White Paper

Cloud-Native AI/ML Edge Applications

High Density Deep Learning Application Orchestration with Nearby One and Intel® Smart Edge Open

Introduction

Adoption of artificial intelligence and the insights driven by the analysis of large data sets is transforming industries and creating opportunities for entirely new solutions. While the initial efforts have leveraged cloud-based resources, there are many solutions that would benefit from having machine learning and artificial intelligence capabilities at the network edge, closer to where the data is created and used. At the same time, the rollout of 5G services is creating new solution opportunities that combine edge computing with high-bandwidth, low-latency, and secure 5G networks to deliver new solutions in areas that previously did not have effective access to cloud compute resources.

With the combination of visual computing and artificial intelligence, solutions can gather real-time intelligence from video feeds to enable new solutions for factory automation, digital security and surveillance, vehicle traffic monitoring and management, and many others. Many of these visual computing solutions require hardware to be located in areas where space restrictions and power constraints limit the power consumption and thermal cooling capabilities of the hardware. The ability to leverage visual processing accelerators, such as Intel® Movidius™ Vision Processing Units (VPU), is critical to deploy high-performance visual computing at the edge.

Intel® Smart Edge Open is a toolkit that brings platform capability enabling high density workload deployment in a multi-tennant edge environment. Nearby Computing has leveraged Intel® Smart Edge Open and its High Density Deep Learning (HDDL) plugin to deliver the Nearby One platform that addresses these key considerations for network edge visual computing.

Nearby One

Nearby One is a cloud-native solution for edge computing, enabling a modular and multi-tenant platform that can be managed from a single pane of glass. Nearby One has been designed to assist with two key aspects of end-to-end service orchestration:

- **Inter-node orchestration:** coordination of multiple, distributed edge platforms and the orchestration of applications, allowing applications to scale and move between nodes based on usage requirements
- **Intra-node orchestration:** management of the hardware and software components of each edge platform, including allocation of hardware resources such as processor cores and memory as well as hardware accelerators to ensure optimal performance

Nearby Orchestration Platform is the main component of the solution. It runs in a central location and oversees the performance of all tasks related to the orchestration of applications and infrastructure.

Nearby Blocks are distributed software components that implement specific applications and network functionalities, allowing edge platforms to easily add capabilities for different applications and workloads.

Table of Contents

Nearby One offers a broad selection of containerized applications and virtual network functions. Nearby One supports industry-standard orchestration platforms, such as OpenShift and other Kubernetes distributions, allowing customers to easily migrate workloads from the cloud to the edge.

As part of the Nearby One ecosystem, specific Intel Smart Edge Open microservices have been integrated into the solution, including HDDL-enabled acceleration of video analytics workloads powered by the OpenVINO™ toolkit.

Intel® Smart Edge Open

Intel Smart Edge Open is a royalty-free edge computing software toolkit that enables highly optimized and performant edge platforms to on-board and manage applications and network functions with cloud-like agility across any type of network. Intel Smart Edge Open uses standardized APIs that enable developers to expand functionality and integrate with existing software. This open platform is designed to foster open collaboration and application innovation at the network and on-premises edge, making it easier to engage with a worldwide ecosystem of hardware vendors, software vendors, and system integrators to develop solutions.

Intel Smart Edge Open offers unique capabilities to accelerate application development at the edge:

- Abstracts the complexity of multiple network access, simplifying the migration of applications from the cloud to the edge
- Enables secure on-boarding and management of applications with an intuitive web-based GUI
- Delivers ready-to-use building blocks for functionalities such as access termination, traffic steering, multi-tenancy, service registry, service authentication, telemetry, and appliance discovery and control

High Density Deep Learning Acceleration

In visual computing solutions where a lot of video data needs to be analyzed, CPU performance can quickly be exhausted. In these demanding applications, hardware acceleration is critical to deliver required performance within acceptable power envelopes.

Intel Movidius VPUs combine highly parallel programmable compute with workload-specific hardware acceleration in a unique architecture that reduces data movement. In addition to delivering a balance of power efficiency and compute performance, Intel Movidius VPUs enable developers to deliver edge solutions in retail, security, and automation that require computer vision and AI workloads to run efficiently.

