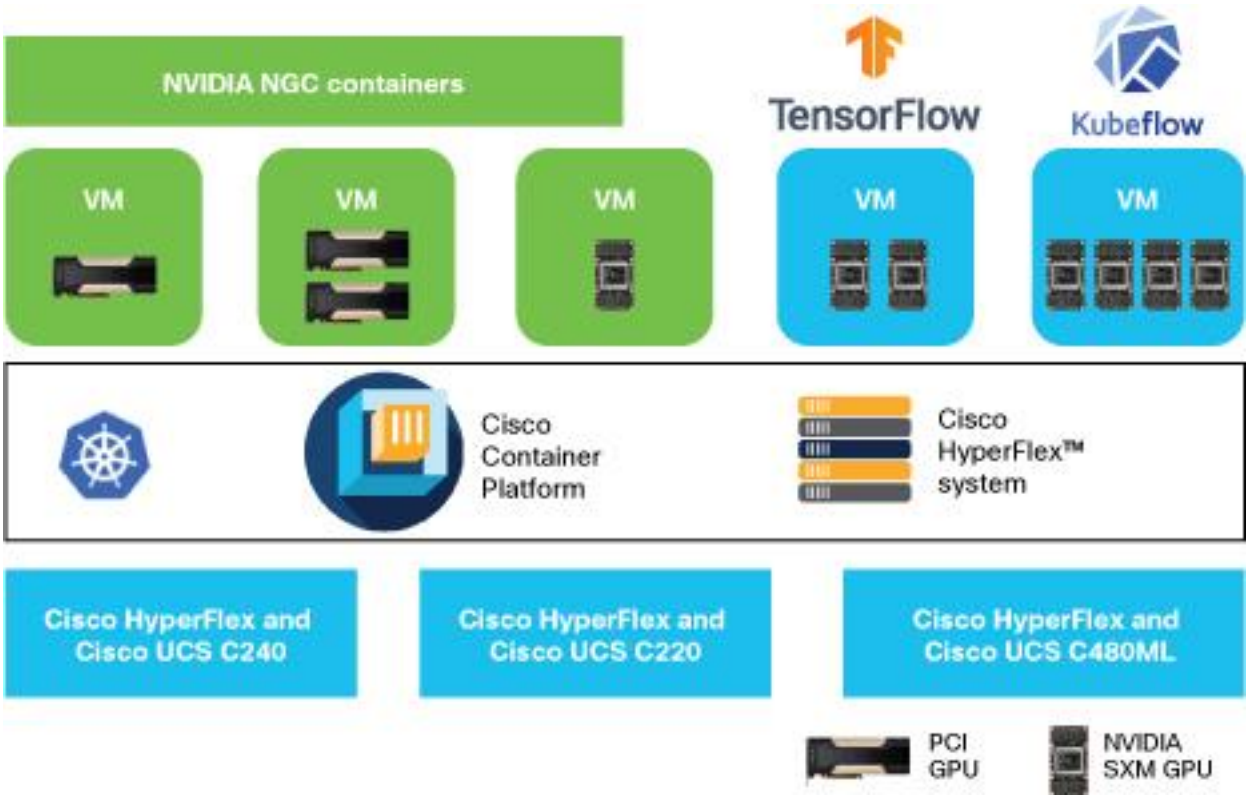


Figure 1.
With Cisco Container Platform, heterogeneous GPUs can be pooled and shared between data science teams

Use case: Automatically provision deep learning software stack

Challenge: Software stack management

Installing the deep learning software stack (Figure 2) can be extremely time consuming for several reasons. First, the tools often have extensive dependencies and interactions between software versions. The installation process is error prone because the installer has to install in specific ways numerous tools in which the installer may have little expertise. Moreover, the software stack must be provisioned and managed across every deep learning application. Finally, because CUDA releases for GPU drivers are frequent, teams must address update and compatibility challenges. Data scientists don't want to deal with installation, compatibility, and update problems.

In addition to these challenges, teams must always consume the latest and greatest software stack because it has a significant impact on deep-learning job performance. As the software tools evolve over time, teams must overcome any deployment challenges so they can make use of those enhancements.

© 2020 Cisco and/or its affiliates. All rights reserved.

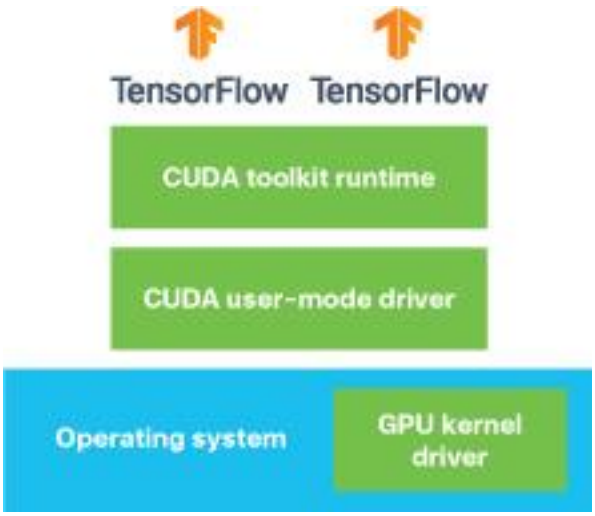


Figure 2.
Deep learning software stack with TensorFlow

Solution: Cisco Container Platform addresses these challenges. When you create a VMware vSphere cluster in Cisco Container Platform, the deep learning stack is provisioned for you. Through monthly releases, Cisco Container Platform keeps up with the frequent CUDA releases, so the deep learning stack that is part of the virtual machine will have the latest CUDA version. Cisco Container Platform also takes care of the associated dependencies of the various tools involved.

