INTERNATIONAL TELECOMMUNICATION UNION

ITU WORKSHOP ON
What rules for IP-enabled NGNs?

Geneva, 23-24 March 2006

Document: NGN/02
23 March 2006

INTERCONNECTION IN AN NGN ENVIRONMENT

BACKGROUND PAPER

Advanced draft to be presented for comments

March 23, 2006

NOTE

This draft paper has been prepared by J. Scott Marcus (Senior Consultant, WIK Consult, Wissenschaftliches Institut für Kommunikationsforschung, Germany, <S.Marcus@wik.org >) to be presented for comments at the ITU New Initiatives Programme workshop on "What rules for an IP-enabled Next Generation Networks?" held on 23-24 March 2006 at the ITU Headquarters, Geneva. The final version of the paper reflecting the comments will be made available at the event's web site in April 2006**.** The views expressed in this paper are those of the authors, and do not necessarily reflect those of the ITU or its membership.

This paper, together with the others relevant for NGN debate and prepared under ITU New Initiatives Programme, can be found at http://www.itu.int/osg/spu/ngn/event-march-2006.phtml. The New Initiatives Project on "What rules for IP-enabled NGNs?" is managed by Jaroslaw Ponder <jaroslaw.ponder@itu.int> under the direction of Rober Shaw <robert.shaw@itu.int>.

ACKNOWLEDGEMENTS

I have had the good fortune to work with many outstanding professionals. The engineers at GTE Internetworking/Genuity were superb, and taught me a great deal about interconnection in the IP-based world. I have also benefited enormously from my associations with prominent economists, including Jean Tirole, Patrick Rey, the late Jean-Jacques Laffont, Donald Stockdale, William Sharkey, Simon Wilkie, David Sappington, Patrick de Graba, Stephen Littlechild, and Justus Haucap. For the paper at hand, my WIK colleague Dieter Elixmann provided a very thoughtful and careful review, and Robert Shaw and Jaroslaw Ponder of ITU's SPU provided helpful guidance and feedback.

TABLE OF CONTENTS

page

TABLES

FIGURES

Executive Summary

This report considers the likely evolution of interconnection arrangements in the context of IP-based *Next Generation Networks (NGNs*).

The NGN represents a synthesis of existing world of the "traditional" *Public Switched Telephone Network (PSTN*) with the world of the Internet. The economic and regulatory arrangements for the two have historically been very different. What should happen when these two worlds collide?

Many of the networks created over the past ten years contain most of the key elements of an NGN. Most, if not all, of the technology necessary for IP-based NGN interconnection has been available for five to ten years. Advanced approaches to interconnection have been slow to deploy, even where the technology has been mature or within hailing distance of maturity.

The NGN interconnection problem is best understood, not as a problem of technology, but rather as a problem in economics. With that in mind, this report seeks to review what is known about interconnection from an economic perspective, in order to reach conclusions about the prospects for deployment going forward and the corresponding implications for policymakers.

A substantial body of economic theory has been developed over the past decade as regards interconnection in the traditional PSTN. A smaller body of solid economic research has emerged in regard to interconnection of IP-based networks. At the level of economic theory, the PSTN and the Internet are *not* worlds apart. Economics provides the necessary bridge between the two worlds, illuminating both the similarities and the differences in these two environments.

This report begins by laying out, for the most part at a non-technical level, the established theory of interconnection, for both the PSTN and the Internet. Wholesale and retail arrangements are considered separately. Most of the observed behavior of these economic networks can be explained in terms of a constellation of known economic effects: market power, the termination monopoly, demand elasticity, network exernalities, transaction costs, service differentiation, price discrimination, and the Coase theorem (which says that private parties can often negotiate arrangements more efficiently than government regulators, provided that necessary preconditions have been met).

With this theory in hand, the report considers the implications for the deployment of differentiated Quality of Service, and of universal service. We also consider the implications of IP-based technology – with the layering, and the changes in industry structure that it implies – service providers become independent of the network, but neither is well equipped to measure or to charge for the other's resource consumption.

The last section of the report represents a hypothetical scenario, a "thought experiment", where the historic wired and mobile incumbent of European country upgrades its networks to an IP-based NGN. We consider the likely results in terms of regulation of the access network, and of interconnection; likely domestic and international interconnection arrangements; and the implications for ubiquitous support of QoS. Key findings include:

- Provided that markets for Internet transit and for consumer broadband Internet access are effectively competitive, a "Coasian" interconnection regime is likely to be more efficient, and more consistent with consumer welfare, than a regulated regime.
- Conversely, where these markets are not effectively competitive, mandates for interconnection at the IP level may prove to be unavoidable, particularly once existing PSTN interconnection is withdrawn. The migration to NGN potentially creates new sources of market power, at the same time that it creates new possibilities for competition.
- Policymakers might consequently be well advised to focus their attention first on ensuring competitive markets, and only secondarily on interconnection.

Current *Calling Party's Network Pays (CPNP*) arrangements contain a number of implicit subsidies. In the world of the NGN, where services providers and networks operators may be different entities, these subsidies need major re-thinking – call termination payments that were intended to finance the terminating network would, by default, flow to independent VoIP service provides who have no network to support. In

the absence of termination fees, independent VoIP providers would tend to compete price levels for telephony service, independent of the network, down to levels not greatly above cost, which would appear to be a societally desirable outcome.

The thought experiment does not flatly preclude the possibility that governments might somehow erect a new system of subsidies to replace the old, but it suggests that any subsidy system will be difficult to sustain over time in the face of new forms of competition enabled by the IP-based NGN – all provided, once again, that underlying markets (especially for wholesale Internet transit and for retail Internet broadband access) remain effectively competitive. A system of Coasian private arrangements, in the absence of vertically integrated competitive bottlenecks, seems likely to lead to unsubsidized arrangements at wholesale and retail price levels not greatly in excess of cost.

1 INTRODUCTION

The English novelist Charles Dickens has a series of ghosts show his miserly and misanthropic protagonist, Scrooge, his past, his present, and a grim future. The chastened Scrooge then asks, "Are these the shadows of the things that Will be, or are they shadows of things that May be, only?"[1]

This report considers the problem of network interconnection in the emerging world of the IP-based NGN from the perspective of established economic theory, and then attempts to "paint a picture" of what might happen if the primary wired and wireless incumbent in a major European country were to migrate rapidly and comprehensively to an IP-based NGN in the near future. It is hoped that this thought experiment sheds light on the likely evolution of interconnection in the evolving NGN world; at the same time, it is important to remember that it depicts *one possible* future, hopefully a plausible future, but not necessarily *the* future.

1.1 The migration to IP-based Next Generation Networks (NGNs)

The global electronic communications industry is experiencing something of a "sea change" as it is integrated to an increasing degree with IP-based services. The plans of British Telecom (BT) to replace outright large parts of its existing over the next few years with a 21st Century Network (21CN) are perhaps the most dramatic example,[2] but the same trend is proceeding, perhaps more quietly, in every developed country. In North America, there is less of the rhetoric of the NGN, but much of the same substance.

1.2 To regulate, or not to regulate?

This migration raises many thorny regulatory questions, especially in the area of network interconnection. The *Public Switched Telephone Network (PSTN*), the existing telephony network, operates under a well established set of interconnection rules that have been more than a century in the making. In the Internet, by contrast, interconnection is generally a matter of private bilateral agreements, usually with no regulatory intervention at all. Both systems seem to work reasonably well most of the time in their respective domains, but how should they be combined?

Inevitably, there have been calls to withdraw regulation altogether. As the number of technical alternatives increases, and competition progressively expands, the regulation of electronic communications should wither away altogether.

In the long run, this is probably the right view. Regulatory best practice argues for withdrawal of regulation once markets have become effectively competitive.

But the long run view may not be the most relevant view. As the English economist John Maynard Keynes remarked, "In the long run, we're all dead." This report focuses on events in an intermediate time frame – the next few years, or perhaps at most the next two decades.

Over that time frame, concerns must be raised over complete withdrawal of regulatory obligations in markets where competition is not yet fully effective. The experience of New Zealand, which attempted for years to avoid putting a traditional communications regulator in place, is particularly relevant – their system proved to be unworkable. In fact, the most serious problems were precisely in regard to interconnection, which is the locus of this report. Starting around 2001, they gave it up as a bad job, and implemented lightweight institutions approximating the function of a traditional regulator.[3]

The scenario analysis in this report suggests that the overarching philosophy that the U.K. regulator, Ofcom, has adopted is much more promising: the focusing of reglulation on areas where there are *durable competitive bottlenecks*, enabling competition at the *deepest level feasible*; and the gradual withdrawal of regulation everywhere else.[4]

1.3 NGN core, NGN access

The migration to Next Generation Networks can be viewed as comprising two distinct threads. On the one hand, current PSTN operators are evolving the *core* of their networks so as to use IP-based technology to carry voice traffic, and other applications as well. On the other, many firms are providing increasingly high speed data *access* to the customer premises.

In a recent document,[5] the European Competitive Telecommuncations Association (ECTA) provided definitions that will serve for purposes of this report:

- The first is the deployment of fibre into the local loop, either to the incumbent's street cabinet (+/- max 1km from the customer premises) in conjunction with VDSL(2) deployment or the deployment of fibre all the way to customer premises (typically apartment blocks rather than individual houses). These will be referred to as *access NGNs*.
- The second is the replacement of legacy transmission and switching equipment by IP technology in the core, or backbone, network. This involves changing telephony switches and installing routers and Voice over IP equipment. These will be referred to as *core NGNs*.

These two threads have somewhat different regulatory implications. In this report, our primary focus in on the NGN core. The adoption of broadband access is very much relevant to this migration, and in this sense the migration to the access NGN can be viewed in regulatory terms as simply being faster broadband.

1.4 A word about the author

I should also say a few words about my own background. We all have a tendency to look at issues through the lens of our own experiences. Before starting work at the WIK, a research institute and consulting firm located in Bad Honnef, Germany, I had been the Senior Advisor for Internet Technology at the FCC (U.S.). Prior to that, I was the Chief Technology Officer (CTO) for GTE Internetworking (Genuity, also U.S.), which at the time was one of the largest Internet backbones in the world.

I am well aware that these issues are complex and contentious. My long experience working with the Internet, with the FCC, and generally in North America inevitably predisposes me toward a Bill and Keep intercarrier compensation model; at the same time, I am reasonably well versed in theory and practice in Europe. The perceptive reader will quickly observe that my personal views on these matters do not strictly follow the lines on which these arguments typically proceed. I have attempted to present the issues and the full range of arguments as clearly and as fairly as I could, and to ground my statements clearly in established economic theory and in documented facts. Only the reader can judge how well I have succeeded.

I should add that, while I know something about economics, I do not regard myself as an economist. I am an engineer by training. Nonetheless, I took an economic perspective in this report, because the interconnection challenges with which this report deals are best understood from that perspective.

1.5 A road map to the balance of the report

The next three sections of the report provide general background drawn from economic theory. Section 2 provides interconnection theory, both for the PSTN and for the Internet. Section 3 provides technical and economic background of differentiated service (IP Quality of Service), and of associated price discrimination. Section 4 talks about market power – its sources, its remedies, and its likely evolution in the world of the IP-based NGN. Section 5 is a brief exploration of the relationship between interconnection arrangements and the funding of universal service in an NGN context. Section 6 considers the interaction between interconnection arrangements and interconnection accounting – what can be measured in an IP-based NGN, and how do measurement constraints translate into constraints on what can be charged for? Finally, chapter 7 uses a hypothetical scenario of an NGN migration in Europe to explore how interconnection arrangements might in practice evolve.[6]

2 UNDERLYING ECONOMIC PRINCIPLES

This section provides background on the underlying economics of network interconnection, in order to motivate the discussion that follows. It attempts to present the economics of the PSTN and that of the Internet in an integrated way, and also to provide a consistent view of the various models that have emerged at the retail and at the wholesale levels.

The interconnection of telecommunications networks has been extensively studied in the literature. Many economists would view the authoritative sources as being Laffont, Rey and Tirole (1998a and 1998b),[7] Armstrong (1998),[8] and Laffont and Tirole (2001).[9] I choose to draw primarily on Laffont and Tirole (2001).

The section seeks to provide non-specialists with a non-technical but thorough grounding in the theory and the literature.[10] It also serves to introduce the economics vocabulary that will be used throughout the balance of the paper. Economists may find this section useful primarily to the extent that it provides a comprehensive and integrated view of what is known of interconnection arrangements in the PSTN and in the Internet.

2.1 The PSTN at the Retail Level

Retail arrangements in the world of conventional telephony are, in a sense, familiar to anyone who uses a telephone. Nonetheless, it may be helpful to put them into a broader perspective, in order to provide a comparative context. Most of us live in a single country, and have only limited exposure to alternative arrangements.

2.1.1 Calling Party Pays (CPP) versus Mobile Party Pays (MPP)

In most countries, the party that *originates* (initiates) a call pays a fee for the call, usually as a function of the duration of the call in minutes, and often also as a function of the distance from the originator to the point at which the call *terminates* (is received). In these same countries, the party that receives the call typically is not charged. These arrangements are collectively referred to as *Calling Party Pays (CPP*).

