Ethernet: The Future of Enterprise AI Connectivity

AI everywhere needs Ethernet everywhere

Current State of AI Connectivity

AI connectivity needs differ based on where AI is used

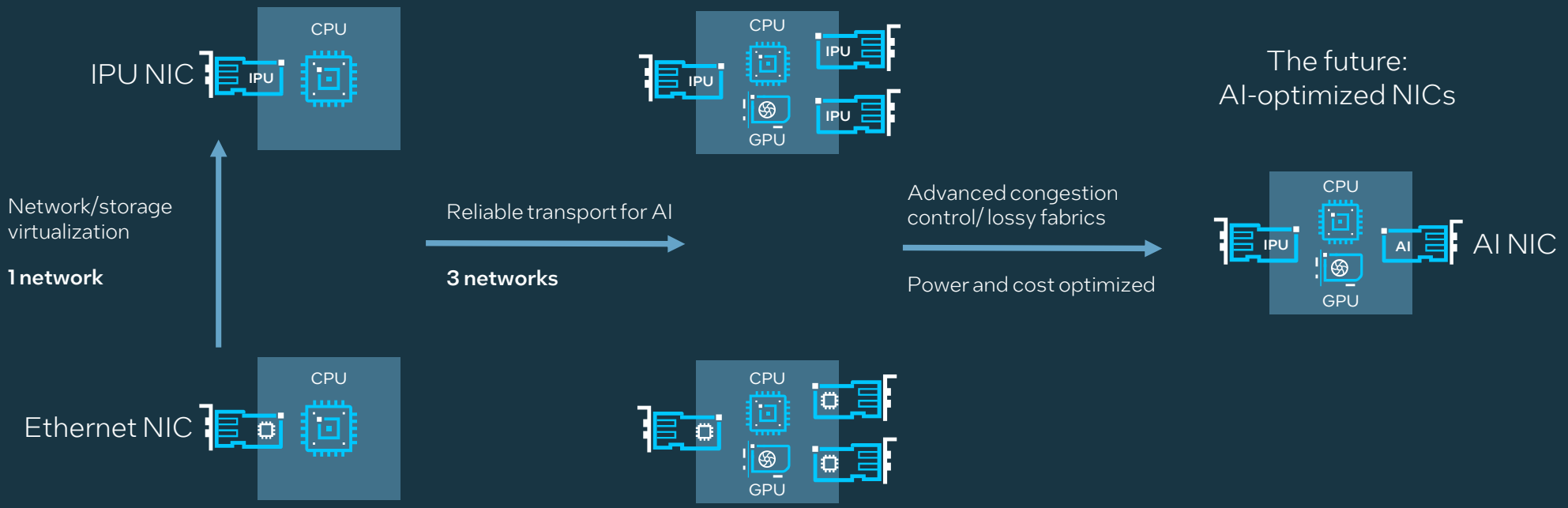
- CPU and/ or GPUs
- Inference vs. Tuning vs. Training

InfiniBand is technology of choice by main GPU vendor

- Only available from 1 vendor
- Closed software
- Expensive
- Different operations stack vs rest of network

High-end cloud training	PODs and MegaPODs Very large clusters (10K GPUs)
Mid-range cloud/ enterprise AI	Multinode rack Cluster of 16-256 servers
Mid/high-end enterprise AI	GPU-based accelerated node
Edge AI	CPU-based accelerated node

Server design evolution



Networks used in AI: Today's solution

Scale-out network (Network 2)

