Network Congestion Management: Considerations and Techniques

An Industry Whitepaper

Contents

Executive Summary

Network congestion occurs when demand for a resource exceeds that resource's capacity. Congestion management equates to achieving cost-savings while preserving subscriber QoE. These dual objectives are often contradictory, and a balance must be struck.

Many factors must be considered when evaluating solutions, including the regulatory environment. Generally, to be compatible with the spirit of network neutrality, a congestion management solution:

- must be narrowly tailored
- must have a proportional and reasonable effect
- must serve a legitimate technical need

A complete congestion management solution has two functional components:

- A mechanism to manage the impact of congestion
- A mechanism to trigger the management policies

Only by linking an accurate trigger with precise management can a CSP ensure that the goals of congestion management are achieved.

Real-time measurements of access round trip time provide the best possible congestion detection mechanism by informing operators precisely where, when, and for whom congestion is manifesting. Other approaches result in misapplication of management policies, in violation of best practices.

Since congestion is a result of overwhelming demand, it is up to the CSP to define:

- What applications or application types will be prioritized during congestion?
- Will congestion management policies take into account subscriber-related attributes?

Version 2.0

sandvine
Intelligent Broadband Networks

Introduction to Access Network Congestion

Network congestion is defined as the situation in which an increase in data transmissions results in a proportionately smaller (or even a reduction in) throughput. In other words, when a network is congested, the more data one tries to send, the less data is actually successfully sent.

Network congestion has been difficult to define quantitatively across the industry, but it's something everyone recognizes when they see it. For subscribers, it means reduced quality of experience (QoE) with video stalls, choppy VoIP communications, a poor web browsing experience and frustrating online gaming performance. For communications service providers (CSPs) it means angry subscribers that are more likely to churn and infrastructure extensions that represent huge capital investments.

The underlying problem is that all access network resources have a finite capacity, and demand can exceed that capacity.

The inconvenience of congestion poses real business threats to CSPs. When a network becomes congested, the operator can expect numerous support calls or high subscriber attrition. By the time the operator is able to respond (often via increased capacity) the damage has already been done, with word-of-mouth compounding the problem.

Throughput and Latency

When subscribers are using the internet, they are readily (if not consciously) aware of two things: throughput (the capacity of the connection, measured in bandwidth terms like megabits per second) and latency (the time it takes traffic to traverse the connection, typically measured in milliseconds).

Throughput is about how much data subscribers can draw from the network over time and is most apparent when transferring large files (e.g., software updates, email attachments, peer-to-peer exchanges, etc.).

Latency refers to how long it takes for a packet to travel from a server to a client, and is most apparent when interactive applications are in use (e.g., web browsing, streaming, gaming, voice, video, etc.). In the example of a VoIP call, high latency introduces unacceptable delay in the conversation. Anyone who has played a latency-sensitive online game doesn't need a study to tell them that a 200ms ping will render the game unplayable and most servers un-joinable. Because online gaming can include paid subscription to a service beyond paying for data access, in such cases subscribers cannot help but feel doubly cheated.

Google and Amazon have both independently released research results showing that latency on their websites has a direct impact on sales. Google, upon introducing an added 0.5 seconds on page loads to deliver 30 results instead of 10, found revenue dropped 20% as a result[1]. Similarly, Amazon conducted an experiment that showed with every 100ms increase in latency, sales dropped by 1%. It's clear that subscribers are affected by latency, and that it doesn't have to be particularly high to have a heavy impact on subscriber QoE.

[1] Interested readers can find the original research in *User Preference and Search Engine Latency*

Congestive Collapse

As throughput increases on a node or router, latency increases due to the growing queue delay[2] and the 'bursty' nature of TCP.[3] At first, the increase in latency is rather marginal, but nevertheless is proportional to the increase in bandwidth. However, as the throughput approaches capacity, latency begins to increase exponentially as it reaches a final tipping point where the element experiences congestive collapse.

