AI without GPUs: A Technical Brief for VMware Private AI with Intel

VMware AI/ML

June 2024

vmware by Broadcom

Table of Contents

Sometimes, You "Can Always get What You Want

You want your AI everywhere, right? In the datacenter, in the cloud, and even out there on the edge. But "everywhere" includes lots of places without GPUs. Well, that's not going to be a problem. This technical brief presents a use case showing how VMware Private AI with Intel (and its 4th Gen Xeon CPU, with built-in AMX acceleration) will let you run your AI apps anywhere you want.

We call this...

AI EVERYWHERE

- 90% — of Enterprise Apps will be with AI by 2025*
- 1M+ — Intel 4th Gen Xeon Install Base*
- 100% — of Fortune 500 Global Companies use VMware Technologies & Services**
- Better Together — 4th Gen Xeon with Built in AI + VMware Products & Services

This technical brief describes VMware Private AI and shows how it works with Intel's built-in accelerator, AMX (Advanced Matrix Extension). We show the architecture of the solution, and we walk through a use case and its performance to help provide customers confidence and a concrete example so they can achieve AI Everywhere leveraging VMware Private AI with Intel.

What is VMware Private AI with Intel

VMware Private AI with Intel is the synergy of VMware Private AI (the VMware base for AI apps, that respects privacy and compliance) and the Intel AMX instructions (for AI acceleration) available of the 4th Gen Xeon CPUs. It also implies compatibility with the Intel AI ecosystem. Let's look at each.

VMware Private AI

VMware Private AI is the architectural approach – the components and expertise – that delivers business gains from AI while meeting the practical privacy and compliance needs of the organization. VMware Private AI is architected with industry partners to create a solution for AI services that ensures privacy and control of corporate data, choice of open source and commercial AI solutions, quick time-to-value, and integrated security and management. The VMware Private AI infrastructure stack is anchored by VMware Cloud Foundation along with acceleration capabilities, such as Intel AMX instructions. It provides an optimized infrastructure solution for delivering Private AI within the Enterprise and allows customers to balance the advantages of AI while governing how and where their data is used for GenAI use cases.

VMware Cloud Foundation

VMware Cloud Foundation (VCF) is a software suite virtualizing (and managing) all three main components of the data center: compute, storage, and networking. You can think of it as the combination of vSphere, vSAN, and NSX.
Implementing demanding ML use cases, such as LLM, atop VCF marks a new era of capabilities and opportunities for enterprises. VCF is a turnkey platform and an ideal choice for deploying AI workloads due to its integrated infrastructure management, which simplifies complex tasks. The platform ensures consistency across various cloud environments, facilitating seamless AI workload deployment. Its scalability meets the resource demands of AI tasks, while its resource efficiency optimizes infrastructure utilization. The platform's automated lifecycle management streamlines updates and upgrades, minimizing disruptions to AI operations. Furthermore, the platform's integration with AI-specific hardware accelerators further amplifies performance potential. The graphic below shows a full representation of the VCF solution and software components.

