SCTE·ISBE CABLE-TEC EXPO 2018 ATLANTA, GA OCTOBER 22-25

2018 Fall Technical Forum SCTE·ISBE • NCTA • CABLELABS

Converging Edge Caching and Computing Power for Simultaneous Mobile and MSO Networks to Handle Latency Sensitive Services Using Co-Operative Caching

A Technical Paper prepared for SCTE•ISBE by

Sandeep Katiyar
Senior Consultant
Nokia Bell Labs Consulting
Bldg. 9A, 7th Floor, DLF Cybercity
Gurugram, India-122002
sandeep.katiyar@bell-labs-consulting.com

Table of Contents

List of Figures

Introduction

Broadcast and demand-based content networks have been pushed to their limits to reduce latency and to provide faster buffering to seamlessly deliver content. From massive data centers to edge based cache servers, caching has followed Multiple System Operators (MSOs) to the cellular edges to fulfill the demand of its subscribers in delivering emerging latency-sensitive services. Mobile and wireline operators have regularly increased bandwidth to meet growing data and new interactive service demands. But bandwidth itself does not address latency challenges. Caching has been used in services such as YouTube and Netflix to reduce video content delivery and web service latency. MSO deep fiber penetration and the future migration of cable hubs to edge clouds to enable virtualized services can be mutually beneficial to wireline and mobile services by bringing better content to mobile subscribers, providing higher quality reduced latency services, and increasing revenue.

On the other hand, with changing user habits and the resulting reprioritization of mobile data over voice services, along with smart device adoption and usage of personalized and enterprise-level mobility applications, mobile network operators face significant challenges related to redesigning the backhaul to support capacity and latency requirements for 5G deployments, If we closely look to the 5G requirements as depicted in Figure 1, densification of mobile networks is required to bring 5G to full use, leading to a dependency and need for high bandwidth access networks and content caching closer to the edge.

Besides raw data management, the low latency signaling required to coordinate and manage application data flow strains the network in terms of its performance. Greater capacity, unencumbered transmission and continuous coverage are needed. To make this happen specially for growing mobile data traffic i.e., video, the mobile network deployment method needs to be changed.

This paper provides an overview of how MSOs can provide caching to reduce latency for mobile networks. We look at cache tradeoffs, define a high-level architecture and finally discuss a new business service/opportunity for MSO edge content aggregation to meet the needs of mobile services,

