The Speed Triangle

speed promises, growing traffic and capacity in balance

An Operational Practice prepared for SCTE by

Robert-Jan van Minnen
Long Range Planning
Liberty Global – Liberty Tech
Boeing Avenue 53
1119 PE Schiphol-Rijk
rvminnen@libertyglobal.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Title	Page Number

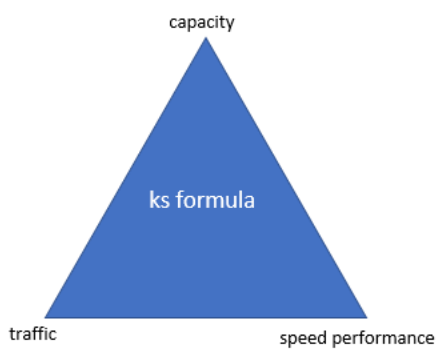
1. Introduction

The 'speed' of internet is typically conceived as the main indicator of the quality of our internet connection. Operators have been using the available top speed as a selling argument. For good reasons because cable could deliver a lot.

As usage grew the race for higher speeds was on. The race could be fueled by adding more and more capacity. But there are limits as spectrum is not free and expansion needs time and investment. Fiber can provide the next order of magnitude but doesn't come for free as well and over time will also require upgrades.

This paper focuses on the question how we can optimize the capacity we deploy in relation to the growing traffic and speed promise.

A simple formula is proposed, tying capacity together with traffic and speed performance. It is an evolution on the k-factor (re. Tom Cloonan, CommScope). This formula is translated into practical applications to plan into the future and to (virtually) measure the actuals of today.



The aim is to support an optimized business balance of the three corners in a measurable way using only the tools we typically already have.

Possible caveats and the current perception of (avoiding) congestion will be discussed as well as practical proof of the results.

Finally, the formula will be applied to possible futures with extreme speeds on DOCSIS or XGS-PON.

2. The 'ks' formula

The proposed formula is essence a combination of two classical methods to determine the amount of capacity needed in a service group (called 'capacity').

- 50% rule; capacity > the highest speed sold / 50%
- k-factor; capacity > utilization + k * the highest speed sold

The 50% rule is extremely simple to use and can be standardize throughout a network, but it discards the actual traffic. As a result, on busy service groups the speed performance is suboptimal. Even up to the point where congestion occurs. In practice an additional planning rule is needed to avoid this. In service groups that are quiet there is more capacity than needed which means inefficiency.

The k-factor must be applied considering the utilization of each service group individually but only provides a factor to the speed. Analyses of the statistical distribution of traffic revealed however that the chance of having sufficient available capacity for a burst of speed is a function of the utilization and not of the planned speed. This means the formula needed an additional factor to modulate the utilization which has been called the 's'. This stands for service level or safety. The result is the ks formula:

capacity = utilization * s + speed * k

Table 1 – ks formula definitions

capacity:	the amount of capacity that is minimally needed [Mbps]
utilization:	the average utilization during the interval that the speed should be reached (e.g. 8-10 pm) [Mbps]
s:	the service level factor to use
speed:	the top speed that is desired [Mbps]
k:	the factor that defines how much of the top speed we want to enable

3. The expectation of 'speed'

The customer perception of speed is influenced by many factors of which some subjective. For network planning a translation into objective parameters is required. In some markets regulators have defined measures to adhere to. Operators may also have defined some themselves. For example:

- *at least 50% of customers at peak time need to achieve headline speed*
- *all customers need to achieve 50% of headline speed at peak*

This in turn requires a definition of the interval for which the rules apply. For example

- Between 8 and 10 pm
- During the busiest hour of the week

It also needs a translation of achievable speed into a network parameter. What a customer would see from a speed test depends on for example:

- The available or unused network capacity during the test interval (typically 10-20 sec)
- Limitations in the customer equipment
- The speed test mechanism

This leads to a basic definition of what 'speed room' is from a network perspective:

- **The amount of capacity in a service group that is available for a burst of traffic**
  - At any given time during the busy hours between 8 and 10 pm
  - For the duration of ten seconds

It is recommended to define 'speed' in this perspective with care since it directly relates to desired performance.

A limited increase of the factor 'k' in the formula can be used to compensate for other factors such as:

- Chance of overlapping speed tests from different users at the same time
- The provisioned speed. Typically, modems are provisioned 5-15% higher than the sold speed. This allows a limited amount of 'catching up' when during a moment in the ten second test there is not enough capacity available.

4. Utilization in a statistical perspective

4.1. Definition

The speed room is defined by the capacity of a service group minus the actual utilization. While the capacity is a constant, the actual utilization in a service group is varying from moment to moment. This calls for a deep dive into this fluctuation.

