CREATING TRANSFORMATIONAL MEDIA SOLUTIONS USING 5G EDGE COMPUTING

B.Bendre, E.Gose, M. Thomas, S.P.K. Prasad, U.Mangla

IBM Corporation, USA & Canada

ABSTRACT

This paper will discuss how to create effective 5G edge media solutions. We will provide a high-level overview of edge computing, key challenges for implementing an end-to-end media edge solution and benefits of such solutions. Media use cases impacted by 5G will be discussed. We will also examine the different layers of the edge architecture including the application and network layer integrated with far edge devices, MEC (Multi-access Edge Compute) and hybrid cloud environments built on a 5G network. 5G media applications integrated with advanced network functions such as slicing with close loop automation to optimize network and application functions for media solutions will be discussed. We will present the above in the context of an implementation and conclude with lessons learned and key challenges that need to be addressed in the future.

INTRODUCTION

5G with edge computing is an emerging area which will revolutionize and transform the media industry through content creation, image recognition, video processing and OTT. The rapidly increasing number of edge nodes, management of sophisticated workloads running AI, variability of edge nodes, distribution of content at the edge and security implications provides challenges to creating end-to-end media solutions. The good news is that edge computing is based on evolution of an ecosystem of trusted technologies. Let us begin by explaining what edge computing is, the challenges and benefits of implementing a 5G edge media solution.

WHAT IS EDGE COMPUTING

Edge computing is a composition of technologies that takes advantage of computing resources that are available outside of traditional and cloud data centers. The workload is placed closer to where data is created such that actions can be taken in response to analysis of that data. By harnessing and managing the compute power that is available at remote premises, developers can create applications that substantially reduce latency, impose lower demands on network bandwidth, increase privacy of sensitive information, and enable operations even when networks are disrupted.

To move the application workload out to the edge, multiple edge nodes may be needed, as shown in Figure 1.

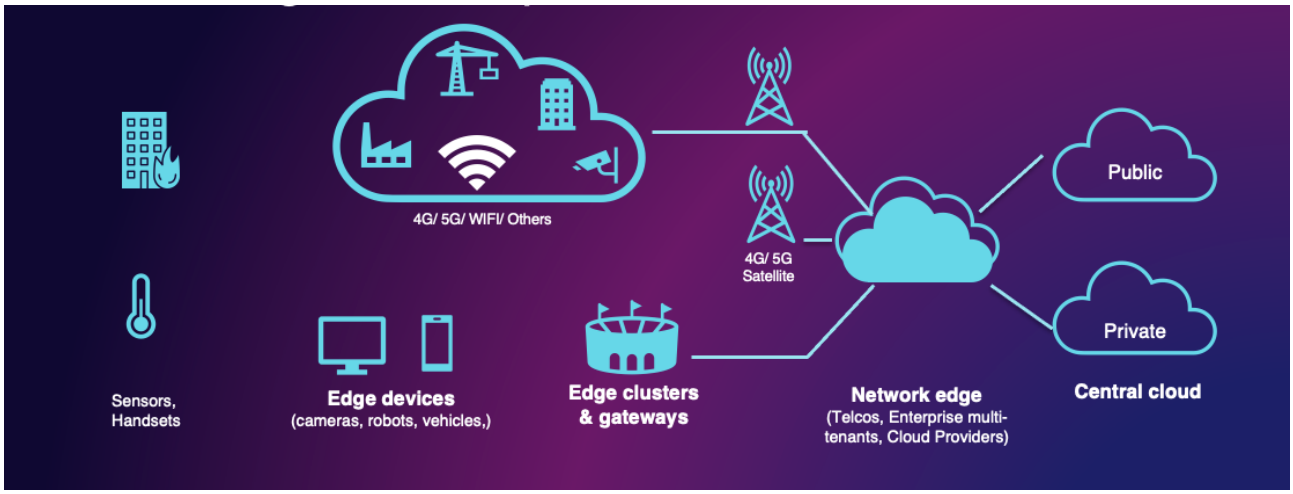


Figure 1: 5G and Edge Landscape

Some key components that form the edge ecosystem are the following:

- **Cloud**: This could be a public or private cloud, which can be a repository for the media container-based workloads including applications and machine learning models. These clouds also host and run the applications that are used to orchestrate and manage different edge nodes.
- **Network Edge:** This is generally part of the Communication Service Providers (CSP) core network which can host larger edge applications and data.
- **Edge cluster/gateway**: An edge cluster/gateway is a multi-edge compute node that is in a remote operations facility such as a factory, retail store, hotel, distribution center, or bank. An edge cluster/gateway is typically constructed with racked computers, enterprise application workloads and shared services.
- **Edge device:** An edge device is a special-purpose piece of equipment that also has compute capacity integrated into the device. Edge devices include assembly machines on a factory floor, ATMs, intelligent cameras, or automobiles.

With this basic understanding of edge computing, let's take a moment to discuss the benefits and challenges around edge computing.

Benefits and challenges of edge computing

Core benefits

The benefits of edge computing technology include the following:

- **Performance:** Near instant compute & analytics at the edge lowers latency, and therefore greatly increases performance. With the advent of 5G, it is possible to rapidly communicate with the edge, and applications running at the edge can quickly respond to the ever-growing demand of consumers.
- **Availability:** Critical systems need to operate irrespective of connectivity. There are many potential points of failure with the current communications flow. With edge computing the communication is primarily between the consumer/requestor and the device/local edge node so there is a resulting increase in availability of the system.
- **Data security:** In edge computing solutions, the data potentially never leaves the physical area where it is gathered and is used within the local edge. This means that

only the edge nodes need to be primarily secured making it easier to manage and monitor which results in the data being more secure.

