Hybrid Cloud to Power Private 5G Networks

Industrial Worker Safety Use Case

A Technical Paper prepared for SCTE by

Mohamed Daoud
Director of Engineering, Emerging Technology
Charter Communications
6399 S. Fiddlers Green Circle, Greenwood Village, CO 80111
720-699-5077
Mohamed.Daoud@charter.com

Ammar Latif
Principal Solutions Architect- Telco
Amazon Web Services (AWS)
2121 7th Avenue, Seattle, WA 98121
732-983-2708
latammar@amazon.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

1. Introduction

Cloud adoption among enterprises continues its strong momentum and many enterprises depend on the scale of the public cloud for their digital transformation journey. These enterprises have adopted a development framework and workloads that utilize public cloud scale and an expansive set of services. The developers utilize as well as innovate cloud services such as databases, analytics, IoT, and artificial intelligence/machine learning (AI/ML) that are available to them with the rich set of cloud APIs. This affords the enterprises the ability to innovate faster without the undifferentiated heavy lifting on managing the underlying infrastructure.

Cloud edge computing allows workloads to benefit from such cloud innovation and provides a consistent experience to enterprise workloads that require on-prem deployment. These applications usually have low latency requirements for interaction between its components as well as handling massive amounts of data processing and storage.

For telco use cases, private 5G (P5G) is an interesting use case for cloud edge computing due to its distributed architecture with a disaggregated set of telco network functions and applications that need to run on-premises due to latency constraints (e.g., gNB and UPF), local data storage (e.g., UDM), or local data processing needs such as applications that require AI/ML capabilities.

In this paper, we provide an overview of P5G and relevant hybrid cloud architectures. We review Amazon Web Services (AWS) edge offerings relevant to the P5G space. We then dive into Charter labs undertaking of building an industrial worker safety application that utilizes a hybrid cloud deployment of P5G. The paper describes the architectural choices for such hybrid cloud deployment as well as the use of cloud AI/ML tools as part of the application development and delivery.

2. Private 5G overview

A P5G network is a cellular-based network that is designed and operated for a specific organization's private use, rather than being accessible to the public. These networks provide dedicated connectivity and services tailored to the organization's requirements. 3GPP defines private 5G networks as 5G non-public networks (NPN). These private networks are particularly beneficial for industries that require high reliability, low latency, and secure connectivity, such as manufacturing, logistics, healthcare, utilities, and smart cities. P5G can support various applications, including industrial automation, remote monitoring, augmented reality, robotics, and mission-critical communications. Here are some key characteristics and benefits of P5G technology:

1. Dedicated Network: Providing organizations with dedicated cellular network infrastructure. This results in greater control and customization over network resources and enhanced performance while benefiting from 5G technology.
2. Low Latency: Offering low-latency connectivity with dedicated deployments, which is critical for applications that require low latency real-time communication, such as industrial automation, autonomous vehicles, and remote surgery.
3. Enhanced Security: Offering increased security measures compared to public 5G networks. This allows organizations to implement their own security protocols, encryption methods and access controls, providing a higher level of data protection.
4. IoT Enablement: The ability to support a large number of Internet of Things (IoT) devices, which allows for the deployment of advanced applications and services that rely on machine-to-machine communications at scale.

5. Industry-Specific Use Cases: P5G technology can be particularly beneficial for industries such as manufacturing, logistics, healthcare, utilities and transportation, where reliable, low-latency and secure connectivity is crucial.

To address the need for P5G networks, 3GPP has introduced specific features and enhancements to support private networks. The key elements of a 3GPP private network include:

1. Flexible Spectrum Allocation: In P5G architecture, private network operators can either use licensed spectrum dedicated to them or utilize shared/unlicensed spectrum for their network deployment.

2. Network Slicing: Enables sharing a single physical network infrastructure into multiple virtual networks (slices), allowing each private network to have its own isolated and customized resources, performance characteristics and services.

3. Security and Isolation: Private networks have enhanced security measures to protect sensitive data and ensure isolation from public networks. This is done through utilizing encryption, authentication and access control mechanisms.

4. Quality of Service (QoS): As P5G networks can use partially or completely separate infrastructure, this architecture allows organizations to define and prioritize certain types of traffic or applications, ensuring the desired quality of service for their specific use cases.

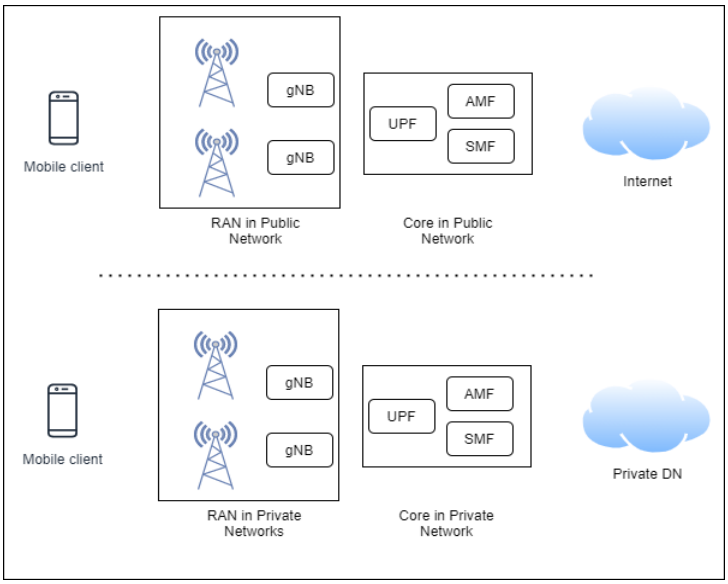

Figure 1: 3GPP NPN Deployment models

3GPP defines two main categories for deploying P5G network:

a) Public Network Integrated NPN (PNI-NPN) where the P5G network relies on some of the network functions that are provided by Public Land Mobile Networks (PLMN). This option is typically provided by PLMN with some of the equipment residing within the targeted enterprise facility.

b) Standalone NPN (SNPS) which by its name is a standalone deployment that does not rely on network functions provided by PMLN. In this option, organizations typically work with network equipment providers, system integrators, and licensed spectrum holders. The solution requires deployment of its own base stations, antennas and network management systems, ensuring dedicated coverage within the targeted premises or desired coverage areas.

SNPNs enable the enterprise customer or the service provider to retain full control of the P5G network. On the other hand, PNI-NPNs can help with introducing P5G networks to enterprises with lower barriers to entry due to reduced initial investment and streamlined ongoing operations that are provided by PLMN. It's important to note that while 3GPP standards provide guidance on architectural options for building private networks, the specific implementation and deployment details may vary depending on the operator, organization and local regulatory requirements.

3. AWS Edge Service For P5G Infrastructure

P5G network requirements vary by industry and use case and thus the architecture for edge placements of network functions vary. For example, use cases that have higher data security or data residency requirements normally would build exclusive UPF to ensure the data flow never gets out of the local network within the targeted zone. Another consideration is latency where industries with lower latency requirements or higher reliability requirements prefer to put user plane function (UPF) in addition to access and mobility management function (AMF) and session management function (SMF) in the targeted zone. On the other hand, deploying NFs in AWS Region offers more versatility of services available to the deployment that are fully managed by the cloud provider.