Leveraging HDDL Acceleration in Nearby One

The Nearby Block catalog contains several applications for visual object detection based on OpenVINO toolkit. These applications automatically detect when HDDL accelerators such as Intel Movidius VPUs are installed on the edge platform, reducing the computation power consumption and improving the predictability of the application's performance without requiring developers to rewrite software to take advantage of hardware acceleration.

HDDL support is provided by the integration of the Intel Smart Edge Open HDDL microservice in Nearby One edge node's provisioning workflow. See Figure 1.

The HDDL microservice provides the backend service to manage VPUs and dispatch inference tasks to VPUs. OpenVINO™ toolkit-based applications that utilize HDDL hardware need to access the device node '/dev/shm' and domain socket under '/var/tmp' to communicate with the kernel and HDDL service. The OpenVINO toolkit-based application can use HDDL acceleration without any changes in its logic.

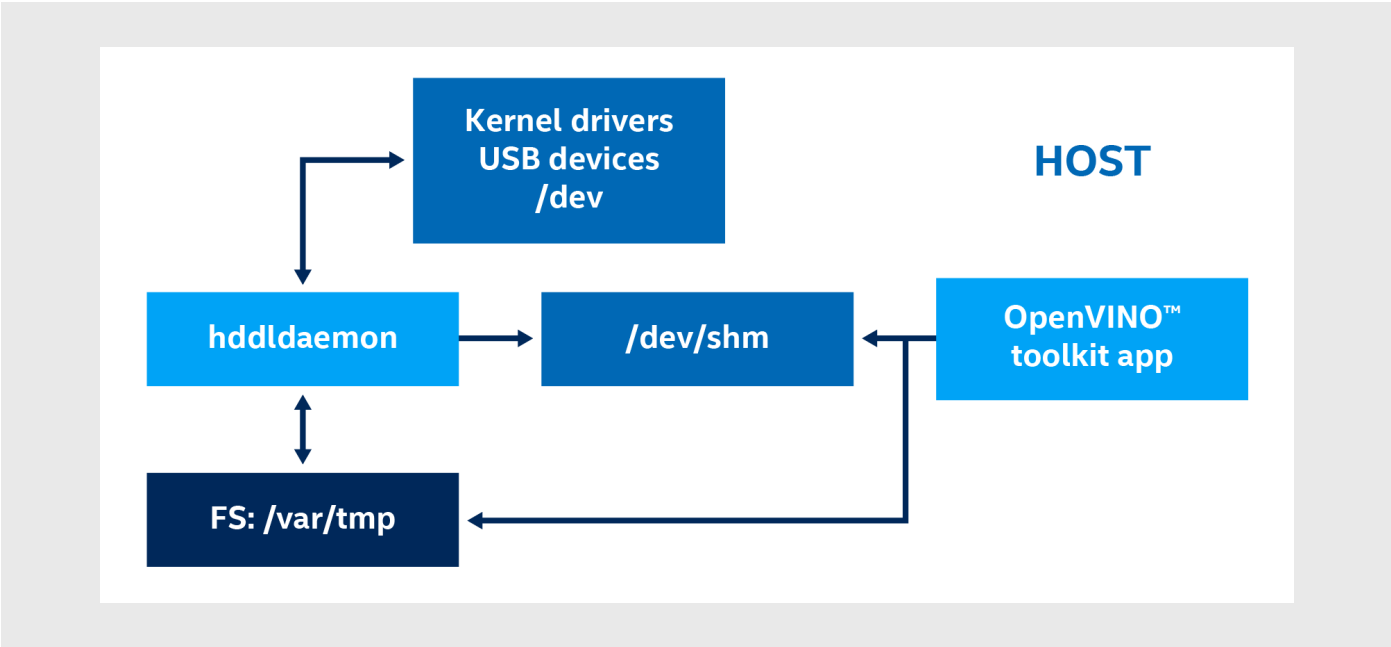

Figure 1. HDDL microservice interaction with the edge node resources and the applications