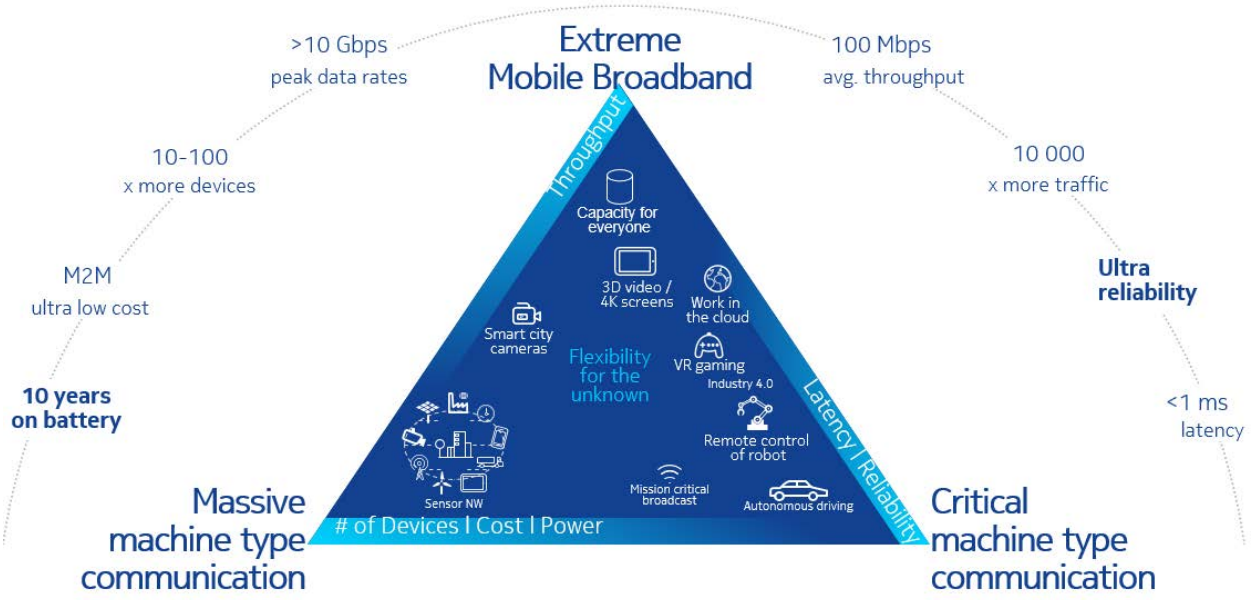

Figure 1 - 5G Requirements

Background

With explosive growth in multimedia traffic, the scalability of Over the Top (OTT) & other video services has become increasingly important. By exploiting the potential cache ability at the edge layer of Mobility and Fixed networks, the performance of multicast delivery can be improved through co-operative caching and realizing the deep reach of MSO networks. This caching technique can help minimize the average bandwidth consumption on the backhaul sides of mobile networks.

A lot of work by researchers and engineers has focused on finding effective ways to reduce duplicate content transmissions. This includes adopting intelligent caching strategies inside mobile networks and enabling edge based caches in MSO networks to access popular content from caches of nearby gateways, using selective Internet Protocol (IP) traffic offload methods. From the MSO's perspective, this also helps reduce traffic exchanged with Internet Service Providers (ISPs) and helps reduce response time required to fetch content. Both MSO and mobile networks face similar problems with respect to the content placement and its delivery, determining the size and location of each cache, and downloading to cache nodes. Co-operative caching addresses the placement issue without compromising Quality of Experience (QoE).

Video is approximately 70-80%[1] of total mobile and fixed network traffic. Given this volume and the need for caching, Co-operative caching can enable MSOs to leverage their network to position caching as a service. Such a service can address QoE and coverage aspects for capacity limited areas, helping reduce video service end-to-end latency, while reducing traffic in core and edge networks.

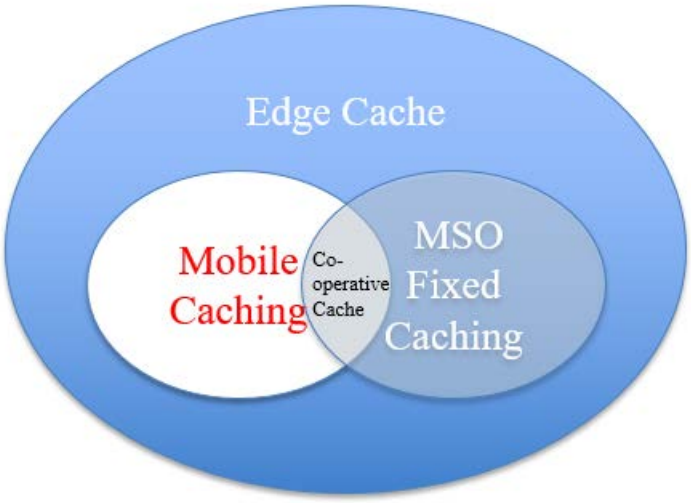

Figure 2 - Co-operative Cache

Caching Strategies

While the potential of co-operative caching within Mobile networks has been evaluated by several recent works [see Reference section], this paper focuses on co-operative caching applied across mobile and fixed

[1] Bell Labs Consulting traffic analysis

networks by providing an overview of the challenges and possible solutions using the edge caches between mobile and fixed network users. For that we need to look at the similarities and the differences between local caching in fixed and mobile edge networks.