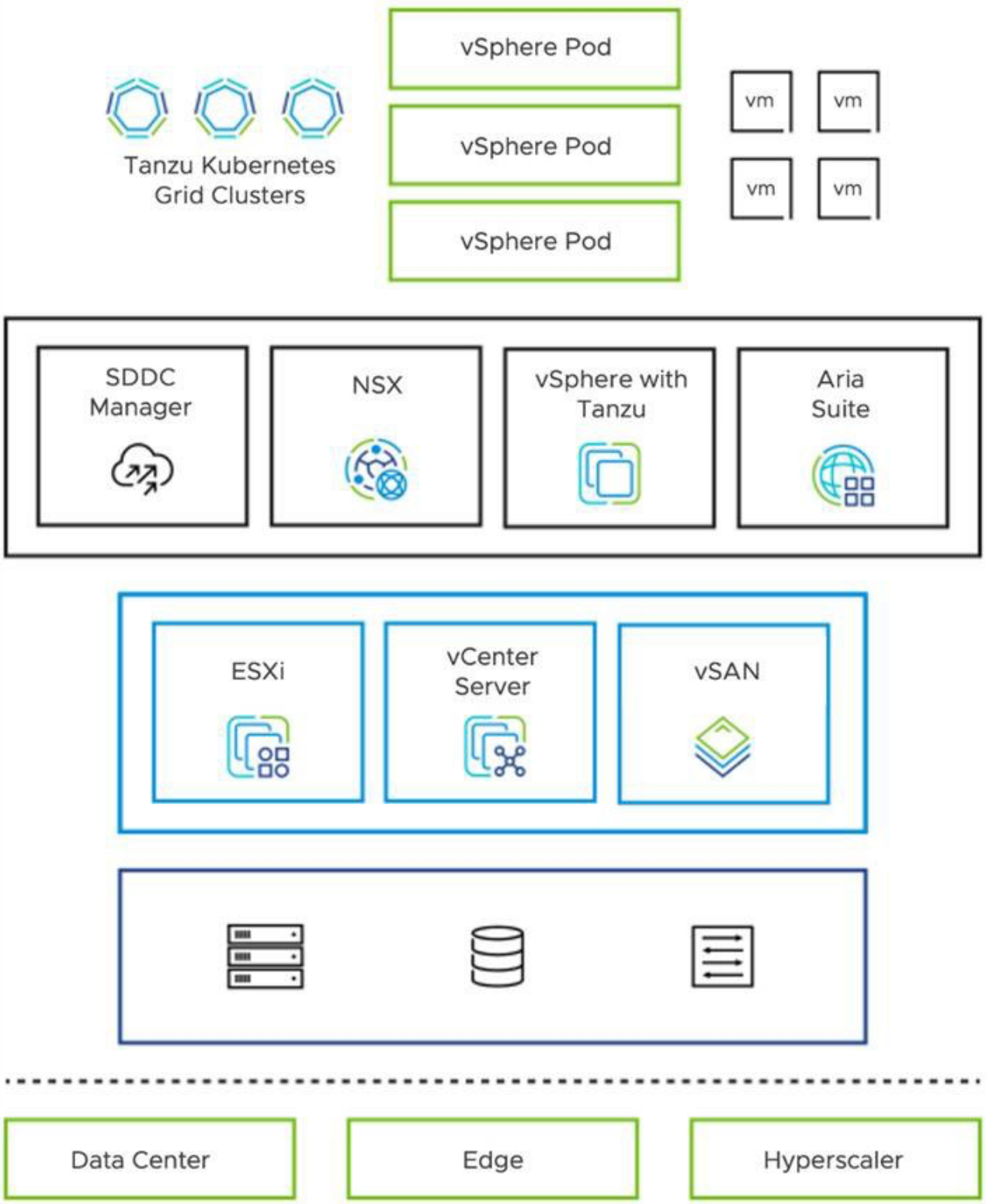

Figure 1: A VCF solution with software components delivering AI Everywhere

Intel 4th Gen Xeon Scalable Processors with AMX

Intel describes AMX as the next big step in AI.
"At Intel we've created Intel® Advanced Matrix Extensions, or Intel® AMX, which is an AI acceleration engine deeply embedded into every core of our 4th Gen Intel® Xeon® Scalable processor."
Intel AMX is a built-in accelerator that enables 4th Gen Intel Xeon Scalable processors to optimize deep learning (DL) training and inferencing workloads. With Intel AMX, 4th Gen Intel Xeon Scalable processors can quickly pivot between optimizing general computing and AI workloads. Imagine an automobile that could excel at city driving and then quickly change to deliver Formula 1 racing performance. 4th Gen Intel Xeon Scalable processors deliver this type of flexibility. Developers can code AI functionality to take advantage of the Intel AMX instruction set, and they can code non-AI functionality to use the processor instruction set architecture (ISA). Intel has integrated the Intel oneAPI Deep Neural Network Library (oneDNN), its oneAPI DL engine, into popular open-source tools for AI applications, including TensorFlow, PyTorch, PaddlePaddle, and OpenVINO.
Intel AMX was designed to balance inference, the most prominent use case for a CPU in AI applications, with more capabilities for training. With Intel Xeon Scalable processors representing 70 percent of the processor units (installed base) that are running AI inference workloads in the data center, selecting 4th Gen Intel Xeon Scalable processors with Intel AMX for new AI deployments is an efficient and cost-effective approach to accelerating AI workloads.
The "Intel" part of VMware Private AI with Intel ensures that AMX is already enabled with vSphere and Tanzu – it's part of the out- of-box experience. This makes it fast and easy to spin up Tanzu Kubernetes Clusters with AMX-enabled CPU workers.

We can do good inference on Skylake, we added instructions in Cooper Lake, Ice Lake, and Cascade Lake. But AMX is a big leap, including for training.
Bob Valentine
Intel Microprocessor Architect

The Architecture

A high-level overview of the solution is depicted below starting with the infrastructure components up to the AI application layer.