In Cisco Container Platform, users can create Kubernetes clusters and attach GPUs to the worker nodes (Figure 3). The worker nodes are preloaded with the latest NVIDIA GPU driver version and the latest CUDA toolkit version. For more details about how to provision vSphere clusters, see the “Creating Clusters on vSphere” section of the Cisco Container Platform User Guide (link).

© 2020 Cisco and/or its affiliates. All rights reserved.

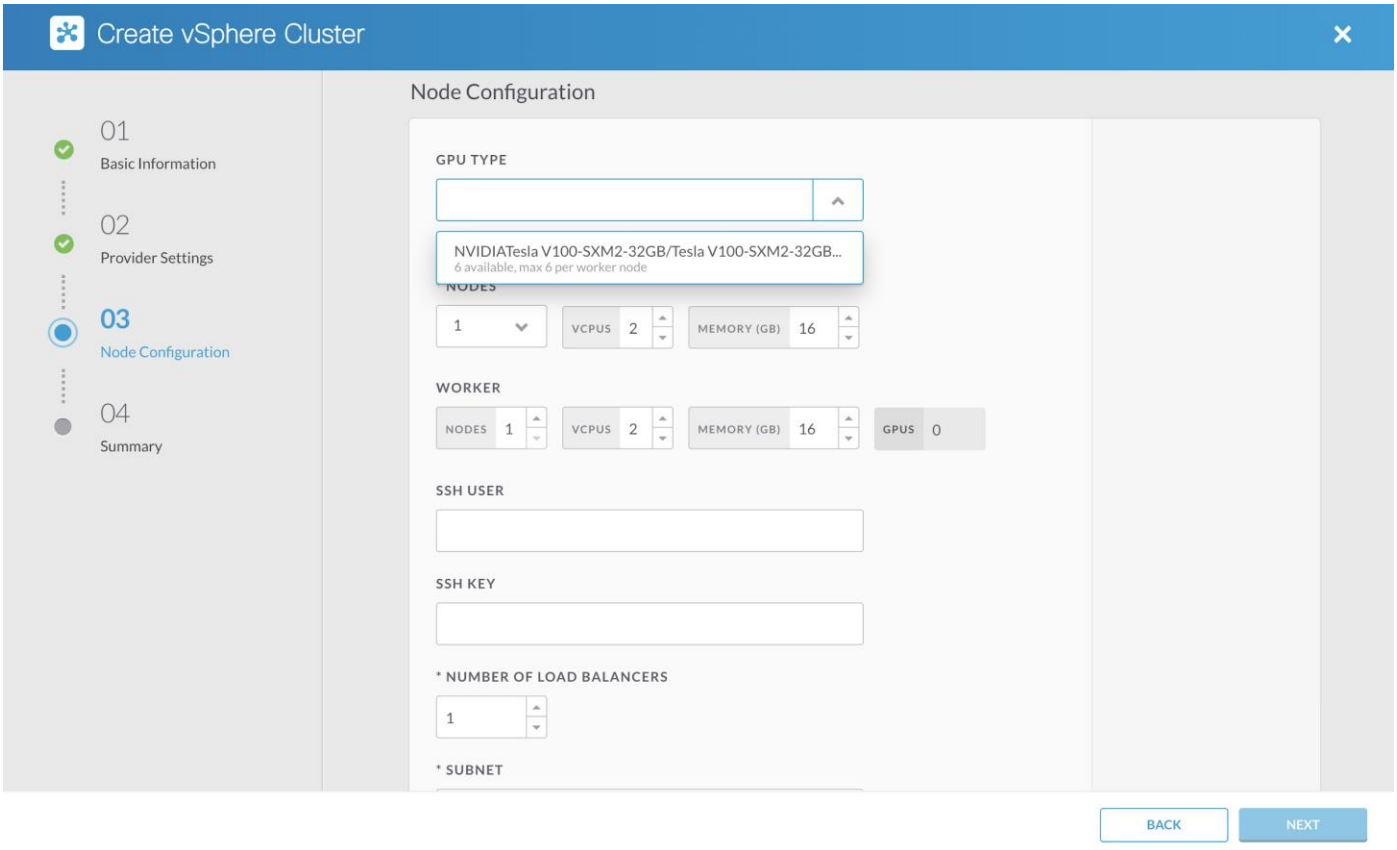


Figure 3.
Cisco Container Platform node configuration screen shows the available GPUs