From a caching strategy point of view, as depicted in Figure 2, mobile caches generally are either placed at edge of mobile networks providing better QoE, low latency and high complexity. or caches are placed in a core data center with lower QoE, high latency and lower complexity as depicted in Figure 4. Local MSO caches are typically available at a distance of 10-15 km from the last mile. These caches provide low latency, higher capacity and better QoE and are suitable to help achieve real densification in terms of content availability with lower transport latency and processing. That is where the relevance of co-operative edge helps to achieve the low latency and better QoE for streaming content by handling such requests directly at the co-operative edge as described further.

For both fixed and mobile networks, the challenges for caching are similar - where should the content be placed and how it should be delivered. This challenge becomes more relevant for the mobile operator, as the user is mobile and the number of users served by a given mobile network operator access node (5G gNb) may vary with time and hence becomes difficult to find efficient caching placement and optimized cost. . Another aspect impacting cache placement is content delivery latency. As the latency gets tighter in 5G networks and as transport bottlenecks appear, the performance degradation of applications that are sensitive to it, such as video calls, voice, or gaming, result in a worse QoE. Figure 3 illustrates examples of applications and latency requirements in the network. BW-efficient 360°video and 4K video streaming are two examples where co-operative caching can be used.

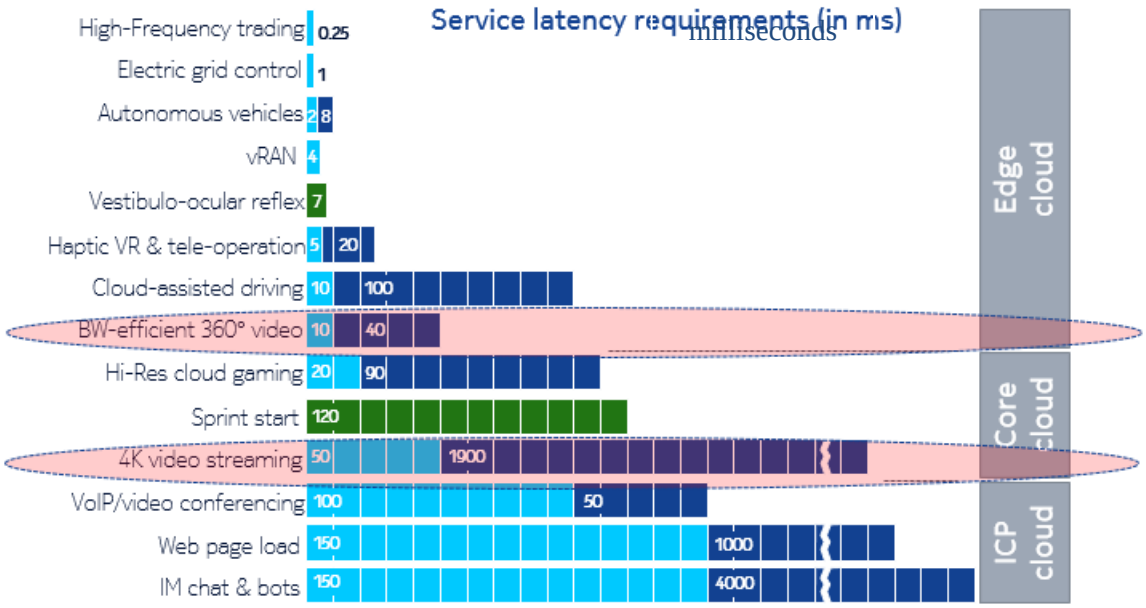

Figure 3 - 5G Application Latency demand

With less storage capacity, the amount of content that can be stored at the local cache is limited.

Processing power can also be limited, and hence it may not be efficient to run some applications from the local cache. End-to-end network complexity increases as network operators deploy, integrate and manage local caches in many locations. Resources may be needlessly duplicated if applications could be efficiently run from an alternate location.

Figure 4 illustrates three caching strategies: mobile Local Cache, Co-operative Cache, and Cloud Core Cache, based on the available latency/storage and processing capacity and distance from the base station, that can be considered by mobile operators.