A few countries – notably, the United States and Canada – use an alternative system referred to as *Receiving Party Pays* (*RPP*). Under RPP, the originating party and the terminating party can each be charged by their respective service providers.

In the U.S. and Canada, CPP arrangements are common for fixed line phones, while RPP arrangements are common for mobile phones. For this reason, some experts prefer to refer to these North American arrangements as *Mobile Party Pays* (*MPP*).

In fact, the system in these countries continues to evolve – the most common arrangements today are for plans that are either *flat rate*, or that are flat rate up to some large number of minutes of use (so-called *buckets of minutes* plans).

Each of these systems has its advantages and its disadvantages, and each has adherents and opponents. Both are in need of a major re-thinking as the world evolves to IP-based NGN arrangements.

2.1.2 Cost Causation

CPP calling arrangements have long been the globally most common set of arrangements. They are extremely logical if one starts from the presumption that the party that originated a call presumably wanted the call to complete, and that the originating party can therefore be considered to be both the prime beneficiary and the *cost-causer* of the call.

Analogously, the receiving party has been thought of as a passive party, involuntarily receiving a call from the originator. Again, under this assumption it is natural to refrain from charging the receiving party.

More recently, a number of economists have challenged this view. The American Patrick deGraba has argued that, "… both parties to a call – i.e., the calling party and the called party – generally benefit from a call, and therefore should share the cost of the call." [11]

A recent paper by Doh-Shin Jeon, the late Jean-Jacques Laffont, and Jean Tirole explores the inherent mirror-image relationship between calling and called party, and find that there is no qualitative difference, as "it takes two to tango." In particular, they consider the implications of *receiver sovereignty* – the notion that

the receiver always has the option to hang up, and therefore should be viewed as playing an equal or nearly equal role in cost causation.[12]

2.1.3 Usage-based pricing versus flat rate

Consumers appear to have a strong preference for flat rate retail pricing arrangements over usage-based pricing. Flat rate arrangements reduce or eliminate the uncertainty as to what the consumer will have to pay.

Customers tend to respond to flat rate plans by making extensive use of the service in question. In an economic sense, this is a normal and predictable demand elasticity response to a perceived marginal price of zero.

If the marginal usage-based cost to the provider were high, this might lead to inefficient use; however, communications services today are characterized to an ever-increasing degree by significant initial costs and low or very low usage-based marginal costs. Under these circumstances, flat rate plans can be efficient for both the consumer and the provider. The high utilization of the service that flat rate promotes can thus be viewed as a gain in consumer welfare.

The U.S.-based mathematician Andrew Odlyzko has argued that pricing structures will tend to gravitate to flat rate whenever the marginal cost is low enough, and purchases frequent enough: "People react extremely negatively to price discrimination. They also dislike the bother of fine-grained pricing, and are willing to pay extra for simple prices, especially flat-rate ones. …[P]rice discrimination and finegrained pricing are likely to prevail for goods and services that are expensive and bought infrequently. For purchases that are inexpensive and made often, simple pricing is likely to prevail."[13]

Flat rate plans are common in the United States, but much less common outside of North America, largely as a function of differences in the underlying wholesale interconnection arrangements – we return to this point in the following section of this paper. Experience in the U.S. strongly bears out the consumer preference for flat rate services.

For example, AT&T Wireless's offer of Digital One Rate in 1998 provided flat rates across the United States. As long as the mobile customer used not more than some fixed (and possibly large) number of minutes of air time, the customer could place or receive calls to and from any point in the continental United States. The customer would incur no per-minute charges, no long distance charges, and no roaming charges.[14]

Digital One Rate proved to be immensely popular. The success of Digital One Rate effectively forced its mobile competitors to provide a competitive response; however, initially they were hampered by their lack of nationwide scale. The net result was a wave of consolidation, alliances and joint ventures that ultimately resulted in a nationwide market for mobile telephone services with multiple carriers, each offering nationwide plans offering a large bucket of minutes for a flat monthly fee.

Today, flat rate plans are becoming increasingly prevalent in the U.S. for all forms of telephony.[15] As dominant local operators were permitted to offer long distance services, they typically offered flat rate plans with unlimited domestic long distance. IP telephony service providers commonly offer unlimited domestic calls at a flat rate.[16]

Analogously, when America Online introduced flat rate pricing of $19.95 per hour for Internet service in 1996, it resulted in an explosion of consumer adoption – so much so, that the company was hard-pressed to deploy new service quickly enough.

At the level of governmental policy, both the U.S. and the U.K. have implemented measures to enable consumers to avoid per-minute charges when using dial-up to access an ISP.[17] These measures are motivated by the same recognition that true usage-based incremental costs are low, and that the societal value and consumer welfare benefits of increased utilization of the Internet are probably substantial.

2.2 The PSTN at the Wholesale Level

Charging arrangements for the PSTN at the wholesale level mirror the arrangements at the retail level, but only loosely.

The most common arrangement by far is often referred to *calling party's network pays* (*CPNP*). In a CPNP regime, the call receiver's operator assesses some predefined charge per minute to the caller's operator for termination. The call receiver's operator pays nothing.[18] Given that, under a pure CPP retail regime, the receiving party does not pay for the call at all at the retail level, the prevailing view has been that the calling party's *network* should compensate the receiving party's *network* (i.e. the terminating network) for its costs with a payment at the wholesale level.

Bill and Keep, by contrast is a United States term of art that denotes the absence of a regulatory obligation to make payments at the wholesale level. Carriers could conceivably choose to voluntarily negotiate compensation arrangements at the wholesale level, but in general they are not motivated to do so.

Most countries use CPP at the retail level, and CPNP at the wholesale level. Indeed, wherever CPNP is practiced with relatively high per-minute termination fees (e.g. in excess of several cents per minute), the use of CPP at the retail level tends to follow as an economic consequence.

By contrast, only a few countries use Bill and Keep, and they tend to use it selectively. The United States, for example, is CPNP for call to fixed incumbent operators,[19] but is generally effectively Bill and Keep for mobile-to-mobile calls and for calls from one non-incumbent fixed provider to another.[20] France used Bill and Keep for mobile-to-mobile calls until 2004, generally with satisfactory results.

Bill and Keep wholesale arrangements make flat rate retail plans possible, but they do not preclude other arrangements at the retail level.

2.2.1 Calling Party's Network Pays (CPNP) versus Bill and Keep

As has been previously noted, a very extensive literature exists on wholesale call termination arrangements in general.[21] A number of papers specifically address the relative merits of CPNP wholesale arrangements in comparison with Bill and Keep.[22]

There is some tendency in the literature to use the terms CPP and CPNP interchangeably, but this can lead to confusion. For our purposes we define CPNP in terms of *wholesale* payments between operators. CPP, by contrast, relates to *retail* payments from end-users to their operators. CPP and CPNP are often found together, but not always. The wholesale arrangements do not invariably dictate the retail arrangements, nor *vice versa*.

2.2.2 The termination monopoly

CPNP termination leads to a problem that is known as the *termination monopoly*. When you attempt to place a call to someone, you may have a number of choices as to how to originate the call, but in general you have no control over how the call is to be terminated – in general, a single operator is able to terminate calls to any given telephone number. This confers a special form of market power on the terminating operator – hence, the term termination monopoly.

The termination monopoly operates even in markets where competition for call origination is effective, and is by no means limited to large players that have market power on the call origination market. Laffont and Tirole speak of "… the *common fallacy that small players do not have market power and should therefore face no constraint on their termination charges.* … A network operator may have a small market share; yet it is still a monopolist on the calls received by its subscribers. Indeed, under the assumption that retail prices do not discriminate according to where the calls terminate, *the network has more market power, the smaller its market share*; whereas a big operator must account for the impact of its wholesale price on its call inflow through the sensitivity of its rivals' final prices to its wholesale price, a small network faces a very inelastic demand for termination and thus can impose higher markups above the marginal cost of terminating calls."[23]

Consequently, and in the absence of regulation, operators will tend in general to set their termination prices well in excess of marginal cost, and at levels that are also well above those that are societally optimal.[24]

The high termination fees can lead to large economic distortions where regulation is asymmetric. For example, the general practice in Europe prior to 2003 was to limit wired incumbent operators to termination fees based on marginal cost plus a reasonable return on capital; mobile operators, however, generally had unregulated termination rates. This resulted in European mobile termination rates that were an order of magnitude greater than fixed termination rates, and in very substantial subsidization of mobile services by

customers of fixed service. A number of economists have argued that these transfer payments constitute an inappropriate subsidy from fixed to mobile services, and a massive economic distortion.[25]

The European Union can be said to generally subscribe to this analysis. Since 2003, the European regulatory framework for electronic communications has in effect treated the termination monopoly as an instance of Significant Market Power (SMP) that national regulators must deal with. In the absence of mitigating factors, all operators – large and small, fixed and mobile – will tend to be assumed to possess SMP. As a result, mobile termination prices have declined somewhat, and are likely to continue to do so in most if not all Member States of the European Union.[26]

Under a Bill and Keep regime, the terminating monopoly problem does not arise. Interconnected operators generally have the opportunity under Bill and Keep to voluntarily negotiate interconnection prices other than zero; however, experience with mobile operators and with non-dominant wired operators (CLECs) in the United States, with[27] mobile operators in France prior to 2004, and with Internet backbones suggests that interconnection prices in the absence of a regulatory mandate will most often be voluntarily set to a price of zero.[28]

2.2.3 The relationship between wholesale intercarrier compensation and retail prices

If traffic is balanced between two operators, and if they were to charge identical termination fees to one another, then there would be no net payment between them. This is true whether the termination fees are low or high. Since termination fees do not change net payments under these conditions, there may be a temptation to think that termination fees do not matter very much.

Laffont and Tirole refer to this as the *bill-and-keep fallacy*. "It is correct that a change in the access charge need not affect the (absence of) net payment between the operators, but the access charge affects each network's perceived marginal cost and therefore retail prices. It is, therefore, *not* neutral even if traffic is balanced."

Each operator views its payments to other operators as a real cost. Other things being equal, operators will tend to be reluctant to offer service at a marginal *price* below their marginal *cost*. For on-net calls – calls from one subscriber of a network to another subscriber of the same network – operators can and often do offer lower prices that correspond to the operator's real costs.[29] For *off-net* calls (calls to a subscriber of another network), however, it is unusual to see retail prices below a "high" wholesale call termination rate,[30] *even where termination payments are likely to net to zero*. This probably reflects the operators' understandable fear of *adverse selection* – if they set their retail price for off-net calls too low, they may attract too many of precisely those users whose calling patterns are such as to cause them to place more off-net calls, thus generating a net payment (an *access deficit*) to other operators.[31]

2.3 Retail prices, subsidies, adoption, and utilization

As we have seen, high termination fees tend to lead to high retail prices for placing calls. (Under CPP retail arrangements, there is no charge for calls that are received, whether termination fees are low or high.) In particular, high call termination rates preclude flat rate or buckets of minutes plans at the retail level. As we might expect, the higher marginal prices at the retail level tend to depress call origination – this is the well-known phenomenon of *demand elasticity* (or the *price elasticity of demand*). As the price of some good or service goes up, we will prefer to purchase less of it if we can.

The American economist Patrick de Graba described these relationships succinctly in a widely read FCC white paper[32]:

> One source of inefficiency is that existing termination charges create an "artificial" per-minute cost structure for carriers that will tend to result in inefficient per-minute retail prices. In unregulated, competitive markets, such as the markets for [mobile telephony] services and Internet access services, retail pricing is moving away from per-minute charges and towards flat charges or two-part tariffs that guarantee a certain number of free minutes. This suggests that few costs are incurred on a per-minute basis, and that flat-rated pricing will lead to more efficient usage of the network. The existing reciprocal compensation scheme, which requires the calling party's network to pay usage sensitive termination charges to the called party's network, imposes an "artificial" per-minute cost structure on

> carriers which, if retail rates are unregulated, will likely be passed through to customers in the form of per-minute retail rates. Such usage sensitive rates thus would likely reduce usage of the network below efficient levels.