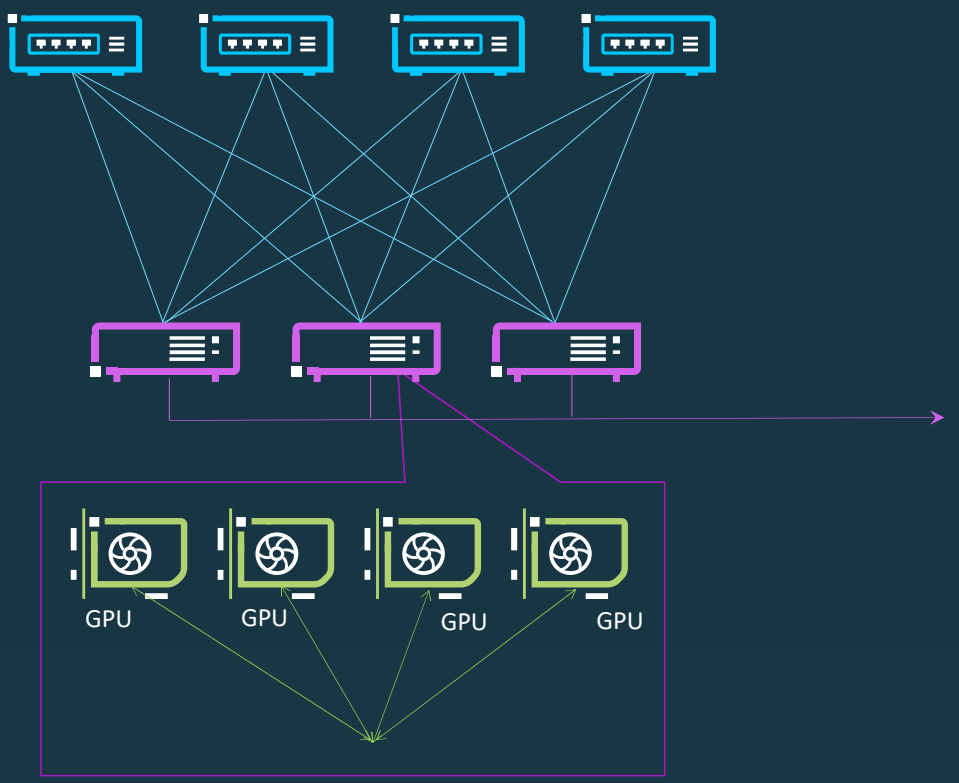
- Network semantics (RDMA/MPI)
- Targets scaling for data parallel
- Large average transactions (8KB-> MBs)
- Needs traditional DC networking (multitenancy)
- Solutions: Ethernet, InfiniBand

Front-end network (Network 1)

Standard data center network on Ethernet

Scale-up network (Network 3)

- Network (RDMA/MPI) and Mem (LD/ST) semantics
- Higher BW scaling for model parallel
- Multiple GPUs in each node
- Average transaction > 8KB
- Solutions: Nvlink, XeLink, Ethernet (Intel), CXL



Intel's Vision for AI Connectivity

Open ecosystem

Enable an open ecosystem for AI networking solutions that eliminates single-vendor tecnologies.

Standard protocols

Replace InfiniBand with open Ethernet based protocols from UEC. Replace proprietary Scale-Up links with an open, standard based protocol on high speed SERDES.

Partnerships

Build partnerships to enable complete, open solutions. Differentiate offerings via cost, power, system integration, and open software.

Connectivity solutions for AI clusters

Ethernet Adapter

1. Standard Ethernet server adapter
2. Broad workload deployments including AI
3. Reliable transport: RoCE v2
4. AI cluster scale: <1,000 GPU nodes per fabric

