ZEDEDA

SECURE ZERO TOUCH KUBERNETES ORCHESTRATION FOR THE DISTRIBUTED EDGE

ZEDEDA Kubernetes White Paper

Introduction

Hosted by the Cloud Native Computing Foundation (CNCF), Kubernetes is a powerful, highly pluggable containerization framework with massive industry adoption. Over the past several years, Kubernetes has become the preferred open source solution for managing containerized applications across multiple hosts, providing foundational mechanisms for deployment, maintenance, and scaling of applications and enabling CI/CD practices for DevOps. There is a clear trend for the majority of workloads to be containerized in the cloud and to increasingly leverage Kubernetes for standardization, redundancy, and scale-out.

The edge is emerging as the next frontier of computing as the number of edge devices and their capacity to produce immense volumes of business-critical data grows. Factors including bandwidth cost, latency, security and privacy are driving the need to analyze data closer to the source and organizations are looking to replicate the benefits of cloud-native principles at the edge for IoT, AI, 5G, network virtualization, and security use cases. It's a natural evolution for Kubernetes to be extended from centralized data centers to edge use cases, however deploying Kubernetes clusters on edge infrastructure is a challenge due to available compute footprint, geographic distribution, security, and the inherently diverse nature of hardware, software and skill sets.

Furthermore, organizations face the challenge of extending cloud-native development principles to the edge while accommodating their legacy software investments and the unique needs of Operational Technical (OT) environments. Modern day cloud-native practices involve continuous development, testing, integration, deployment, and monitoring of software applications throughout its development life cycle, however these practices are often not compatible with the need for uptime and stability in industrial processes.

This white paper explores the benefits and unique challenges of deploying Kubernetes at the edge, including the unique needs of both IT and OT professionals. It then outlines how ZEDEDA's subscription-based orchestration solution greatly simplifies provisioning, managing and securing Kubernetes clusters at the distributed edge on choice of hardware at scale.

Contents

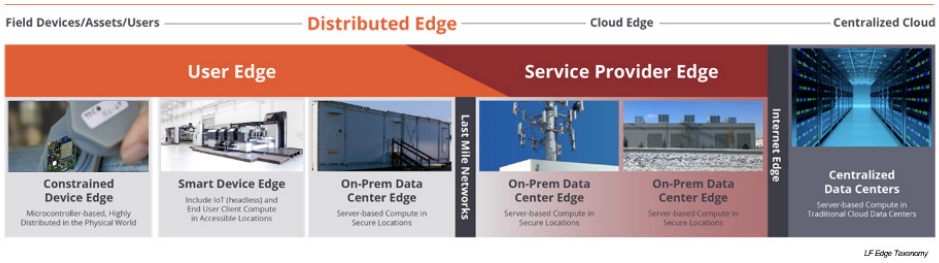

Fig. 1: The Edge Continuum

Defining the Edge Continuum

When considering the utilization of Kubernetes at the edge, it's first important to recognize that the edge is a continuum of deployment locations spanning devices in the physical world to the centralized cloud. The LF Edge taxonomy breaks this continuum down into several key paradigms based on inherent technical tradeoffs: 1) Whether a piece of hardware has the resources available to support abstraction in the form of virtualization or containers or not, 2) Whether it is on a LAN or a WAN relative to the users/processes it serves (compute for time-critical workloads will always be deployed on a LAN), and 3) Whether the edge compute hardware is deployed in a physically-secure data center or has no defined security or network perimeter (Fig. 1).

Examples of distributed edge computing nodes include an industrial PC embedded in a connected machine, an IoT gateway on a factory floor, and a small cluster of servers at the fringes of the data center. These deployments can be in locations such as a drone, truck, wind turbine, oil rig, factory floor, retail store, and beyond. Extending Kubernetes to these environments is a natural evolution, however it's important to carefully consider implementation details. Where processing and applications are deployed across the continuum is based on a balance of performance, uptime, safety, and cost.

Why Kubernetes at the Edge

As with the cloud, the primary reasons to leverage Kubernetes at the distributed edge is clustering hardware and applications for redundancy and scale-out. This is key for many edge use cases in which uptime is paramount, for example on a factory floor where a pause in operations can mean tens of thousands of dollars a minute in lost productivity. Additional Kubernetes features (e.g., aspects of High Availability / HA and clustered storage) will make their way into distributed edge use cases over time as use cases evolve, however there will be an inherent tension between the value of these features at the edge and the resource constraints imposed by available infrastructure.

Related to this, it's important to realize that Kubernetes is not a panacea for powering edge infrastructure, rather an important tool in a collection of various deployment models that must be matched to the target use case. ZEDEDA's scale-up approach is a critical enabler for this because the solution supports any combination of application deployment models. This ensures that enterprises don't need the overhead of Kubernetes for more simple use cases such as deploying a single Virtual Machine (VM) or Docker container, or when the chosen edge hardware simply can't support the Kubernetes footprint.

Considerations for Deploying Kubernetes at the Edge

In a perfect world, we'd have one orchestration solution capable of deploying and managing infrastructure and applications spanning the entire edge continuum, however the inherent technical tradeoffs dictate necessarily different tool sets with similar principles applied. While it may be tempting to extend your existing data center tools outside of the data center, they quickly start to break for reasons including being too resource intensive, lacking an adequate Zero Trust security model, requiring specialized skill sets for deployment and management that are not commonly found in the field, not comprehending the scale factor or pre-supposing a near-constant network connection to the central console which often isn't the case in remote edge environments. Our goal is to eliminate such security concerns for the deployment of hardware and applications at the distributed edge by providing a holistic approach built from the ground up.