Some challenges

These benefits come with some challenges. Examples are provided below:

- **Managing at scale**: Workload locations shift when you incorporate edge computing, and the deployment of applications and analytics capabilities occur in a more distributed fashion. The ability to manage this change in an architecturally consistent way requires the use of orchestration and automation tools to scale.
- **Making workloads portable**: To operate at scale, the workloads being considered for edge computing need to be portable. Moving to the edge will mean less compute resources to run the workload. In addition, adopting a set of standards can be difficult with many varied workloads to consider. Work will need to be done on how best to break workloads up into sub-components to take advantage of the distributed architecture of edge computing.
- **Security**: While data security may be easier as the data can be limited to certain physical locations for specific applications, overall security is an additional challenge when adopting edge computing. Physical edge devices may not have the ability to leverage existing security standards or solutions due to their limited capabilities. In addition, with multiple device edges, the security is now distributed and more complex to handle.

Let's next look at some of the main media use cases that are emerging with the advent of 5G edge computing.

5G EDGE MEDIA AND ENTERTAINMENT USE CASES

5G-PPP [1] funded projects in the areas of remote production and field-based production have opened new areas for innovation. These new capabilities will provide the broadcast community and other verticals with production and distribution/consumption workflows.

One use case is remote production, sending compressed real-time in-sync multi-camera feeds (including 4K) from the field (venues, events sites, outdoor sports locations…) into the cloud or to the production facility, rather than sending to an outside broadcasting unit in Electronic Field Production with all the equipment and staff. This is a complete remote production scenario, including in extreme cases multi-room distributed production, such as multiple production staff operating remotely from one another working, collaboratively, on the same live content.

Improving the fan experience during an event is another use case being developed that requires careful planning of how the media applications is architected and deployed with the underlying network functions. Improving the fan experience will include providing a second screen experience at the event where fans at the live event can further interact with the event and have a more immersive experience using 5G and edge computing. Remote production in the field using private 5G networks (i.e., Non-Public-Networks – NPNs), with uncompressed, or slightly compressed, feeds from cameras sent to the on-

site production truck which would be cableless, non-line-of-sight and high-quality field production will also benefit from edge computing. Another use case would be a future wireless studio - a vision where an all-IP based 5G NPN cableless wireless studio(s) is used, with all A/V devices connected over a 5G network which will all be IP-based

To cover live content, enhanced news gathering for live and recorded coverage benefiting from the additional uplink capacity and enhanced user density support, will become a reality with 5G.

On the distribution side, we see mass HD content distribution for consumer consumption, both live and non-live. More users, watching more content, at higher quality, with no buffering will be common. For live content, 5G broadcasting has the important potential to reduce network load, enhance viewer experience and reduce operators' costs.

Consumers viewing and experiencing AR/VR on mobile devices will be supported by 5G edge technology. The AR/VR content may be live or pre-arranged. The requirement for high bandwidth at very low latency is expected to be resolved by 5G. Similarly, eGaming and eSports which also require multi-player synching with very low latency will benefit from 5G edge computing

In addition, we see possible use cases in other verticals, where video or other media is required, making very good use of 5G. These include telemedicine – high-quality, remote access to medical experts, home treatments etc – where user can interact with media content. Additional use cases include remotely operated, or assisted, medical robots and machinery. Very low latency, high uplink bandwidth and very high reliability are key to this remote point-to-point telemedicine use. COVID is expected to further boost the exploration of these use cases, including remote patient monitoring in ambulances or other out-of-hospital locations. There are also AI-driven media uses cases where back-office AI is used to analyse and work on high-quality video coming in from mobile field devices for various outputs.

There are many media use cases that will benefit from 5G edge computing as described in brief overview above. Let us now look at the overall architecture and key components of a media 5G edge solution which could support some of the above use cases and then take a dive into implementing the use case to improve fan experience.

OVERALL ARCHITECTURE

The diagram below illustrates the key components of a 5G edge solution.

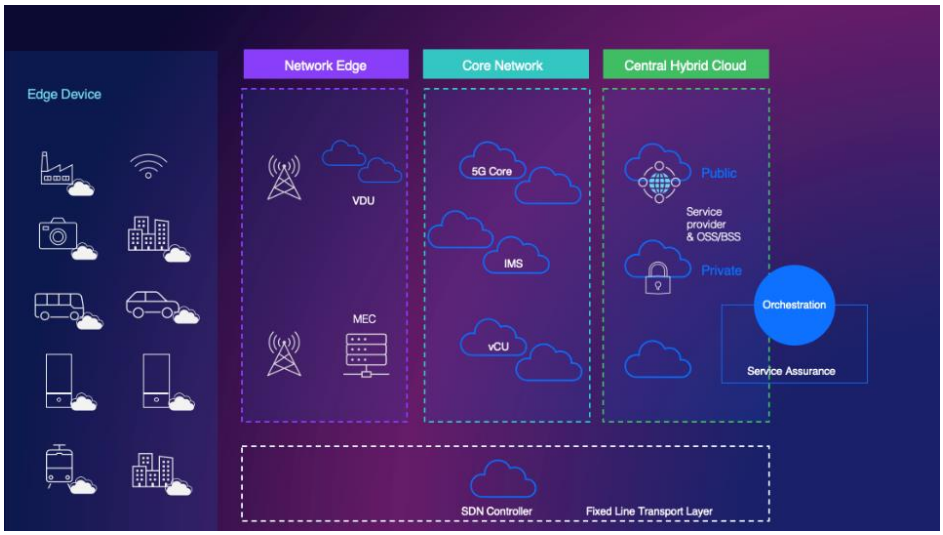


Figure 2: Architecture depicting key components of a 5G edge solution

The architecture can be broken into the following key components:

