Managing latency-sensitive applications

A conversation with Neeraj Pandey, Associate Vice President, Vodafone Shared Services, Vodafone

Monica Paolini, Senza Fili

Managing latency-sensitive applications

A conversation with Neeraj Pandey, Associate Vice President, Vodafone Shared Services, Vodafone

By Monica Paolini, Senza Fili

There is much talk about latency right now, fueled by 5G and its requirements for ultra-low latency. I had the pleasure of talking with Neeraj Pandey, Associate Vice President at Vodafone Shared Services in India, about the role of latency in ensuring high QoE in our networks today and in future 5G networks.

Monica Paolini: Neeraj, can you tell our readers what your role is at Vodafone?

Neeraj Pandey: I have a team that takes care of the network engineering for different Vodafone OpCos in Europe.

The team looks after the design requirements for the large group of Vodafone data centers, which provide different hosting services to Vodafone OpCos. We also take care of deployment, which is part of network engineering.

Recently, we also got involved, as are other leading operators, in various transformation activities around software-defined networking, network function virtualization, and telco or cloud activities.

The team is involved in designing the solutions, selecting the vendors, and conducting the PoCs. It's a collaborative team, and part of the activity is conducted from India, the other part resides onshore in Europe.

As you know, virtualization and cloudification of the network, with leading-edge technologies such as network slicing, cloud RAN, and mobile edge computing, are prerequisites for the rollout of 5G. All those new initiatives require virtualization of the packet core, as well as the RAN.

Monica: Latency is a very hot topic today, because low latency is crucial for many real-time applications. How is the role of latency changing from a few years ago, to now and into when we move to 5G?

Neeraj: The transition has already happened, particularly with the wide adoption of smartphones. A lot of applications are being developed for smartphones. Predominantly, OTT players are using the existing mobile network as infrastructure to stream various services to the end users.

Another reason why latency will become critical is that the end user not only is the content consumer but has become the content producer as well. There's a lot of traffic on the upstream side, in addition to the downstream side.

For real-time applications, voice is still predominant, but interactive video applications and real-time gaming are critical to providing the end-user with real quality of experience.

Now there's more and more talk about latency. The entire ecosystem is going to change with the coming of IoT, where every user is supposedly going to have at least ten connected devices. There's going to be a massive surge of connected devices.

There are various estimates of the number of devices, which may be on the high side. According to a survey from Ericsson, by 2020 there will be 50 billion devices. If we take the active mobile user population into account and assume that each user will have ten connected devices on average, the figure still goes up to 20 billion connected devices.

These devices will have different latency requirements. Low latency may not be critical for smart meters and some other IoT applications. But applications such as autonomous control of cars and other vehicles that can be virtually assisted by the network are highly sensitive to latency.

There are new applications like AR and VR – augmented reality, virtual reality – as well as tactile internet, remote surgery, and manufacturing processes, where the latency requirement could be as low as 1 ms. Latency is becoming very critical.

Monica: When the user becomes the content generator, latency becomes important not only in the downlink but in the uplink too.

Neeraj: Exactly. A challenge for the operators is the asymmetric nature of latency. The latency in the upstream and the latency in the downstream will be different, and it has to be measured, tracked and managed differently.

The monitoring and measurement tools have to be very precise in measuring the latency each way. We cannot typically calculate the one-way latency simply by dividing the RTT, the round-trip time, by two. Latency needs to be calculated in each direction to identify the precise bottlenecks and address them.

Various protocols are available. TWAMP, Two-Way Active Measurement Protocol, is one of them. All OEM equipment, whether you're talking about routers, switches, or other equipment, have the capability to measure the latency based on timestamps for this protocol. This is a protocol used a lot by operators for the latency measurements in one direction.

Another problem with latency comes into play because of the different behaviors of protocols such as the Transmission Control Protocol – TCP – and the User Datagram Protocol – UDP.

80% of the applications use TCP, but TCP, as a reliable protocol, has its own characteristics. It works on a self-clocking mechanism, where the center or the source adjusts the sending speed depending on the feedback from the receiver.