Deploying Kubernetes at the distributed edge involves some unique logistical requirements as well. For starters, when deploying Kubernetes in centralized cloud data centers, the physical infrastructure is managed by the cloud provider. However, at the distributed edge, both the physical infrastructure (e.g., compute, network and storage) and the virtual Kubernetes clusters are often provisioned and managed by the end user.

How this infrastructure is deployed is also very different. In a cloud model, an administrator can simply spin up a multi-node Kubernetes cluster, not caring about the physical location of these resources. Meanwhile, the physical location of nodes in a given edge cluster is a major consideration and admins need to dictate the specific edge nodes they want to form clusters with. Related to this, edge applications also have a location affinity based on connectivity requirements and available hardware footprint.

Security, as we have discussed above, is largely an IT concern. However, secure infrastructure is paramount for hosting OT applications in order to maximize uptime and safety and ensure data provenance from critical processes. While not covered in this white paper, the ZEDEDA solution is architected to maximize uptime through mechanisms for automatic failback and recovery after failures or configuration mistakes, combined with state-of-the-art availability approaches in ZEDCloud.

Location Affinity for Edge Apps, Driven by Connectivity Requirements

Edge apps process data in the field to make rapid decisions and filter data before backhaul to centralized resources. These apps often ingest data from sensors and systems that communicate through non-IP based connections (e.g., serial, BLE, LoRa, USB), making it necessary for these apps to be scheduled on edge hardware with the required connectivity.

Location Affinity for Edge Apps, Driven by Hardware Requirements

Some edge apps might need some special arrangements with respect to the capability of the underlying node. For example, an Edge AI application that needs to be scheduled on a node with a GPU accelerator. As with connectivity, it's important that the infrastructure layer collaborates with the Kubernetes orchestration layer to make this happen.

Importance of Concurrent Support for Legacy Applications

Additionally, it's important to be able to have concurrent support for legacy applications deployed in VMs alongside Kubernetes at the edge. Resource availability is practically unlimited in centralized data centers, hence VMs and Kubernetes workloads can be deployed and managed in different sets of environments that are networked together. Meanwhile, at a given edge location there are often a limited number of nodes that are not always easily networked together. Therefore it is important that the infrastructure allows users to consolidate VM-based applications and Kubernetes workloads on the same edge hardware.

Finally, the typical pricing for commercially-supported Kubernetes orchestration solutions optimized for the data center doesn't work for the scale of deployments at the distributed edge. Pricing must be aligned to the cost and capability of the underlying edge infrastructure.

With the right edge strategy, providers can tap into the inbuilt recoverability and dynamic scheduling that Kubernetes brings to the table to offer resilient and available edge applications, while meeting the unique requirements for deployments outside of traditional data centers.

Starting with the Right Foundation: EVE-OS, the Android of the Edge

Compared to orchestration solutions suited for the data center, ZEDEDA has built its subscription-based orchestration solution from the ground up to be optimized for deploying and managing distributed edge computing hardware and applications. ZEDCloud - our easy-to-use, cloud-based orchestration UI - leverages the bare metal EVE-OS from Project EVE deployed on distributed edge nodes. EVE-OS is a lightweight, secure, open, universal, and Linux-based distributed edge operating system with open, vendor-neutral APIs for remote lifecycle management. The operating system can run on any hardware (e.g., x86, Arm, GPU, FPGA) and leverages different hypervisors and container runtimes to ensure policy-based isolation between applications, host hardware and networks.

EVE-OS is optimized to run on the lightest compute footprint possible while still supporting application abstraction in the form of virtualization and containerization (currently 512MB of memory and two CPU cores). The result is a lowest-common denominator resource footprint for Kubernetes with target hardware spanning a single edge gateway in the field to a small server cluster at the fringes of an on-prem data center, meeting the Kubernetes trend as it extends down from centralized data centers in the cloud.

ZEDEDA's state-of-the-art and market-leading Zero Trust security architecture (Fig. 2) assumes that edge nodes distributed in the field are physically accessible, in addition to not having a defined network perimeter. Features include support for silicon-based root of trust, measured boot, remote attestation, crypto-based ID (eliminating local device login as a threat vector), full disk encryption, remote disablement of I/O ports (e.g., USB, serial), distributed firewall, and more. Distributed firewall capability enables secure routing of data between edge applications and both on-prem and cloud resources based on network-wide policies.

Fig. 2: EVE-OS Zero Trust Security Components

EVE-OS also features volume instance encryption. This enables edge application data to be encrypted at rest, with the decryption keys locked with measurements from hardware-based root of trust (e.g. TPM). This means that unless EVE-OS establishes proof of its software state, the trust anchor will not release the decryption key, and EVE-OS will need to go through remote attestation with ZEDCloud to get the key back. This mechanism protects app data from being compromised, which is especially critical at the distributed edge where edge nodes can often be accessed physically.

For more detail, refer to ZEDEDA's Whitepaper on Zero-Trust Security at the Distributed Edge.