When utilization is measured it is typically:

the total amount of bits transported during a sample and expressed in Megabits per second [Mbps]

In regular measurement tools the sample time varies from five minutes to one week. For an entire network this yields a vast amount of data. Normally this data is aggregated to averages over a certain amount of time and 'peak' levels.

But if the aim is to measure ten second intervals for speed room calculations the challenge will become soon just to handle this data. Since traffic appears to be generated in a semi-random pattern, the working assumption is that it can be described as a bell curve.

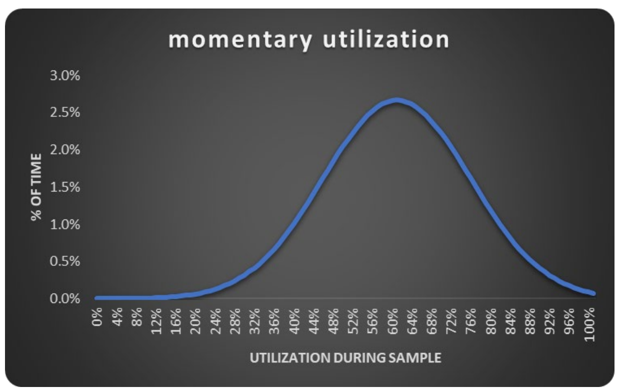


Figure 1 – theoretical distribution of momentary utilization

With this assumption only a few parameters are sufficient to determine the utilization during the interval and the chances of having sufficient speed room.

- Average utilization
- Median utilization (50th percentile)
- Width of the curve (measured at 20th, 80th percentiles)

In this example the average utilization is 60%. Because the bell curve is symmetrical, 50% of the samples have a higher utilization, the other 50% lower. The symmetry of the curve implies that:

median utilization = average utilization

4.2. The width of the bell curve

In practice the expected bell curve may be differing between service groups and can also change over time. It may be narrow when many users are active and wider in more relaxed service groups.

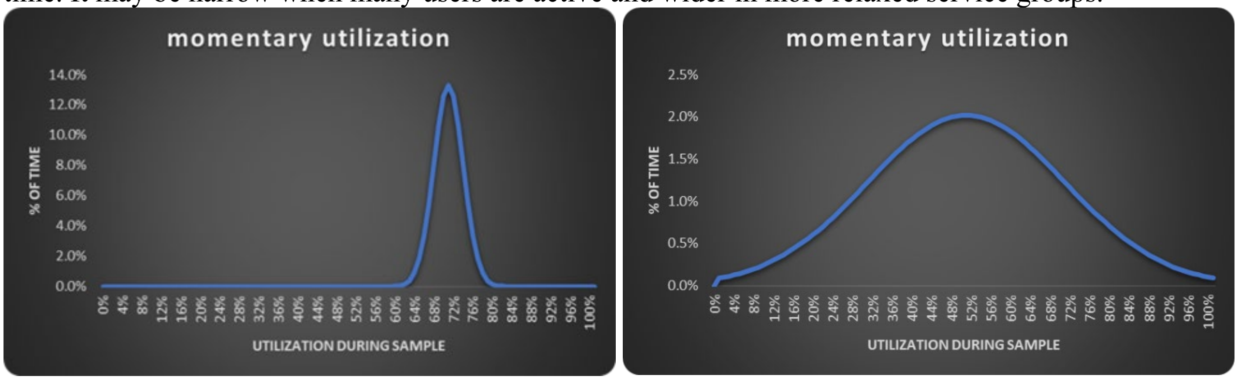


Figure 2 – examples of busy and quiet service groups

The duration of the samples has an impact on the width of the curve. Short samples will give a wider curve. The longer the sample are, the narrower the curve will be. However,

for the median speed room, it is not relevant how wide a symmetrical distribution is

4.3. The symmetry of the bell curve

So far it has been assumed that the curve is symmetrical. What if it is not?

These two examples of a skewed distribution visualize a non-symmetric distribution.

