The Obsession with Bandwidth (Part 1)…

Over the last 20 years the number one recommendation to resolve network experience problems has been “get more bandwidth”. However, despite bandwidth increasing by several orders of magnitude, a poor user experience still remains the no. 1 complaint.

Today it is therefore not surprising that our global fixation with bandwidth fuels an almost immediate willingness to invest in more bandwidth when encouraged to do so. More importantly, our desire for so called ‘speed’ drives an unnatural passion to persistently measure our bandwidth as if our life depended on it! Literally several million bandwidth tests, commonly referred to as speed tests, are done every day just in the USA.

It is clear that the world has bought into the marketing message that more bandwidth is faster and that faster is clearly better than slower. To be more specific, it is a common belief that 100Mbps is faster than 50Mbps and will clearly deliver a better user experience. Is this true?

Consider a new PC that is ordered from Amazon -- most will agree that ‘next day’ delivery is faster than ‘second day’ delivery. For the purchaser, faster is therefore defined by the delivery time for the PC to arrive. It is hard to debate that arriving sooner is slower than arriving later.

To measure the purchaser experience, does it help to know that Amazon’s maximum shipping rate is 20,000pph (packages per hour)? The answer is ‘No', as *a package rate limit has no tangible correlation to delivery time*. What about distance, is distance to the customer material to the experience? For example, will a package being sent to Denver incur a different delivery experience than a package being sent to Sydney from the same shipping point? Most probably ‘Yes’.

Like packages per second, bandwidth is also defined as a rate, namely bits per second (bps). Therefore, should it be expected that bps and pph share similar traits? For example, is sending a package to a customer not unlike sending a network packet to an online user? In particular, as world geography relates to packages, would you expect a greater distance to the online user to consume more time than a shorter distance? If the answer is yes, then it should also be expected that a network connection can incur variations in delivery time based on network related factors such as wireless versus wired or the performance of a cheap $60 router versus one costing $10,000. Interestingly, network time -- which is known as latency -- is well defined, but unlike bandwidth it is seldom considered or referenced. Latency is a measure of time taken for a packet to travel to the destination and back (round trip time, RTT) and is expressed in milliseconds (ms).

Latency defines one of the most important aspects of packet delivery. Consider a ‘cloud’ application service provider and two customers. The connection latency to customer ‘a’ is 10ms and the connection latency to customer ‘b’ is 20ms. If a packet is sent at the same time to both ‘a’ and ‘b’, which customer should get the packet first? Obviously the 10ms connection should arrive first because 10ms is shorter in time than 20ms. That being accepted, would you now expect a different result if the 10ms bandwidth was 50Mbps and the 20ms bandwidth was 100Mbps? Most probably not and yet our passion for bandwidth says otherwise. In short, time is directly material to the user experience whereas bits per second is open to question.

To clarify the time/rate disconnect further, consider the following. A car comes to a fork in a road and both left and right directions are signed 'To Airport'. If the left fork is a 70mph 8-lane freeway and the right fork is a 30mph single lane shared road, should the car go left or right?

If 'speed' is the criteria for selection, then the 'left fork' to freeway is clearly over two times faster. However, if distance is added into the equation this choice may change. For example, if the left sign shows 'to airport, 35 miles' and the right shows 'to airport, 10 miles', the time assessment is shorter for the 'right fork'. The airport is a 30min journey via freeway versus a 20min journey via the single lane road. Does the 'car per hour rate' of the road directly relate to the experience? No, it does not! The *'time to destination'* on the other hand defines a material parameter for the driver's choice. As with the PC, arriving sooner is faster than arriving later, or put in 'experience terms': arriving later can lead to a missed a flight, which is not a good experience.

Why measure bandwidth at all?

Bandwidth essentially defines capacity. The bandwidth rate is one of two parameters that defines a finite limit of a pipe. Under the principles of supply and demand, the stability and performance of a connection is wholly dependent on demand never exceeding supply. We all know what it is like to drive in the rush hour on a road where demand exceeds supply. In comparison, consider a 50Mbps connection that has a 20ms trip latency. This equates to a supply capacity of 1 million bits (20 x 50,000) end-to-end.

Now, what of demand? Turning bits to bytes, 1 million bits is only 125,000 bytes or 125KB -- about the size of a single, low quality, photographic image. A typical web page is comprised of many such images, even higher quality images, graphic files, an abundance of text, tables, animation objects, including videos with sound. In short, the demand of a single user's web browser can easily exceed the capacity of a 50Mbps pipe at 20ms for a period of time, during which, no one else can use the connection.

Understanding, 'how big a delay' and, 'how many users' quickly become crucial questions that demand an answer. The bandwidth rate and latency define the application transaction time for a user but it is the simultaneous user demand that defines the extent and severity of the supply delay penalties that result. Therefore, if bandwidth is simply measured as bits per second without including the assessment of capacity, concurrency and demand, do not be surprised if users complain of unreasonable response times or worse, disconnects.

