Reference Design
September 2024

Intel® Gaudi® 3 AI Accelerator Cluster Reference Design

Accelerate your AI solutions with the latest Intel Gaudi 3 accelerator-based systems—built for scale and expandability with all-Ethernet-based fabrics and support for a wide range of industry AI models and frameworks

Contents

1. Introduction

AI's growing popularity is driven by improved usability and a broadening selection of vertical solutions that are tailored for nearly every industry, such as healthcare, legal, transportation, manufacturing, energy, and more. However, the high cost of AI infrastructure and concerns about being locked into vendor-specific solutions can slow AI adoption. Fortunately, the market now offers more open industry solutions such as those based on the Intel® Gaudi® AI accelerator product line.

Intel Gaudi accelerators are architected for deep learning (DL) and Generative AI, excelling at large language model (LLM) and multi-modal model training and inferencing. Intel Gaudi AI accelerator-based clusters are purpose-built for running DL workloads of all sizes across multi-tenant data centers. Intel Gaudi accelerators have proven to be a viable alternative to the competition in Generative AI compute capability, pricing, energy efficiency, and market availability.[1]

Most enterprise AI solutions for training and inference require multiple accelerators or GPUs to be interconnected across multiple chassis and often employ several racks of compute, network, and storage equipment. While most AI GPU clusters have been deployed on proprietary fabrics like Nvidia's NVLink or InfiniBand, Ethernet-based solutions are gaining momentum.

This document is designed to help enterprise IT operations, developers, and infrastructure leaders specify and deploy multi-node AI infrastructure using Intel Gaudi 3 AI accelerator-based systems.[2]

2. Reference Design Overview

Figure 1 represents a high-level view of key components in the Intel Gaudi 3 AI accelerator cluster Reference Design. Intel Gaudi 3 accelerators support a wide range of industry AI models and frameworks and the breadth of options is beyond the scope of this paper. This Reference Design defines an "AI-ready" cluster infrastructure including required node-level software, system management software, and provisioning and management tools. It also recommends specific compute, network, storage, and control plane hardware. A more comprehensive summary of software components is included in Section 5.

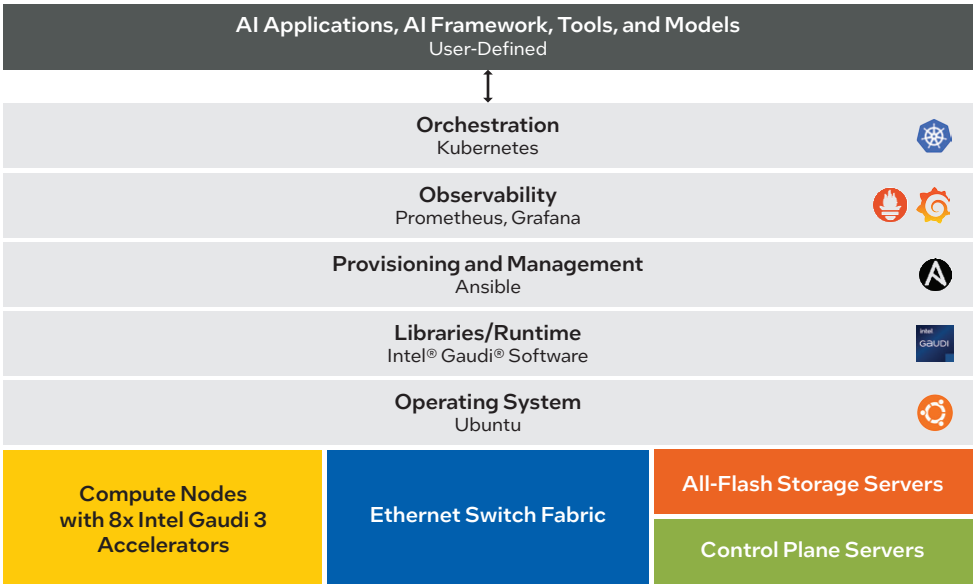

Figure 1. Overview of Intel® Gaudi® 3 AI accelerator Reference Design requirements.