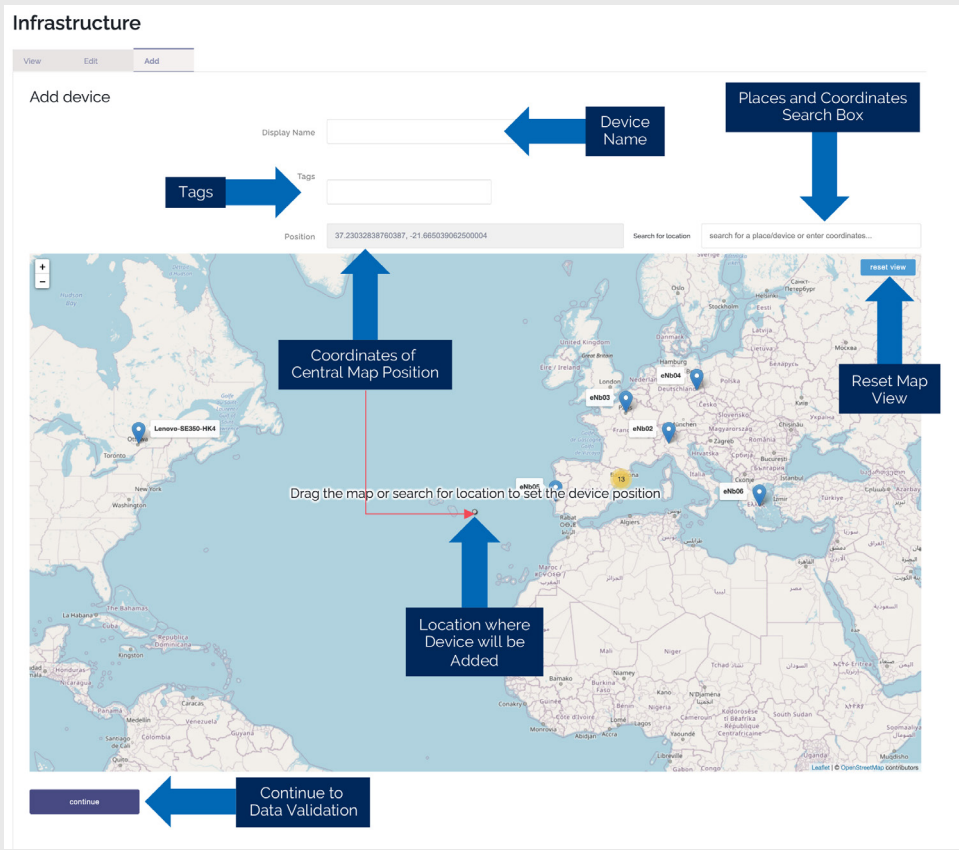

Figure 2. Infrastructure module–Add View–Layout

Deployment of HDDL-ready Nearby Blocks

When desiring to use hardware acceleration with OpenVINO toolkit media analytics, in addition to ensuring an Intel Movidius VPU is installed on the node, the node must also have the HDDL microservice incorporated in its provisioning workflow.

The HDDL edge node must first be registered in the Nearby One device catalog. Information about the device is provided before the device can be added. When the user is satisfied with the information, the process continues with a data validation form before registering the device. See Figure 2.

The device-related information to be added is:

- Geographical location
- Device name
- Search tags

Because the device must be provisioned with Nearby One agents after registration, one of the required tags will be 'nztp' (Near Zero-Touch Provisioning). Once this tag is added, two new form fields will be required before finishing the registration process:

- **MAC address:** Primary device address used for device provisioning. For secure provisioning, the edge device must boot from the Nearby One edge ISO image. The edge ISO, during the boot process, will capture the MAC address of the primary device.
- **Workflow:** The user is presented with a list of existing workflows as configured by the System Administrator in a drop-down list. The user can choose among the existing options defined by the System Administrator. Each of these workflows defines a full stack of OS and software (i.e., Red Hat Enterprise Linux + DPDK + Intel Smart Edge Open building blocks). Nearby One uses Ansible as the mechanism to configure devices after the OS is provisioned, and therefore, virtually any device configuration is possible after it has been correctly bootstrapped.

