HEAVY
READING

White Paper

Fine-Grained Backhaul Monitoring for High-Quality Real-Time 4G Services

Prepared by

Patrick Donegan
Chief Analyst, Heavy Reading
www.heavyreading.com

on behalf of

VIAVI

www.viavisolutions.com

April 2015

How Today's Mobile Network Delivers – Sort Of

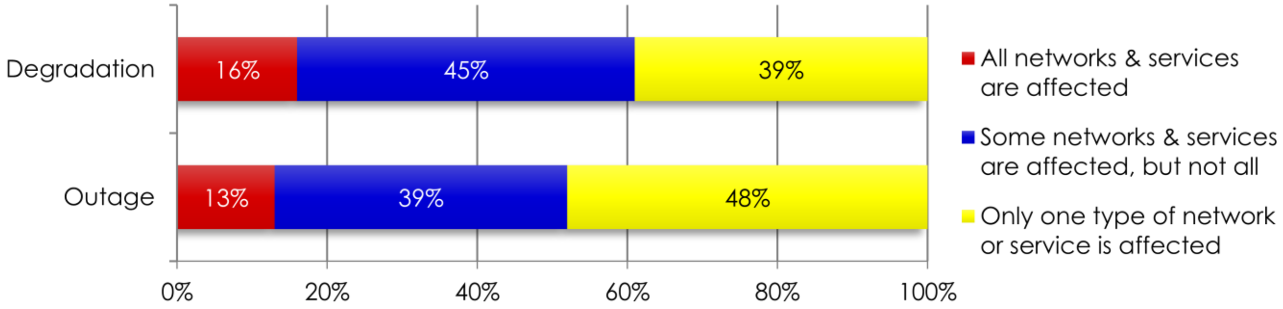
In terms of their ability to manage colossal volumes of data traffic, most of today's mobile networks are unrecognizable from when the operators first rolled out HSPA in earnest, or first started deploying LTE. The sheer capacity that now exists in the network across the radio access network (RAN), core, transport and backhaul is staggering. Equally, the sophistication that is there today – in terms of bearers, schedulers, quality-of-service (QoS) levels, traffic management and other policy controls – is an order of magnitude or two greater than five or six years ago.

In a slower-moving industry, the teams that plan, engineer and operate mobile networks would be celebrated for what mobile networks are capable of these days. Yet the focus of C-level executives within operators today isn't so much on what the network *can* do in terms of supporting huge volumes of data, but on what it still *cannot* do, and the drag that is creating relative to the demands of customers and the new goals that are being set for the business.

Why the Mobile Network Still Isn't Meeting Expectations

There are a number of reasons why the mobile network still isn't meeting expectations, including the perennial customer issues of coverage and price. Increasingly, however, operator executives as well as their high-end customers are unhappy with the quality and rapidity of new service innovation in the mobile network. They are also unhappy with the end-to-end performance of the mobile network when it comes to being able to scale to support highly delay-sensitive applications such as VoIP, Voice over LTE (VoLTE), interactive video and gaming. This is because, as shown in **Figure 1**, mobile networks suffer partial service degradations much more often than headline-making, full-scale outages.

Figure 1: Fewer Than One in Five Outages & Degradations Impact All Networks & Services



Source: Heavy Reading survey of 76 mobile operators, October 2013

From a business perspective, mobile operators have no choice but to be less and less tolerant of poor network performance. This is because the ability to drive faster delivery of new, advanced real-time applications and services that perform superbly is critical. Arguably, it is the single biggest weapon that network operators have in seeking to retain or enhance their position in the value chain for communications services, in the face of increasing competition from over-the-top (OTT) players.

The mobile network's ability (or inability) to consistently deliver delay-sensitive applications to a high standard is the focus of this white paper. Focusing on the backhaul domain, this paper explains why it is that despite the progress that has been made in mobile networks, consistently superb delivery of delay-sensitive applications remains elusive for most, if not all, mobile operators today. This paper endorses continued investment in tried and trusted techniques and standards for testing the mobile network as a basic platform for improving application-specific performance. It also examines some new ideas and techniques that have potential to augment existing testing standards with additional monitoring and reporting metrics.

The Need to Focus on the Backhaul Network

This paper focuses on testing requirements in the backhaul for good reason: While the delivery of high-quality delay-sensitive services necessarily requires a focus on end-to-end network latency, there are specific testing and monitoring issues in each discrete domain. In the case of the backhaul domain, several factors combine to make accurate testing and monitoring particularly challenging.