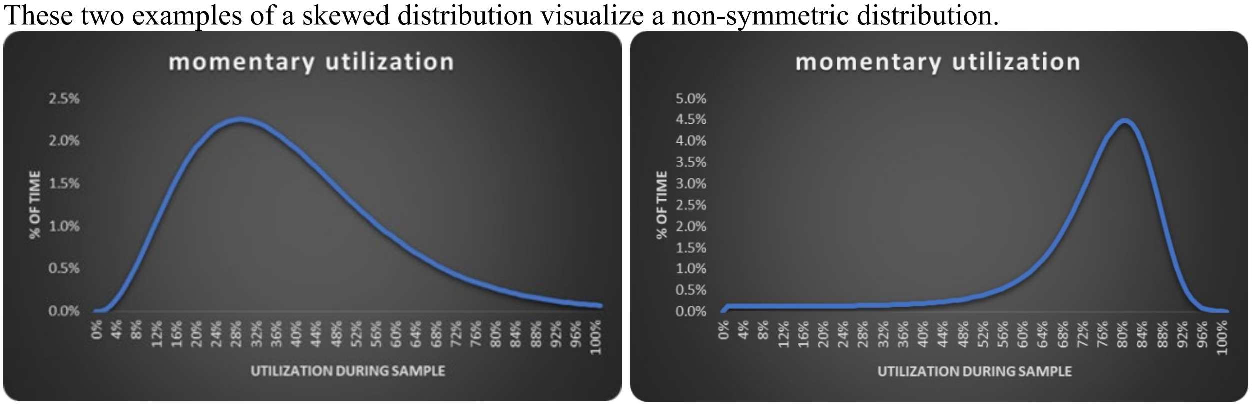


Figure 3 – examples of assymetrical distribution

Calculations on extreme examples have revealed that the median utilization can differ up to 10% from the average utilization. In terms of the formula this means that the median speed is reached when

$$0.9 < s < 1.1$$

This brings in uncertainty in the result. In the left example, the median will be 8% lower than the average. This means the results will be worse than expected if s=1 is used. In the right situation we the median is 6% higher than the average. This can be compensated by using s' = 1.08 (left) / 0.94 (right).

The uncertainty caused by the translation of average utilization into median can be overcome by:

- Measuring the median utilization in ten second samples which may give additional complexity
- Use a default s' = 1.1 and accept a possible overestimation of the traffic
- Calibrate with actual speed test results

4.4. The bell curve in reality

Conversion of actual utilization metrics into a distribution chart can be quite challenging. Part of the challenge is the fact that besides the randomness there are also diurnal, seasonal and long-term trends. On top there are single events such as big downloads from software releases, video events and even the pandemic.

The working assumption has been tested in various set-ups. Regardless of the sample time, network and time of day, a histogram yields a bell-type curve.

In this example of a long period of two years where the busiest hour of each week is shown, the distribution is narrow. The median is very close to the average (0.3% difference).

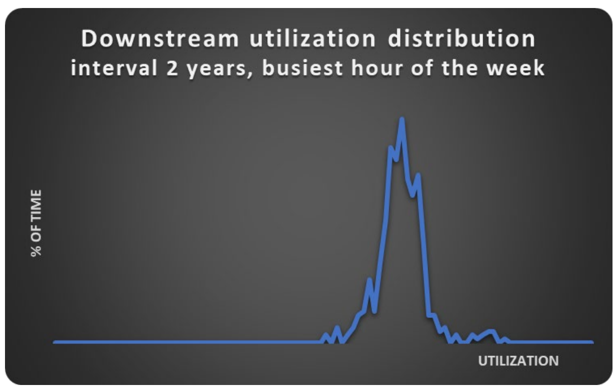


Figure 4 – long term utilization distribution of an entire network

The following example shows the correlation between ten second- and fifteen-minute samples. Both yield a bell-curve which follows the diurnal and weekly trend and correlate strongly. Because it is measured over two weeks, the results include busy and quiet times.

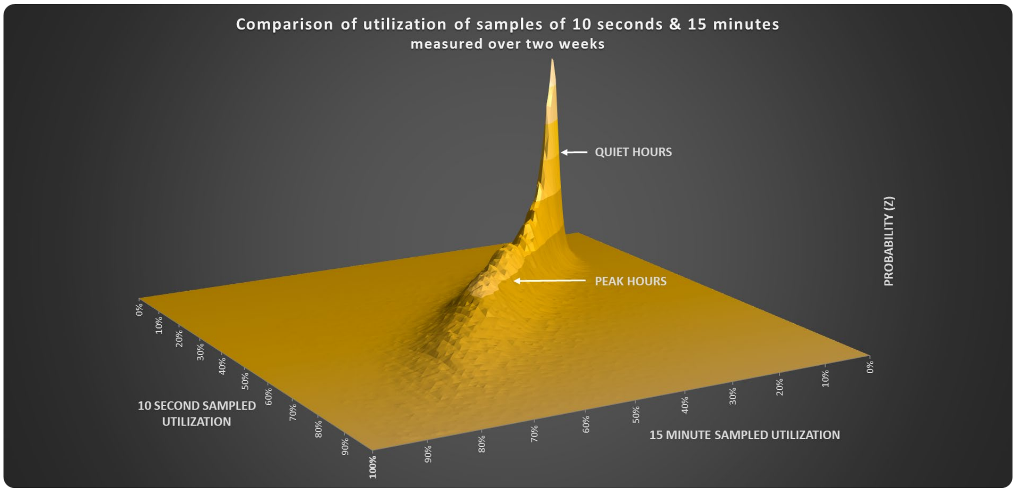


Figure 5 – short-term utilization of a service group

4.5. The cumulative curve

The distribution from figure 1 can be shown as a cumulative curve. This shows for each level of relative utilization which percentage of time the actuals will be lower. If we depict a speed as a percentage of the service group capacity (e.g. 1 Gbps in a 2Gbps service group is ½ = 50%) we can read the percentage of time the utilization samples will be lower and a speed test would be successful.