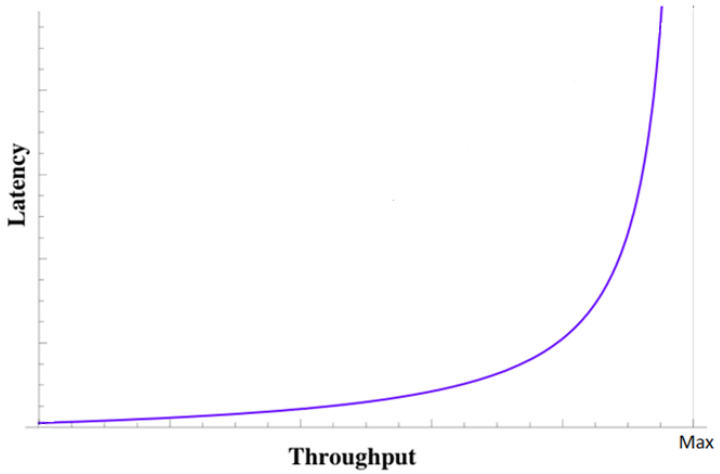

Figure 1 - Relationship between throughput and latency; the exponential increase in latency represents congestive collapse

The tipping point often occurs somewhat before the resource's theoretical maximum bandwidth threshold is actually reached. Plus, since both fixed and access network resources experience dynamic capacity (this phenomenon is explained in greater detail in a later section), there is no fixed point at which the congestive collapse acceleration begins – the point varies every moment.

For subscribers, when an access node is near capacity, the rapid increase in latency causes a substantial deterioration of QoE for latency-sensitive applications.

Congestion Management

At a high level, a congestion management solution has two functional components:

- A mechanism to alleviate the impact of congestion; the actual 'management'
- A mechanism to switch on the alleviation mechanism; the detection 'trigger'

A complete congestion management solution requires both these components, but this simple characterization betrays enormous complexity; available solutions vary enormously in effectiveness and complexity.

[2] In this case, the queue delay is the time a packet waits in a queue until it can be processed. As the queue begins to fill up due to packets arriving faster than they can be processed, the amount of delay a packet experiences increases. The problem is particularly pronounced when TCP is in its 'ramp-up' stage. More information is available here: http://en.wikipedia.org/wiki/Queuing_delay

[3] This study looks at the phenomena of TCP burstiness: http://www.caida.org/publications/papers/2005/pam-tcpdynamics/pam-tcpdynamics.pdf

Considerations and Techniques

Considerations

There are many factors that must be considered and understood to critically evaluate congestion management systems, including:

- Defining the goal: what is a CSP trying to achieve?
- Network neutrality: does the solution fit within a particular regulatory environment?
- Topology awareness: does the solution rely upon a complete understanding of the network's structure?

Defining the Goal

Congestion management is a general solution that actually addresses several potential objectives, from the perspective of the CSP:

- Defer capital investments by extending infrastructure lifetime: congestion management can provide short term relief that buys time before additional capacity can be physically added to the network
- Extend the lifetime and utility of legacy network equipment that will not receive expansion: some earlier generation access technologies will not receive additional investment, but are still serving many subscribers
- Reduce or contain transit costs: a congestion management solution can be applied to cap peaks on expensive links or get the most utility out of a saturated link
- Enhance or protect the subscriber quality of experience: make sure that even during times of congestion, the subscriber base is happy; put another way, this goal could be stated as "to maximize the quality of experience, for the maximum number of subscribers, for the maximum amount of time"
- Implement part of a fair use or fair access policy: many fair use policies include provisions about traffic management to protect the integrity of the network during busy periods

Cost Savings and Subscriber QoE: Finding the Right Balance

In the end, congestion management comes down to seeking the optimal balance between achieving desired cost-savings while preserving subscriber QoE during times of congestion. Doing nothing (i.e., making no investment in congestion management or additional capacity) will provide the greatest cost-savings at the complete expense of the subscriber experience, whereas seeking to satisfy any and all demand at all times is not financially viable.

Regardless of the particular motivation, the dual components (i.e., detection and management) are both present, but the factors that must be considered while evaluating solutions can vary. For instance, a fair use policy could imply the requirement of subscriber awareness, while transit costs can be addressed without this knowledge.

Therefore, it is important that a CSP consciously decide what the objectives and use cases are, both at the present and in the future, so that potential solutions can be evaluated in an educated context.

Network Neutrality

The term "network neutrality" in this context breaks down to mean "what a CSP is permitted (whether by explicit regulation or suggested guidelines) to do when managing traffic."

Ultimately, a key requirement of any solution should be having sufficient flexibility, versatility, and configurability in all components to be adapted to fit a particular regulatory environment (both now and in the foreseeable future).