- **Edge device:** These have already been discussed in the earlier section and are small factor devices where small applications which can be containerized are deployed. Edge computing analyzes the data at the device source.
- **Network Edge:** There usually are two important components
  - Multi-access Edge Compute (MEC) [2]: Applications such as AR/VR and video recognition that need to run at the edge will be hosted here.
  - vDU [3]: The Radio Access network access is split into the Central Unit (CU) and Distributed Unit (DU). The vDU resides on the network edge and most vendors are making it available in a cloud native environment
- **Core Network:**
  - vCU [3]: The central Unit of the Radio Access network is usually placed in the core network
  - Evolved Packet Core and 5G Core: The 5G Core consists of the user plane, control plane, service-based architecture, and management is usually placed in this layer
  - IMS: Delivers IP Multimedia services to the solution
  - Transport layer. This layer provides switching, routing, and firewall security in a more scalable fashion to provide secure protection across private, public, and hybrid clouds. This creates the transport network to connect the 5G network components in our deployment.
- **Central Hybrid Cloud:** The key media enterprise systems to provision, manage and monitor the 5G components run here. The components here include:
  - Orchestration: Enables automated operations by managing the end-to-end lifecycle of virtual network services, from release management of third party Virtual/Container Network Function (xNF) software packages through to the continuous orchestration or running of xNF and service instances.
  - Service Assurance: Service assurance provides a consolidated view of events and network topology across local, cloud, and hybrid environments. It delivers actionable insight into the performance of services and their associated dynamic network and IT infrastructures. AIOps is also included to enable advanced explainable AI to be deployed across the ITOps toolchain. This AI helps in

assessing, diagnosing, and resolving incidents. It can group events, detect anomalies, and localize incidents so that one can diagnose problems faster.

- Other components: Various other components will also run here to manage and deploy the edge applications. Examples include the systems that trains and containerizes the AI models and additional systems the applications on the MEC will be integrated with. For example, an application at the MEC may need video content, but all the content cannot be stored at the MEC. In such cases, the content will reside at the hybrid cloud and the MEC applications will access the required content.

SAMPLE IMPLEMENTATION OF A 5G MEDIA SOLUTION

We have implemented multiple 5G edge media solutions using slicing, closed loop automation and Media workloads built on AI. We will now describe one of those implementations by describing the use case and implementation details.

Use Case: In a soccer game, fans may be seated in different parts of a large stadium. Each part of the stadium might be best suited for a particular view, but unless the fan looks at another screen, he or she might not get the best experience. For example, when a key event occurs, fans in the stadium must often look at the big screen in the stadium, if it exists, or might have to watch the game on their cell phones. To get the best experience, we created an app which will alert the fan in the stadium whenever there is an exciting moment in the game and provide a video stream from the best view possible. We also provide additional content related to the exciting moment such as a similar goal that was scored at another game. Premium subscribers of the app will guarantee a better experience, even if the network gets overloaded with data. We will now discuss how we achieved the best fan experience using 5G edge computing.

The assumption is that the stadium is working with a service provider and the MEC, cameras and necessary 5G equipment are deployed at the stadium. The stadium can conduct a variety of events such as different sporting events, concerts, and public events. The additional content related to the moment of interest will not be stored at the MEC but at the central cloud, as there is limited storage capacity at the MEC.

The high-level architecture was provided earlier. We will specifically focus on the following implementation aspects:

- Training of visual models
- Automated provisioning of software infrastructure
- Edge application deployment
- Closed loop automation and 5G slicing

Below is a more detailed architectural diagram of our implementation of the application portion of the use case.