Compute Nodes

Several leading OEMs offer AI computing systems featuring Intel Gaudi 3 accelerators. For this Reference Design, the compute element (referred to as a node) is an air-cooled server chassis featuring eight Intel Gaudi 3 accelerators. This compute node is purpose-built to enable rapid development, training, and deployment of large Generative AI models. It can be configured into scalable clusters of many nodes to handle vast data sets for AI and is well-suited for LLMs, multi-modal models, recommendation engines, and neural network applications.

Intel Gaudi 3 accelerators have proven to be a viable alternative to the competition in Generative AI compute capability, pricing, energy efficiency, and market availability.[3] One key advantage is in networking: Intel Gaudi accelerators are designed for excellent scalability. Every accelerator is interconnected to every other accelerator within a node in an all-to-all configuration. Each accelerator integrates 24 RDMA over Converged Ethernet (RoCE) ports, of which 21 are used for scale-up connectivity within an eight-card universal baseboard, and three are used for scale-out connectivity. To scale outside a node, 800 Gbps OSFP Ethernet switches are used for the accelerator fabric. These switches directly interconnect all the AI accelerators in the cluster and deliver up to 25.6 Tbps throughput.[4]

Cluster Networking Overview

Generative AI clusters typically have three primary high-speed networks that connect the compute, storage, and control plane servers and a fourth network for out-of-band management.

Compute

A group of Intel® Xeon® processor-based accelerator nodes connects to a dedicated accelerator fabric. The accelerator fabric is an Ethernet network used by the Intel Gaudi AI accelerators to directly communicate with other Intel Gaudi AI accelerators in other nodes. This is implemented as a three-ply full Clos Ethernet fabric using OSFP4x 200 Gbps links to provide 800 Gbps per connector.

Storage

A group of Intel Xeon processor-based server nodes with SSDs and a scalable storage application connects through a dedicated full Clos 100 Gbps Ethernet fabric, providing high-performance storage for the compute nodes.

Control plane

The control plane consists of Intel Xeon processor-based server nodes that facilitate and streamline the deployment, management, and operation of a scalable AI cluster. Control plane functions include provisioning, configuration, network setup, upgrades, monitoring, system health, performance metrics, and user management. The control plane fabric is an Ethernet fabric that connects the control plane servers to the compute nodes and the rest of the cluster infrastructure. It is implemented as a full Clos 100 Gbps Ethernet fabric.

Management

The out-of-band management fabric connects the control plane servers, switch management ports, baseboard management controllers (BMCs), and power distribution units (PDUs). It also handles all switch management communication. It is implemented as a layer 2 fabric with 25 and 1 Gbps connections with 25 Gbps aggregation links.

Figure 2 depicts a cluster with independent storage and control plane fabrics.

