ARITARI
THE LATENCY DILEMMA

Stuart Hardy – EOH Group

INTRODUCTION

In almost every conversation I have about overcoming the issue that latency causes in global application delivery, I'm met with cynicism and outright disbelief that there is a solution for such a problem. Yet by the end of each conversation the opposite is true.

To understand the answer to the ever growing latency problem for global companies centralising applications into global Cloud infrastructures, accessed by remote branches around the world, we first need to look at the problem. Now the problem in itself is technically complex, which is why I have introduced a way to communicate it to the business side of the organisation, in a non-technical way. For the technical side refer to below references and blogs.

THE PROBLEM

Global application delivery and performance has been an issue for many years, but is relatively new and growing problem for specific regions. We'll use South African companies to illustrate this. Most South African companies have and still do, consume their business applications from inside the network or from local Cloud providers, and therefore inside South Africa. And as a South Africa application user you will probably note that when you access your company application in HQ, it's faster than when you access it from a regional office. There is no doubt that the role of bandwidth comes into play, but there is another problem which is far greater and which you will only become aware of in more extreme scenarios of application delivery. The problem of latency. So here goes:

Latency represents the time that it takes to send data (in this case an application) between one point and another (The application and the user). Ideally the lowest latency is always sought.

But not many people really understand what the effect of introducing more and more latency between these two points is, until now. As companies move their business applications from South Africa to global locations such as Azure, 365, AWS etc (over the Internet or private network) they are noticing a significant and harmful effect on user experience. And throwing bandwidth at this international problem is both unaffordable and mostly futile. That's because the issue is actually to do with latency.

It's quite simple to explain. The problem is in the TCP protocol. This protocol is responsible for supporting the transportation of your application from your user to the server. If the latency is low, the TCP Protocol determines a large session or window size for the transmission (Calculates window size in relation to latency). This translates into lots of bandwidth for a user.

But if the latency is higher, it determines a smaller window size or session which reduces the amount of bandwidth for the user. The more latency, the worse the user experience.

Here is an illustration:

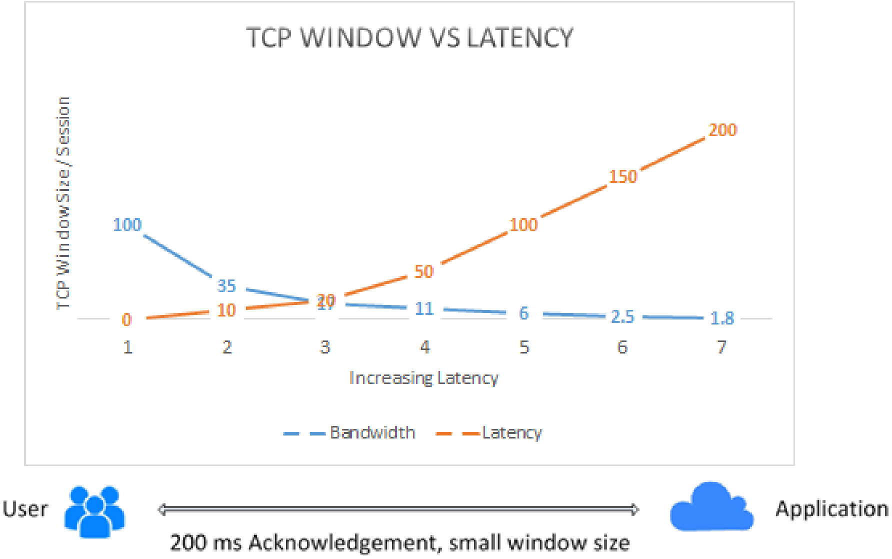
In the graph below, we have a 100 Mbps network link, used to access an application.

Lets assume there is only one user on that link to remove general congestion. When the latency is 1ms, the user session (TCP Window size) will be roughly 97 Mbps which is evident on the far left hand side of the graph. SA Cloud services are generally within 20ms of users and therefore perform well.

But what you notice as more latency is introduced between the user and the application, is that the window size reduces, as indicated in the moving from left to right on the graph. If you consider that most globally situated applications are accessed in excess of 200ms, you can note that your available bandwidth as a user will reduce to only a few Mbps or even Kbps. This has nothing to do with your provider, but is simply and purely based on the way TCP calculates the session and TCP window sized based on latency.