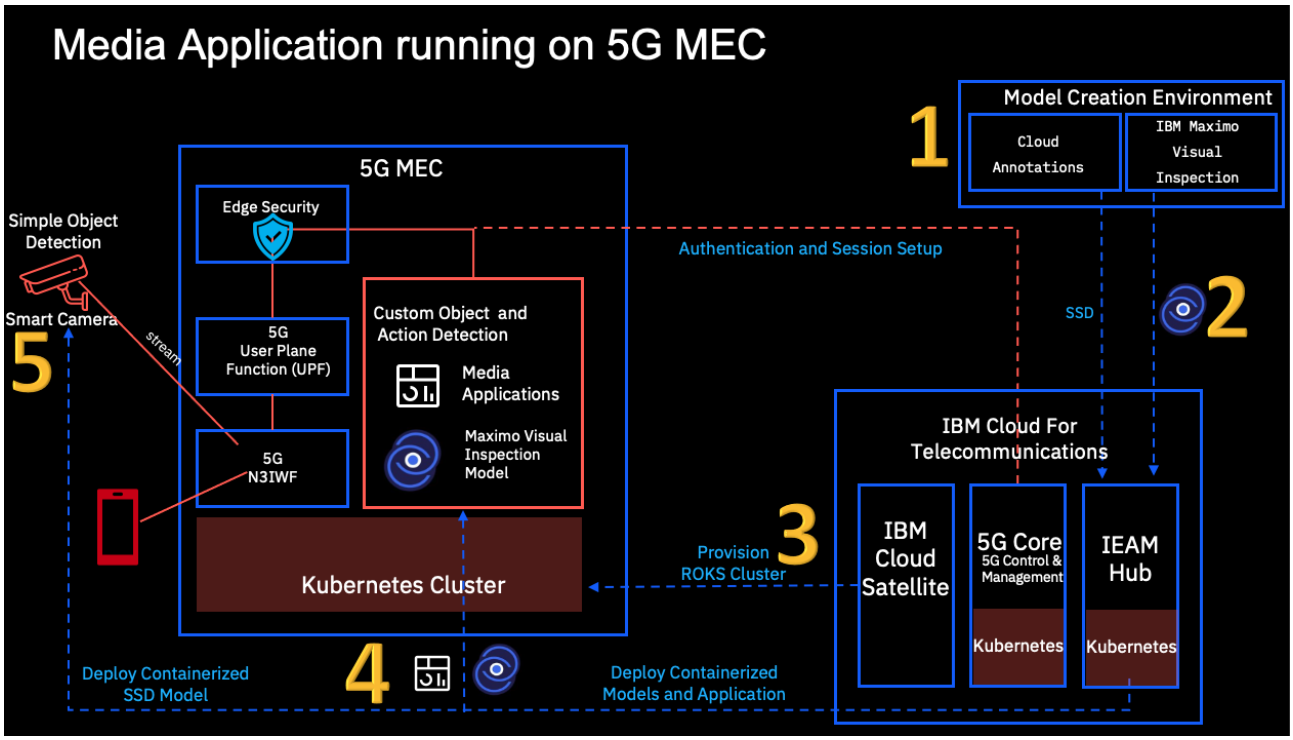


Figure 3: Architecture diagram of the Application layer

1. We trained a model to identify items of interest in a soccer game. Examples include actions such as a striker scoring a goal. a goalkeeper making a save or a player making a foul and getting a card. To recognize these actions, we trained deep learning models to identify these events. We also trained a model to identify specific objects of interest such as identifying a particular player based on identifying a player's jersey number and other objects such as a red or a yellow card. We used visual inspection tools [4] to create these models. These tools can either run on public cloud or, as in our case, an on-prem location. These video and image analysis platforms make it easy for subject matter experts to train and deploy image classification, object detection and action detection models. These models can be optimized using a variety of options. For example, the object detection models can be trained using models such as faster R-CNN (FR-CNN), Tiny YOLO, YOLO, Detectron, high resolution or Single Shot Detector (SSD). Each of these different model types have their own specific specialty and function. For example, the Faster R-CNN model is optimized for accuracy. The YOLO models are optimized for speed. The Detectron Mask R-CNN models can use objects that are labeled with polygons for greater training accuracy. High resolution model is optimized for accuracy and is suitable for training and inferencing on high resolution images. The SSD model type is used for real-time inferencing and can run on edge devices. SSD models are often as fast as YOLO but not as accurate as FR-CNN. In our specific case, we trained the object detection models using SSD and FR-CNN. We trained a Structured Segment Network (SSN) for video action detection models.
2. Once the model is trained, we containerize it and make it available on edge management platform such as IBM Edge Application Manager (IEAM). IEAM

platform enables autonomous management of applications deployed to distributed fleets of edge clusters and devices without requiring on-premise administrators. IEAM platform can run on premise, or in our case, on a Public cloud.

3. We then extend cloud services to edge clusters running at the stadium. In our case, we used IBM Cloud Satellite service to provision a managed Redhat OpenShift Cluster on an x86 system with GPU.
4. Using Deployment / Business policies, we ensure the deep learning models trained in step 1 and the media applications are deployed onto the edge cluster as well as onto smart cameras located throughout the stadium. The deployment will only happen shortly before the event happens as these models are trained specifically for a soccer game.
5. Once the game starts, the model running on each camera detects an important on-field action and starts streaming video to the Multi-access Edge Compute (MEC) located in the stadium. The streaming to the MEC only happens when the event of interest happens so there is minimal bandwidth consumed and the MEC is not constantly processing the data. Additional processing occurs on the MEC device, such as identification of the players, statistics about the identified players, real-time and historic predictions etc.

   The attendees in the stadium get a notification asking them to open the app to view the video stream from the best possible view, if they are interested. An example is shown below of what the fans will see being processed in real time.