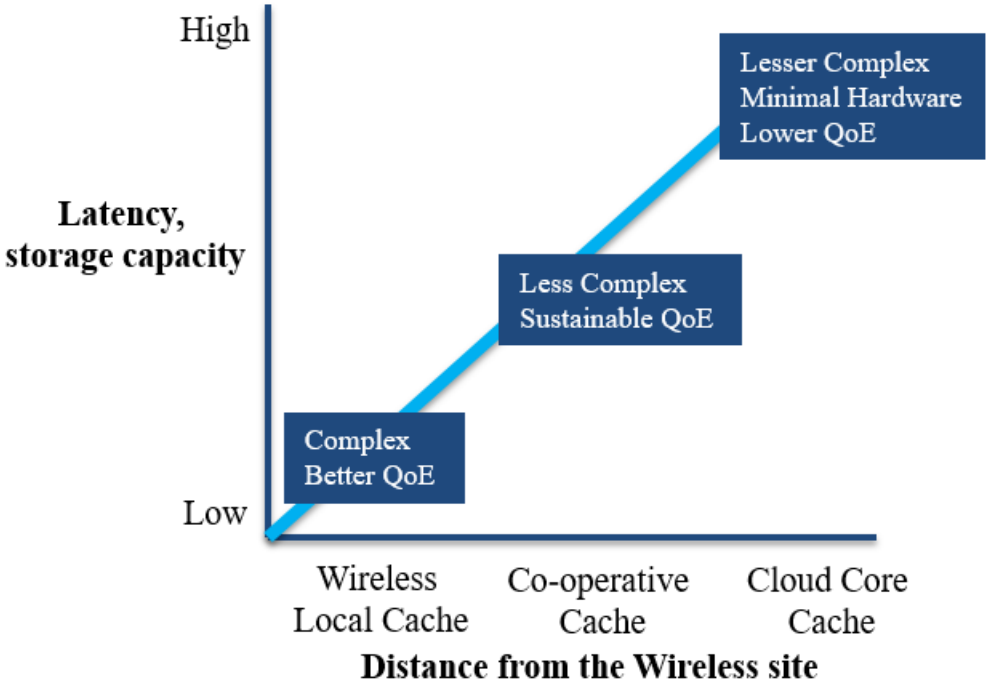

Figure 4 - Cache Tradeoffs

As described earlier, edge caching involves both content placement and content delivery. With mobile operators this becomes more challenging due to densification of the mobile cell sites in a 5G use case. Earlier studies [2]show that two factors which affect a mobile cache in comparison to a wired network are:

1. Low cache-hit probability: When proactive and reactive caching policies designed for the Internet are not effective for caching at the node and result in insufficient utilization of caches, sharing the cache among the nodes, or redirecting the streaming requests to the co-operative cache can be used as suggested in this paper.

2. Topology uncertainty: Fixed networks generally have well known node topologies for subscriber connections, while in a mobility case a user request (i.e., the user connectivity to the base station for processing a request) will always be undetermined due to its mobile nature. This further complicates the determination of expected content and bandwidth. It can be overcome with the provisioning of co-operative edge at different points of networks to cater a certain % of mobile subscribers, together with a deterministic approach of caching the popular content at the MSO co-operative edge cache.

[2]Caching at the Wireless Edge: Design Aspects, Challenges, and Future Directions. Dong Liu, Binqiang Ch en, Chenyang Yang, and Andreas F. Molisch. IEEE Communications Magazine. 2016

Deployment Considerations

Before moving ahead with deployment considerations, we need to understand some attributes of emerging 5G networks. First, 5G radio access is dependent upon the fiber-based fronthaul due to low latency service requirements. Second, due to massive densification of mobile sites specially in urban areas, there is a need for more capacity in terms of throughput and cache. In Figure 5, the Next Generation NodeB gNBs (the 5G mobile base station) are equipped with Mobile Edge Computing (MEC) servers which can be used for local cache and deliver frequently requested content.

