Myths and Fallacies about Usage Based Billing (UBB)

Bill.St.Arnaud@Gmail.com

March 18, 2011

Introduction

This paper was commissioned by Netflix Inc to serve as a background document on the issues related to what is called Usage Based Billing (UBB) that is now being implemented by many telephone and cable companies as part of their Internet service offerings.

This paper is not intended to addresses the specific issues of UBB as it pertains to the recent "wholesale UBB" ruling and upcoming hearings by the Canadian Radio and Telecommunications Commission (CRTC) but to the more general issue what are the true costs of delivering streaming content over the Internet, particularly over the cable-TV and telephone last mile local network infrastructure.

It is important for regulators and policy makers to understand the mechanics behind (UBB) and whether streaming video and other services imply an actual true increased cost for the service providers. More importantly it will be useful to understand whether or not infrequent or occasional users of the Internet are in fact subsidizing heavy users and if there is any true cost correlation with usage.

UBB has been proposed as an economic "Internet traffic management practice" intended to reduce network congestion by controlling heavy users rather than using technical processes to achieve the same effect. It has also been suggested that UBB is necessary to provide telephone and cable companies with revenues and incentives to make sufficient investment in their broadband infrastructure in order to meet the future tsunami of data that is predicted to be delivered over the Internet.

In this paper we will demonstrate three important facts:

(a) Internet video streaming services actually **reduce** costs for Internet backbone networks operated by telephone and cable companies, even as traffic volume grows;

(b) There is no correlation between volume of Internet consumption and costs for telephone and cable company last mile providers and that congestion, if any, is more of an artifact of design assumptions made by the operators; and

(c) Cable and telephone companies operate competing video streaming services over the same last mile infrastructure used for Internet access services, which generally are not priced based on usage, and yet somehow seem to able to avoid congestion as well as provide the service for fraction a price of what they charge for delivery of the same video content delivered over the Internet.

These three facts call into question the whole rationale and validity of the need for UBB. More importantly, in other jurisdictions outside of Canada, Internet service is often provide without UBB and so one must question what is the real purpose of UBB in Canada?

The former monopoly cable-TV and telephone companies control the local wireline infrastructure duopoly and they offer on-demand regulated and unregulated video streaming services over this last mile broadband infrastructure that is also used to provide competitive Internet access. As well, several of these firms have extensive investments in broadcast content. This suggests that UBB may serve other purposes for these incumbent network operators. UBB can be a means for cable and telephone companies to recoup any decline in revenues as users shift to Internet services. It can also serve to constrain the competitiveness of unaffiliated on-line video streaming services by increasing end-user costs for those services. When passed-on at wholesale, UBB also greatly reduces the pricing flexibility of independent ISPs, making them less effective competitors. Moreover, and paradoxically, UBB puts downward pressure on demand for Internet services which reduces the need for, rather than creating an incentive for, network investment.

What is Usage Based Billing (UBB?)

UBB is a pricing mechanism implemented by many service providers that provides a number of tiered services where a customer is allowed to download or upload a certain amount of Internet traffic to specified maximum “cap”. If the user exceeds the cap, measured in Gigabytes (GB) of data then a surcharge is imposed of $x per GB for each additional Gigabyte of data the user consumes in a given month. The number Gigabytes of data used in the cap and in the surcharge is the sum of both the upload and download data transfer.

For example Rogers Cable UBB plan [ROGERS] typical low end offering costs $27.99 per month which would entitle the user to consumer up to 2GB of data per month. The user is charged an additional $5 for every Gigabyte of data they consume that exceeds the 2GB cap. The next tier of usage is priced at $35.99 per month with a cap of 15 GB and a surcharge of $4 per GB for data volumes that exceed the cap. Each successive tier has progressively higher monthly prices with larger caps and smaller per Gigabyte charges for data volumes that exceed the cap.

The streaming of one movie typically is around 1 GB of usage. The streaming of one high definition movie typically consumes 3 GB of data. Of course actual data volumes may very widely depending on the length of movie. A cap of 2GB data per month is not a lot of data and can easily be consumed through routine use of e-mail and web browsing.   The bottom line is that costs at least $5 to download an average movie through Roger’s lowest price offering, for the data transmission alone. This does not include any costs that might be applied by Netflix or Hulu to view the movie itself.

UBB is a mechanism that is often justified by telephone and cable companies to ensure fairness amongst users. The argument often used is that light Internet users should not be subsidizing heavy network users. On the surface, this may seem like a reasonable and rational pricing policy – but it is based on a big presumption that network costs increase with consumption, or that heavy users in some way disenfranchise light users from their fair share of the network. As we shall see in the next section this assumption that costs are related to usage is not necessarily true especially with respect to the Internet.

UBB is also often justified by network operators as an economic traffic management practice to control congestion on the network. In fact, regulators have encouraged cable and telephone companies to institute economic measures to control usage such as charging consumers rates based on how much bandwidth they use each month, or offer discounts during off-peak hours [CRTC]. Regulators believe that this is a more effective and less intrusive mechanism than technical means to manage traffic, such as traffic shaping, The belief is that usage based billing is the most transparent traffic management practice because it is clearly identified on monthly bills. With this information, consumers can compare between different internet services and match their bandwidth needs with the amount they are willing to pay.

Not every service provider implements UBB. Some service providers advertise “unlimited” downloading and uploading of data up to the maximum possible amount of data that can be delivered over the network. In reality “unlimited” can mean a number of different things where intrinsic or deliberate design limitations in the network prevent users from consuming data anywhere approaching the theoretical capacity. Often “unlimited” means there is an actual physical limit to the amount of data that can be consumed, but the user is not assessed any additional charges beyond the basic monthly fee for the data they download.

In other cases, service providers announce a “throttling” cap, at which point they will only throttle or slow down transmission speeds of the data transfer for an user who has exceeded their cap in the event of congestion on the network. This is done to allow users with smaller data flows who have not exceeded their caps to get fair access to the network. No additional charges are applied to either the heavy or light user.