IPU Adapter

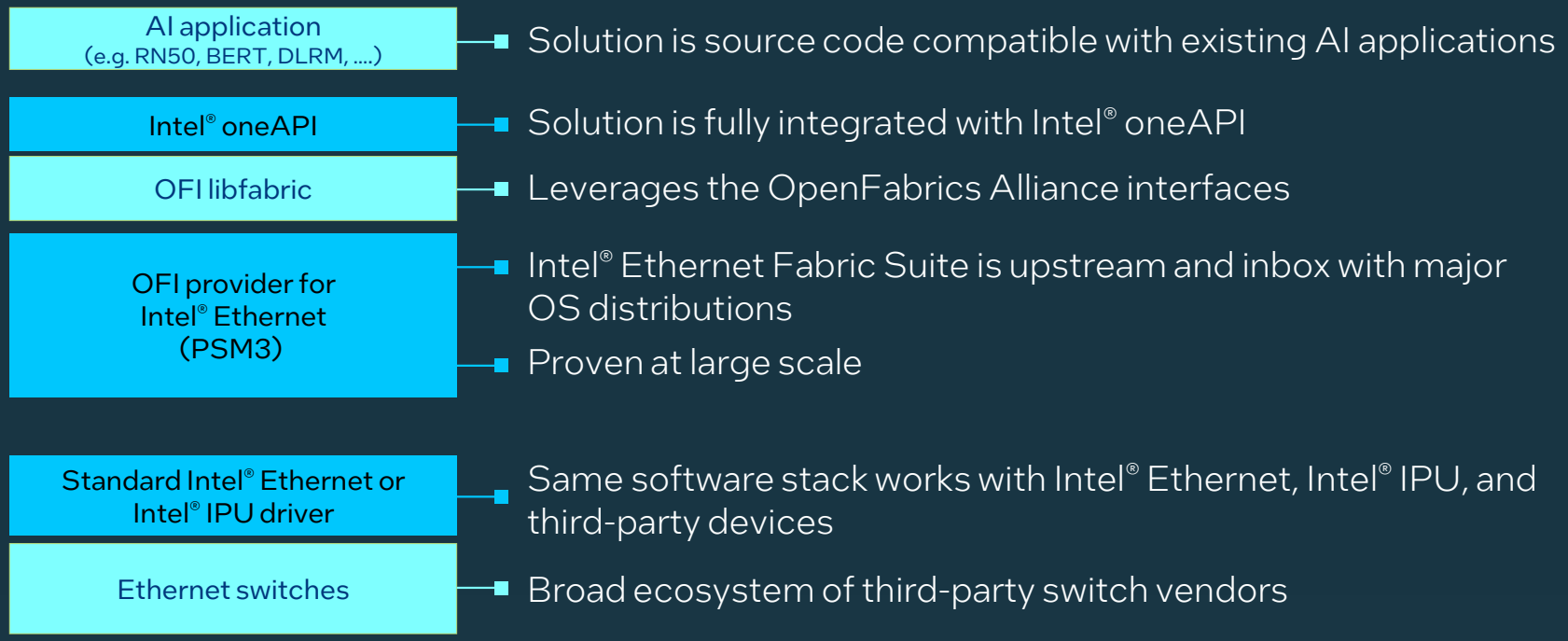
1. Premium server adapter that's highly customizable with premium features
2. Broad workload deployments including AI
3. Reliable transport: Falcon RT and RoCE v2
4. AI cluster scale: Small to large clusters