Although formal guidelines for network neutrality are rare, some best practices have nevertheless emerged. Generally, to be considered compatible with the spirit of network neutrality, a congestion management solution:

- must be narrowly tailored (i.e., accurate in detection and enforcement, and only apply when and where congestion manifests)
- must have a proportional and reasonable effect (i.e., it has to be fair and justifiable)
- must serve a legitimate and demonstrable technical need (i.e., target actual network congestion, and not simply seek to reduce overall bandwidth consumption)

There is obviously much room for variation and interpretation within those best practices, and careful consideration must be given to them. The first two in particular provide powerful criteria by which market alternatives can be examined. For instance, the best practice of being narrowly tailored can imply a long list of requirements, including: subscriber awareness, to precisely choose which subscribers are impacted; application awareness, to precisely manage only latency-insensitive applications or application types; location and topology awareness, to only manage traffic on a congested link; real-time congestion detection, rather than a rudimentary time-based system; etc.

Topology Awareness

Both to have the most precise determination of where congestion is within a network, and to apply congestion management as precisely as possible, a solution must have complete knowledge of a network's topology.

Congestion Detection

It seems almost too obvious to state that congestion management should only be applied when there is actual congestion on the network (suggesting a requirement for real-time detection), particularly as it relates to the best practice of being narrowly tailored. However, it bears detailed examination because there are many solutions available that actually apply traffic management for the supposed purpose of congestion management when there is no congestion.

Time of Day

The fact that time of day isn't technically even a detection mechanism (i.e., it's not detecting anything on the network) should be sufficient to dismiss it from consideration or at least cast a healthy shadow of skepticism, but many vendors rely on this approach because it is trivial to implement and seems plausible to those unfamiliar with the complex realities of network congestion.

The premise is simple: the network is busier at certain times than others, so managing traffic during those times should prevent congestion. However, this generalization hides the fact that congestion of a particular link is unpredictable with any precision: while the network as a whole may be carrying more traffic, the utilization of individual links varies enormously and dynamically. The reality is that there is little or no correlation between time of day and network congestion (at a precise link level), as shown by multiple studies on the subject.[4]

[4] For further reading and references to studies, refer to TechCrunch and Fierce Wireless.

The fundamental flaw of the time of day trigger is that it assumes congestion is occurring without actually confirming the case. As a result, management policies are implemented without technical justification (potentially in violation of best practices), are certainly not narrowly tailored, and are arguably out of proportion (to the non-existent congestion).

Concurrent User Thresholds

This approach counts the number of users concurrently on a resource or link, and triggers congestion management when a particular threshold has been breached.

As opposed to time of day triggers, measuring concurrent users at least relies upon measuring *something*, but it makes an assumption that the network is congested without actually verifying the case. In fact, the correlation between the number of users and the presence of congestion is extremely weak, and certainly so weak as to be useless (or at the least, dangerous) as to serve a proxy for the presence of congestion. As such, it violates the important principle of only managing traffic when congestion is verifiably occurring.

Additionally, as will be seen to be the case with bandwidth thresholds, establishing concurrent user thresholds has severe drawbacks in terms of operational requirements that are seemingly at odds with the otherwise simplistic approach. In particular, every single resource or link has to have a threshold defined (which may or may not require some sort of analysis to determine an 'appropriate' threshold[5]); these thresholds must be reexamined whenever the network capacity or topology changes.

Bandwidth Thresholds

In this approach, congestion is triggered when particular levels of bandwidth are exceeded. The general premise applies to solutions functioning at a network level (e.g., "apply management when network bandwidth exceeds 800 Gbps") and at a link level (e.g., "apply management when traffic on this link exceeds 72 Mbps"). In the latter case, individual thresholds must be set for every link to be managed – a fairly onerous operations task. In both cases, the bandwidth threshold is set just below the point at which congestion collapse begins.

Like a measurement of concurrent users, bandwidth thresholds have in their favor reliance upon actual network behavior, but, similarly to the two approaches examined previously, this one suffers from some flawed assumptions.

First, there is an assumption that bandwidth automatically causes a degradation in quality of experience. While overwhelming bandwidth does have a high correlation with lowering QoE, it is entirely possible for a link or resource to be at maximum capacity without causing subscribers to suffer lowered QoE.