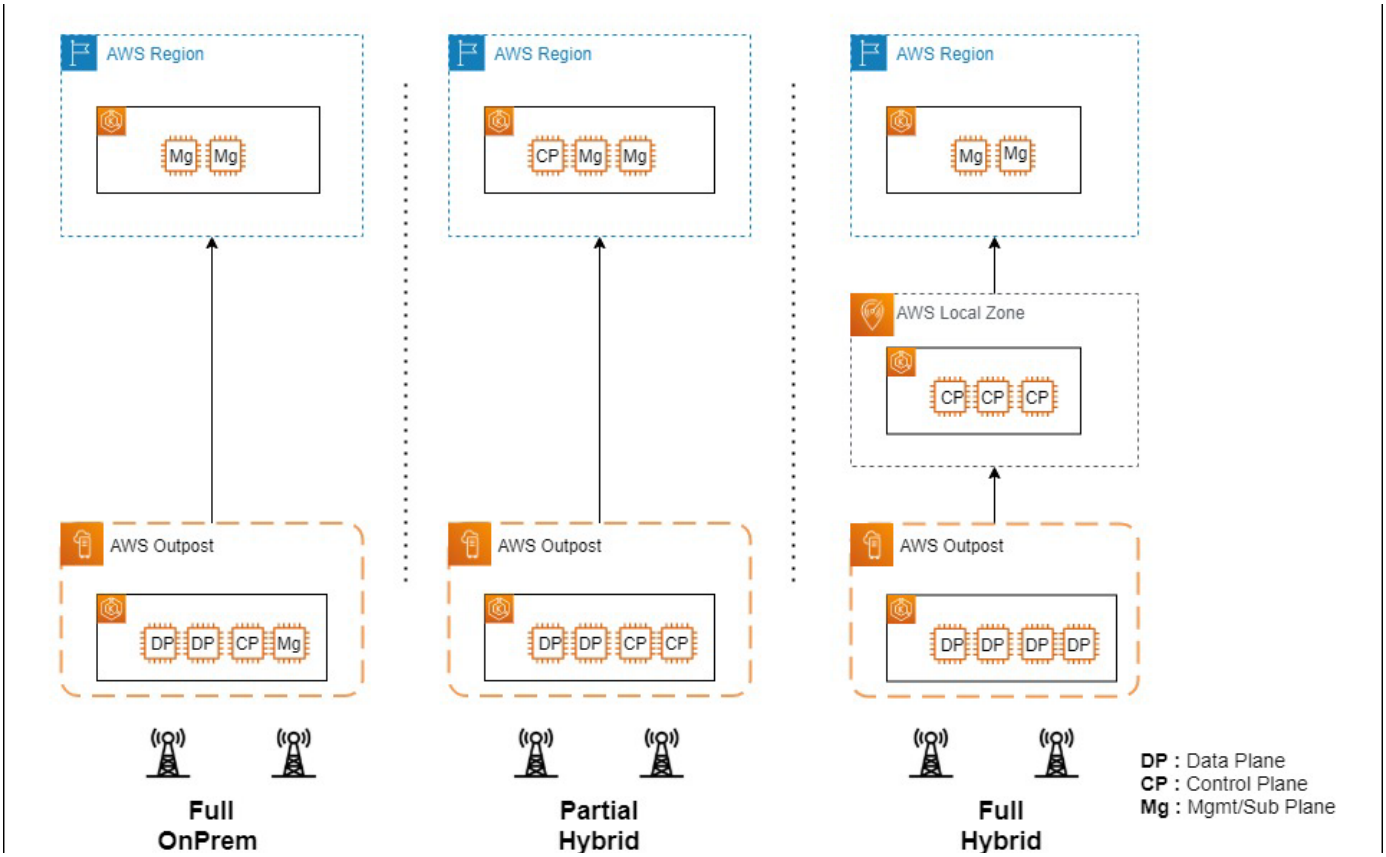

Figure 2: P5G Cloud Deployment Scenarios

AWS offers several edge services that can be utilized by telecom companies to enhance their services and improve the user experience. Here are a few AWS edge services relevant to the telecom industry:

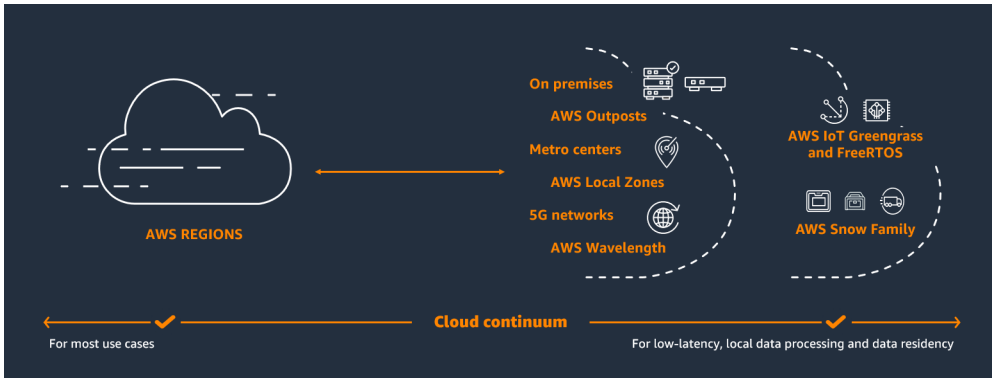

Figure 3: AWS Edge Services

1. AWS Outposts Family: AWS Outposts is a fully managed service by AWS that extends AWS infrastructure, services, APIs and tools to customer on-premises data centers or co-location facilities. It enables organizations to run AWS services locally, thereby providing a consistent hybrid experience with a unified management console. With AWS Outposts, customers can enjoy the same APIs, control plane, and management tools as in the AWS cloud while keeping their data and applications on-premises. This is particularly useful for organizations that have specific

data residency requirements, low-latency application requirements, or legacy applications that cannot easily be migrated to the cloud.

AWS Outposts brings AWS infrastructure and services to on-premises data centers or edge locations. This can be beneficial for telecom companies that require low-latency access to AWS services for specific applications or data processing tasks at the edge.

2. AWS Local Zones: AWS Local Zones are an extension of AWS infrastructure designed to bring AWS services closer to end-users and data centers in specific geographic areas. AWS Local Zones provide low-latency access to select AWS services for applications and workloads that require single-digit millisecond latency. AWS Local Zones are located in metropolitan areas separate from AWS Regions and are connected to the parent AWS Region via a high-bandwidth, low-latency network link. They are typically established to address specific needs, such as providing services to customers in a particular city or region with high local demand or latency-sensitive applications.

   By leveraging AWS Local Zones, telecom companies can deploy applications closer to end-users and process data in a local environment, reducing the latency between the application and its users. This can be beneficial for applications that require real-time responsiveness, such as gaming, media streaming, and financial services.

3. AWS Wavelength: Wavelength brings AWS compute and storage services to the edge of telecom networks, specifically at the edge of 5G networks. It allows developers to build applications that can run directly on Wavelength zones, reducing latency and enabling ultra-low latency use cases like real-time gaming, machine learning inference, and interactive video streaming.

4. AWS Snow Family: The AWS Snow Family helps customers that need to run operations in austere, non-data center environments, and in locations where there's lack of consistent network connectivity. The Snow Family, comprising AWS Snowcone and AWS Snowball, offers a number of physical devices and capacity points, most with built-in computing capabilities. These services enable customers to access the storage and compute power of the AWS Cloud locally and cost effectively in places where connecting to the internet might not be an option.

These services are designed to help telecom companies leverage the power of edge computing, improve performance, reduce latency and enhance the overall user experience for their customers.

4. Charter's Private 5G Lab Setup

In Charter Communications' lab, the team explored diverse P5G network architectures tailored to suit various enterprise use cases. Among the architectures investigated, one notable implementation involved constructing a P5G network for remote location scenarios utilizing AWS Snowcone. This setup incorporated a 5G standalone (SA) core network alongside 5G small cells, offering a robust and versatile solution.

Additionally, the team delved into a hybrid approach that employed AWS Snowcone to host the UPF. In contrast, the SMF, AMF, and Unified Data Management (UDM) components were hosted in the AWS region. Although this hybrid scenario presented intriguing possibilities, the focal point of this paper revolves around Charter's lab architecture, centered explicitly on a P5G network built using the AWS outpost.

By focusing on the AWS Outposts-based architecture, the team aimed to demonstrate this particular setup's unique capabilities and advantages. Through careful design and implementation, Charter's lab succeeded in creating a P5G network that leveraged AWS Outpost infrastructure, showcasing the potential for efficient and reliable connectivity within enterprise environments.

4.1. AWS Outposts

AWS Outposts represents a comprehensive suite of fully-managed solutions designed to seamlessly bring AWS infrastructure and services to on-premises or edge locations, ensuring a unified hybrid experience. These innovative AWS Outposts solutions empower users to extend and operate native AWS services directly on their premises, offering unrivaled flexibility and control. The range of AWS Outposts form factors caters to diverse requirements, spanning from compact 1U and 2U servers to expansive 42U racks and multi-rack deployments.

With the integration of AWS Outposts, operators and enterprises can deploy select AWS services locally while retaining seamless connectivity to an extensive array of services available within their local AWS Region. This empowers them to run applications and workloads on-premises effortlessly, leveraging the familiar ecosystem of AWS services, tools, and APIs. AWS Outposts are optimized to support workloads and devices necessitating low-latency access to on-premises systems, local data processing, data residency compliance, and application migration with local system interdependencies.

