ActualTech Media

PLATFORM9

Distributed Edge with Managed Kubernetes

Joep Piscaer

CONTENTS

IN THIS PAPER

Edge computing unlocks many opportunities, but has its own technical challenges to do right. The Platform9 distributed architecture of its Managed Kubernetes service helps telcos, retailers, manufacturers, enterprises, ISVs, and SaaS vendors unlock the potential of their applications across edge computing use cases.

This paper dives into the distributed architecture of the Platform9 Managed Kubernetes control plane, showing you how to use the platform to manage edge locations, from one to many, in a policy-based way, reducing operational cost, and making onboarding of new edge locations seamless and consistent.

Each generation of networking presents unique challenges that are met by unique solutions. Current application architectures connect small computers at the edge—mobile apps, Internet of Things (IoT) devices, retail point-of-sale systems, and so forth—with hubs in the cloud. These environments are characterized by:

- Compute-intensive requests sent by mobile devices into the cloud, where a vendor is expected to handle the requests and return results quickly
- Cellular networks to handle most requests
- Modular, scalable worker nodes that handle requests and are controlled by Kubernetes, the most popular tool currently for distributing requests across worker nodes

This network architecture calls on companies hosting applications to place worker sites on network endpoints, so they can quickly accept requests from mobile device users, calculate the answers to the requests, and return them to the mobile devices.

A New Architecture for Responding to Compute-Intensive Applications

During the past 10 to 15 years, companies have gotten used to accepting data from mobile users or edge devices into a centralized data center. This data center run by the company owning the app determines the proper response and returns it to the edge.

> Edge computing brings compute power closer to end users and their devices, essentially decentralizing some of the compute capabilities of centralized public cloud offerings.

But this simple model has turned out to be inadequate for delivering the performance needed as the complexity of calculations grows. Cellular networks, especially those employing some form of 5G, have alleviated bandwidth limitations between the cellular tower and the edge, whether the edge is an end user on a mobile device or an IoT device reporting conditions in the field. But a significant bottleneck remains between the cell tower and the company's systems.

A new architecture has therefore developed, based on distributing the work to intelligent systems at the collection points provided by enterprises. Much of the information and work is never seen by the sites run by app developers. Instead, enterprise companies launch local containers to handle requests. Kubernetes instances are also launched at the endpoints to start up and monitor worker processes.

> With an architecture that requires less hardware at the edge, savings scale linearly with the number of edge locations.

This paper describes the new architecture and how to take advantage of it to offer applications that respond gracefully to user requests or reports from devices in the field.

Bandwidth and Compute Power in a Distributed Architecture

Edge computing brings compute power closer to end users and their devices, essentially decentralizing some of the compute capabilities of centralized public cloud offerings.

Recent developments in connectivity, such as the increased bandwidth of 5G networks, have increased the opportunities for communication between edge locations and end users. This increase in connectivity allows an enormous growth of data and unlocks many new use cases, from image and video processing and voice recognition to running factories and retail locations over 5G instead of Wi-Fi. Fast response time is critical in the new applications. Mobile users are impatient, and IoT devices must make real-time decisions.

While the connection between the end user and the edge location enjoys ample bandwidth, responses from the centralized cloud or data center can't keep up in speed. Edge locations in many cases are remote and have limited connectivity, but require complex processing for huge

amounts of locally generated data. That in turn creates an architectural challenge to bring compute resources where they can exploit the increases in bandwidth at the edge.

A DISTRIBUTED ARCHITECTURE TAILORED TO KUBERNETES FOR THE EDGE

Transporting data from the edge to the central cloud or core data center for processing doesn't make sense, especially if there's a large amount of data to be transferred and fast response times are required. Processing at the edge is more cost-effective. In this architecture an edge location, such as a 5G radio tower, runs one or more clusters of worker nodes. The edge location sends only processed data that's useful for business-related analytics to the central repository.