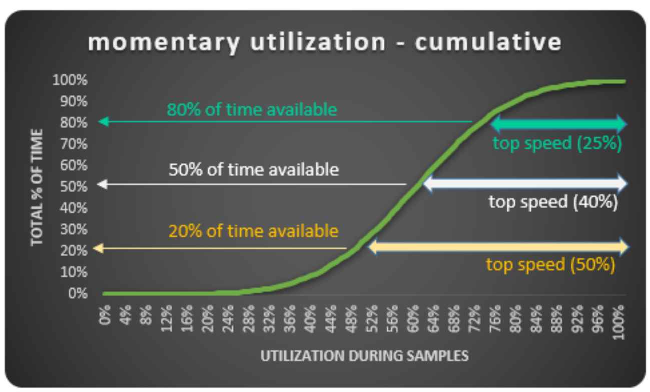


Figure 6 –cumulative momentary utilization

In this example the average utilization is 60%. It is a symmetrical distribution, so the median utilization is the same. If this is a 2Gbps service group, the chance of a successful speed test for 1Gbps is only 20%. This is an example that the traditional planning with a 50% rule will give a relatively low performance in a busy service group.

Through mathematical manipulation using the ks formula, any distribution (measured or assumed) can be altered to show which percentage of time success belongs to which s if we measure or calculate it.

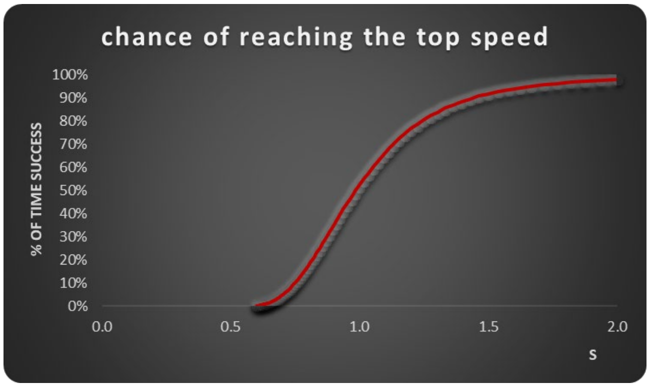


Figure 7 –s and the percentage of success

Following the example above, the 1Gbps top speed has an s = (2 -1 Gbps)/ (60% * 2 Gbps) = 0.83. The corresponding % of time success is 20% as is also found in figure 5.

Note that this is just another representation of the same distribution data. It is useful for:

- Determination of other percentiles than the median
- Sensitivity analyses (what if the growth is higher than expected)
- Reconstruction of the distribution from actual speed test data

5. Planning capacity

5.1. Basic calculation

Calculation of required capacity in a service group to obtain a certain speed room has become simple.

$$\textbf{capacity} = \textbf{utilization} * \textbf{s} + \textbf{speed} * \textbf{k}$$

Example 1: for a service group with an average utilization of 400 Mbps during peak hours and a top speed of 500 Mbps, the minimum capacity = 400 * 1 + 500 * 1 = 900 Mbps. In that case the median speed room will be the top speed.

Example 2: for a service group with an average utilization of 250 Mbps during peak hours and a top speed of 300 Mbps, 80% of speed tests must be successful. To achieve an 80th percentile of success, s = 1.25 must be used. This value is obtained from the cumulative distribution in the previous chapter. The minimum capacity = 250 * 1.25 + 300 * 1 = 625 Mbps. In that case the 80th percentile of speed tests will be equal to the top speed.

5.2. Calculating a speed upgrade

When the top speed of a network is upgraded the capacity must be expanded accordingly.

As an example, the top speed is upgraded from 1Gbps to 2Gbps. In the traditional calculation with a 50% rule this would require 1Gbps / 50% = 2Gbps of additional capacity. The ks approach requires the addition of k * 1Gbps of capacity. Supposed that k = 1.2 this is 1.2Gbps of additional capacity. Since the term s * utilization remains does not change, the performance of the 2Gbps product after the upgrade will be the same as the 1Gbps product before. Instead of 2Gbps additional capacity, only 1.2Gbps is added with the same performance as before the upgrade!

5.3. No standard capacity across the network

In essence the required capacity is different for every service group because the utilization is different. This could pose an operational challenge. In practice it will be desirable to have some degree of standardization. A method can be to classify all service groups on their utilization and/or spectrum availability. Then plan for the busiest in each group. This ensures sufficient capacity while avoiding too much idle capacity, unnecessary spectrum occupation or node splits.