AI NIC

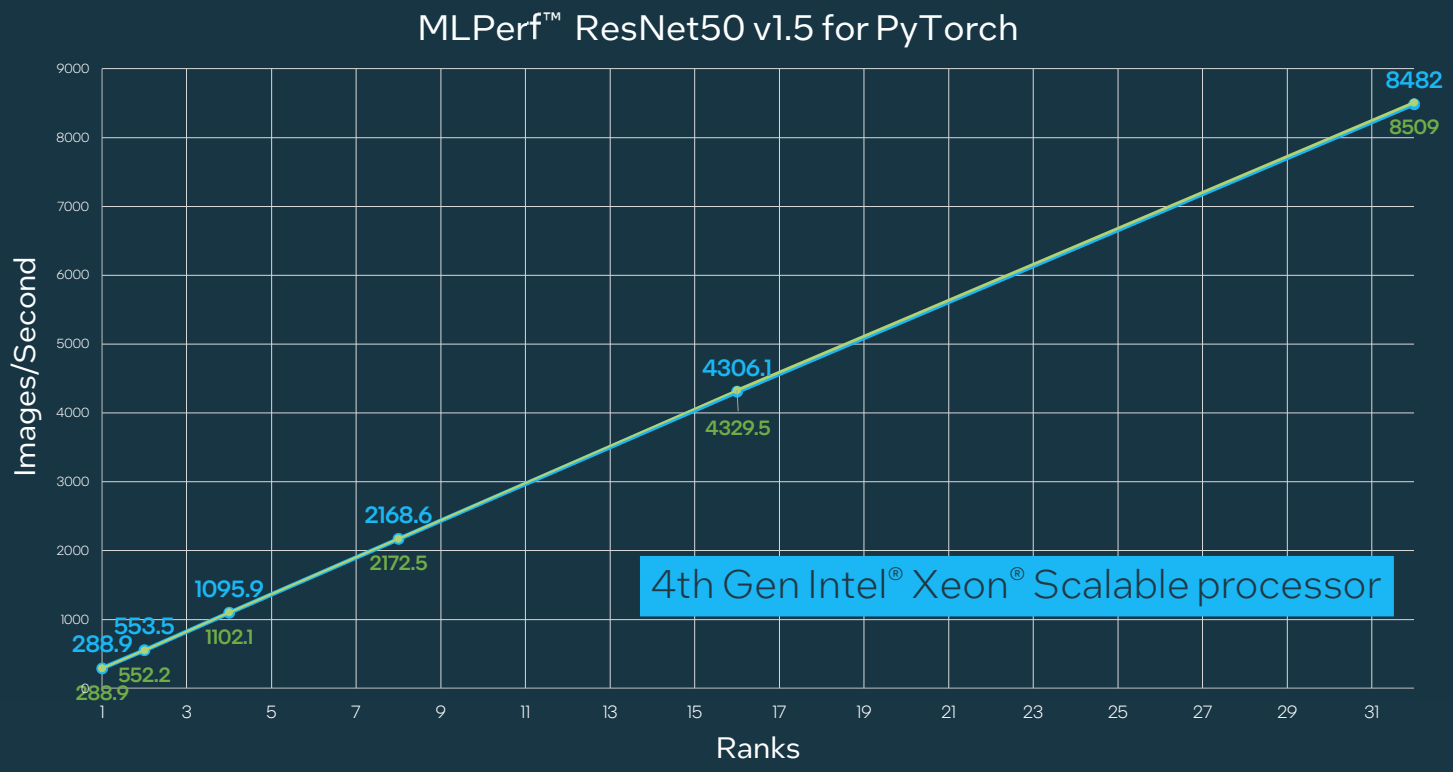
1. AI-optimized server and NIC design
2. Optimized for AI workloads
3. Reliable transport: AI-optimized RoCE v2
4. AI cluster scale: 64K+ GPU nodes per fabric

It takes an open ecosystem

Solution includes Intel® oneAPI, Intel® Ethernet Fabric Suite, third-party switches



Intel® Ethernet AI training performance



- Dual Intel® Ethernet Network Adapters E810-CQDA2 with RoCE v2 and Intel® Ethernet Fabric Suite
- Nvidia HDR200 (InfiniBand)

2 oneCCL ranks per node, --batch-size=102 --model=resnet50 --device=cpu --num-warmup-batches=20 --num-batches-per-iter=10 --num-iters=200

* See backup for workloads and configurations. Results may vary.

Enabling scaling beyond RoCE v2

Commonly used for clusters and storage

Scaling becomes an issue for standards-based RoCE v2

Problem: Customers tell us they want the same technology for clusters and storage at every scale

Developed by Google, released publicly through OCP

Addresses the Ethernet RoCE v2 scale issue

Intel® IPU E200 Series is the world's Falcon compatible devices

We're not alone. The industry agrees.