This new architecture processes data close to where it's generated, with a few core data centers or cloud regions acting as the brains of the operation.

In this scenario, the edge locations themselves need sophisticated task management for thousands of simultaneous processes that are set up and torn down quickly. Kubernetes is the current industry standard for this kind of process management. That means starting up Kubernetes worker nodes at the edge locations to run the data processing applications locally. By running only the absolutely necessary workloads at the edge, companies can reduce costs associated with maintaining centralized data centers and transferring data from the edge.

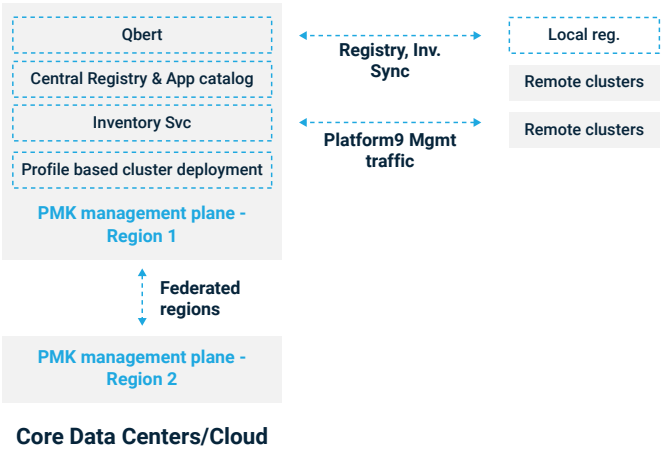

Figure 1: Distributed edge platform overview

Application providers are now dealing with hundreds to thousands of edge locations, or even more. With an architecture that requires less hardware at the edge, savings scale linearly with the number of edge locations.

Of course, this concept applies to more than just data-processing applications. 5G, as well as faster and more cost-effective endpoints (consumer and industrial IoT devices alike), are creating new use cases that generate far more data than ever before, such as video feeds from CCTV systems, telemetry information from industrial IoT devices, and interactions generated by apps and games on consumer phones.

Central Management Is Still Needed

Although the modern, distributed applications described in this paper process data at the edge, application providers still need central control and visibility into the processing. Platform9 Managed Kubernetes (PMK) delivers Software-as-a-Service (SaaS) Kubernetes cluster management and simplicity of operations like native public cloud services but using upstream open source stacks that are deployed and operated on a wide range of on-premises (VMware, bare metal), public clouds (AWS and Azure), and edge infrastructures.

The federated, distributed architecture of PMK provides a consistent experience across regions, while being centrally managed and resilient against connectivity and bandwidth issues. The architecture supports multiple regions in a hub-and-spoke model. **Figure 1** shows an overview of the federated architecture supported by PMK. Multiple regions work independently. Each management plane region acts as the central hub and brains of a region. The management plane defines policies for all the edge processors, with the federation of configuration templates and apps.

The management plane is where DevOps engineers manage the entire operation. There, they store container images and inventory caches of remote locations. Synchronization ensures eventual consistency to regional and edge locations automatically, regardless of the number of locations.

PROFILE-BASED MANAGEMENT

Platform9 facilitates the scaling of edge computing to thousands of edge data centers by grouping them so that similar data centers can be managed centrally through a single policy known as a *profile*. This relieves IT staff from managing each data center individually. Instead, the staff just defines a small number of profiles and indicates exceptions to policies where needed for a particular data center. Each edge location, such as radio towers, warehouses, and retail locations, runs its own worker nodes and containers.

Profile-based cluster management makes it easier to deploy identical remote clusters and configurations, from a single profile, instead of managing each remote cluster separately. That minimizes configuration drift, while still being able to apply unique configurations where needed.

This feature helps standardize application and container deployment across clusters, making the onboarding of new edge regions easy and consistent. This allows the day-to-day operations work to scale non-linearly, making the most of each engineer's time. The feature allows administrators to manage a large number of edge locations without additional work.