Within the Charter team, a full AWS outpost rack boasting 42U capacity has been procured, comprising a balanced mix of CPU (Central Processing Unit) and GPU (Graphics Processing Unit) resources. The entire ordering process for the outpost was easily executed through the user-friendly AWS console, ensuring transparency and enabling real-time tracking at each acquisition stage.

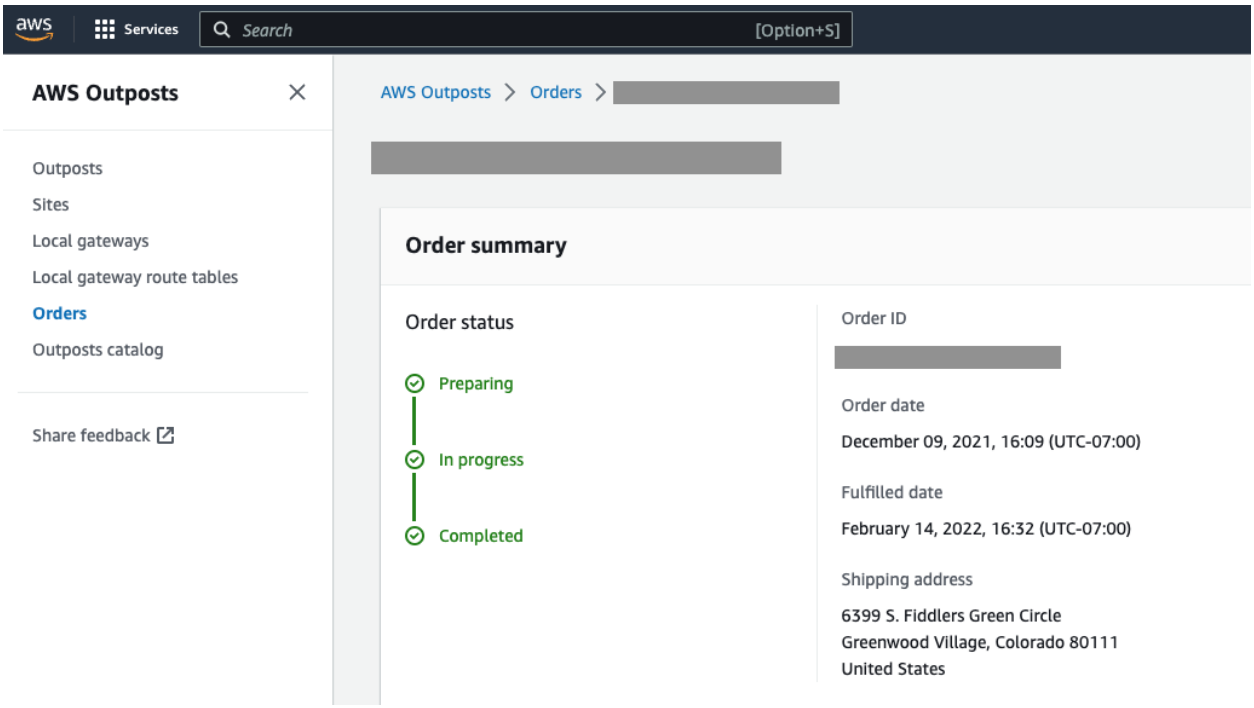

Figure 4: AWS Outpost Ordering and Tracking via the Console

4.2. Charter Lab AWS Outpost Configuration

The AWS Outpost deployed within the Charter lab establishes a secure connection with the AWS US West region through a robust site-to-site VPN (Virtual Private Network). This enables the extension of the Virtual Private Cloud (VPC) from the region directly to the Outpost's on-premises infrastructure, ensuring seamless connectivity between workloads in both environments. While the AWS Outpost offers the alternative of connecting back to the region via AWS Direct Connect, the Charter lab team deliberately chose the VPN option to meet their specific requirements.

Following the successful arrival and configuration of the Outpost, it seamlessly integrated into the AWS ecosystem, accessible and manageable through the same familiar AWS console interface utilized for other resources within the AWS region. This cohesive experience enhances ease of use and ensures a consistent operational environment for the Charter lab team.

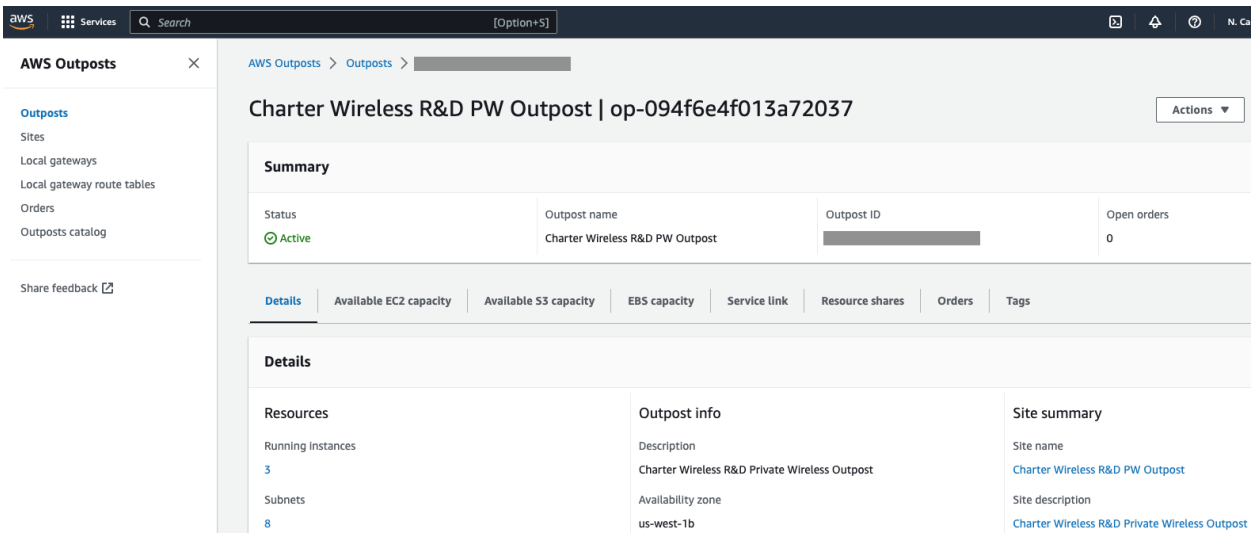

Figure 5: Outpost Management via AWS Console

The outpost in Charter lab has GPUs in the form of four G4dn.2xlarge and four G4dn.4xlarge and CPUs in the form of six M5.4xlarge instances.

Table 1: Charter's AWS Outpost Compute Configuration

Instance Type	Quantity
G4dn.2xlarge	4
G4dn.4xlarge	4
M5.4xlarge	6

The distinction between G4dn.2xlarge and G4dn.4xlarge instances lies in their vCPU count and CPU memory capacity. Both instance types, powered by NVIDIA T4 GPUs and custom Intel Cascade Lake CPUs, offer exceptional performance. Notably, the GPUs are consistent across both types, featuring a single T4 GPU equipped with 16 GB of GPU memory.

Instance	GPUs	vCPU	Memory (GiB)	GPU Memory (GiB)	Instance Storage (GB)	Network Performance (Gbps)***	EBS Bandwidth (Gbps)
g4dn.xlarge	1	4	16	16	1 x 125 NVMe SSD	Up to 25	Up to 3.5
g4dn.2xlarge	1	8	32	16	1 x 225 NVMe SSD	Up to 25	Up to 3.5
g4dn.4xlarge	1	16	64	16	1 x 225 NVMe SSD	Up to 25	4.75
g4dn.8xlarge	1	32	128	16	1 x 900 NVMe SSD	50	9.5
g4dn.16xlarge	1	64	256	16	1 x 900 NVMe SSD	50	9.5
g4dn.12xlarge	4	48	192	64	1 x 900 NVMe SSD	50	9.5
g4dn.metal	8	96	384	128	2 x 900 NVMe SSD	100	19

All instances have the following specs:

- 2.5 GHz Cascade Lake 24C processors

Figure 6: GPU-based Instances on Outpost

The configuration mentioned in Table 1 is what Charter decided to have. However, the Outpost compute instances are configurable via a slotting mechanism. Once slotted, the configuration can't be changed unless all EC 2 instances are terminated.

The original Outpost configuration is two g4dn.12xlarge and one m5.24xlarge, as shown in Figure 7 but the Charter team decided to slot it as in Table 1 to run a 5G SA core and enterprise application simultaneously on the outpost.