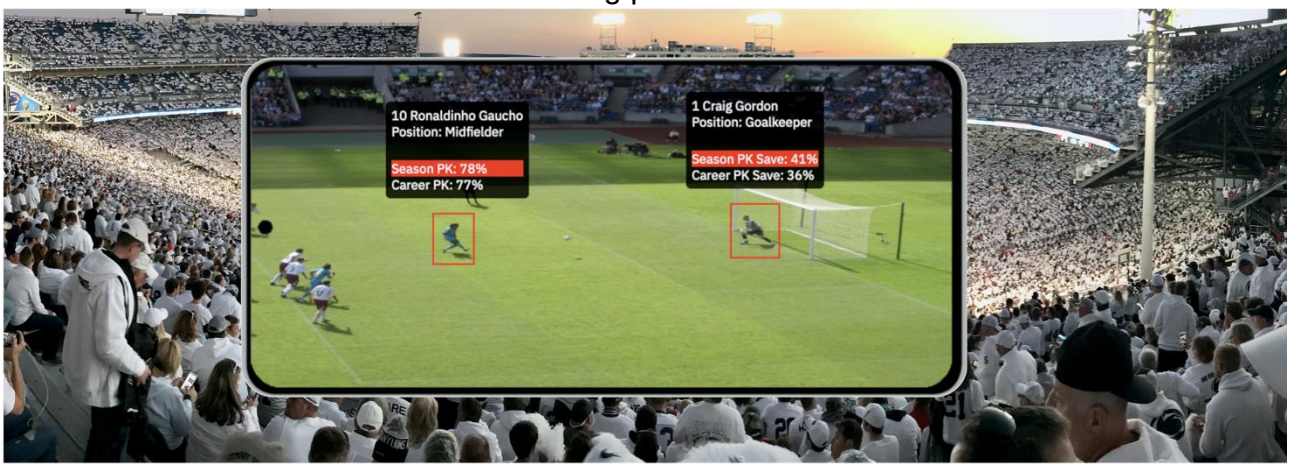


Figure 4: Video stream from the best placed camera with additional information [5]

Subscribers of the app also get exclusive access to additional videos related to the event that just happened. Below is a diagram showing how a premium subscriber of the app can swipe up to see related recommended videos. These related videos are typically stored and streamed from a media server located in the central cloud.

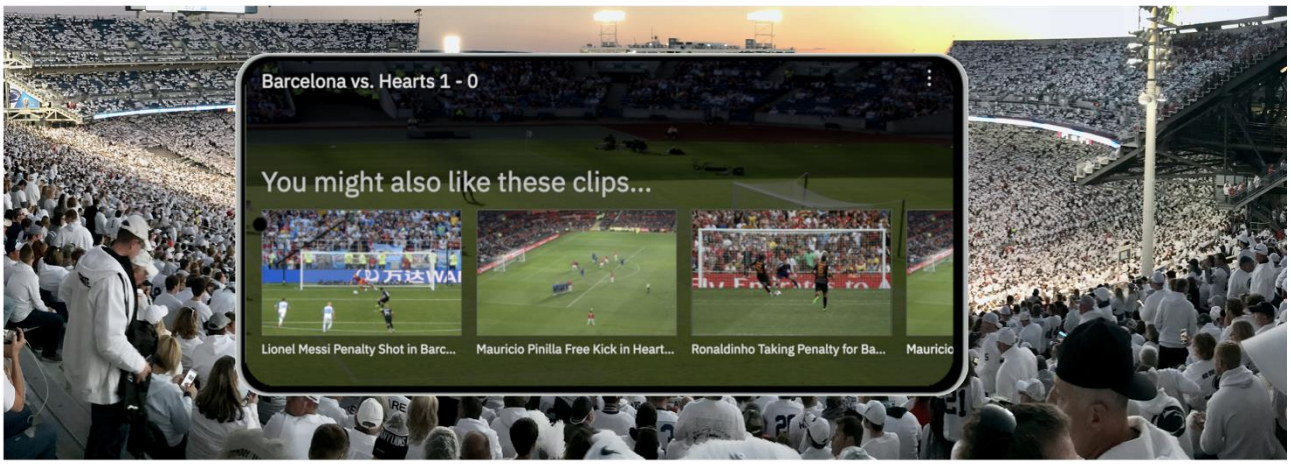


Figure 5: Additional videos related to live on field action [6][7][8]

To do this, the application on the MEC sends the results of its analysis of what happened (e.g., Ronaldo scored a goal using his head) to the central cloud. The central cloud processes this data and locates related content that will be of interest and displays them as shown above.

The above use case mentioned that premium subscribers of the app will be guaranteed a better experience, even if the network gets overloaded with data. To do this we need to consider the network layer and how it is impacted by traffic the workloads generate. For example, if a larger than expected number of fans stream the additional related videos, this can then lead to sudden bursts of high network traffic to the cloud. This often leads to poor response times, higher latency, and an overall poor experience for the fans.

To handle these bursts of sudden network load and prevent higher latency, we implemented an AI Driven closed loop automation [9] system as shown in the figure 6. This system consists of AI models that are trained on large amounts of historic data and can predict higher traffic more accurately. During the game, network probes collect network data from the various network devices and pass it onto the machine learning models in real time. These machine learning models are then able to predict higher traffic ahead of time. When such a prediction of higher traffic occurs, the AI system automatically triggers an automatized task to orchestrate the deployment of a 5G slice on the network as illustrated in figure 7. Once the new 5G slice is provisioned, the high paying premium customers are automatically moved to this slice and are thus guaranteed higher quality of service. An overview of the close loop automation and slicing implemented is provided below

- Closed Loop Automation can address issues of varying complexities at the application and network layer. Closed-loop Automation ensures the 5G network and related media / broadcasting applications operate efficiently with minimal human intervention by using AI to detect anomalies, determine resolution and implement the required changes within a continuous highly automated framework. The diagram below is closed loop implemented for the above use case.

