WHITE PAPER

SUPERMICRO® ARTIFICIAL INTELLIGENCE / MACHINE LEARNING READY SOLUTION

Meet all your AI/ML application needs with Supermicro optimized GPU server solutions

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The rapid expansion of Artificial Intelligence (AI) and Machine Learning (ML) applications into all aspects of business and everyday life is generating an explosion in Big Data. This advancement comes with a price, however the need for frequent training, retraining, and hyperparameter tuning longer times than are now the norm. In addition, AI/ML also requires enormous amounts of processing power for model training.

Compute-intensive Machine Learning algorithms take extended times to complete when using hardware without acceleration features, resulting in overall poor application performance and reduced ROI. With this growing demand for AI/ML applications, enterprise data centers accommodate budget, space, and IT resources, while also shortening this training time bottleneck.

With no end in sight to expanding datasets, nor to compute and memory-intensive applications, data center managers must rapidly secure the necessary processing horsepower and matching AI/ML platforms to satisfy their business needs. With the proper selection of vendors, these hardware-plus-application solutions will help users to identify trends and patterns, improving throughput and training times, thus leading to a positive cycle of advancement. This paper describes one such AI/ML solution from Supermicro.

Super Micro Computer, Inc.
980 Rock Avenue
San Jose, CA 95131 USA
www.supermicro.com

KEY CUSTOMER BENEFITS

- **Pre-validated reference architectures**
- **Certified components**
- **Scale-out to multiple racks**
- **TCO Optimization for best performance / watt /$ /ft2**
- **Start as professional by leveraging expertise, support and service**
- **Supports TensorFlow, Kubeflow, Kubernetes**
- **Sharable resources with higher utilization**

SUPERMICRO AI / ML SOLUTION

GENERAL DESCRIPTION

As Artificial Intelligence and Machine Learning solutions become more accessible and more mature, global organizations will come to realize the value that these solutions can deliver to solve the advanced business challenges.

The Supermicro AI/ML solution features a best-in-class hardware platform with the enterprise-ready Canonical Distribution of Kubernetes (CDK) and software-defined storage capabilities from Ceph. The solution through its reference architecture integrates network, compute, and storage. The recommended starting implementation includes a single rack with capabilities to scale to many racks as required.

AI / ML REFERENCE ARCHITECTURE

The reference architecture is ready to deploy end-to-end AI / ML solution that includes AI SW stack, orchestration, and containers. The optimized reference design fits machine learning training and inference applications. The architecture on a high-level comprises software, network switches, control, compute, storage, and support services.

The reference design shown in Figure 1 contains two data switches, two management switches, three infrastructure nodes that act as foundation nodes for MAAS / JUJU, and six cloud nodes. It is built on the Kubernetes platform and provides Canonical hardened packages for Kubernetes containers and Ceph. Kubeflow provides a machine learning toolkit for Kubernetes.

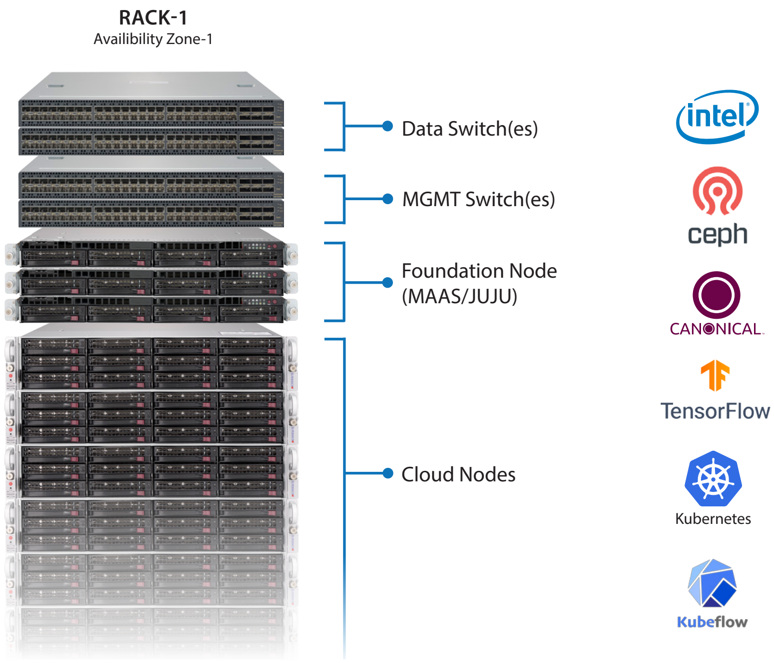


Figure 1. *Supermicro AI / ML Reference Architecture*

SOLUTION CONFIGURATION

- **Up to 216 compute cores**
- **Up to 3072 GB system memory**
- **Up to 36 TB storage**
- **Up to 40 Gbe data networking**
- **19U height**
- **High performance caching utilizing NVMe flash storage**