Ultra Ethernet Consortium

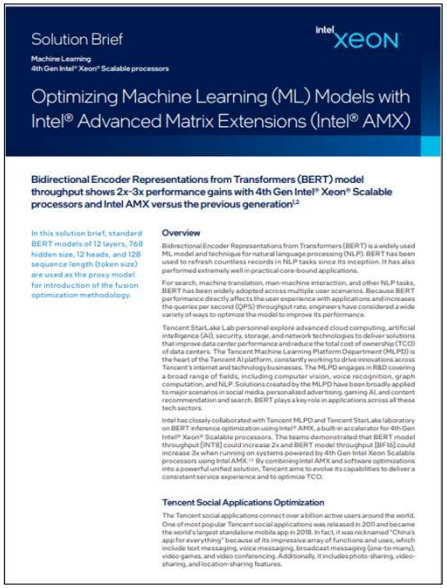
Our mission

Deliver an Ethernet-based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI and HPC at scale

Ethernet Adapters for AI/HPC solutions

Intel® Ethernet 800 Series

Machine Learning Inference Optimization. Natural Language Processing - Tencent


Solution Brief

Machine Learning
4th Gen Intel® Xeon® Scalable processors

Optimizing Machine Learning (ML) Models with Intel® Advanced Matrix Extensions (Intel® AMX)

Bidirectional Encoder Representations from Transformers (BERT) model throughput shows 2x-3x performance gains with 4th Gen Intel® Xeon® Scalable processors and Intel AMX versus the previous generation[1,2]

In this solution brief, standard BERT models of 12 layers, 768 hidden size, 12 heads, and 128 sequence length (token size) are used as the proxy model for introduction of the fusion optimization methodology.

Overview

Bidirectional Encoder Representations from Transformers (BERT) is a widely used ML model and technique for natural language processing (NLP). BERT has been used to refresh countless records in NLP tasks since its inception. It has also performed extremely well in practical core-bound applications.

For search, machine translation, man-machine interaction, and other NLP tasks, BERT has been widely adopted across multiple user scenarios. Because BERT performance directly affects the user experience with applications and increases the queries per second (QPS) throughput rate, engineers have considered a wide variety of ways to optimize the model to improve its performance.

Tencent StarLake Lab personnel explore advanced cloud computing, artificial intelligence (AI), security, storage, and network technologies to deliver solutions that improve data center performance and reduce the total cost of ownership (TCO) of data centers. The Tencent Machine Learning Platform Department (MLPD) is the heart of the Tencent AI platform, constantly working to drive innovations across Tencent's internet and technology businesses. The MLPD engages in R&D covering a broad range of fields, including computer vision, voice recognition, graph computation, and NLP. Solutions created by the MLPD have been broadly applied to major scenarios in social media, personalized advertising, gaming AI, and content recommendation and search. BERT plays a key role in applications across all these tech sectors.

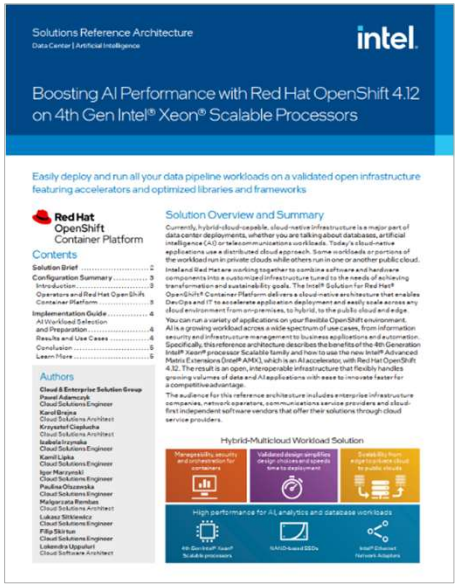
Intel has closely collaborated with Tencent MLPD and Tencent StarLake laboratory on BERT inference optimization using Intel® AMX, a built-in accelerator for 4th Gen Intel® Xeon® Scalable processors. The teams demonstrated that BERT model throughput [INT8] could increase 2x and BERT model throughput [BF16] could increase 3x when running on systems powered by 4th Gen Intel Xeon Scalable processors using Intel AMX.[1] By combining Intel AMX and software optimizations into a powerful unified solution, Tencent aims to evolve its capabilities to deliver a consistent service experience and to optimize TCO.