While most orchestration solutions targeting distributed edge use cases only support containers, the bare metal EVE-OS architecture supports any combination of Virtual Machines (VMs) and native Docker containers while integrating with customers' existing CI/CD workflow. Support for VMs in addition to containers and Kubernetes clusters is critical because it enables deployment of any combination of legacy Windows-based applications (e.g., SCADA, HMI, Historian, VMS, PoS), monolithic Linux-based images, and container runtimes such as Docker/Moby, Azure IoT Edge and AWS Greengrass.

Other solutions for orchestrating distributed edge computing are typically agent-based, touting that they can be used with any edge operating systems. However, without tight integration between the agent and underlying OS, and significant hardening of the OS itself, these solutions are vulnerable to hacks and there is a high likelihood of bricking the device in the field during an update. EVE-OS effectively integrates the agent into the bare metal foundation that has deep access into the hardware below.

Furthermore, the EVE-OS architecture features A/B partitions to ensure maximum uptime and an eventual consistency model that ensures autonomy in the field regardless of connectivity status to the cloud. Compared to data center solutions that assume a constant connection between controller and server, EVE-OS assumes that this connection will be periodically lost and works to update to the latest software revisions whenever connectivity is regained. Another key difference is that edge nodes always call home to their centralized controller (including through NATs and firewalls) to ask for updates, compared to the data center solutions in which the controller reaches out to the servers under management.

Finally, the hosting of Project EVE within the Linux Foundation's LF Edge organization provides vendor-neural governance. This is important because it provides a high level of transparency and creates an open anchor point around which to build an ecosystem of hardware, software and services experts. The open, vendor neutral API within EVE-OS prevents lock-in and enables anyone to build their own controller. The open EVE-OS API also provides an anchor point to unify an ecosystem of edge computing hardware and software. In this regard, you can think of EVE-OS as "the Android of the Edge".

For more detail on the overall ZEDEDA solution please refer to the ZEDEDA Architecture Whitepaper.

Selecting the Right Kubernetes Distribution for the Edge

It's important to choose the right Kubernetes distribution when deploying applications at the distributed edge. While K8s is the standard for centralized data centers, there are a number of additional Kubernetes variants emerging - for example K0s, K3s, MicroK8s, and KubeEdge. ZEDEDA selected the K3s distribution, pioneered by SUSE-Rancher, as an initial focus because it strips out features that are unnecessary for deployments outside of centralized data centers. In this white paper, we outline how ZEDEDA integrates with K3s as a baseline implementation, however ZEDEDA can support any Kubernetes distribution and 3rd party workload management solution equally (Fig. 3).

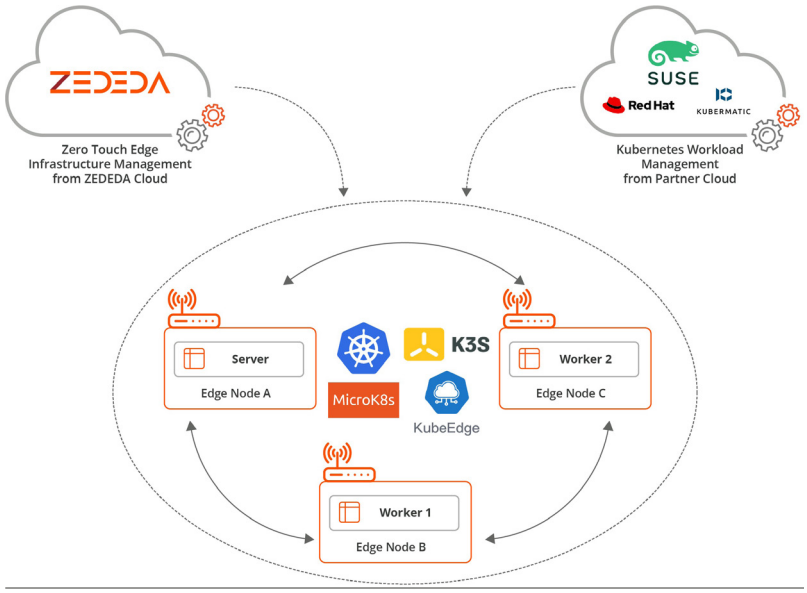

Fig. 3: Zero Touch Provisioning of Kubernetes Clusters by ZEDEDA

ZEDEDA's Orchestration Solution for Kubernetes at the Distributed Edge

ZEDEDA's orchestration solution greatly simplifies Kubernetes infrastructure management, security and visibility as customers look to deploy Kubernetes clusters outside of centralized data centers. ZEDCloud has a simple and intuitive UI along with comprehensive APIs that abstract all the complexities of provisioning Kubernetes without requiring specialized IT skills in the field. ZEDEDA provides a rich set of visibility for both Day One and Day Two management of edge nodes (e.g., CPU, memory, disk, and network usage, network flow visualization), clusters and applications. This is both by single edge node or application and across a fleet deployment.