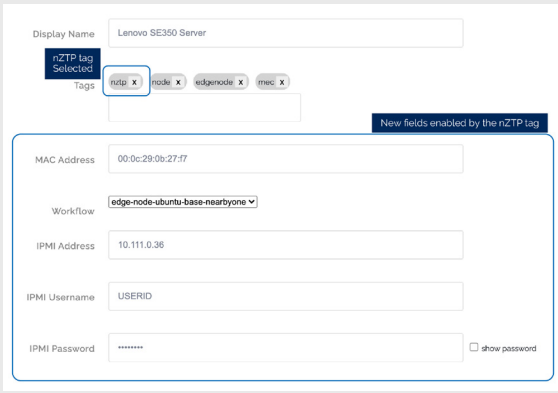

Figure 3. Infrastructure module–Fields for nZTP devices

Once the edge node has been registered and the proper workflow selected, the platform is ready to boot. For this purpose, either a bootable USB stick or an ISO image provided by Nearby Computing is pre-loaded into the server. When powered on, the edge platform will run the automated installer and configure the Intel Smart Edge Open building blocks.

Nearby One includes a designer module that allows users to:

- Define service chains, including the Nearby Blocks that are part of the deployment
- Establish the relationship between the various Nearby Blocks to be deployed, at both the networking and capabilities level
- Provide deployment parameters for the specific instances of the service to be deployed
- Select the deployment type (cloud native, VMs, hybrid)
- Define and apply business policies
- Define networking parameters associated to the deployment, including VLANs and shared networks between tenants

When adding an OpenVINO toolkit-based block, the designer allows properties such as inference model configuration, preferred hardware, and video input URLs to be selected.

