Latency as a service

A conversation with Mansoor Hanif, Director of the Converged Network Research Lab, BT

Monica Paolini, Senza Fili

Sponsored by EXFO

Latency as a service

A conversation with Mansoor Hanif, Director of the Converged Network Research Lab, BT

By Monica Paolini, Senza Fili

Operators have always measured latency and jitter as key KPIs, but with 5G, latency has become a hot topic. Can operators meet the stringent requirement of 1 ms latency in the RAN – and if so, when and where?

I had the pleasure of talking with Mansoor Hanif, the director of Converged Network, Research at BT, about latency's increasingly prominent role in wireless networks, and about how low-latency service may become a revenue generator in its own right.

Monica Paolini: What's new about latency? Why are we all of a sudden talking so much about latency?

Mansoor Hanif: Latency has become one of the key characteristics that distinguish 5G from 4G or any other technology. The reduction of latency that's being made possible with 5G is quite extensive and variable.

There is a range of different latency values that 5G makes possible. They allow us to broaden the services we offer into areas where we wouldn't have been capable before. It helps us expand into new markets.

I would add, though, that at BT Research we are focusing on ultra-reliable low latency, which is one of the three main use case categories of the initial version of 5G.

Reliability is key, and this leads to my main point here, which is that it's not so much trying to get the latency as low as possible, but it's about being able to manage the latency at a specific, reliable level, and being able to maintain that latency target for a specific user. That's why we talk about ultra-reliable low-latency communications: because we need to be able to specify latency for a certain customer, and then guarantee that latency over a specific period.

That's what we're going to do with 5G. People sometimes forget that it's not just about giving the lowest latency, it's about being able to manage the latency consistently.

People are talking about latency a lot because they haven't got their heads around that second point, managing latency. There's a whole debate about what we should aim for as a target for low latency, whether it should be 1 ms or below that.

It should be below 1 ms. But what are the costs involved? And what do we need latencies below 1 ms for? People have been struggling to define the services or applications that require those levels of latency. To be honest, there aren't many today, but you can be absolutely assured there will be many, many coming up in the future. There are already certain services that require 1 to 2 ms latency.

The debate has been focused on how low latency needs to be and what type of services will be able to make use of latency that low. As I said before, this is missing the point. It's more about if you can manage

latency for a customer. And what services can you enable with the managed latency you provide?

Monica: Before we discuss how to manage latency, can you tell us what is your definition of ultra-reliable low latency, and what are the main use cases we see today?

Mansoor: Ultra-reliable low-latency communications, uRLLC, require latencies that are anywhere from below 1 ms to around 7 or 8 ms.

That's quite a big range. Once you get below the 5 to 2 ms mark, you're talking about robotic and machine-critical use cases. The really, really low-latency use cases are for high-end industrial systems that require that type of low latency to make them work.

That's why I believe it's a very niche market that wants 1 ms. The cost for a customer to apply a 1 ms latency, compared to a 6 or 7 ms latency, can be prohibitive, because you have to optimize a lot of the resources on the end-to-end chain to satisfy the latency requirements for that customer. That can be quite an astronomical cost.

That's where network slicing is key. I believe we should be able to provision latency as a service. If somebody wants low latency for a specific period, there should be a fair cost associated with that, related to the amount of resources we need to satisfy that latency requirement.

It's an interesting case for network slicing and SLA agreements, in terms of how we monetize this service. If people don't need that latency, they can pay for standard latency. Standard latency could be 6 or 7 ms, which has a lower relative cost. That's pretty much doable in most cases in 5G.

Other use cases might require a latency of 3 ms. We should be able to provision that for those customers. We should be able to manage that latency and guarantee it for those customers. And we should be able to have the service assurance wrap around that.

Again, I would go back to my original comment, which is that it's about a managed latency. If customers have several types of robots or industrial devices around their site or around the country or even around different countries, in many cases, they need those devices to be synchronized very closely.

What that means is measuring the end-to-end latency from a central point or from each of the local points to those devices, and then, making sure the delta latency between those devices is managed to a certain extent.

A lot of the applications that we're looking for in terms of automation are looking for that type of managed latency, so they could coordinate devices and compensate for latency that might be higher on one device than on another. I think that there is a huge demand for these services.

Synchronized drone operation is a good example. Where you need the drones to act in synchronous mode, you need them to be aligned in latency. Now, if those drones were going over two separate networks with different latencies, what we need to do is have an intelligent system that can calculate the lagging latency, and then manage both latencies to the point that they're synchronized. Otherwise, you're never going to have dancing drones, unless their dancing means crashing into each other.