DeGraba also notes that "…[t]he ISP market illustrates the importance of rate structure on usage. When AOL changed from usage sensitive rates to a flat charge for unlimited usage in late 1996 the number of customers and the usage per customer rose dramatically and other competitors soon followed. … Similarly, the introduction by [mobile operators] in the United States of pricing plans that include 'buckets' of minutes appear [sic] to have contributed significantly to the growth in wireless usage."[33]

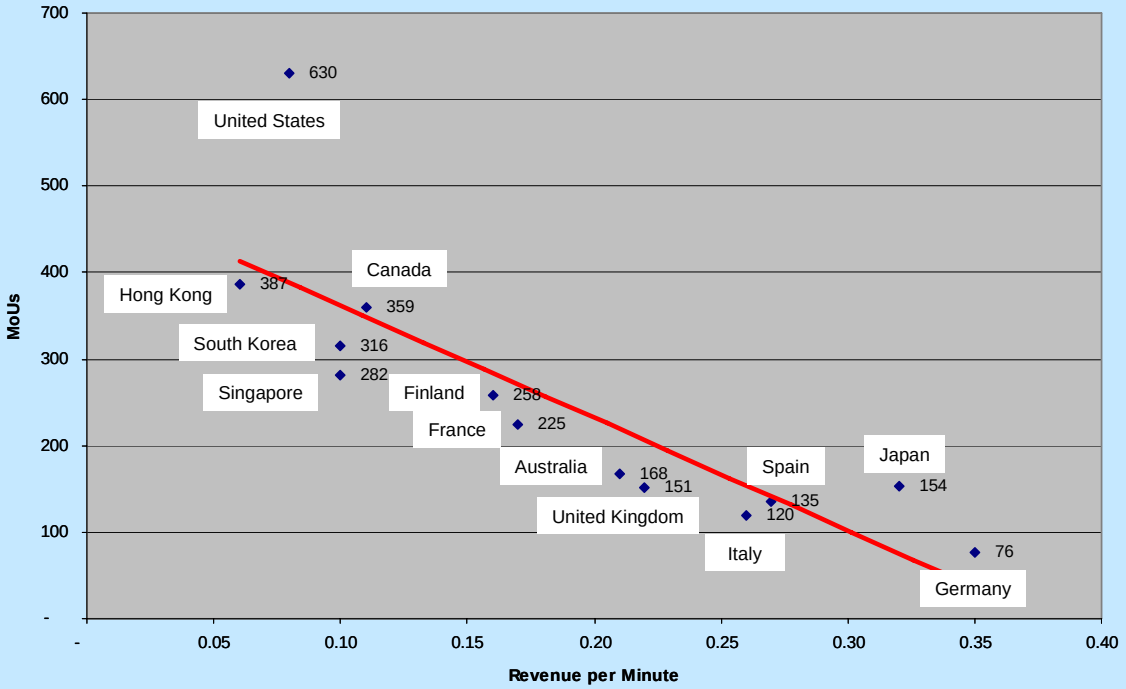
The relationship between termination fees, retail prices, and usage of the service by consumers can more readily be appreciated in regard to the mobile sector, since termination fees and in some cases retail prices are often regulated for fixed incumbents.[34] The investment firm Merrill-Lynch provides an annual analysis of the mobile sector in a number of countries that the U.S. FCC routinely quotes in their annual reports on competition in the U.S. mobile industry,[35] and that other economists also find it convenient to quote.[36] This data is shown in Figure x. For this purpose, we can take the *revenue per minute* for all carriers in a country as being a reasonable proxy for retail price, and a proxy that avoids the complexity of dealing with a plethora of different pricing plans and promotional offers. The *minutes of use* includes minutes of both origination and termination, whether charged or not. Based on this data, Figure 2.1 below depicts the relationship between revenue per minute and minute of use for a number of countries.

Table 2.1: Revenue per minute versus monthly minutes of use for mobile services.

Country	Revenue per Minute ($)	Minutes of Use
USA	0.08	630
Hong Kong	0.06	387
Canada	0.11	359
South Korea	0.10	316
Singapore	0.10	282
Finland	0.16	258
France	0.17	225
Australia	0.21	168
Japan	0.32	154
UK	0.22	151
Spain	0.27	135
Italy	0.26	120
Germany	0.35	76

Source: *FCC,* Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 10th Report (10th CMRS Competition Report*), July 2005, Table 10, based on Glen Campbell et al.,* Global Wireless Matrix 4Q04, *Global Securities Research & Economics Group, Merrill Lynch, Apr. 13, 2005.*

Figure 2.2: Minutes of use versus revenue per minute for mobile services.



Source: The data derive from FCC, *Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 10th Report (10th CMRS Competition Report*), July 2005, Table 10, based on Glen Campbell et al., *Global Wireless Matrix 4Q04*, Global Securities Research & Economics Group, Merrill Lynch, Apr. 13, 2005.

The data clearly suggest that lower retail prices will tend to be associated with significantly higher utilization, expressed in minutes of use per month, and *vice versa*. The United States – with per-minute revenues of just $0.08 per minute, but with a marginal price that many users perceive (somewhat inexactly)[37] as zero – experiences more than eight times as much consumption, expressed in terms of minutes used per month, as a country like Germany, where average revenue per month is about $0.35 per month.

Strictly speaking, what is depicted is not demand elasticity – these are not the same customers, and the mobile services that they are using are not mutually substitutable, because they exist in different countries. But the data strongly suggest that demand is elastic, which is to say that a lower price will lead to notably higher utilization.

Thus, Bill and Keep arrangements make possible retail plans with flat or bucketed rates that are perceived as having zero marginal price, and that consequently generate heavy and efficient usage; however, these same plans tend to be associated with slower adoption of mobile services by consumers. The more common CPP/CPNP arrangements generate effective subsidies to mobile operators. Portions of these subsidies are returned to consumers[38] in the form of low or zero commitment periods, subsidies on handset purchase, and low or zero fixed (monthly) fees. CPP/CPNP systems also may be more hospitable to pre-paid arrangements than are Bill and Keep arrangements.

The low fixed fees and low monthly price make it very easy for a consumer to procure a new mobile service. The consumer need make only a small initial investment and commitment. To the extent that the consumer intends primarily to receive calls, rather than to originate them, the total cost will remain low. Conversely, the operator benefits from termination fees in excess of marginal cost whenever the consumer receives calls. The low, subsidised initial price is a clear case of "giving away a razor in order to sell the blades".

The combined effect is to encourage consumers to initially adopt mobile service.[39]

In Europe, there is a growing sense that it is no longer necessary to subsidize the adoption of mobile services.[40] A number of European countries have penetration rates in excess of 100%.[41] Conversely, Crandall and Sidak argue persuasively that mobile phone penetration in the United States (currently at 65%, and growing by about five points per year) is within just a year or two of reaching European levels, and that Canada is following the same pattern but trailing by a few years.[42] Thus, countries that have buckets of minutes arrangements, based on Bill and Keep wholesale arrangements, tend to experience slower take-up, but can in time achieve reasonably high adoption rates.

In particular, these termination arrangements effectively subsidize mobile operators at the expense of fixed operators and fixed customers. This subsidy is arguably irrational and inappropriate.

To re-cap, what appears to be known is:

- Bill and Keep wholesale arrangements enable low or zero retail per-minute usage fees, but higher initial and fixed per-month fees;
- CPNP wholesale arrangements tend conversely to preclude flat rate or buckets of minutes retail arrangements, leading instead to low initial and per-month fees but high per-minute usage fees;
- Countries with buckets of minutes retail arrangements tend to experience high and efficient utilization, but slower adoption of mobile services;
- Countries with conventional CPNP/CPP arrangements tend to experience lower utilization, but faster adoption of mobile services.

An obvious implication is that countries where the market for mobile services is already mature or saturated might want to consider changing to Bill and Keep arrangements. Conversely, developing countries anxious to foster the widespread initial adoption of mobile services might prefer CPP/CPNP.

2.4 The Internet

2.4.1 Peering versus Transit

The two most prevalent forms of Internet interconnection are *peering* and *transit.* For a definition of these terms, we turn to a publication of the Network Reliability and Interoperability Council (NRIC), an industry advisory panel to the U.S. FCC:

> *Peering* is an agreement between ISPs to carry traffic for each other and for their respective customers. Peering does not include the obligation to carry traffic to third parties. Peering is usually a bilateral business and technical arrangement, where two providers agree to accept traffic from one another, and from one another's customers (and thus from their customers' customers). …
>
> *Transit* is an agreement where an ISP agrees to carry traffic on behalf of another ISP or end user. In most cases transit will include an obligation to carry traffic to third parties. Transit is usually a bilateral business and technical arrangement, where one provider (the transit provider) agrees to carry traffic to third parties on behalf of another provider or an end user (the customer). In most cases, the transit provider carries traffic to and from its other customers, and to and from every destination on the Internet, as part of the transit arrangement. In a transit agreement, the ISP often also provides ancillary services, such as Service Level Agreements, installation support, local telecom provisioning, and Network Operations Center (NOC) support.
>
> Peering thus offers a provider access only to a single provider's customers. Transit, by contrast, usually provides access at a predictable price to the entire Internet. … Historically, peering has often been done on a bill-and-keep basis, without cash payments. Peering where there is no explicit exchange of money between parties, and where each party supports part of the cost of the interconnect, … is typically used where both parties perceive a roughly equal exchange of value. Peering therefore is fundamentally a barter relationship.[43]

In the literature, there is some tendency to assume that peering is invariably free, but this is not necessarily the case. Peering is a technical rather than an economic matter; the economic consequences then follow. When the author was in charge of peering policy for GTE Internetworking (at the time one of the five largest

Internet backbones in the world), about 10% of our peering relationships involved payment. These payments were not a function of the relative sizes of the participants; rather, they were a reflection of traffic imbalance. For Internet backbones interconnected at multiple points by means of shortest exit routing, the traffic *received* from another network must on the average be carried further, and must therefore cost more, than the traffic *sent* to the other network ; consequently, when traffic is unbalanced, the network that *sends* more traffic incurs lower cost than the network that *receives* more traffic.[44]

2.4.2 Roughly hierarchical structure

It is impractical for every ISP to directly peer with every other ISP.

A few years ago, *Boardwatch Magazine* listed more than 7,000 ISPs in the United States alone.[45] I am aware of no current reliable data on the number of distinct ISPs in the world, but the number of *Autnonomous System Numbers (ASNs*) currently assigned sets an effective upper limit, since it represents the maximum number of distinct networks that could be using BGP routing to exchange IP data. According to data maintained by the IANA, the responsible global assignment authority, this number might be somewhere between 30,000 and 40,000 networks.[46]

A few years ago, the author was in charge of peering policy for one of the largest Internet backbones in the world at the time. As of 2001, we had perhaps 50 peering relationships. At the same, my staff felt that technical constraints would limit the firm to perhaps a couple of hundred peering relationships at the maximum.

Aside from any remaining technical constraints, the number of peering relationships will in practice also be limited by:

- The costs of providing connections to each of a large number of peering partners; and
- The significant administrative costs associated with maintaining peering agreements with a large number of organizations.

For all of these reasons, the maximum number of peers that an organization could cost-effectively accommodate is perhaps two orders of magnitude less than the number of independent IP-based networks in the world.

This is why the system that has evolved uses a combination of peering and transit relationships to connect to all Internet endpoints in the world. In practice, the Internet can be viewed as a very roughly hierarchical system, comprising (1) a very few large providers that are so richly interconnected as to have no need of a transit provider, and (2) a much larger number of providers who may selectively use peering with a more limited number of partners, and use one of more transit providers to reach the destinations that their peering relationships cannot.[47]

Milgrom et. al. analyzed these peering and transit relationships in depth. Their "… economic analysis of Internet interconnection concludes that routing costs are lower in a hierarchy in which a relatively small number of core ISPs interconnect with each other to provide full routing service to themselves and to non-core ISPs."[48]

2.4.3 Incentives to interconnect

A body of economic theory that first appeared twenty years ago analyzed incentives of firms to conform standards when participating in markets characterized by strong network externalities.[49] Economic analysis suggested that a firm that had a large or dominant customer base would not wish to adhere perfectly to open standards, because full adherence (and thus full fungibility with competing products or services) would limit the ability of the dominant firm to exploit its market power. Some years later, it was recognized that substantially the same analysis applied to network interconnection.

The issue came up in the context of a number of major mergers, and was analyzed at length in Cremer et. al.[50] Again, the conclusion was that, in a market for Internet backbone services characterized by strong network externality effects, if one backbone were to achieve a very large share of the customer base, it would have both the ability and the incentive to disadvantage its competitors. Conversely, as long as the largest backbone had not too large a share of the customer base, and as long as the disparity between the largest

backbone and its nearest competitors were not too great, incentives to achieve excellent interconnection would predominate.

Milgrom et. al. studied backbone peering and reached similar conclusions: "A simple bargaining model of peering arrangements suggests that so long as there is a sufficient number of core ISPs of roughly comparable size that compete vigorously for market share in order to maintain their bill-and-keep interconnection arrangements, the prices of transit and Internet service to end users will be close to cost."[51]

The thresholds at which the potential anticompetitive effects might dominate have not been rigorously determined.[52] What can be said today is that Internet interconnectivity is near perfect, and that peering disputes are, in a relative sense, quite rare. It is reasonable, based on these indicia, to conclude that the global Internet is operating well below the thresholds where the anticompetitive effects would predominate.

2.5 Internet interconnection and PSTN interconnection

In this section, we seek to compare and contrast interconnection in the PSTN world with peering in the world of the Internet. First, we briefly review some results from economic theory. Second, we consider the significance of the absence, in general, of regulation of Internet peering. Third, we draw parallels between the largely unregulated mobile telephony sector in the U.S. and the Internet.

2.5.1 Economic theory and the "missing payment"

Interconnection in the world of the Internet evolved independently from interconnection in the PSTN. There is some tendency, due in part to differences of culture and orientation of the respective market participants, to assume that these are different worlds, with little or no commonality.

In fact, the economic models for intercarrier compensation in the two worlds are closely linked. The definitive works on intercarrier compensation in the world of the PSTN are generally considered to be Armstrong (1998)[53] and Laffont, Rey and Tirole (1998a)[54]. In Laffont et. al. (2005)[55], we compared Internet backbone peering with these economic analyses of the PSTN and found:

> A key difference with this telecommunications literature is that in the latter there is a missing price: receivers do not pay for receiving calls … The missing price has two important implications:
>
> **Pricing.** The operators' optimal usage price reflects their perceived marginal cost. Comparing the two perceived marginal costs of outgoing traffic with and without receiver charge, for given access charge and market shares, the price for sending traffic is higher (lower) than in the presence of reception charges if and only if there is a termination discount (markup). … In sum, the missing payment affects the backbones' perceived costs, and it reallocates costs between origination and reception.
>
> **Stability in competition**. When networks are close substitutes, and receivers are not charged, there exists no equilibrium unless the access charge is near the termination cost.

