Executive Summary

A flexible AI processing environment makes it easy to incorporate AI into IT workflows. Many applications involve small to large AI models and different data batch sizes. Simultaneously, AI product development and IT deployment require a wide range of AI processing capabilities on demand. NVIDIA® AI Enterprise with Red Hat® OpenShift® running on Supermicro’s NVIDIA-Certified Systems provide the flexibility, cost effectiveness, and responsiveness to run AI workloads for developers and deployment.

This solution brief describes a Red Hat OpenShift reference architecture running on Supermicro’s systems to support a broad range of AI workloads. Three configurations use AMD CPUs, describing small, medium, and large deployments. The Red Hat OpenShift environment provides a highly available infrastructure to support developers running NVIDIA AI Enterprise. In addition, the infrastructure can easily scale with demand to support expansive deployments.
Red Hat OpenShift

Red Hat® OpenShift® is an enterprise-ready Kubernetes container platform built for an open hybrid cloud strategy. It provides a consistent application platform to manage hybrid cloud, multi-cloud, and edge deployments. Using GPU Operators and other Operators, Red Hat OpenShift enables easy setup and robust operations running NVIDIA AI Enterprise workloads.

NVIDIA AI Enterprise Software Suite

The NVIDIA AI Enterprise software suite includes AI tools and frameworks, cloud native deployment, and infrastructure optimization software to enable rapid AI development and deployment.

By providing minimal risk and a simple approach to integrating AI into the existing enterprise container environment, NVIDIA AI Enterprise enables an end-to-end software stack approach to start using AI in the enterprise. Enterprise developers can initially run small trials until they feel comfortable expanding to more extensive deployment. At that point, the solution is very scalable to deployment in multiple racks.

Enterprise Support Services

Using the NVIDIA AI Enterprise software suite, enterprise customers get enterprise-grade support for the entire system, from AI software to the virtualization and system hardware, including NVIDIA data center GPUs and network accelerators, optimized in Supermicro systems. As a solutions provider, Supermicro offers and supports the entire Supermicro systems with Red Hat OpenShift and NVIDIA AI Enterprise software.

Management & Security

Supermicro systems provide out-of-band and in-band monitoring. Using out-of-band IPMI and Redfish management, the health and operation of each server in the cluster can be managed. The servers also come with optional TPM 2.0 and Root of Trust security features.
Supermicro Reference Architecture for NVIDIA AI Enterprise running Red Hat OpenShift

Supermicro Reference Architecture for NVIDIA AI Enterprise and Red Hat OpenShift provides a scalable architecture. As a result, enterprise AI developers can quickly develop AI solutions to increase efficiency and enable new services, using AI models, pre-trained or otherwise, that can be easily deployed on Red Hat OpenShift’s orchestrated container environment. Enterprise support is available on the Supermicro systems that are NVIDIA-Certified, and Red Hat certified, including the entire software stack.

### Red Hat OpenShift

- SuperCloud Composer
  - Bootstrap/Ansible
- 3 x OpenShift Master Nodes
- 1 x AMD EPYC 7343 (16 core)
  - 64GB
  - 256GB M.2
  - 2 x 1TB SSD
  - Dual 10GbE
- 1 x AMD EPYC 7443 (24 core)
  - 256GB
  - 256GB M.2
  - 2 x 1TB SSD
  - Dual 10GbE
- 2 x AMD EPYC 7543 (32 core)
  - 1024GB
  - 2 x 1TB M.2
  - 2 x 4TB SSD
  - 4 x 25GbE
- 2 x AMD EPYC 7713 (64 core)
  - 2048GB
  - 2 x 1TB M.2
  - 2 x 1TB SSD
  - 8 x 200GbE