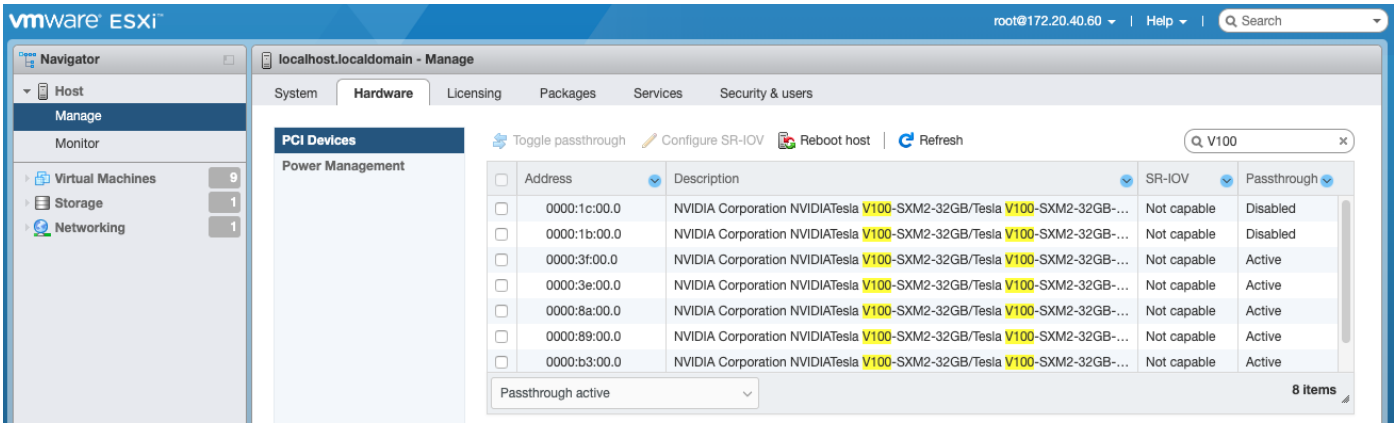
Figure 3 shows all the available GPU models under GPU Type. To make GPUs visible in Cisco Container Platform, as a one-time configuration the user will need to enable PCI passthrough for GPUs in the VMware vCenter Server.

Passthrough devices provide the means to more efficiently use resources and improve performance in your environment. Enabling PCI passthrough allows a virtual machine to use a host device as if the device were directly attached to the virtual machine and offers near-native performance.

Follow these steps to enable PCI passthrough:

1. Log in to vCenter Server.
2. Enter maintenance mode on the node on which the GPUs are installed. To enter maintenance mode, right-click the node and choose Maintenance Mode > Enter Maintenance Mode.
3. In a new browser window, log in directly to the VMware ESXi node.
4. Click Manage.
5. On the Hardware tab, click PCI Devices. A list of available passthrough devices appears.
6. Search for the PCI device and select the device you want to enable for passthrough. Click Toggle Passthrough.

© 2020 Cisco and/or its affiliates. All rights reserved.



7. Reboot the host to make the PCI device available for use.
8. After the reboot completes, verify that the node is out of maintenance mode.

Use case: Perform testing and development on premises with cloud-like simplicity

Challenge: Easy on-premises testing and development and cloud-like simplicity

Because of the simplicity of the cloud, data scientists often start in the cloud with a smaller data set and switch to on-premises resources when they need access to larger, more detailed granular data sets. Data scientists and data engineers seek cloud-like deployment simplicity and easy on-premises test and development processes. Today, bringing applications back on premises often requires rearchitecting the application.

Solution: Cisco Container Platform's multiple-GPU-as-a-service capability and preinstalled deep-learning stack address these challenges. You can also integrate Cisco Container Platform into your existing Cisco UCS computing cluster in a simple way.

Conclusion

With Cisco UCS and Cisco Container Platform, IT teams can provide data scientists with a containerized, scalable, and self-service AI platform.

With the Cisco Container Platform multi-GPU-as-a-service capability, you can:

- Provision GPUs using the multi-GPU-as-a-service capability
- Pool and share GPUs between data science teams simply by enabling GPUs in the nodes of a cluster
- Achieve higher GPU utilization
- Support the full range of deep-learning applications and workloads

Cisco UCS and Cisco Container Platform enable teams to get better return on their investments in machine-learning infrastructure. IT teams can provide a cloud-like experience for individual data scientists or teams while maintaining the same security benefits as if each individual or team had built its own separate deployment. The developer has the benefit of a customized environment, and IT has the benefit of the capability to manage a unified deployment across teams.

© 2020 Cisco and/or its affiliates. All rights reserved.

For more information

For more information about Cisco® solutions for AI/ML workloads visit: https://www.cisco.com/go/ai-compute

Cisco® Container Platform documentation can be accessed here:

https://www.cisco.com/c/en/us/support/cloud-systems-management/container-platform/series.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: https://www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

Printed in USA 220167.4 09/20

© 2020 Cisco and/or its affiliates. All rights reserved.