Definition of duopoly and oligopoly

Throughout this paper we refer to the telephone and cable company duopoly or oligopoly based on the definitions used by the FCC and US Department of Justice using the Herfindahl Index [FCC]. The Herfindahl- Hirschman Index (HHI), is a measure of market concentration the degree to which a few firms control the pricing and output in a market. A market’s Herfindahl Index is the sum of the squares of the market shares of each firm in the industry. For example, if there are only four firms in a market and two firms each have 30 percent of market sales, and the other two firms each have 20 percent of market sales, then the Herfindahl Index is:

$30^2 + 30^2 + 20^2 + 20^2 = 2600$

This market would be considered highly concentrated by the antitrust division of the U.S. Justice Department, which considers any market with a Herfindahl Index of less than 1800 to be competitive.

In North America where most communities are serviced by only the cable and telephone company the HHI ranges from 5000 to 10,000, which ranges between duopoly and monopoly. It should be noted that although there may be multiple competitors a monopoly situation can still exist as measure by the HHI.

Some basic concepts about consumption and usage

Charging a consumer based on usage or consumption seems to be elementary and a straight forward concept. But in reality there are many products and/or services which we use every day where we are not charged by the amount we consume.

Traditional broadcast radio over the air is a good example. Radio stations do not charge listeners for the amount of time they spend listening to the radio. In fact they encourage listeners to spend as much time as possible tuned to their station. The reason that a listener can spend as many hours as he or she wants is that no listener deprives any other from the same ability to listen to that same station. One single listener can spend as much time, or as little time, as he or she wants listening to a favorite station without impacting anyone else's ability to do the same thing.

Economists refer to this as "non-rivalrous" consumption or usage. In essence non-rivalrous consumption means that my use or consumption of a product or service does not impede or deprive anyone else from enjoying that same product or service. Watching broadcast TV or listening to a radio are good examples of this type of consumption.

On the other hand there are many products or services that are deemed "rivalrous" in that my consumption or usage of a product or service prevents or deprives others from using that same product or service. The consumption of most physical goods such as food, gasoline, etc is considered rivalrous consumption.

Rivalrous and non-rivalrous are the two bookends of consumption and usage. Clearly there are many products or services that are combination of both. But generally most products and services tend to predominate in one end of the spectrum or the other and their pricing reflects that nature of the product. The pricing for non-rivalrous products or services is rarely usage based, which would be clearly counterproductive as in the case of most TV or radio. Over-the-air TV and radio earn their revenues selling audiences to advertisers so audiences pay nothing.

Broadcasting distributors (both telco and cableco) bundle TV channels for sale to subscribers at prices that have little to do with either the bandwidth they occupy or the number of viewers they attract. The exceptions are pay-per-view and on demand channels, where copyright fees, rather than UBB for bandwidth consumed, drives usage-based pricing. Yet it is interesting that these same network operators seek to charge by the byte for access to all Internet services.

Although pricing for rivalrous products tend to be usage or consumption based, another important factor on whether a product is priced based on its volume of consumption is to look at its production costs. If the production costs are relatively low, some clearly rivalrous products have prices that are not based on consumption. A simple example is the free grocery bag you get when you purchase goods at a grocery store.

Economists breakdown production costs into two different basic categories: high capital with low operating cost versus low capital with high operating costs. A good example of the former is a nuclear power plant where bulk of the costs that must be passed onto consumers is not from inputs such as the purchase of uranium or labor but the interest on the loans required to build the plant. Conversely an example of a production with low capital but high operational costs is a fast food restaurant where virtually all the costs are food supplies and labor.

In a capital intensive industry like nuclear power the amount of output such as electrical power is largely independent of the operational costs. As such in these types of industries it often makes sense to maximize production output regardless of market demand in terms of usage or consumption. And this often happens where nuclear power plants or steel mills will dump product on the market and drive down prices because it is cheaper to keep the plant operating rather than shut it down, regardless of the price of the final product. A good example of this recently was in Ontario Canada, where an excess of power from nuclear power forced the operator to pay customers to consume the power. Of course in non-intensive capital industry like food services it makes much more sense to limit output to match demand.

The big challenge for capital intensive industries, as opposed to those with high processing costs is deciding when to invest in new production capabilities in order to meet demand. In non-intensive capital industries like food service increased demand can be simply met by increasing purchase of raw materials and hiring more labour. For capital intensive industries, investors must be assured there is sufficient cash flow to pay for the interest on the capital over the coming years. As mentioned before these interest costs are often the largest component of pricing that is passed on to consumers. Often, these investment costs, including interest, are passed on through prices linked to consumption so the consumer has the impression that there is a direct correlation between consumption and costs. While this may be a convenient pricing model for the supplier, in reality there is very often little correlation between consumption and the cost to deliver the product. This is the case with the Internet.

Is the Internet a rivalrous or non-rivalrous service?

Telecommunications is a capital-intensive industry with significant up-front costs and relatively low operational costs. Traditionally telecom products were considered highly rivalrous. In the old days of telephone circuits which consumed a fixed amount of bandwidth, a single phone call consumed resources that were not available for others to use throughout the duration of the phone call. Voice telephony was clearly a rivalrous service. Telephone companies therefore had to invest significant capital to ensure there was enough capacity to handle a reasonable number of phone calls at any given time. For example, historically, call volumes on Mother's Day risked bringing the network down.

The Internet changed all that. The Internet is based on a technical concept called "statistical multiplexing" where multiple users can share the same bandwidth which previously would be consumed by one user making a single phone call. In theory, with "statistical multiplexing" there is no limit on how many users can share the bandwidth represented by the circuit for a single phone call. (Practically there are some limitations which will be discussed later). But almost over night, thanks to the Internet, telecom bandwidth changed from a highly rivalrous service to one that is highly non-rivalrous. A further, and related, characteristic of telecom networks is that, unlike other utilities such as water, electricity, gas or oil, where the product is actually "consumed" by end users, network bandwidth is never truly "consumed": bandwidth is infinitely re-usable and therefore only temporarily in use (or "consumed") at any given time.

This fundamental change in bandwidth usage brought about by Internet technologies revolutionized the telecom industry and led to the dramatic drop in prices we have seen over the past several years. But not only did the Internet change the nature of consumption – from dedicated to multiplexed usage – it also radically reduced the operational costs of delivering telecom services (although full production costs include investment in infrastructure which remains up front capital intensive).