Zero Touch Kubernetes Orchestration

ZEDCloud features powerful capabilities for remotely provisioning and managing edge hardware and applications at hyperscale. The built-in app marketplace features a catalog of popular edge applications for industrial connectivity, networking, security, analytics, data management, and DevOps. The catalog also includes Kubernetes distributions, for example, a pre-validated K3s Virtual Machine (VM) image, built on Ubuntu with all the required binaries pre-installed. As with any other app in the marketplace, a Kubernetes VM can be bulk-deployed to target edge hardware with a single click. Upgrading to a newer app version in the marketplace is as simple as pointing the deployment manifest to a new binary. Meanwhile, Kubernetes instances deployed on a fleet of distributed edge hardware can be updated with a single click through the bulk app management workflows in ZEDCloud.

ZEDEDA's focus is initial cluster bring-up and management of both Kubernetes distributions and the hardware below. This separation of underlying infrastructure management and security from the application plane (Fig. 4) is critical for consistency when deploying clusters in heterogeneous edge environments including a mix of hardware, additional software (both cloud-native and legacy) and stakeholders (e.g., DevOps, network admins, field technicians, 3rd-party service providers).

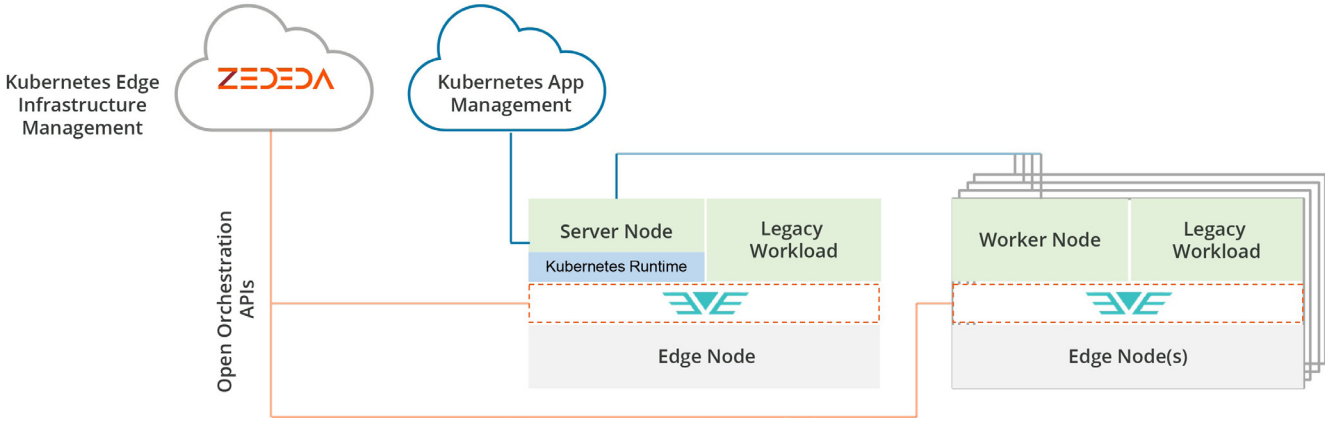

Fig. 4: Kubernetes Orchestration for the Distributed Edge

Cluster Configuration

Part of ZEDEDA's Zero Trust security architecture is prohibiting users from directly logging into edge hardware running EVE-OS in order to prevent unauthorized tampering. As such, Kubernetes cluster configuration is entirely performed through ZEDCloud.

Configuring a cluster within ZEDCloud simply requires selecting the target edge hardware and a few provisioning parameters. The entire process of deploying K3s and bringing up the cluster is then automated in the background.

The configuration UI also provides the option of specifying the URL of a third-party Kubernetes deployment manifest, for example the SUSE import URL, in the Cluster GUI to import their K3s distribution. This facilitates automatic enrollment of the newly created cluster on their management portal for subsequent deployment and management of Kubernetes applications.

Selecting Cluster Provisioning Options

This section covers the various parameters to specify when configuring a cluster within ZEDCloud, including screenshots of the cloud-based UI.

Edge Hardware Selection

Administrators use a flexible tagging mechanism in ZEDCloud to match and select the edge hardware for deploying in the cluster. An example would be to add a tag "ClusterPool: SJC_Store_1" when provisioning new edge hardware located in a retail store, and mention the same tag while creating the cluster (Fig. 5).

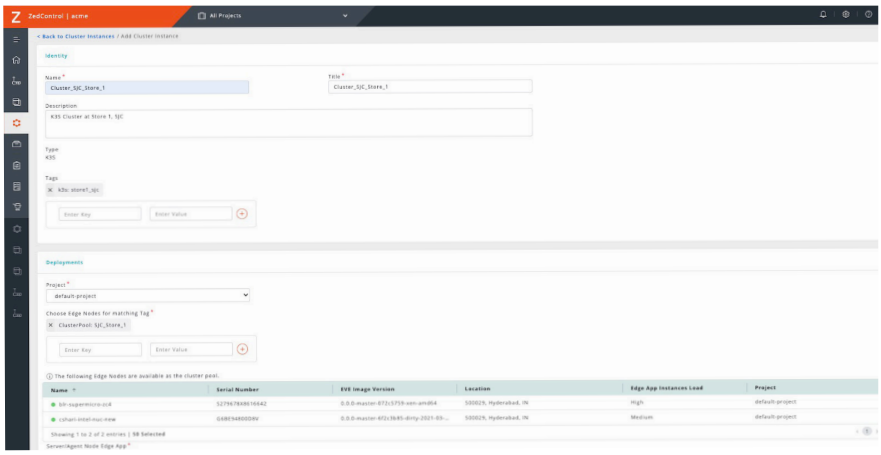

Fig. 5: List of Edge Nodes Selected by Tag "ClusterPool:SJC_Store_1"

ZEDCloud automatically selects edge hardware matching this tag, and brings up the cluster on that infrastructure. Adding edge nodes to this cluster is as simple as tagging the new hardware with this tag value, and ZEDCloud will automatically add the new hardware to the cluster - greatly simplifying pre-provisioning of the cluster.

