White paper

PLATFORM9

# Guide to optimizing Kubernetes for 5G environments

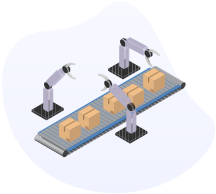
# Contents

# 5G is driving new use cases across industries

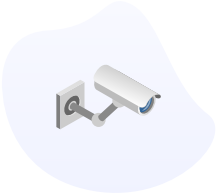
The emergence of 5G technology is enabling a plethora of use cases for enterprises, mobile telco operators and their hardware and software suppliers.

Here are a few examples of industries that are leveraging 5G ultra-wideband speeds (up to 1 Gbs) to deliver innovative solutions:

**Connected cars:** Future cars will come equipped with multi-sensor IoT devices that communicate with other cars, sense road & traffic conditions, and provide near- instantaneous feedback plus provide safety to the driver. 5G speeds will be needed for this. As a result, 5G usage in cars will increase from 74% by 2023.



**Smart factories:** With private 5G, factory automation can cut the cables, eliminate spotty Wi-Fi connections, reduce or remove the cost of networking gear, and drive automated guided robots remotely and wirelessly.

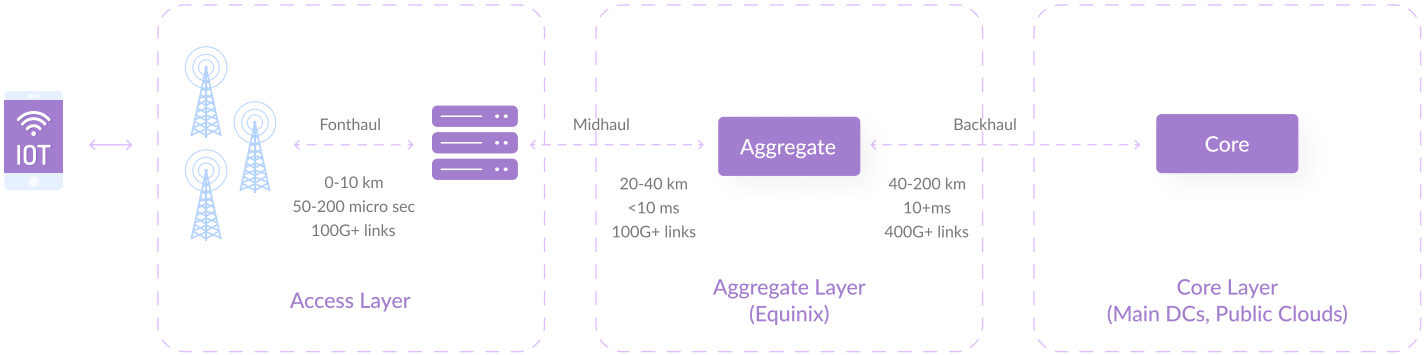


**Video surveillance:** With 5G-enabled cameras, provide near-instantaneous AI-based detection and response to events for vastly improved premises security.

Other areas we see 5G solutions include Augmented and Virtual Reality, Smart Stadiums, eHealth, supply chain, warehouse management, and the list goes on. Fundamentally, 5G is enabling faster digital experiences with higher bandwidth and taking over some of the work that was done by traditional networking and Wi-Fi gear

# High-level 5G architecture view

From a compute and network infrastructure viewpoint, the large mobile operator 5G use case is the most complex and highly distributed. The other use cases are less complex, but the deployment and technical challenges are similar. Let's look at a 10,000 ft high-level infrastructure architecture for a 5G mobile operator deployment and discuss the technical challenges associated with implementing this in practice




A typical 5G architecture has four components:

1. Devices/sensors (Edge)

   These are the mobile devices, sensors, automobiles, factory robots etc that are the endpoints that send and receive data wirelessly over 5G to the nearest cell towers.

2. Access Layer (Edge)

   This is where the servers for handling the front-haul network traffic are deployed, These could be right below the radio tower in a kiosk or within a 10 Km distance from the towers connected over high-speed links (100 G+). Sometimes referred to as "dark fiber", these networks are not connected to the Internet.