5.4. Efficiency gain

Compared to a standardized capacity across all service groups, a deployment based on the ks-formula gives an efficiency gain. The amount depends of course on the distribution of busy and quiet areas and the already used standard capacity. In an example busy network the actuals were compared with the requirements following the ks formula. In the quieter areas 17.5% less capacity was needed. In the busiest areas (4.4% of service groups) the standard capacity was insufficient. To reach the median speed as defined 0.2% of additional capacity would be needed. This would give a net gain of ~ 17%, provided there is spectrum available. If this is not available this should drive segmentations and cost 4.4% more capacity. Still a net gain of ~13% and ensured speed performance across the network.

6. Virtual measurement

The formula has a mathematical form and can be altered to virtually measure both s and k from real life data. This can be used to verify the calculation or even manage the actual operation of a network. There are two reporting options which are explained here:

- Customers reaching the defined speed
- Median available speed room

6.1. Customers reaching the median speed

In essence this is counting the customers that are on service groups with s >= 1.

In an existing network the capacity, utilization and sold top speed are known. This means if k=1 (or another value of k that is desired) s can be calculated for each service group. The ks formula is transformed into:

$$s = \frac{\text{capacity - speed * k}}{\text{utilization}}$$

The resulting dimensionless s provides information on the propensity of reaching the top speed for a random speed test as explained. For s = 1 this is the median result.

When all customers are counted that are connected to a service group that has s >= 1 and divided by the total customer count we have an indication of the percentage of customers that have sufficient speed room.

6.2. Customers reaching the minimum speed

When there is a minimum speed guarantee the calculation can be repeated with a different s. In the example the rule applies for the busiest hour. This means a second s_{max} must be calculated for the utilization at the busiest hour which of course must be available.
A 100% guarantee is not possible, in the example 95% is chosen. So 95% of random tests must have a minimum of 50% of speed room available. The graph in figure 6 shows that the related s = 1.7

Again, the customers are counted. This time from service groups with s_{max} >= 1.7.

6.3. Median speed room

To calculate the median speed room the s in the formula is fixed: s = 1. Another transformation of the ks formula yields:

$$k = \frac{\text{capacity - utilization * s}}{\text{top speed}}$$

The resulting k gives the speed room / top speed ratio. The actual median speed room is then the found with k * top speed. This value is not directly usable to report on a customer level because CPE are provisioned for a certain top speed. Normally 5-15% more than the sold top speed. This means the median top speed from network perspective must be limited to this value to obtain a value that is achievable for customers. If the overprovisioning is 10%, all results above 110% must be capped at this level before calculating the average for a network.

6.4. Median speed in real life

The virtual measurement of the median speed has been put to the test in a multi-million homes network with several thousands of SamKnows' probes. The virtual measurement theory dictates that if k>=1, the median speed room equals the top speed. With k < 1, the median speed goes proportionally down.

The results from hourly tests from ~1.500 probes are compared here with the results from the virtual measurement with network data as described in section 6.3. The expected relation is clearly visible and the found k to meet the median speed in practice is ~ 1.2 for downstream and ~ 1 for upstream.

