Optical Networks in Edge Clouds: Energy and Application Dimensions

Daniel C. Kilper[1], Tasha A. Adams[1], Fatma AlAli[2], Ronald D. Williams[2], Malathi Veeraraghavan[2]
College of Optical Sciences, University of Arizona, 1630 E. University Blvd., Tucson, AZ 85721, USA
Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904
{dkilper, tadams}@optics.arizona.edu; {fha6np, rdw, mv5g}@virginia.edu

Abstract: *Optical networks provide capacity and low latency that can enable new edge cloud computing based applications, including 5G baseband processing, personal computing, and connected vehicles. The role of optical networks in these applications is examined and the efficiency of the associated networks is reviewed.*
Keywords: *Optical core/metro/data-center network architecture, design, virtualization, slice, control and management*

I. INTRODUCTION

Edge computing has been gaining interest as a technology platform to support emerging 5G wireless applications. Concepts such as edge clouds or Fog computing have emerged as an approach to achieve many of the efficiencies and benefits of cloud computing at the edges of the network. Working at the network edge is considered to be essential due to the need for deterministic, ultra-low latency (~ 1 ms or less) together with high network speeds (> 1 Gb/s to users) at increased reliability [1]. End-to-end optical links from the antenna to the server provide the best performance against each of these requirements. The difficulty is in the cost and complexity of optical networks, particularly when optical switching is used to gain cost efficiency. The use of optical switching can introduce signaling and control plane delays that compromise the latency benefit. In fact, the control in passive optical networks (PON) can introduce delays that create non-deterministic round trip latencies exceeding the 1 ms target. Research to address this problem generally takes one of two directions: (1) increasing capacity and reducing the latency of PONs, (2) reducing cost and complexity of optical circuit switching, typically in the wavelength domain. These are essentially attempts to adapt the primary access and metro network architectures to meet in the middle. Each can be used to support edge computing. Here we primarily look at approach (2), however, much of what is considered can be readily applied to approach (1).

A central question to how optical networks will be used in edge cloud environments is where does the processing take place? Today, processing generally takes place both in the user equipment and in the network. Significant processing capabilities are required in the user equipment because the access and core network round trip latency is > 20 ms and lacks the reliability and availability that is generally needed for most use cases (see Fig. 1, left). Furthermore, the cell baseband processing is performed at the cell tower for similar reasons. With improvements in reliability, availability, capacity, and latency expected with 5G wireless networks, the situation has the potential to see dramatic architectural changes (Fig. 1, right). Most notably, processing may be performed at any location within the network edge, which here we define as extending from the user equipment to the edge cloud infrastructure, entirely encompassing the access network. Maintaining round trip delay of 1 ms or less, is sufficient for virtually any human interface, including tactile controls such as touch screen navigation or precision remote controls.

In recent years, there has been much focus on making use of this new edge environment to move the 5G baseband processing from the cell site to nearby edge cloud computing infrastructure. Within an edge cloud, better economies of scale can be achieved for the very large number of cell sites anticipated for 5G. However, a similar opportunity exists for personal computing and IoT device computing.

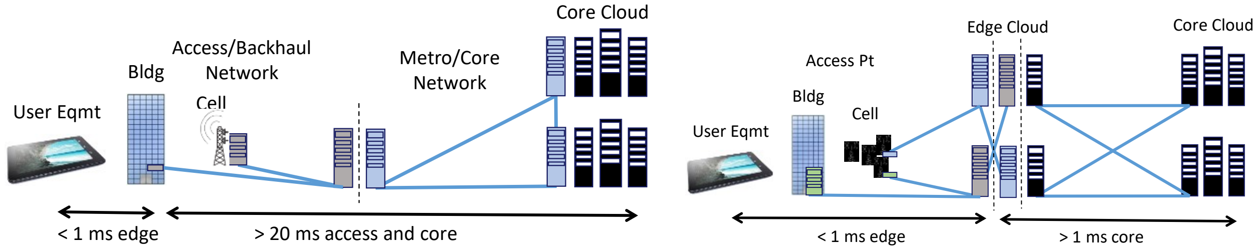

Fig. 1. (left) Current hierarchical network model; (right) edge-cloud network model.

Zero-client computing is a remote, virtual desktop platform that enables disaggregation of the user controls and interface from the processing and storage. Unlike thin clients, a zero client does not provide an operating system for the user and instead provides simple network and security functions to extend the link to the user interfaces over long distances from the computing environment. The zero client is essentially a peripheral device to the cloud based servers.