What I'm trying to say is that the managed latency – for instance for drones, which is another good application for managed latency – is the key to monetizing the service.

Monica: You're right – traditionally, we've been measuring latency as a KPI, but not managing it. This is an entirely new way to deal with latency. Latency is something you cannot avoid. It's always been there, always will be there. But how do you manage it?

Mansoor: In the simple case, where you've got several devices that are on the same network, and it's a 5G network, if our customer wants an SLA of 4 ms latency to this specific set of devices across the UK, then we will provision that over a network slice.

The network slice would be configured in a way that it will provision just for those devices. The customer might only need that latency from 9 am to 5 pm, during factory hours, and it may need it in five factories.

That's the kind of network slice that we believe, as a service, we can monetize and offer. We can give a value to our customer that they can't get anywhere else, because there isn't any other way you could do that.

Typically, that's a relatively simple case, because you're managing the whole infrastructure in which those devices are running. It's simply a question of putting in the right network capabilities, the right orchestration and the right slicing infrastructure, applying the SLA, and then providing service assurance to check that you're meeting that SLA.

If that slice is coming into conflict at any time with other slices, obviously that's why we need the slice orchestration to manage the conflicts within the system.

The more complex case is when you're only controlling part of the infrastructure. We're doing some interesting tests. I can't reveal too much right now, but we're doing some international tests of federated slicing capability across multiple operators, to see to what extent we can cooperate with other operators through federated slicing by offering this on an international, global level.

In that case, what's happening is that you're operating across several countries, and you have different infrastructures over which the devices are running. If, in the same way as roaming today, you have a federated slicing agreement with the operator on the other side, the case becomes quite similar to the first case I talked about.

It could be that in some of the countries, you don't even have a 5G network. You may be running over Wi-Fi, where there's no managed latency. In that case, what we need to offer is a service layer on top, which measures the latency in real time, calculates the "long pole in the tent" (i.e., the lagging device) in terms of the highest latency, and then compensates for delays to all of the devices to bring everybody within a certain delta of the highest latency. That would be another type of managed latency.

All of this should be capable under a single platform. Adding a layer of software on top to manage the latency is something we're already looking at with some of our partners, such as Unmanned Life, in their autonomy-as-a-service software.

Monica: Basically, you'd be able to charge for service much like how you charge for capacity. If in one location you can have only that much capacity, you charge less than the place where you have more. Latency becomes a metric on which not only can you have an SLA, but also you would get revenues, right?

Mansoor: Yes, you're absolutely right. We see latency as one of the parameters, one of the variables that we can offer as a service. Again, it's managed latency, which means we can set a target and meet that target.

If we're not managing the whole infrastructure on a global level, what we can agree to is a delta, but we can compensate what's in our control to match what's not in our control.

This is why it's important for people to understand that slicing is going into the domain of offering things like latency as a service, a managed service. People are still fixed in the old mindset of just slicing the core, just slicing this bit or that bit, which is boring stuff. We're way beyond that now.

We're talking about doing what Amazon did when they created Web Services. They said,

"OK, how can we slice and dice computer systems, and what are the bits we can sell as a service?" The first thing they came up with was storage, and the second thing they came up with was compute.

Then, they came up with other services: "This makes sense that we can slice and dice computer infrastructure and offer this as a service." It's exactly the same with network slicing.

We haven't even begun to explore all the dimensions of network slicing, but latency is just a nice way for people to get this into their heads. This is what we're talking about.

Another example is beams in the RAN. If you've got 16 beams, in the future of beamforming, we could take one beam and allocate that to a customer. What type of customer would benefit from that? You could have a terrestrial antenna, and a drone flying at 300 feet; we could allocate two beams in a certain area to track that drone over the same infrastructure.

That's another slice. That's another service. The dimensions of network slicing are much, much bigger than we expect. Over the next year, you're going to see, I believe, a much bigger understanding of the full dimension of what can be offered by network slicing.

Latency is a nice way into that, because once you understand the concept of managed latency, you start to understand the overall concept of network slicing.

Monica: Edge computing is also going to have to play a big part in managing latency. It makes management of latency much more powerful.

Mansoor: Absolutely. MEC is a key enabler and a building block, because once you have a good service to offer, like managed latency or a private LTE network or industrialized private networks, you don't want to wait until you have full national coverage before you start to monetize that.