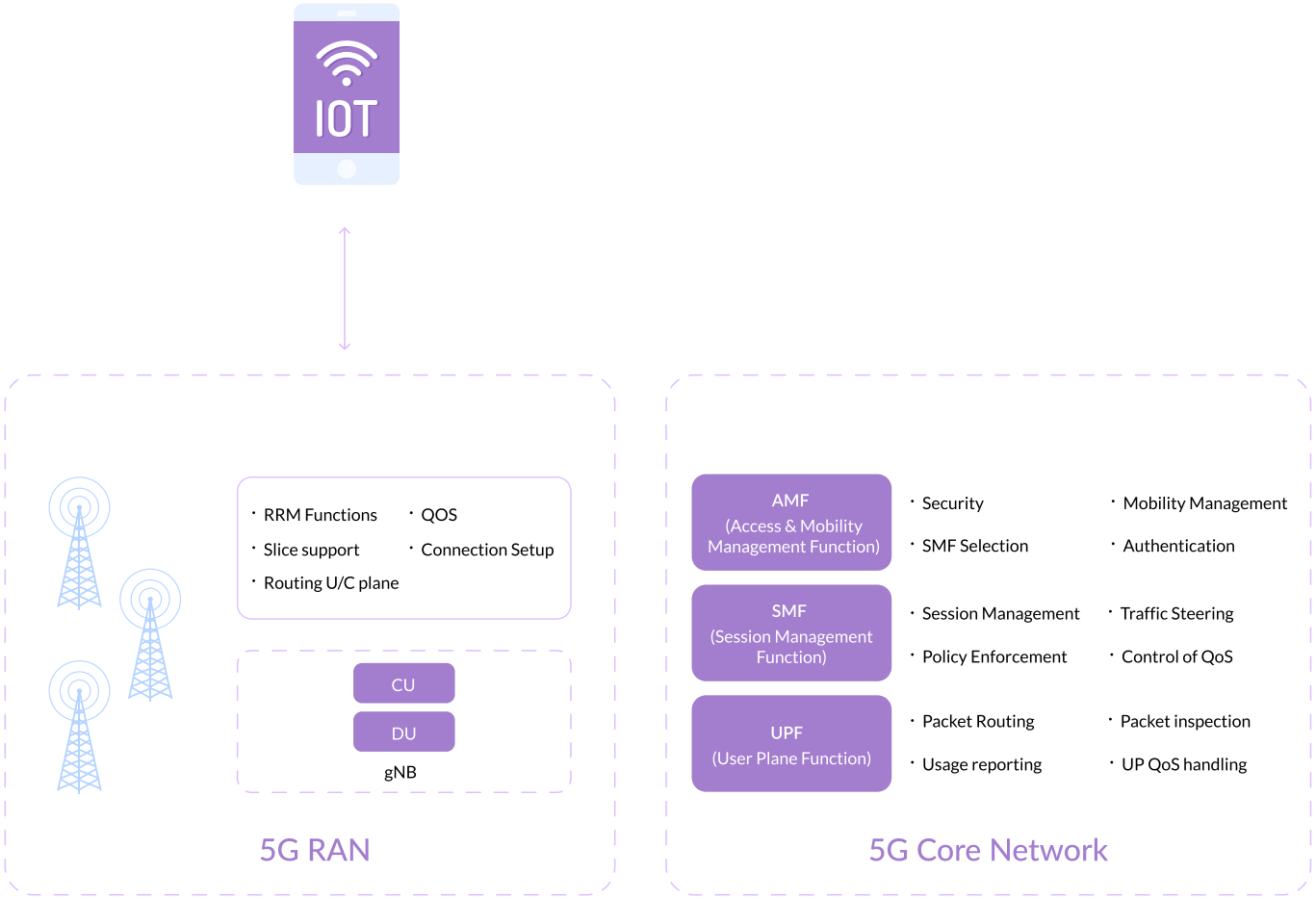
3. Aggregate Layer (Edge)

   Aggregate edge is where the access edge servers are connected & aggregated for further processing and routing. These are typically present in co-location facilities (such as equinix) or telco central offices which is where the public internet connectivity is available

4. Core Layer (Data centers, Public Clouds)

   The aggregate layer then connects through the back-haul networks (typically the internet or 4 dedicated lines) back to the core data center or public clouds. This is where workloads such as Packet Core, dev/test, operational, billing, backup, and others are executed. Workload type, latency, and performance requirements determine the scale of deployment, location of these layers and the distances between them, and compute and storage capacity needed.

