

Lethal Cocktail: Internet Traffic Offloading and Traffic Shaping Don't Mix Well

By Manish Singh, VP Product Line Management

Published in *telecomasia.net*, December 11, 2009

The introduction of flat rate \$99 all-you-can-eat data plans was, if not the most important, a very critical factor that catalyzed mobile broadband adoption. At the same time, these plans have set consumers' expectations for getting more and more bandwidth for less and less money. In fact, the decoupling between traffic growth and revenue growth is the single biggest challenge that mobile network operators face today.

To understand the magnitude of this challenge, let's look at some recent numbers. In Q2 2009 Apple sold 5.2M iPhones globally, and AT&T alone activated 2.4M iPhones. AT&T claims that each iPhone typically drives 30X more traffic than any other feature phone – yet a close look at AT&T's data pricing plan indicates that iPhone users are not even paying 3X more ARPU when compared to feature phone users (much less 30X more), yet they are driving 30X more traffic.

To meet the growing subscriber demand, operators need to add significant capacity to their networks, yet flat revenue limits their ability to fund the required infrastructure investments. This widening gap between traffic growth and associated revenues has triggered operators to look at ways to manage growth while simultaneously reducing the cost per bit of bandwidth delivered.

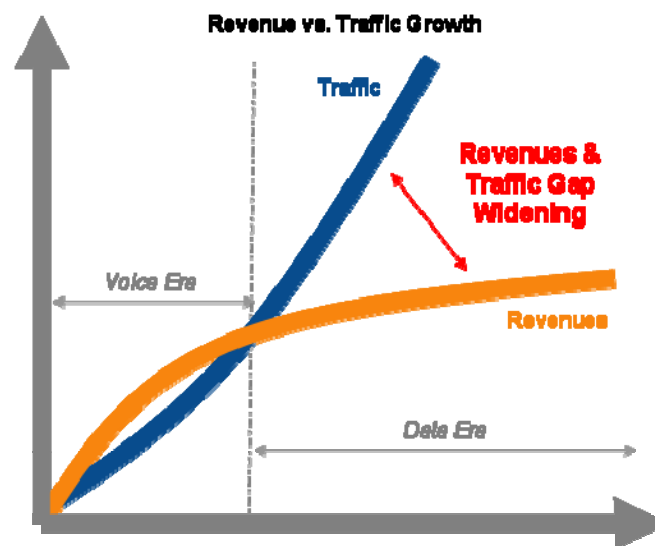


Figure 1: Mobile data traffic is increasing fast yet corresponding revenues are tapering off

Traffic Shaping

Traffic Shaping allows operators to manage traffic growth and Quality of Service (QoS) in their networks. 3GPP defines four main QoS classes for Universal Mobile Telecommunications Service (UMTS):

- Conversational – voice, video
- Streaming – webinars, online classroom
- Interactive – web browsing, file downloading
- Background – emails, database updates

The main characteristic distinguishing QoS classes is delay sensitivity – with “conversational” being very sensitive to delays and “background” being the least sensitive.

However, today’s mobile networks are carrying complex combinations of traffic and the problem is bigger than just classifying traffic into four broad buckets. With the rapid evolution of the Internet, application developers are burying data deep inside packets and things like voice mash-ups significantly increase traffic classification complexity. As a result, simple classification and prioritization no longer works.

Another problem that operators face is scale. As the sheer volume of data traffic continues to multiply, operators need to enforce policies that rate limit or, in extreme cases, block certain types of traffic that might otherwise overwhelm the network. For example, AT&T recently decided to block Slingbox traffic from iPhones on the AT&T 3G networks, indicating that this traffic would use large amounts of their 3G network’s capacity. Similarly, the rapid adoption of data cards, dongles and netbooks is quickly opening the door for bandwidth-hogging peer-to-peer (P2P) traffic to migrate to wireless networks. Since the capacity of a wireless network is inherently spectrum limited, network operators are scrambling to deploy traffic shaping functionality to rate limit said P2P traffic and ensure adequate QoS for all subscribers.