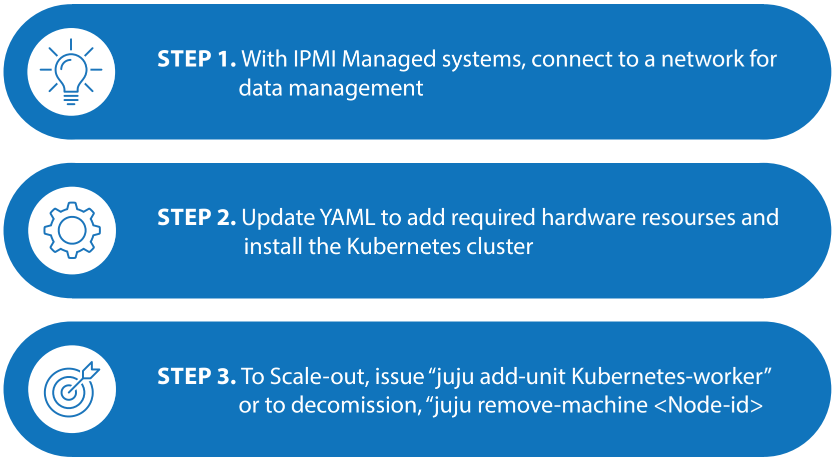
The key highlights include a certified reference architecture with validated and tested components, racks that scale-out from one to many, green Resource Saving servers for the Cloud saving hundreds of dollars per server, industry-leading performance, optional consulting and support services, and an optimized solution for certified Intel AI partners.

This solution is built and validated on Supermicro server families: Ultra and BigTwin™. It also utilizes Supermicro ethernet switches such as SSE-G3648B (management/IPMI traffic switch), SSE-X3648S (10 GbE data network switch), SSE-F3548S (25 GbE data network switch), and SSE-C3632S (40 GbE data network switch). The solution is optimized for performance and designed to provide the highest levels of reliability, quality, and scalability.

HOW SUPERMICRO SOLUTION IS DEPLOYED

The supermicro solution can be deployed by IT administrators, Data scientists, and DevOps. The deployment process is explained below in Fig 2 and Fig 3.

DEPLOYMENT FOR IT ADMIN



STEP 1. With IPMI Managed systems, connect to a network for data management

STEP 2. Update YAML to add required hardware resourses and install the Kubernetes cluster

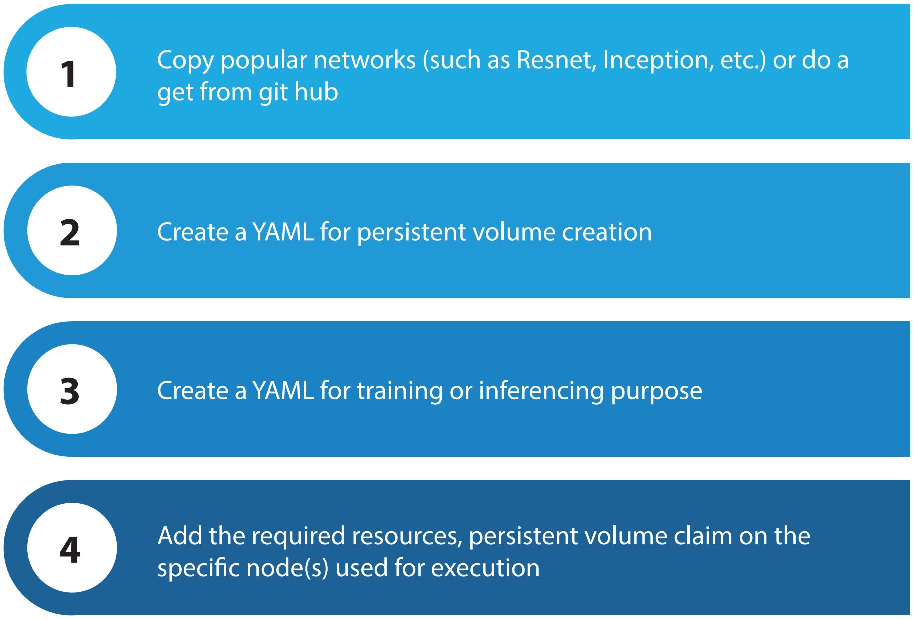
STEP 3. To Scale-out, issue "juju add-unit Kubernetes-worker" or to decomission, "juju remove-machine <Node-id>