# 5G - RAN and Core components




The 5G software stack consists of two main components: Radio Access Network (RAN) and Core

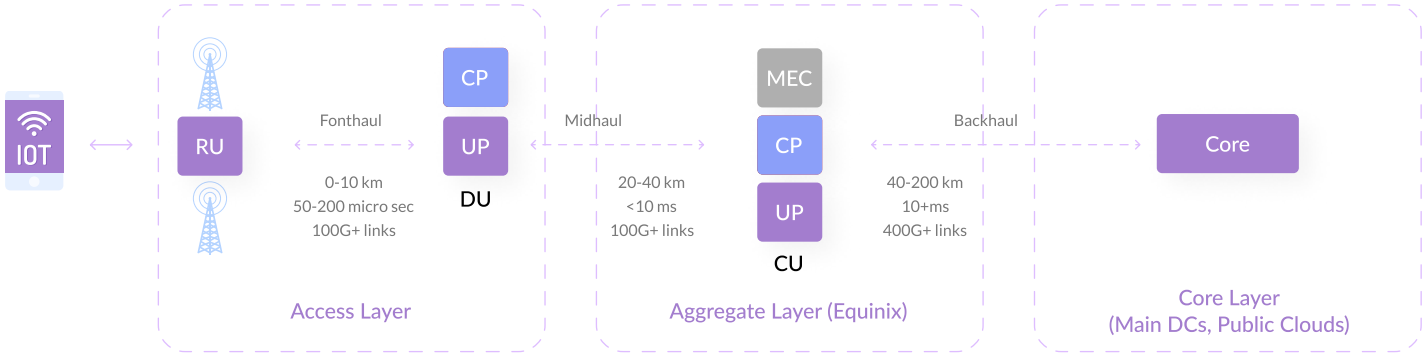
## 1. Core

Core handles numerous functions around security, authentication, mobility management, session management, and more complicated routing.

Core functions resolve many subscriber-related questions like "Is this subscriber the right subscriber? Can this subscriber use this carrier? Has the subscriber paid her bill so she's okay to go on and make a request?" "Is it even in the quota of 10GB plan that she has purchased?" so on and so forth.

## 2. Radio Access Network (RAN)

RAN's main functionality is to make sure the end device wireless signals are converted to packets and routed appropriately. The routing involves complex processing and relies on a number of factors includes quality of service (QoS), carrier priority, and other functions related to radio resource management.




- RU: Radio Unit
- DU: Distributed Unit (RT or near RT)
- CU: Central Unit (no n RT)
- MEC: Mobi

RAN functionality is typically distributed into various "Units" based on where they are located as shown in the picture above

### Radio Units (RU)

These are deployed close to the towers.

### Distributed Units (DU)

These are either close to the towers or close (within 10 Km) of the towers. Bandwidth requirements, latency, and signal strength is what determines how many of these DUs are deployed at what distance.

### Central Units (CU)

Central Units are farther away and are part of the aggregate layer. Typical large-scale 5G rollouts will have the number of DUs that are at least one order of magnitude more than CUs. The specific scale again depends on the user density, amount of traffic and bandwidth needed, and many other factors.

Furthermore, these DUs and CUs have control plane (CP) user plane (UP) components within their architecture. The control plane deals with the radio signaling and control aspects, while the user plane performs the actual data transfer. In order to maintain strict latency (50-200 Microseconds), these RAN components is typically deployed very close to where the mobile users are and the connectivity is provided via dedicated high-speed fiber link that can handle the bandwidth and latency needs.

# Five technical challenges to optimize Kubernetes for 5G deployments

As described in the previous chapter, Operators of 5G mobile broadband networks deal with large-scale, complex, dynamic, and highly distributed infrastructure requirements. They have to deploy and operate thousands of radio towers and networks while managing software applications across the access layer, aggregate layer, and core data centers. Furthermore, they have to satisfy stringent specifications for latency and network performance of their applications and infrastructure. And finally, operators need the flexibility to dynamically relocate services to optimize network performance, improve latency, and reduce costs of operations.

As a result, 5G architectures need to be services-based with hundreds and thousands of network services in the form of VNFs (Virtual Network Functions) or CNFs (Container Network Functions) that are deployed in geographically distributed remote environments. Kubernetes addresses a part of this challenge being able to manage CNFs. However, it has several limitations when it comes to managing 5G services across distributed locations with stringent latency and performance requirements.

Let us look at the top 5 challenges and what needs to be done to optimize Kubernetes for 5G deployments.