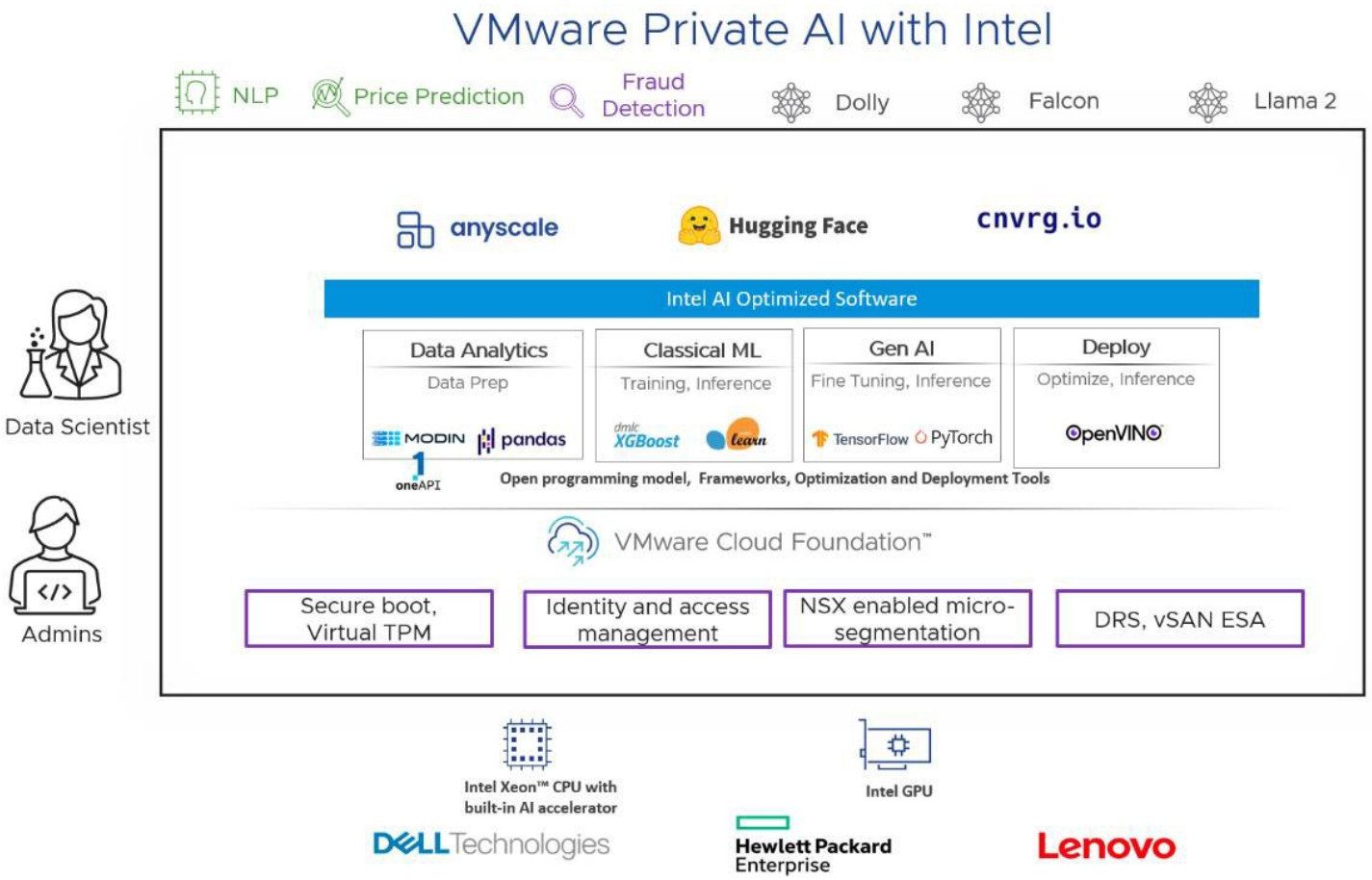

Figure 2: The architecture with VMware Private AI and Intel AMX

Figure 2 illustrates the stacking of the software components used in various applications starting from the AI and MLOps Platform and all the way up.
Customers can use the solution for many workflows. With LLM, for example, the two primary workflows are possible – customization (fine-tuning, prompt-tuning, and others) and inference at scale. Both workflows demand more computing capacity than traditional deep learning workloads regarding LLMs. Inference also requires accelerated resources depending on application needs or number of users. By leveraging VCF and distributed computing frameworks like Ray, unused resources in an environment can be repurposed for machine learning (ML) workflows. This improves infrastructure utilization and boosts productivity for ML overall.

The Hardware

When embarking on your journey to AI Everywhere, it is important to pick the right CPU for the job. Below we have provided some guidance on which CPU version we selected, and some other hardware details used to build out the environment that validated the use case outlined later in the paper. The specifications are based on a per node configuration.

Resource	QTY	Specification
CPU	2	Intel(R) Xeon Gen4 Scalable processors 8480+, 56C/112T per CPU
RAM	8	32GB 4800 MT/s DDR5 (256GB total per host)
Storage	6	P5620 6.4TB Mixed Use NVMe PCIe4.0x4 HS SSD
Network	1	Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter

Table 1: Hardware used in use case

VMware Tanzu and Tanzu Kubernetes Releases (TKR)

VMware vSphere with Tanzu transforms traditional virtualization infrastructure into a robust platform for containerized workloads. VMware Tanzu Kubernetes Grid™ facilitates the creation and management of Tanzu Kubernetes Cluster (TKC) natively within vSphere, seamlessly integrating Kubernetes capabilities with the reliable features of vSphere. With vSphere's networking, storage, security, and VMware vSphere High Availability (vSphere HA) features, organizations achieve better visibility and operational simplicity for their hybrid application environments.
VMware vSphere 8 and Tanzu now support AMX instruction sets in the latest Tanzu Kubernetes Releases (TKR) out of the box. Making it fast and easy to spin up Tanzu Kubernetes Clusters with AMX-enabled CPU worker nodes.
vSphere with Tanzu also enables organizations to run application components in containers or VMs on a unified platform, streamlining workload management and deployment. By converging containers and VMs, IT teams can leverage the benefits of both paradigms while maintaining a cohesive and scalable infrastructure.

Intel Ecosystem

Private AI, Accelerated by Intel® AI with VMware is a robust suite designed to boost AI workloads in various industries. In partnership with VMware, Intel delivers the virtual drivers and AMX instruction sets that enable the Tanzu Kubernetes Resources and Virtual Machines for optimal performance of AI workloads running on VCF and Tanzu Kubernetes Grid.
To access AMX instructions for accelerating AI/ML workloads your vSphere environment will require: Hardware with Intel Xeon Sapphire Rapids CPUs.
Guest VMs running Linux kernel 5.16 or later. Kernel 5.19 or later recommended.
Guest VMs using HW version 20 (requires ESXI 8.0u1).
Tanzu Kubernetes worker nodes running Linux kernel 5.16 or later and HW version 20, which are both supported in the latest TKR images (requires vCenter 8.0u2).

Deploy a Tanzu Kubernetes Cluster (TKC)

First verify that you are using vCenter 8.0u2 (8.0.2) and that your ESXi hosts are running ESXi 8.0u1 or later.
Assuming that you have already deployed a Tanzu Supervisor Cluster, to stand up a Tanzu Kubernetes Cluster (TKC) that supports AMX you need to make sure that you are using a Tanzu Kubernetes Release (TKR) image that contains kernel 5.16 or later and that uses HW version 20, and that's it.
The latest TKR that supports this kernel is in a separate content library from the mainstream Tanzu TKRs, so first you will need to add a new content library (vSphere Client > Content Libraries > Create) with the subscription
URL: https://wp-content.vmware.com/v2/labs/hwe/lib.json. Once you have created this content library the TKR will be available for deploying new TKCs.

