Managing the Mobile Bandwidth Explosion: A Solution for the Client Edge

A White Paper from GoS Networks

Contents

Executive Summary

The relentless growth of network bandwidth, and the associated increase in traffic volume, provides an on-going challenge for Mobile Network Operators everywhere – particularly when it comes to the management and control of CAPEX and OPEX. Significant proportions of MNO CAPEX and OPEX are consumed by growing and managing network infrastructure, and the problem is growing with traffic volumes.

Despite continued heavy investment, they are still failing to meet capacity demand – data from Cisco suggests a CAGR of 108% for mobile data traffic between 2009 and 2014 (Cisco, 2010). With this growth rate MNOs simply cannot upgrade capacity fast enough.

Insufficient bandwidth provision leads to poor network performance. This has a negative effect on subscriber experience, leading to dissatisfaction with the network provider. The challenge for MNOs is to manage CAPEX and OPEX investments to ensure return on investment is maximised, and to deliver the levels of enhanced performance that are now being demanded by users.

In addition, MNOs need to understand the 'nature' of the traffic sent to and from subscribers in order to provision existing network resources effectively and appropriately, and to capture data on subscriber behaviour.

Unfortunately, there are numerous problems that currently reduce the efficient allocation of existing bandwidth resources, the benefits of future investments, and the ability of MNOs to fully understand user application demands.

Current solutions leverage Deep Packet Inspection (DPI) a technology that allows Internet Protocol (IP) packets to be inspected as they traverse paths between end points. However, DPI-based solutions are limited as they do not provide full visibility of all traffic from the end points or user devices.

MNOs need to solve the congestion problem and secure greater visibility of application performance, usage and demand across their networks. A new solution promises to address these issues: namely, Guarantee of Service (or 'GoS™') from GoS Networks.

Key Points

1. With GoS, Mobile Network Operators can defer CAPEX investment in bandwidth and capacity by up to 9 months
2. GoS enables Mobile Network Operators to improve video download efficiency by up to 10 x.
3. GoS helps Mobile Network Operators save 20% of overall bandwidth requirements
4. GoS can be deployed in mobile devices, dongles and internet keys to prioritise and manage bandwidth requirements for both downstream and upstream traffic, helping reduce the total network load.
5. GoS can restrict bandwidth consumption to what is really necessary, helping promote fair usage between users in any given cell.
6. GoS enables bandwidth to be used more efficiently, allowing MNOs to defer investment in capacity upgrades, help meet financial targets and manage CAPEX and OPEX more effectively.
7. GoS allows an MNO to rapidly deploy and guarantee revenue-generating data services on an existing data connection, reducing OPEX consumption

Guarantee of Service, or “GoS”, is a unique solution from GoS Networks that can be embedded as a software client within mobile devices and dongles. GoS delivers multiple real-time classes of service, enabling multiple, parallel real-time (and near real-time) applications by providing independent control of traffic.

GoS provides the means to manage bandwidth demands more efficiently by maintaining traffic characteristics even when the bandwidth is fully utilised and by allowing up to 90% of the bandwidth to be used for real-time and critical applications. This can lead to a significant reduction in network congestion. If congestion can be relieved so that available bandwidth is used more efficiently, then investments in capacity increases can be deferred.

Streaming video is an application that is experiencing strong growth from mobile users, but it can cause significant downstream congestion when downloading takes place at the maximum possible bandwidth. Estimates by Cisco and others suggest that half of all traffic in mobile networks is derived from video delivered to mobile broadband clients. Calculations suggest that applying GoS to streaming video traffic could reduce the overall peak load by nearly 50%, which is equivalent to 9.5 months traffic growth at current rates, estimated at 108% CAGR. Further, additional calculations based on streaming rates indicate that use of GoS can result in a 20% reduction in the amount of video data consumed.

GoS can be deployed in mobile devices and dongles to prioritise and manage bandwidth requirements for both downstream and upstream traffic, helping to reduce the total network load. By restricting bandwidth consumption to what is really necessary, GoS promotes fair usage between users in any given cell. By using bandwidth more efficiently, MNOs can defer investment in capacity upgrades in order to meet financial targets and manage CAPEX and OPEX more effectively.