Most critically, though, there is an underlying assumption that link/resource capacity has a fixed maximum, but that is not the case either for mobile access networks or for fixed access networks. As a practical result, it is impossible to pick a threshold that is guaranteed to be below the point at which congestive collapse begins[6].

[5] Which begs the question: if one can measure the presence of congestion via some other means (so as to then determine how many subscribers are online at that point, and subsequently use that number as the threshold), then why not use the actual presence of congestion as the trigger?

[6] Well, one could pick something trivially low, but that would cause congestion management to be triggered for no good reason, violating the best practices.

Dynamic Capacity in Mobile Access Networks

In mobile networks, the point at which a cell begins to dive rapidly into congestive collapse always varies over time because the capacity of the resource changes based on variable external conditions (e.g., cloud cover for a satellite network, and general interference and signal variability in mobile networks).

This reality makes configuring static bandwidth thresholds for mobile network resources a best-guess proposition that cannot accurately determine when the resource is actually congested, and will inevitably lead both to unnecessary management (when capacity is high) and lack of management when it is required (when capacity is low).

Dynamic Capacity in Fixed Access Networks

It comes as a surprise to many, even seasoned industry veterans, that capacity of fixed access network resources is dynamic, but it is true nonetheless and warrants explanation.

Consider the example shown by Figure 2, which demonstrates the practice of daisy-chaining DSLAMs to provide coverage for rural areas. In this case, setting a throughput threshold for the subtended DSLAMs offers no benefit in terms of congestion detection since the maximum capacity available is determined by the sum of subscriber demand linked back to the master DSLAMs. It is possible to set a bandwidth shaper that artificially reduces the maximum available throughput for the subtended DSLAMs far below its actual maximum. The problem with this approach is that it results in an access resource that cannot be used to its full potential. Furthermore, any change in network topology requires the operator to manually reconfigure the shapers for every affected node[7].

To enforce properly in this scenario, all traffic must be correctly mapped to all resources/DSLAMs/links the traffic traverses. When resources are shared, and/or traffic goes through multiple DSLAMs, assigning traffic to the correct "chain" of resources is very challenging, if not impossible, especially when the topology is changing.

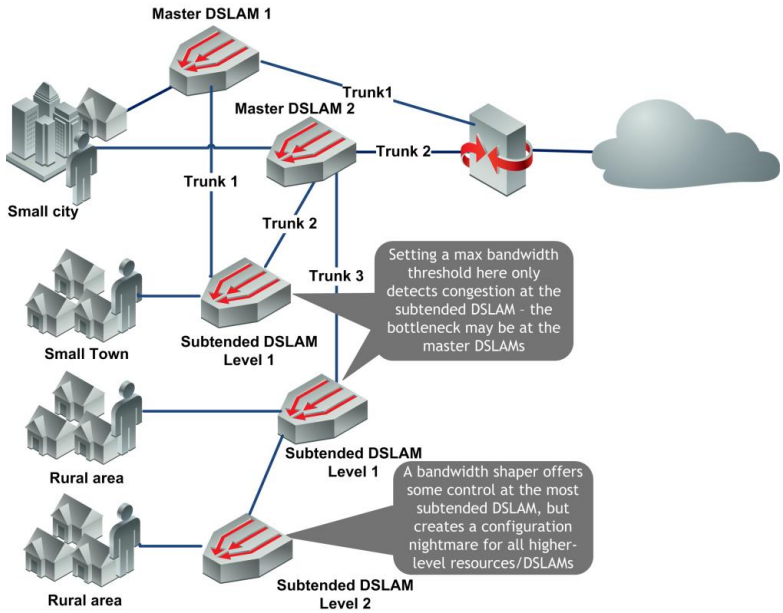

Figure 2 - Effect of DSLAM daisy-chaining on congestion detection and enforcement

[7] Another general strike against fixed bandwidth thresholds

Examining Figure 2 more closely, we see a clear example of overlapping trunks extending from Master DSLAM 2 to simultaneously serve two different subtended DSLAMs, and we can also see that two different Master DSLAMs are linking to one subtended DSLAM, both very common in practice. No matter how patient and willing an operator may be, it is simply not possible to configure static bandwidth thresholds in this case. Traffic demand on the overlapping links will vary constantly.