Factors particular to the backhaul include the fact that the network may be comprised of different technologies with different performance and latency profiles (e.g., fiber and microwave); different Layer 2 and 3 (L2/L3) protocols with different QoS settings; and different hub, spoke and even mesh architectures in aggregation and pre-aggregation layers. Not forgetting, of course, that some or all of the backhaul network may be outsourced to a third-party wholesaler, with the result that the mobile operator itself doesn't directly control that part of the network.

With mobile operators often spending billions on radio spectrum, backhaul requirements have often tended to be overlooked or taken for granted as compared with the RAN. There has certainly been a step change with the focus on adding huge backhaul capacity in the transition from TDM to IP backhaul. However, that focus on the backhaul now needs to ratchet up another gear, as operators prepare the network for mass-market deployment of VoLTE and other delay-sensitive services.

A New Need for Order in the Mobile Network

The fundamental reason why mobile networks aren't yet delivering user experience outcomes to the standard that operators and consumers increasingly want is that the demand for order in the network on the part of the operator is keeping pace with, if not being outstripped by, the supply of chaos into the network in the form of congestion, including traffic bursts, as outlined below.

Put simply, in the battle between order and chaos in the mobile network, order is doing well enough when it comes to applications such as email, Web browsing and video streaming; but when it comes to real-time, delay-sensitive applications, chaos still has the edge. Even with all of the new tools, network entities and features that have been deployed over the last few years for accurately monitoring and reporting on – and then mitigating – the chaos in the network, the operators still don't have enough to ensure that order wins out with a big enough margin to allow delay-sensitive services to be deployed at scale.

The Role of the Network Guys

It's the role of the network guys in the operator to convert the promises committed to consumers by the business into an ability to receive packets of data that are originated in the end-user device and launched into the mobile network with unique missions or QoS settings. In 4G LTE, for example, there are nine QoS Class Identifiers (QCIs) for the RAN, ranging from QCI 1 down to QCI 9: The former is for delay-sensitive voice conversation, while the latter is for non-delay-sensitive TCP-based services, such as Web browsing and email. Each QCI value has a unique profile in terms of the maximum delay and packet loss.

As they transit across the network, these QCI values also have to be mapped against the L2 or L3 QoS settings present in other domains, such as the core, transport and backhaul. As previously stated, mapping in the backhaul may not only be required with the RAN and core. They may also be required within the backhaul itself, for example across different L2 and L3 segments of the network. Put simply, the network team's job is then to ensure that each packet is able to fulfill its specific mission across these many individual domains and individual hops.

Based on the test and measurement solutions they currently use, the network team derives data on the continuous performance of the network. It then leverages that data to direct decision-making with respect to capacity planning as well as traffic management and policy control settings to ensure optimal overbooking ratios and contention points in the network. Key configuration parameters that the network team has at its disposal to ensure the performance of the network include the Committed Information Rate (CIR), QoS Queues and Committed Burst Size (CBS) buffers.

VoLTE is the Lead Latency-Sensitive Service Requiring Greater Order

With so many operators around the world now preparing to launch it, VoLTE is a particularly good example of a service that requires preferential treatment of packets, without which it will inevitably deliver a poor user experience. Operators are rolling out VoLTE because it offers faster set-up times, HD voice quality, a platform for other real-time multimedia applications, the ability to communicate with legacy phone networks as well as other VoLTE devices, as well as an opportunity to save costs by reducing dependency on legacy SS7 infrastructure.

For VoLTE, network planners need to ensure that there is enough capacity in the network, and that the prioritization accorded to each audio packet is enforced end-to-end and across each domain. This enables a stream of VoLTE packets to be transmitted with sufficiently low latency, jitter and delay variation to ensure a high-quality user experience of a VoLTE call. This can be measured empirically by the conventional Mean Opinion Score (MOS) standard, which measures the quality of received audio signals as perceived by the human ear. Once again, the backhaul is unique among all domains in that packet prioritization – hence application performance – has to be enforced across an often highly heterogeneous network environment.

Traffic Bursts & Microbursts Introduce Chaos

The operator's task of correctly and consistently enforcing order – enforcing end-to-end packet prioritization across many billions of packets – is rendered incredibly challenging by the chaos that is now being introduced into the mobile network every day – particularly as the 4G LTE network scales up.

User traffic can vary dramatically either at specific cell sites, regional clusters of sites, or across the network, reflecting responses to specific events. The type of applications that are used will also vary according to the specific event. In their network planning, many operators take account of – and plan for – these fluctuations based on historical patterns, but these assumptions about traffic patterns need to be constantly reviewed to keep pace with changing usage patterns.