As you saw in **Figure 1**, the "core" registries and catalogues are synced with remote locations, which cache this information to reduce bandwidth and remove any connectivity dependencies they might have to the core data centers.

> Profile-based cluster management makes it easier to deploy identical remote clusters and configurations, from a single profile, instead of managing each remote cluster separately.

As a result, deploying and scaling applications at the edge isn't dependent on the core data centers, but can be handled locally while still receiving periodic policy-based changes to keep configurations consistent and in sync with the centrally defined policies.

The architecture deep dive in **Figure 2** shows the entire architecture, from endpoints to core data centers. We'll dive into each area of this diagram in the next sections.

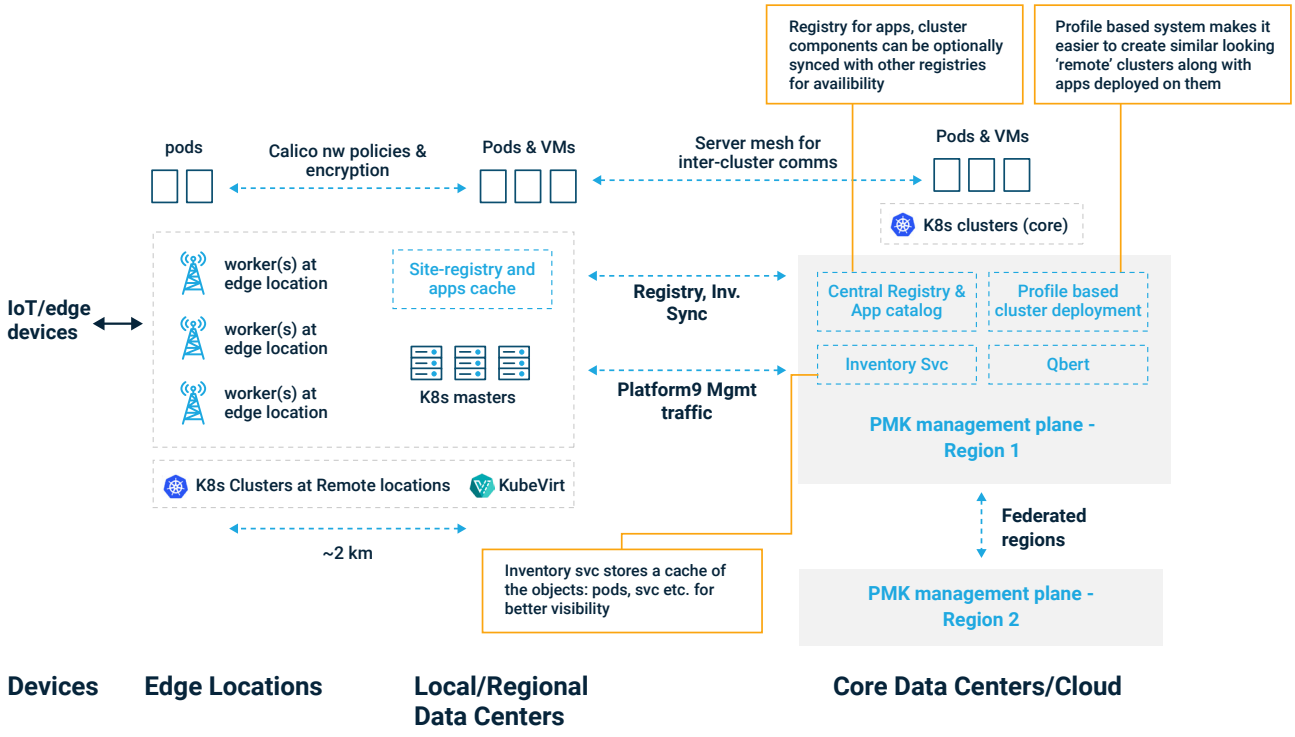

Figure 2: Distributed edge platform deep dive

CORE DATA CENTERS ARE THE BRAINS OF THE OPERATION

The core data centers run the management plane, central container registry, and App Catalog, as well as inventory services and policy-based deployment and configuration tools for cluster management.