2.5.2 The unregulated Internet

An important difference between PSTN interconnection and Internet interconnection is that the latter has generally not been subject to regulation. Bilateral negotiations for Internet interconnection have in most cases led to very satisfactory arrangements for all parties concerned.[56] This outcome is best understood in terms of (1) the Coase Theorem, and (2) issues of market power.

The Nobel-prize-winning economist Ronald H. Coase has argued, most notably in a famous 1959 paper,[57] that private parties could in many cases negotiate arrangements to reflect economic values far more accurately and effectively than regulators, provided that relevant property-like rights were sufficiently well defined. The generally positive experience with Internet peering appears to bear this out.

If one party to a bilateral negotiation had significant market power, and the other lacked countervailing power, then one might expect that the Coasian negotiation might either break down or might arrive at an outcome that was not societally optimal. In general, this does not appear to be the case at present. To date, it has been widely if not universally recognized that Internet backbones do not possess significant market power.

The migration to IP-based NGNs is one of several interrelated trends[58] that have the potential to change this assumption in a number of ways. On the one hand, as wired incumbent telephone companies and, in some countries, cable companies evolve into vertically integrated enterprises that are also significant Internet backbones, it is entirely possible that they might leverage the market power associated with last mile facilities into their Internet role. Whether this is actually the case for a specific firm or a specific country would need to be evaluated based on market developments in that country, and also through the lens of that country's regulatory and institutional arrangements. Some countries are well equipped to deal with market power; others are not.

At the same time, market power may be mitigated by the emergence and deployment of technological alternatives. Broadband Internet over cable television already has some tendency to mitigate the market power of telephone incumbents. To the extent that broadband over powerline, broadband wireless and other alternatives achieve widespread deployment, they could go a long way to ameliorating or preventing the emergence of market power.

All things considered, this author is of the opinion that:

- unregulated, Coasian Internet interconnection arrangements continue to work well today in most cases, but that
- regulators will need to pay *more*, not less, attention to potential problems in this regard for some years to come.

2.5.3 Analogy of Internet peering to US mobile-mobile interconnection

In the United States, mobile operators have generally been under no regulatory obligation to interconnect with one another; nonetheless, privately negotiated Coasian wholesale interconnection arrangements have worked well. The sector has tended to operate on a Bill and Keep basis. [59] Retail pricing arrangements are completely unregulated, but operators and consumers have increasingly chosen flat rate (buckets of minutes) plans.

The parallels to Internet peering are striking. This experience reinforces the notion that the predicted economic outcome, in a market characterized by strong network externalities, a lack of market power, and no regulatory constraints, is (1) for good interconnectivity and interoperability, and (2) for Bill and Keep arrangements. Moreover, this experience reinforces the notion that these results flow from the underlying economics, and not from any unique technological property of the Internet.

3 QUALITY OF SERVICE

The IP-based NGN is envisioned as providing different levels of *Quality of Service (QoS*), each perhaps offered at a different price, in order to support applications such as real time voice and video on the same IP-based multi-purpose network as data.

In this section, we consider the economics of QoS service differentiation, the technical QoS requirements of applications such as real time voice, the implications of network externalities for adoption of QoS service differentiation, and the implications for long term widespread adoption of QoS differentiation.

3.1 The economics of service differentiation and price discrimination

The basic notion of service differentiation is not new,[60] and the underlying economics have been well understood for many years.[61] Service differentiation recognizes that different consumers may have different needs and preferences, which translate in economic terms into a different *surplus* (the difference between perceived benefits and cost) deriving from the purchase of one service versus another. Service providers can choose to offer tailored products that will be preferred only by certain consumers, or not.[62] In practice, they general target their distinct offers at different *groups* of consumers (second order price discrimination) rather than targeting different individual consumers (first order price discrimination).

We experience service and price differentiation every day. We drive into a gas station, and choose to purchase regular gasoline or premium. We purchase a ticket for an airplane or train, and choose to purchase either economy or first class. To the extent that the amenities offered in first class have value to us, they increase our surplus, which in turn increases the price that we are willing to pay. The airline charges a higher price because they recognize that those customers that value the amenities are willing to pay the higher price.

Even though the benefits of service differentiation are obvious, it enjoys only mixed public acceptance in the context of industries that have historically provided *common carriage*. A long-standing tradition, particularly in England and in the United States, is that certain industries should serve the public *indifferently*. This indifference is taken to imply that *price discrimination* is not allowed. It is largely as a result of these attitudes that airline prices, for example, were regulated for many years.

Today, economists would generally agree that deregulation of the airline industry in the United States and elsewhere (which permitted the airlines to price discriminate) has provided greater consumer choice, and prices that are on the average lower than they would have been had the industry remained regulated.[63] Consumers have had to adjust to the fact that the person sitting in the adjacent seat may have paid a much higher, or a much lower price than they did; nonetheless, overall consumer welfare has improved.

The airline experience in the United States demonstrates both the opportunities and the risks associated with price discrimination. As the economist Alfred E. Kahn (both a proponent and a primary implementer of airline deregulation in the U.S.) has observed, competition on many air routes proved to be limited to only one or two carriers. "In such imperfect markets, the major carriers have become extremely sophisticated in practicing price discrimination, which has produced an enormously increased spread between discounted and average fares, on the one side, and full fares, on the other. While that development is almost certainly welfare-enhancing, on balance, it also raises the possibility of monopolistic exploitation of demand-inelastic travelers."[64] In other words, those consumers with limited flexibility in their travel requirements could be charged a high premium with impunity. In markets with effective competition, service differentiation and associated price discrimination will tend to enhance consumer welfare. In markets characterized by significant market power, price discrimination could detract from consumer welfare. The airline industry in the U.S. represents an intermediate case, characterized by imperfect competition.

Laffont et. al. (2003)[65] provides a fairly detailed analysis of Internet backbone peering from an economic perspective. In it, we considered possible service differentiation in terms of the mean and variance of packet delay, and in terms of network reliability. We assumed distinct costs for sending and receiving traffic, each proportionate to the total volume of traffic, and we also assumed access charges (either symmetric or asymmetric) proportionate to the volume of traffic, but independent of any consideration of distance. Under these assumptions, symmetric access charges lead to stable competition. In the absence of service differentiation, the backbones would tend to compete away their profits; however, service differentiation between networks can enable the backbones to earn a positive profit.

3.2 Technological considerations for IP/QoS

We now turn to the technological underpinnings of differentiated QoS in an IP network. First, we touch briefly on communications protocol issues; then, we consider application requirements as regards the mean and variance of packet delay. With that established, we consider protocol performance, and discuss the implications for the prospects of widespread adoption.

3.2.1 DiffServ, RSVP, MPLS

By the early Nineties, it had already become obvious to the engineering community that real-time bidirectional voice and video communication could potentially benefit from delivery guarantees on delay. This led to a series of standards efforts – first, the *RSVP*-based Integrated Services Architecture, and then to *Differentiated Services (DiffServ)*.

RSVP provided a comprehensive end-to-end QoS management architecture. Over time, it came to be viewed as hopelessly complex,[66] and was effectively abandoned in favor of DiffServ. DiffServ provides a simple means of specifying, on a hop-by-hop basis, the desired performance characterstics – it is then up to the network to meet those requirements as well as it can.

DiffServ should thus be viewed as a *signaling* mechanism. Technically, it is trivial. The implementation of QoS *within* an IP-based network, with or without DiffServ, has been straightforward with or without DiffServ for at least a decade. Implementation of QoS *between or among* independently managed IP-based networks has never gotten off the ground. Given that the technology is fairly simple, the answers clearly lie in business and economic factors.

3.2.2 Application requirements for bounded delay

Some readers might perhaps assume that all voice and video traffic requires assured quality of service; in reality, however, assurances on the mean and variance of delay are required only for services that involve bidirectional (or multidirectional) voice and video in real time.

The receiving application typically implements a jitter buffer that can be used to smooth the variability in end to end delay. For streaming (one way) audio or video, most users will tolerate a delay of a few seconds when the application starts up. After that, a jitter buffer can typically deal with a considerable amount of variable delay.

For real time bidirectional voice and video, however, users will tend to "collide" if the end to end delay exceeds about 150 to 200 milliseconds. They will both start speaking at roughly the same time, because neither can initially discern whether the other is speaking.[67] This imposes a practical ceiling on the delay that the jitter buffer can allow.

3.2.3 Analysis of delay

This delay in turn imposes limits on both the mean and the standard deviation of delay for the traffic. In an IP-based network, the traffic is composed of individual packets. The delay for these packets can be viewed as comprising a fixed component (based primarily on the speed of signal propagation along the path from send to receiver, and thus dependent primarily on the distance along the path, and also on the deterministic delay to "clock" the packet onto each outbound data transmission link) and a variable component (based on queuing delays in each router through which the packet must pass, especially those associated with gaining access to the outbound transmission link). For a given traffic flow, the unidirectional delay can thus be viewed as a probability distribution with a mean and a standard deviation.

The ability to achieve a round trip delay of not more than 150 milliseconds depends on both the mean and the standard deviation of delay. It is a classic statistical confidence interval problem – it is necessary that the "tail" of the distribution in excess of about 150 milliseconds be suitably small. Note that an occasional outlier is generally permissible – as an example, the *codecs* (coder-decoders) used for Voice over IP (VoIP) services typically interpolate over missing data, and the human ear does a surprisingly good job in compensating for very short data losses. Human speech presumably incorporates a great deal of redundant information that can be used to fill in the gaps.

Fixed delay can be viewed as comprising propagation delay (which is a consequence of the large but finite speed of light) and clocking delay (which is a function of the speed of the transmission link).

We often forget that the speed of light *is* a meaningful constraint. In vacuum, light travels about 300 Km in a millisecond. Signal is not quite as fast when propagating through wires or fiber; moreover, transmission paths (e.g. fiber runs) do not proceed in a geometric straight line. For intercontinental calls, propagation delay can consume a significant fraction of the 150 millisecond budget.

Clocking delay is a function of the speed of the transmission link. Over a dial-up connection to the Internet, clocking delay poses a serious constraint. Over broadband media, it is much less of an issue. In the core of the Internet, the links are very fast indeed, so the deterministic clocking is correspondingly small.

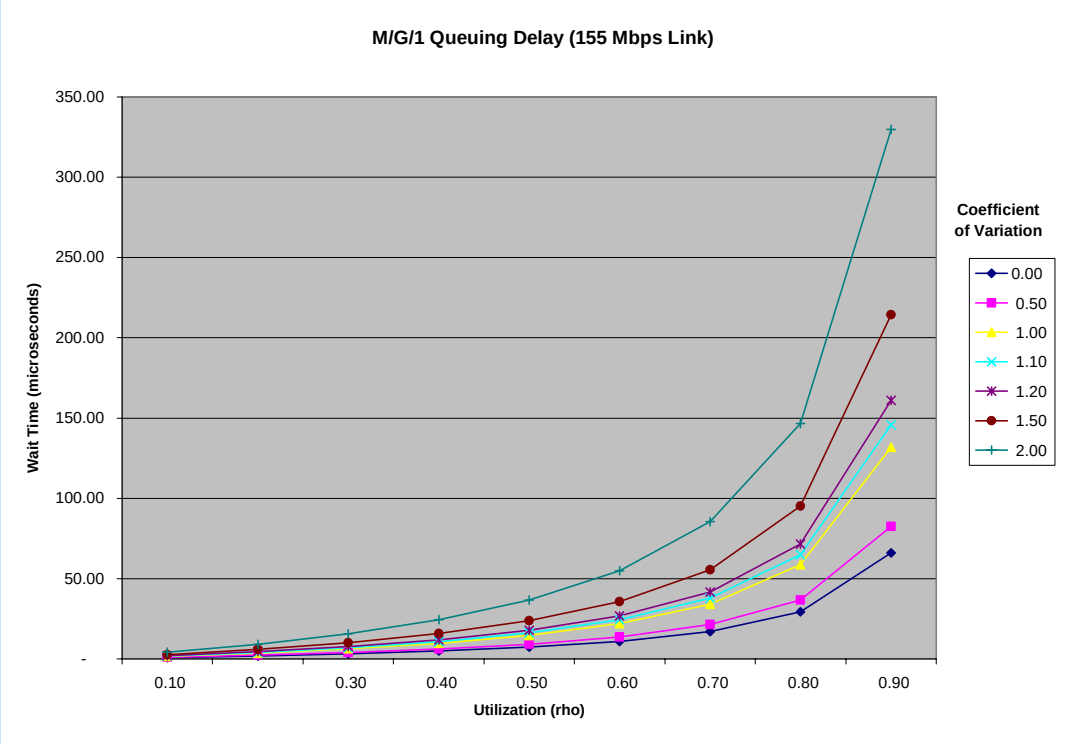
Variable delay is best modeled and analyzed on a hop by hop basis. At each hop, it primarily reflects the queuing delay waiting to clock the traffic onto an outbound link. (Queuing delay for the processor of the router is also possible, but unless the processor is saturated it is generally small enough to ignore.) This variable delay can be analyzed using a branch of mathematics known as *queuing theory* – the science of waiting lines.[68]

Queuing theory tells us that average variable delay reflects three things:

- The average service time (in this case, the deterministic clocking delay);
- The load on the server, which we can think of as the percent of time that it is busy; and
- The variability of the service time, expressed as a coefficient of variation (the standard deviation divided by the mean).