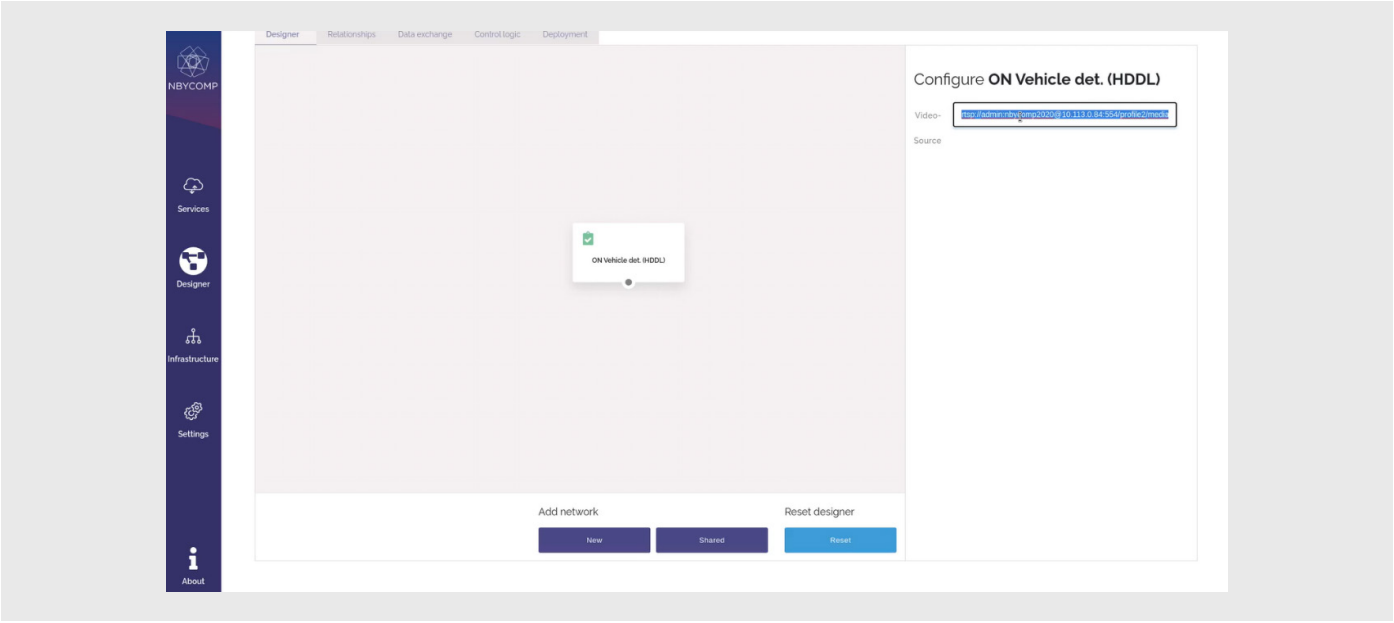

Figure 4. HDDL sample block configuration

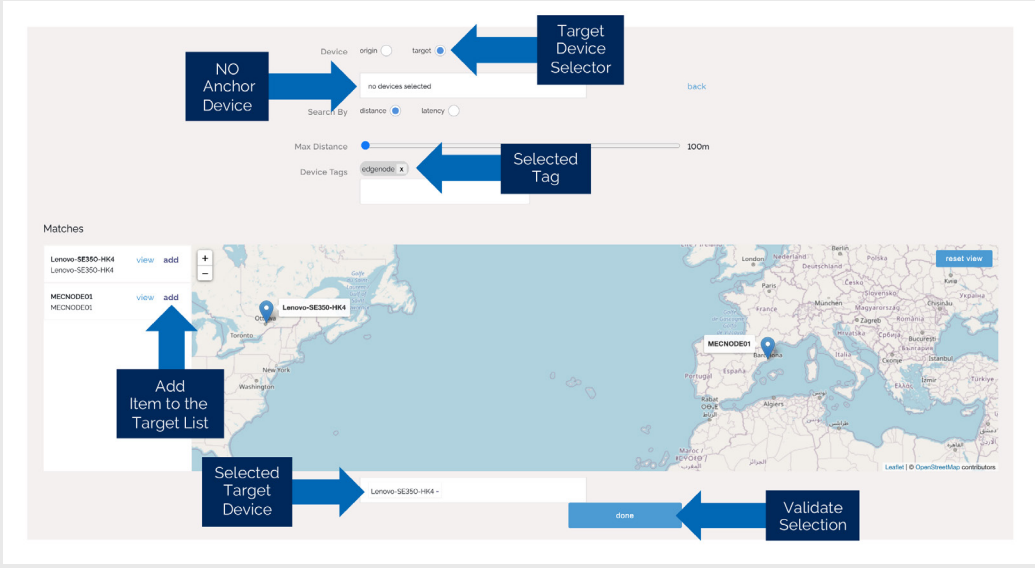

Figure 5. Direct device selection form

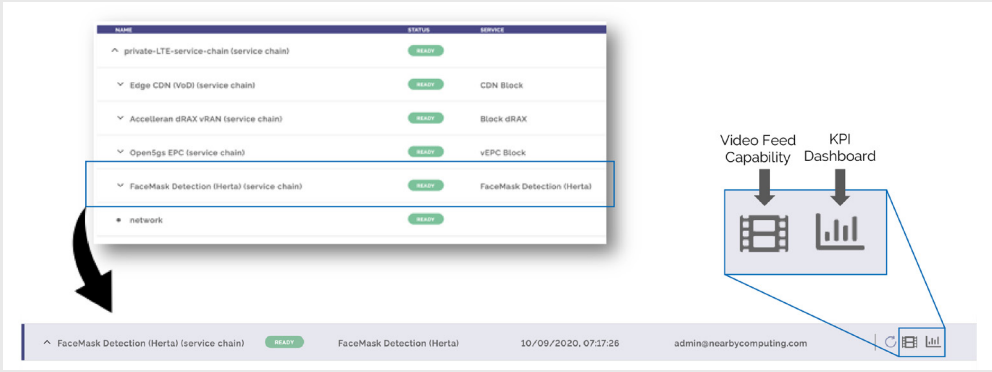

Figure 6. Services module: Block deployment status

Nearby Blocks can export performance KPIs to the orchestration layer. This process is automatically handled by Nearby One engine and accessed in the UI next to the service instance descriptor in the Service module.