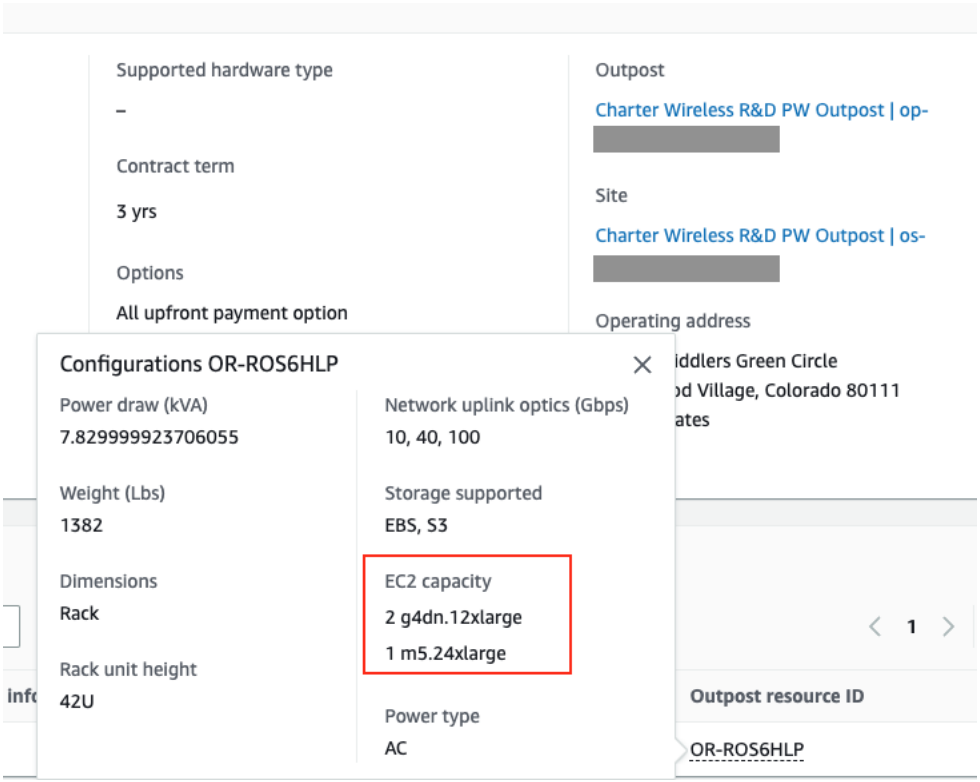

Figure 7: Outpost Original Instance Capacity before Slotting

Figure 8 illustrates various g4dn slotting options, each offering distinct configurations in terms of CPU, GPU, and memory utilization. The Charter team engaged in several productive discussions with the AWS team to carefully determine the optimal slotting requirements for the project. These discussions took into account the specific use cases, computing needs, capacity requirements, as well as potential future workloads anticipated to run on the outpost. Through collaborative deliberation, the team arrived at a well-informed decision on the most suitable slotting configuration.

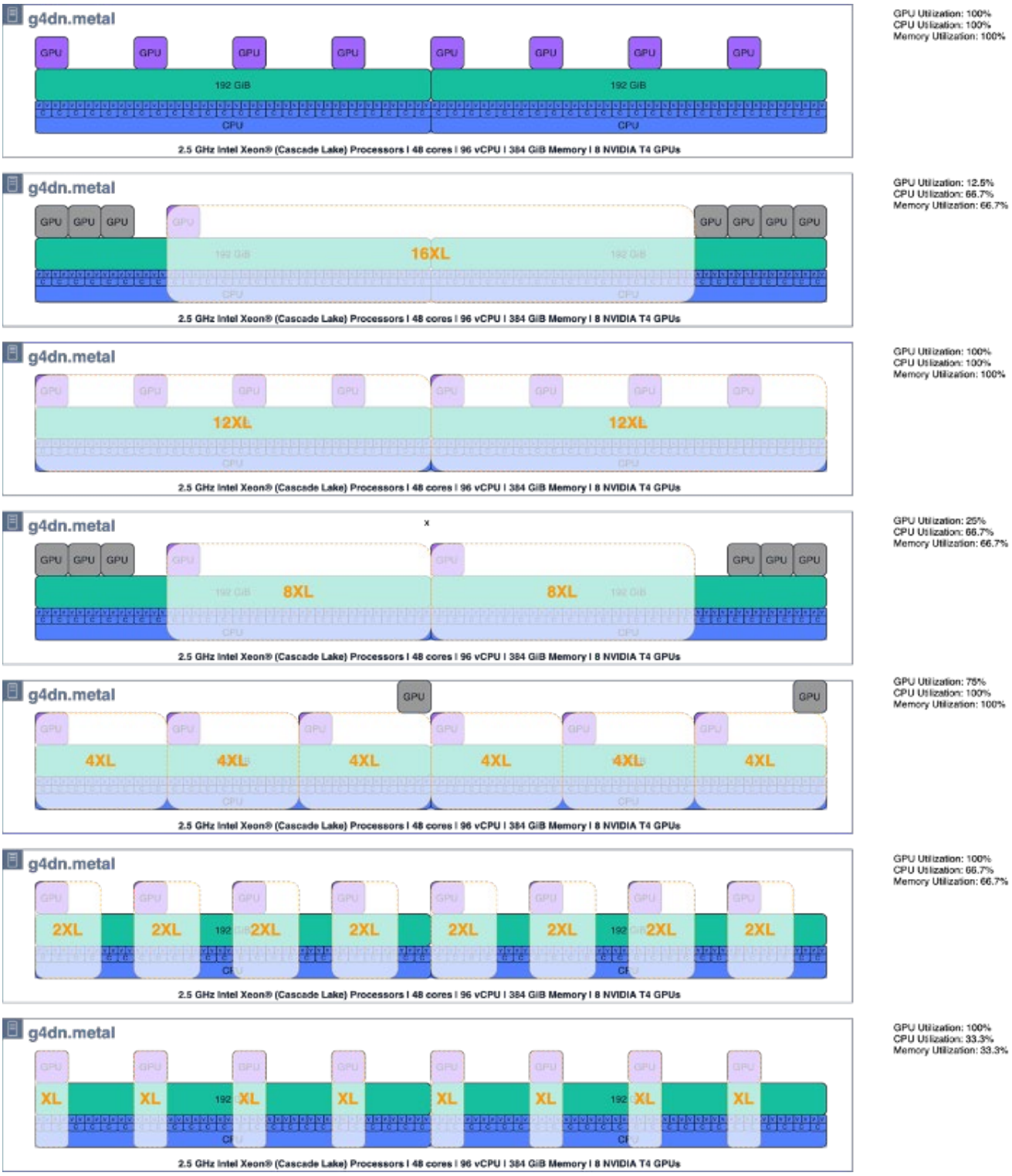

Figure 8: Example of g4dn.metal Slotting Options

Considering future scalability, the Charter team deliberately selected an AWS Outpost configuration allowing for expansion. As depicted in Figures 9 and 10, the Charter AWS Outpost is designed with ample physical space to accommodate additional compute and storage resources as required. This strategic decision ensures that the infrastructure can quickly scale and adapt to evolving demands, providing the flexibility needed for future growth and optimization.