Introduction

The relentless growth of network bandwidth, and the associated increase in traffic volume, provides an on-going challenge for Mobile Network Operators everywhere – particularly when it comes to the management and control of CAPEX and OPEX. Significant proportions of MNO CAPEX and OPEX are consumed by growing and managing network infrastructure, and the problem is growing with traffic volumes. These challenges are aptly illustrated by the following passage, which describes “The Red Queen’s Race” in the classic book “Through the Looking Glass” by Lewis Carroll:

"Well, in our country," said Alice, still panting a little, "you'd generally get to somewhere else — if you run very fast for a long time, as we've been doing."

"A slow sort of country!" said the Queen. "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"

While investment in bandwidth is unavoidable, current growth rates mean that MNOs face a continually moving target. Despite continued heavy investment, they are still failing to meet capacity demand – data from Cisco suggests a CAGR of 108% for mobile data traffic between 2009 and 2014 (Cisco, 2010). With this growth rate MNOs simply cannot upgrade capacity fast enough. Like the Red Queen, they are running just to stand still.

Of course, insufficient bandwidth provision leads to poor network performance. In turn, this has a negative effect on subscriber experience, leading to dissatisfaction with the network provider. The challenge for MNOs is to manage CAPEX and OPEX investments to ensure return on investment is maximised, and to deliver the levels of enhanced performance that are now being demanded by users.

In addition, MNOs need to understand the ‘nature’ of the traffic sent to and from subscribers in order to provision existing network resources effectively and appropriately, and to capture data on subscriber behaviour.

Unfortunately, there are numerous problems that currently reduce the efficient allocation of existing bandwidth resources, the benefits of future investments, and the ability of MNOs to fully understand user application demands.

Solving these problems will ensure more effective utilisation of bandwidth deployments and create conditions for greater returns on planned network capacity upgrades. If existing bandwidth can be used more efficiently, investments can be deferred, generating significant savings in CAPEX and OPEX (depending on the model used by the MNO for bandwidth deployment: owned or leased capacity). Furthermore, a better understanding of subscriber behaviour will help MNOs to create new revenue streams through more targeted offers and to enhance customer satisfaction with existing offers.

Current solutions leverage Deep Packet Inspection (DPI) a technology that allows Internet Protocol (IP) packets to be inspected as they traverse paths between end points. However, DPI-based solutions are limited as they do not provide full visibility of all traffic from the end points or user devices.

This white paper will explore the problems faced by MNOs in meeting bandwidth demands and highlight a solution that complements traditional DPI equipment and can:

- Help investments go further, and allow deferment of capacity investment;
- Lead to a more efficient evolution of network infrastructure; and
- Provide greater visibility of traffic usage for mobile devices and dongles.

Bandwidth Consumption: A Multi-Sided Problem

Current limitations with mobile broadband capacity effectively constrain application performance – and hence, user experience. From the user perspective, there are two aspects of bandwidth consumption that are important:

- Upstream traffic; that generated at subscriber devices; and
- Downstream traffic; that delivered to client devices.

Quite simply, consumers expect bandwidth capacity for downstream traffic to be sufficient to power the applications they want at an appropriate performance level. Traditionally, MNOs have focused on achieving sufficient downstream traffic performance. However, user-generated traffic can be equally important when addressing capacity concerns, but it is often overlooked by MNOs. In a typical network, greater capacity is available in the downstream direction – with the result that upstream bandwidth is more easily saturated.

Increasingly, mobile users are generating greater volumes of traffic, all of which needs to be delivered by the network. This is associated with, among other things, the rise of user-generated content, increased use of Voice over IP (VoIP) clients for communication, and 'Peer to Peer' (P2P) traffic. However, this presents an additional problem for MNOs: some of these applications create revenue, while others simply burden the network. In addition, the combination of different traffic sources may cause problems for certain applications. For example, heavy P2P traffic (which does not generate revenue) may limit the effective transmission of revenue-generating services, such as VoIP, or cause disruptions to other real-time services.

The continued rise in downstream and upstream traffic leads to congestion in the network that inhibits the performance of applications in both directions. Currently, this can only be effectively relieved by increasing network capacity, which *pace* the Red Queen, is an ultimately unsatisfactory approach.