01 Inefficient and expensive siloed Management of VNFs, CNFs, and 5G sites

02 Bare Metal Orchestration is Manual and Error-prone

03 High-Performance Networking Options are difficult to configure and operate

04 Standard Resource Scheduling Is Not Suitable for Latency Sensitive CNFs

05 Lack of consistent central management of 5G sites

## Challenge 1: Inefficient and expensive siloed Management of VNFs, CNFs, and 5G sites

It took a better part of a couple of decades for telcos to transition over from physical big iron network deployments to Virtual Networking Functions (VNFs). This improved operational flexibility and dramatically accelerated the speed of deployments and management for the current 3G/4G environments.

But 5G rolls outs are an order-of-magnitude larger scale due to the need for more towers and need much higher levels of flexibility to deliver dynamic networking provisioning and management. For example, providing on-demand bandwidth and variable pricing requires the agility and the speed of Container Network Functions (CNFs).



By 2024, 5G is expected to handle 25 percent of all mobile traffic which will, in turn, drive faster adoption and deployment of CNF's. However, a vast majority of current networks will continue to rely on VNFs.The fact is VNFs and CNFs will have to coexist. This means telco operators need to maintain two siloed management stacks to run this 5G and the older networks adding to the 7 operational burden and cost. Multiply this by the number of sites that need to be managed, you end up with control plane proliferation and inefficiency of siloed management.

One elegant solution is to run both VNFs and CNFs using Kubernetes as the infrastructure control fabric. This then can function as the VIM layer in the MANO stack. Using KubeVirt, an open-source project, that enables VMs to be managed by Kubernetes alongside containers, operators can standardize on the Kubernetes VIM layer eliminating the operational silos. This also eliminates the need to port all of the applications to containers or managing two entirely separate stacks. You can now get the best of both worlds

## Challenge 2: Bare Metal orchestration is manual and error-prone

A large-scale 5G network roll-out involves thousands of access layer sites, hundreds of aggregate sites, and possibly dozens of core data centers. All of these sites have bare-metal servers that need to be provisioned, configured, and managed throughout their lifecycle.End-to-end bare-metal orchestration requires a number of steps, most of which are manual:

- Provisioning bare-metal servers that just expose an IPMI interface over a network
- Deploying OS images on them, running applications on them
- Updating those applications based on network demands
- Upgrading software on the servers to keep them up to date with security patches and bug fixes
- Guaranteeing the availability of those servers in case there is an outage
- Re-provisioning servers when there are performance glitches or other issues



The sheer number of manual steps involved, the complexity of prerequisite knowledge required, and the risks associated with server downtime, and the large number of 5G sites, make it difficult to manage and operate bare metal servers efficiently. Flexibility and agility are impacted when bare metal servers need to be manually provisioned, upgraded, and scaled. 5G telco operators running large environments with combinations of bare metal, VNFs, and CNFs need a simpler, selfservice, automated, remote operating model.

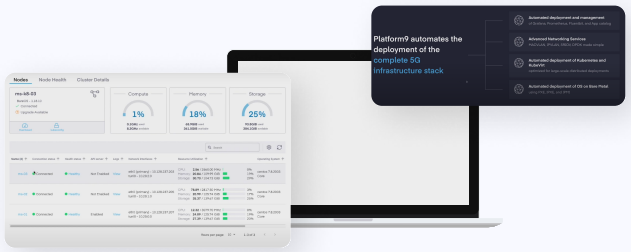
Platform9 brings cloud agility to your bare metal infrastructure providing a centralized pane of management for all of your distributed 5G locations. Using a unique SaaS delivery model, Platform9 automates and offloads all of your manual bare metal life-cycle management tasks. This enables 5G operators and developers to unleash the full performance of the physical hosts — no matter where they are located in a 5G network — as an elastic and flexible bare metal cloud, where they can rapidly deploy and redeploy CNFs or VNFs at moment's notice.

## Challenge 3: High-Performance Networking Options are difficult to configure and operate

The number of end points (mobile devices, IoT sensors, nodes etc) that 5G will inter-connect will exceed hundreds of billions in the next few years: there are simply not enough IP addresses to go around with the current IPv4 standard. IPv6 fundamentally solves this problem and is a must-have for 5G deployments. Infrastructure stacks driven by Kubernetes must handle IPv6 from the ground-up and must support API-driven automated IP address management (IPAM) out of the box.