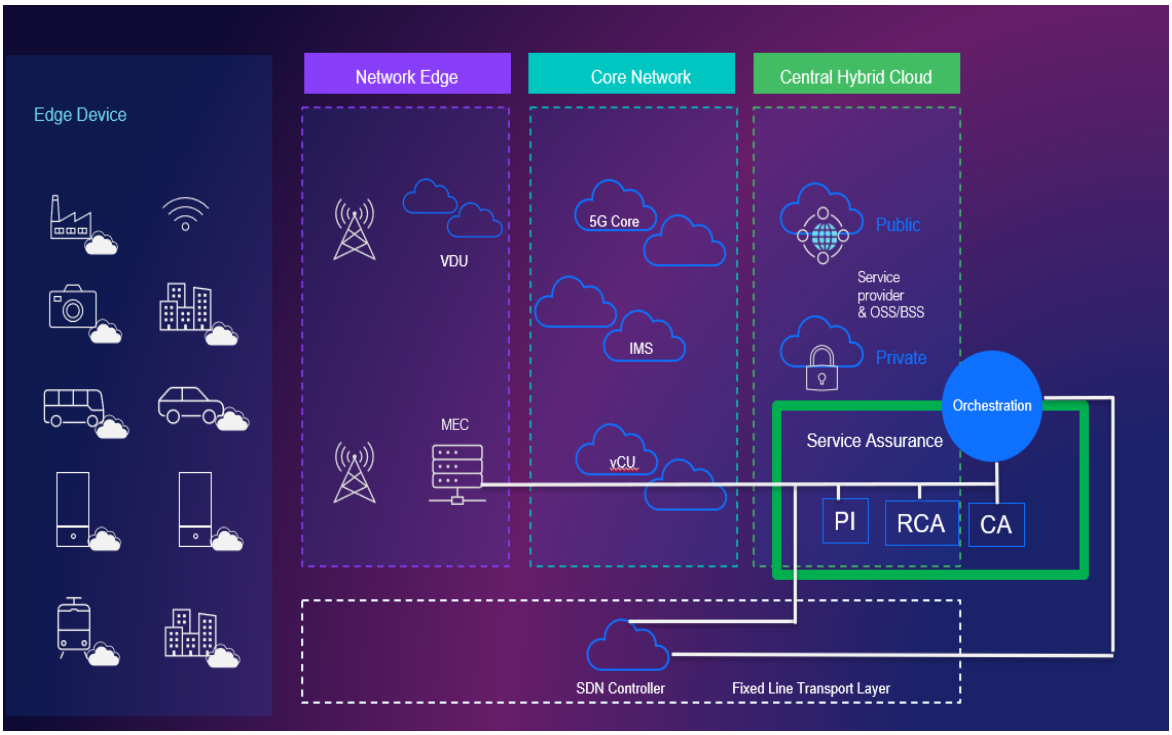


Figure 6: Closed Loop Automation we implemented

Data is gathered from various sources including the transport layer (time series network performance data from Juniper SDN in our implementation) and media applications running at the MEC (performance data about the applications, number of users connecting to the application, system performance of the MEC, various performance metric from Red Hat OpenShift Cluster). Our data consisted of logs and time series data which were used to determine the state of the network functions and applications running on the MEC and are fed into the Service Assurance Layer. AIOps is part of the Service Assurance system and includes Predictive Insights (PI) component, to predict a potential issue with the network and application, Root Cause Analysis component (RCA), to perform root cause analysis on issues and the cognitive automation (CA) system which will determine the solution to the problem. The AI system will then determine what the root cause of the issue is. It has also been trained to resolve issues and will use the orchestration tool to make the appropriate changes to the network or application layer to correct the issue. The training consists of using trouble tickets, run books and other document which the AI systems learns from. If the issue cannot be automatically resolved, a trouble ticket will be issued so a human can resolve the issue.

- When a slice is created, the orchestration layer modifies the underlying network components so that the appropriate slice is created. This will impact the vDU, vCU, transport layer and core layer. For example, a tunnel will be created in the transport layer and the core layer will be configured with the relevant parameters such as the IMSI ID of the participants who can create a slice. Please note that the full creation of a slice is quite complex and only a summary is provided here. We used network orchestration tools to provision the slice across the relevant

components. The diagram below provides a high-level overview of the slice created for the above use case