The traditional telephone network needed very expensive technology in order to set up and complete end to end telephone circuits. The Internet on the other hand uses packets to transmit and does require the complex end to end circuits communicating information. Instead each node in the Internet, called a router, only needs to know the location of the next router in the forwarding path in order to transit data. The deployment of routers and Internet networks is significantly less costly than traditional telephone and cable technologies. The question that needs to be asked is whether the capital costs have dropped so significantly that other mechanisms to pay for the infrastructure can be found?

Even if we conclude that Internet is largely non-rivalrous with low operating costs, the telephone and cable companies still need to be compensated to undertake the up front capital costs to deploy the last mile infrastructure. As well, given the anticipated ongoing exponential growth in Internet the telephone and cable companies need to be provided with incentives to continue to invest in the necessary infrastructure to meet the demand and to replace ageing equipment. The question is whether usage based billing (UBB) the most appropriate mechanism for doing this?

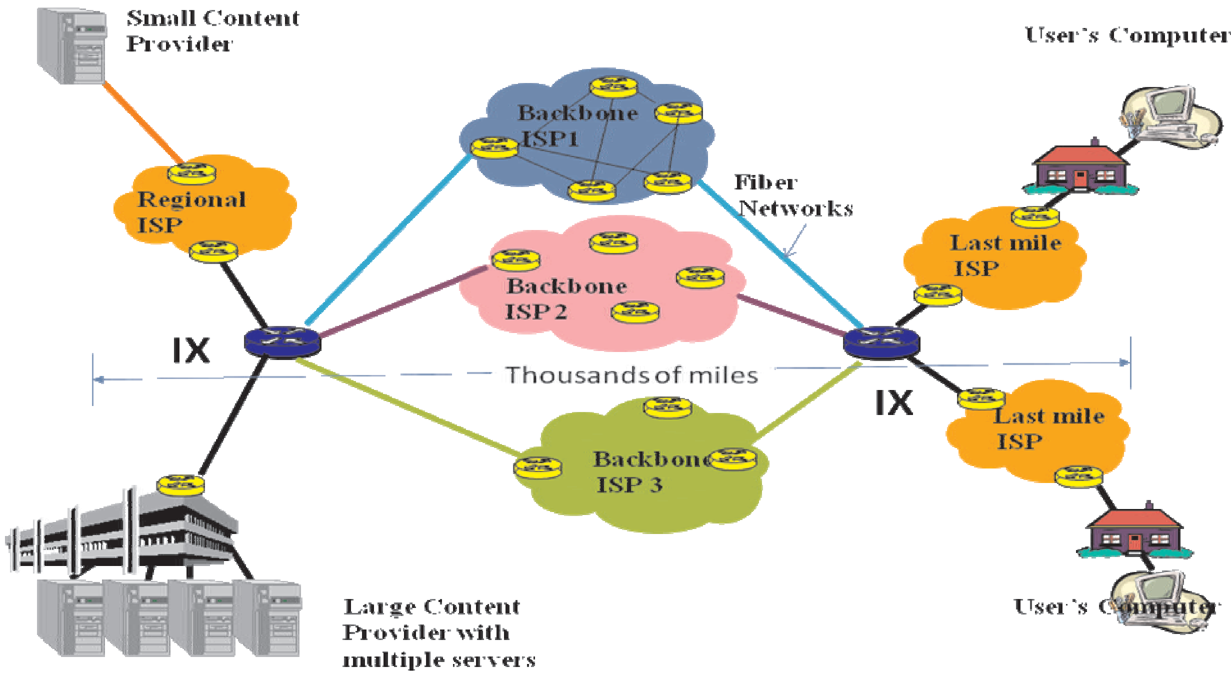
To delve into this question further we need to understand some of the basic components of the Internet, especially related to the delivery of streaming video to the home consumer.

It should be stressed that the Internet is not the only mechanism for steaming movies to users. The cable and telephone companies also have their own proprietary systems for streaming movies such as on demand digital cable and IPTV services. These services use virtually the same technology and last mile infrastructure as that used by the Internet to deliver except that they have the added cost of requiring the network operators to also host and store the content locally. As we shall see later they also consume more of the bandwidth on the last mile infrastructure than that used by Internet based streaming services and are often given segregated capacity and/or a higher priority. Interestingly, although these services use the same last mile technology and infrastructure they are priced completely differently from how the operators charge for the Internet and never employ UBB for bandwidth consumed.

Evolution of the Internet and content distribution

The Internet famously is a "network of networks," but many do not stop to better understand what exactly that means. The "networks" most visible to people are the on-ramps and off-ramps to and from the rest of the Internet -- namely, the local broadband communications links that enable users to reach all the Net's content and applications and other services. Consumers get access to the Internet through last-mile broadband providers, but how do Internet application and content companies make their services available online? These companies invest many billions of dollars to have their services distributed and hosted throughout the Internet. These investments, for the most part, far exceed those made by the telephone and cable companies in their own infrastructure.

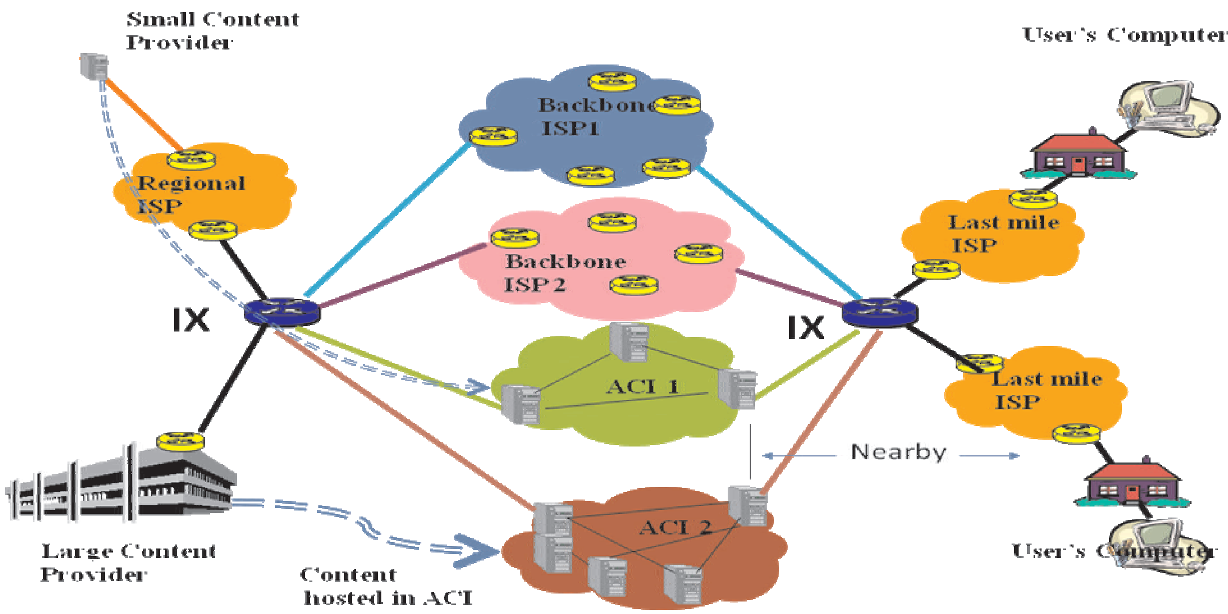
The "textbook" example of content and application hosting and routing is still employed today by regulators and policy makers which often leads to misunderstandings on how the costs of delivering content are allocated across the entire network. Traditionally content and application companies deploy servers to host their content in a single location and then connect to a telephone or cable company Internet Service Provider (ISP), who carries the traffic across the global Internet to the end-user. The following diagram illustrates this arrangement:



The original Internet was made up of several different components and identified as "last mile", "regional Internet Service Provider (ISP)", Internet Exchanges (IXs) and "backbone ISPs. The link from the user's home or residence is identified as the last mile and is usually provided by the local telephone or cable company. In many cases these days the last mile telephone or cable provider is also the regional ISP and/or backbone ISP. From the telephone or cable company ISP's hub or integration point the Internet traffic is carried to a local or regional IX. IXs are located in just about every major city in North America. It is at the IX where the regional ISP exchanges traffic with one or more national backbone carrier ISPs. The backbone carrier ISPs carry the traffic to another IXs elsewhere in the county or around the world.

Historically large content providers would need to deploy tens if not hundreds of servers on their premises in order to meet the demand for their content or applications. Even smaller providers would need a number of servers if they wanted to provide any reasonable level of service. In either case, the content or application provided would also have to over provision their server capacity in anticipation of periods of heavy demand. But as a result there were often long periods of idle capacity.

As new multimedia content such as video and network applications have evolved, so too have the forms of delivering the content. Broadcast and content providers are now taking advantage of a constellation of new infrastructure, including data centers, content distribution networks (CDNs), distributed computing and storage services (aka "clouds"). This infrastructure is then connected with the public Internet at major Internet Exchange (IX) points to the last-mile ISPs, rather than carried across the Internet through backbone providers as shown in the following diagram:



Last-mile broadband telephone and cable network operators are still the unique gateway between users and the rest of the Internet -- that still has not changed. However, what is changing is the way that broadcast and content companies' services are hosted and served before they reach the last-mile provider's network. The big difference these days is rather than the content and application company hosting the servers they are instead distributed across the Internet and located much closer to the end user.

There are a wide variety of content distribution networks (CDNs) that offer many different services to third-party application and content providers. For instance, Akamai, Limelight, Voxel, CDNetworks are among the many providers of CDNs and other services. In addition, some large companies like Google and Amazon have built and operate their own infrastructure to deliver these services.

One principal driver behind the use of CDNs to host and serve content and applications from a location that is close as possible to end-users. For example, video content providers take advantage of CDNs to host and "cache" popular content, so that it is available from many distributed locations. This increases responsiveness and decreases latency, due to the fact that the data packets have less distance to cover and less networks to traverse. In addition, CDNs can be further optimized by deploying their services deep within the last mile provider's network to improve performance relative to delivery through backbone and provider's internal networks. Of course, this improved performance helps provide a better user experience.

Most significantly the CDN networks reduce costs for telephone and cable network operators because the traffic is not carried on their backbone or internal networks. Instead the traffic from a content provider is carried over the CDN's private network to the closest IX nearest to the customer or right to the last mile operator's headend or central office, and in some cases to the network node closest to the consumer.

How CDN's improve delivery of broadcast content

One of the original and still critically important drivers for CDNs is to improve the user's Internet experience by reducing latency and enhance application responsiveness. As most congestion occurs in the last mile networks operated by the telephone and cable companies the impact of such congestion is significantly reduced by moving content and applications as close as possible to the consumer. Keeping data local and thereby reducing transmission time can significantly enhance a users' perception of the quality of the responsiveness of an application or quality of a broadcast – even under congestion and packet loss. As long as there is sufficient bandwidth to prevent egregious congestion in the last mile virtually all new innovative content and applications can be delivered over a best effort last mile network if they originate from an CDN. For example, only a short time ago it seemed unimaginable that one could deliver HD quality movies over the Internet – but this is done routinely today by many services delivered using CDN, such as Hulu, Netflix, and YouTube.

On the Internet, each time an application on an originating computer sends a packet of data to a distant user it must be acknowledged by the recipient's computer. (The recipient sends a return packet – called an "ACK" – before another packet from the originating computer is sent. If the originating computer does not receive this acknowledgement packet within a specified time it assumes it has been lost somewhere in the network and it retransmits the original packet.) This simple "handshake" protocol ensures that no packets are lost in transmission across the network.

Packet loss also signals to the originating computer there may be congestion on the network. Presumably the original packet got dropped on the floor somewhere where there was too much data for the network to handle. If you look closely at Internet traffic you will see it is made up of billions and billions of these types of data and acknowledgement packets back and forth across the network.

If the originating computer does not receive an acknowledgement packet within a specified time frame not only does it retransmit the original packet it also immediately slows down its transmission rate. Then very slowly the originating computer speeds up the transmission of packets as long as it continues to receive ACKs for every transmitted packet. But if at any time it does not receive an ACK it immediately throttles back the transmission of packets.

The time to reach maximum transmission speed is very important and is directly related to the distance between the originating and receiving computer – the greater the distance the longer it takes to reach maximum transmission speed. There are a number of other factors that may control transmission speed, but a basic rule of thumb is that for every doubling of distance, the time to reach maximum speed of transmission is reduced by half. This speed reduction is necessary because greater distances mean it takes a much longer time to send packets and receive ACKs.