What queuing theory tells us about variable delay in the core of the large IP-based networks is that, *in a properly designed network and under normal operating conditions*, variable delay plays only a very minor role. Figure xxx below depicts the average packet wait time for a 155 Mbps data link, which is the *slowest* link that one would expect to find in the core of a modern Internet backbone.

Figure 3.1: Packet wait time on a 155 Mbps link



Among the family of curves shown, the one corresponding to a coefficient of variation of 1.20 is the one the accords most closely with observational experience around 2001, the most recent date on which this author had access to industry statistics.[69]

The computed average wait time per hop, even at a utilization of 90%, is about 150 *microseconds*. Note that this is *three orders of magnitude* less than the delay budget of 150 *milliseconds*. Beyond this, consider that many backbone links today are one or two orders of magnitude faster than 155 Mbps, with predicted delays correspondingly smaller.

This is not to say that delay could never be a problem. The same queuing theory analysis tells us that, as utilization approaches 100%, predicted mean wait time increases with no upper bound. But no network should be *designed* to operate routinely at those levels. Saturation will occur either as a result of (1) poor planning or forecasting on the part of the network designer, or (2) substantial failures elsewhere in the network that necessitate re-routing of traffic.

3.2.4 Implications for market prospects for QoS

The analysis in the preceding section has significant implications as regards the willingness of customers to pay a surcharge for QoS (in the sense of statiscally bounded delay).

DiffServ-based QoS capabilities cannot speed up a network; they can only prevent it from slowing down (for certain packets) under load. They generally determine (1) which queued packets are served first, and (2) which queued packets are discarded when there is insufficient room to store them.

Under most circumstances, these effects will be too small for the end user to perceive.

It should come as no surprise that end users are unwilling to play a large surcharge for a performance improvement that is not visible to them.[70]

This is not to say that there is no commercial opportunity for inter-provider QoS; rather, it argues that the opportunities will not necessarily be found in the core of the network, which is the place where most people tend to look for them.[71] Instead, QoS will tend to be commercially interesting:

- Within a single provider's network, where the costs of implementation are also low;
- For slower circuits at the edge of the network;
- For shared circuits to the end user (e.g. cable modem services);
- When one or more circuits are saturated;
- When one or more components have failed;
- When a *force majeure* incident (a natural or man-made disaster) has occurred; and especially
- Where more than one of these factors is present.

Providers may also find that offering QoS provides a competitive advantage in attracting new customers, even if those customers are unwilling to pay a large premium.

3.3 Network externalities, transaction costs, and the initial adoption "hump"

The technological capability to deploy differentiated QoS capability at reasonable cost has existed for at least ten years, and has in fact been deployed within many networks. Why has there been so little deployment between or among networks?

The explanation has very little to do with technology, but a great deal to do with economics – specifically, with the economics of *network effects (*or *network externalities)*. An economic market is said to experience network effects when the service becomes more valuable as more people use it. Differentiated QoS is typical of capabilities that take on value only as more networks and more end-users adopt them.

The economist Jeffrey H. Rohlfs has written extensively on the subject of network effects, noting that many new high technology services encounter difficulty in achieving sufficient penetration to get past an initial adoption hump.[72] A certain number of end-users might take up a product or service based solely on its intrinsic value, but that is likely to be far fewer end-users than the number that would take up the service if everybody else did. The market can easily settle into equilibrium at a number of end-users that is far less than the level that would be societally optimal.

The initial adoption hump is often exacerbated by complementarities. A service cannot get launched because it depends on supporting upstream or downstream products and services. CD players could not have succeeded in the marketplace without a substantial inventory of music to play on them. Television sets could not have succeeded without programs to watch. Personal computers could not have succeeded without software to run on them.

Different successful offerings have met this challenge in different ways. In some cases, government intervention has been required. Ubiquitous telephone service is explicitly or implicitly subsidized in many countries – this is referred to as *universal service*. The initial adoption of CD players was facilitated by the fact that the companies that made the players – Phillips and Matsushita – also had interests in studios, could profit on both sides of the market, and were consequently highly motivated to ensure that both players and content were available. The deployment of VCRs in the United States was facilitated by an initial deployment for time shifting of programs – a market for the rental of videos did not emerge until enough devices had worked their way into the hands of consumers.

Certain Internet capabilities have deployed effortlessly – for example, the worldwide web. In many cases, the successful capabilities benefit from the end to end principle – they can be implemented by end-user organizations or consumers, without requiring any action at all on the part of the providers of the underlying IP-based network.

Conversely, other capabilities have tended to deploy at a glacial pace or to stall for reasons not necessarily related to technology, notably including IP version 6 (IPv6), DNS security (DNSSEC), and multicast. A common characteristic among the stalled capabilities is that, rather than being end to end features independent of the network, the stalled capabilities require concerted action and concerted change to the core of the network. Regrettably, inter-provider QoS seems to clearly fit the profile of the stalled capabilities.

Common characteristics among the slow-deploying capabilities include:

- Benefits that are in some sense insufficient: too limited, too difficult to quantify, too slow to appear, too difficult for the investing organizations to internalize.
- Limited benefits until the service is widely deployed.
- The need for coordination among a large number of organizations, leading to high economic *transaction costs* (the cost for a network or an end-user to adopt the service).

If the tangible economic benefits were well in excess of the costs, the services would deploy effortlessly. There are services where the benefits to the organizations that would have to make the investments do not clearly exceed the costs – consequently, the investments are made slowly if at all. The unfavorable relationship between costs and visible benefits hinders initial deployment, and thwarts attempts to reach critical mass and thereby to get beyond the initial adoption hump.[73]

3.4 Prospects for inter-provider QoS in an NGN world

For inter-provider QoS, the benefits in most cases may not be compelling for reasons outlined in Section 3.2 of this paper – in the absence of differentiated QoS, the performance of best efforts traffic will tend to be perfectly adequate in most networks most of the time, and consumers are unlikely to perceive a difference that they are willing to pay for. Moreover, the benefits are limited by the number of other providers that support QoS – the benefits to the first few providers are quite limited.

Conversely, the number of parties that would have to come to agreement to achieve a globally interconnected QoS-capable world is very large.[74] If every pair of providers requires a contractual agreement in order to put QoS in place, then a world with thousands of independent providers will require literally millions of agreements – and complicated agreements at that, for reasons that are explained in section 6 of this report. This will not happen. It is safe to predict that a comprehensive, global and universal system of QoS-capable interconnection will not happen without some kind of help.

It might nonetheless be possible to get inter-provider QoS to deploy. Anything that can reduce the associated transaction costs will tend to increase the likelihood of getting a decent deployment. Some initiatives that might possibly reduce transaction costs include:

- Experiments and pilot projects among pairs or small groups of cooperating service providers.
- Once the problem is better understood, model agreements for inter-provider interconnection including QoS support.[75]
- The continued enhancement of commercial monitoring and measurement tools that could serve as standardized building blocks for service provider operational support systems (OSS).
- Possible emergence of organizations that could gain acceptance as trusted third parties to capture statistics and/or to mediate billing and accounting disputes.

In addition, it is worth noting that the service providers are unable to require that the entire world implement QoS. Most providers will find that the majority of their traffic is exchanged with a limited number of "trading partners in bits", perhaps a dozen or two. Any realistic provider deployment plan will have to simply accept that some providers will offer QoS-capable interconnection, while others will not.

4 MARKET POWER AND NGN INTERCONNECTION

At the regulatory and policy level, interconnection has always been closely associated with questions of market power. It has been a general article of faith that governments must be prepared to intervene to address such abuses of market power as might exist.

Telecommunications networks were initially presumed to be natural monopolies, industries where initial costs were so high as to preclude competition between two providers in a single geographic area. In most countries, the government itself provided these services, usually through a Post, Telephone and Telegraphy (PTT) authority. In a few, notably including the United States and Canada, equivalent services were historically provided by highly regulated firms that were *de facto* monopolies with significant *de jure* privileges and protection.

With liberalization, services that were previously provided by the government have been privatized, and competitors have been encouraged to enter these markets. In most cases, the established incumbents have resisted competitive entry, either by price-based or by non-price-based discrimination.[76] This behavior is conditioned and shaped by legal and regulatory institutions in each country, but similar underlying economic factors tend to encourage similar incumbent behaviors in all countries.[77]

Once competition is established and effective, it is generally accepted that regulation should be withdrawn. At that point, market forces will channel service provider behavior more effectively than any regulator could hope to.

At the same time, it is important that regulation not be withdrawn *before* competition is effective. Reform-minded New Zealand attempted for many years to operate without a conventional sector-specific regulator. In 2001, they gave it up as a bad job and implemented lightweight institutions approximating the function of a sector specific regulator. Interminable interconnection disputes were the primary reason.[78]

4.1 Sources of market power

Market power most often arises as a result of control of some asset that represents a competitive bottleneck, and that cannot easily be replicated by competitors. In telephony, the primary concern has usually been with "last mile" facilities, which are discussed in the next sub-section. There are other potential bottlenecks that might manifest themselves in specific circumstances, or perhaps more generally in the future – we consider those as well in the subsequent sections.

4.1.1 Last mile considerations

Wired access to the customer premises (e.g. to the consumer's residence) tends in to be a durable competitive bottleneck throughout the world, but more so in some countries, and in portions of some countries, than in others.

The emergence of NGN access networks may mitigate these concerns, but it is unlikely to eliminate them for the foreseeable future.

In some developed countries, cable television service is sufficiently widespread, and is sufficiently ubiquitously upgraded to carry data and/or telephony, to significantly mitigate the market power of the wired telephony incumbent. Mobile services may also serve as a counterbalance against the market power of the incumbent, including to an increasing degree wireless broadband services. Satellite must also be considered, but it tends to play less of a role for reasons of cost and scalability. Emerging technologies, including broadband over powerline, may play a significant role in the future.

Nonetheless, last mile bottlenecks are likely to be significant for many years to come, and at least portions of most countries are likely to lack effective competition on the last mile. Wherever last mile competitive bottlenecks exist, established operators are likely to find it profitable to restrict or prevent interconnection. Governments and regulators will need to remain alert to this possibility, and must be prepared to intervene if necessary.

4.1.2 Network externality considerations

Last mile bottlenecks tend to be the most commonly noted concern as regards competitive bottlenecks, but they are not the only possible concern.

A body of economic theory argues that, in markets characterized by strong network externality effects, firms with a strong market share of customers will be motivated to have less-than-perfect interoperability and less-than-perfect interconnection.[79]

These concerns have occasionally been relevant to policy in significant ways. They played a large role in the evaluation of the WorldCom-MCI merger and the attempted WorldCom-Sprint merger.[80]

Economic theory does not provide any clear indication as to how large a market share is needed for these effects to motivate action, i.e. to be profitable. At the same time, there is good reason to believe that the world is generally well below that threshold – Internet interconnection today is nearly perfect worldwide, and interconnection disputes are rare.[81]

4.2 Addressing market power

Different countries will have developed different methodologies for addressing market power as it relates to interconnection. In the view of the author, the approach that the European Union adopted in 2003 reflects a particularly forward-looking way to deal with migrations such as that to the NGN.

Under the European regulatory framework for electronic communications, regulators (1) clearly identify a set of relevant markets that could be of interest; (2) determine, using tools borrowed from competition law and economics, whether any firm or group of firms has Significant Market Power (SMP) on such a market; (3) applies a minimally adequate set of *ex ante* (in advance) remedies only to the firm or firms that possess SMP; and (4) removes any corresponding obligations that might have previously existed from firms that do not possess SMP. The framework is technologically neutral – whether a service is delivered using a traditional network or an IP-based NGN is irrelevant. A relevant market is determined based on the service or services delivered to the user, and considering the degree of substitutability for other services, consistent with competition law.

Properly implemented, a regulatory framework of this type enables a regulator to address such market power as may still exist in an NGN world, and also provides a natural and organic method for withdrawing regulation when it is no longer needed.[82]

4.3 Remedies for market power, or a "regulatory holiday"?

In Europe and in North America, a key question has emerged: What is the most appropriate role for government in ensuring that necessary investments are made in new network infrastructure? The debate has largely focused on broadband Internet access, which can be viewed as the access portion of the NGN, but similar issues can be raised about the NGN core.

In a perceptive essay[83], Nicholas Garnham observed that regulatory policy is confused to the extent that it tries to follow multiple economic theories at once, without a way to prioritize or to choose among different and mutually contradictory implications. One of these models is the classical view of competition law and economics, which argues that governments must address such market power as may exist. Another is the Hayekian view, which argues that government must refrain from favoring one solution over another, in order to enable the best to survive – a sort of Darwinian economics. A third is the view of Schumpeter, which argues that progress comes from "creative destruction", and that supracompetitive profits are necessary in order to motivate investment.