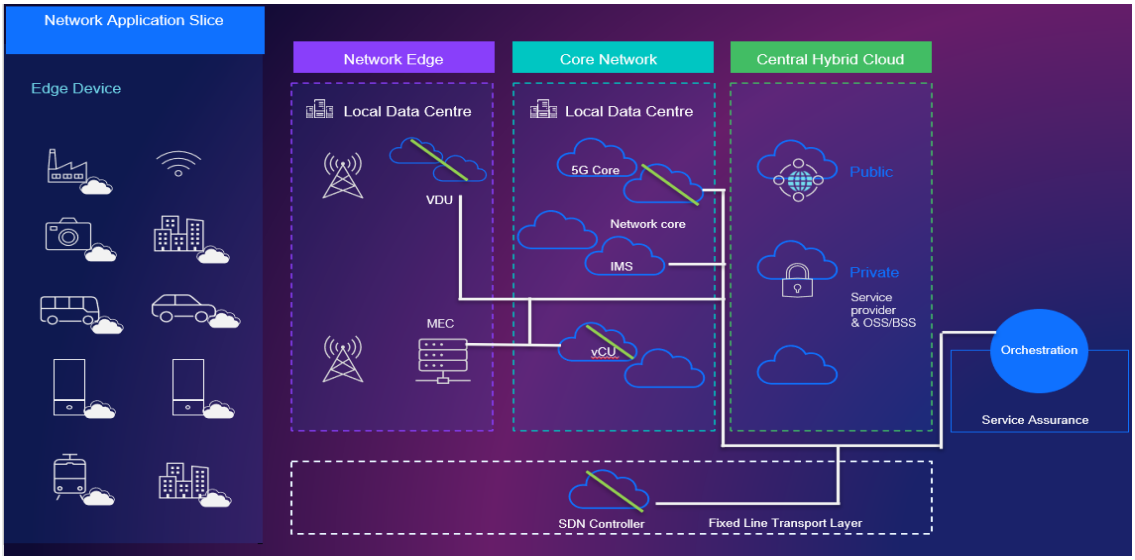


Figure 7: Creation of a 5G network slice

In summary, we have demonstrated how media workloads can be created and containerized to run at the edge to enable fans to get a better experience at an event. The processing can occur at various edge nodes including at devices such as a smart camera or at the MEC which has greater compute capacity. The different edge nodes need to be integrated and containers distributed as quickly as possible and, in some cases, the systems at the stadium will interact with systems in the cloud and other locations. Issues can occur at any of these edge nodes, and it is important to predict such issues and fix them before they deteriorate network and application performance. Closed loop automation enables the identification of such issues and remediation using AI. Such remediation can include the creation of networks slices which will automatically be created to ensure network throughput is greatly improved.

KEY LESSONS LEARNED

Some of the key lessons we learned from multiple 5G edge media solutions include:

1. Infrastructure: The underlying infrastructure and mechanism to deploy it needs to be carefully considered. The applications should ideally run in a cloud native environment but that may not be easy to do this given existing legacy applications. In our case there are multiple MEC's distributed across different stadiums and manual deployment of the Kubernetes cluster is slow. We therefore used Cloud Satellite to automate the deployment and management of the infrastructure. Media companies may outsource this activity, but they need to be actively involved as the decisions made here will impact the applications that can be deployed at the edge
2. AI: Training the models with correct data, confirming the model accuracy, and ensuring biases are not introduced is key
    - To have an accurate model, a lot of training data is required. In general, the more iterations the model is trained, the more accurate the model will be. However, in many cases the test accuracy will plateau at some point beyond which further

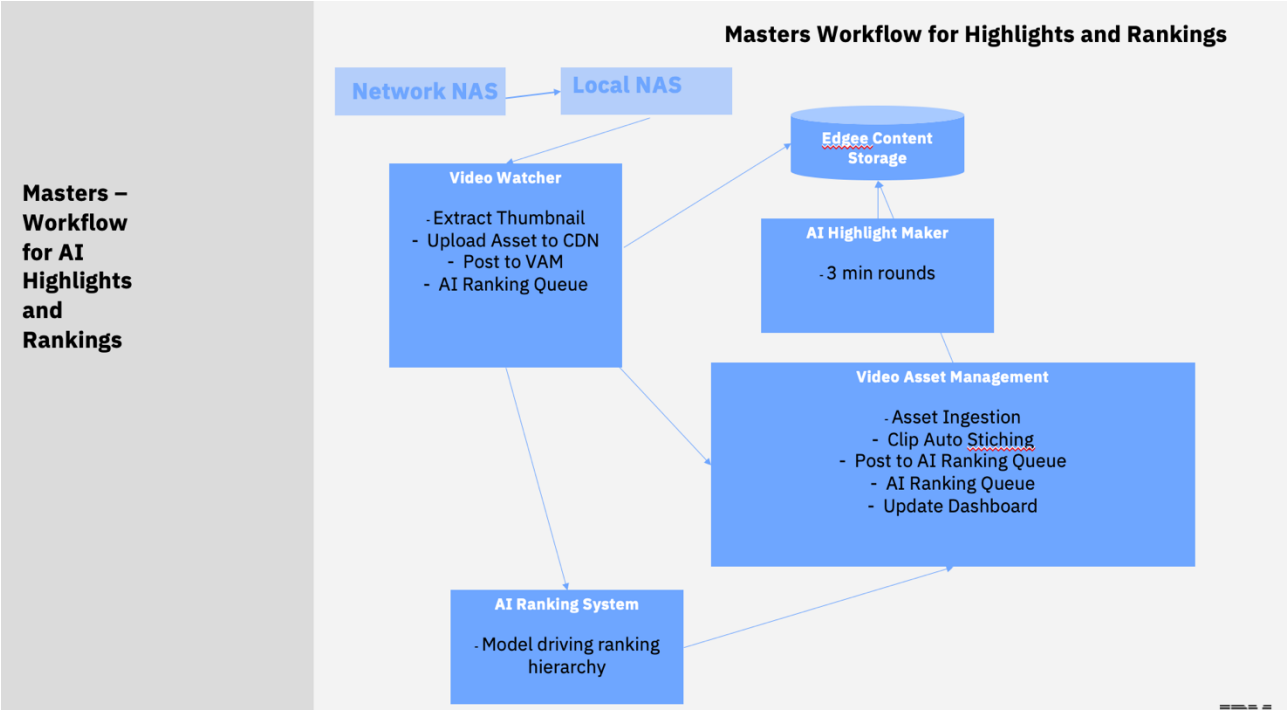
iterations do not result in a significant improvement in the accuracy of the model. For example, our tiny YOLO v2 model required more iterations to achieve a class and IoU (Intersection over Union) accuracy comparable to other SSD and FR-CNN object detection models that we ended up using. In our case, the various models had a recall between 67% to 99%.

- Depending on training factors, such as the number of data set samples, number of classes, inter and intra similarity between classes/objects and target accuracy, the optimal number of iterations may vary. One strategy for determining the optimal number of iterations is to choose a target accuracy and train the model until the desired accuracy is reached or the loss rate no longer is improving.
- Sometimes there is not enough data to accurately train the model. In our case, we improved the model by using data augmentation to add modified images to the data set, then retraining the model. Data augmentation is the use of filters, such as blur and rotate, to create new versions of existing images or frames.