Figure 2. *Steps for deploying AI / ML solution for IT ADMIN*

Figure 2 explains the solution deployment process for IT administrators. IPMI is used for connecting to the network for data management. Once it is connected, steps 2 and 3 allow hardware resource additions and usage of JUJU commands for scaling and, later, for decommissioning out-of-service equipment.

DEPLOYMENT FOR DATA SCIENTISTS AND DEVOPS

Figure 3 explains the solution deployment process for data scientists and DevOps. It outlines for users the different steps involved in utilizing the ML solution and popular networks for performance-driven Machine Learning training and inferencing.



1 Copy popular networks (such as Resnet, Inception, etc.) or do a get from git hub

2 Create a YAML for persistent volume creation

3 Create a YAML for training or inferencing purpose

4 Add the required resources, persistent volume claim on the specific node(s) used for execution

Figure 3. *Steps for deploying AI / ML solution for Data scientists and DevOps*

HOW SUPERMICRO SOLUTION FLOW WORKS

Kubeflow is an open-source project dedicated to providing easy-to-use Machine Learning (ML) resources on top of a Kubernetes cluster. With Canonical MAAS and Juju in place, setting up a Kubernetes/Kubeflow environment becomes relatively simple. The Juju controller makes it easy to deploy the Kubernetes cluster on a single node and multi-node cluster based on the infrastructure supported. Kubeflow eases the installation of TensorFlow, and with the addition of Supermicro systems containing the appropriate accelerators (Intel MKL) it can provide accelerated performance for the submitted ML jobs. Lastly, Prometheus is used for event monitoring and alerting

KUBEFLOW

- Built from an upstream source and clean Kubeflow
- Security updates spanning from Kernel to Kubernetes
- Guaranteed upgrades with the option to consume latest version
- Robust encryption for all control plane components
- Readily available training, certification, and support services

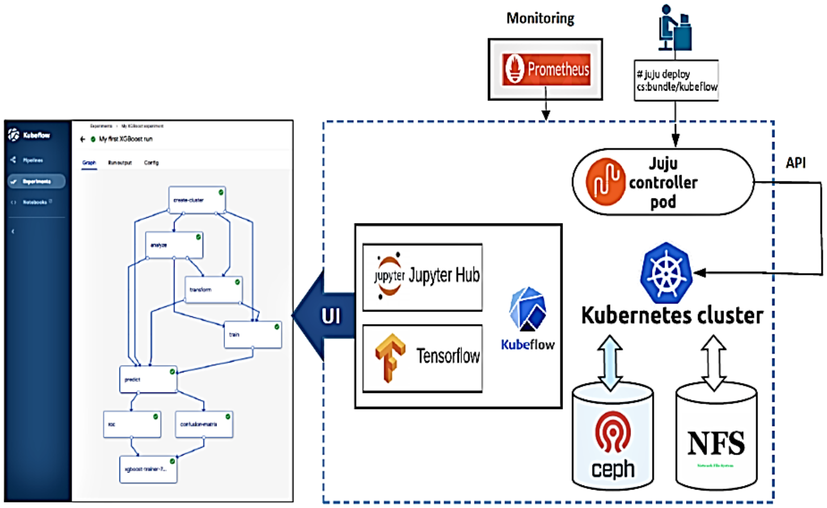


Figure 4. *Pictorial description of Supermicro AI / ML flow*

SUPERMICRO SYSTEM

- **Second generation Intel® Xeon® Scalable processors**
- **Intel optimized models and packages**
- **Canonical Ubuntu OS**
- **Canonical distribution of Kubernetes**
- **Ceph**
- **Supermicro Network Switches**

SYSTEM DETAILS

SOLUTION ARCHITECTURE

Supermicro AI / ML configurations are optimized for Cloud and scale-out architectures, based on high-density computing resources and scale-out software-defined storage.