Figure 9: AWS Outpost in Charter Lab

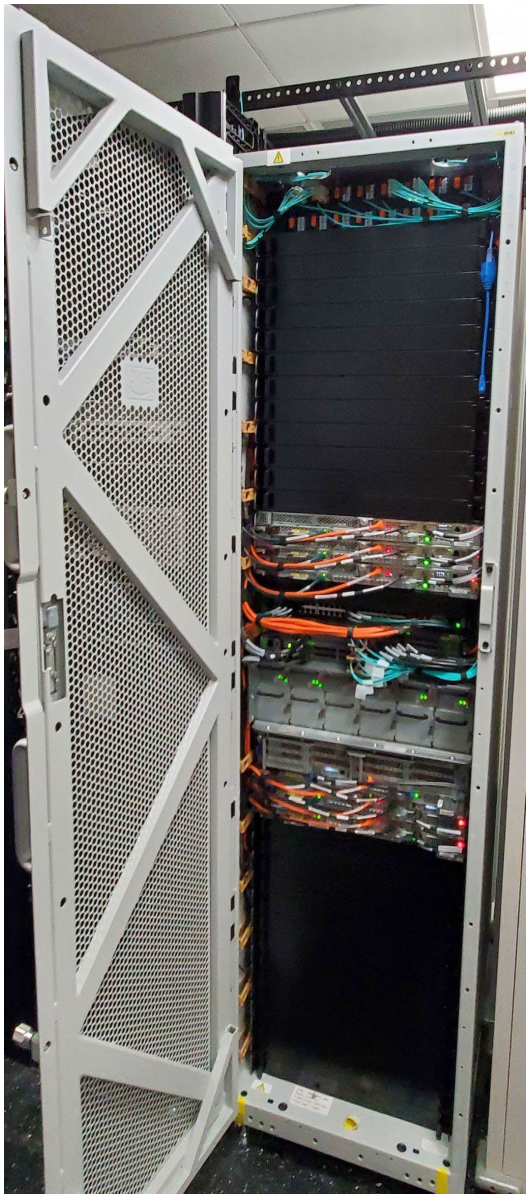

Figure 10: AWS Outpost Inside View Showing Installed RUs

4.3. Charter Private Wireless Lab Setup

In the Charter lab, a robust 5G system has been successfully implemented by leveraging a combination of on-premises servers and cloud infrastructure provided by AWS Outpost. To ensure efficient deployment of specific components, three Linux-based hosts have been utilized: one for the 5G core, another for the Radio Security Gateway (R-SGW), and the third for the OAM NAT (Network Address Translation).

The 5G core and R-SGW have been deployed on-premises to optimize performance and reduce latency. This strategic placement near the 5G small cell enables efficient security processing, resulting in lower latency during data transmission. Conversely, the OAM component responsible for network management and monitoring has been deployed in the AWS cloud region (us-west-1). This cloud-based approach allows for centralized management and monitoring of the 5G network infrastructure, eliminating the need for physical presence at each network site. Leveraging the cloud-based OAM offers advantages such as streamlined management and effective control over the 5G networks.

The Charter team has implemented a site-to-site VPN to establish secure communication and seamless data transfer between the on-premises servers and cloud resources. This VPN ensures the confidentiality and integrity of the communication between the on-premises infrastructure and the cloud-based resources.

For reliable communication between the 5G core and the 5G small cell, a high-speed 1 Gbps Ethernet connection has been employed. This connection utilizes an Ethernet cable connected to a gigabit switch, facilitating efficient and robust communication between these crucial components of the 5G system.

The Charter team chose 5G Citizen Broadband Radio Service (CBRS) small cells for the lab setup. CBRS is the 150 MHz of spectrum in the 3550 MHz to 3700 MHz range (3.5 GHz to 3.7 GHz) that the United States Federal Communications Commission (FCC) has designated for sharing among three tiers of users: incumbent users, Priority Access License (PAL) users, and General Authorized Access (GAA) users.

Charter Communications owns a significant amount of PAL across major U.S. markets.

Collaborating with the AWS team, Charter has deployed a 5G SA core supplied by a major 5G core provider and two small cells from two different vendors capable of transmitting on n48 band. This combination of cutting-edge technologies ensures the successful operation of the 5G system, meeting the requirements of Charter's implementation.

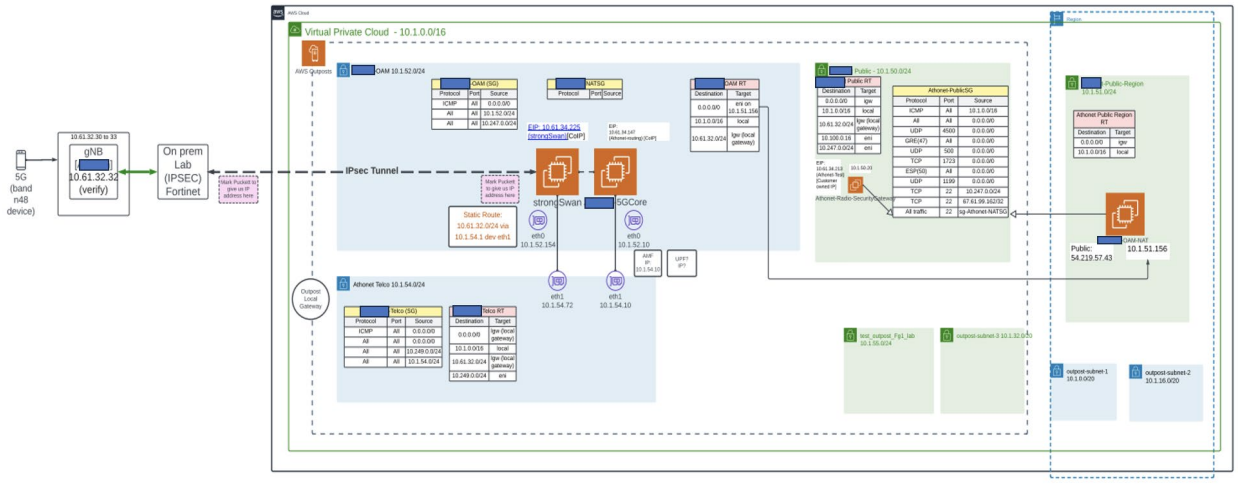

Figure 11: Charter's Lab Private Wireless on AWS Outpost Architecture

Implementing the 5G Core network in the Charter lab demonstrates a well-designed architecture that effectively utilizes AWS Outpost, as depicted in Figure 11. The core network is deployed on-premises

using the "m5.4xlarge" server instance provided by AWS. This powerful server instance ensures optimal performance and resource allocation for the 5G Core software.

In the us-west-1 region, a dedicated "t3.micro" instance operates the Operations, Administration, and Maintenance (OAM) service. This instance is vital in comprehensive management, monitoring and fault detection within the 5G Core network. By leveraging this cloud-based OAM instance, Charter can efficiently manage and maintain the network infrastructure remotely.

The "m5.4xlarge" servers are intelligently configured with virtualized network functions specifically designed for the 5G Core. These servers connect to the local network through a Layer 3 (L3) switch, directly linking with the 5G small cell. The 5G small cell supports Band n48, allowing 5G phones to establish connections and benefit from the high-speed network.

This carefully planned setup enables the deployment of a robust 5G Core network on-premises, providing greater control and minimizing latency throughout the network infrastructure. The powerful "m5.4xlarge" server instance ensures smooth operation of the 5G Core software. In contrast, the OAM instance in the us-west-1 region enables efficient management and maintenance of the network infrastructure from a remote location.

The 5G SA core setup was easy and quick since the 5G core was based on containers and predefined templates. Using the AWS console and with just a few clicks, the 5G SA core was instantiated and configured successfully. This easy and quick 5G Core setup results from AWS and telecom vendors' work to ensure compatibility and a hassle-free setup.

After standing up the 5G core and connecting the 5G small cells, Charter and AWS collaborated closely to conduct comprehensive testing of the lab setup, ensuring that devices could successfully attach and achieve the maximum throughput, validating the reliability and performance of the implemented 5G Core network.

5. Lab Private Wireless Use Case

The Charter team has successfully leveraged the AWS Outpost and a P5G network within their lab to implement an enterprise use case, showcasing an end-to-end network and enterprise application integration on the same hardware. Specifically, they have utilized a portion of the outpost compute as Multi-Access Edge Computing (MEC), enhancing the capabilities of their infrastructure.

The highlighted enterprise use case focuses on worker safety, featuring an innovative application that detects non-compliant workers and takes immediate action to ensure their safety. In unsafe scenarios, the application promptly intervenes by stopping the machine being operated by the worker.

This integration of AWS Outpost, P5G network and the worker safety application demonstrates the Charter team's commitment to harnessing cutting-edge technologies to create a secure and efficient working environment. By utilizing MEC functionality, they have effectively brought compute capabilities closer to the edge, enabling swift and intelligent decision-making at the point of action, thereby prioritizing worker safety and well-being.

5.1. Worker Safety Use Case

The concept of MEC has revolutionized the way traffic and services are handled by moving them from the cloud to the network's edge. This shift brings significant improvements in terms of latency and security. MEC is advantageous for applications such as Data and Video Analytics, Augmented Reality, and various Internet of Things (IoT) use cases. Recognizing the strengths of MEC, the idea emerged to host a low-latency safety application on this platform, specifically designed to monitor and prevent accidents in a factory environment involving a drill press.