Through multi-access edge computing, MEC, you can go to a corporate customer and check what latency they're getting now. One option is to build an early 5G bubble of service around their site. That's doable. Another option is to put in the MEC capability on site to guarantee, already, certain latency requirements for certain automated operations, as you build out over the wider network.

It gives you that flexibility of being able to offer consistent service through various tool sets, much earlier than having to wait for the full 5G network to be rolled out in the whole country.

Monica: How does the need to manage latency change the way we measure it?

Mansoor: It's becoming much more demanding, number one. The type of SLAs we'll be looking to implement are much more demanding than ever before, just in pure performance terms. What we're asking for from our partners in the test and measurement, monitoring, and service assurance world is that extra level of precision, number one.

Number two, as we've just discussed, the variables are expanding for the services we're looking to assure. They're going to be much wider and much more diverse than ever before. We need the same flexibility in the tools we use to assure those services.

On both fronts, it's quite a big challenge, but it's interesting to see that there're quite a few companies in the service assurance field who have understood this is a major opportunity for them, as well. They're already facing that challenge by expanding their capabilities.

Monica: The traditional way of measuring latency is two-way latency. Is this sufficient when you're trying to use latency as a service?

Mansoor: No, the dimensions of what we offer as a service of latency would be quite different. It's not good enough just to have

a single point-to-point measurement of round-trip latency or even one-way latency up and down.

What we'll be looking for is a way to provision various checkpoints across many parts of the network, where our customers will be. Then to measure the aggregate latency to those end points and be able to define a delta at which we expect all our latency to be managed. If we're falling out of that latency, the measurement tool needs to send us alerts. If we're getting close to the edge, send us alerts, which would then need to trigger the network slicing algorithm, in certain cases, to allocate more resource to bring that latency down for that particular customer.

Monica: You need to know where in the network you are accumulating latency, so you can manage the SLA. But also, you need to have it at the network slice or application traffic level. You're going to have multiple measurements of latency, and you would have the flexibility to look at them all and act on them, right?

Mansoor: Exactly. As I said, I think the SLA for latency is likely to be relevant to corporate customers in the beginning, or mission-critical service providers. Their devices will be spread out across the network in various locations.

We need a way to make sure our service assurance tools can be spread out virtually, following a customer to those locations, so we manage the aggregate latency – what we're guaranteeing them – and identifying where latency is failing, getting close to the SLA.

It's a very different type of problem than what we've had in the past. We're not only measuring more variables, but the variables are getting more and more demanding – latency getting lower or throughput getting higher, capacity getting higher. And the number of variables is getting wider.

We mentioned latency, but also power, throughput, beams, etc. What we're offering as a service is going to be expanding, but also how we measure that needs to be more and more tailored to the individual customer. And it needs to be flexibly provisioned across wide areas of our network in a very specific way, which is pretty neat.

Monica: How long is it going to take before you reach the vision you're outlining?

Mansoor: At BT Labs, we're starting up with speeding up our Converged Digital Infrastructure Lab with a range of partners. We've been working on slicing for quite a while now, with some of the big tier-ones. Our focus for this year is going to be on multi-vendor end-to-end slicing.

We're looking to bring in some of these service assurance partners to work with us, because that multi-vendor capability is key, and we don't think there's been enough done on that. Some of the big tier ones have been holding back too much, because they're quite protective about some areas of the network, and therefore they don't want to input too much into things like RAN slicing.

We feel that it's absolutely fundamental, so we'll be pushing a lot on that. BT is co-chairing the new End-to-End Slicing Project Group in TIP. We've gotten a lot of traction already, and we'll be using that to get a lot more momentum behind end-to-end slicing.

We'll be doing our own work in the labs, too. We'll be hooking that in to what we do with the TIP Community Labs and the End-to-End Slicing Project to make sure what we do is shareable across the industry with other operators and other entities, so we can accelerate that end-to-end slicing momentum.

This year is that focus on multi-vendor end-to-end slicing. We see a lot of interest from other operators right through from Asia, to Europe, to the US on exactly the same thing. It's something that does need to be led by operators this year, and that's what we're hoping to do.

I would hope that by the end of this year, we would have several community labs up and running with versions of end-to-end slicing, including the RAN, the transport, the core, and even potentially going into some of the verticals and IT domains while up and running with demonstrations and with multi-vendor cases this year.

Monica: You mentioned that operators need to take a bigger role, but verticals do too, because they need to say what their requirements are.