Particularly if the network has some bottlenecks or congestion, or if any of the packets drop, the source exponentially backs off. This has a lot of implications for the overall throughput in the network. We need to take into account what applications are running.

This is a significant change from four or five years back, when the network was agnostic about what applications were riding on the network. Now things have changed. Most of the interactive or real-time applications use UDP as a protocol, whereas the other web-based applications use TCP.

If you do not have some queuing algorithm, maybe AQM or any other random early detection, then normally, when there is congestion, the tail-drop effect takes place. Then you cannot give end users a good quality of experience.

You have to be very much aware of what kinds of applications are riding on your network, and you have to implement the quality of service or some queuing mechanism so that your UDP traffic gets precedence, because it's much more sensitive than the TCP traffic.

Monica: Can you resolve these issues with TCP or UDP?

Neeraj: Yes. For example, in gaming applications, they call this is the ping effect. A latency or delay of 1 s could spoil the enjoyment of a game. It becomes that critical.

The network calibration has to be done meticulously to avoid any congestion in the network. Because of the nature of TCP, you provide buffers at various devices where they can store the data.

If the buffer size is very large, there's a phenomenon called buffer bloat, where all the packets are queued, including all UDP traffic. It will spoil the enjoyment of watching a live video, of a VoIP communication, or of real-time gaming.

You have to be very, very intelligent to find out how to calibrate your network so that your end user – whether using a real-time application or an ordinary web application that runs on HTTP – receives a fair amount of bandwidth.

Monica: In the end-to-end network, from the RAN to the core, where do the latency problems arise from, in your experience?

Neeraj: First of all – and this is an interesting development, particularly with 4G and in the future with 5G – the latency in the RAN is shrinking. That means the latency in the typical RAN is coming down vis-à-vis the backhaul or the core network.

In both the core and the backhaul, there's a law of physics at work: the propagation of light.

There are two factors which add to the latency in the backhaul and the core. One is, of course, the propagation time, and second is the processing time, because if you have multiple devices, then every device takes time to process a packet.

Of course, buffers or serialization delays and other factors add to the latency. In 4G, RAN latency is 10 ms or less, whereas the latency on the backhaul or the core could be significantly higher, maybe in the range of 30 to 40 ms.

Similarly, you have to take into account what kind of latency should be there for the control plane, for the user plane, for the synchronization plane, and for the management plane.

All these latencies need to be defined and measured across the network. And before benchmarking the network, you should be very sure in what areas you can minimize the latency.

As a result, nowadays more and more content is moved to the edge so the subscriber need not traverse all the way to deep in the core to access content that is latency sensitive.

Monica: As you said earlier, there're different dimensions to latency. The first step is to understand them. The second step is to optimize your network by managing the different dimensions of latency.

Neeraj: That is where we talk about having a very good measuring tool. Old tools like ping or a trace route do not work to benchmark your current network adequately.

You have to have a timestamp mechanism, and monitoring or measuring has to be in real time. You have to be very watchful about latency, particularly on the transmission side.

You want to know whether there are any network elements – routers, switches, or any other transmission equipment – that add a lot of latency to the network. That means you have to do the calculation on the hop-by-hop basis.

Wherever you find any rogue element, you need to take care of it. Sometimes the backbone was created earlier, and it carries not only mobile traffic but other traffic as well.

You have to have strict quality of service. If you cannot afford to replace the equipment, you need to give precedence to mobile traffic. And the quality of service has to be managed end to end.

LTE has been methodical and meticulous in defining different quality of service class indicators, or QCIs – 0 to 15. There are QCIs for real-time applications, which are different from user-plane IMS signaling ones. They deal with guaranteed bit rate, non-guaranteed bit rate, and all that. It's the job of the operator to ensure that the same quality of service is defined end to end. That is the only way you can manage the latency.

Secondly, you have to identify, on an almost daily basis, the nature or the type of the traffic: how much your end consumer or the end operator is downloading one particular stream, how much time they spend accessing Facebook, or accessing Google or maybe Netflix videos or any other application.