## Platform9 Advanced Networking solutions for Kubernetes in 5G environments

Watch video →




Another major requirement in 5G deployments is high-performance networking. Performancedriven VNFs and CNFs require near line-rate for network packets. Hypervisors and Docker bridge/networks introduce overhead and performance bottlenecks. Depending on the use case, one or more of these options would be needed: SR-IOV, DPDK, PCI-passthrough, MACvLAN, IPvLAN etc.

However, configuring a host and its physical network and “plumbing” all the necessary components and configuring them within a Kubernetes cluster for a particular use case and particular network requirement can be very time consuming and complicated. The following diagram illustrates the complexity of this network configuration: configuring two CNI plug-ins using Multus, deploying a SR-IOV plug-ion, creating SR-IOV virtual functions, configuring the physical NICs and connecting them to the appropriate eth interfaces on the pod etc.

## How to configure advanced networking options with Kubernetes

Learn more →

The solution is to provide dynamic and automated ways to remotely configure these type of advanced networking settings based on use case and business requirements

## Challenge 4: Standard Resource Scheduling Is Not Suitable for Latency Sensitive CNFs

Latency-critical CNFs need guaranteed access to CPU, memory, and network resources. Pod scheduling algorithms in Kubernetes are designed to enable efficient CPU resource utilization and multi-tasking. However, the negative consequence of this is non-deterministic performance, making it unsuitable for latency-sensitive CNFs. A solution to this problem is to "isolate" or "pin" a CPU core or a set of CPU cores such that the scheduler can provide pods exclusive access to those CPU resources, resulting in more deterministic behavior and ability to meet latency requirements.

CPU-pinning, NUMA-aware scheduling, HugePages, Topology Manager, CPU manager and many other open source solutions are starting to emerge. However, many of these are still early in their maturity and are not entirely ready for large scale production use.

**CPU manager and topology manager which are designed to provide advanced resource bindings in Kubernetes**

Learn more →

Platform9 provides advanced resource bindings required for 5G implementations as part of their SaaS delivery model. Telco providers and operators can offload the overhead of deployment, support, upgrades, monitoring, and general production-readiness to Platform9. The result will be a cloud-native, self-service, declarative model of operations that will be consistent across CNFs, VNFs, and the variety of performance and latency-sensitive requirements.

## Challenge 5: Lack of consistent central management of 5G sites

It's quite a challenge to deploy, manage, and upgrade hundreds or thousands of distributed 5G sites that need to be managed with low or no touch, usually with no staff and little access. Given the large distributed scale, traditional data center management processes won't apply. The edge deployments should support heterogeneity of location, remote management, and autonomy at scale; enable developers; integrate well with public cloud and/or core data centers. While Kubernetes is great for orchestrating microservices in a cluster, managing thousands of such clusters requires another layer of management and DevOps style API-driven automation.

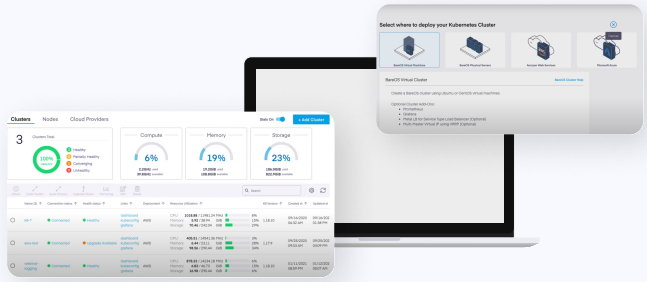
This management plane is where DevOps engineers manage the entire operation. There, they store container images and inventory caches of remote locations. Synchronization ensures eventual consistency to regional and edge locations automatically, regardless of the number of locations.

Fleet management is an approach to grouping sites so that similar configurations can be managed centrally through a single policy known as a profile. This relieves operators 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 site Each 5G site, such as radio towers, access layer, or core data centers runs its own worker nodes and containers.

Additionally, troubleshooting issues and keeping all the services up to date is an ongoing operational nightmare, especially when there are hundreds of these services deployed at each site.

## Platform9 simplifies 5G services on Kubernetes

Watch video →



Platform9 provide a SaaS based centralized management that provides the following capabilities:

- Single sign-on for distributed infrastructure locations 11
- Cluster profiles to ensure consistency of deployment across large number of clusters and customers
- Centralized management of tooling, APIs, and app catalog to simplify application management at scale
- Cluster monitoring and fully-automated Day-2 operations such as upgrades, security patching, and troubleshooting.