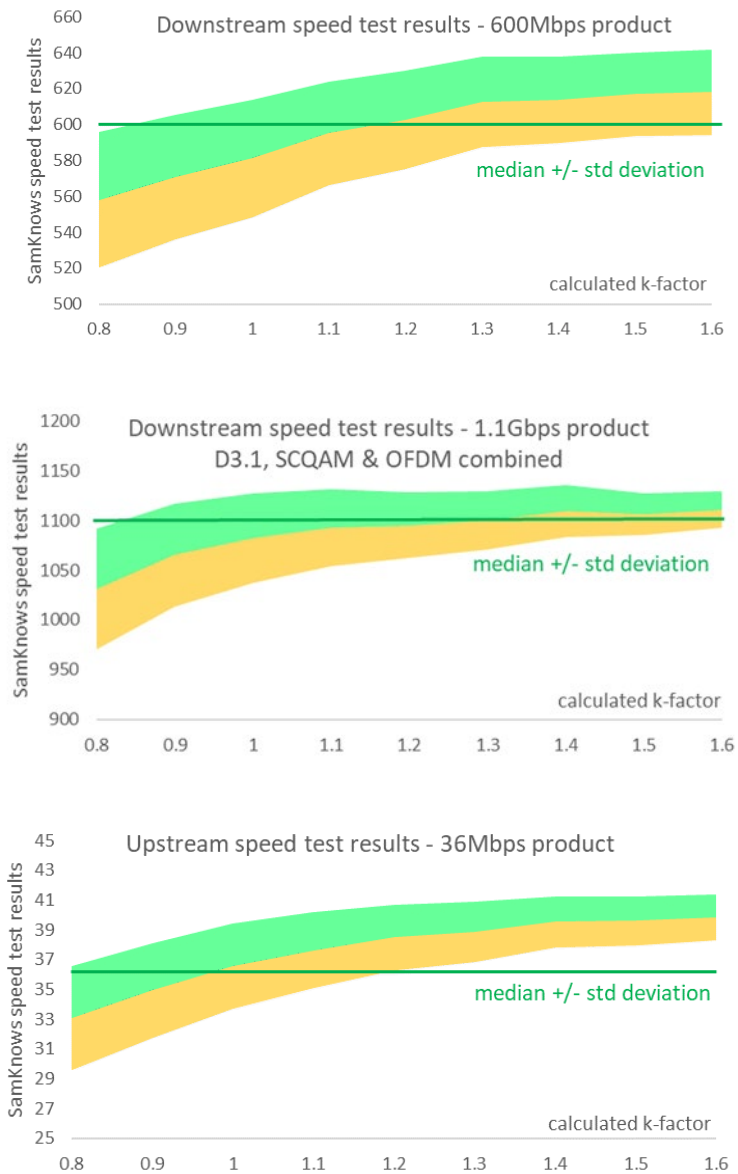


Figure 8 –virtually measured speed compared with actual test results

7. Congestion

Network congestion is a thing operators want to avoid. The classical approach is to provide sufficient capacity to keep the utilization under a certain percentage. The knowledge from the ks approach can be used to create a different view on congestion and discover opportunities for improvement.

7.1. What is traffic congestion

First step is to define what congestion is and what should be avoided. For this general road traffic theory is used which defines three stages of traffic flow:

1. Free flow of traffic. There are no obstacles to reduce the speed.
2. Viscous flow of traffic. It is busy and the speed reduces on occasions but still flowing.
3. Jammed traffic. Speeds drop to zero on occasions. This leads to a build-up of even more traffic and longer stand-stills

7.2. Congestion versus ks

The three stages of traffic flow can be defined as speed rules. For example:

1. Free flow: the median speed room is achieved at evening peak hours
2. Viscous: at the busiest hour at least 50% of the top speed must be available in 95% of cases
3. Jammed: the speed room is < 10% for > 1% of time

The cumulative traffic curve(s) in use can be used to find the related s for each stage. These values of s can be put in the formula to add additional rules for planning. At the same time the actual network data can be virtually measured to determine the percentage of customers in each stage.

7.3. Never 100%

The stochastic behavior of traffic predicts that at any time short peaks of traffic will utilize the full capacity, even in quiet service groups. If the samples are short enough, we will 'hit the ceiling' sometimes. This is not a problem as there are buffers and IP has ways of resolving lost packets. Customer impact is expected however when the ceiling is hit too long or often. This is exactly what the cumulative distribution curve can be used for to detect.

8. Traffic events

The ks formula works from the assumption that traffic has a semi-random character. The practical tests have confirmed this but in recent years we have seen an increasing number of events that can change the traffic volume suddenly.

The graph below shows the surges of the combined Liberty Global network. The percentage is relative to the expected volume and compensated for normal growth and seasonal fluctuation. In other words, a sudden spike in traffic for one or more hours.

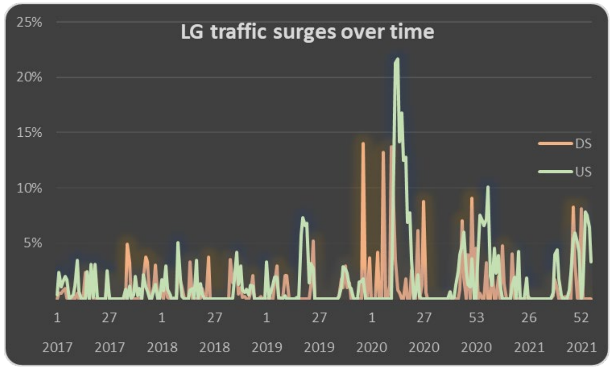


Figure 9 –growing traffic surges

Qualitative analysis showed that the surges are mainly caused by:

- Video broadcast over IP of popular games
- Releases of large software updates Upstream surges from:
- Work and school from home

Because the events are not always plannable, may last short and data collection is complex there is to date insufficient data to determine the impact on the cumulative distribution. A desk examination on two examples can give some hints on what to expect.

8.1. Many video streams

If many customers are watching a video stream, the traffic per customer is relatively low but the number of users is high. For game events the beginning and end are approximately the same for every user. This means a constant amount of traffic is added during the event with little variation. Note that a significant amount of the video traffic will replace regular traffic so the net addition will be less than calculated by the number of users * traffic per user. This 'replacement' is estimated at ~50% of the calculated amount.