For traffic management, wireless networks are increasingly using Deep Packet Inspection (DPI) technology. Simply put, DPI looks “deep” inside packets and classifies them according to a variety of criteria – all in real time. Once classification is accomplished, different policies can be applied to the packet and its associated stream such as prioritization, rate limiting, blocking, etc. Today’s DPI gear can easily identify thousands of Internet-based protocols and traffic types while shaping traffic in real-time up to 80Gbps throughput from a single system.

Different carriers are deploying such DPI equipment differently. Some carriers deploy DPI-based traffic shapers as a standalone box beyond the GPRS Gateway Support Node (GGSN) on the Gi interface as a so-called “bump-in-the-wire”. Other operators are deploying GGSNs that have DPI-based traffic shaping functionality pre-integrated in the GGSN itself. Irrespective of the physical deployment strategy, what is relevant and important is that DPI-based traffic shaping gear sees 100% of asymmetric traffic flowing over the operator’s network in order to do its job correctly.

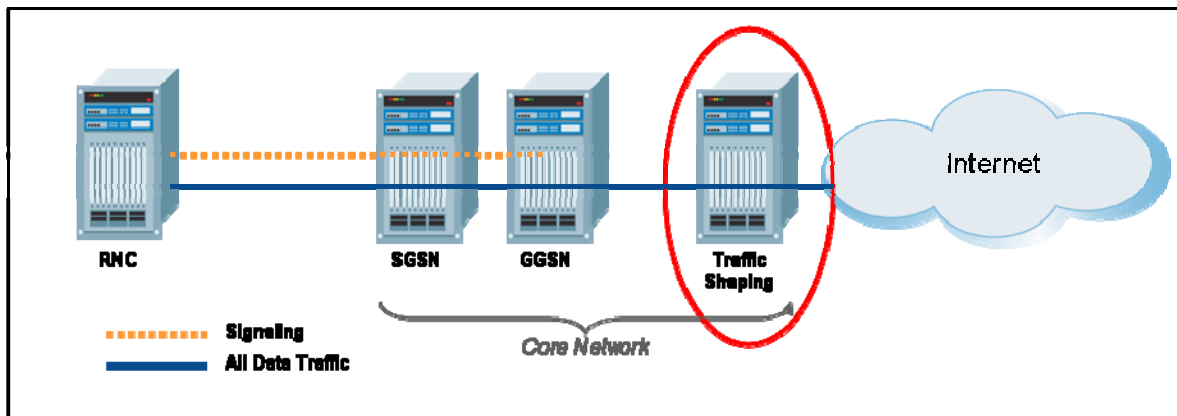


Figure 2: DPI-enabled “bump-in-the-wire” traffic shaper

Internet Traffic Offloading

The de-coupling of traffic growth and associated data revenues has suddenly peaked operators’ interest in Internet traffic offloading. The concept is straightforward: network operators want to offload all Internet traffic at the earliest possible opportunity so that don’t have to keep investing to increase capacity in the core (i.e., add more SGSNs and GGSNs).

Internet Traffic Offloading gear needs DPI capabilities, too. For example, two Real-time Transport Protocol (RTP) streams might look alike and thus a simple QoS-based UMTS qualification would not suffice in differentiating them. Let’s say an operator offering Voice over IP (VoIP) services decides that it would be wise to transport those VoIP RTP streams via its core network to guarantee QoS. That same operator might also want to offload other VoIP RTP streams (i.e., non-revenue-generating traffic from over-the-top service providers such as Skype or Vonage) to the Internet. Without DPI, the two sets of RTP streams would be indistinguishable.

There is much more complexity than what is summarized here, but for brevity’s sake it suffices to say that traffic off-loaders classify traffic streams using DPI and then, based on the operator’s policies, offload part of the traffic directly onto the Internet while sending the remaining traffic to the core network. Traffic off-loaders typically will be deployed as “bump-in-the-wire” boxes between the radio access network (RAN) and the core network (CN); in the future, some Radio Network Controllers (RNCs) might include this functionality inside the box.