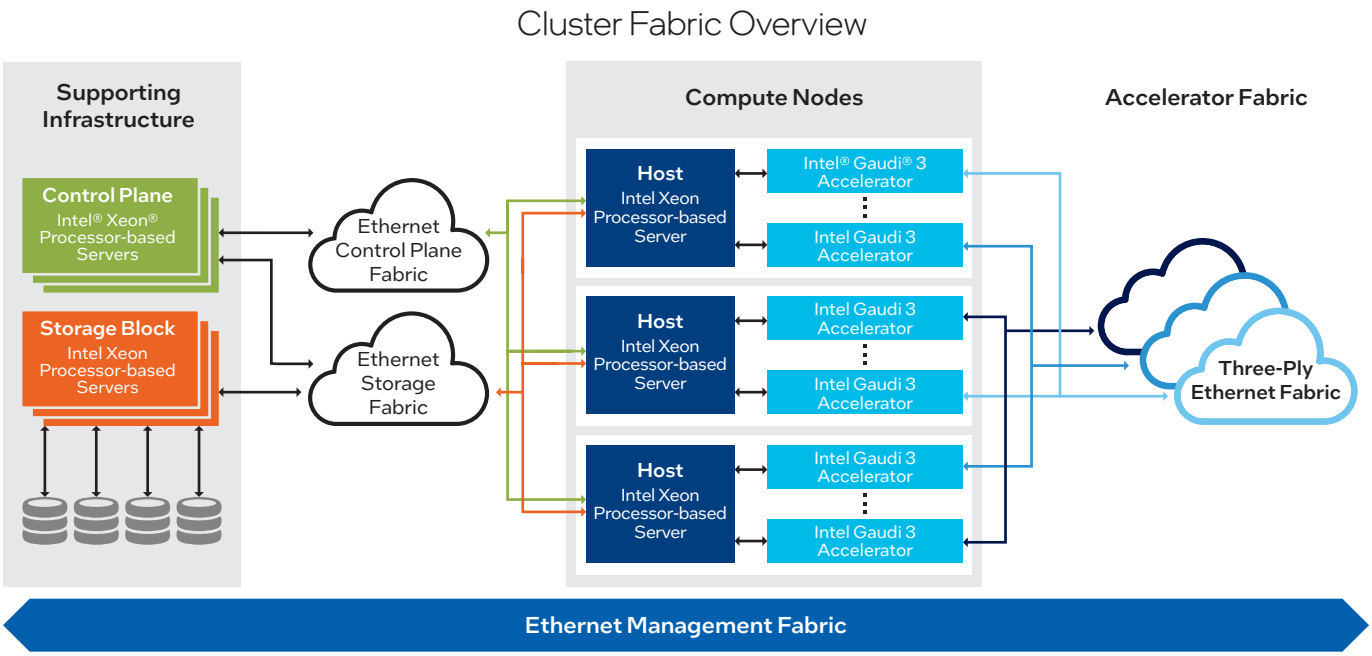

Figure 2. Cluster design using independent storage and control plane fabrics.

Scaling Intel® Gaudi® 3 AI Accelerator Clusters

This Reference Design illustrates how to build a 32-node cluster. Defining modular blocks that can be quickly and repeatedly instantiated and tied together by a core infrastructure is key to building at scale.

Even larger clusters with 2K, 4K, or 8K Intel® Gaudi® 3 AI accelerators are possible by scaling the industry-standard Ethernet-based accelerator fabric. There are many ways to achieve this. The simplest way is to increase the radix of the spine switches. Systems featuring 16K or even 32K accelerators are possible with commercially available Ethernet switch products. Of course, all the fabrics and storage will need to be increased proportionally, and a few additional control plane servers will likely be needed as more demand is placed on operational services. Use the following scaling table to guide cluster design.

Cluster Size (Nodes)	Number of Intel® Gaudi® 3 AI Accelerators	FP8 AI Compute[a]	Number of Spine Switches (64 Ports)	Number of Leaf Switches (64 Ports)	Recommended Storage Bandwidth
256	2048	3.76EF	24	48	0.8–1 TB/s Read 0.5 TB/s Write
512	4096	7.52EF	48	96	1.5–2 TB/s Read 1 TB/s Write
1024	8192	15EF	96	192	3–4 TB/s Read 2 TB/s Write

[a] Peak projected performance varies by use, configuration, and other factors. Results may vary.

3. Expandable 32-Node Cluster

System Architecture

Figure 3 depicts the high-level system diagram for an expandable 32-node cluster comprising 256 individual Intel Gaudi 3 AI accelerators. This section includes a bill of materials (BOM) for the cluster infrastructure and configuration details.