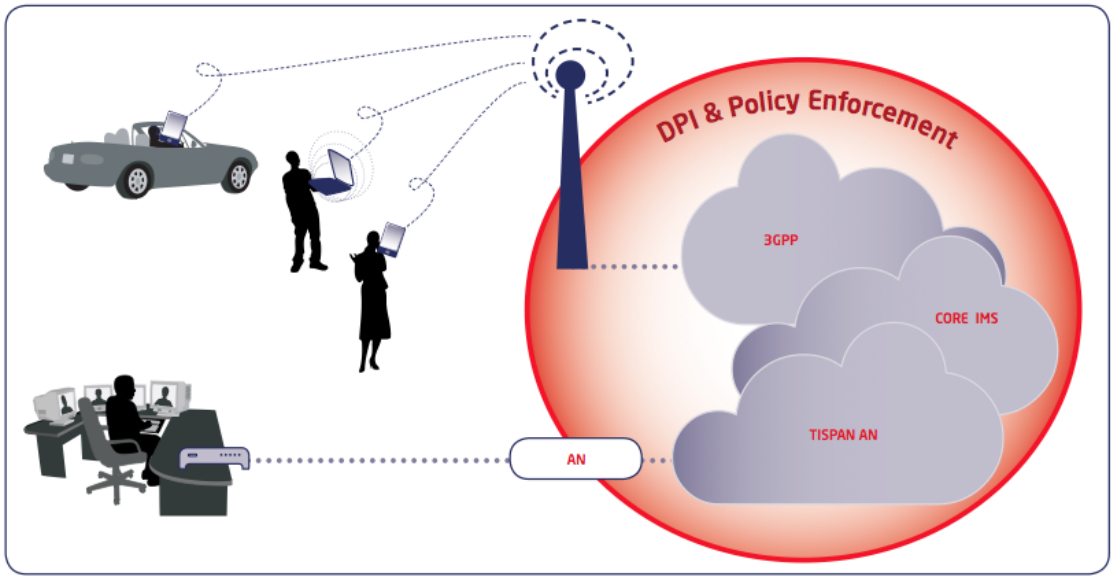
An additional problem with bandwidth consumption is related to the 'visibility' of the traffic consumed by users. For MNOs, it is essential that they can monitor, in real time, the types of applications being used by subscribers, for several reasons. First, MNOs have to ensure that applications perform at the appropriate level, so that user experience is maintained. This ensures that their basic service offer is of an acceptable standard and is an important aspect of the fight against subscriber churn.

Secondly, MNOs need to constantly monitor their network to ensure that it performs consistently across all cell sites. For example, some areas will experience more demand than others, which typically relates to fluctuations in user density, or to spikes in demand caused by localised or irregular events. It is therefore not enough for MNOs to be aware of specific applications, they also need to consider the entire network infrastructure, the nature of the traffic within it, and how to cope with expected surges in demand.

Thirdly, MNOs have historically faced continued downward pricing pressure and now wish to address this by offering differentiated pricing bundles and services. In order to achieve this most efficiently, MNOs have to increase their knowledge of subscriber behaviour and application demand. Awareness of application usage and demand allows MNOs to understand subscriber behaviour and create more targeted offers both to all users and, critically, on an individual user level. Such knowledge can enable MNOs to offer differentiated charging plans that go beyond "all you can eat" bundles, and to create new revenue opportunities.

These problems are illustrated in Figure 1, which highlights the lack of visibility of application demand and consumption at the edge of the network.

Figure 1: Application Visibility



MNOs need to solve the congestion problem and secure greater visibility of application performance, usage and demand across their networks. A new solution promises to address these issues: namely, Guarantee of Service (or 'GoS') from GoS Networks.

Solving Congestion Problems

Guarantee of Service, or "GoS", is a unique solution from GoS Networks that can be embedded as a software client within mobile devices and dongles. GoS delivers multiple real-time classes of service, enabling multiple, parallel real-time (and near real-time) applications by providing independent control of traffic. These service classes concern:

- Throughput;
- Loss; and
- Delay / Jitter.

In this way, GoS provides fine-grained prioritisation of different traffic types, ensuring acceptable performance for real-time applications even when the upstream bandwidth is saturated. GoS also provides the means to manage bandwidth demands more efficiently by maintaining traffic characteristics even when the bandwidth is fully utilised and by allowing up to 90% of the bandwidth to be used for real-time and critical applications. In turn, this can lead to a significant reduction in network congestion. If congestion can be relieved so that available bandwidth is used more efficiently, then investments in capacity increases can be deferred.