The advantages of zero clients are many including, reduced cyber-attack surface, low power user equipment, shared computer use and management, and high performance computing using a low cost interface. In today's networks, cloud based zero clients suffer from the long and tenuous network connection to the cloud. However, zero clients are routinely used with high-performance graphics machines over local area connections, typically within the confines of an enterprise campus or building. Deterministic, millisecond latency is clearly needed in order to avoid a noticeable degradation of the zero client computing experience, for example in touch screen applications [1]. We recently found that the zero-client user experience also falls off with packet loss rates and noticeable quality of experience degradation occurs for as little as 0.5% packet loss [2]. Improved network performance through an edge cloud, 5G wireless network has the potential to support high-performance zero client use across an entire community. This extends to IoT applications such as video analytics on surveillance video, using zero clients. The entire community served by the edge cloud becomes similar to an enterprise campus.

The full spectrum of computing in this edge cloud environment, from user equipment based computing to zero client cloud based computing will have differing merits depending on the application use case. For example, frequent video frame transmission will place a greater load on the network and therefore favor shorter distance transport such as edge cloud or user equipment computing. On the other hand, infrequent, low rate applications are likely to be better served by more centralized cloud computing. Here we review several key application cases and survey recent studies based on their energy efficiency. Applications include radio baseband processing, personal computing, and connected vehicles.

II. SHARING AT THE EDGE

One of the key motivations for moving to the cloud for 5G baseband processing is the potential efficiency improvements. The primary benefit in this regard comes from the sharing of processing infrastructure and for some applications the centralization of processing resources. The trade-off is typically the cost of transporting the raw and unprocessed data. For 5G radio signals, this corresponds to transport of either digitized or analog I-Q radio signals, delivered over fiber optics directly to/from the antenna. For personal and IoT computing applications the transported data can vary widely. Many sensor applications, referred to as narrowband IoT (NBIoT), use very low bandwidth and therefore specialized low bandwidth transport mechanisms such as low power wireless access (LPWA) have been proposed. Other applications, however, such as video surveillance and personal computing may require the transport of streaming video signals. Streaming of high definition video formats can require large data rates, especially when there is limited processing available to compress the video data. When personal computing is disaggregated across the network, the display video stream accounts for most of the data. Furthermore, the video quality is related to the amount of data transmitted to the terminal device, as lost packets are retransmitted [2]. Similar to the efficiency benefit with 5G cloud processing, there is an energy efficiency benefit to sharing the access network infrastructure [3].

III. SELECTED EDGE-CLOUD APPLICATIONS

A. 5G Baseband Processing

Here we treat radio baseband processing as a cloud application, although it can also be considered a service upon which various applications are delivered. Baseband processing can be divided into multiple 'split-PHY' processing jobs, which perform the various baseband processing operations. Each PHY split processing step reduces the transport capacity requirements of the digitized radio signals through the network. The process is similar to data compression in that there is a processing cost to reducing the data that must be transported over the network. We recently examined the impact of split-PHY processing on routing and wavelength assignment in a WDM based optically switching edge cloud network [4]. As the split PHY processing reduces the amount of traffic carried in the network, it can be more advantageous to route to the nearest edge data center for split PHY processing rather than directly to the final destination site where the application is processed.

The placement of edge cloud resources for baseband processing within a network is also an important consideration. We recently studied the optimized placement of such resources [5]. Larger optical network capacity and path diversity generally decrease the number of edge data centers that are needed to obtain high throughput, and latency requirements will force data center placement near more isolated nodes.

B. Personal Computing

The energy efficiency of cloud computing was previously investigated for public and private clouds, where the private cloud model is essentially an edge cloud model [6]. This analysis showed that the largest efficiency gains come from the degree of sharing of the computing resources. Additional benefit, however, does come from processing at the edge (in the private cloud) versus the public cloud. Efficient optical transport at the edge is particularly important for this bandwidth intensive application. Larger energy efficiency benefits to using both edge and public clouds occur when processing intensive computing is involved.

Centralization of personal computing also provides benefits in terms of centralized management. More efficient computing infrastructure can be used and better maintained for performance and security. With a zero client, there is no need for a firewall between the zero client and the edge cloud and therefore the extra delay and energy of firewalls can be avoided. Firewall appliances tend to be among the least efficient network elements [7]. The entire community can be protected by data center grade firewalls to the outside world.