3. Close Loop and slicing: Ensuring that the applications running on all edges perform well, is key. There are many nodes in an edge environment and correcting issues at these nodes needs to be automated as much as possible. In addition, it is not easy to determine if an issue is really at the network level or the application level. One solution we have implemented is for the application layer to invoke the network API and log some key network performances metrics when the application is not running well so the network and application logs can be more easily correlated. This data can be used by the AI models in the closed loop automation to perform the right root cause analysis. Creating the right slice at the right time for the right users can be a challenge. The application layer will know who the high paying customers are which need to be related to the different slices. Customers cannot be assigned to multiple slices and may need to be removed from a slice. We had to create a slice management layer that interact with the media application layer to effectively manage slices.
4. Media Workload placement: With edge computing, there can be hundreds of servers and thousands of edge devices, deployed in thousands of remote locations, with new locations or edge devices and servers being continuously added. It's very difficult for an administrator to understand the topology and relevant differences, which is critical when attempting to deploy new applications to the edge.
   Therefore, an autonomous management approach is required. This is achieved by asynchronous communication between software agents on edge endpoints and a management hub that constitute the autonomous management software. The actions carried out are based on the administrator intent (mentioned through JSON format Policy files) without his or her intervention. In our case, we used simple deployment / business policies. The constraints mentioned on these policies matched the edge server and edge device properties. For example, a business policy mentions that the models should get deployed to edge servers that have a GPU. Therefore, all edge servers that have a GPU would autonomously pull the models as mentioned in the business policy.
5. Media Application Selection and Development: One needs to be careful about selecting which applications to run on the edge. Simply running the application on the edge will not improve the customer experience. It is important to identify the right applications which require low latency and ensure that deploying the application on the MEC will fulfill the requirements. In addition, the application needs to be architected and implemented correctly. Just because an application runs in a cloud native environment does not mean it will run effectively on the edge. For example, we had to update our application to generate the logs and time series data which could be used by closed loop automation as well as integrate network correlation data so that the closed loop automation systems could do an

end-to-end root cause analysis of issues across the network and application layer. In addition, we found that our media application running on the MEC was transferring a lot of data to/from the cloud, so we used coresets [10] to transmit the required data between the cloud and MEC

6. Distributed edge nodes: Another use case is rendering highlight of a major event such as the Masters. Content is stored then quickly analyzed and made available to fans as illustrated below. Pushing the workflows to relevant edge nodes to do highlight clipping or AI ranking and sending to CDN cache was critical to meeting the customer's demands of performance on a mobile device. The architecture needs to be robust enough to support such complex distribution of the application and network nodes across different edge locations.



CONCLUSION

5G combined with edge offers a range of key technologies to enable and support key media solutions. We discussed how media solution built on 5G and edge can be created and key considerations include identifying the applications that will truly benefit from 5G and edge, integration between the application and network layers to capitalize on 5G functionality such as slicing and the management of the of the edge environment through closed loop automation.

The journey to 5G and edge to deploy various mainstream media solutions is still in its early stages. For example, a tier-1 North American CSP is already piloting a 5G edge-driven Stadium based fan experience. Their goal is to attract premium gaming subscribers on the consumption end. Remote production & rich 4k content creation built on a 5G network as part of production workflows reduce the current heavy workflow overhead. There are implementation challenges that need to be addressed including the gap

between the promised performance and what can be delivered with current technologies. This requires research and validation – technological, operational, and economic. Capital investment will be needed to support basic infrastructure, optional advanced infrastructure, and more expensive end-user technologies.

Ultimately investment priorities and Return on Investment analysis will drive the actual realization of these benefits. IoT, smart cities, vehicle to vehicle communications and gaming will compete for the same investment funding. Service delivery at promised levels and any mission critical media related use cases will continue to take priority to ensure seamless experiences for the end customer. It is therefore essential that media companies capitalize on the recent development with 5G and edge by prioritizing their investments and develop a business and technical strategy to incorporate 5G and edge into their solution portfolio.

References

[1] 5G-ppp, https://5g-ppp.eu/
[2] Multi-access Edge Compute, https://www.etsi.org/technologies/multi-access-edge-computing
[3] Prepare for 5G Networks, https://www.redhat.com/rhdc/managed-files/ve-virtualized-radio-access-network-altiostar-partner-solution-brief-f17782-201906-en.pdf
[4] IBM Maximo Visual Inspection https://www.ibm.com/docs/en/maximo-vi
[5] Justin Brockie https://commons.wikimedia.org/wiki/File:Ronaldinho_Bar%C3%A7a-Hearts.jpg
[6] Oleg Bkhambri https://commons.wikimedia.org/wiki/File:FWC_2018_-_Group_D_-_ARG_v_ISL_-_Messi_penalty_kick.jpg
[7] Duncan Hull https://www.flickr.com/photos/dullhunk/3193984762/in/photostream/
[8] Ronnie Macdonald https://commons.wikimedia.org/wiki/File:Drogba_taking_penalty_for_Galatasaray.jpg
[9] AI Closed-Loop Automation & Anomaly Detection & Resolution, https://www.tmforum.org/resources/toolkit/ai-closed-loop-automation-anomaly-detection-resolution/
[10] Robust Coreset Construction for Distributed Machine Learning, https://arxiv.org/abs/1904.05961