With current published growth rate estimates, it is possible to calculate the time factor for CAPEX deferment that a given reduction in network congestion will deliver for particular use

cases. It is further possible to calculate bandwidth savings based on the application of GoS to particular applications.

Capex Deferment Estimates

Streaming video is an application that is experiencing strong growth from mobile users, but it can cause significant downstream congestion when downloading takes place at the maximum possible bandwidth. It is often assumed that to achieve appropriate levels of customer satisfaction for streaming video it is essential to maintain maximum download rates at all times. However, independent experiments undertaken by GoS Networks have shown that customer experience is actually unaffected if the maximum download rate is constrained to about 1.25x the rate of consumption. Since the maximum download rate can be anything up to 16[1] times the minimum required, this represents a substantial saving. A typical streaming rate is 440kbps, so 660kbps is more than adequate to prevent playout pauses. Restricting a video download to this rate – rather than the full HSDPA rate of 7.2Mbps – reduces the peak bandwidth consumed by video downloads by a factor of more than 10. Estimates by Cisco and others suggest that half of all traffic in mobile networks is derived from video delivered to mobile broadband clients, so applying GoS in this case could reduce the overall peak by nearly 50%, which is equivalent to 9.5 months traffic growth at current rates, estimated at 108% CAGR (Cisco, 2010).

For MNOs faced with pressure on CAPEX budgets, being able to gracefully manage network enhancements and plan with more granularity can be a significant benefit. This can be achieved by deploying GoS in mobile devices and dongles to actively manage download services by restricting download rates while ensuring fair access to available services and resources.

Active Bandwidth Saving

Device clients that stream video download content do so at a faster rate than it is played out in order to fill buffers. Experiments show that a 25% excess rate is sufficient to avoid playback pauses. However, a client in an HSPA (High Speed Packet Access) cell may use the full 3G rate, thereby denying service to other users. Furthermore, statistics show that users typically abandon a video clip before reaching the end, so the capacity used to download and buffer the end of the clip is wasted. A recent study by Visible Measures examined short-

[1] This figure is derived from independent calculations by GoS Networks, that compare the minimum download rate with the typical maximum downlink capacity of HSPA 3G. Bmin = 440k, Bmax = 7.2M, => Bmax/Bmin = 16.36

form videos of less than 300 seconds (5 minutes) in duration, and made a number of key findings (Visible Measures, 2010).

- 20% of the audience will abandon viewing the video in the first 10 seconds;
- 33% will click away before the 30 second mark; and
- 44% of the audience will have left by 60 seconds.

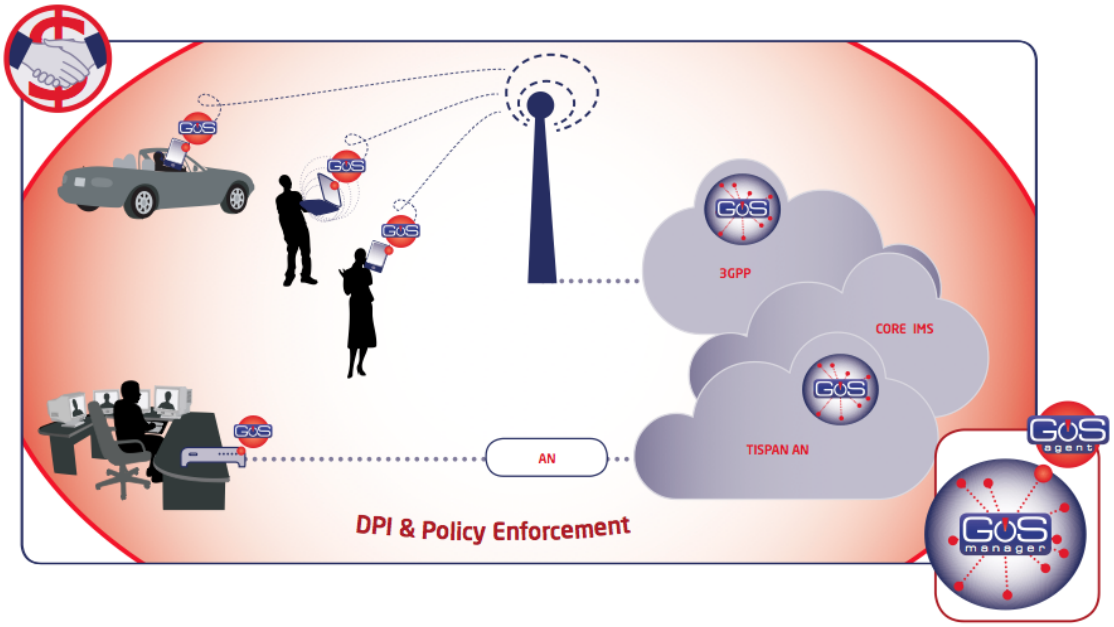
This data can be used to calculate potential bandwidth savings that can be achieved by deploying GoS in mobile devices and dongles. Calculations undertaken by GoS Networks suggest a 20% reduction in the amount of video data consumed [see Appendix 1].