Cluster Composition

Next the admin selects the number of server and agent nodes required in the cluster, with a potentially different minimum and maximum values for each (Fig. 6). ZEDCloud will then bring up cluster nodes on the edge hardware and match this composition.

If a minimum number of edge nodes for a cluster is specified, ZEDCloud will wait until it collects this number of nodes before bringing up the cluster. Similarly, a maximum number can be specified to cap the number of edge nodes selected for that cluster.

Third-Party Configuration for Kubernetes Management

A fully functioning Kubernetes cluster running application workloads requires two parts to be fulfilled:

1. Cluster orchestration
2. Workload orchestration

ZEDEDA's Kubernetes integration can support any third-party Kubernetes cluster management platform (such as SUSE-Rancher), with the initial configuration being provided during the initial creation of edge clusters. Through the combined infrastructure and application orchestration capabilities, users have the ability to remotely manage the entire lifecycle of both the Kubernetes runtime and the underlying hardware at scale. A significant benefit of this architecture is that ZEDEDA provides a consistent experience for securing and managing Kubernetes infrastructure at the edge regardless of choice of cluster management solution (Fig. 3).

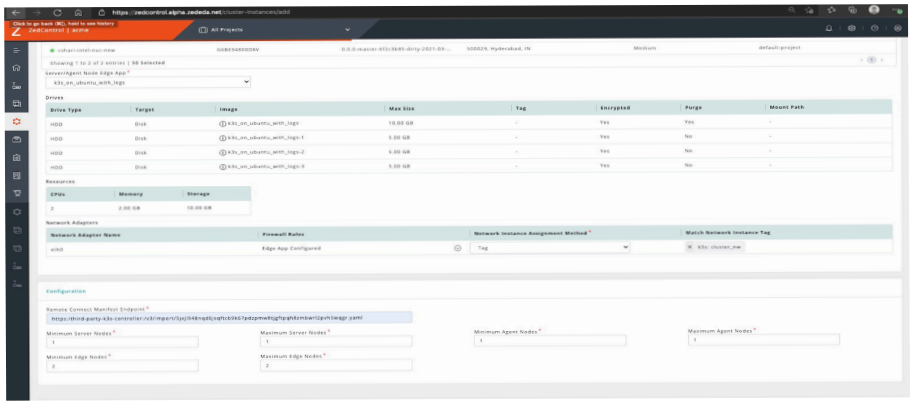

Fig 6: Policy Parameters for Creating a New Edge Cluster

Granular, Software-Defined Network Connectivity

Depending on the edge hardware capabilities and desired network segmentation, ZEDEDA offers flexibility where admins can specify member nodes of a cluster to be on the same network used to communicate with the cloud, on an air-gapped network together with downstream devices like sensors and machines, or on both networks. To enable these different scenarios, ZEDCloud provides a comprehensive list of network connectivity options. An admin simply creates "Network Instances" in ZEDCloud which direct EVE-OS running on each edge node to route traffic based on policy, which in turn maintains proper network segmentation.

The default network instance that is automatically provisioned on EVE-OS provides unique IP addresses to each edge application, along with suitable network address translation (Fig. 7).