### Switch Infrastructure

- Data Switches
- Management Switches

### Runs NVIDIA AI Enterprise Applications

#### AI Worker Nodes (GPU Servers)

- **Small**
- **Medium**
- **Large**

### Red Hat OpenShift Master Nodes, SuperCloud Composer Node

<table>
<thead>
<tr>
<th>Server</th>
<th>Small AI Worker</th>
<th>Medium AI Worker</th>
<th>Large AI Worker</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS-1114CS-TNR 1U</td>
<td>AS-2114GT-DNR 2U (w/ 2 server nodes)</td>
<td>AS-4124GS-TNR 4U</td>
<td>AS-4124GO-NART+ 4U</td>
</tr>
<tr>
<td>Number of Servers</td>
<td>4</td>
<td>1 to 256</td>
<td>1 to 256</td>
</tr>
<tr>
<td>Server Configuration</td>
<td>1 x AMD EPYC 7343 (16 core)</td>
<td>Per Node, 1 x AMD EPYC 7443 (24 core)</td>
<td>2 x AMD EPYC 7543 (32 core)</td>
</tr>
<tr>
<td></td>
<td>64GB</td>
<td>256GB M.2</td>
<td>1024GB</td>
</tr>
<tr>
<td></td>
<td>256GB M.2</td>
<td>256GB M.2</td>
<td>2 x 1TB M.2</td>
</tr>
<tr>
<td></td>
<td>2 x 1TB SSD</td>
<td>2 x 1TB SSD</td>
<td>2 x 4TB SSD</td>
</tr>
<tr>
<td></td>
<td>Dual 10GbE</td>
<td>Dual 10GbE</td>
<td>4 x 25GbE</td>
</tr>
<tr>
<td>GPU</td>
<td>-</td>
<td>1 to 2 x A30 per node</td>
<td>1 to 8 x A100</td>
</tr>
<tr>
<td>BMC Switches (per 32 worker nodes)</td>
<td>-</td>
<td>2 x SSE-G3648B</td>
<td>2 x SSE-G3648B</td>
</tr>
<tr>
<td>Data Switches (per 32 worker nodes)</td>
<td>-</td>
<td>2 x SSE-X3648SR</td>
<td>2 x SSE-F3548SR</td>
</tr>
</tbody>
</table>
Example Applications

Here are example applications using containerized machine learning infrastructure. Specific customer solutions need to be adjusted to match customer needs.

<table>
<thead>
<tr>
<th>Number of Simultaneous Users</th>
<th>CPU Cores</th>
<th>System Memory</th>
<th>Storage</th>
<th>NVIDIA GPU</th>
<th>GPU System</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI Development – Smaller Jobs</td>
<td>Up to 12 (on 2 servers)</td>
<td>24</td>
<td>256GB</td>
<td>100TB</td>
<td>1 x A30 per server</td>
</tr>
<tr>
<td>AI Development – Medium Jobs</td>
<td>Up to 100</td>
<td>64 - 128</td>
<td>1024GB</td>
<td>200TB</td>
<td>4 x A100</td>
</tr>
<tr>
<td>AI Development – Large Jobs (multiple GPUs)</td>
<td>Multiple 100's</td>
<td>128</td>
<td>2048GB</td>
<td>400TB</td>
<td>8 x A100</td>
</tr>
<tr>
<td>AI Inference</td>
<td>Continual</td>
<td>128</td>
<td>1024GB</td>
<td>100TB</td>
<td>4 x A100</td>
</tr>
</tbody>
</table>

Conclusion

Supermicro NVIDIA-Certified Systems support NVIDIA AI Enterprise running on Red Hat OpenShift to enable AI developments and delivery to run small to large AI workloads. The reference architectures with specific small, medium, and large configurations provide a robust framework for customers to start using NVIDIA AI Enterprise, running in a robust orchestrated container environment provided by Red Hat OpenShift.

Supermicro offers these as integrated solutions, including systems, software, and support. Please call your Supermicro representative for more information.

References

1. NVIDIA AI Enterprise
2. NVIDIA-Certified Systems
3. Supermicro GPU Servers supporting the Ampere architecture
4. Red Hat OpenShift

© 2022 Copyright Super Micro Computer, Inc. All rights reserved. Supermicro, the Supermicro logo, Building Block Solutions, We Keep IT Green, SuperServer, Twin, BigTwin, TwinPro, TwinPro2, SuperDoctor are trademarks and/or registered trademarks of Super Micro Computer, Inc. All other product names, logos, and brands are property of their respective owners.