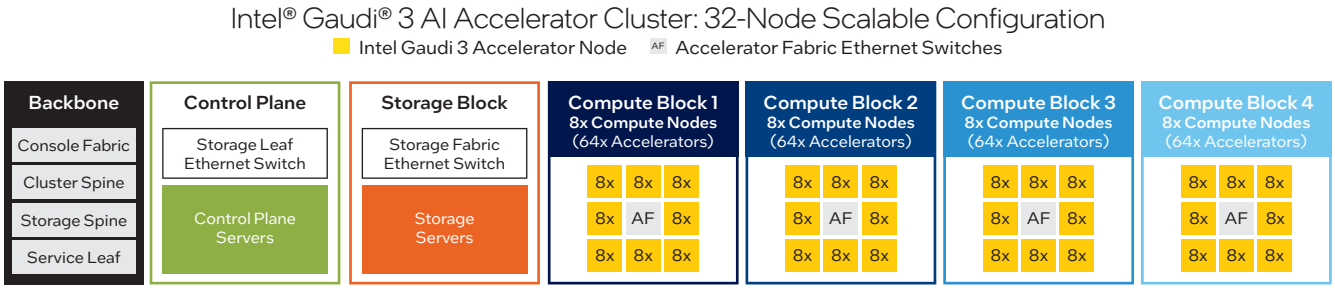

Figure 3. High-level design of a 32-node Intel® Gaudi® 3 AI accelerator cluster.

Cluster Bill of Materials

The BOM for this Reference Design reflects prototypical and validated components. In order to give customers infrastructure options, additional components such as switches and storage appliances are being evaluated and validated.

Table 1. Cluster with Accelerator Interconnect BOM

Qty	Description	Part Number
32	Intel® Gaudi® 3 AI accelerator node	OEM Gaudi 3 system
12	32x 800 Gbps OSFP Arista Ethernet switch	Arista DCS-7060PX5-64E-F
96	Credo 800G active Cu 3.0m	CAC83X301A1A-C1-HW
96	Credo 800G active Cu 3.5m	CAC835301A1A-C1-HW
384	Intel® Silicon Photonics 800G DR8 OSFP optical transceiver with MPO optical connector, 500m reach	SPTSRS3PNCDF 800G OSFP
48	Intel Silicon Photonics 400G DR4 QSFP-DD optical transceiver with MPO optical connector, 500m reach	MMID 99A694 - 228083
96	100G AoC cables	Length TBD based on layout
216	Anixter LYNN Elec MPO fiber cables	Length TBD based on layout
1	Digi ConnectPort 32-port serial expander	EZ32-A100-US
70	Cat 6 RJ45 management cables	Length TBD based on layout
18	Cat 6 RJ45 switch console cables for LTS-32	Length TBD based on layout
8	25G AoC cables	Length TBD based on layout
1	Arista 7050X4, 24x 100 Gbps + 8x 400 Gbps, cluster	7050CX4-24D8
1	Quad storage leaf DCS-7358X: 64x100G QSFP28; 16x400G QSFP-DD	DCS-7358X-BND-F, 4xDCS7358-16C, 4xDCS7368-4D
4	Arista 7010 48x 1G	DCS-7010TX-48-F

Note that corresponding accelerator fabric switches are listed in the accelerator spine BOM (see Table 2).

Table 2. Accelerator Spine BOM

Qty	Description	Part Number
6	32x 800 Gbps QSFP Arista Ethernet switch	Arista DCS-7060PX5-64E-F
1	Arista 7010 48x 1G	DCS-7010T-48-F
1	LTS-32	N/A
2	25G AoC cables	Length TBD based on layout

Table 3. Storage Block BOM

Qty	Description	Part Number
4 - 32	Storage servers with sufficient bandwidth and capacity per user workload	OEM storage server
1	WEKA license, capacity-based	N/A
3	Arista 7050X4, 48x 100 Gbps + 8x 400 Gbps	7050CX4M-48D8

Table 4. Control Plane BOM

Qty	Description	Part Number
18	Infrastructure servers	OEM control plane server
4	Arista 7060DX4 32x 400 Gbps	N/A
7	Arista 7050SX3 48x 25 Gbps	N/A
6	Arista 7050X4, 24x 100 Gbps + 8x 400 Gbps	7050CX4-24D8
4	Arista 7010 48x 1G	DCS-7010T-48-F
1	LTS-32	
30	100G AoC cables	Length TBD based on layout
36	Credo 400G active Cu	
48	Anixter LYNN Elec 400 Gbps fiber cables	Length TBD based on layout
8	25G AoC cables	Length TBD based on layout

Rack Configuration Overview

The cluster rack elevation consists of four sections: the compute blocks, the storage block, the control plane, and the fabric spines. In the elevation in Figure 4, the control plane and fabric spines are combined into a common set of racks to conserve space. The graphic shows 32 Intel Gaudi accelerator nodes and the leaf switches for all the required fabrics. The storage block is a set of servers running a scalable storage application. Storage can be scaled in both capacity and performance simply by adding more SSDs or storage nodes. Scaling to a larger cluster is extremely simple—just add additional compute and storage blocks and expand the spines to accommodate the additional leaf switches.