Tencent Social Applications Optimization

The Tencent social applications connect over a billion active users around the world. One of most popular Tencent social applications was released in 2011 and became the world's largest standalone mobile app in 2018. In fact, it was nicknamed "China's app for everything" because of its impressive array of functions and uses, which include text messaging, voice messaging, broadcast messaging (one-to-many), video games, and video conferencing. Additionally, it includes photo-sharing, video-sharing, and location-sharing features.

Learn more: Solution Brief

Analytics, Database, Edge Computing - RedHat


Solutions Reference Architecture
Data Center | Artificial Intelligence

Boosting AI Performance with Red Hat OpenShift 4.12 on 4th Gen Intel® Xeon® Scalable Processors

Easily deploy and run all your data pipeline workloads on a validated open infrastructure featuring accelerators and optimized libraries and frameworks

Red Hat OpenShift Container Platform

Contents

Solution Brief 2
Configuration Summary 3
Introduction 3
Operators and Red Hat OpenShift Container Platform 3
Implementation Guide 4
AI Workload Selection and Preparation 4
Results and Use Cases 4
Conclusion 5
Learn More 5

Authors

Cloud & Enterprise Solution Group
Pawel Adamczak
Cloud Solutions Engineer
Karol Blazejewski
Cloud Solutions Architect
Krzysztof Ciepluszka
Cloud Solutions Architect
Isabela Inzynska
Cloud Solutions Engineer
Kamil Lipka
Cloud Solutions Engineer
Igor Marzynski
Cloud Solutions Engineer
Paulina Olszewska
Cloud Solutions Engineer
Malgorzata Rembas
Cloud Solutions Architect
Lukasz Sitkiewicz
Cloud Solutions Engineer
Filip Skirtun
Cloud Solutions Engineer
Lokendra Uppuluri
Cloud Software Architect

Solution Overview and Summary

Currently, hybrid-cloud-capable, cloud-native infrastructure is a major part of data center deployments, whether you are talking about databases, artificial intelligence (AI) or telecommunications workloads. Today's cloud-native applications use a distributed cloud approach. Some workloads or portions of the workload run in private clouds while others run in one or another public cloud.

Intel and Red Hat are working together to combine software and hardware components into a customized infrastructure tuned to the needs of achieving transformation and sustainability goals. The Intel® Solution for Red Hat® OpenShift® Container Platform delivers a cloud-native architecture that enables DevOps and IT to accelerate application deployment and easily scale across any cloud environment from on-premises, to hybrid, to the public cloud and edge.

You can run a variety of applications on your flexible OpenShift environment. AI is a growing workload across a wide spectrum of use cases, from information security and infrastructure management to business applications and automation. Specifically, this reference architecture describes the benefits of the 4th Generation Intel® Xeon® processor Scalable family and how to use the new Intel® Advanced Matrix Extensions (Intel® AMX), which is an AI accelerator, with Red Hat OpenShift 4.12. The result is an open, interoperable infrastructure that flexibly handles growing volumes of data and AI applications with ease to innovate faster for a competitive advantage.

The audience for this reference architecture includes enterprise infrastructure companies, network operators, communications service providers and cloud-first independent software vendors that offer their solutions through cloud service providers.

Hybrid-Multicloud Workload Solution

Manageability, security and orchestration for containers

Validated design simplifies design choices and speeds time to deployment

Scalability from edge to on-prem cloud to public cloud

High performance for AI, analytics and database workloads

4th Gen Intel® Xeon® Scalable processors

NAND-based SSDs

Intel® Ethernet Network Adapters

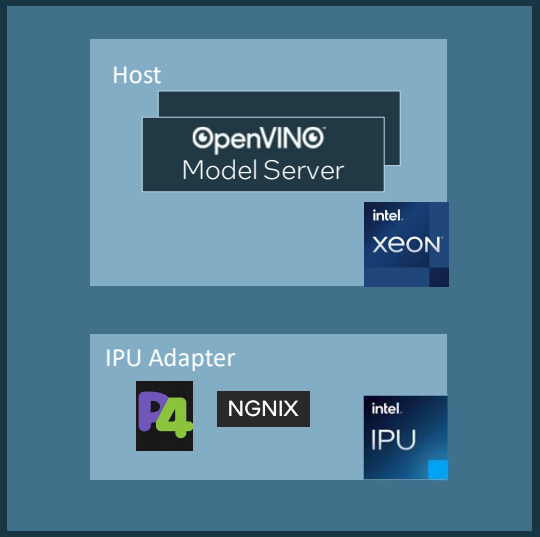
Learn more: Solution Reference Architecture and at Red Hat.com