The Schumpeterian view is sometimes invoked in support of radical deregulation. The competition law view implies instead that, in problematic markets characterized by non-replicable assets, procompetitive regulation may be needed until effective competition has emerged.

Justus Haucap has characterized this tension of objectives as reflecting a confusion of *deregulation* with *liberalization* – both are much praised, sometimes in the same breath, but they are not the same thing.

Liberalization is a matter of *enabling market entry*, which in some cases implies to need to impose or maintain regulation, not necessarily to eliminate it.[84]

4.3.1 Incentives for providers to deploy

In North America, we have seen the rapid withdrawal of regulation. In Europe, the debate has been expressed in terms of the need for a *regulatory holiday* – a deferral or forbearance from regulation for some period of time in order to spur investment. On both continents, there is support in the law for the sensible notion that regulation should not prematurely be imposed on nascent or emerging services. What is not so clear, unfortunately, is the proper balance between the conflicting Schumpeterian and competition law objectives. Beyond that, what exactly is an emerging service? How long should regulation be deferred? When can an emerging service be said to have emerged?

This debate is likely to be with us for some years to come. Both sides will have adherents, and those adherents are likely to be well funded. It may be some years before the effects can be seen to clearly favor one approach or another.

My personal view is that, in markets that are well established, and where one or more market participants continue to have durable and significant market power, that premature withdrawal of procompetitive regulation is likely to do much more harm than good. Deregulation under those conditions might possibly spur investment by the incumbent operator in the near term, but it will also depress investment by competitive operators. Over time, it seems to me that it is likely to lead to less competition, less innovation and less investment than an effectively regulated system.

4.3.2 Return on Investment (ROI) under conditions of risk

Whatever one's views about deregulation of markets that are not yet competitive, it is clearly appropriate for service providers to make a reasonable return on reasonable investments. For a firm that is subject to regulation, this generally implies a need to compute the *Return on Investment (ROI*) that will be considered to be acceptable for regulatory purposes. Greater risks – as might be expected in connection with migration to the NGN – should be associated with greater expected returns.

Regulators typically determine an appropriate ROI by computing an appropriate *Weighted Average Cost of Capital (WACC*) for the firm. The Weighted Average Cost of Capital (WACC) reflects the cost of equity, the cost of debt, and the company's gearing (a measure of the company's ratio between debt and equity).

The Capital Asset Pricing Mechanism ("CAPM") is a widely used and theoretically well grounded methodology for reflecting risk and its impact on the returns that shareholders should expect. In CAPM, the cost of equity capital is rolled up from three components: (1) the risk free rate; (2) the expected market equity risk premium; and (3) the value of beta for the company in question. The *Risk Free Rate (RFR*) is simply the return that an investor would expect on a risk free investment. The *Equity Risk Premium (ERP*) is a stock-market factor, rather than being company specific, that reflects the degree to which investors expect a higher return for putting money into equity instruments (stocks) than into risk free investments. The beta is a relative measure of the risk that is relevant to the specific firm.

Ofcom, the UK regulator, recently conducted a detailed analysis of the appropriate WACC for British Telecom (BT).[85] Their consultation document provides a very lucid overview of the determination of a WACC for an incumbent provider that is on the verge of a rapid migration to an NGN. They chose to *disaggregate* BT's beta – instead of using a single beta for all of BT, they associated a somewhat lower beta with BT's relatively low risk local loop activities, and a somewhat higher beta with the rest of BT's activities. These different betas then led Ofcom to compute two different WACCs and thus to permit different levels of ROI for different parts of BT.

Ofcom considered various options, but they did not finally resolve the ROI that might be appropriate when BT migrates to an NGN (which BT intends to do on a very accelerated schedule. Ofcom has indicated that BT's risk might be slightly higher for next generation core networks, and significantly higher for next generation access networks, than for BT's current network. Ofcom might address this through further refinements to BT's beta; alternatively, they have raised the possibility of addressing these different levels of risk through a modeling mechanism known as Real Options[86].

4.4 The "network neutrality" debate

A debate has raged in the United States over the past several years over the degree to which providers of broadband Internet access service should be obliged to provide nondiscriminatory access to all content[87] available on the Internet, using any equipment and any application and any protocol that does not harm the network.

In essence, there is increasing concern that new forms of market power might emerge and might be exploited by broadband providers. The concern is exacerbated by the movement of phone companies to also provide video programming, thus offering a vertically integrated service that competes with cable television.

A number of very different concerns have been raised under the banner of network neutrality, mostly in connection with local telephone incumbents or cable TV operators that are vertically integrated with an Internet Service Providers (ISP):

- The possibility that an integrated ISP might offer better performance to some Internet sites than to others;
- The possibility that an integrated ISP might assess a surcharge where a customer wants better-than-standard performance to certain Internet sites;
- The fear that the integrated ISP might permit access only to affiliated sites, and block access to unaffiliated sites;
- The fear that the integrated ISP might assess surcharges for the use of certain applications, or of certain devices;
- The fear that the integrated ISP might disallow outright the use of certain applications, or of certain devices, especially where those applications or devices compete with services that the integrated ISP offers and for which it charges; and
- The fear that the integrated ISP might erect "tollgates" in order to collect unwarranted charges from unaffiliated content providers who need to reach the integrated ISP's customers.[88]

The perceptive reader will have already observed that a number of these concerns (but not all) relate to conduct that, in the absence of market power, would clearly tend to enhance consumer welfare. In a fully competitive market, demanding a surcharge for better performance or for the ability to use highly valued applications would be unobjectionable. With effective competition, the potential for abuse – for example, in the form of assessing chargest that exceed cost to an unreasonable degree – would tend to be contained by the likelihood that competitors would find it profitable to steal customers by offering equivalent services at prices that were less elevated, or under terms and conditions that were less onerous.

As an example, some net neutrality advocates have complained because their provider would offer static (i.e. permanent) IP addresses only in connection with higher-priced services. They complained that they were effectively being prevented from running web servers and other services. In an economic sense, however, this "blockage" is not necessarily problematic. Running a web server will, on the average, result in more traffic for the provider's network, which will in turn tend to result in increased cost to the provider. Aside from that, it represents increased utility to the consumer, and thus an increased surplus and an increased willingness to pay on the part of the consumer. In economics, one of the key properties associated with a service that can be offered for sale is *excludability* – the ability to prevent its use by those who have not paid for it. In this sense, providing static IP addresses only in connection with a higher priced service would, in a competitive marketplace, be viewed as entirely normal and appropriate.

It is also worth noting that there are a great many legitimate reasons to block access to specific Internet addresses – most notably, concerns about security or SPAM. Beyond this, no Internet provider is able to guarantee access to all Internet addresses at all times.

All of this suggests, first, that there is enormous confusion and ambiguity as to what conduct is truly objectionable, and second, that it would be exceptionally difficult to craft a meaningful and enforceable *ex ante* rule to prevent abuse.

4.4.1 Developments in the U.S.

On March 3, 2005, the FCC announced that it had reached a consent decree with Madison River, a small local telecommunications incumbent.[89] Madison River agreed to make a payment, in effect a fine, in recognition that it had blocked access to VoIP services offered by Vonage.

The FCC has not published supporting details,[90] but one might reasonably infer (1) that Madison River customers had little or no ability to choose another broadband provider, and (2) that Madison River chose to block Vonage in order to prevent competition with its own conventional PSTN voice services. If these conjectures are true, then Madison River's conduct was indeed problematic – its actions could be viewed as a leveraging of last mile market power into an otherwise competitive market.

The net result of the FCC's actions, however, must be said to be very confused. The action was, in a sense, probably appropriate, but it left no clear ground rules going forward. Normally, a firm can be fined for willfully violating an FCC rule; however, that implies that there was a rule to violate, and that the company knew or could reasonably infer the rule. The FCC has published no rule, and it is difficult to see how any company could reasonably infer what conduct is permitted and what conduct prohibited today.

Meanwhile, the issue continues to churn in the United States. In recent days, a number of senior telephone company and cable TV executives have spoken of the need to charge content providers such as Yahoo and Google (who are not necessarily customers of th integrated ISP in question) for their use of the ISP's network to reach the integrated ISP's customers. This is not a new idea – it was tried in the past, with no success. In a competitive market, the content providers will simply refuse to pay. An open question is whether recent changes in the U.S. broadband and Internet marketplace, in terms of consolidation and of the collapse of the wholesale market for broadband services,[91] have now made this a profitable strategy.

4.4.2 Policy implications

My view is that there has been very little real abuse of this type to date, and moreover that much of the abuse that has been alleged should not be viewed as problematic. At the same time, there is good reason to believe that problematic behaviors would be both feasible and profitable in the context of a sufficiently concentrated marketplace for broadband Internet access, especially as providers become increasingly vertically integrated.

If these behaviors were to become solidly entrenched, it would be difficult if not impossible to prevent them by means of *ex ante* rules. It is simply too difficult to distinguish between appropriate and inappropriate behavior.

What this strongly suggests is that most countries would be well advised to ensure that they maintain robust competition for broadband Internet services. Competition must be the first, and most critical, line of defense. It is worth noting that the competition need not be facilities-based – service-based competition could be perfectly adequate, as long as the underlying facilities provider cannot constrain the competitive provider's connectivity.

A second implication is, in countries where competition law provides an *ex post* complement to sector-specific regulation, that isolated abuses of this type might be most appropriately addressed *ex post* as violations of competition law, rather than by *ex ante* regulation. My belief is that the truly problematic abuses generally represent inappropriate exploitation of market power.

5 UNIVERSAL SERVICE AND NGN INTERCONNECTION

Charges associated with interconnection are often used as a means of financing universal service – the availability of basic electronic communications to all, at affordable prices. Section 5.1 explains the rationale, in terms of network externalities, economic distortions, and consumer welfare. Section 5.2 explains the use of implicit interconnection-based subsidies within a developing country, while Section 5.3 explores subsidization mechanisms among independent nations. Section 5.4 expands on the implications for policy.

5.1 Network externalities, economic distortions, and consumer welfare

In section 3.x, we explained that markets characterized by network externalities may have a tendency to reach stable equilibrium at levels of service adoption that are much lower than those that are societally optimal. Most countries have felt that voice telephone service was so important that the government should subsidize the service where necessary in order to ensure that the service is available to all, and even to those of limited means. In some cases, this has meant a commitment to universal access (e.g. availability in a nearby school, library or post office) rather than in the home.

Different countries generate these subsidies in different ways. Most economists would argue that it is best to take the funds from general revenues (i.e. overall taxation), because doing so ensures that the cost is spread as widely and as equitably as possible, and thus minimizes economic distortions; however, this is very rarely done in practice.

Some countries simply expect the incumbent local carrier to provide universal service, and to someone extract enough profit from other customers to cover the cost. Still others provide a specific universal service fund, with all providers of electronic communication services contributing.

The relevance of this discussion to interconnection arrangements is that intercarrier compensation is often used as an alternative, implicit means of generating the necessary subsidies.

5.2 Intercarrier compensation as a funding mechanism for ICT development

Domestically, access charges can provide a funding vehicle in the form of implicit subsidies. Network costs will tend to be greater in those areas that pose universal service challenges due to low teledensity or unfavorable geography. Some countries find it convenient to set access charges to higher levels in those areas in order to generate a net influx of money.

The World Bank has generally been supportive of the use of access charges as means of subsidizing telecoms deployment to rural or remote areas of developing countries.

At the same time, this technique is by no means limited to developing countries. It continues to generate implicit universal subsidies in a number of developed countries, including the United States. The U.S. has attempted to phase out these implicit subsidies for years, but they persist.

A number of concerns must be raised in connection with these subsidies. They represent an economic distortion. They are subtle, and not likely to be understood by the public – there can thus be a notable lack of transparency. And they can easily turn into "slush funds".

5.3 Traffic imbalance – the "Robin Hood" effect

In section 2 of this report, we explained that traditional PSTN intercarrier compensation in most countries is paid according to the Calling Party's Network Pays (CPNP) principle. It turns out that inhabitants of developed countries tend to place far more calls to inhabitants of developing countries than *vice versa*; consequently, these international termination fees (technically referred to as *settlement fees*) generate a net transfer of money from developed countries to developing countries.

This mechanism has the rather strange property of transferring money from richer countries to poorer ones. As such, one could draw a certain parallel to the mythical English folk hero Robin Hood, who robbed from the rich in order to give to the poor. The system functions as an inadvertent form of foreign aid.

Not surprisingly, developing countries have generally wanted to keep per-minute wholesale termination fees[92] at very high levels, well in excess of real cost, in order to maximize the transfer of funds. Equally unsurprisingly, a number of developed countries, most notably the United States, have wanted to drive these payments down to levels approximating real termination costs.

In one recent incident, the government of Jamaica imposed a levy on international call termination payments, in order to explicitly generate subsidies to fund universal service.[93] The U.S. FCC complained, saying that "… universal service obligations must be administered in a transparent, non-discriminatory and competitively neutral manner, and that hidden subsidies in settlement rates and subsidies borne

disproportionately by one service, in the case of settlement rates, by consumers from net payer countries, are not consistent with these principles and cannot be sustained in a competitive global market."[94]

5.4 Policy implications

The migration from today's world of the PSTN to tomorrow's world of the IP-based NGN probably implies that all of these implicit subsidy mechanisms will gradually either be explicitly phased out, or else will become irrelevant over time.