Each rack is a standard size: 42U 600mm x 1200mm. It is equipped with 2x APC NetShelter Rack PDU, Switched Metering, 3-phase 415V 60A with 21 C19 & 21 C13 outlets, and 34.6KW/rack pre-power factor correction.

	Control Plane			Storage Block		Compute Blocks							
						Block 1			2	3	Block 4		
U	B01	B02	B03	B04	B05	06	07	08	...	...	15	16	17
	600 mm	600 mm	600 mm	600 mm	600 mm	600 mm	600 mm	600 mm	...	...	600 mm	600 mm	600 mm
42	7010	7010	7010	7010			7010					7010	
41													
40													
39	**Storage Leaf**	**Storage Leaf**	**Storage Leaf**										
38	7050CX4-24	7050CX4-24	7050CX4-24										
37	**Cluster Leaf**	**Cluster Leaf**	**Cluster Leaf**	7050CX4M-48	7050CX4M-48		7358 **Storage Ply**					7050CX4M **Cluster Ply**	
36	7050CX4-24	7050CX4-24	7050CX4-24										
35	**CP1**	**CP2**	**CP3**										
34	Type 2 Server	Type 2 Server	Type 2 Server										
33				2U Storage Server	2U Storage Server								
32	Type 2 Server	Type 2 Server	Type 2 Server										
31				2U Storage Server	2U Storage Server								
30	Type 2 Server	Type 2 Server	Type 2 Server										
29				2U Storage Server	2U Storage Server								
28						Intel® Gaudi® Accelerator Node	Intel Gaudi Accelerator Node	Intel Gaudi Accelerator Node	...	...	Intel Gaudi Accelerator Node	Intel Gaudi Accelerator Node	Intel Gaudi Accelerator Node
27	Type 3 Server	Type 3 Server	Type 3 Server	2U Storage Server	2U Storage Server								
26													
25	Type 3 Server	Type 3 Server	Type 3 Server	2U Storage Server	2U Storage Server								
24													
23	Type 3 Server	Type 3 Server	Type 3 Server	2U Storage Server	2U Storage Server								
22													
21				2U Storage Server	2U Storage Server	Intel Gaudi Accelerator Node	7060PX5-64	Intel Gaudi Accelerator Node	...	...	Intel Gaudi Accelerator Node	7060PX5-64	Intel Gaudi Accelerator Node
20							**Ply 3**					**Ply 3**	
19	**Management Backbone**			2U Storage Server	2U Storage Server		7060PX5-64					7060PX5-64	
18	7050SX3	7050SX3					**Ply 2**					**Ply 2**	
17	**Console Fabric**		**Ply 3**	2U Storage Server	2U Storage Server		7060PX5-64					7060PX5-64	
16	7010 Digi	LTS-32	7060PX5-64				**Ply 1**					**Ply 1**	
15	**Cluster Spine**			2U Storage Server	2U Storage Server								
14	7060DX4	7060DX4	7060PX5-64			Intel Gaudi Accelerator Node	Intel Gaudi Accelerator Node	Intel Gaudi Accelerator Node	...	...	Intel Gaudi Accelerator Node	Intel Gaudi Accelerator Node	Intel Gaudi Accelerator Node
13	**Storage Spine**		**Ply 2**	2U Storage Server	2U Storage Server								
12	7060DX4	7060DX4	7060PX5-64										
11	**Service Leaf**			2U Storage Server	2U Storage Server								
10	7050CX4-24	7050CX4-24	7060PX5-64										
09			**Ply 1**	2U Storage Server	2U Storage Server								
08			7060PX5-64										
07				2U Storage Server	2U Storage Server								
06			7060PX5-64										
05				2U Storage Server	2U Storage Server								
04													
03				2U Storage Server	2U Storage Server								
02													
01													
RACK	1	2	3	4	5	6	7	8	...	...	15	16	17

CP = Control Plane

Figure 4. Cluster rack elevation.