The example above is actually very simple – for DSL networks serving rural areas the chain can be quite long and complex. It is simply not realistic to expect a service provider to expend effort on determining the best shaping value for thousands of DSLAMs every time the topology changes.

The same concepts and challenges can be extrapolated for cable networks. The DSL scenario depicted above is similar to "nested channel bonding" in cable, where traffic may be assigned to an individual channel or bonded group, where bonded groups may be nested, and where traffic must be assigned to the correct "chain" of bonded groups/channels to which it contributes. The difference with cable is that the complex channel/group hierarchy is internal to the CMTS and the topology can be fetched using SNMP. In DSL networks the hierarchy is not easily retrieved anywhere – it must be provisioned.

Finally, the concept of "overlapping bonded groups" in cable networks is similar to the overlapping trunks shown by Figure 2 in that the overlapping bonding in cable obfuscates the "chain" of bonded groups to which traffic is assigned. Static bandwidth thresholds simply cannot be relied upon in such dynamic environments.

Subscriber Quality of Experience

The best way to trigger congestion management is to directly measure congestion itself (as opposed to conditions assumed to exist in correlation). However, as mentioned in the introduction, there is currently no standardized metric to measure detection. The next best thing is to measure a metric that is proven to have direct correlation with network congestion – ideally a metric that is caused by the congestion, rather than vice versa.

Access round trip time (aRTT) is such a metric. Not only is aRTT known to increase dramatically when congestion is present, but it also has high correlation with subscriber assessments of quality of experience.

The access round trip time indicator measures the time from when a packet enters the operator's access network to when the response packet leaves the access network via the same point[8]. Specifically, aRTT is the measure of time between the SYN-ACK and ACK packets (time between T1 and T2 in Figure 3) on a TCP flow when the subscriber is the client of a TCP connection. The smaller the aRTT value, the better the quality of the link; the larger the aRTT value, the worse the quality of the link.

These measurements can be taken for every TCP connection and rolled up on a per-subscriber and per-access network resource basis, to yield incredible precision about when, where, and for whom congestion is manifesting.

[8] The Internet round trip time (iRTT) is similarly defined, but on the opposite 'side' of the network, and measures the latency of things leaving the network and the responses returning.

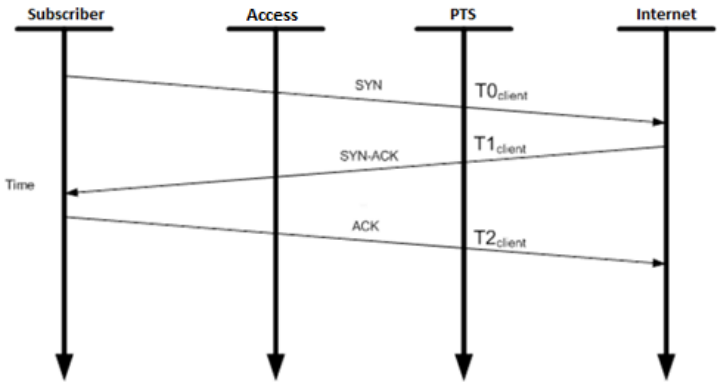

Figure 3 - Visualization of access round trip time (aRTT) calculation

The access network is where operators have the most influence over latency; for this reason, when measuring per-resource subscriber QoE for congestion management it is important to isolate aRTT from total round trip time (RTT).

It must be noted that this calculation is not a direct measure of latency; in fact it is the sum of the following times:

- Downstream latency
- Client device processing time
- Upstream latency
- Retransmissions

Furthermore, it is a fact that unknown conditions such as weather (for mobile networks) or inter-NAT routing can impact measurements for retransmission and client processing. However, a healthy statistical distribution of aRTT sampling measurements across many subscribers and flows will cancel out the skew introduced by these variables, making aRTT a very good measure of access network latency.

Latency is the key value affecting QoE for the majority of subscribers, and therefore is the most accurate measure of when congestion is actually occurring.

Like some of the approaches discussed previously, aRTT only works as a trigger when the measured values are compared against a threshold, and this comes at an operational cost of determining and configuring these thresholds. However, doing so for aRTT is no more onerous than for bandwidth or concurrent users, and since aRTT is directly tied to congestion it is actually possible to automate this process for each and every link on a network (provided the solution has topology awareness).