In essence, GoS allows MNOs to constrain excess bandwidth that might otherwise be consumed by the client devices, freeing capacity for other users and applications. As users tend to terminate sessions prematurely, the total amount of available bandwidth is also deployed with greater efficiency.

Bandwidth Visibility at the Network Edge

For MNOs to effectively manage their networks, it is essential for them to be able to view bandwidth demand and performance at the extreme edge of the network – that is, in the devices that are actually involved in sessions, such as mobile terminals and dongles. However, current solutions – such as DPI equipment – cannot extend beyond the network boundary. Consequently, DPI products need to be complemented by software that can be deployed directly into mobile devices and dongles. GoS is a unique solution that can achieve this. With GoS, MNOs can obtain full visibility of the packets that are being sent downstream to devices, as well as the packets that are being demanded by upstream applications. This is because GoS is responsible for managing congestion on the link at the point of origin and so performs the function of selectively discarding excess packets. In turn, MNOs can determine which applications are suffering packet loss – data that cannot be obtained using DPI equipment. The effect of deploying GoS is illustrated in Figure 2.

Figure 2: Application Visibility with the Deployment of GoS



Armed with this information, MNOs can determine which applications are being utilised and by whom and to review application performance and to optimise their networks to ensure that the right applications receive the right priority. GoS also provides information about the real-time link rates at the client side, which enables control mechanisms to be enforced at the point of traffic origin or the source of demand, rather than deeper in the network. In other words, MNOs can act before the problem becomes apparent in the network core.

Collecting this information has additional benefits. For example, MNOs can be instantly alerted when quality and performance drop below expected levels. Understanding why things *do not* work is equally important for delivering successful and profitable services. Although many services do not attract incremental revenue, they can be key benchmarks for user satisfaction. Moreover, there are many new service possibilities that can generate additional revenue and their numbers are expected to grow as MNOs move from basic service packages to more differentiated offers. If a service is to capture revenue, it has to perform to the levels that consumers demand.

Another challenge for MNOs is that, as more devices switch between 3GPP and non-3GPP modes of access, such as WiFi, there is a greater risk of losing visibility of the applications that are being consumed across alternative networks. Again, GoS provides a solution. Since GoS is embedded in mobile devices and dongles, it is still able to capture valuable data regarding subscriber behaviour, irrespective of whether traffic traverses the MNO network or an alternative WiFi access point. This data can be used to enhance offers to subscribers and also promotes network optimisation.

Conclusion

In the face of continually soaring user demand, MNOs are fighting a losing battle – throwing capacity at the problem is not the solution, as it places increasing strain on CAPEX and OPEX budgets. Therefore, MNOs should explore solutions that help defer CAPEX and OPEX investment through more efficient usage of deployed capacity. However, to be effective, any solution must act at the network edge – that is, in user devices. Current solutions, such as DPI, are deployed in the network core and cannot provide visibility of user application demand. A unique solution to this problem is offered by GoS Networks.

GoS can be deployed in mobile devices and dongles to prioritise and manage bandwidth requirements for both downstream and upstream traffic, helping to reduce the total network load. By restricting bandwidth consumption to what is really necessary, GoS promotes fair usage between users in any given cell. By using bandwidth more efficiently, MNOs can defer investment in capacity upgrades in order to meet financial targets and manage CAPEX and OPEX more effectively.