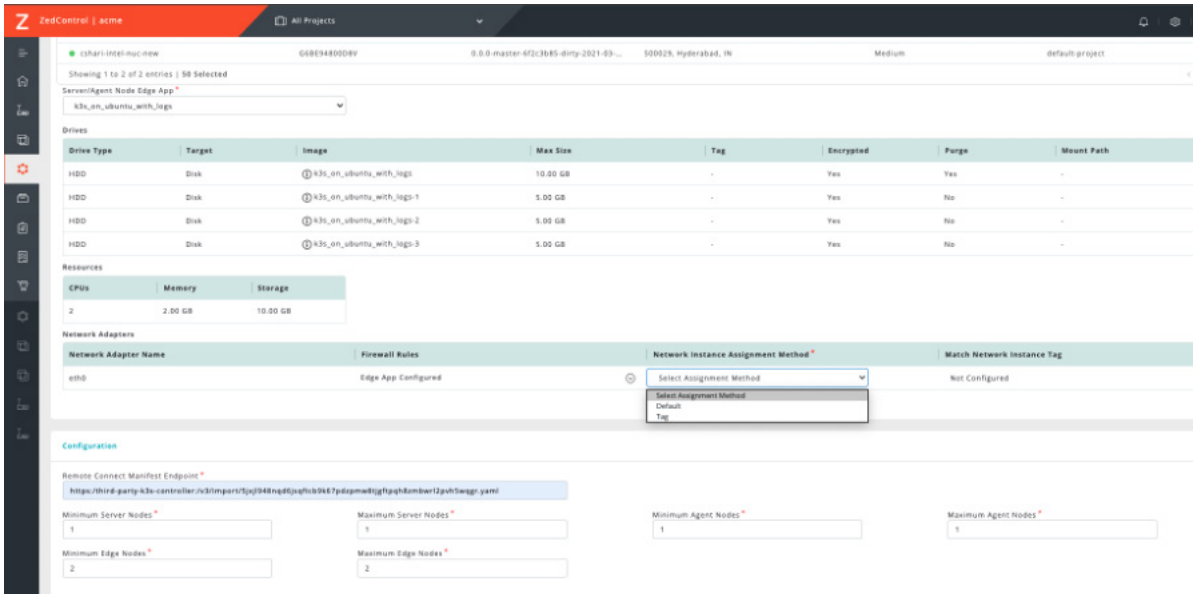

Fig 7: Network Connectivity Options for a New Edge Cluster

If desired, an admin can create network instances for node-to-node communication, locally on the same box or via a switch on different edge nodes. During cluster creation the admin simply specifies the tag value used by these network instances (Fig. 8). ZEDCloud will take care of matching and finding the right network instance, and the cluster members will be placed on that network automatically.

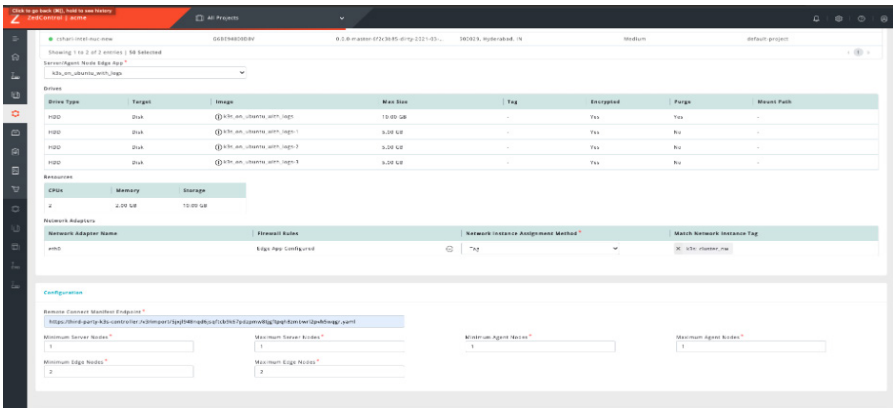

Fig. 8: Configuration of Custom Network Connectivity for a New Edge Cluster

Monitoring and Managing Kubernetes Infrastructure at the Edge

ZEDEDA provides deep visibility into the state of Kubernetes clusters deployed in the field, spanning the underlying hardware to the resource-utilization of the Kubernetes nodes. Using the ZEDCloud UI, admins can easily see the current state of any edge cluster, the roles of the Kubernetes members and map them to all the edge nodes they are deployed on.

The admin also has access to granular telemetry data for network traffic, CPU and memory usage of all individual nodes making up clusters distributed in the field. ZEDCloud leverages the powerful distributed firewall capabilities of EVE-OS to enable admins to open up only the edge hardware ports required for Kubernetes operation, thereby blocking any unwanted traffic. ZEDCloud records and displays a historical chronology of events observed on cluster instances and admins have the option of recording and displaying the logs from the Kubernetes applications running on these instances. Finally, admins have the ability to deactivate and activate a specific edge cluster at any time.

Details for each of these capabilities are illustrated in Figures 9-13.