When considering each of the congestion detection mechanism against the congestion management best practices, one reaches the conclusion that real-time measurements of access round trip time provide the best possible congestion detection mechanism. When implemented in real-time, on a per-link basis, such an approach tells an operator precisely where, when, and for whom congestion is manifesting.

Congestion Management

Once congestion has been detected, congestion management seeks to control the situation. Since congestion is a result of overwhelming demand, the congestion itself cannot technically be managed

away[9]. Instead, CSPs are actually defining how the congestion manifests on the network. Left alone, greedy protocols[10] will compete in a free for all, regardless of relative tolerance for latency, importance to subscriber, criticality to services, etc.

It is up to the CSP, then, to determine and define:

- What applications or application types will be prioritized during congestion?
- Will congestion management policies take into account subscriber-related attributes, like recent usage and service plan?

In practice, the management policy will take into account a combination of factors.

Defining Application Priorities[11]

There are several ways to define relative application priority and value, but the one most-aligned with the many goals of congestion management is to define priority based on tolerance for latency.

Some applications are more tolerant than others regarding latency – that is, they continue to deliver a high QoE even when latency is present - and this gives a convenient framework by which to define what gets prioritized access to limited network resources during times of congestion. For example:

- Low Tolerance: Voice-over-IP, Gaming, Web Browsing – essentially any Internet activity where even the slightest increase in latency results in a significantly degraded online experience.
- Medium Tolerance: Video streaming – here, increased latency can have an impact, but when managed carefully and intelligently (e.g., by taking into account the video characteristics like delivery mechanism, resolution, etc.) some increase in latency can be accommodated while preserving the overall online QoE.
- High Tolerance: Peer-to-Peer and bulk downloading (e.g., software updates, email attachments, FTP transfers, newsgroups) – here the impact of latency has the least effect on subscriber experience, since the intention is to draw data from the network and experience the benefit later.

By defining these rankings of application tolerance, CSPs can decide what traffic is subjected to congestion management policies initially (i.e., the high tolerance categories), then what gets managed next (i.e., the medium tolerance categories). Realistically, management of the low tolerance categories should be avoided unless the integrity of the network is in danger.

Preventing Starvation

One issue that must be addressed is how far to take congestion management of particular 'things' (e.g., applications, application categories, subscribers, etc.). No CSP wants to do more damage to QoE than the congestion itself would do, and one important mechanism to ensure QoE is acceptable for even managed traffic is to prevent starvation – that is, to manage to a certain minimum amount and then to go no further.

[9] It is prudent to note that 'peak-shifting' business practices, like allowing unlimited usage in overnight hours actually can reduce network demand and will help to prevent congestion, but in practice few subscribers take CSPs up on these offers

[10] For instance, protocols that burst up or that open many concurrent connections

[11] In regards to network neutrality, some jurisdictions forbid any incorporation of application identity whatsoever, others allow management of application categories but not individual applications, and others permit anything

Minimizing Negative Subscriber Impact

In areas where regulations forbid congestion management to incorporate application identity, the entire congestion management solution might be based upon managing particular subscribers; in other areas, choosing particular subscribers is just a small part of a multi-factor solution.

In either case, since congestion management seeks to achieve a perfect balance between maximizing an access resource's lifetime and maximizing QoE for the greatest number of subscribers, the typical goal is to impact the fewest subscribers possible during a management period. Alternatively, the goal could be to impact subscribers 'fairly', which of course is subjective.

This approach reflects the reality that subscribers are often both contributors to and victims of congestion, with their precise role changing very suddenly and randomly.

To deliver the maximum positive impact to the network's capacity to deliver service while negatively impacting (i.e., managing) the fewest subscribers, the optimal strategy is to target the largest contributors to congestion; this translates into managing the traffic of those subscribers who are using a disproportionate amount of bandwidth at the time an access resource is congested.

The most effective predictor of congestion contribution is not long term usage[12], but very short term usage data. Studies on short-term subscriber usage show that during 15-minute periods of congestion, 1-5% of subscribers use up to 80% of an access resource's bandwidth. Using the short-term usage history of users (15 minutes) targets the true contributors to congestion for a fair, proportionate, and application-agnostic solution (and it can also be combined with policies that are aware of application category).