C. Connected Vehicles

The current focus of autonomous vehicles is to place sufficient processing on board to handle all necessary processing for vehicle control and navigation. This is largely driven by the fact that sufficient network connectivity cannot be relied upon under all operational environments. Nevertheless, there is widespread agreement that enhancement to the control and navigation of vehicles through network connections is desirable and has potential for significant benefits. These benefits may come from sharing rich data with nearby infrastructure and vehicles or accessing additional processing external to the vehicle. For certain vehicles, such as small, lightweight drones, it may not be possible or desirable to provide sufficient processing capabilities onboard and therefore zero or thin client approaches may be needed.

Connected vehicles also provide an important use case for capacity management based optical switching. The flow of automobile traffic and congestion events will create scenarios in which the network capacity will need to follow such events, which evolve over minutes and hours. We recently showed how machine learning based data traffic prediction can be used to anticipate similar evolving capacity demands on diurnal time scales and allow for advance reconfiguration of the optical network to vary the capacity to nearby edge cloud infrastructure, for example through software defined network controls [8]. Optical capacity management can reduce provisioned capacity needed for a given level of performance, reducing energy and cost. Of course, the 5G capacity will still be needed at the access points and that is expected to be there. The ability to dynamically scale the optical capacity and make use of split-PHY processing will allow for further overall efficiency improvements that will be needed in scaling 5G networks.

IV. CONCLUSIONS

Optical networks will play an important role in edge cloud computing used to support a wide range of new applications. High speeds and deterministic low latency from optical connections are expected to open up many new applications. In particular, it will enable a disaggregation of baseband processing for 5G and a disaggregation of personal computing within the edge cloud service area. Disaggregation of the user interface from the computational hardware is already occurring within enterprises. Edge clouds provide an environment in which this zero client approach might be used across an entire community. Connected vehicles may also benefit from efficient capacity management in optical networks supporting 5G cloud based processing.

ACKNOWLEDGMENT

The authors acknowledge Rider Foley, Yao Li, Clayton Banks, and Bruce Lincoln for their insightful discussions related to the ideas and development of the work described here. This research was supported by a grant from the National Science Foundation, #CNS-1737453 and #ECC-0812072.

REFERENCES

[1] G. Fettweis and S. Alamouti, "5G: Personal Mobile Internet Beyond What Cellular did to Telephony" IEEE Comm. Mag. 140-145, February 2014.

[2] F. Alali, T. Adams, R. W. Foley, D. Kilper, R. D. Williams, M. Veeraraghavan, "Objective and subjective evaluation of zero client computing under network packet loss" submitted for publication.

[3] C. Gray, R. Ayre, K. Hinton, R. S. Tucker, "Power Consumption of IoT Access Network Technologies" IEEE ICC Workshop on Next Generation Green ICT 2818-2823, 2015.

[4] J. Yu, Y. Li, M. Bhopalwala, S. Das, M. Ruffini, D. C. Kilper, "Midhaul Transmission Using Edge Data Centers with Split PHY Processing and Wavelength Reassignment for 5G Wireless Networks" Proceedings of the Optical Network Design and Modeling (ONDM) Conference 2018, 178-183.

[5] Y. Li, M. Bhopalwala, S. Das, J. Yu, W. Mo, M. Ruffini, D. C. Kilper, "Joint Optimization of BBU Pool Allocation and Selection for C-RAN Networks" Proceedings of the Optical Fiber Communication (OFC) Conference, paper Th1B.5, 2018.

[6] J. Baliga, R. W. A. Ayre, K. Hinton, R. S. Tucker, "Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport" Proc. of IEEE 10.1109/JPROC.2010.2060451, 2010.

[7] D. C. Kilper, G. Atkinson, S. K. Korotky, S. Goyal, P. Vetter, D. Suvakovic, and O. Blume, "Power Trends in Communication Networks" IEEE J. Sel. Top. Quantum Electron. 17, 275, 2011

[8] W. Mo, C. L. Gutterman, Y. Li, G. Zussman, D. C. Kilper, "Deep Neural Network Based Dynamic Resource Reallocation of BBU Pools in 5G C-RAN Networks" Proceedings of the Optical Fiber Communications (OFC) Conference, paper Th1B.4, 2018