To list the release names of available TKRs:

```
kubectl config use-context \
      my-tanzu-kubernetes-cluster-namespace
kubectl get tanzukubernetesreleases
```

Only the names that have READY=True and COMPATIBLE=True can be used to deploy a cluster.
VMware Tanzu Configuration – configure a VMClass with the following specifications: 26 vCPUs with 100% reservation (i.e., 26 cores/1 thread per core) and 50GB of memory (you can reserve the memory but not required), and HW version 20 — compatible with ESXi 8.0 and later. (The flow to select the HW version for a VMclass will be available in a future release of vCenter)

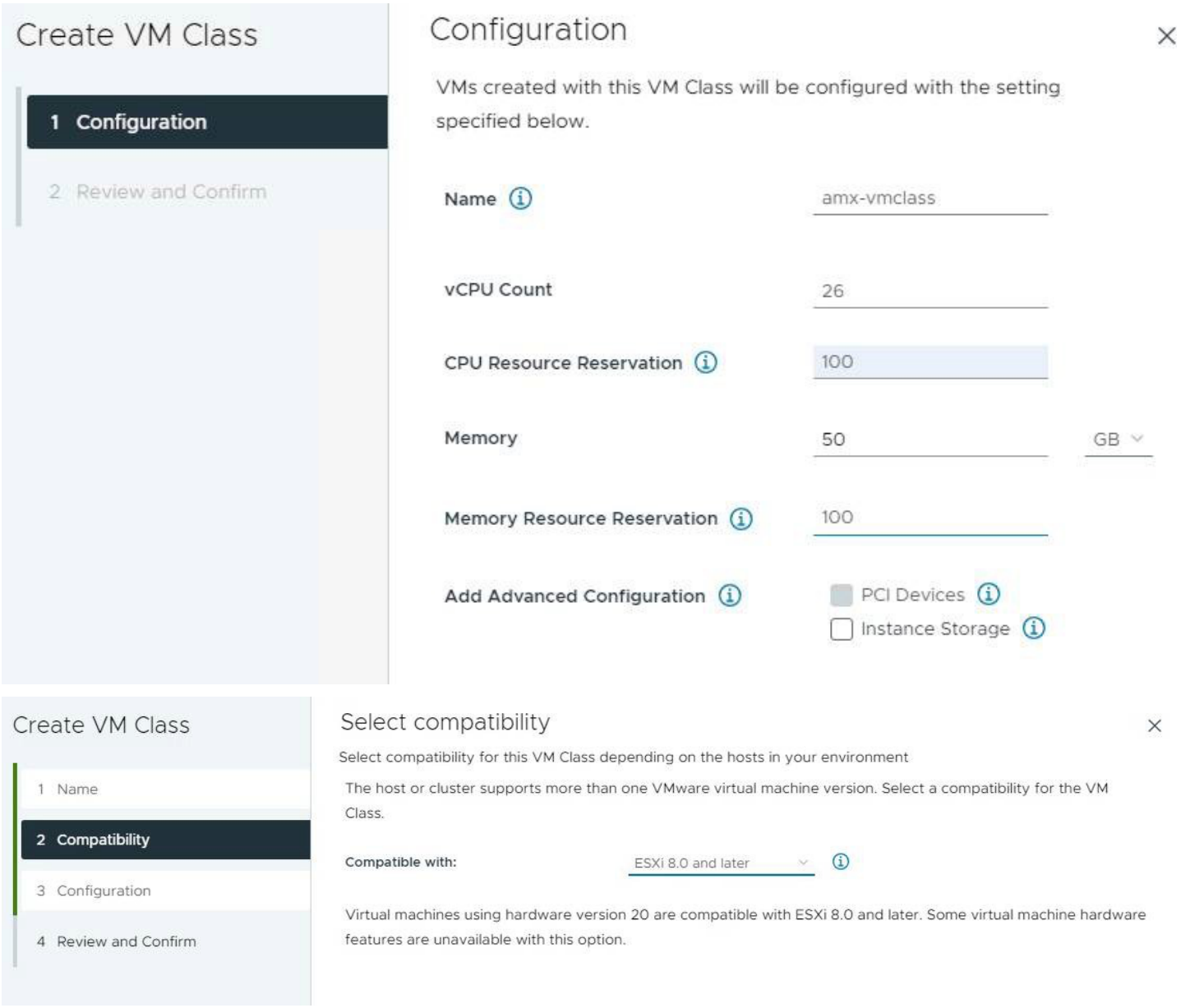

Figure 3: VMware Tanzu configuration for Llama 2 use case

This can also be achieved by using the vCenter Rest API, and leveraging the Developer Center within the vCenter UI users can create a VMclass without having to ssh into the vCenter and use dcli. The screenshot below shows this capability for reference:

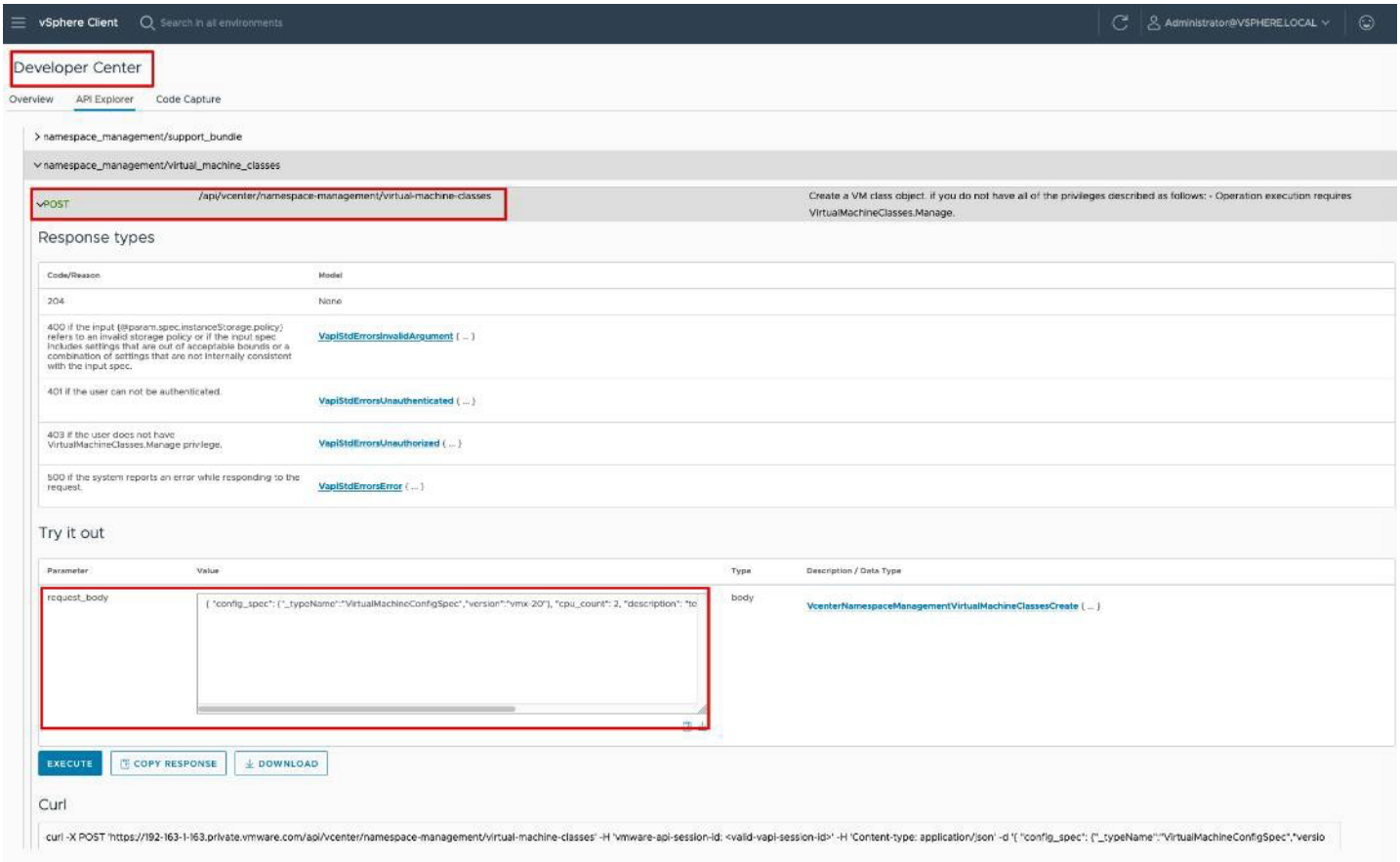

Figure 4: Using vCenter Rest API for VMware Tanzu configuration

Below is a graphical representation of what your environment will look like after you have deployed a Tanzu Kubernetes Cluster (TKC):