To be effective in the face of dynamic demand, this "short-term heavy user" category of subscribers must be updated as the clock advances, so that management is always directed towards those who are most likely to be contributing to congestion *right now*.

Maximizing Precision

As stated previously, congestion management will be most effective (i.e., will have the largest positive impact while avoiding managing areas where congestion is not appearing) when the solution is completely topology-aware.

In combination with application and application category criteria, and precise selection of subscribers, topology awareness ensures that the highest positive impact is gained at the cost of impacting the fewest number of subscribers for the shortest amount of time.

In mobile networks, however, there is one additional nuance that must be addressed.

Mobility Awareness in Mobile Networks

A congestion management solution must have visibility of which subscribers are attached to which access network resources at any particular point in time. Unlike fixed networks, where subscribers stay on the same access segment until they move residence or a service provider makes changes, mobile data networks are defined by subscribers who travel from one network access segment to another while accessing data. Subscriber mobility awareness is needed to maintain the correct association between a mobile subscriber and the access segment to which they are attached. Without it, policy

[12] That's why solutions that target the month, week, or day's heavy users have trivial impact on congestion, achieved at the expense of managing many subscribers

control consists of blind swings and best guesses with unverifiable impact and questionable regulatory compliance.

A very recent study[13] measured the movement of over 20,000 mobile data subscribers in a major North American city. Over the course of one hour, just over 20 percent of users were still engaged in the same data session but had moved to a second cell sector. Another study looking at a tier-1 network in Asia showed mobility of about 27% within the same data session every minute.[14] No two mobile sectors have an identical congestion profile at any moment in time. Without subscriber mobility awareness, the solution "loses" visibility of 20-30% of potential congestion contributors and applies policy to the wrong subscribers and resources.

Policy Enforcement

The congestion management policies themselves take many forms (e.g., prioritization/de-prioritization, shaping and rate-limiting, weighted fair queues, etc.) and can also be enforced by multiple devices working towards the same goal.

For instance, an intelligent inline device can perform traffic classification activities and make the decisions about what traffic should get managed, but leaves the management part for other inline devices. Traffic is simply marked as it flows past, and the management devices read these marks.

Alternatively, instead of marking traffic, the intelligence device can use a PCRF to communicate to other enforcement devices.

Additionally, different traffic can be managed by different devices: in one popular approach, an inline PCEF applies downstream congestion management, but signals to the gateway for upstream congestion management.[15]

The point is that there are a range of strategies that can be part of an effective congestion management solution, and informed CSPs can work with a vendor to determine the best approach for a particular network.

[13] Sandvine internal study examining subscriber movement patterns for a tier-1 mobile service provider in North America over the course of 1 hour.
[14] Sandvine internal study examining subscriber movement patterns during the same data session for a tier-1 mobile service provider in Asia every minute.
[15] This approach is described in detail here: http://tools.ietf.org/html/rfc6057

Conclusions

There are many factors that must be considered and understood to critically evaluate congestion management systems, including:

- Defining the goal: what is a CSP trying to achieve?
- Network neutrality: does the solution fit within a particular regulatory environment?
- Topology awareness: does the solution rely upon a complete understanding of the network's structure?

In the end, congestion management comes down to seeking the optimal balance between achieving desired cost-savings while preserving subscriber QoE during times of congestion.

At a high level, a congestion management solution has two functional components:

- A mechanism to alleviate the impact of congestion; the actual 'management'
- A mechanism to switch on the alleviation mechanism; the detection 'trigger'

These components work together to ensure that the goals of the congestion management solution are achieved in accordance with the CSP's operating and regulatory parameters.

However, the accuracy of detection mechanisms varies widely, and many factors determine how effective the management policies themselves are in alleviating the congestion. Only by linking an accurate trigger with a precise management policy can a CSP ensure that the goals of congestion management – whatever they are – are achieved.

Summary of Detection Techniques

Congestion management should only be applied when there is actual congestion on the network. Consequently, an effective congestion management solution relies upon an accurate detection mechanism to trigger the actual management policies.