**Platform9 gets 5G sites operational in days instead of weeks or months.**

# Approaches to solving the K8s operational challenges

Telco operators and providers need to evaluate various approaches to solve these issues. The most common alternatives include: DIY, commercial Kubernetes distributions, and Managed Kubernetes as a service.

## 1. DIY

- While appealing on the surface, the DIY approach usually means hiring and maintaining a fairly large team with the requisite skills and expertise. But Kubernetes expertise is hard to come by and retain. Higher Kubernetes engineering personnel costs also often consume budget that could have been used to create, perfect and deploy new revenue generating services in the retail environment
- The DIY approach can also involve significant complexity and implementation time delay as Kubernetes is inherently a complex technology to deploy and initial schedules often do not take this into account.
- Furthermore, developing a centralized management plane, advanced networking integration, and 5G edge monitoring requires additional time and resources which further slows down roll-out schedules and increases costs.

## 2. Commercial Kubernetes distributions or public cloud

- Outsourcing Kubernetes to a commercial distribution or public cloud implementation often involves getting locked-in to proprietary technology stacks that add unneeded costs and services over time.
- There are often minimum package costs and mandatory fees plus proprietary technology used that is not easy to migrate away from. Integration fees for combining services purchased can also be expensive.
- These approaches can often slow to provide continuous open source innovation and so do not take advantage of the latest technology innovations available to the Kubernetes community.

## 3. Managed Kubernetes as service

A managed Kubernetes service, especially one that uses a SaaS deployment model, and one that is infrastructure agnostic provides the best of all the worlds. An example of such a service is Platform9 software as a service (SaaS) Kubernetes cluster management that delivers an experience like native public cloud services but on a wider range of on-premises, cloud, and edge infrastructures. This experience is made possible due to the use of a SaaS management model, which entails automating the entire lifecycle of managing Kubernetes deployments, offloading all the operational complexity that ISVs do not have to deal with. The benefits of this model include:

**Kubernetes operations SLA**

By automating health monitoring, runbook driven resolution of common problems, and streamlining upgrades; it is possible to provide an operational SLA for distributed Kubernetes clusters. Hitherto, this operations icon
only available via hyperscale public clouds, but ISVs who desire the simplicity and peace of mind of managed service can now get a similar SLA on any infrastructure of their choice.


**Rapid, repeatable multi-cluster, edge deployments at diverse 5G sites**

In contrast to public cloud Kubernetes services, deploying Kubernetes on-premises and edge environments, and bare metal is complex and slow. What is needed is the automation of the initial deployment and configuration and ongoing management of multi-cluster Kubernetes environments, reducing production implementation to hours (from weeks or months) and reducing cost of operations.

## Conclusion

Getting the most out of Kubernetes for 5G deployment is very complex and challenging. Despite all of the built-in automation features that Kubernetes offers for managing CNFs intelligently, achieving optimal performance, cost and reliability at 5G distributed scale requires careful planning and tuning of Kubernetes environments.

The operational complexity increases exponentially with large-scale distributed sites that 5G rollsout have to deal with. Challenges increase when you add requirements for advanced networking, stringent latency and performance needs, central/remote management, and bare metal orchestration.

Kubernetes distributions and public clouds partially solve some of these challenges leaving a lot of heavy lifting still needed to be done by the 5G implementers.

A SaaS based approach with everything you need for 5G available out of the box simplifies the journey, greatly accelerates the time to get 5G sites up and running from months to days, and reduces operational costs by 90% as a result of hyper-automation at all layers of the 5G infrastructure stack.

PLATFORM9

# About Platform9

Platform9 empowers enterprises with a faster, better, and more cost-effective way to go cloud native. Its fully automated container management and orchestration solution delivers cost control, resource reduction, and speed of application deployment. Its unique always-on assurance™ technology ensures 24/7 non-stop operations through remote monitoring, automated upgrades, and proactive problem resolution, Innovative enterprises like Juniper, Kingfisher Plc, Mavenir, Redfin, and Cloudera achieve 4x faster time-to-market, up to 90% reduction in operational costs, and 99.99% uptime. Platform9 is an inclusive, globally distributed company backed by leading investors.

Follow us on

Headquarter: 800 W El Camino Real, #180, Mountain View, California 94040, US

Phone: 650-898-7369 | Phone: https://platform9.com/contact/