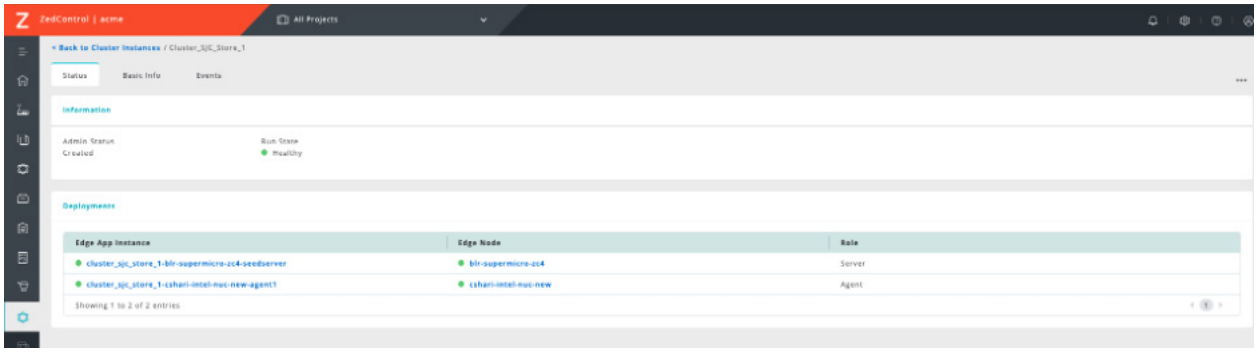

Fig. 9: Status of a Kubernetes Cluster

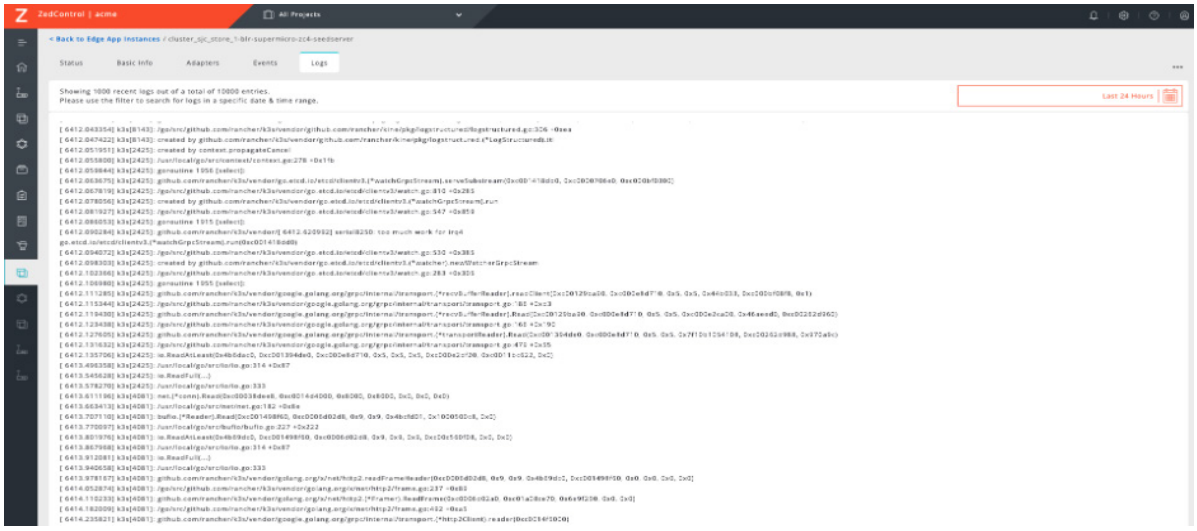

Fig. 10: Logs from Deployed Kubernetes Instance

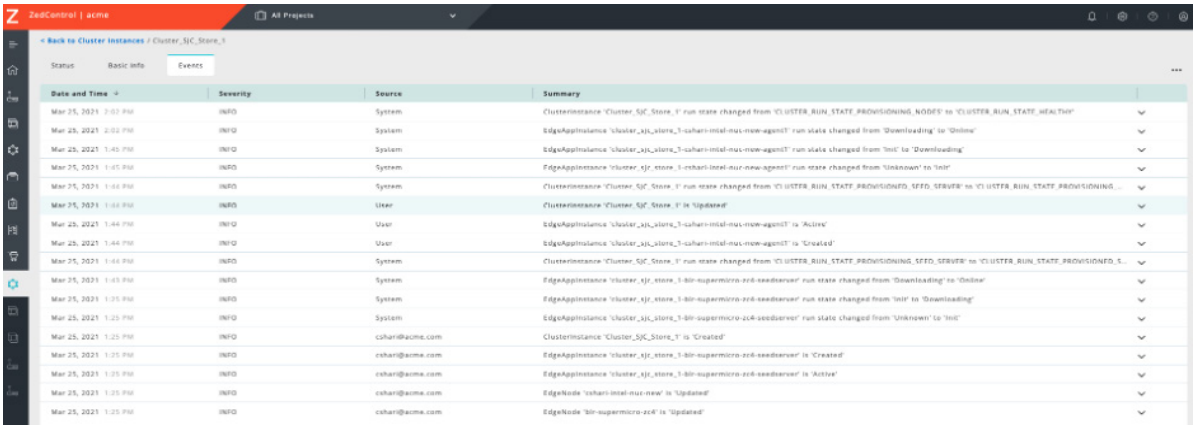

Fig. 11: Events from Kubernetes Clusters

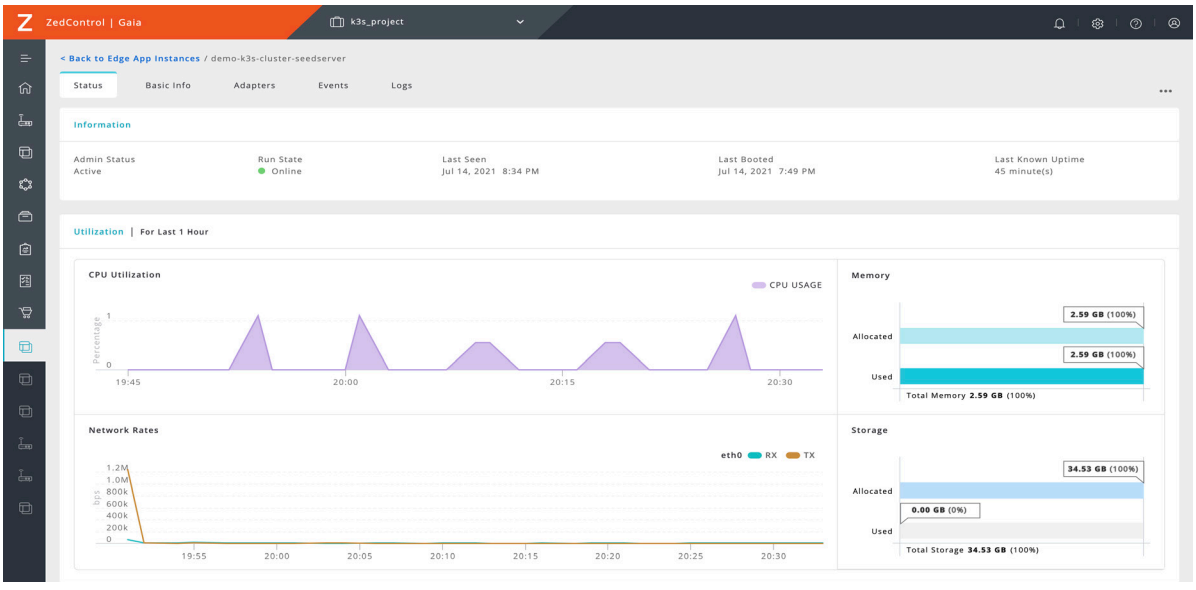

Fig. 12: CPU, Memory, Storage, and Network Usage for a Kubernetes Node

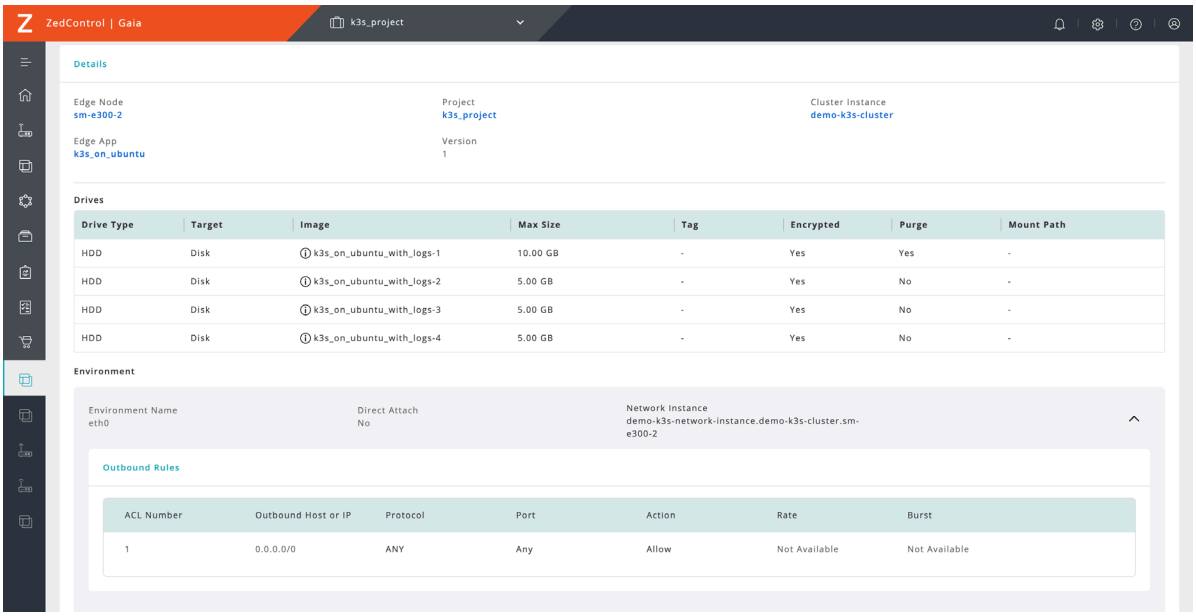

Fig. 13: Outbound Firewall Policies for a Kubernetes Node

Summary

As edge computing deployments grow at a rapid pace, along with the capacity to produce immense volumes of business-critical data, the need for lightweight applications to analyze data closer to the source has become ever more important. The edge is emerging as the next frontier of computing and organizations are looking to extend cloud-native principles to the field, however there have been several significant barriers. Solutions initially built for the data center cannot seamlessly be retrofitted to meet the unique needs of the distributed edge in areas of available compute footprint, security, deployment, scale and required skills - including Kubernetes expertise.

ZEDEDA offers a simple and secure Zero Touch edge orchestration solution that's built on an open source foundation and supports any Kubernetes distribution and 3rd-party cluster controllers, in addition to deploying and managing any combination of additional VMs and native Docker containers. This provides customers with maximum flexibility to scale and future-proof their operations, without risk of vendor lock-in and while also continuing to leverage legacy apps (e.g., Windows-based) in parallel.