It's also important to understand that when TCP IP transmits data, it does so by breaking down the data or message into many smaller packets. It sends a few packets between the user and application (or two points), and then waits for a response (or receipt) from the destination before sending more. Each time this receipt takes 200ms to be received, and each time the TCP protocol then uses that information to calculate the session or window size (handshake+buffering).

Diagram 1.1 TCP IP Latency Problem



Now most people would believe that because latency is based on the physical properties of the network, in this case fibre cable system between SA and Europe, that it is not possible to resolve this issue as it's physically impossible to reduce the latency. In that determination they are correct. But the problem is not actually about the latency, the problem is about how TCP calculates session sizes using latency.

THE SOLUTION

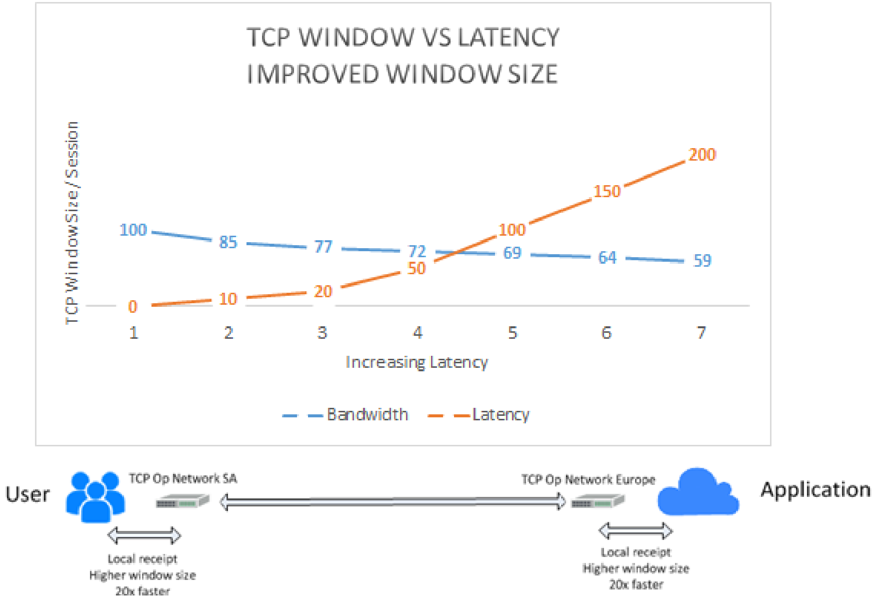
Now that you have a better view of the problem, we can explain how to resolve it using technology. This is known and marketed as TCP IP Acceleration.

The solution to high latency networks and application performance is actually as simple as the problem itself. If you want to improve the session and window size over a high latency network, you need to add a mechanism that provides the TCP Protocol an acknowledgement or receipt without having to travel all the way to the destination (200ms).

The below diagram (1.2) illustrates by passing the TCP session through localised technology, you can give a local receipt to the TCP protocol in less than 5ms. The result of this change is that the TCP protocol uses the faster receipt to calculate a larger window size which has a direct translation in session size and user experience (Perceived latency of less than 10ms).

This requires the company to place Optimisation technology in the network between these locations, or to leverage a network that has built this ability into their local POPs. In this case their POP in UK and their POP in SA would have TCP Optimisation.

Diagram 1.2 Solving TCP Latency Problem



TCP IP Acceleration is a critical consideration for customers leveraging SaaS, IaaS or PaaS on a global level. Some of these technologies are also accompanied by compression, caching and de duplication to further enhance the user experience and cost of international bandwidth.

The alternative to these options is to keep your applications in a local Cloud network in country which is challenging for global companies that have a centralised application strategy.

TCP Window Size

http://apmblog.dynatrace.com/2014/08/12/understanding-application-performance-network-part-tcp-window-size/

Latency

http://whatis.techtarget.com/definition/latency

CONCLUSION

There are a range of software providers that have developed TCP Acceleration as part of a larger and generally more expensive software stack. Only Aritari has made this service available in isolation of other capability.

Companies that offer TCP Acceleration include Riverbed, Aryaka, and Aritari.