The profile-based system makes it easier to create identical remote clusters and to deploy applications to remote clusters consistently, keeping configuration drift to a minimum and ensuring maximum application compatibility.

> Compliance can be enforced even in untrusted physical environments through service mesh and micro-segmentation technologies.

Core data centers are federated across regions for consistent deployment in large scenarios. A single management plane region is able to handle up to 100 Kubernetes clusters with up to 100 nodes per cluster. The central inventory service caches a representation of the remote sites for better visibility.

CONSISTENT NETWORKING AND GRANULAR SECURITY

Networking and security policies are deployed consistently from core to edge, making sure application deployments are secure and compliant. Compliance can be enforced even in untrusted physical environments through service mesh and micro-segmentation technologies. A distributed Kubernetes architecture combines all of these locations—public cloud VPCs, core data centers, edge data centers, and edge locations—in a single, interconnected mesh.

LOCAL DATA CENTERS ENSURE INDEPENDENT OPERATION

Each local data center can deploy and scale applications independently of the core data centers, creating geographically separated "cells" that can run without a continuous connection to the core data centers. Each local or regional location runs one or more Kubernetes master cluster nodes, which manage worker nodes across that location. This way, each edge location runs only the absolute necessary hardware—often low-cost, low-power, and low-maintenance machines—while regional hubs coordinate application deployment and resilience across their local region.

Applying This Architecture to Retail, Manufacturing, and SaaS

A large number of use cases can benefit from the distributed Kubernetes architecture described in this paper. Many enterprises see the same growing need for edge computing as their revenue streams become more and more digitally focused. The enterprises see more need for connecting cloud services to where their users are, regardless of whether those users are consumers, other businesses, or IoT-enabled devices.

Retailers are innovating and transforming the shopping experience to be seamless across online and in-store visits. This requires connecting cloud and on-premises locations to work together, offering buyers a consistent experience. The new architecture unlocks new revenue streams and increases existing value streams by accelerating the rollout of digital store concepts and by increasing in-store automation and the use of innovative retail software.

The distributed Kubernetes architecture helps retailers deploy new stores quickly and consistently. It reduces per-store operational IT costs both for onboarding new stores and for continuous operation, including lifecycle management and seamless software upgrades of running containers.

Manufacturers are replacing Wi-Fi and legacy wired networks with 5G wireless connectivity for factory floors and manufacturing plants to connect IoT and edge devices. That means they need to connect 5G endpoints with worker nodes for data processing, central management, and factory process engineering. Putting the compute power at the endpoints minimizes costs and operational burden.

Similarly, SaaS and independent software vendors (ISVs) are starting to use edge computing to improve their users'

experience, decrease time to market, and reduce costs and operational burdens. The PMK distributed architecture helps them deploy their software to edge locations, decreasing latency and bandwidth requirements, while consistently deploying the application to many locations simultaneously.

> The distributed Kubernetes architecture helps retailers deploy new stores quickly and consistently.

Especially with many single-tenant application deployments across edge locations (such as customer sites), upgrades and other operational and lifecycle tasks are automated and consistently executed across the board. This reduces support costs and effort, and allows developers to spend more time on delivering new features and software.

Solve Your Kubernetes-at-the-Edge Challenges

Platform9 Managed Kubernetes is suitable for any kind of application at the edge across telco, retail, manufacturing, enterprise, ISV, and SaaS use cases. Its ability to manage deployments on any infrastructure based on centralized policies is a huge time saver. It lowers time-to-fix when outages occur, lowers support costs, and improves customer satisfaction.

Make sure to look at Platform9 Managed Kubernetes and its distributed architecture to solve your Kubernetes-at-the-edge challenges. Try it for free or download the updated buyer's guide.