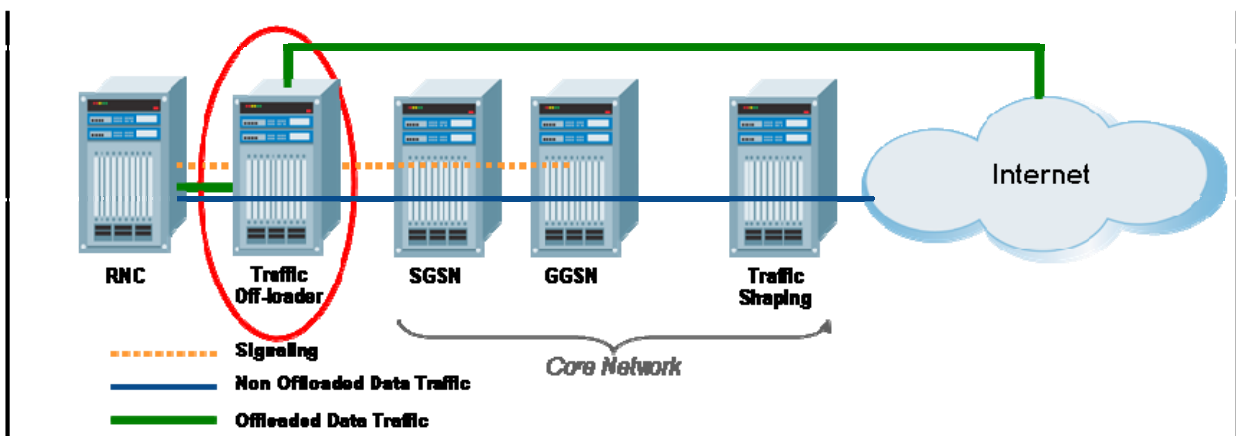


Figure 3: DPI-enabled “bump-in-the-wire” Internet Traffic Off-loader

(Potentially) Poisonous Combination

Internet traffic offloading and traffic shaping don't mix well because as soon as the off-loader is deployed in the network, the traffic shaper sees only part of the overall network traffic.

The traffic off-loader's sole job is to offload the Internet traffic so that operators don't have to invest to add more capacity in their core networks. By virtue of this very requirement the traffic off-loader has to be located either in the RAN or at the RAN / CN edge to minimize traffic volume to the core.

On the other hand, the traffic shaper's main purpose is to rate limit or block certain types of traffic in order to avoid congestion in the RAN – either on the backhaul or on the air interface. Network operators currently deploy traffic shapers in the CN and the shaper is responsible for shaping the traffic of the entire network. For shaper to be effective, all the network traffic should then flow through it.

Here's the conundrum: as soon as a network operator deploys an Internet traffic off-loader in the RAN, the traffic shaper – being un-aware of the off-loader – will start to prioritize traffic incorrectly. Conversely, the Internet traffic off-loader will break a network operator's traffic shaping policy enforcement.

Extending the AT&T Slingbox example above, the newly deployed traffic off-loader will offload Slingbox traffic to the Internet. Since this traffic is offloaded, the traffic shaper never sees it and cannot take any action, thereby completely breaking enforcement of the operator's traffic shaping policies. Similarly, for P2P traffic the traffic off-loader will shunt this traffic directly to the Internet, and since the traffic no longer flows through the traffic shaper, no one is rate-limiting it – thereby inviting P2P-caused congestion in the RAN.

The end result of introducing an Internet traffic off-loader with a currently-deployed traffic shaper is that traffic shaping will stop working properly. The bottom line is that network congestion will get worse and, worst of all, most of the congestion will move to the RAN as the CN stops throttling traffic even under high load conditions. Subscribers suffer degraded service delivery and poor QoE.