The Supermicro solution systems feature the latest second-generation Intel Xeon scalable processors along with Intel optimized ML models and packages. It further utilizes Ubuntu OS, Kubernetes, Kubeflow, Ceph storage, and Supermicro network switches to ensure the solution provides a scale-out infrastructure, better performance, throughput, and faster training times.

NETWORK ARCHITECTURE

As data grows exponentially on the order of terabytes and petabytes, a network infrastructure requires a reliable scale-out storage solution. Ceph is the preferred storage system to achieve that stable, robust network infrastructure. The highly scalable fault-tolerant storage cluster transforms the network into a high-performance infrastructure by handling users' data throughput and transaction requirements.

Additionally, the AI/ML solution comprises dual management switches (IPMI and Kubernetes), dual data switches, three infrastructure nodes, and six cloud nodes. The management switch supports 1Gbps connectivity and is common to all three networking options, which are 10Gbps, 25Gbps, and 40Gbps. In addition, the data switch supports 10Gbps, 25Gbps, and 40Gbps as well. The 10 GbE and 40 GbE data switches require a Cumulus OS, whereas the 25 GbE data switch requires a Supermicro (SMIS) OS.

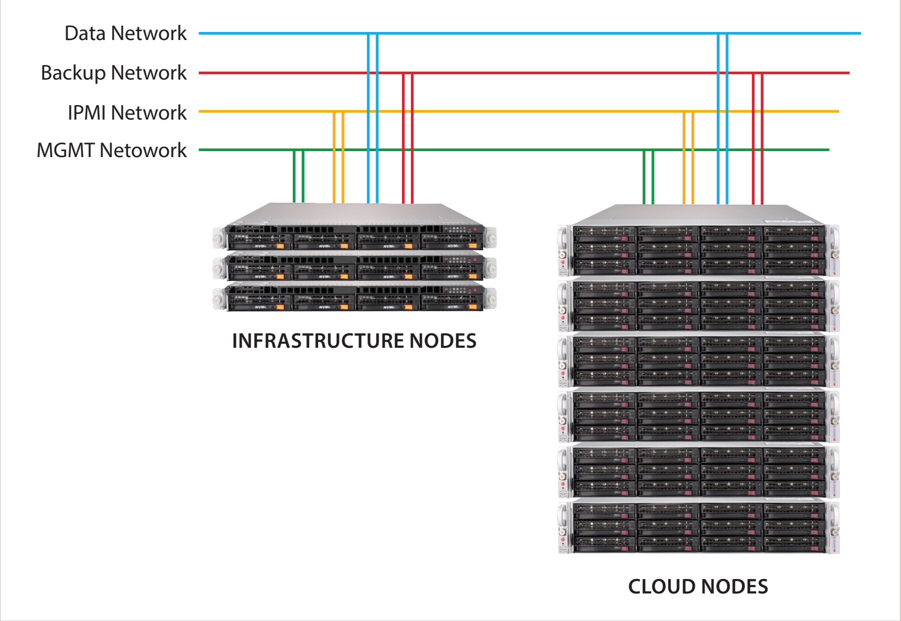


Figure 5. *Network Architecture diagram*

CONFIGURATION

AI / ML components include three Supermicro Ultra infrastructure nodes (SYS-6019U-TN4RT), six Ultra cloud nodes (SYS-6029U-TR4T), and twelve cloud node data disks (U.2 NVMe drives). The configuration also includes Ubuntu Advantage Advanced and Ubuntu Kubernetes Discoverer licenses. Optional services include Datacenter design validation and bootstrap services, Supermicro rack integration services, and Supermicro onsite support.