Signaling storms also continue to cause substantial fluctuations in mobile network traffic. There were some very high-profile signaling storms that caused major network outages at Verizon Wireless, NTT Docomo and Telefónica O2 UK between the second half of 2011 and the first half of 2012. These were generated by a smartphone OS update and misconfigured network elements, respectively. As indicated in **Figure 1**, these high-profile, full-scale outages might be comparatively rare, but user-impacting service degradations are a lot more common.

From the perspective of network engineers, these types of traffic patterns are commonly termed "bursty." The exposure of the mobile network to bursty traffic in the transition to 3G and 4G is well known. TCP/IP traffic is a particular culprit driving this kind of chaos into the mobile network, and there's no sign of any let up. Despite some progress arising from better dialogue among operators, equipment/device vendors and application developers, increasing volumes and varieties of applications continue to be made available from app stores with widely varying signaling behaviors, often taking little or no account of their impact on the network.

Worse, leading operators are increasingly focusing on the phenomenon of so-called "microbursts." Rather than causing traffic to spike for minutes or seconds, microbursts cause traffic volumes to spike for no more than a few milliseconds or tens of milliseconds. Microbursts already have a very high profile in the financial sector. In this environment, latency requirements are much more exacting than they have traditionally been in the mobile network. Lost packets can also lose a financial institution a lot more than they do a network operator.

Nevertheless, some leading mobile operators in developed markets report concern about the greater impact microbursts can potentially have as delay-sensitive services such as VoLTE are rolled out. And they report concern that there is a risk of these incidents increasing, not just with increasing volumes and varieties of smartphone traffic generated by people, but also with the often very bursty behavior of devices deployed as part of the emerging Internet of Things.

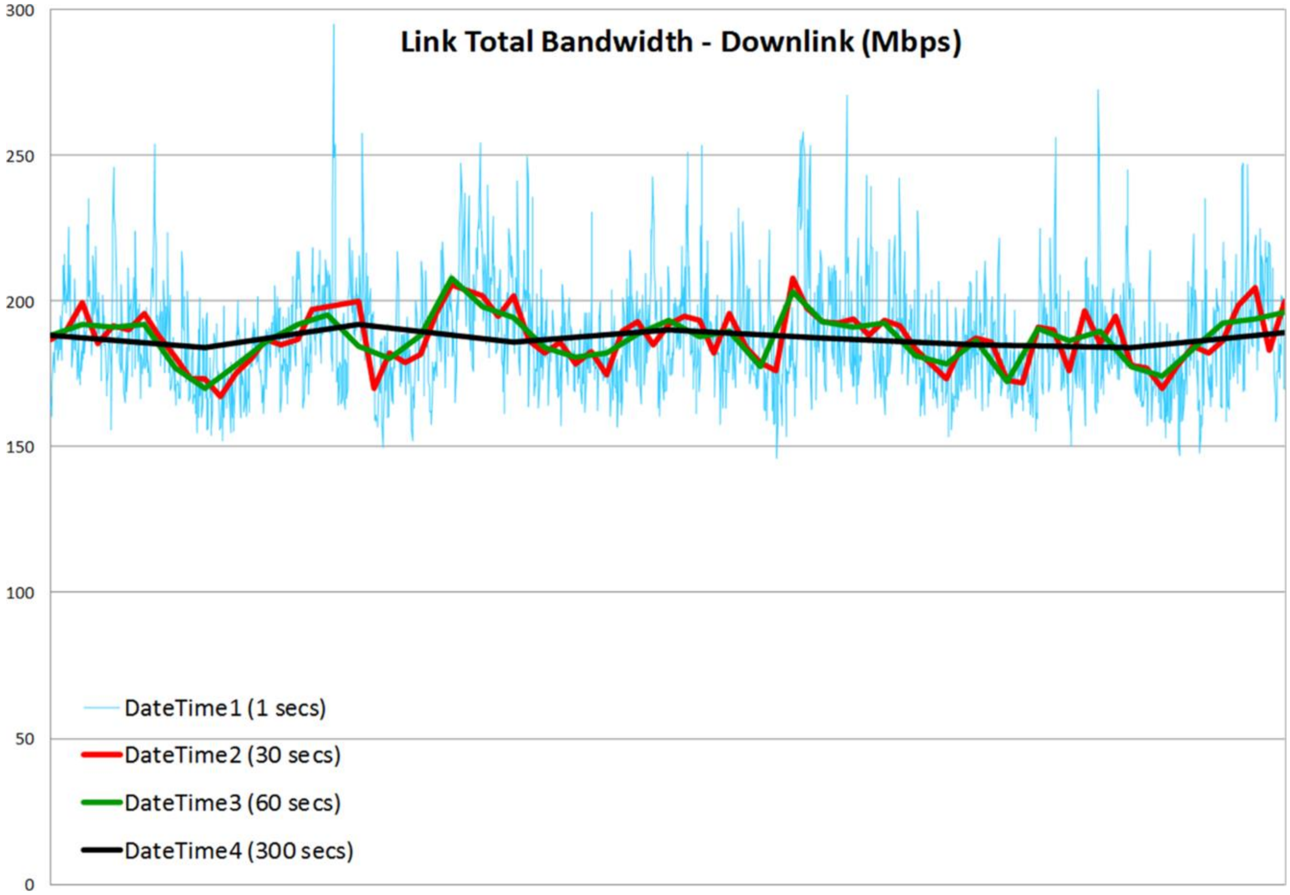
Data Sampling to Smooth Traffic Flows Across Bursts & Microbursts