These termination payments are assuredly not an ideal subsidy mechanism; nonetheless, the fact remains that they have transferred funds to developing countries, and that portions of those funds may have served to fund telecoms development projects to remote or rural areas. The funding vehicle is likely to go away, but the development needs that it addressed, however imperfectly, will remain.

6 BILLING AND ACCOUNTING IN AN IP-BASED WORLD

Up to this point, we have primarily considered possible intercarrier compensation arrangements from an economic perspective. These arrangements interact with the underlying IP technology in complicated ways, and have business implications that are perhaps unobvious. In this section, we explore some of the interactions between technology and economics.

6.1 Protocol layering, services, and the underlying network

In an IP-based environment, applications such as Voice over IP (VoIP) operate over an IP-based core network. Protocols are layered in the interest of simplifying the network, and facilitating its evolution over time. These properties have profound implications, not only for usage accounting and billing, but also for the structure of the industry.

Historically, it was generally the case that a single organization would provide both the public telephony *service* and the *network* used to deliver that service. In the world of the IP-based NGN, the network provider will still in most cases still be a service provider, but it will not necessarily be the *only* service provider. Vonage, Skype and SIPgate are examples of competitive firms that provide services without operating a network of their own. For the foreseeable future, integrated and independent service providers are likely to coexist, and to compete for the same end-users customers. Moreover, this competition between integrated and independent service providers is a useful thing, that should be preserved – it tends to enhance consumer welfare.

This separation of function has profound implications for both the network provider and the service provider.

In theory, the network provider in an IP-based world does not know or care about the nature of the application traffic that it is carrying – and in this context, voice is just another application. The network is aware of the Quality of Service that the application has requested for any particular packet, but it should not concern itself with the application itself.

Conversely, the application provider – for example, the independent VoIP provider – will have little or no visibility into the networks that it is traversing. In fact, the application will not necessarily be able to predict which networks its traffic will traverse, and in general the application should not care. The networks collectively provide a path for the application's data traffic, but little more. The application can request a particular Quality of Service for its traffic, but without absolute certainty that its request will be honored.

This lack of awareness has in general proven to be a valuable quality, but it has implications. The independent application provider cannot guarantee the quality of transmission, because it does not own the underlying networks and may not know or care which networks are involved.

The application service space, for example for VoIP, will tend to be a highly competitive market segment unless regulation or anticompetitive actions on the part of network operators (see the discussion on Network Neutrality later in the section) dictate otherwise. The competitiveness of the segment will tend to restrict prices to competitive levels, generally reflecting marginal cost plus a reasonable return on investment. This

same competition will tend to constrain the price that the network operator can charge for its integrated service.

All indications are that the marginal cost of the VoIP-based telephony service, independent of the underlying network, is very low.[95] If independent VoIP service providers indeed maintain a competitive market for the service, then the low marginal cost should lead to a low marginal consumer price for the service.

At the same time, the network operator may have (absent regulation) some degree of market power associated with last mile broadband access. To the degree that this is so, the network operator could be said to have market power on one market segment (network access, especially last mile access) that is vertically related to another market segment that is competitive. Under those circumstances, the network operator is likely to exploit its market power, and may try to extend it to the otherwise competitive segment. The simplest and most likely strategy is for the network operator to take a high mark-up (a monopoly profit if it is the only network operator) on the last mile network access, while pricing the voice application at competitive levels.

For this reason, many countries will find it necessary to maintain regulation that seeks to address durable bottlenecks associated with last mile access, to the extent that effective competition has not yet emerged for the last mile. Countries will see these needs through the lens of their own experience and their own institutions, but many or most will find it necessary to retain regulatory measures, or to institute them if they do not exist, in order to enable competitive entry and to sustain it over time, and to limit the exploitation of market power where competition is not yet effective.

6.2 Point-to-point versus end-to-end measurement

The technology and economics of these systems interact in complicated ways.

The underlying network economics strongly influence the nature of the things that operators and service providers will want to bill for; however, those bills will have to be justified and reconciled based on some kind of accounting data. Billing needs largely determine accounting system needs.

Conversely, not all of the data that might be desired can be acquired at reasonable cost, so the capabilities that can reasonably be achieved by accounting systems necessarily reflect back and influence what metrics could potentially be used for billing.

In the wired PSTN, the points of origination and termination are generally known or knowable when the call is initiated. Once the call is initiated, these points remain stable for the duration of the call. The traffic during the call is not relevant to the bill. Typically, the only accounting datum needed after the call has been originated is the time at which the time at which it ends.

In the Internet, some things are known at the level of the *application* or *service*, while very different things are known at the level of the *network*. For VoIP, a server that implements a protocol like SIP will know the time at which a session is initiated, and may know that time at which it ends, but will know next to nothing about the network resources consumed in the interim. The *topological* location (the logical location within the network) of the originating and terminating end points will be known, but not necessarily the *geographical* location.[96]

Beyond this, an IP-based network will be dealing with a far broader array of applications than just traditional voice. The notion that the *call originator* should be viewed as the *cost causer* breaks down in the general case. In the general case, there is no obvious “right answer” to the question of how to allocate costs among end-users.

The underlying *network* knows very different things. In an IP-based environment, each IP datagram is independently addressed, and could in principle be independently routed (although routing in practice is much more stable than this implies). Relatively simple applications can generate a very large number of IP datagrams. For accounting purposes, it is necessary to summarize this data – otherwise, the accounting systems will be deluged with unmanageable data volumes.

For analogous reasons, it is trivial to measure the traffic over a given point-to-point data transmission link, but expensive and cumbersome to develop an overall traffic matrix based on end-to-end traffic destinations.

For all of these reasons, billing and accounting arrangements in the Internet have historically tended to reflect huge simplifying assumptions. For individual consumes and for enterprise customers, billing has most often been on a flat rate basis, as a function of the maximum capacity of the access link from the service provider to the customer (i.e. the price is based on the size of the "pipe", which sets an upper limit on the amount of traffic that the provider must carry).

At an enterprise level, prices have sometimes reflected the total traffic carried over the pipe, most often based on some percentile of data transmission rates (for example, a 95th percentile of rates sampled at 15 minute intervals, which will correspond roughly to average traffic for the busiest hour of the day).

It is important to note what is *not* charged for. Network operators do not assess usage-based charges for things that they cannot measure (at reasonable cost). Retail prices do not generally reflect either the distance that IP-based traffic is carried, or the degree to which international boundaries are crossed. It is simply too difficult and too expensive to measure these things. Wholesale arrangements between providers might take account of distance to some extent – the providers know the circuits between them, and can measure the point-to-point traffic over those circuits.

6.3 Reconciliation of statistics

To the extent that billing reflects usage, occasional issues and disagreements are inevitable. It is important that providers be able to reconcile their usage statistics, and that they be able to reach agreement at reasonable cost.

At the retail level, providers often choose to avoid this issue entirely by avoiding usage-based prices. At the wholesale level, the use of Bill and Keep peering arrangements also serves to reduce if not eliminate the need to reconcile statistics.

Where two providers charge one another based on traffic sent in both directions, reconciliation will be necessary. One might well imagine that, where provider A measures the traffic over a particular transmission link to provider B, that that measurement should correspond exactly to B's measurement of traffic from A to B over the same transmission link. My experience during my time in industry suggests, unfortunately, that disputes will occasionally occur, even where both parties are (most likely) acting in good faith, and even where it would seem that both parties should be measuring the same thing.

There are steps that can be taken to reduce, but not prevent, misunderstandings. Coordinating reporting start times and intervals can help. This is particularly important if the usage charges between providers depend on a percentile measure of traffic – the mean of traffic is independent of sampling interval, but the standard deviation is not. Sampling a given stream at more frequent intervals will lead to a "lumpier" distribution – a fundamental consequence of the Central Limit Theorem. If two organizations want to reach the same conclusions about a percentile, they should sample with identical frequency.

An approach that has sometimes been used – for example, in the U.S. mobile industry at one point – is to have a trusted intermediary collect and analyze the statistics. In general, the intermediary cannot itself be a competitor in the same market – otherwise, it will not be trusted.

6.4 Accounting for Quality of Service[97]

If two providers want to compensate one another for carrying their respective delay-sensitive traffic at a preferred Quality of Service, each will want to verify that the other has in fact done what it committed to do.

In the case of QoS, this would seem to imply measurements of (1) the amount of traffic of each class of service exchanged in each direction between the providers; and (2) metrics of the quality of service actually provided. Measuring the volume of traffic by class is, once again, trivial – it is no harder than measuring the overall traffic for the same transmission link. Measuring the QoS is much more complex, both at a technical level and at a business level.

For QoS, commitments between providers would presumably be primarily in terms of the mean and variance of delay. One can measure delay with primitive tools such as PING[98], or with more sophisticated tools such as IPPM probes.[99] One could imagine a pair of providers who mutually agree to instrument their networks to support one or more of these measurement tools, and to mutually measure delay between their respective

networks. One might imagine that this should be easy – one would need to agree where the probe points should be physically situated, and what measurement metrics should be employed, and one might imagine that nothing more should be needed. The reality is much more complex.

First, it is important to remember that this measurement activity implies a degree of cooperation between network operators who are direct competitors for the same end-user customers. Each operator will be sensitive about revealing the internal performance characteristics of its networks to a competitor. Neither would want the other to reveal any limitations in its network to prospective customers.

Second, there might be concerns that the measurement servers – operated within one's own network, for the benefit of a competitor – might turn into an operational nightmare, or perhaps a security exposure, within the perimeter of one's own network.

Again, there might possibly be scope for a trusted and independent third party to perform this function.

6.5 Gaming the system

If the arrangements between providers were such as to make it attractive to carry delay-sensitive traffic, then it is safe to predict that some providers will attempt, absent countermeasures, to benefit from the arrangements. Whether this should be viewed as fraud, as arbitrage, or simply as creative entrepreneurship might depend on the specific circumstances, and might be difficult to judge in practice.

For example, a network operator might discount its retail connectivity prices to end-user enterprises that operate call centers, on the theory that the resulting traffic would enable it to capture more revenue from other operators for carrying high-QoS traffic. This would seem to be a legitimate business option.

On the other hand, one could imagine an operator creating, or causing to be created, a software robot that would generate a great deal of otherwise unnecessary traffic that the operator would then have to be paid to deliver. This would seem to be a matter of arbitrage or worse, with no redeeming value.

In practice, distinguishing between appropriate and inappropriate arrangements is likely to be difficult. The actual forms that abuse might take cannot be predicted with confidence.

7 A HYPOTHETICAL SCENARIO: INTERCONNECTION IN AN NGN WORLD

In this section, we consider possible consequences of the migration to an IP-based NGN. It is a thought experiment that seeks to shed light on possible developments.

We develop a scenario, premised on the assumption that the primary incumbent in a country that operates within the regulatory framework of the European Union migrates to an IP-based NGN core.

The country is assumed, on the eve of migration, to have:

(1) an incumbent wired and wireless operator that had previously been the country's PTT, and that still has substantial market share and market power;

(2) various wired and wireless competitve operators;

(3) various independent providers of broadband Internet services, some facilities-based, some providing service competition based on procompetitive regulation (LLU, bitstream, and shared access);

(4) several independent providers of VoIP; and

(5) a number of local providers of Internet content, both web and video.

Our focus here is on IP-based NGN core migration. The characteristics of NGN access migration are, for these purposes, assumed to be possibly different in scale but similar in concept to the broadband deployment that we see today.

I have attempted to sketch a number of plausible scenarios, but I must emphasize at the outset that this is a highly speculative and perhaps controversial business. As the American baseball coach Yogi Berra once said, "It's hard to make predictions, especially about the future."

7.1 The scenario

During an extended transitional phase, the historic incumbent (BigCo for purposes of this discussion) operates traditional PSTN-based voiced services, traditional broadband and dial-up Internet access, and new integrated IP-based NGN capabilities. The NGN-based capabilities are first offered opportunistically in those areas where demand is expected to be highest and most concentrated, or in areas that required significant upgrades independent of the migration to NGN.

In the longer term, the migration to NGN will enable BigCo to achieve not only faster time-to-market for new services, but also cost savings through integration. In the near term, however, unit costs may tend to be stable or possibly to *increase*, for two reasons. First, it is unlikely to be cost-effective to decommission much of the current network until the migration is quite far advanced; and second, the need to operate two kinds of infrastructure in parallel during the transition implies increased operational expense for engineering, training, spare parts, support and operations staff, and the maintenance of software operational support systems.

Assuming a competitive retail market, BigCo is unlikely to increase prices in response to any short term increase in unit costs. They will not want to lose hard-to-replace customers to competitors. A more likely scenario is that they will hold prices steay or reduce them slightly, effectively subsidizing current customers by borrowing from anticipated future savings.

BigCo's traditional competitors will respond to perceived competitive pressure by initiating their own migration to NGN core networks, if they have not already done so. This will be prompted in part by the need to achieve economies of scale and scope closer to those of BigCo, and partly by the fear that they will otherwise be unable to compete when BigCo is eventually permitted to withdraw regulatorily mandated traditional PSTN interconnection in favor of NGN interconnection.