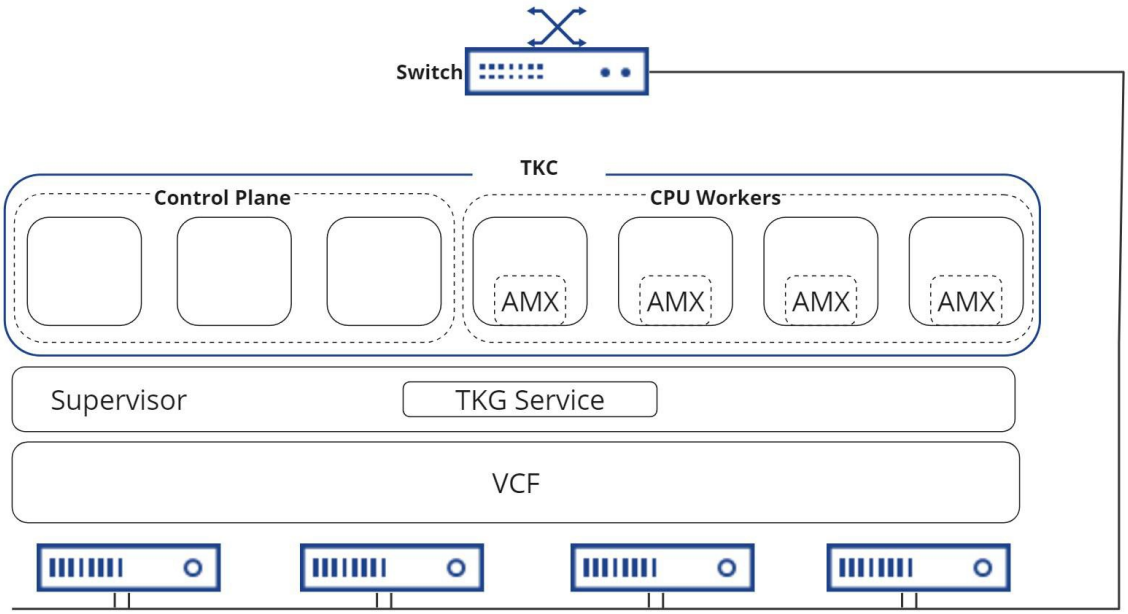

Figure 5: Environment for Llama 2 use case after deploying TKC

Next create a cluster definition yaml file. Here is an example you can use:

```
    apiVersion:
    run.tanzu.vmware.com/v1alpha3 kind:
    TanzuKubernetesCluster
    metadata:
      name: my-tkc-name
      namespace: my-tanzu-kubernetes-cluster-
      namespace annotations:
        run.tanzu.vmware.com/resolve-os-image: os-
    name=ubuntu spec:
      topology:
controlPlane
          :
 replicas: 3
          vmClass: guaranteed-small
          storageClass: vsan-default-
          storage-policy tkr:
            reference:
              name: v1.26.5---vmware.2-fips.1-
        hwekernel.1 nodePools:
        - name:
          worker
          replicas:
          3
          vmClass: amx-vmclass
          storageClass: vsan-default-
          storage-policy volumes:
            - name: containerd
              mountPath:
              /var/lib/containerd
              capacity:
                storage: 160Gi
          tkr:
            reference:
              name: v1.26.5---vmware.2-fips.1-hwekernel.1
```

Finally, deploy the cluster:

```
kubectl config use-context \

my-tanzu-kubernetes-cluster-namespace

kubectl apply -f my-yaml-filename
```

Verify AMX Capability

To verify that a VM or Tanzu worker node can use AMX instructions you need to ssh to the VM or worker node and run some commands.

You can get the IP address of a Tanzu worker node from vCenter by going to Hosts and Clusters, then drilling down through my- datacenter > my-cluster > Namespaces > my-tanzu-kubernetes-cluster-namespace > my-tanzu-kubernetes-cluster-name > my- worker-node-name. The node's IP addresses are listed on the right panel on the Summary tab.

On the same Summary tab, just above the IP addresses for a VM or Tanzu worker node, check the value of Compatibility. It must say "VM version 20" (or greater) if you want to access AMX instructions.

To ssh into a Tanzu worker node you need to get the password for the "vmware-system-user". To get the password run these kubectl commands:

```
kubectl config use-context \
    my-tanzu-kubernetes-cluster-
namespace kubectl get secret \
    my-tanzu-kubernetes-cluster-name-ssh-password \
    -o jsonpath='{.data.ssh-passwordkey}' \
    -n my-tanzu-kubernetes-cluster-namespace |
base64 –d ssh -o PubkeyAuthentication=no \
    vmware-system-user@vm-ip-address
```

Once you have ssh'ed into the VM or worker node:

- Run "uname -a". The kernel version must be 5.16 or greater to use AMX.
- Run "grep amx /proc/cpuinfo | head –1". The line returned should list amx_bf16, amx_tile, and amx_int8 as supported instructions. If nothing is returned, the VM cannot run AMX.

The Use Case

Large Language Models are at the forefront of innovation. They empower organizations to streamline customer service, extract insights from mountains of data and make informed decisions. For learners, they provide personalized guidance and answers at fingertips. In healthcare, LLMs assist in diagnosis and research, potentially saving lives. They unlock gates to global communication, breaking down language barriers and making information universally accessible. LLMs are not just tools, their benefits are limitless and have the potential to unlock new forms of creativity and insights, as well as inspire passion in the AI community to advance the technology.
Intel is committed to democratizing AI with ubiquitous hardware and open software. Intel offers a portfolio of AI solutions that provide competitive and compelling options for the community to develop and run models like Llama 2. Intel's rich AI hardware portfolio combined with optimized open software provides alternatives to mitigate the challenge of accessing limited compute resources.

The App, Tools, and Libraries

Below we list the Intel Optimized Frameworks and tools required to instantiate your own environment so you can validate the real- world LLAMA 2 use case and have confidence in this joint VMware and Intel solution.

Figure 6: Configuration of a single node in a Llama 2 use case

Framework: https://github.com/intel/intel-extension-for-pytorch/tree/master/examples/cpu/inference/python/llm

The Model

Our use case is Llama 2-7B: Verified "meta-llama/Llama-2-7b-hf"

Please refer to the model card on the site referenced below to learn more about this model and how it was trained.
https://huggingface.co/meta-llama/Llama-2-7b-hf