However if there is a single packet drop then transmission is immediately reduced by up to 50% and then slowly allowed to ramp back up. It is obvious that the probability of packet loss will increase the greater the distance between the originating and recipient computers as there are many network links and nodes to traverse – any one of which could be briefly overloaded and cause congestion.

Packet loss can occur anywhere along a network from the user's home, in the last mile and/or across the carrier backbone network. However numerous studies indicate that most congestion occurs in the last mile.

Regardless of where the congestion and packet loss occurs in the network the effect is immediate in terms of the responsiveness of the application as perceived by the user. The data transmission rate is immediately reduced which is further aggravated if the originating and recipient computer are a great distance apart. As applications incorporate more multimedia elements such as images and video the impact can be quite dramatic in terms of the speed of which the data is received. The same application hosted on a distant server will take orders of magnitude longer to download than if it is delivered locally given typical last mile congestion and packet loss. Internet users experience this phenomenon every day.

To understand the impact that CDNs are having on the evolving Internet one only has to look at a recent study by Arbor Networks [ARBOR]. Arbor Networks – in collaboration with University of Michigan and Merit Network presented the largest study of global Internet traffic since the start of the commercial Internet at recent Internet engineering conference in October 2009. The study included analysis of Internet traffic across 110 large and geographically diverse cable and telephone operators' networks, international end-to end carrier backbones, regional networks and content providers. Results from the study show that over the last five years, Internet traffic has migrated away from the traditional Internet core and now flows directly between large content and broadcast providers and last mile networks (and then to their end-users).

CDNs have also significantly reduced the cost of Internet infrastructure of telephone and cable companies by adopting new network architectures that simplify and reduce the complexity of the traditional backbone ISP facilities. Because CDNs provide load balancing and redundancy through the wide distribution of servers at various nodes around the world they do not need complex and highly redundant network facilities between these nodes.

As a result the costs of Internet transit have dropped dramatically for all the parties in the Internet eco-system. Major media and content now often deliver their content directly to third party CDNs at major IXs where most traffic is exchanged on a settlement free basis. Telephone and cable companies have then been able to reduce the amount of Internet transit they purchase from international backbone Internet providers by directly connecting to a CDN at an IX. Backbone ISPs have also benefited in that their costs are also significantly reduced as they can also access the same traffic delivered by CDNs rather than carrying it on their backbones.

Many CDNs have also extended their content networks deep within last mile provider's own networks, at no cost to the last mile provider. Again this is done for similar reasons of improving user experience by bypassing any congestion that might occur in the middle mile between the provider's DSLAM or CMTS and their head end. This is particularly important for wireless applications. **It also relieves the last mile provider the necessity of upgrading their own internal network to meet increased demand for video streaming and other applications.**

Since the CDN traffic is delivered on a settlement free basis there is no cost to the telephone or cable company to receive this traffic. **More importantly as traffic volumes grow the telephone and cable company Internet backbones do not incur any additional costs. So, to the extent that cable and telephone companies are, *in fact*, experiencing levels of Internet traffic congestion that need to be managed by economic or technical measures, this congestion appears to be an artifact in the duopoly local wireline broadband network infrastructure.**

An overview of the architecture of the last mile infrastructure

The last mile infrastructure is the main battleground for debates on UBB, congestion control, network neutrality and other issues. Most Internet consumers in North America are only serviced by a duopoly of cable or telephone wireline local network infrastructure for Internet access. Only a small percentage of consumers have access to high speed fiber or wireless services. But even in these situations there are rarely more than one or two facilities based providers.

Not only is the infrastructure provided by cable and telephone companies used to provide Internet access service it also carries other specialized services such as voice, cable TV or IPTV and video on demand services. Although these latter services share the same infrastructure as that used to deliver Internet access, they are entirely under the control and management of the last mile network operator. As opposed to the Internet, third party application and service providers cannot deliver services over the segregated infrastructure capacity used for these services without first getting permission (which is rarely given) by the telephone or cable company. In most cases this "parallel infrastructure capacity" is reserved for the exclusive use of the telephone or cable company to deliver their proprietary applications and services to the consumer.

Currently the two primary applications delivered over the telephone and cable proprietary infrastructure are telephony – Voice over Internet Protocol (VOIP) – and video on demand streaming services. In cable parlance this proprietary video streaming service is known as Video on Demand (VOD) and on telephone networks it is often referred to as Internet Protocol TV (IPTV).

The delivery of these propriety services as well as the Internet is done in a similar manner for both the telephone and cable companies, the only difference being the underlying enabling technology. On cable systems the enabling technology is the partitioning of the cable spectrum into parallel circuits called QAM (Quadrature Amplitude Modulation) channels. On telephone systems the enabling technology is PPPoE (Point to Point Protocol over Ethernet).

QAM channels on cable systems are generally of fixed capacity and cannot be shared amongst many users or different services such as VOD or Internet and are therefore considered rivalrous. PPPoE on the other hand uses statistical multiplexing and therefore the DSL bandwidth can be shared amongst different applications, but the telephone company usually employs techniques to guarantee bandwidth to different services or tags VOD traffic with a higher priority on the same DSL service.

Telephone and cable companies both deploy what is called a trunk and branch architecture where the trunk is usually a high speed fiber network to nodes scattered throughout the serving area. At the nodes the high speed optical signal is converted to an electronic signal for delivery over the last mile infrastructure – coaxial cable in the case of cable companies and copper wire for telephone companies. It is at these nodes where the cable and telephone companies usually locate Customer Aggregation Equipment (CAE) to convert the signals carried on the copper or coax infrastructure into digital optical signals and vice versa.

At the CAE a variety of technologies are deployed to distribute data on both the proprietary VOD systems used by the last mile providers as well as for delivery of Internet traffic. In the telephone system the same technology is usually used for both services and it is called a Digital Subscriber Line Access Multiplexer (DSLAM). In cable systems generally separate technologies are used for the cable operator's proprietary system versus delivery of Internet traffic. In the latter case a Cable Model Termination System (CMTS) is used to aggregate and convert Internet traffic from the coaxial cable facility to the fiber trunk back to the head end.