In addition to efficient bandwidth usage, GoS also provides visibility of application and subscriber behaviour, allowing MNOs to create more targeted offers and achieve more consistent network performance. As MNOs seek to move from basic, “all you can eat” data plans, such knowledge can help to create valuable new revenue streams.

GoS Networks works with mobile network operators to calculate potential CAPEX and OPEX savings based on deferred network capacity upgrades. Using data that is specific to an individual network, the benefits of deploying a GoS solution in client-edge devices can be calculated, and in many cases identify the potential for significant cost savings. Contact GoS Networks now to find out how much GoS can save you.

References

Carroll, L. (1871). *Through the Looking Glass.*

Cisco. (2010). *Visual Networking Index.*

Visible Measures. (2010). *Understanding Viewer Abandonment Trends in Short-Form Online Video Content.*

Abbreviations

3GPP *Third Generation Partnership Project*

AN *Access Node*

CAGR *Compound Annual Growth Rate*

DPI *Deep Packet Inspection*

IMS *IP Multimedia Subsystem*

IP *Internet Protocol*

MNO *Mobile Network Operator*

P2P *Peer to Peer*

QoE *Quality of Experience*

QoS *Quality of Service*

TISPAN *Telecommunications and Internet converged Services and Protocols for Advanced Networking*

VoIP *Voice over IP*

Appendix 1

Parameter	Notation	Notes
proportion of traffic due to mobile dongle users	P_D	Cisco estimate this at 70% by 2014
proportion of traffic due to streaming video	P_V	Cisco estimate this at 66% by 2014
minimum bandwidth consumed by a streaming video client	B_{Min}	Typically about 440kbps
maximum bandwidth consumed by a streaming video client	B_{Max}	Upto 7.2Mbps in HSPA
bandwidth consumed by a streaming video client when controlled by GoS	B_{GoS}	E.g. 660kbpsbps: $B_{Max} > B_{GoS} > B_{Min}$
average proportion of a video clip that is actually watched	P_W	With an average clip length of 150s, based on the figures from Visible Measures, this is 64%

Then we can calculate:

The proportion of total traffic that is streaming video to mobile broadband users:

$P_{DV} = P_D \times P_V$ (= about 50% based on Cisco estimates).

When we apply GoS to control such streaming video this traffic is reduced by a fraction $\frac{B_{GoS}}{B_{Max}}$.

Thus the peak traffic load is reduced to: $T_{GoS} = T_{Original} \times \left(1 - \frac{B_{GoS}}{B_{Max}}\right) P_{DV}$.

Cisco predicts mobile data growth of 108% CAGR: in mathematical terms this means that traffic *T* is an exponential function of time, *t*:

$T = Ae^{\alpha t}$

If we measure *t* in years, then Cisco's prediction makes $\alpha = \ln(2.08) = 0.732$. By reducing traffic by a factor $F_{TR} \equiv \frac{T_{GoS}}{T_{Original}} = 1 - P_{DV}\left(1 - \frac{B_{GoS}}{B_{Max}}\right)$ then the time at which the traffic reaches a particular threshold T_{max} is delayed from time *t* to time *t'*:

$T_{max} = Ae^{\alpha t} = F_{TR}Ae^{\alpha t'}$, from which $t' = t - {\ln(F_{TR})}/{\alpha}$. Using the figures in the text gives a delay of 9.5 months.

Now consider the amount of data downloaded. The time to download a complete clip of size S at bandwidth B is $t_D = \frac{S}{B}$; and the maximum acceptable time is the duration of the clip, $t_C = \frac{S}{B_{Min}}$. The typical time at which the clip is abandoned is $t_A = P_W \times t_C$.

If $t_D \leq t_A$ then we download the whole clip before abandoning it, which wastes data. If $t_D > t_A$ then we give up before downloading the whole clip, thus saving some bandwidth on average. This condition is equivalent to $P_W < \frac{B_{Min}}{B_{GoS}}$; in this case we save a proportion $\frac{B_{GoS}}{B_{Min}} \times P_W$ of the data that would otherwise have been downloaded.

If we use the figures:
$B_{Min} = 440kbps$, $B_{GoS} = 660kbps$, then $\frac{B_{Min}}{P_W} = 1.55 > B_{GoS}$ so we meet the condition to save data, and the proportion saved is $\frac{B_{GoS}}{B_{Min}} \times P_W = 0.8$ i.e. a 20% reduction in the amount of video data consumed.