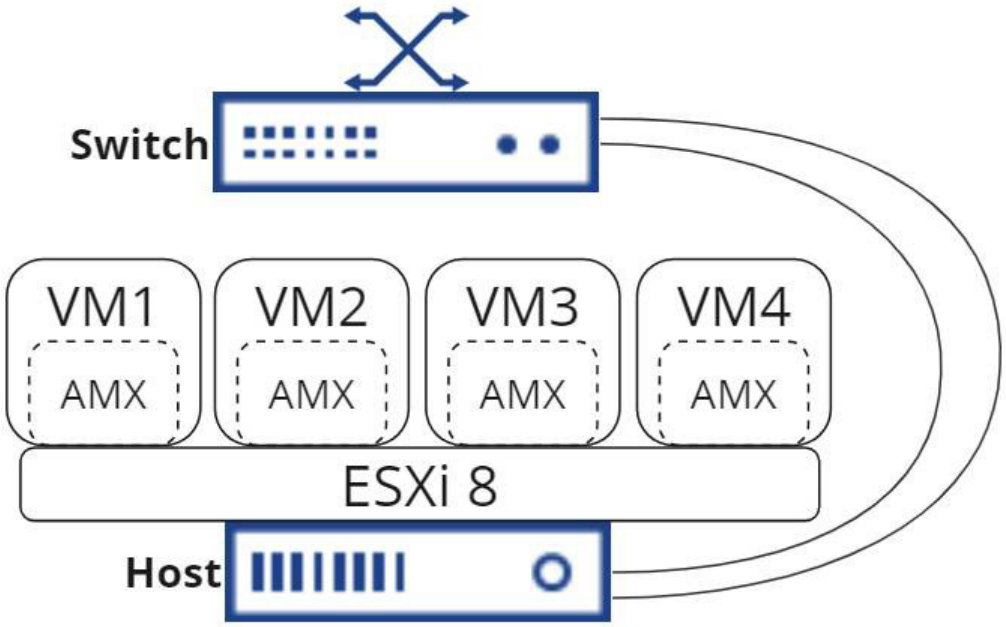

Figure 7: Configuration of a single node in a Llama 2 use case

If we look at the individual components that are being deployed into each VM the following diagram shows the key elements:

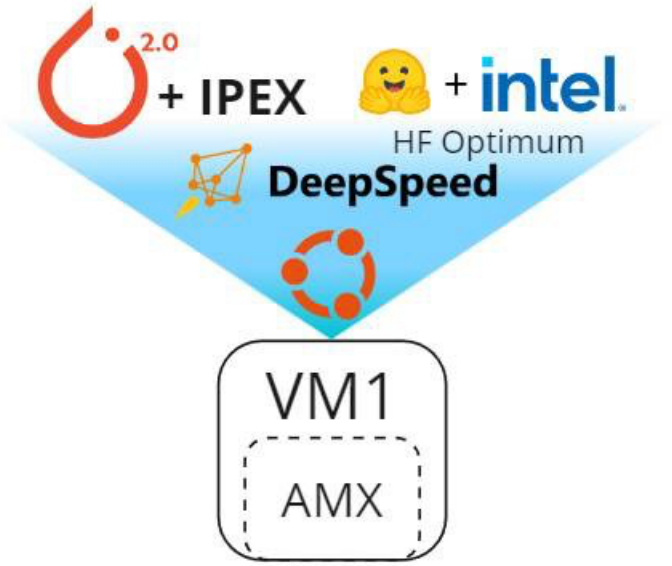

Figure 8: The components on each VM in a Llama 2 use case

Performance

We are excited to share the initial inference performance of 7B and 13B parameters LLAMA 2 models, on Intel(R) 4th Gen Xeon Scalable Processors, with VMware vSphere 8.

LLAMA2_7B INT8/BFLOAT16 (on one socket, output tokens: 32)

- Across input tokens (32, 128, 1K, 2K), INT8 1 instance/ socket can deliver inference with avg. latency under 50ms.
- Across input tokens (32, 128, 1K, 2K), INT8 2 instances/socket can deliver inference with avg. latency under 100ms.
- Across input tokens (32, 128, 1K), INT8 3 instances/socket can deliver inference with avg. latency under 100ms.
- Across input tokens (32, 128, 1K, 2K), BF16 1 instance 1 socket can deliver inference with avg. latency under 100ms.
- Across input tokens (32, 128, 1K, 2K), INT8 speed up is up to 1.8x of BF16 model.

LLAMA2_13B INT8 (on one socket, output tokens: 32)

Intel(R) 4th Gen Xeon delivers <100ms latency (on 1 socket), for Llama-2 7B (INT8) model across various tokens input ranging from 32 bytes to 2K, tokens output is 32.