Disclaimer: The following example involving Personal Protective Equipment (PPE) and Mobile Edge Computing (MEC) is purely for illustrative purposes to showcase the potential benefits of MEC in a private network edge scenario. It is not intended to represent or comply with real-life safety standards, safety codes, or regulations. For accurate and up-to-date information on safety standards and regulations related to Personal Protective Equipment or any safety-related matters, it is essential to consult and adhere to the guidelines provided by relevant safety administrations and authorities. This example should not be used as a reference for establishing safety protocols or procedures. Always prioritize safety by following official safety guidelines and seeking advice from appropriate experts and regulatory bodies.

The lab worker safety application developed for this purpose focuses on ensuring compliance with Personal Protective Equipment (PPE) requirements, including helmets, safety vests, gloves and goggles. By verifying the presence of the necessary PPE, the application categorizes workers as either "green" or "red" based on compliance. Its primary objective is to prevent accidents caused by the tips of drill press bits. The application incorporates a computer vision model and algorithm responsible for recognizing and localizing detections to achieve this.

The architecture of the overall worker safety application is depicted in Figure 12, highlighting the interconnected components and their functionalities.

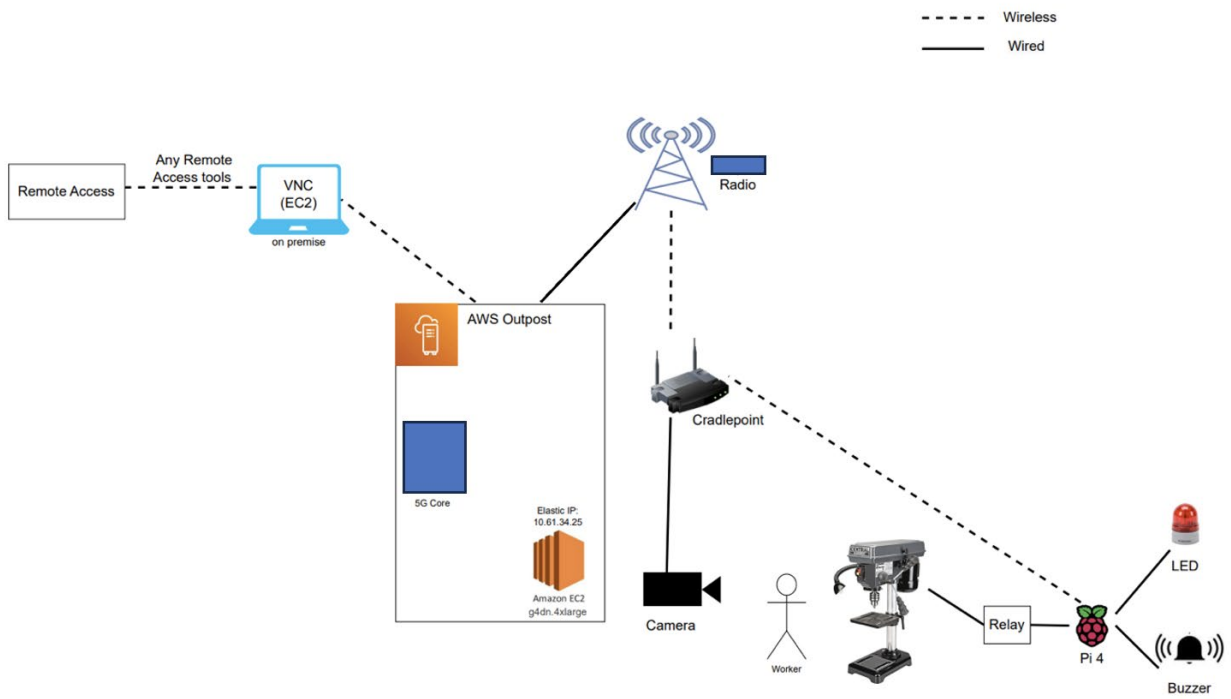

Figure 12: Charter Lab Use Case Architecture Deployed on AWS Outpost

To implement this application, the setup utilizes AWS Outpost as the MEC platform, hosting the 5G SA core, 5G radio components, a Cradlepoint router, a camera, a drill press machine, a Raspberry Pi, a relay switch, LEDs and a buzzer.

This configuration creates a Radio Access Network (RAN) ecosystem, with the 5G core hosted on the AWS Outpost and the radios strategically placed throughout the workplace, establishing connections with routers and integrating all the components. The camera, which operates using Power over Ethernet (PoE), is positioned to focus on the drill press. The camera leverages the Real-Time Streaming Protocol (RTSP) to transmit camera frames over the network.

A GPU processes each frame captured by the camera within the AWS Outpost, and the resulting output responses are sent back over the radio to the corresponding Raspberry Pi. The Raspberry Pi is wirelessly connected to the router and controls the relay switch, which is responsible for turning the drill press on/off and managing the LEDs and buzzer.

Integrating these components in the MEC architecture ensures real-time safety compliance monitoring and swift preventive action, effectively reducing the risk of accidents in the factory environment.

5.2. Use Case Computer Vision Model

Implementing a computer vision model is essential to detect workers, PPE, and machines accurately. The chosen model for this purpose is YOLOv4, a well-known and widely used deep neural network computer vision model. YOLO, which stands for You Only Look Once, has gained popularity as an open-source solution.

The development cycle of the computer vision model for object recognition tasks follows a sequential process comprising several crucial steps: data collection, data annotation, model training, model validation and model deployment.

The data collection process involves creating a custom dataset by capturing images specific to the application scenario and utilizing publicly available datasets online. We employed both approaches to develop our model, resulting in the collection of approximately 10,000 images. Once the data is gathered, the next step is annotation.

These 10,000 images needed to be annotated to be fed into the machine learning model to train it. Annotating is the practice of assigning labels to an image or parts of it. In this case, annotating was especially hard as each image included several objects to annotate, like a vest, helmet, gloves, worker, and more. Annotating the 10,000 images manually would have taken hundreds of hours to do. The Charter team had to find a solution for that.

5.2.1. Annotating Images using AWS Mechanical Turk

The Charter team initiated the annotation process by meticulously labeling a subset of images from the dataset, gaining valuable insights into the requirements, time consumption and obtaining examples of annotated images.

To ensure accurate and efficient annotation of the entire 10,000-image dataset, the team leveraged AWS SageMaker's Ground Truth, a cloud service that simplifies data labeling and provides streamlined annotation jobs. Ground Truth facilitated the utilization of human annotators through various means such as Amazon Mechanical Turk, third-party vendors, or an in-house workforce. Amazon Mechanical Turk functions as a crowdsourcing marketplace, enabling the delegation of tasks to a distributed global workforce. These tasks range from simple data validation and research to more subjective assignments like survey participation and content moderation. AWS Mechanical Turk empowers businesses to tap into a diverse workforce's collective intelligence, skills, and insights, enhancing operational efficiency, data collection, and analysis and expediting machine learning development.

By combining the manual annotation efforts of the Charter team with the assistance of AWS SageMaker's Ground Truth service, the dataset received comprehensive annotation, providing the crucial training data required for the YOLOv4 model. This approach ensures accurate and reliable object detection and classification within the computer vision system for workers, PPE, and machinery.

Remarkably, the annotation of the 10,000 images was completed in under 24 hours, delivering highly satisfactory results.