Even the leading operators – or perhaps *especially* the leading operators, due to their exposure to the dynamic cutting edge of unpredictable new applications – are still grappling with formulae for modeling data traffic in the mobile broadband era. In the meantime, the data-averaging science behind today's capacity dimensioning and traffic management practices typically used in mobile operators is based on what is now outdated Erlang Theory. An Erlang is a unit of traffic measurement developed in the telephony era. It was originally adopted for use in dimensioning 1G TACS, NMT and AMPS networks and then 2G GSM and CDMA networks.

The foundation of a lot of mobile network planning and traffic management assumptions today for 3G and 4G LTE remains rooted in 2G voice. Network teams in the mobile operators today typically dimension the network and apply traffic management and policy control to cope with congestion by relying on a process of data averaging derived from data on traffic patterns from their network. This typically involves using traffic samples taken at intervals of a few minutes, half an hour, or in some cases several hours as the basis for adjusting network parameters.

The common practice of relying on data sampling to give a picture of the network every 5, 15 or 60 minutes is flawed, because smaller bursts going down to microbursts end up being averaged out in the resulting data. As a result, the real impact of these microbursts on traffic peaks ends up being obscured, as shown in **Figure 2**. This means not only that the network team doesn't know what impact these smaller bursts are having and how to fix them; even worse, it means the network team doesn't even know that these microbursts are happening at all.

Figure 2: Long Traffic Sample Intervals Disguise Smaller Bursts



Source: Viavi Solutions

The Importance of Existing Standards

All of the traditional activation testing or turn up testing that operators have been using in the LTE network for several years continue to be very important in testing the network as latency-sensitive services are rolled out. These tests ensure that before the operator first turns up a circuit to live traffic, it has been thoroughly tested to ensure that it is configured correctly from a QoS perspective relative to the application and network behavior that is expected at the outset.

Mature performance monitoring standards that measure frame delay, frame delay variation, frame loss and sometimes throughput also continue to be relevant as mobile operators prepare to launch latency-sensitive services at scale. The primary standards here are the ITU's Y1731 for performance monitoring at L2 and the IETF's Two-Way Active Measurement Protocol (TWAMP) for performance monitoring at L3. These standards consist of engineering into the traffic the periodic sending of a synthetic traffic pattern every second or millisecond across the network to a specific link. This traffic pattern has the characteristics of a specific QoS type, to ascertain whether or not the link is performing as intended.

The Limitations of Existing Standards

While existing approaches to using existing standards continue to be necessary, it is becoming increasingly clear that they are not going to be sufficient for an era of mass-market VoLTE and other real-time services. Some limitations of existing standards-based approaches are listed below.

Operators are increasingly being driven to test separately at L2 and L3

The backhaul network environment in which mobile operators conduct performance monitoring has been changing a lot because of the growing use of L3 in backhaul networks. Most of the first generation of Ethernet backhaul deployments were L2 only, based on Carrier Ethernet. However, with the acceleration of HSPA, and particularly LTE, IP/MPLS has become firmly established in many backhaul networks – certainly in the aggregation layer, but increasingly in the access layer as well. In some cases, it is also driven by operators wanting to double up and deliver L3 VPNs to enterprises from the same cell sites that host mobile backhaul equipment.