Increasing bandwidth is fundamental to managing the *concurrency of use*. To address a rush hour problem, the addition of more lanes to a 60 mile (60mph) highway increases the width to allow more cars to share the highway. However, *this should not change the end-to-end time of one hour* (60mph = 1mile per minute) unless congestion delay was known to be the problem. The *delay time* that results from a poor supply/demand ratio, is a crucial metric that is ignored by almost every so-called bandwidth/speed test service. As a result, nearly all the bandwidth measuring solutions fail to assess bandwidth correctly.

Why are most Internet bandwidth tests wrong?

There are three core reasons.

Strike 1, most users accept a bandwidth measurement believing it to be a speed. In reality, if increasing the bit rate does not reduce latency time then any expected improvement will be questionable. What if opposite happens, latency time increases?

Strike 2, many bandwidth measurement solutions, including the most popular, report the bandwidth rate as a speed with no reference to capacity or time. Besides being wrong this naturally aggravates the 'speed/capacity' disconnect.

Strike 3, many bandwidth testing solutions do not measure data correctly. Any bandwidth test method that relies on the principle that the pipe must be continuously kept full of data for the entire duration of the test has flaws. Filling the pipe with certainty requires the bandwidth value to be known when bandwidth measure is the objective of the test. The approach of simply adding data and dividing by time without reference to latency delay delivers a meaningless result that is accepted as meaningful.

What is important about the 1-hour road trip example is arrival time, the 1-hour time defines the true user experience of the highway. To measure bandwidth correctly is no different, the test must assess the bandwidth within its latency time, meaning did the data get to the end of the road in one hour or not? Yes, is good, no is bad.

Example, if a highway test was run for 1 hour and 10 cars report that they each covered 6 miles, the bad test result will report, 60 miles covered in 60 minutes is 60mph. It is not… In reality, by including time equilibrium, the correct result should be reported as 60 miles results in 10 hours to reach the destination i.e. 6mph.

Put another way in terms of the user experience focus, 60mph will be accepted as great, whereas 6mph will be accepted as disaster, which it is. For this reason, it is not surprising that bad bandwidth tests are therefore very popular, especially with providers, and a poor user experience remains the no.1 complaint. Such tests make a bad network look good.

How should bandwidth be measured?

Keep in mind that any bandwidth test that relies on the principle of reading bytes of data continuously (no gaps allowed) for a period of time has 5 fundamental problems.

1. The pipe must always be kept 100% full of data for the entire duration of the test.

2. The amount of data to achieve item 1 is bandwidth x latency, therefore to be accurate the test must know the bandwidth to be able to run the correct test of the bandwidth.

3. Any test demanding 100% of a pipe will threaten the user experience.

4. Any test demanding 100% of a pipe will compromise accuracy unless the pipe is 100% empty

5. Saturating high bandwidth pipes to a 100% will harm the experience of a broad set of users. E.g. 10Mbps and 10Gbps are very different in that 10Gbps will accommodate a much larger number of concurrent users.

If the method for measuring bandwidth threatens the user experience, it stands to reason that limiting the frequency of such tests is essential to minimize the fall out, even if it means running the test in the middle of the night. Does running a test in the middle of the night validate the user experience? Would you expect driving on a highway in the middle of the night to provide an accurate assessment of your commute to work? I suspect not!

However, if the approach is changed to focus on the assessment of the user experience, then the more meaningful result is attained *not by measuring bandwidth less frequently,* but by *measuring the equilibrium throughput (being on time) more frequently*. Such an approach delivers many tangible benefits:

1. A time equilibrium test always reports the user experience because the test is a single user.

2. A single-user test is nondestructive and can be run frequently, including during busy working hours, without affecting the user experience.

3. Non-destructive tests improve the effectiveness of a performance management strategy because the baseline of data is more representative.

4. Frequently run tests quickly identify any network delay variances that threaten the user experience because of the increase in data samples.

5. Destructive bandwidth tests can be initiated to assess capacity specific concerns (supply and demand problems) only when problems are evident.

In conclusion

Acknowledging that a connection's bandwidth defines capacity and concurrency is fundamental to a sound network measurement strategy focused on underwriting a good user experience. The focus of the assessment strategy changes from intangible values, such as the bits per second, to 'user experience' values such as data being 'on-time' and efficiency.

Changing focus to the 'user experience' does not in itself guarantee a good experience. However, ensuring the metrics that harm the user experience are the priority means that threats to the user experience can be recognized early and resolved proactively. Establishing a proactive approach will, without doubt, improve the quality and consistency of the customer experience overall. Equally, establishing an effective early warning assessment of the network experience establishes a framework to understand experience trends and improves the planning and effectiveness of future network enhancements. The old approach of 'we need more bandwidth' is a flawed (and costly) solution without a clear understanding of the experience factor.

In short, elevating the importance of latency and bandwidth (as a capacity, not a speed), will significantly improve the benefits of a sound network assessment strategy by positioning the user experience front and center.

Latency is important because time is everything.