In essence, a standalone traffic off-loader coupled with a standalone traffic shaper will break a functioning network. This is bad.

The Antidote: Embedding Shaping within Off-loading

Since the traffic off-loader has to be located in the RAN, the solution is to move the traffic shaping function further up in the network and embed it *within* the Internet traffic off-loader itself. This is essential to achieve a properly functioning network. Moving the traffic shaping function in the RAN also nicely paves the way to add [Adaptive Traffic Shaping](#) functionality in the future to do RAN-aware traffic shaping.

The combined off-loader and shaper will have to be placed in the RAN or at the RAN / CN edge. By doing so, the traffic shaping function in the box can see full asymmetric traffic flowing to/from the RAN and shape it first according to the network operator's policy. After that, the traffic flow is subjected to the Internet traffic offloading function which can then correctly categorize the traffic and offload Internet traffic.

Revisiting the Slingbox example again, now with the combined shaping and off-loading functionality collocated and in the right sequence, as soon as the Slingbox traffic is passed through the shaping function it will be blocked, hence the off-loader never sees this traffic. Similarly, in case of P2P traffic, the traffic shaping function will continue to rate-limit the traffic based on network congestion and then as this traffic flows through the off-loading function it will be shunted to the Internet (which is the desired result).

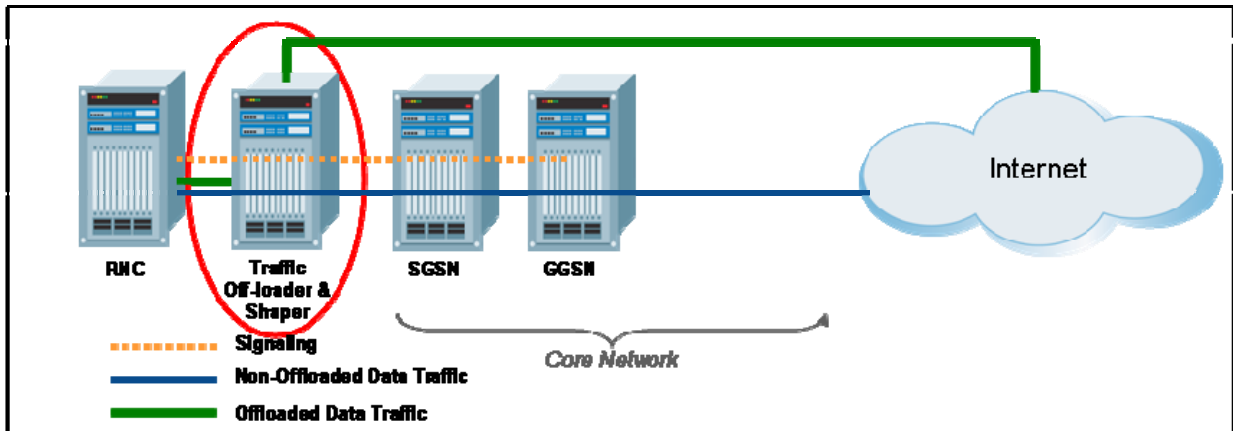


Figure 4: Internet traffic off-loader needs to do traffic shaping as well

Some of the key system requirements for Internet traffic off-loaders are:

- Stateful load balancing for GTP-based load distribution across payload blades
- DPI capabilities for complex traffic classification
- Fast path GTP and SCTP solutions to handle high data rates
- Anchoring GTP tunnels for offloaded traffic
- Support for Inter-RNC handoffs
- Re-targeting GTP packets' TEID for offloaded traffic
- High port counts to connect the off-loader to RANs, CN, and Internet highways
- Iu-PS signaling stack in passive mode, as it is deployed as bump-in-the-wire
- Scalability from 10Gbps to 80Gbps

To add traffic shaping capabilities to Internet traffic off-loaders requires additional design requirements on the system:

- Stateful load balancing to support complex traffic flows
- Higher packet processing capabilities
- DPI capabilities for traffic classification
- Prioritizing, rate-limiting and blocking flows based on traffic types and policies
- Fail-to-wire in case of system failure
- High port counts for I/O
- Support for 1G, 10G, and 40G ports
- Scalability from 10Gbps to 80Gbps

This combination of requirements needs to be delivered while achieving the lowest price/bit, of course. With its uniquely architected FlexTCA Mobile DPI platform, Continuous Computing solves these challenges using switch-based load balancing that leverages the power of FlexPacket™ ATCA-PP50 10G packet processing, FlexCore™ ATCA-FM40 10G base and fabric switching, and field-proven Trillium 3G Iu-PS protocol software to deliver the highest performing DPI-enabled system for combined traffic shaping and Internet traffic offloading.

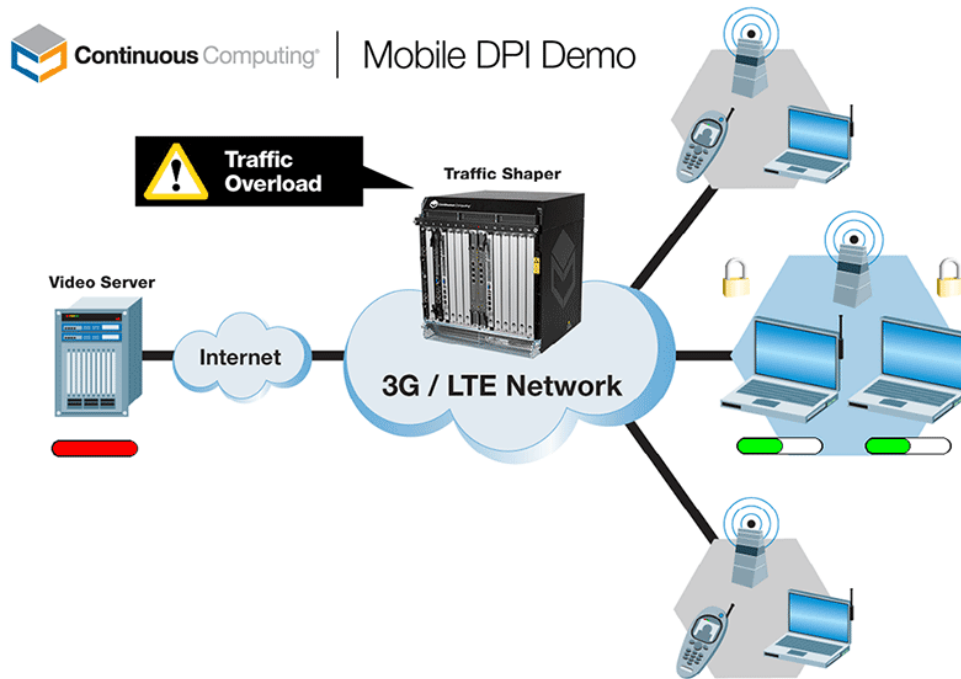


Figure 5: High-level view of Mobile DPI system from Continuous Computing

About Continuous Computing

Continuous Computing is the only company deploying uniquely architected systems comprised of telecom platforms and Trillium software. Leveraging more than 20 years of innovation, the company enables network equipment providers to rapidly deploy carrier-class LTE, DPI, and femtocell applications with reduced risk, cost, and complexity. Only Continuous Computing combines open-standards systems, Trillium protocol software, and expert professional services to create fully-integrated solutions that empower more than 150 customers worldwide to accelerate new product delivery and maximize return on investment. www.ccpu.com.

Continuous Computing is an active member of 3GPP, CP-TA, ETSI, Femto Forum, Intel ECA, and the SCOPE Alliance.

Continuous Computing, the Continuous Computing logo, and Trillium are trademarks or registered trademarks of Continuous Computing Corporation. Other names and brands may be claimed as the property of others.