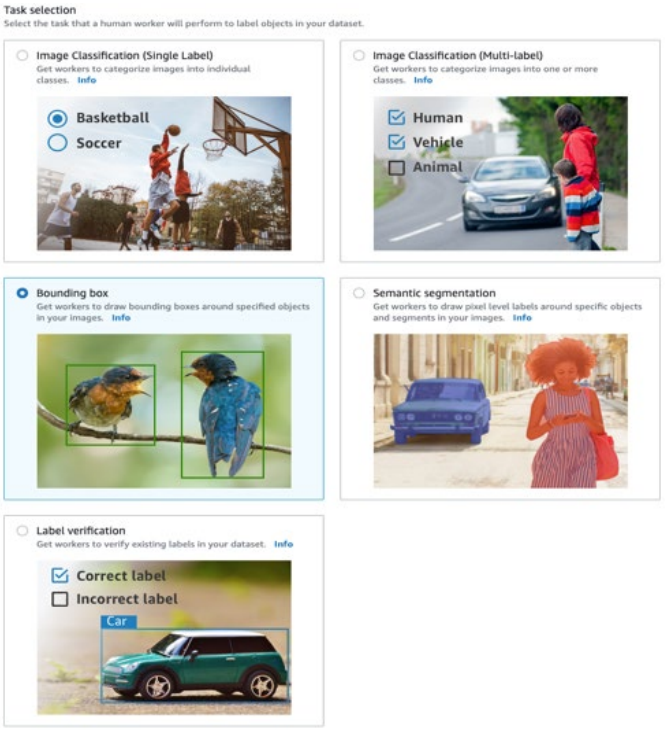

Figure 13: AWS Mechanical Turk Labeling Task

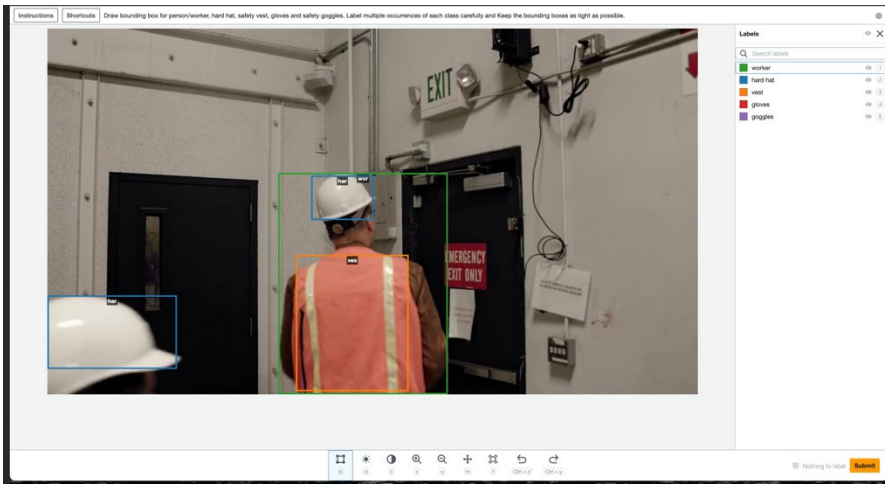

Figure 14; AWS Mechanical Turk - Charter's Worker Safety Task

5.2.2. Training Machine Learning Model on EC2

The computer vision model employed for this project is highly complex, comprising more than 60 million parameters. If the training of this model were to be carried out on a conventional computer, it would take

several months to complete. The model training was executed on an EC2 instance within the AWS regions to expedite the training process.

Specifically, the training took place on a p3.16xlarge instance, utilizing the Deep Learning AMI. The p3.16xlarge instance boasts eight Tesla V100 GPUs, collectively offering a GPU memory capacity of 128 GB. Leveraging the computational power of this setup, the training task was significantly accelerated. Instead of taking months, it was completed within hours.

The training process effectively utilized the CUDA platform, which was pre-installed on the AMI, further optimizing the training efficiency. The model's input picture size was set at 416 x 416, and 85% of the dataset was randomly selected for training. Important training parameters, such as the batch size and number of epochs, were determined as 64 and 25000, respectively. The model training yielded an average loss estimate of 1.5387, indicating a satisfactory training outcome.

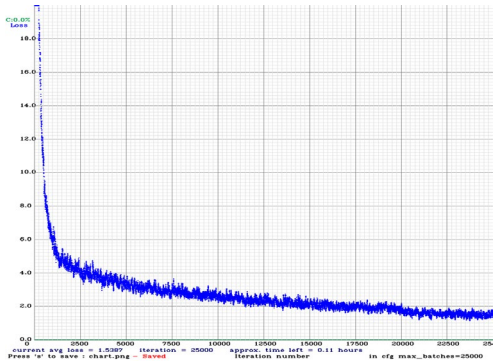

Figure 15: Training Loss Curve

To ensure the reliability of the trained model, it underwent validation using the remaining 15% of images from the dataset. The model's performance was evaluated using the mean average precision (mAP) metric, which comprehensively assesses its behavior.

For the specific trained model with an input size of 416x416, benchmarking revealed an impressive frames-per-second (FPS) rate of 72 FPS. Additionally, the mAP at an Intersection over Union (IoU) threshold of 0.75 was measured at 0.7238 or 72.38%. These results demonstrated the model's ability to detect and effectively classify objects within the images.

During the benchmarking process, it was observed that altering the input size impacted the model's performance. Increasing the input size resulted in improved mAP scores but reduced the frame rate, highlighting the trade-off between accuracy and computational efficiency.

To facilitate the training model's integration and deployment, the binary weight file was converted to the ONNX framework, streamlining its conversion to the TensorRT framework. TensorRT is a model acceleration framework that optimizes weights to a lower resolution format while preserving the model's behavior. As a result, the accelerated model achieved an impressive 96 FPS frame rate.

The inference of the accelerated model was conducted on an AWS Outpost g4dn.4xlarge instance, utilizing the Deep Learning AMI. This particular instance is equipped with a single T4 GPU and 64 GB of GPU memory, providing sufficient computational resources for efficient model inference.

In the Charter lab, the trained model was successfully implemented and executed using the AWS Outpost. The provided pictures in Figure 16 are samples of the model's inference, demonstrating its ability to detect objects within the given images accurately.

Figure 16: Charter Developed Model Inference

5.3. Private 5G Network and Uplink Capacity

It was necessary to explore solutions enabling efficient video uplink to accommodate the demands of multiple cameras in enterprise use cases. Two approaches were considered for addressing this challenge.

The first approach focuses on optimizing the 5G network configuration. One way to achieve this is by modifying the Time Division Duplex (TDD) frame configuration, allowing for the allocation of more uplink slots. This enables an increased capacity for camera uplink transmissions. Additionally, the uplink carrier aggregation feature can be employed to augment the available bandwidth for uplink purposes. Leveraging these network-level adjustments can enhance the capacity for handling multiple camera streams.

The second approach involves optimizing the video feed itself. Various techniques can be employed to achieve this goal. One such technique is utilizing advanced encoding methods to reduce the size of the video data without compromising its quality. Lower frame rates and resolutions can also be employed to minimize the bandwidth requirements of each camera stream. Implementing these optimizations makes it feasible to support multiple camera feeds simultaneously on the 5G private wireless network.

In the subsequent section, we will delve deeper into the specifics of optimizing the video camera feeds, exploring methods and strategies that enable the efficient operation of multiple cameras within the 5G private wireless environment.

5.3.1. Camera Configuration

Typically, cellular networks prioritize downlink traffic and possess a restricted capacity for uplink transmission. This inherent limitation significantly impacts the number of cameras that can be accommodated within the network and their resolution and frame rate configurations. The encoding algorithm employed by the cameras directly influences the instantaneous upload bitrate required for transmitting the video data over the network.

Over time, encoding techniques have evolved, becoming more dynamic in nature. The H.264 encoding standard was employed in the cameras utilized for testing. This particular encoding standard offers distinct advantages over other encoding types, especially when dealing with dynamic field-of-view scenarios. The utilization of H.264 ensures efficient compression of video data while maintaining reasonable image quality, thereby optimizing the utilization of the limited uplink capacity in cellular networks.

5.3.2. Private 5G Uplink Capacity, Camera FPS, and Resolution

In this scenario, a crucial aspect was finding the right balance between uplink bitrate/throughput, measured in megabits per second (Mbps), and the camera feed's desired FPS and resolution. Adjustments to the FPS and resolution were made within the camera's stream settings to achieve the desired outcome.

To assess the uplink throughput, the router was configured to estimate the data transmission rate over a 30-minute period, gathering 100 samples per hour. However, it should be noted that certain tests could not be conducted due to the cameras used for testing lacking support for 60 FPS at higher resolutions.

It's important to recognize that higher resolutions necessitate a greater uplink bitrate. Consequently, a delicate balance needed to be struck between the FPS settings that would ensure accurate operation for the specific use case and the inherent limitations of the uplink capacity.

By carefully considering these factors, an optimal configuration was sought to enable effective camera performance while working within the constraints of the available uplink capacity.