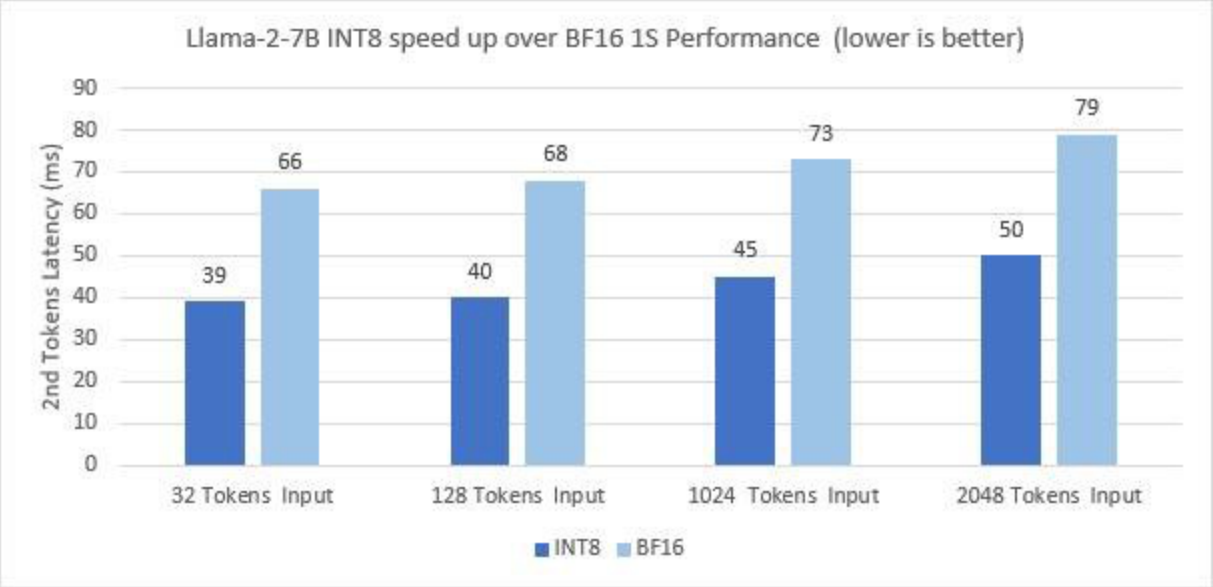

Figure 9: LLAMA2 – 7B: One Instance per Socket

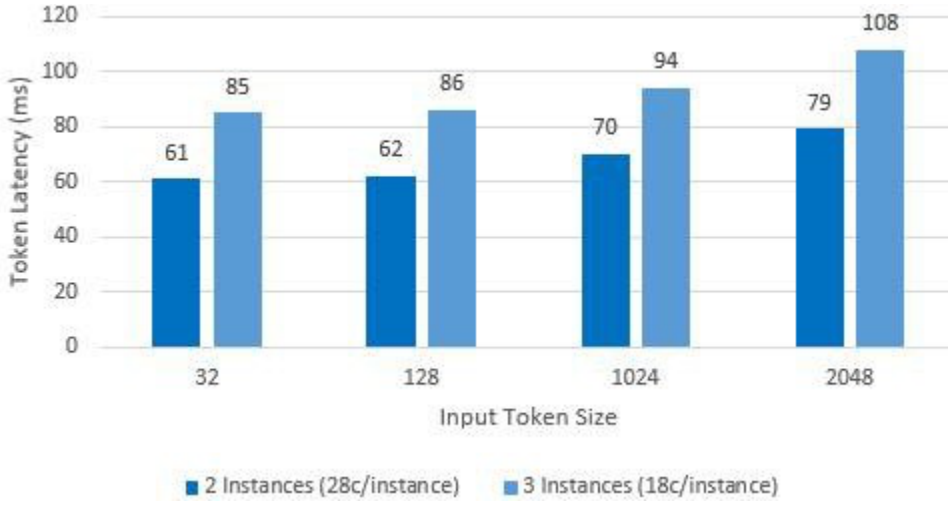

Figure 10: LLAMA2-7B INT8: Multi-instance Per Socket

Two instances running on 1S delivers <100ms latency for Llama-2 7B model, up to tokens input of 2K, tokens output is 32 (batch- size=1)

Three instances running on 1S delivers <100ms latency for Llama-2 7B model, up to tokens input of 1K, tokens output is 32 (batch- size=1)

Figure 11: LLAMA2 – 13B INT8: One Instance per socket

Latency of <100ms one single socket, for Llama-2 13B (INT8) model with 2K tokens.

You Got What You Need

In conclusion, the collaboration between VMware and Intel is a powerful combination that enables AI Everywhere by bringing the value of 4th Gen Intel Xeon processors to the forefront. These processors offer advanced features such as Intel AMX (Advanced Matrix Extensions) and support for DDR5 that are essential for AI workloads. With VMware's software-defined approach to AI infrastructure, deploying and managing AI workloads becomes much simpler, allowing organizations to focus on developing and deploying AI models that drive business outcomes.

The collaboration between VMware and Intel simplifies the deployment of AI infrastructure by providing a seamless and efficient experience for customers. By leveraging the power of 4th Gen Intel Xeon processors, organizations can unlock the full potential of AI and achieve better business outcomes. With VMware and Intel working together, customers can accelerate their AI journey and achieve AI Everywhere.

Find Out More

For more information on accelerating AI/ML workloads on vSphere 8 and Intel hardware please read:

Intel and VMware Partnership - https://www.vmware.com/partners/strategic-technology-partners/intel.html

Intel and VMware - https://www.vmware.com/partners/strategic-technology-partners/intel.html

Intel Developer Catalog- https://www.intel.com/content/www/us/en/developer/tools/software-catalog/overview.html

AI without GPUs: Accessing Sapphire Rapids AMX instructions on vSphere - https://core.vmware.com/blog/ai-without-gpus-accessing-sapphire-rapids-amx-instructions-vsphere

How to create a Local TKR Content Library

https://docs.vmware.com/en/VMware-vSphere/8.0/vsphere-with-tanzu-tkg/GUID-19E8E034-5256-4EFC-BEBF-D4F17A8ED021.html

Accelerate AI Workloads on VMware vSphere / vSAN Using 4th Gen Intel® Xeon® Scalable Processors with Intel® AMX — Solution Design Brief -

https://www.intel.com/content/www/us/en/content-details/780611/accelerate-ai-workloads-on-vmware-vsphere-vsan-using-4th-gen-intel-xeon-scalable-processors-with-intel-amx-solution-design-brief.html

To learn more about data analytics, classical machine learning, deep learning training, and inference & optimization - https://www.intel.com/content/www/us/en/developer/tools/oneapi/aikit-config-beta.html

ITREX - https://github.com/intel/intel-extension-for-transformers

* Forbes: Forbes

: https://www.forbes.com/sites/gilpress/2019/11/22/top-artificial-intelligence-ai-predictions-for-2020-from-idc-and-forrester/#4fef9821315a

** VMware: https://www.vmware.com/files/pdf/VMware-Corporate-Brochure-BR-EN.pdf