ZEDEDA's solution is unique and purpose-built to address the needs of distributed edge computing by starting with a lowest-common-denominator foundation that scales up to bridge the edge to the cloud paradigm.

Visit www.zededa.com for more information.

About ZEDEDA

ZEDEDA, the leader in orchestration for the distributed edge, delivers visibility, control and security for edge computing deployments. ZEDEDA enables customers the freedom of deploying and managing any app on any hardware at scale and connecting to any cloud or on-premises systems. Distributed edge solutions require a diverse mix of technologies and domain expertise, and ZEDEDA provides customers with an open, vendor-agnostic orchestration framework that breaks down silos and provides the needed agility and future-proofing as they evolve their connected operations.

Customers can now seamlessly orchestrate intelligent applications at the distributed edge to gain access to critical insights, make real-time decisions and maximize operational efficiency. ZEDEDA is a venture-backed Silicon Valley company, headquartered in San Jose, CA, with teams based in Bangalore and Pune, India and Berlin, Germany. For more information, contact info@zededa.com.

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current ZEDEDA product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from ZEDEDA and its affiliates, suppliers or licensors. ZEDEDA products or services are provided "as is" without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of ZEDEDA to its customers are controlled by ZEDEDA agreements, and this document is not part of, nor does it modify, any agreement between ZEDEDA and its customers.