Then content caching needs to be done at the location nearest the end consumers so they get a good experience.

Monica: You mentioned that you want to know what your subscribers do, on an application basis, in real time. It's no longer just looking at historical data. Tracking all of this, it's a big challenge. How do you do that?

Neeraj: You can do tracking in various ways, and now we have the availability of big data and good analytics tools. You can get insight into consumer behavior by analyzing that data.

In fact, the data on consumer behavior is not only an investment but could also be a potential revenue generator. Because you can pinpoint subscriber behavior: at what time do they access Facebook, or when are they watching the TV or video, when are they downstreaming the video?

Also, you can analyze the data by segment. What are the youths' preferences? What are the working professionals' preferences? What are the housewives' preferences?

You have to be intelligent enough to catch hold of what the end user is consuming. Then you can customize your network or other contents according to the customer preference.

Monica: You mentioned network slicing earlier. Is network slicing going to help you to reduce or manage latency?

Neeraj: Yes, very much, Monica. There are two killer technologies. One is edge computing. It was previously known as Mobile Edge Computing – now they call it Multiple-access Edge Computing. The second is network slicing.

Edge computing has a number of advantages. The first is that it gives you the power to virtualize your RAN so that you can have your baseband unit, your BBU, on a virtualized platform.

Also, the edge is the place where you can host different applications or content that an end user wants to access with the least latency. Also, you can host applications for augmented reality and virtual reality, and greatly reduce latency.

You can tailor the edge computing to give latency of 1 ms, as well, or tactile internet. That is one area.

Also, because of the virtualization, you can have your control plane or user plane accordingly placed in the network, rather than everything placed in the core. You can systematically put it or plan it to go on that edge. That is the edge computing.

On the network slicing, as we mentioned, the entire ecosystem of business models is changing. The 5G ecosystem will cater mainly to three different applications. The first is enhanced mobile broadband. The second is ultra-reliable, low-latency communications. The third, and most important, is massive machine-type communications.

There's one core that is catering to all these segments. That one core, you either have to over-provision or under-provision it. It is definitely not meant to cater to multiple segments that have different requirements for latency, throughput, bandwidth, customer experience.

With network slicing, you've got one physical infrastructure available, and you can divide it into different logical networks. One slice of this network can be dedicated to taking care of one application.

For example, in smart metering, the requirement of data throughput or latency will not be very high, but your signal should reach to the basement. You can create a particular slice that will take care of that.

Suppose you've got users who want to download a streaming video or enjoy a 4K movie or ultra-high-definition TV. For them, you have to create a separate slice, so they can enjoy their content.

Suppose you have an MVNO, a mobile virtual network operator, which is riding on your network. You can create a different slice for it.

Network slicing is going to be helpful in addressing the latency, bandwidth, and other requirements of different segments of the business.

Monica: That's very important, because that allows you to lower the latency where you need to because you cannot have extremely low latency for all the applications.

Neeraj: Exactly. Operators are facing this challenge because increasing data volume is not necessarily getting translated into increasing revenue. They have to protect their revenue, and also offer good quality of experience or differentiated services to different segments of the business.

Monica: With 5G, there's a lot of talk about ultra-low latency. You mentioned 1 ms. How much of a challenge is it for the networks to get to that level of latency?

Neeraj: It's a significant challenge because we are talking not only about a stationary subscriber, but also about the moving subscriber with applications such as autonomous driving or assisted driving. That means your network should be agile enough to deliver this ultra-reliable low latency. Latency has to be both ultra-low and reliable. Otherwise, it could be catastrophic.

For that, as I have said, edge computing and network slicing are important.

Another important area is network densification, because the network should cover with cells each nook and corner of a particular area, and should have a footprint with adequate coverage to give these applications around them.

It's going to be exciting and a challenge for an operator to deliver 5G.

Monica: Absolutely. It's also going to be a huge opportunity to have a network that delivers the service that the subscribers want.