To understand how Co-operative caching works, let's take an example where applications like 4K streaming and virtual reality streaming initiate data requests via the gNB in the 5G network. If a request for un-cached content is initiated, the co-ordination server works with MEC to first check available content sources at the MSO cache and/or the mobile operator cache server in the core network, and the latency over the paths to these content sources. This is achieved by providing feedback to MEC about the results of microburst latency results at regular time intervals. Once the requested content is cached in the co-ordination server, the content can be delivered to the user from the local cache. If the requested content is not cached in the local gNB, but is available in a nearby gNB, the content can be delivered from there.

This also opens a new way to offload all the video based content towards the MSO edge cache, thus relieving the backhaul for other bidirectional latency sensitive services.

The coordinator server plays a key role in co-operative caching, by maintaining the state of the edge nodes, path latency information, and decides the content delivery paths. Together with the MSO edge cache, which has powerful computation and large storage, they deliver co-operative caching for MSO and mobile operators.

The proposed co-operative caching framework is built on top of the gNB's MEC framework, and the MSO Edge Cache and co-ordination server potentially running on a virtualized platform in a MSO edge node as illustrated in Figure 5. The MSO edge node also consists of a headend and an edge router. Together the MSO edge node provides rich computing resources, storage capacity, connectivity, and access to cached contents. The MSO edge node interacts with the mobile operator's gNBs to handle streaming requests via the co-ordination server. To support the coordinated approach, the mobile edge and the MSO coordination server interconnects the set of service functions dynamically so that the request packets traverse the system and get processed by each service function.

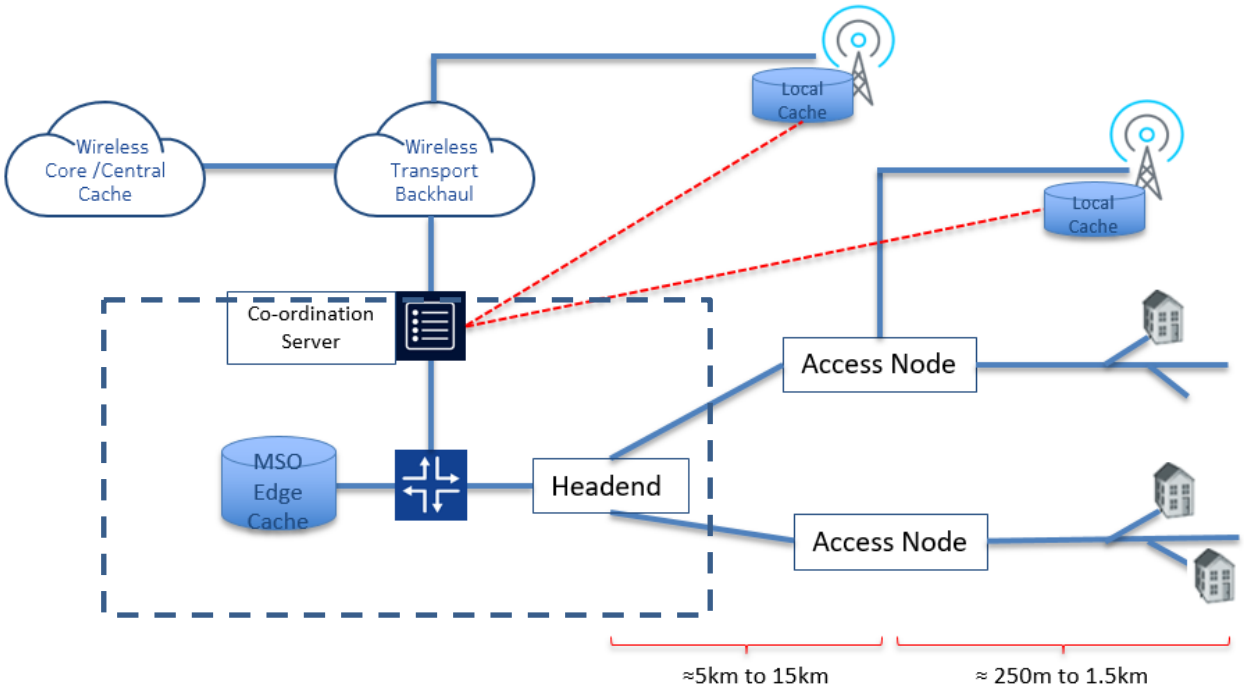

Figure 5 - Framework for Co-operative Caching

The proposed architecture thus provides two ways to implement the co-operative caching in the network:

1. Either we rely on the mobile operator backhaul as the breakout point for the streaming data, or,
2. In case the mobile operator decides to deploy a site over the MSO provided transport wherein all the streaming data is carried over the MSO network and rest of the services pass through aco-ordination server toward the mobile operator core.

The co-ordination server platform hosts the functionality required to run mobile streaming applications on top of the virtualization infrastructure. The platform hosts a set of services that can be consumed by the authorized applications. Some typical services provided by the platform include transport latency calculation function, location, and bandwidth manager. The co-ordination server platform provides visibility of the services available to the applications. If a service is provided, it can be registered in the list of services on the MEC platform so that it may redirect the request from the mobile node directly to the MSO cache server. The applications communicate with the services through well-defined application programming interfaces (APIs).

Co-operative caching between the mobile edge cache and the available fixed network MSO cache is one of the ways to support the required 5G densification while maintaining the latency, and to reduce the video traffic over the mobile backhaul. As discussed earlier, the MEC platform can provide the radio network information, latency and user location collected from the RAN (Radio Access Network) to the co-ordination server. This information is essential for the co-ordination caching server to make an optimized decision on the caching policy and resource allocation for the service type.

Figure 6 below depicts a caching framework defining the functional blocks of the coordinated edge caching system, However, to fully realize the concept, these functional blocks need to be interconnected in a sequential order, the service chaining within the system aims at interconnecting a set of network/service functions (multicast, server load balancers, HTTP header manipulation, etc.) to support

network applications. With service chaining, an operator is able to define and configure customized "service chains" in software without change at the hardware level. The service chaining helps addresses the requirement for both optimization of the network, through better utilization of resources and monetization, through the provision of services that are tailored to the service requirement context, Typically, these chains are applied to Layer 4-7 services. Here, service chaining, uses Software Defined Network (SDN) capabilities to create a chain of connected network services and connects them in a virtual chain, wherein it helps to dynamically apply or tear down single or multiple applicable services to the traffic. This capability can be used by network operators to set up suites or catalogs of connected services for use by different customers with different service and characteristics.