Mansoor: Absolutely, verticals are a key partner in this. The end-to-end slicing, for us, doesn't stop just at the end of the network. It does go into the verticals. We see a lot of interest from the verticals. They're in danger of not knowing where to express their requirements. That's the biggest danger.

This year is the time for action, because we have not only the momentum behind what we're doing in the labs and the End-to-End Slicing Group in TIP, we've also got a number of potential UK government-funded 5G vertical projects running this year as part of the UK government's funding for vertical trials.

We put in a number of bids on that with consortium partners. Some of the vertical players are involved in those consortiums. That's a great medium in which to bring the end-to-end slicing to life. Simply having something that's up and running on a trial case, where we can bring in the verticals and show them what slicing looks like, is worth a million specifications.

Cooperation through these trials is going to accelerate getting the verticals on board. It's very difficult for us to go to a vertical and ask them to specify what they want in terms of an SLA. They simply can't answer that on their own.

Through these trials we can do an iteration and say, "OK, now we get it. This is what we're looking for. This is the reason we're going to implement it. This improves our cost base," and iterate two or three times. That's the only way to get this done.

Monica: You mentioned TIP. What is TIP doing in this area?

Mansoor: Just before Christmas, at the TIP Summit in November, we launched the End-to-End Network Slicing Project Group within TIP. It's chaired by Andy Corston-Petrie, who is in our team, and Marie-Paule Odini's team at Hewlett-Packard Enterprise. We're chairing it. We have HP as co-chair.

We've had a huge amount of interest already, across many, many operators, including DoCoMo and several others. We're in the process of collecting all the different use cases that people would like to test out, and we're prioritizing, structuring them into a work plan for this year.

We expect that at Mobile World Congress 2018 in Barcelona we'll have a working session to agree on the prioritization for those use cases. As you know, in TIP everybody can get involved, but we will be prioritizing the use cases that have clear inputs coming in from our partners and have clear use cases that we can implement across the lab.

I would be expecting, in a way similar to the vRAN project, at least two or three community labs being set up this year, with three or four use cases each, and various partners coming out of that work group. That should be finalized in February or March.

Glossary

4G	Fourth generation
5G	Fifth generation
IT	Information technology
KPI	Key performance indicator
LTE	Long Term Evolution
MEC	Multiple-access Edge Computing
RAN	Radio access network
SLA	Service level agreement
TIP	Telecom Infra Project
uRLLC	Ultra-reliable low-latency communications

About BT

BT's purpose is to use the power of communications to make a better world. It is one of the world's leading providers of communications services and solutions, serving customers in 180 countries. Its principal activities include the provision of networked IT services globally; local, national and international telecommunications services to its customers for use at home, at work and on the move; broadband, TV and internet products and services; and converged fixed-mobile products and services.BT consists of six customer-facing lines of business: Consumer, EE, Business and Public Sector, Global Services, Wholesale and Ventures, and Openreach. For the year ended 31 March 2016, BT Group's reported revenue was £19,042m with reported profit before taxation of £3,029m. British Telecommunications plc (BT) is a wholly-owned subsidiary of BT Group plc and encompasses virtually all businesses and assets of the BT Group. BT Group plc is listed on stock exchanges in London and New York.

About Mansoor Hanif

Mansoor joined EE in November 2011 and led the technical launch of the 1st 4G network in the UK and was also accountable for the integration of the legacy 2G and 3G Orange and T-mobile networks. Until 2016 he led the team who plan, design, rollout, optimise and operate all EE radio access networks, including Mobile Backhaul and Small Cells, and was accountable for the coverage aspects of EE's Emergency Services over LTE programme. He was also a board member of MBNL (the joint venture of EE with H3G) until 2016. During the acquisition of EE by BT, Mansoor led the EE network Integration team and is currently Director for Converged Networks and Innovation in BT R&D. He is a member of the BT Technology Steering Board , Scottish Innovation Programme, Intel CommSP Technical Advisory Board and alternate board member of the Telecom Infra Project (TIP).

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, strategy, business plan support, and due diligence. Our client base is international and spans the entire value chain: clients include wireline, fixed wireless, and mobile operators, enterprises and other vertical players, vendors, system integrators, investors, regulators, and industry associations. We provide a bridge between technologies and services, helping our clients assess established and emerging technologies, use 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.

About Monica Paolini

Monica Paolini, PhD, founded Senza Fili in 2003. She is an expert in wireless technologies and has helped clients worldwide to understand technology and customer requirements, 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 she has written many reports and articles on wireless technologies and services. She has a PhD 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 contact Monica at monica.paolini@senzafiliconsulting.com