MACHINE LEARNING PART DESCRIPTION	SKU	QTY
COMPONENTS USED		
Infrastructure Node	SYS-6019U-TN4RT	3
Cloud Node	SYS-6029U-TR4T	6
Cloud Node Data Disks	U.2 NVMe Drives (2 TB)	12 (2 per node x 6)
	HDS-IUN2-SSDPE2KX020T8	
SOFTWARE LICENSES		
Ubuntu Advantage Advanced (3 Years)	SVC-CNC-SVR-AS	9
Infrastructure Node	SVC-CNCFC-FOB	1
SERVICES (OPTIONAL)		
Data Center design, validation and boot-strapping services		9
Supermicro rack integration service		1
Supermicro onsite support		9

Figure 6. *Configuration describing ML part, SKU, Quantity*

BENCHMARK RESULTS

The second-generation Intel® Xeon® scalable processors showed approximately 25% higher performance results over previous generation systems in both training and inferencing with CNN benchmark testing.

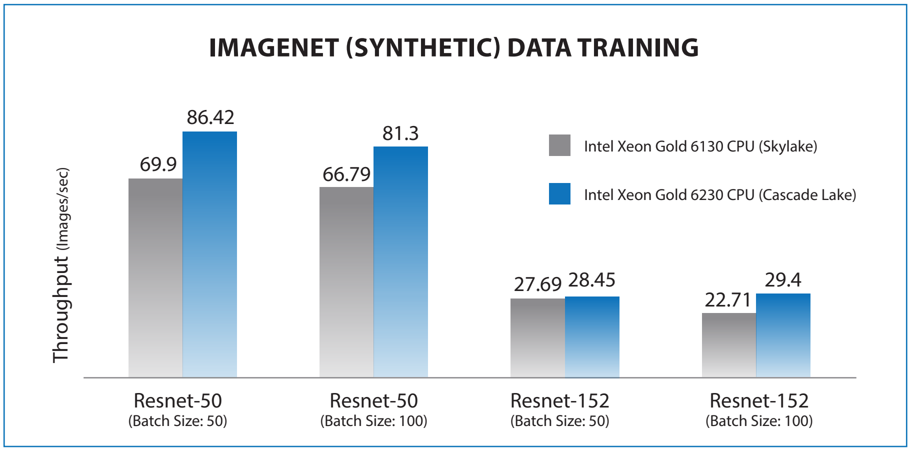


Figure 7. *Training benchmark comparison of Skylake and Cascade lake , using SYS-6029U-TR4 with 2 Intel Xeon Gold 6130 CPUs*

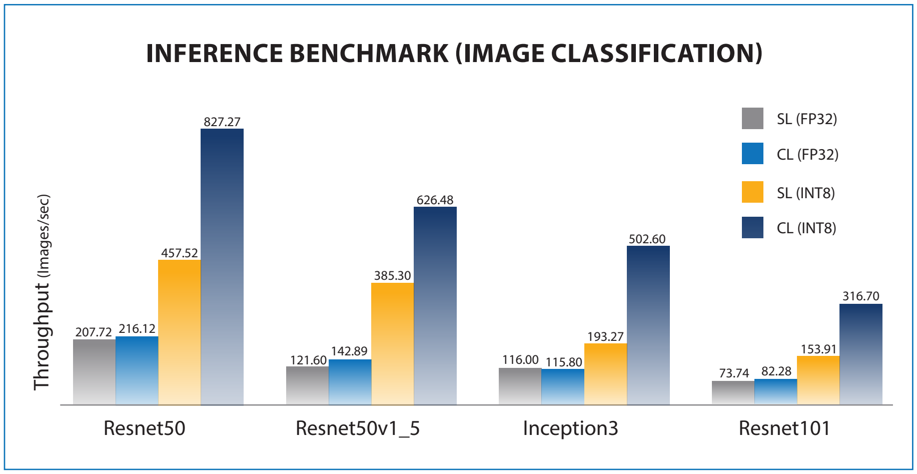


Figure 8. *Inference benchmark comparison of Skylake and Cascade lake, using SYS-6029U-TR4 with 2 Intel Xeon Gold 6130 CPUs*

BigTwin with Intel Xeon Platinum 8260L scalable processor showed improved throughput for both training and inference.