If viewed from a ten second sample perspective, the distribution curve shifts to the right during the event. Average and median will also shift with the same amount. To maintain the performance this traffic must be included in the planned capacity as discussed.
If it is not included, the performance will go down. This is shown in the graphs below. The s-curve shows the chance of success according to the initially planned traffic. If we planned for an s=1 and temporarily include 20% of traffic, the percentage of successful speed tests will drop from 50% (median) to 25%.

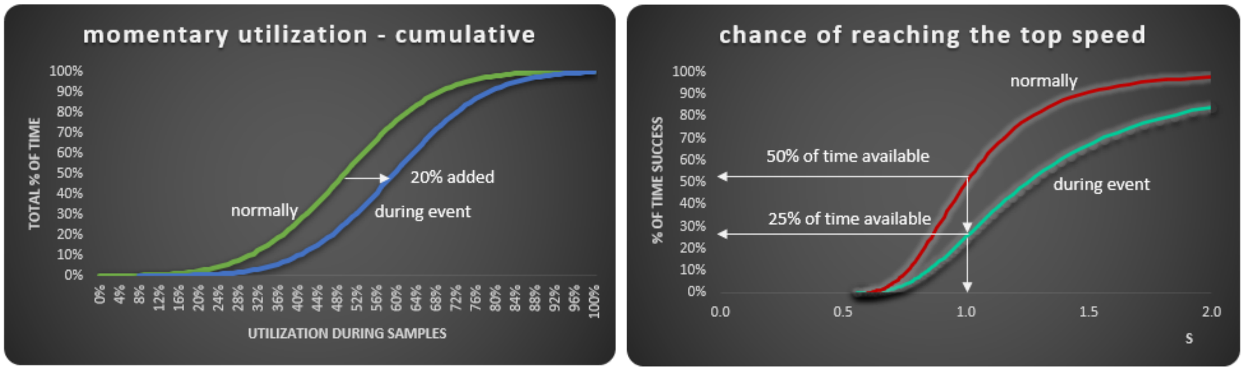


Figure 10 –impact of video event

8.2. Big downloads

An example is used from an observed surge of 25% linked to a game update of 40 GB. The amount of data corresponds with ~1.3% of all customers downloading the update in the same hour.

A 1Gbps customer would be able to download the update in ~5.3 minutes, a 500Mbps customer 11 minutes and for 100Mbps it would take 53 minutes. If these downloads start at the same time, for the first five minutes the traffic surges +78% after which the surge drops in size. If all customers would have a top speed of 1Gbps, for five minutes the traffic would increase by 281%.

The graph shows the hypothetical traffic increase if all downloads start at the same time and customers have a top speed as they are normally distributed.

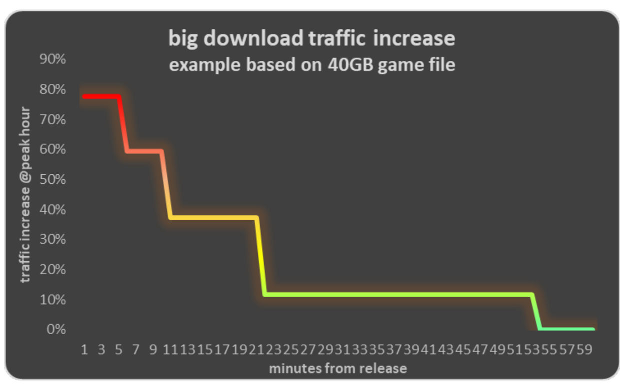


Figure 11 –impact of simultaneous software download

In reality traffic increased with 25% for the full hour and beyond.

- downloads do not start at the same time (planned)
- peering or IP network capacity is limited
- access capacity is temporarily at its limit

Chances are that the limit of capacity in terms of this document is reached for some time during the downloading. As a result, the speed room reduces to zero during those times.

Simply adding 25% of traffic in the formula (multiply s with 1.25) may not be sufficient to cater for these events. The reason is that the surge would become higher but still the ceiling would still be hit for some time, only shorter. It is a matter of choice to what extent the speed room must be available during these releases. Managing the release times may be more effective and cheaper option. Or accepting that the performance is not compliant for the short duration of these events.

8.3. Speed tests

Performing actual speed tests from the customer's equipment is of course the ultimate proof of speed performance. For testing the network speed SamKnows probes in modems are available. This test rules out any limitation in the customer equipment and wiring. One could argue to use these systematically to have a reliable indicator of the network performance. But performing tests has an impact on the utilization of the network and in the end influences the performance for regular traffic.

Based on the average provisioned speeds an estimation of the additional traffic at peak hours can be made if all would perform a scheduled hourly test of ten seconds. For downstream the addition would be 19%, for upstream 30%. This means that just because of the proofing of the speed room, the capacity must be increased by these amounts to actually achieve the speed room.