Detection Technique	Accuracy	Explanation
Time of Day	Extremely Low	The fundamental flaw of the time of day trigger is that it assumes congestion is occurring without actually confirming the case. As a result, management policies are implemented without technical justification (potentially in violation of best practices), are certainly not narrowly tailored, and are arguably out of proportion (to the non-existent congestion).
Concurrent User Thresholds	Low	Measuring concurrent users at least relies upon measuring *something*, but it makes an assumption that the network is congested without actually verifying the case. In fact, the correlation between the number of users and the presence of congestion is extremely weak.
Bandwidth Thresholds	Medium	There is an assumption that bandwidth automatically causes a degradation in quality of experience. While overwhelming bandwidth does have a high correlation with lowering QoE, it is entirely possible for a link or resource to be at or close to maximum capacity without causing subscribers to suffer lowered QoE. Critically, though, there is an underlying assumption that link/resource capacity has a fixed maximum, but that is

		not the case either for mobile access networks or for fixed access networks. As a practical result, it is impossible to pick a threshold that is guaranteed to be below the point at which congestive collapse begins.
Subscriber Quality of Experience	Extremely High	The theoretically best way to trigger congestion management is with a direct measurement of congestion itself. However, there is currently no standardized metric to measure detection. The next best thing is to measure a metric that is proven to have direct correlation with network congestion – ideally a metric that is caused by the congestion, rather than vice versa. Access round trip time (aRTT) is such a metric. Not only is aRTT known to increase dramatically when congestion is present, but it also has high correlation with subscriber assessments of quality of experience. Today, access round trip time provides the best possible congestion detection mechanism. When implemented in real-time, on a per-link basis, such an approach tells an operator precisely where, when, and for whom congestion is manifesting.

Summary of Management Considerations

Since congestion is a result of overwhelming demand, the congestion itself cannot technically be managed away. Instead, CSPs are actually defining how the congestion manifests on the network.

It is up to the CSP, then, to determine and define the factors that are taken into account by a management policy.

Management Consideration	Explanation
Defining Application Priorities	Some applications are more tolerant than others regarding latency – that is, they continue to deliver a high QoE even when latency is present - and this gives a convenient framework by which to define what gets prioritized access to limited network resources during times of congestion. To prevent application starvation (and the negative impact to QoE that result) the solution should be support a minimum rate capability.
Minimizing Negative Subscriber Impact	Since congestion management seeks to achieve a perfect balance between maximizing an access resource's lifetime and maximizing QoE for the greatest number of subscribers, the goal here is (typically) to impact the fewest subscribers possible during a management period. The most effective predictor of congestion contribution is very short term usage data. Using the short-term usage history of users targets the true contributors to congestion for a fair, proportionate, and application-agnostic (where desired or required) solution. To be effective in the face of dynamic demand, this "short-term heavy user" category of subscribers must be updated as the clock advances.
Maximizing Precision	In combination with application and application category criteria, and precise selection of subscribers, topology awareness ensures that the highest positive impact is gained at the cost of impacting the fewest

	number of subscribers for the shortest amount of time. In mobile networks, the solution must be subscriber-mobility aware; that is, it must have a real-time (or near real-time) knowledge of subscriber location, otherwise management policies will be needlessly applied and will have little positive effect.
Policy Enforcement	The congestion management policies themselves take many forms (e.g., prioritization/de-prioritization, shaping and rate-limiting, weighted fair queues, etc.) and can also be enforced by multiple devices working towards the same goal. There are a range of strategies that can be part of an effective congestion management solution, and informed CSPs can work with a vendor to determine the best approach for a particular network.

Related Resources

In addition to the resources footnoted throughout this document, please consider reading the Sandvine technology showcase *The QualityGuard Congestion Response System*.

You also might be interested in the whitepaper *Reasonable Network Management: Best Practices for Network Neutrality*.

Finally, you can learn more about Sandvine's activities regarding network neutrality, including our public commentary, at https://www.sandvine.com/trends/network-neutrality.html.

Invitation to Provide Feedback

Thank you for taking the time to read this whitepaper. We hope that you found it useful, and that it contributed to a greater understanding of the complexities of real-time congestion management.

If you have any feedback at all, then please get in touch with us at whitepapers@sandvine.com.

Headquarters
Sandvine Incorporated ULC
Waterloo, Ontario Canada
Phone: +1 519 880 2600
Email: sales@sandvine.com

European Offices
Sandvine Limited
Basingstoke, UK
Phone: +44 0 1256 698021
Email: sales@sandvine.co.uk

sandvine®
Intelligent Broadband Networks