The cable propriety system for streaming video is far less standardized. Different operators use different technologies. In general operators use a combination of switched video and proprietary caching architectures for their VOD services which in reality look very similar to the CDN architectures used in the Internet by Netflix, Hulu and others to stream video.

In both the telephone and cable systems, the last mile operator plans its distribution capacity based on the number of digital homes served. Because QAM channels are rivalrous and cannot be shared amongst consumers the calculation of the number of channels to assign to VOD or Internet can make a big difference in terms of congestion and contention for either service. In general a cable network operator aims to assign enough QAM channels to ensure a very low contention or congestion rate on its proprietary VOD service. The exact congestion or contention ratio is usually a safely guarded secret.

Most cable operators have 40 to 60 digital QAM's. Usually only 5 QAM channels are reserved for Internet services (DOCSIS 3.0) and the rest are earmarked for their VOD and regular broadcast services. With the increasing demand for more HDTV channels there is incessant pressure to assign more QAM channels. To address this need cable operators are moving many of their less popular specialty channels to the VOD/caching service described previously thus freeing up additional QAM channels. These channels can be either assigned to the operator's Internet service to reduce Internet congestion or to the operator's private VOD service. There is very little, or no capital cost to reassign these QAM channels.

Telephone IPTV systems have more bandwidth constraints because of the limitations of DSL. As a consequence far fewer telephone companies have deployed IPTV in comparison to the widespread deployment of VOD by the cable companies. AT&T's Uverse is a good example of large scale IPTV deployment. AT&T will reserve up to 7 Mbps on the DSL circuit for this service, the balance being allocated to Internet and VOIP services. But again the telephone company, as with the cable company, has considerable flexibility in the assignment of various PPPoE channels and their priority to support various services. If congestion occurs on a given service additional bandwidth or priority can be reallocated on that service at very low or negligible cost, understanding, in general, DSL has less overall capacity than cable and therefore may hit capacity limits sooner.

However, as we shall see in the next section, both telephone and cable network operators have other tools in the toolbox for addressing congestion that require little capital investment.

Internet congestion management by cable and telephone last mile providers

Telephone and cable companies have often complained that they need both economic and technical tools to handle large volumes of traffic and periodic congestion. Video streaming services such as Netflix, Hulu, etc are increasingly singled out as the primary culprit of traffic growth and congestion.

While it is convenient to blame “heavy” users for this type of traffic, in fact, most of the elements that affect congestion are under control of the local network operator and have little to do with network volume or usage. So while they complain they need economic tools like UBB to control congestion and utilization, surprisingly they do not apply these same mechanisms to their own proprietary video streaming networks.

One of the particular tools of congestion management that are in hands of the cable operators is the number of QAM channels that are assigned to Internet versus VOD for cable companies and in the case of telephone companies the ratio of bandwidth allocated to IPTV versus Internet. As we have seen in the previous section both last mile operators allocate far more spectrum to their in house services such as VOD, VOIP and IPTV then they do to Internet. Up to 60 QAM channels may be assigned to these specialty services versus only 5 for delivery of the Internet.

Some of this allocation is responsive to broadcasting regulatory requirements, including priority carriage of Canadian services; the author has no way of assessing to what extent this is in fact the case. To be clear, this paper takes no issue with such requirements. Rather, the focus is on the efficient application of network management practices in relation to proprietary services and public Internet services, and the impact that these practices, notably UBB, may have on the Internet ecosystem.

The other primary cause of congestion on last mile networks is oversubscription ratios, particularly in reference to Internet services. Without awareness of subscription ratios one cannot fully understand the nature of the problem, and whether traffic billing techniques may appear to be reasonable and fair.

The last mile providers have succeeded in justifying UBB as an alternative to altering their oversubscription ratios, or in the case of cable allocating QAM channels in large part because of a lack in transparency in these ratios for either their Internet or proprietary service offerings. Without such information, customers are not aware of the extent to which their internet provider is able to deliver the bandwidth that it advertises or if it is favoring bandwidth allocation to its own proprietary network versus its Internet offering. Conversely, to the extent that proprietary broadcasting services are being favoured, it is impossible to know whether this realistically reflects regulatory requirements or may be an indirect means to influence the market for Internet access and online services.

Both cable and DSL network architectures manage traffic in essentially the same way. Although any of the network interconnection points mentioned above can be sources of congestion, the most frequent congestion control point is located at the DSLAM or CMTS.

Typically, one or more users will be aggregated at one port on a DSLAM or CMTS and under heavy traffic conditions this will lead to congestion at that CAE. In addition to customer inbound traffic congestion at the CAE point, there can be congestion on the CAE outbound port, which connects to the headend or central office.

There are several causes of congestion as described above. The amount of traffic generated by users through any application, including downloading video, is certainly a factor. However, a more significant factor is the telephone and cable company practice of "oversubscription" (i.e., selling more bandwidth that it can actually provide at any given time). The CAE uses statistical multiplexing and as such the traffic is largely non-rivalrous and therefore many users can share the same bandwidth. However this is practical only up to a certain point where significant congestion occurs to be noticeable to the end user. Congestion does occur quite often at the CAE, but rarely noticeable by the user.

This key factor underlying traffic management is often under-emphasized, largely because of the lack of transparency surrounding it. It is common practice for last mile operators to sell more aggregate bandwidth to their customers at the CAE than their networks are capable of handling. This is, to some extent, efficient and acceptable, because no customer uses their full allotment of bandwidth at all times. However, if oversubscription ratios are too high, congestion results. (The ratio between how much bandwidth a telco/cableco sells to its customers and how much it actually provisions for in its network is referred to as the telco/cableco's "oversubscription ratio".)

A typical example of an oversubscription ratio would be as follows: Assume that 50 customers in a neighbourhood share a single CMTS or DSLAM port, and that each customer has been sold a 1 Mbps service and has the physical capacity to send 1 Mb of traffic towards the CAE port in any given second. This arrangement would potentially allow for 50 Mbps of inbound traffic to that CMTS or DSLAM port. However, since it is unlikely that all 50 customers will be using their connection at full capacity at the same time, the ISP may only provision that CMTS or DSLAM port so as to handle 10 Mbps at any one time without congestion occurring. This would result in an oversubscription ratio of 5:1.