Neeraj: Exactly, because this is going to be the key for the incumbent operators to differentiate themselves from the normal over-the-top players and the competing operators. It's going to be the key differentiator for them.

Glossary

4G	Fourth generation
5G	Fifth generation
AQM	Active queue management
AR	Augmented reality
BBU	Baseband unit
HTTP	Hypertext Transfer Protocol
IMS	IP multimedia subsystem
IoT	Internet of things
IP	Internet Protocol
LTE	Long Term Evolution
MEC	Multiple-access Edge Computing
MVNO	Mobile virtual network operator
OEM	Original equipment manufacturer
OpCo	Operating company
OTT	Over the top
PoC	Proof of concept
QCI	QoS class indicator
QoS	Quality of service
RAN	Radio access network
RTT	Round-trip time
TCP	Transmission Control Protocol
TWAMP	Two-Way Active Measurement Protocol
UDP	User Datagram Protocol
VoIP	Voice over IP
VR	Virtual reality

About Vodafone

Vodafone Group is one of the world's largest telecommunications companies and provides a range of services including voice, messaging, data and fixed communications. Vodafone Group has mobile operations in 26 countries, partners with mobile networks in 48 more, and fixed broadband operations in 19 markets. As of 30 June 2017, Vodafone Group had 523.5 million mobile customers and 18.5 million fixed broadband customers, including India and all of the customers in Vodafone's joint ventures and associates. For more information, please visit: www.vodafone.com

About Neeraj Pandey

Neeraj Pandey is Associate Vice President of Vodafone Shared Services, based in India. His team of Certified Solution Architects is responsible for delivering solutions involving software-defined networking, Network Functions Virtualization, orchestration, and software-defined data centers. A 20-year technical veteran, Neeraj has extensive experience in building up long-haul optical networks. His assignments have included the planning, rollout, optimization, and operation of cellular networks – RAN, backhaul, packet backbone, circuit core, and packet core.

About EXFO

EXFO develops smarter network test, monitoring and analytics solutions for the world's leading communications service providers, network equipment manufacturers and webscale companies. Since 1985, we've worked side by side with our customers in the lab, field, data center, boardroom and beyond to pioneer essential technology and methods for each phase of the network lifecycle. Our portfolio of test orchestration and real-time 3D analytics solutions turn complex into simple and deliver business-critical insights from the network, service and subscriber dimensions. Most importantly, we help our customers flourish in a rapidly transforming industry where "good enough" testing, monitoring and analytics just aren't good enough anymore—they never were for us, anyway. For more information, visit EXFO.com and follow us on the EXFO Blog.

About Senza Fili

Senza Fili provides advisory support on wireless technologies and services. At Senza Fili we have in-depth expertise in financial modeling, market forecasts and research, white paper and report preparation, business plan support, strategic advice, and due diligence. Our client base is international and spans the entire value chain: clients include vendors, system integrators, investors, regulators, and industry associations. We provide a bridge between technologies and services, helping our clients assess established and emerging technologies, leverage these technologies to support new or existing services, and build solid, profitable business models. Independent advice, a strong quantitative orientation, and an international perspective are the hallmarks of our work. For additional information, visit www.senzafiliconsulting.com or contact us at info@senzafiliconsulting.com or +1 425 657 4991.

About Monica Paolini

Monica Paolini, Ph.D., is the founder and president of Senza Fili. She is an expert in wireless technologies and has helped clients worldwide to understand new technologies and customer requirements, create and assess financial TCO and ROI models, evaluate business plan opportunities, market their services and products, and estimate the market size and revenue opportunity of new and established wireless technologies. She frequently gives presentations at conferences, and writes reports, blog entries and articles on wireless technologies and services, covering end-to-end mobile networks, the operator, enterprise and IoT markets. She has a Ph.D. in cognitive science from the University of California, San Diego (US), an MBA from the University of Oxford (UK), and a BA/MA in philosophy from the University of Bologna (Italy). You can reach her at monica.paolini@senzafiliconsulting.com.