If tests are performed randomly during the hour there is a chance they overlap. In case two top speed tests overlap it is likely both will fail as they compete for the same speed headroom. This chance of failure is the highest for the top speeds but can also happen with other speeds or if more than two overlap. In the example the chances of overlap are:

Table 2 - overlap and failure of random speed tests

speedtest for	overlaps with another for	overlap	failure
1000 Mbps	1000 Mbps	2%	4%
1000 Mbps	500 Mbps	13%	24%
all	all	53%	78%

The percentages indicate that a significant amount of speed tests would be impaired just because of the fact that the tests are performed. Similar percentages occur in the upstream direction.

This self-inflicted performance reduction can be mitigated by scheduling the tests so that they do not overlap. The feasibility of this scheduling is not investigated but assumed possible (SamKnows is supposed to avoid multiple tests at the same time on one node). In the example during 46 minutes of every hour there would be a downstream speed test going on (38 in upstream).

Implementing scheduled speed tests means adding capacity to the network to carry the added amount of traffic. If this is added following the ks formula, the statistical availability of speed room would still be as expected because both consider 10 second samples.

There could be however an impact on the customer perception for normal traffic since this is built up from much shorter bursts. To understand the impact in detail, the utilization distribution of shorter samples would be needed. This is not available to date.

9. Long term outlook

The ks formula works from metrics that are familiar from the DOCSIS world. In essence the metrics relate to properties of the aggregated internet traffic from customers, regardless of the technology. This implies that the same method can be used for other means of internet access.

9.1. XGS-PON

As an exercise the calculations are done for hypothetical DOCSIS4.0 and XGS-PON networks.

Table 3 - DOCSIS4.0 and XGS-PON example calculations

	D4.0 1.8/204		XGS-PON	
	Down	Up	Down	Up
capacity	12280	1228	10000	10000
traffic/customer	10	1	10	1
customers/service group	180		24	
traffic/service group	1800	180	240	24
for median (s = 1)				
speed room (k * top speed)	10480	1048	9760	9976
for k = 1.2				
ensured median top speed	**8733**	**873**	**8133**	**8313**
for an 8Gps call out, s =	1.4	1.5	0.7	6.7
speed test success	**81%**	**89%**	**20%**	**~100%**
utilization	**15%**	**15%**	**2%**	**0%**

The expected traffic is subtracted from the available capacity to obtain the speed room. When divided by the k of choice (here k = 1.2), the ensured median speed room is obtained. In case the market call-out is 8Gbps (0.8 for upstream in DOCSIS), the resulting s is calculated (s = (capacity - k * speed) / utilization). With the sample s-curve from figure 2, the speed test success ratio is determined.

9.2. Service group size

Over time the traffic per customer grows and as a result the number of customers per service group reduces. Metcalfe's law predicts that with less users, the volatility of traffic increases. Investigations have confirmed this in practice. As a result, the cumulative distribution and s-curve will get steeper. This means that smaller variations in traffic or s will result in changed speed room results. For the median it is not relevant how wide the distribution is. But for a guaranteed minimum (95% of time minimum 50%

speed room) a higher s is needed, leading to more capacity. For long-term planning it is recommended to include this in the long term s factor.

10. Conclusions

The capacity that a network needs depends on the traffic and desired speed room. The semi-random behavior of traffic can be modelled with a certain accuracy with a simple formula:

$$\textbf{capacity} = \textbf{utilization} * \textbf{s} \quad + \quad \textbf{speed} * \textbf{k}$$

The formula is most accurate and easiest to implement when the performance is defined for the median speed room. It can be enhanced with multiple planning rules to suit the business or market regulations.

Application of the formula for planning provides the option to diversify the capacity and with that save ~15% of capacity.

The same model provides a measurement tool for speed performance based on regular network metrics. Refinement of the k and s variables can be performed with these virtual measurements or with actual speed test results when available.

Congestion avoidance can be refined because the formula can be used predictively and divide congestion in phases, like in road traffic theory.

Traffic events can disturb the traffic distribution for a short period of time. Video events can be predicted and modelled with the formula or accepted with a reported temporary lower speed room. Big downloads can have a significant and prolonged impact on the performance.

Speed tests are useful to verify the planning and virtual measurement but testing in high volumes should be avoided since the tests influence the result and increase the need for capacity.

The ks formula is a general tool, independent of technology or time. It is recommended for short- and long-term planning and quick speed performance evaluation.

Abbreviations and Definitions

service group	aggregation of channels to serve several customers
Capacity [Mbps]	Total capacity in a service group in Megabits per second
Megabits per second [Mbps]	Megabits per second [Mbps]
GB	Gigabyte
SCTE	Society of Cable Telecommunications Engineers

Bibliography & References

k-factor, T. Cloonan; CommScope

Metcalfe's law; Wikipedia