If the total traffic generated by the 50 customers at any given time is generally less than 10 Mbps, than the last mile operator has set a reasonable oversubscription ratio. If, however, the 50 customers regularly produce 15 Mbps, then congestion will occur at that CAE, and may possibly impact on the experience of the 50 customers.

Generally, this type of congestion problem is easily managed. The last mile operators regularly measure utilization at any given link in their network as a proxy for congestion. For each such link on its network, a telco/cableco will have calculated a provisioning threshold. This threshold is typically based on the level of utilization a given link experiences, which in turn is presumably based on an estimate of what would lead to an unacceptable level of congestion at that link.

In theory, once the provisioning threshold of a link is reached, the telephone or cable company responds to reduce congestion on that link so as to ensure that the 50 customers connected to it do not experience inferior service. This response typically involves an expansion of capacity at the link in question or reducing the number of subscribers on a given port. Another solution is to encourage CDN companies to colocate their caching and content distribution boxes next to the DSLAM or CMTS. There is virtually no additional cost the last mile operator other than perhaps a one time small port charge to add an additional CAE port and perhaps add incremental bandwidth to the upstream bandwidth from the CAE to the head end. Alternatively, cable operators can reassign QAM channels, or free up additional QAM channels from their analog broadcast service, again with little or no additional capital cost.

Many of the telcos/cablecos claim that the growth in video stream traffic is such that it would require unreasonable amounts of investment to respond to congested links with provisioning alone. This claim is impossible to validate without data on oversubscription ratios and targeted congestion levels, or to understand why they have not encouraged collocation of CDN services in their CMTS or DSLAM nodes.

However it should be noted that the last mile providers routinely change the over subscription ratios and upstream bandwidth on their proprietary VOD and IPTV services without passing on the charges to the consumer. They often also re-architect their network to meet demand, for example by converting less popular digital specialty channels to VOD. In many cases they also install their own caching boxes for their own video streaming services at neighbourhood nodes which further alleviates congestion on the their proprietary services. Again this is done without any incremental or usage based charges to the end consumer. All these solutions are equally applicable to their Internet service delivery.

If oversubscription ratios and bandwidth assignment comparisons between the operator proprietary services and the Internet thresholds were made a matter of public record, customers would be able to properly compare and make informed choices among ISPs. **This would likely force ISPs to compete on provisioning, and thus push them towards network expansion, or enabling deep deployment of competitive CDN services, instead of UBB as a means of addressing growth in traffic. It would help consumers make more informed choices among ISPs. It would also help consumers, and regulators, to determine if the last mile operators are engaging in anticompetitive behaviour by using UBB to give competitive advantage to their proprietary services and/or to make the services of independent ISPs and online services less competitive.**

In Canada, the monopoly cable and telephone network operators that control the local broadband duopoly wireline network infrastructure are also regionally dominant in the broadcasting distribution, local telephone, Internet access and mobile telephone markets. Several of these dominant players also are heavily invested in broadcast and/or production facilities – notably

Bell Canada, Shaw, Rogers and Quebecor/Videotron. Services and content delivered over the Internet, including through video streaming, are in direct competition with this aspect of their business.

This situation inevitably gives rise to potential conflicts of interest. As operators of the duopoly local broadband bottleneck facilities, these companies have considerable incentive and, as has been shown, both the technical and the pricing capability to engage in anticompetitive discriminatory practices. Little, if any evidence has been filed on the public record to support the claims of cable and telephone companies that UBB is an efficient or effective means to manage the network congestion that it purports to address. No clear causal relationship has been established between UBB pricing and bandwidth usage. This paper calls into question both suggestions.

Given the potential for conflict of interest between the telephone companies' and cable companies' proprietary, regulated services and independent services delivered via the public Internet, consumers and regulators need to be vigilant in questioning practices such as UBB and whether they address their stated purpose. The CRTC's net neutrality framework distinguishes between technological Internet Traffic Management Practices (ITMPs) and economic ITMPs. While the CRTC concluded that the latter may have some advantages, it is clear that both are capable of serving anticompetitive ends that undermine competition, consumer choice and even investment in the infrastructure needed to meet growing Internet traffic.

Network Costing of Last Mile Networks

One of the most thorniest issues is determining the real costs of delivering Internet bits over last mile networks and what incremental costs, if any, are incurred when delivering new services and applications like streaming video.

As discussed previously the Internet is largely a non-rivalrous technology which provides lots of flexibility in terms of architectural solutions and costing models. For example many applications there is often a trade-off between deploying distributed caching or content distribution boxes versus increasing bandwidth. Most application and content providers have chosen the former approach as the more cost effective solution as alternative to purchasing Internet transit. This not only reduces their costs but that of the last mile provider as well. Deploying these types of CDN services deep into the last mile provider's network reduces the last mile provider's costs, but also implies a significant competitive threat to the last mile provider's own VOD and IPTV services. Often last mile provider's defer deploying such services (even if there is no cost to them) because of the threat they represent to their own proprietary services**. In fact Comcast in the United States charges CDN providers to install their caching and content distribution**

boxes deep within their cable network. It is rumoured that Canadian last mile providers are looking to adopt this strategy as well.

The International Telecommunications Union has done an extensive analysis of the costs of a DSL network [ITU]. Broadband Internet networks tend to be very capital intensive with relatively small operating costs [OPS]. It should be noted operational costs are declining on a per bit or per subscriber basis with each new generation of technology.

The operating costs of the equipment are largely power consumption, maintenance, management overhead, etc. The way that most equipment is designed there is little variation in either power consumption, maintenance or management overhead regardless of the number of bits that traverse the last mile network. **So even though video streaming or other applications may increase the volume of traffic it has little or zero impact on the cost of operating the network.** It is only at a point where frequent congestion occurs that the network operator actually might incur a cost to upgrade the network or implement other solutions.

The biggest network cost is the amortization of the physical plant made up of switches, customer premise equipment, broadband routers etc. From the ITU data we see the ADSL costs are approximately 150 Euro ($200) per subscriber using existing copper plant, while next generation DSL (ADSL 2) is about double that price at 400 Euro ($550) for a net difference of $350. Cable DOCSIS deployments tend to significantly cheaper averaging about $100- $140 per subscriber to upgrade from DOCSIS 2.0 to DOCSIS 3.0 service [CISCO]

The biggest challenge in coming up with any costing data is to get data on the mix of these services in Canada, and understand the appropriate amortization and depreciation schedules used by the operators. As well the underlying copper or duct infrastructure is shared with other services such as traditional voice telephony and cable-TV. Only a portion of that cost can be allocated to Internet service. However, assuming services like Netflix due cause congestion and the last mile network operator decides to upgrade the network, as opposed, to alternate cheaper solution as outlined in this report then the cost per bit is still relatively trivial.