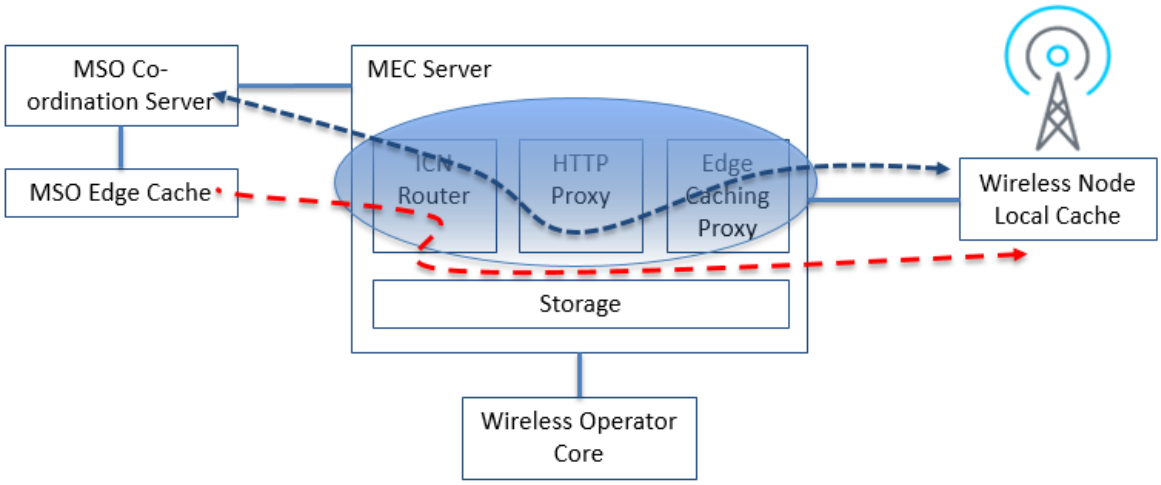

Figure 6 - Service Chaining in Co-operative Caching

High Level Business Considerations

Providing Cache as a Service can potentially be offered to mobile operators. An MSO as an streaming data redirector for mobile operators can upsell content and cache, also from a mobile operator point of view, based on current global traffic forecasts, we see an average of 70% traffic dominated by video on mobile networks refer to Figure 7. Even if 50% of that traffic is redirected to a co-operative edge, mobile operators can reduce the backhaul traffic by 30-40% and potentially improve the QoE and support densification sites by utilizing the MSO fiber connectivity at last mile.

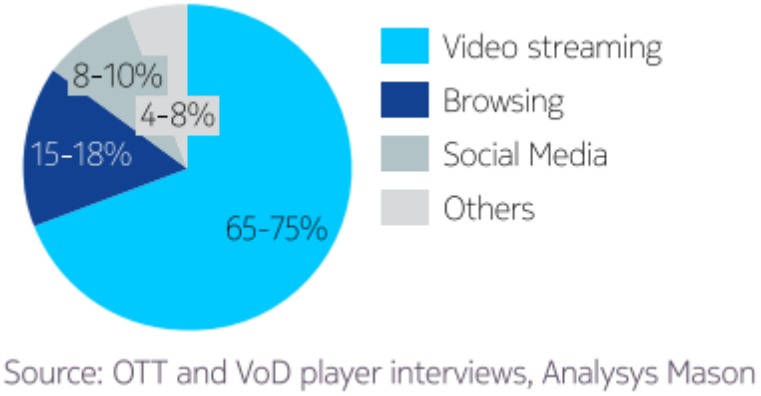

Figure 7 - Mobile Traffic by Content Type

Co-operative Edge Stakeholder	Benefits
MSO Operator	Sell Cache as a Service
	Position as OTT Traffic redirector for Mobile Operator
	Leverage unique position with the Content and application providers
Mobile Operator	Better QoE
	Better Network Resource Utilization
	Offload Streaming data to MSO for better capacity availability for bidirectional real-time services.

Figure 8 - Stakeholder Benefits

Mobile operators will continue to deploy MEC as integral parts of their network infrastructure. Although given the continuous pressure on CapEx and operational challenges, new approaches such as discussed in this paper may be welcomed.

In the co-operative cache space as depicted earlier, some new business models may arise that have a more direct and active role for MSOs, on the periphery side, and for content/application providers, on the cloud side. This in any case can be win-win situation for both from the co-operative infrastructure as provided in Figure 8, which provides the benefits which can be leveraged by stakeholders.

For example, a co-operative cache server that supports industrial applications in some localities may be better positioned with MSO networks than with mobile operator networks. The MSO may see a compelling business opportunity, and offer its services to mobile operators while a mobile operator might struggle to see a positive ROI or might not be able to assess the revenue potential. Similarly, a content or application provider may be willing to locate some of the infrastructure it needs at the co-operative edge of the network where it can serve to both and that is more effective – and potentially more cost effective – than a remote cloud location.

Smaller data analytics companies may be willing to locate processing and storage functionality at the edge in a combined environment to leverage the best of mobility and fixed access, where they do not need to own a host server but might pay only for the services they need. In this model, the MSO may deploy and pay for the initial edge hardware, but then it can monetize the investment by renting access to it to a mobile operator.