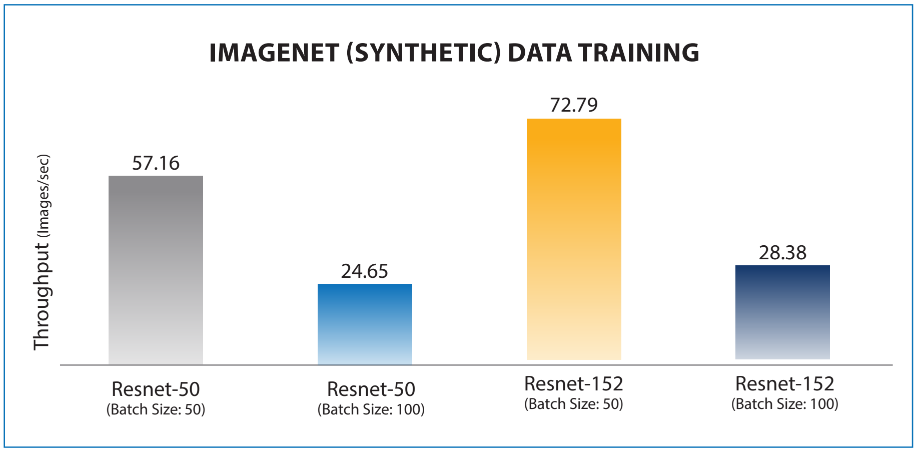


Figure 9. *Training benchmarks using BigTwin (SYS-2029BT-HNC0R) with 2 Intel Xeon Platinum 8260L CPUs*

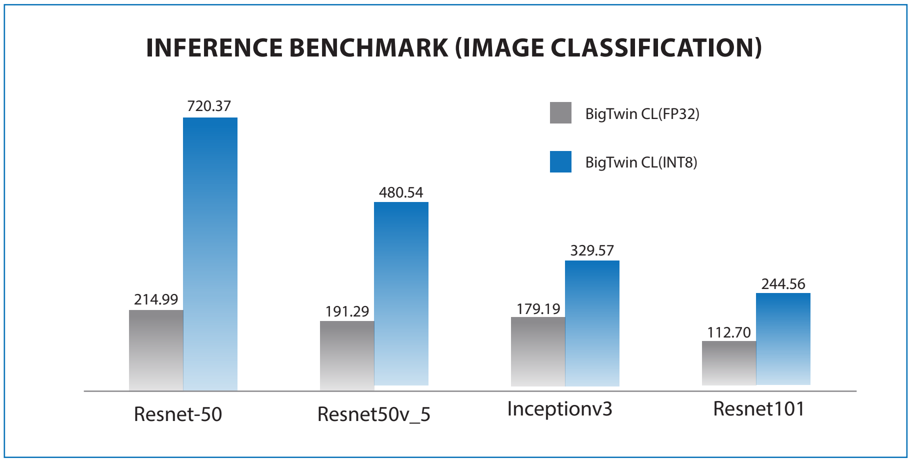


Figure 10. *Inference benchmarks using BigTwin (SYS-2029BT-HNC0R) with 2 Intel Xeon Platinum 8260L CPUs*

SUPPORT AND SERVICES

Canonical and Supermicro in a joint partnership provide enterprise support for the Canonical Distribution of Kubernetes. This partnership offers a discovery and design service, scaling the infrastructure to the required size and specifications and helping customers gain access to a global pool of knowledge and expertise.

CONCLUSION

The Supermicro AI / ML ready end-to-end solution is easy to deploy and handles low-level implementation details so developers, data scientists, and IT administrators can be more productive. This certified solution serves as a perfect platform for machine learning training and inferencing needs. It provides competitive throughputs and reduced training/inferencing times by accelerating compute and memory intensive AI / ML application workloads.

Supermicro, with its optimized server/storage/networking hardware and high-performance compute power, along with the matching AI / ML infrastructure, can identify trends and patterns from Big data and take appropriate action for machine learning workloads for better throughput and training times, leading to successful business results.

To learn more, please visit **https://www.supermicro.com/en/solutions/tensorflow-canonical**

ABOUT SUPER MICRO COMPUTER, INC.

Supermicro® (NASDAQ: SMCI), the leading innovator in high-performance, high-efficiency server technology is a premier provider of advanced server Building Block Solutions® for Data Center, Cloud Computing, Enterprise IT, Hadoop/Big Data, HPC and Embedded Systems worldwide. Supermicro is committed to protecting the environment through its "We Keep IT Green®" initiative and provides customers with the most energy-efficient, environmentally-friendly solutions available on the market.

www.supermicro.com

Printed in USA ♻ Please Recycle 01_AI-ML-Ready_2020_01-6