Compute Rack Elevation Overview

The 32-node cluster has four groups of eight compute nodes (see Figure 4 on the previous page). Each group shares three Arista 7060PX5-64 switches used as Intel Gaudi AI accelerator interconnect leaf switches. The eight systems are arranged in three racks, in a 3-2-3 configuration, encircling the three leaf switches. This minimizes the length of 800 Gbps active Cu cables between the compute nodes and the Arista 7060PX5-64 switches. Racks are configured with either two or three compute nodes per rack. The Intel Gaudi AI accelerator interconnect leaf switches are placed in the two-system rack. Each group of eight systems also has a 1 Gbps management switch for connecting BMCs, PDU, and switches into a management fabric.

Each compute node has 2x 100 Gbps connections to the storage fabric and 1x 100 Gbps connection to the control plane fabric. The storage leaf switch is an Arista 7358 configured to have 64x 100 Gbps ports and 16x 400 Gbps ports. Two connections of 100 Gbps per compute node support network-attached storage, while the 16x 400 Gbps connections are connected to the storage spines. The cluster leaf switch is an Arista 7050CX4M-48 and is also shared across the compute blocks. One 100 Gbps connection is routed to each compute node, and the 400 Gbps connections are routed to the cluster spine. The last shared function is a serial port expander. The serial port expander provides an RS-232 serial connection to each network switch as an alternative control path. This path is important to support minimal-touch data centers when switch infrastructure errors or corruption occur. The Digi LTS-32 should be connected to an independent data center access network.

4. Design Considerations

Network Fabric

Scale-out cluster sizes beyond what can be supported with a single-level switch require building a non-blocking fabric or Clos fabric. A simple Clos fabric has two levels of switches: leafs and spines. The leaf switch ports are split; 50% are connected to Intel Gaudi AI accelerators and 50% to spine switches. The spines connect only to leaf ports; therefore, using a same-size switch, there are twice as many leafs as spines in a full Clos configuration.

The management fabric, or out-of-band fabric, provides isolated access to core data center infrastructure. The three main areas are BMCs, network switches, and PDUs. All of these devices are connected to the 1 Gbps management switch associated with each eight-system group. Although there are many ways to configure this network, three VLANs are implemented in the four management switches to isolate BMC, PDU, and switch access.

The control plane fabric is built as a full Clos fabric using an Arista 7050CX4M-48 as a leaf across the compute blocks. One 100 Gbps link is connected to each compute node.

The storage fabric is similar to the control plane fabric in that it is a full Clos fabric but with two 100 Gbps connections to the leaf. The leaf switch is an Arista 7358 with four blades of 16x 100 Gbps ports and four blades of 4x 400 Gbps ports. A pair of 100 Gbps connections from blade 1 (ports 1 and 2) are connected to the compute node in rack 1 U09, and incrementally through the 32 systems.

The accelerator fabric is built as a full Clos fabric similar to both the control plane and storage fabrics, but this fabric is based on 800 Gbps connections configured as 4x 200 Gbps connections. Each compute node has six 800 Gbps connections; each pair contains a single 200 Gbps connection from each of the eight Gaudi 3 accelerators in the system. Since the Intel Gaudi architecture can support three independent plys (for cluster scale-out), this is the most efficient topology. As described earlier, there are three Arista 7060PX5-64 leaf switches surrounded by eight compute nodes.

Storage Considerations

Storage is the most customer-dependent aspect of the cluster. For the 32-node cluster Reference Design, the storage block is based on 32 2U servers with a minimum of 8x NVMe SSDs and 2x 100 Gbps network connections per server. This configuration can deliver about 20 GB/s large-block random read performance per server or approximately 0.5 TB/s random read per cluster. Storage capacity is very user- and workload-dependent and can be configured both by the size and the number of SSDs used per storage server.