A model of this type can be mutually beneficial, in case to optimize network resources and performance.

Conclusion

This paper surveyed and provided a framework for the co-operative caching based on breaking out the streaming data traffic from the mobile edge to the fixed edge which is a one step towards integrating computing, caching and communication resources.

The issues of co-operation between the two edges and as well as some existing edge caching and computing platforms are presented. Co-operative cache edge goes beyond the centralized cloud model, which combines centralized and distributed processing, storage and control. Operators can leverage network flexibility to find the best edge location to maximize QoE and optimize network resource utilization, the main drivers for edge computing.

New business opportunities will accelerate a move to the edge, with an increased role of fixed asset owners, enterprises, and application and content providers. Traffic optimization at co-operative edge encourages a tighter co-operation of mobile operators with MSOs with clear benefits for the mobile operators.

Abbreviations

API	application program interface
gNB	next generation NodeB
HFC	hybrid fiber-coax
HTTP	hypertext transfer protocol
ICN	information centric networking
IP	internet protocol
IOT	internet of things
ISBE	International Society of Broadband Experts
ISP	internet service provider
MEC	mobile edge computing
MSO	multi service operator
OTT	over the top
QoE	quality of experience
QoS	quality of service
RAN	radio access network
ROI	return on investment
SCTE	Society of Cable Telecommunications Engineers
SDN	software defined networking

Acknowledgments

- Martin Glapa, Partner & Bell Labs Fellow, Bell Labs Consulting, USA.
- Bill Krogfoss, Principal, Bell Labs Consulting, USA.
- Ben Tang, Principal, Bell Labs DMTS, Bell Labs Consulting, USA
- R.J Vale, Principal, Bell Labs Consulting, USA.

Bibliography & References

1. Online Edge Caching and Wireless Delivery in Fog-Aided Networks with Dynamic Content Popularity. Seyyed Mohammad reza Azimi, Osvaldo Simeone, Avik Sengupta and Ravi Tandon. IEEE Journal. 2018.
2. Modeling Operational Expenditures for Telecom Operators. Sofie Verbrugge, Sandrine Pasqualini, Fritz-Joachim Westphal, Monika Jäger, Andreas Iselt, Andreas Kirstädter, Rayane Chahine, Didier Colle, Mario Pickavet and Piet Demeester. Conference on Optical Network Design and Modeling. 2005.
3. Power at the edge. Monica Paolini, Senza Fili. 2017.
4. A Distributed Caching Architecture for Over-the-Top Content Distribution.Rui Dias∗†, Adriano Fiorese†‡, Lucas Guardalben†, Susana Sargento∗†. 14th annual conference on WONS. 2018
5. Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems. Xiaofei Wang, Min Chen, Tarik Taleb, Adlen Ksentini, Victor C. M. Leung. IEEE Communications Magazine. February 2014.
6. Toward Smart and Cooperative Edge Caching for 5G Networks. Haitian Pang∗y, Jiangchuan Liuy, Xiaoyi Fany, Lifeng Sun∗. IEEE Journal on selected ares in communications. 2018.
7. Caching at the Wireless Edge: Design Aspects, Challenges, and Future Directions. Dong Liu, Binqiang Chen, Chenyang Yang, and Andreas F. Molisch. IEEE Communications Magazine. 2016.
8. A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications. SHUO WANG1, XING ZHANG1, YAN ZHANG2, LIN WANG1, JUWO YANG1, AND WENBO WANG1. IEEE Access. 2017.
9. Edge Computing and the Role of Cellular Networks. Guenter Klas, Vodafone Group. The IEEE Computer Society. 2017.
10. Content-Exchanged Based Cooperative Caching in 5G Wireless Networks. Shu Fu, Peng Duan, and Yunjian Jia. IEEE. 2017.
11. Collaborative Edge Caching through Service Function Chaining: Architecture and Challenges. Lei Lei, Xiong Xiong, Lu Hou, and Kan Zheng. IEEE Wireless Communications. June 2018.