If we make a worst case assumption of only a one year depreciation and an annual download of 350 GB with a historical annual average increase in bandwidth consumption of 30-40% (150 GByte) then the additional cost per Gigabyte for on result costs in a cost increase of about $.30 for GByte for DSL service and $.10 per GB for cable. This is significantly less than what the last mile providers are charging under their UBB programs. And it is important to note that these costs only apply where the operator may be experiencing congestion.

UBB as tool for congestion management

One of the often cited reasons for UBB is that can be used as an economic tool to throttle heavy users and avoid congestion on the network. While UBB may indeed change the behavior of heavy users to reduce traffic volumes it may not have any impact on congestion issues.

Congestion is a momentary phenomena where the number of packets entering or exiting the network exceed the capacity of the network much like traffic jams occur on expressways at rush hour. Charging drivers an extra monthly fee if they drive an excess number of kilometers per year does not guarantee that these drivers will still not use the freeway at rush hour.

Unless UBB contains a time of day billing feature and some immediate feedback on congestion it is hard to imagine how it can be used as a congestion management tool. The only exception is the case where all drivers are charged an exorbitant price for mileage which would result in an overhaul reduction of using freeways.

UBB as a general usage suppression tool may be the intended purpose of such high UBB fees and a clear incentive to push consumers into higher priced bands.

Conclusion

Understanding the complexity of relationship between usage and cost is fraught with many challenges. Most of us, who have grown up in a world of physical products understand there is a clear relationship between consumption and costs. But as we move into the future information society, dominated by network-based services, this traditional relationship between consumption and costs starts to break down.

Because most services are virtual and consume very little physical material or energy they can be shared by many users at the same time. The digitization of all sorts of traditional physical products such as movies, books and music is revolutionizing the way we think of their costs. Young kids who download music have probably intuitively grasped this concept long before us older supposedly wiser adults. In this age of the Internet and digitization consumption of a service such as listening to a song over the Internet does not deprive anyone else from listening to that same music. In the old days when music was delivered by a physical medium such as a CD or vinyl album, one needed that physical manifestation in order to enjoy that music, which deprived anyone else from enjoying that same music unless they had the same physical representation.

With video streaming we are still seeing regulators and policy makers locked into antiquated thinking about goods delivered in the physical world where there was a clear relationship between consumption and costs. While some small vestige of that phenomenon still exists in the digital world it is almost negligible in comparison. Also inapt are comparisons between broadband networks and older utilities like water, electricity, gas and oil in which products are physically "consumed" by end users to the exclusion of others.

These analogies betray notions of scarcity and exclusivity that simply do not apply in the Internet ecosystem. Rather the broadband Internet is characterized by abundance, non-exclusive access and use, and rapid innovation in services and network capacity. Beyond the local broadband duopoly wireline bottleneck, these characteristics are driven by dynamic innovation and robust competition. Regulators should take note.

If there are genuine problems with traffic volume and congestion because of the downloading of streaming videos from Netflix and similar services then last mile network operators have a suite of tools for dealing with the problem, which are trivial to implement. The fact that they regularly implement these solutions for their proprietary video streaming services without passing costs to the consumer through UBB, underscores the message that perhaps UBB applied against Internet consumption is being used for some other purpose.

Given the fact that the Internet consumes so little bandwidth on cable or telephone networks, because it is largely a non-rivalrous technology, compared to the proprietary rivalrous video streaming solutions deployed by the cable or telephone company itself. Yet, in terms of bandwidth allocation, it is sold at significant premium to those same services – on top of which are applied surcharges for bandwidth overage that are not clearly related to cost.

Moreover as we have shown video streaming services such as Netflix, Hulu and others delivered through CDN networks actually reduce the costs of Internet transit and last mile networks for the telephone and cable company. Any inordinate bandwidth consumption or congestion can easily be handled at virtually no cost by reallocating the bandwidth or channels assigned from the operator's proprietary video streaming service or enabling deep deployment of competitive content distribution services.

In this context, the effectiveness of UBB in dealing with the issues it purports to address, and its impact on the end-to-end Internet value chain demand close examination. This practice reduces consumer demand for broadband services, may undermine competitors and may create disincentives for network investment. So who benefits from UBB? The significant risk is that UBB is really a mechanism to protect the market dominance of former monopoly local network operators and to help them leverage that dominance into the Internet ecosystem.

References

[ARBOR] C. Labovitz, et al "ATLAS Internet Observatory 2009 Annual Report" http://www.nanog.org/meetings/nanog47/presentations/Monday/Labovitz_ObserveReport_N47_Mon.pdf

[CISCO] DOCSIS 3.0 TCO Analysis http://www.ciscoknowledgenetwork.com/cable/documents/3CKN_4-22_Show_DOCSIS_3-0_LRP_and_TCO_Analysis-A_Cisco_Value_proposalv3.pdf?utm_source=&utm_medium=&utm_campaign=&PRIORITY_CODE=194542_20

[CRTC] http://www.cbc.ca/news/technology/story/2009/10/20/crtc-net-neutrality-ruling.html

[FCC] Herfandahl Index and Telecom Competition https://www.fcc.gov/14report.pdf

[ROGERS] Rogers Cable Price Plans https://www.rogers.com/web/link/ptvBrowsePackagesFlowBegin?forwardTo=landing

[ITU] Cost Analysis of DSL networks http://www.ictregulationtoolkit.org/en/PracticeNote.aspx?id=2899

[OPS] Operating vs Capital cost comparison http://www.itu.int/ITU-T/worksem/asna/presentations/Session_2/asna_0604_whitepaper_brouse.doc

[STREAMCENTER] Cataltepe, Z. Moghe, P., Characterizing nature and location of congestion on the public Internet, 30 June-3 July 2003, p 741- 746 vol.2, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1214206