Each storage server has 2x 100 Gbps links to a leaf switch at the top of each storage block rack. Like all the installed systems, the storage servers are connected to redundant power. The BMCs are linked to the 7010 management switches; the 100 Gbps connections to the storage leafs are also linked to the 7010 management switches.

Control Plane Considerations

The control plane consists of several functions. Control plane servers run the control stack, management fabric backbone, control plane fabric spine, storage spine, service leaf, and console fabric. They also provide highly reliable storage for the control plane.

The control plane configuration can be flexible depending on customer requirements, such as the level of fault tolerance required. For smaller clusters, as few as two control plane servers are required. For larger configurations, such as the 32-node cluster in this Reference Design, up to six control plane servers may be needed. In a configuration including three sets of six control plane servers, a triple-redundant system for the control plane is recommended:

- Each control plane server resides in an independent rack with dedicated leaf switches on each fabric.
- The control plane servers are connected to a redundant power supply.
- The BMC is connected to the 7010 management switches. The 100 Gbps connections to both cluster and storage leafs in each rack are also linked to the 7010 switches.

The minimum requirements for control plane servers are as follows:

- 2x 5th Gen Intel Xeon processors
- 1 TB DRAM
- 1x 2 TB SSD for storage
- 4x 200 GbE connections per client
- 2x 100 GbE connections per client

5. Software

The software stack for an Intel Gaudi 3 AI accelerator cluster is designed to streamline the cluster deployment and operation. The deployment and operations include cluster provisioning, configuration, upgrades, monitoring, and visualization; network configuration; performance and system health monitoring; and user management. Table 5 lists the software components for the 32-node cluster Reference Design.

Table 5. Software Components

Category	Function	Tool
Virtualization, Containerization, and Orchestration	Container foundation	Kubernetes (K8s)
	K8s cluster setup	Kubespray
	K8s device plugin	Gaudi K8s plugin
	K8s virtualization	KubeVirt
Authentication and Key Management	Key storage and management	HashiCorp Vault
	AAA, LDAP, DNS, key management	Freeipa
Network	K8s networking and security	Calico
	Load balancing	MetalLB
	Network source of truth/IP address management (IPAM)	NetBox
Storage	Storage	WEKA
Provisioning and Management	Provisioning and configuration management	Ansible
	Network Time Protocol (NTP)	Chrony
	MLOps	Kubeflow
Monitoring	Real-time metrics	Prometheus
	Visualization	Grafana
	Centralized log management	Loki
	Log aggregation	Promtail
Operating System	Libraries, compiler, runtime	Intel® Gaudi® software
	Node OS	Ubuntu Linux

6. Summary

As Generative AI becomes steadily embedded in every industry, Intel Gaudi 3 AI accelerators can provide the high performance that is necessary for rapid training and inference at the scale and efficiency sought by companies of all sizes. Using open-source software and industry-standard Ethernet connections, these accelerators can help drive down the cost of AI solutions. This Reference Design simplifies cluster deployment by providing prescriptive guidance for selecting and configuring system components. Using this reference design, enterprises can quickly innovate and take their AI journey to the next level.

Begin exploring Intel Gaudi accelerators in the **Intel® Tiber™ Developer Cloud**.

[1] Intel® Gaudi® 3 accelerator training vs. NVIDIA H100; average performance measured across models: Llama 2 70B & Llama 3 8B. Gaudi 3 inference vs. NVIDIA H100; average performance projected across multiple models, multiple configurations: Llama2 7B & 70B, Falcon 180B, https://github.com/NVIDIA/TensorRT-LLM/blob/main/docs/source/performance/perf-overview.md. Intel results obtained in September 2024. Results may vary. NVIDIA H100 GPU (900 GB/s closed NVLink connectivity) vs. Intel Gaudi 3 accelerator (1200 GB/s open standard RoCE).
Source: https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi3.html

[2] The recommended system components in this Reference Design have been tested and validated with Intel® Gaudi® 2 AI accelerators. However, testing with Intel Gaudi 3 AI accelerators is not yet complete.

[3] See endnote 1.

[4] Source: https://www.arista.com/assets/data/pdf/Datasheets/7060X5-Datasheet.pdf