The developer of each block determines the relevant performance indications for that function. Figure 6 shows the location of the KPI dashboard link for a service block that exposes such metrics. In this release, the KPIs are exposed as a Grafana dashboard that collects and visualizes all the metrics associated to a specific block.

Furthermore, deployed devices also have an assigned Grafana dashboard that shows its telemetry, such as CPU usage or power consumption.

When execution is augmented with HDDL-acceleration, power consumption and CPU usage drops significantly, as shown in Figures 7, 8 and 9.

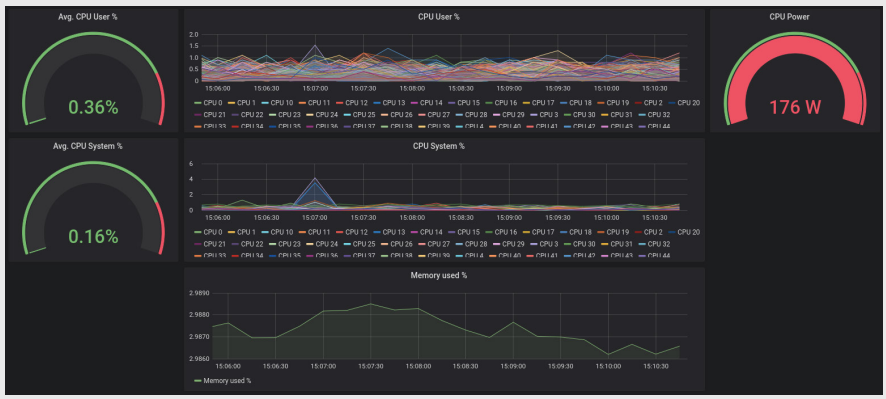

Figure 7. Telemetry KPIs from the idle node

Figure 8. Telemetry KPIs from node running object detection app in

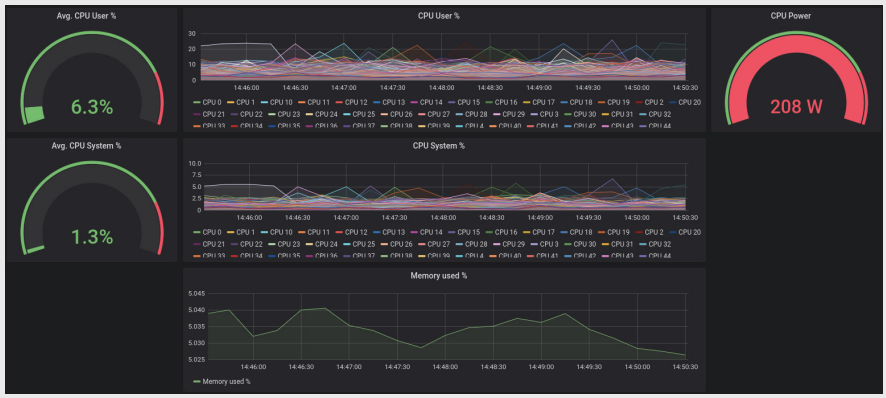

Figure 9. Telemetry KPIs from node running object detection app in HDDL

Conclusion

By leveraging Intel Smart Edge Open to deploy hardware and software resources using the industry-standard OpenVINO toolkit, Nearby One can easily support artificial intelligence and machine learning workloads while addressing run time cost and power considerations. Nearby One includes a marketplace of pre-integrated applications and virtual network functions that are easily deployed and managed, including support for hardware accelerators such as Intel Movidius Vision Processing Unit (VPU) for applications that require acceleration of video analytics workloads.

The result is a scalable platform that enables AI/ML workloads to be deployed at the edge of the network, resulting in lower latency and bandwidth costs. Industry-standard container and application orchestration allows workloads to be easily deployed and scaled as required. These features, combined with hardware acceleration capabilities, result in a platform that addresses the specific requirements for emerging augmented reality and visual analytics solutions.