intel. VISION

IPU Adapters for edge/ enterprise AI use cases

Intel® IPU 2000 Series

Secure, Multi-tenant AI Edge inference

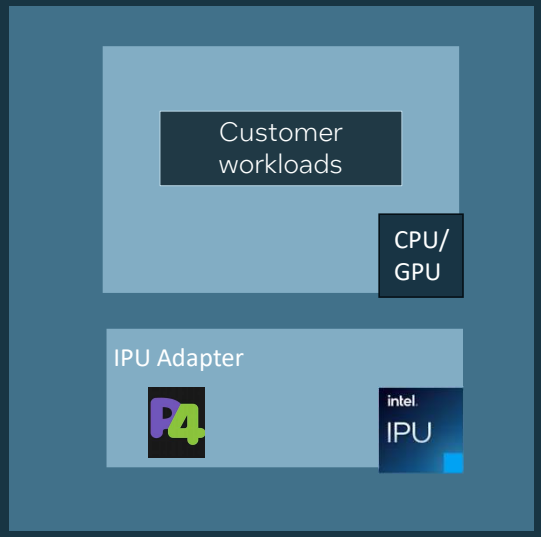


NGINX: secure add-on for authentication / proxy function

IPU:

- physical segmentation of infra and user space
- Customizable data path filters
- Encryption accelerators

General Purpose Reliable Transport



Falcon Transport: support for AI and storage over lossy fabrics

IPU:

- physical segmentation of infra and user space
- Encryption accelerators
- ARM core complex for custom datapath or controller functions

Jason Carolan

Chief Innovation Officer

Flexential

Andrew Thorstensen

Chief Engineer

IBM

Get ready to deploy Ethernet-based AI products now

Intel, with the industry, is enabling **Ethernet for AI everywhere**

Intel has a broad portfolio of AI connectivity products both available now and coming soon

Contact your Intel sales rep for more info

intel. VISION

Notices and Disclaimers

For notices, disclaimers, and details about certain performance claims, visit www.intel.com/PerformanceIndex or scan the QR code:


intel
VISION
Thank You!

Backup

Configuration for MLPerf™ Resnet runs – E810 vs. HDR

Tests performed 11/28/2022 on 2-socket Intel® Xeon® Platinum 8480+ CPU @ 3.0-3.8GHz. Intel® Hyper-Threading Technology enabled. Intel® Turbo Boost Technology enabled with Intel Pstate driver. Red Hat Enterprise Linux 8.6 (Ootpa). 4.18.0-372.9.1.el8.x86_64 kernel. 16x32GB, 512 GB total, 4800 MT/s. Intel® MPI 2021.6. irdma version 1.9.30. ice version 1.9.5. DDP version 1.3.30.0. Intel® Ethernet E800 Series firmware-version: 4.00 0x800117e9 1.3236.0, 104 TxRx queues. pfc_enable: 0x1. Intel Ethernet 800 Series tf2 firmware-version: 4.00 0x800117e9 1.3236.0, 104 TxRx queues. pfc_enable: 0x1. Intel® Ethernet Fabric Suite 11.4.0.0.78. Intel® Ethernet E810-CQDA2 network adapter: PFC enabled, DCB on switch, willing mode on NICs. Accton/Edgecore x86_64-accton_as9516_32d-r0 (Intel® Tofino™ 2 switch). HDR: MLNX_OFED_LINUX-5.4-1.0.3.0. MLPerf™ ResNet50 v1.5 with PyTorch. Results may vary.


intel