Table 2: Uplink Bit Rate vs. FPS and Resolution

FPS	*15*			*30*			*60*		
Resolution	*Uplink Bitrate (Mbps)*			*Uplink Bitrate (Mbps)*			*Uplink Bitrate (Mbps)*		
	Min	*Avg*	*Max*	*Min*	*Avg*	*Max*	*Min*	*Avg*	*Max*
640x480	0.37	0.59	4.2	0.29	0.97	1.34	1.20	1.62	2.30
1280x720	0.62	1.03	1.63	1.52	1.86	2.54	3.24	5.03	6.93
1920x1080	2.29	2.96	3.85	3.30	4.13	6.64	NA	NA	NA
2992x1680	5.15	6.96	9.32	3.89	8.12	11.13	NA	NA	NA

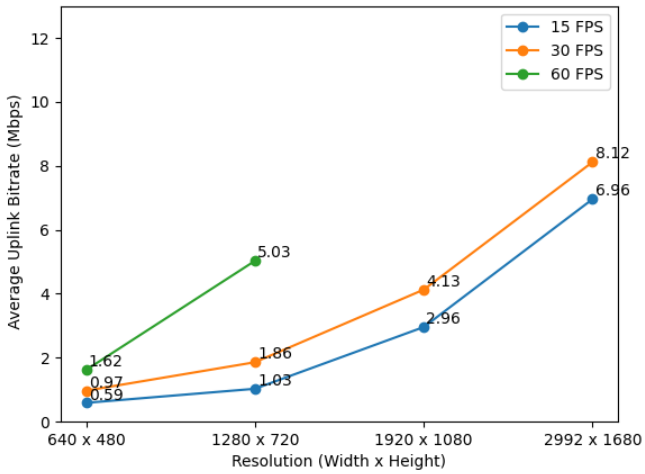

Figure 17: Resolution vs. Average Bitrate

5.3.3. The Video Streaming Choice for the Lab Use Case

To ensure the scalability of the worker safety solution, careful consideration is given to the tradeoff between the resolution and frame rate of the video stream. This decision enables the connection of multiple cameras to a single CBRS 5G router, facilitating a comprehensive monitoring setup.

When analyzing the uplink bitrate usage, it becomes evident that lower-resolution streams consume less bandwidth than higher-resolution streams. Conversely, the uplink bitrate increases with higher frame rates. While lower-resolution images may exhibit graininess and pixel loss, they tend to be less effective in challenging lighting conditions.

Since this application pertains to safety, lower frame rates are a disadvantage as they capture fewer frames per second, resulting in less information. A higher frame rate is preferred to maximize the potential for incident detection and prevention. According to latency data, a frame capture latency of approximately 9ms is achieved at 60 FPS, allowing for more frequent updates and providing more information to the algorithm. This advantage positions higher frame rates as capable of preempting unfortunate incidents more effectively than lower frame rates.

Considering the camera capture latency and uplink bitrate considerations, the application was run at a resolution of 1280x720 and a frame rate of 60 FPS. This configuration strikes a balance between capturing crucial details and minimizing bandwidth requirements.

A test was conducted using up to three cameras to assess the solution's scalability, examining the system's ability to handle multiple camera inputs simultaneously.

Table 3: Uplink Bitrate for Multiple Cameras

Number of Cameras	Uplink Bitrate (Mbps)		
	Min	Avg	Max
2	6.48	7.37	7.76
3	9.04	11.7	14.61

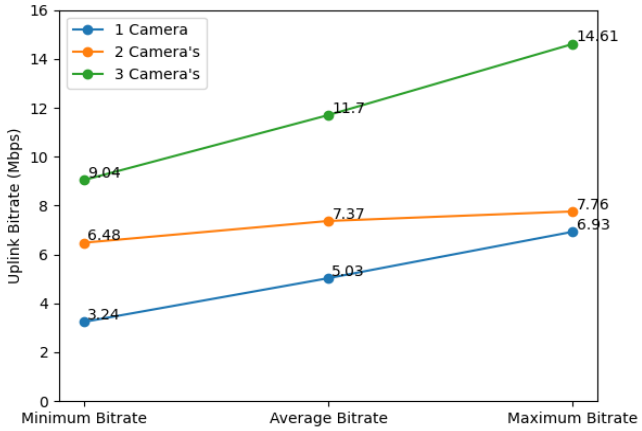

Figure 18: Uplink Bitrate for Multiple Cameras

5.4. Private 5G Use case Latency

The end-to-end latency measured for the use case was below 50 msec. This 50 msec is measured once the camera captures the first video frame until the drill press stops. All delays in the system are mentioned in Table 4.

Having both the network and use case deployed on edge helped reduce the latency significantly. The Charter team believes that further optimization could be done to reduce the overall latency further.

Table 4: Latency Causing Delays in the Worker Safety Use Case

System Delay	Definition
Camera Fetch	Time between the EC2 request a single frame from the camera till the single frame is received by the EC2
Processing	Time taken by the EC2 to run a single frame through the neural network model (YOLOv4) detect PPE objects
Code Algorithm	Time taken by EC2 to determine if the worker is compliant or not and what kind of non-compliance is happening
Code Processing per Frame	Time taken to run the overall algorithm for a single frame. Code Processing per Frame = Camera Fetch + Processing + Code algorithm.
Downlink	Time to send a command from outpost EC2 to RPi via wired, cellular, and WIFI links and RPi taking action
Relay trigger time	Time taken by relay to turn on/off
LED light change	Time taken by LED to change colors
RPi	Delay added in raspberry Pi to properly receive payloads over wireless link. At the current rate of 60FPS the RPi was unstable so we added a delay/buffer in the RPi to prevent crashes
Buzzer	Delay added for proper functioning of buzzer, without adding this delay the buzzer was out of sync with the LED and Relay
Overall Delay per frame	Time taken to receive a frame and translate the frame to give output at the RPi (end to end delay)

Finally, the worker safety use case and Private 5G network functions ran on the same hardware AWS outpost and the team demonstrated how to quickly and efficiently spin up a private network along with a useful enterprise use case.

6. Conclusion

Cloud edge computing allows telcos to provide their customers private 5G services while taking advantage of all services/offerings provided by public cloud. The Charter emerging technologies team in partnership with AWS telco team successfully implemented a P5G network hosted in AWS Outpost. The team deployed a worker safety application in this P5G environment. This project is a showcase for establishing a fully functional end-to-end 5G enterprise network, incorporating a practical video analytics application, all within a single lab hardware setup. The potential for expansion and further development is evident.

Looking ahead, the future work will explore the possibilities of deploying multiple worker safety applications across various locations, all powered by a centralized AWS outpost within the Charter network. This approach entails consolidating computational resources from individual locations into a central hub, enabling multiple enterprises to deploy their applications without the need for owning individual outposts. This centralized approach promises increased efficiency and cost-effectiveness for enterprises, streamlining their operations and fostering collaboration within the network.

Finally, CBRS proved very useful for private networks and can be leveraged to provide enterprise, government, and business-oriented 5G networks and low latency use cases.

Abbreviations

3GPP	The 3rd Generation Partnership Project
SA	Standalone
P5G	Private 5G
PLMN	Public Land Mobile Networks
NPN	Non-Public Network
PNI-NPN	Public Network Integrated NPN
SNPS	Standalone NPN
UPF	User Plane Function
SMF	Session Management Function
AMF	Access and Mobility Management Function
UDM	Unified Data Management
gNB	Next generation NodeB
CPU	Central Processing Unit
GPU	Graphics Processing Unit
VPC	Virtual Private Cloud
VPN	Virtual Private Network
OAM	Operations, Administration, and Maintenance
NAT	Network Address Translation
L3	Layer 3
MEC	Multi-Access Edge Computing
IoT	Internet of Things
PPE	Personal Protective Equipment
RAN	Radio Access Network
PoE	Power over Ethernet
RTSP	the Real-Time Streaming Protocol
YOLO	You Only Look Once
mAP	Mean Average Precision
FPS	frames-per-second
IoU	Intersection over Union
Mbps	Megabits per second
CBRS	Citizen Broadband Radio Service
FCC	Federal Communications Commission
PAL	Priority Access License
GAA	General Authorized Access

Bibliography & References

Non-Public Networks (NPN). Dongwook Kim, 3GPP MCC https://www.3gpp.org/technologies/npn

Fully Automated Athonet 4G/5G Core Management and Orchestration on AWS *https://aws.amazon.com/blogs/apn/fully-automated-athonet-4g-5g-core-management-and-orchestration-on-aws/*

AWS for Edge https://aws.amazon.com/edge/

AWS Outposts Family Overview https://aws.amazon.com/outposts/

AWS Local Zones Overview https://aws.amazon.com/about-aws/global-infrastructure/localzones/

AWS Wavelength Overview https://aws.amazon.com/wavelength/

AWS Snow Family Overview https://aws.amazon.com/snow/