The new challenge where latency services are concerned lies in the way the different services now have to be mapped across multiple Differentiated Services Code Point (DSCP) bits for classifying traffic and providing QoS at both L3 and L2. Hence in some networks the operations teams are being driven to run performance monitoring tests generating two different synthetic traffic patterns across the network at different points in time: one with L2 traffic characteristics, the other with L3. In the quite recent past, most operators only needed the one test pattern at L2. While it is critical to test these multi-technology networks, test access points usually have access at one layer, making it challenging to run these individual tests.

The necessary one-way tests needed to support delay-sensitive services are difficult to implement using existing standards

When testing an LTE network's ability to support email and Web browsing, it doesn't matter if three quarters of the latency is experienced in one direction of a link and one quarter in the other. However, when testing for the network's ability to support VoLTE, it does – and the network team needs to have visibility of that in order to fix it. In testing terms, that means that while round-trip tests have tended to be adequate for LTE testing until now, the network team increasingly needs visibility of the performance in each direction. While most test standards do support one-way testing, the issue is that synchronizing the timing at both ends of the test has proven difficult.

Performance testing is typically done by injecting a test packet on fixed timeslots per second or per hundred milliseconds

Given the potential susceptibility of the mobile network to microbursts lasting just a few milliseconds, this risks creating a false picture of what's really going on in the network, for example if packets are injected at favorable points, when the performance of the network is looking unrepresentatively good.

Performance test traffic must be engineered into the network

Because performance monitoring methods that use synthetic traffic take up bandwidth themselves, they must be carefully engineered into the network, which has even greater complexity given the presence of events such as service-impacting microbursts. If non-visible bursty traffic exists in the network, running synthetic tests at the wrong time can end up exacerbating the problem.

The Case for Augmenting Existing Standards

In light of the traffic trends, testing issues and operator business objectives discussed throughout this paper, the attention of network planners and operations teams in some leading mobile operators is starting to shift.

Leading operators are starting to recognize that in today's increasingly heterogeneous environment it is not enough to know that there is a problem with network performance. They need to know which specific hop is causing that problem. They need to know whether it is a capacity or a configuration issue, what specific capacity or configuration issue it is, and by how much they need to adjust it.

Attention is therefore starting to shift away from traditional frame delay, frame delay variation, frame loss and throughput measurements, and toward additional capabilities that can provide a finer-grained, real-time view of network performance. This can be used to supplement the data from traditional testing methodologies and network performance data that can miss potentially service-impacting traffic events.

A real-time monitoring solution focused on viewing, understanding and responding to microbursts in data traffic promises a number of potential benefits:

- Avoid the complexity of simultaneously testing L2 and L3 traffic by monitoring using real live user traffic.
- Augment the end-to-end network performance view provided by current standards with a segmented hop by hop view of performance metrics to quickly pinpoint the source of problems.
- Provide additional reports on the delay being seen in each direction of a round trip test.
- Fill in the gaps in periodic performance measurements, which may miss microbursts, by measuring performance of real traffic measured in units of ten milliseconds or as a low as per millisecond.

Automated or Manual Responses to Performance Test Inputs

Armed with this greater granularity of information, operators would ideally like to be able to respond to it automatically, in real time. That will become possible going forward, with the advent of self-organizing networks (SON) and software-defined networks (SDN). However, there is still substantial value in the network teams having access to these more fine-grained reports to enable better decision-making today. This kind of granularity has the potential to support better decision-making in regard to capacity dimensioning and traffic management. Hence it can enable the operator to bridge the gap in network performance between what has been good enough for non-delay-sensitive services up to now, and the emerging requirement for lower-latency services at scale.

Conclusion

Real-time, delay-sensitive services such as VoLTE require a step change in the performance of LTE networks. With contention for resources already very challenging to manage, the new requirements of delay-sensitive services suggest that operators will need more granular performance data if these new services are to have any chance of scaling with a level of quality that will deliver a consistently good end-user experience.

Real-time visibility into microbursts of traffic has the potential to fill in the gaps in conventional testing standards that are only now emerging, as operators prepare to scale VoLTE and other delay-sensitive applications. Network teams in some leading operators are already showing interest in how this capability can support their performance objectives going forward.

About Viavi Solutions

Viavi (Nasdaq: VIAV) software and hardware platforms and instruments deliver unprecedented end-to-end visibility across physical, virtual and hybrid networks. Precise intelligence and actionable insight from across the network ecosystem optimizes the service experience for increased customer loyalty, greater profitability and quicker transitions to next-generation technologies. Viavi is also a leader in anti-counterfeiting solutions for currency authentication and high-value optical components and instruments for diverse government and commercial applications.