IP-based competitors will not perceive the need to make radical changes to their operations – they are, for the most part, already there. They will perceive a need to anticipate forthcoming IP-based NGN interconnect offerings.

As the transition phase comes to a close, BigCo will phase out traditional services on a large scale. From this point forward, the traditional services and traditional models of interconnect become less relevant.

7.2 Regulatory implications for last mile access

During the transition phase, existing regulatory oblgations for access to last mile facilities, both for traditional PSTN-based competitors and for broadband providers, will likely need to be maintained. In the near term, the last mile will continue to represent a durable competitive bottleneck in most (but not all) regions of most countries. In the near term, neither the migration to an NGN core nor the incumbent's deployment of NGN access will obviate the need for competitive access. In other words, BigCO will most likely continue to possess whatever last mile market power it had prior to the migration to NGN. In the European context, this implies the continuation of some combination of local loop unbundling (LLU), shared access, bitstream access, and resale.

For countries, or regions of countries, where three or more effective facilities-based alternative broadband options are available, and to the extent that competition appears to be effective and sustainable, it may be appropriate to eliminate or phase out these last mile obligations.

When migration is well advanced, it is possible that broadband competition will be the only meaningful last mile competition that is meaningful. There may be no further need to enable resale or LLU as an enabler for PSTN-based competition.

7.3 Regulatory implications for interconnection

During the transition phase, BigCo will still be obliged to maintain traditional PSTN interconnection capabilities. Assuming that it is possible for competitors to reach BigCo's NGN-based end-user customers through traditional interconnection, there will not necessarily be a regulatory obligation to provide new NGN-based interconnection capabilities.

BigCo will offer IP-based interconnection at some point during the transition phase. As the transition phase draws to a close, they will want to withdraw traditional interconnection. To the extent that they still possess market power, they will almost certainly be under regulatory obligations to provide NGN interconnection at cost-based prices. To the extent that the NGN implies lower forward-looking unit costs, the cost-based interconnection prices will be lower than those that pertain today.

In Europe today, all or nearly all operators that provide publicly available telephone service (PATS) are subject to regulatory obligations to interconnect, because all – even small operators, as we have seen in section 2 of this report – have significant market power in regard to the termination of telephone calls.

7.4 Peering versus transit

As we have seen, in the world of the Internet, the great majority of interconnection take the form either of peering or of transit. In our hypothetical scenario, will market participants prefer peering, transit, or some other model of interconnection? Recall that peering offers exchange of traffic only between BigCo's customers and those of its peer, but does not provide either with access to third parties. In a typical transit relationship, by contrast, the transit customer can use the transit provider's network to reach destinations anywhere on the Internet.

7.4.1 Peering versus Transit for international interconnection

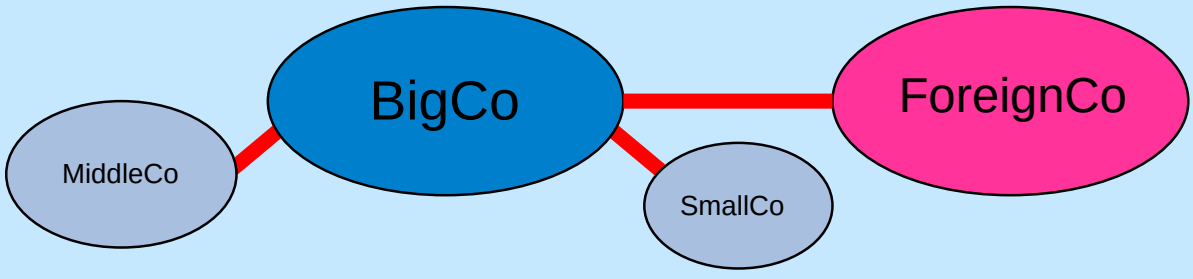
We start by considering BigCo's relationship to similarly situated operators in other countries. Experience to date strongly suggests that these arrangements will tend to be peering relationships. Historically, peering arrangements have usually been on a Bill and Keep basis; however, in an NGN world that supports differentiated QoS, it is possible that BigCo and its peer might agree to one level of charges for conventional best efforts traffic and another, higher level of charges for traffic with preferred QoS. In fact, there could be more than two levels.

On the other hand, BigCo is unlikely to agree to peer with tiny competitive operators, either in other countries or for that matter in BigCo's own country. This implies that tiny, competitive operators will generally need to contract with some transit provider (but not necessarily BigCo).

There is likely to be an extended period of coexistence, where BigCo interconnects with some operators (especially foreign operators) by peering, with others by transit, and with quite a few others by means of traditional PSTN interconnection. Internationally, traditional PSTN interconnection will surely persist.

There is also a matter of transaction costs – each interface migration from a PSTN basis to an NGN/IP basis implies certain real transition costs, as well as transaction costs associated with creating and managing new interconnection agreements. Overnight mass migration cannot be cost-effective. This implies that BigCo will, other things being equal, first seek out IP-based interconnection arrangements with those operators with which the agreements provide it with the greatest benefit, which might tend to be those similarly situated operators with which it exchanges the largest volume of traffic.

Figure 7.1: Hypothetical peering arrangements



The transition costs pose a regulatory challenge as well. To the extent that BigCo unilaterally chooses to massively re-shape its network in the NGN world, possibly withdrawing network interconnection points, what are its obligation to competitive providers with which it has existing arrangements? It seems inappropriate that competitive providers should be involuntarily burdened with new costs that are not of their making; at the same time, BigCo should not be forced to maintain obsolete interconnection points indefinitely. These complicated trade-offs have been a central theme in several Ofcom (UK) public consultations on the migration to the NGN.[100]

Finally, we note that an incentives problem could easily arise that could slow or prevent the migration to next generation international interconnections. The existing arrangements tend to transfer significant sums of money from one operator to another, either because mobile rates are much higher than fixed, or because far more calls are initiated from developed countries to developing ones than *vice versa*. The migration to peering is likely to result either in Bill and Keep or in cost-based arrangements, which would either reduce or eliminate the subsidies. This means that two operators that contemplate a migration from current arrangements to IP-based peering are likely to perceive the change as a zero-sum game – one provider will benefit from the change, and one will suffer. Under those assumptions, the provider that is negatively impacted can reasonably be expected to refuse to make the transition, or, if somehow compelled to upgrade, to delay the transition as long as possible.

7.4.2 Peering versus transit for domestic interconnection within BigCo's country

As previously noted, BigCo is unlikely to be motivated to offer peering arrangements to tiny competitive operators in its own country. It might offer peering arrangements to just a few of its largest domestic competitors.

A difference between this case and the international case is that these competitive operators will be highly motivated to have good connectivity to BigCo's customers. (To the extent that BigCo's customer base is much larger than that of its competitors, it will tend to prefer less-than-perfect interconnection with small competitors. This is a straightforward application of the Katz-Schapiro result discussed in section 2 of this paper.[101])

At that point, small domestic competitors have limited options:

(1) As long as traditional PSTN interconnect is offered, and to the extent that it is sufficient for the competitor's needs, they might stick with PSTN interconnect.

(2) They can purchase transit service from BigCo.

(3) They can purchase transit service from some provider other than BigCo.

My prediction is that many of the small domestic providers would choose to purchase transit service from BigCo (perhaps in addition to service from some other transit provider) as long as BigCo's price is competitive.

As long as the market for wholesale transit services is reasonably competitive (and assuming that BigCo also faces an effectively competitive market for broadband Internet access), this should lead to quite reasonable domestic outcomes. BigCo's wholesale price for transit service will be constrained by competition from third parties. BigCo's competitors need access to BigCo's customers, and will prefer the best connection that they can afford, *but they can reach BigCo's customers perfectly well through a third party transit provider*.

This is an important distinction between the NGN world and the PSTN world. In the IP-based world, indirect interconnection is perfectly reasonable.

To the extent that peering arrangements with domestic competitors either are on a Bill and Keep basis, or that they reflect roughly balanced net payments,[102] and to the extent that underlying facilities are available on a competitive or a nondiscriminatory basis, the competitors' costs to reach BigCo's customers should not greatly exceed those of BigCo itself (except to the extent that BigCo enjoys advantages of scale). Consequently, competition from these domestic competitors should appropriately constrain BigCo's behavior, and prices are likely to be competed down to levels not greatly in excess of marginal cost.

Foreign peers would experience somewhat higher costs in competing for BigCo's domestic end user customers, but only to the extent that their costs are impacted by lacking a local base of operations.[103]

Potential competition from foreign service providers thus provides a second (albeit looser) constraint on BigCo's pricing power. If BigCo were to attempt to price well in excess of cost, these foreign providers might be motivated to establish a presence in BigCo's country so as to compete directly.

To re-cap, this implies that the likely domestic pattern is:

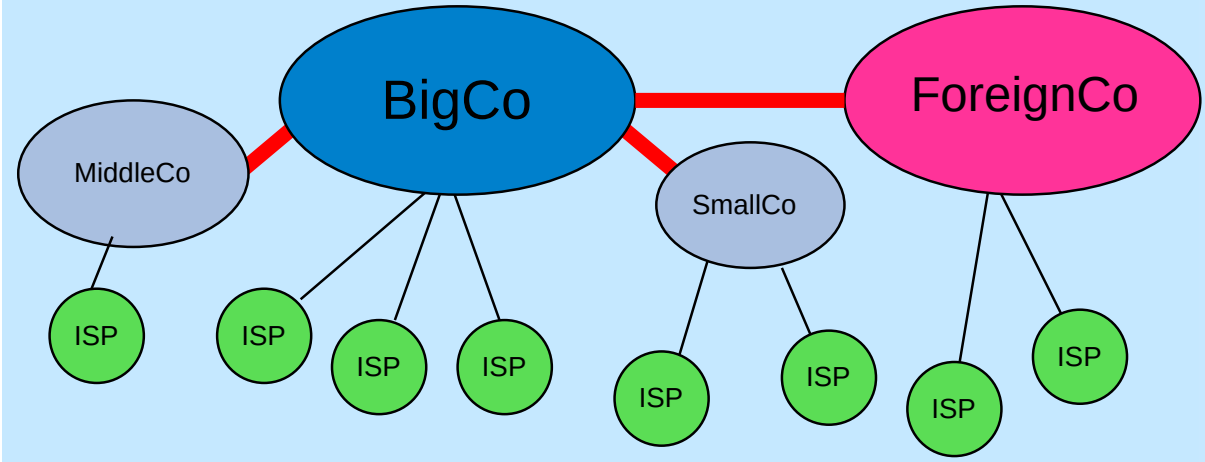
(1) a few of the largest competitors might peer with BigCo;

(2) small competitors will purchase transit from BigCo if they can;

(3) small competitors will supplement or replace BigCo transit with transit from third parties; and

(4) small competitors may choose, as an economic optimization, to peer with one another whenever the traffic that they can exchange reduces their transit costs sufficiently to pay for the cost of any peering circuits and infrastructure.[104]

Figure 7.2: Hypothetical peering and transit arrangements



This returns us to a key question of regulatory policy. It is natural to assume that BigCo's existing PSTN market power as regards interconnection will automatically confer market power as regards interconnection in the NGN world, and that any interconnection remedies therefore need to automatically carry over to NGN interconnection; I would argue, however, that making this presumption today would be greatly premature. For the reasons outlined above, it is entirely possible (given adequate competition or effective regulatory access to necessary underlying facilities such as leased lines, wholesale transit and broadband Internet access) that unregulated IP-based interconnection will lead to a perfectly satisfactory Coasian solution – a solution which would likely be superior to anything that a regulator could craft.

7.5 Network provider versus application servce provider

In the world of the NGN, the terminating monopoly requires some re-thinking. The end-user may get his or her broadband connection from BigCo, or from a competitive broadband Internet access provider. He may get his voice telephony service – assuming that the service continues to look much as it does today – from BigCo, or he may get it from an alternative VoIP service provider. For telephone calls, if anyone possesses a termination monopoly, it is the VoIP service provider, not the provider of the broadband pipe.

Who, then, should collect the termination charge? It is important to remember that termination costs exist to recompense the terminating carrier for the incremental usage-based costs imposed on its network. An independent VoIP provider has no network, and experiences very little incremental usage-based cost.

Recall, too, that the network provider has only limited visibility into the traffic that it is carrying. The network provider could, however, assess a surcharge for packets where the user explicitly requests preferred

Quality of Service; however, if the charge is high, the user will probably prefer services that operate with standard best-efforts QoS (which will, as previously noted, still provide perfectly adequate voice quality in general). The network operator could conceivably attempt to monitor the user's service in order to assess a surcharge for voice traffic (leaving aside for the moment the possible invasion of privacy that this implies), whether associated with preferred QoS or not; however, if the surcharge were large, users might again respond by encrypting their traffic to prevent the network provider from inspecting it. Technology could conceivably close any or all of these holes, but there is no obvious social benefit in doing so. To the contrary, consumer welfare would appear to be maximized by giving consumers as much latitude as possible to do what they want to do, with as few restrictions as possible.

It also bears noting that it costs the network no more to carry a VoIP packet (on a best efforts basis) than it does to carry a WorldWide Web packet, or any other data packet for that matter. Moreover, the marginal usage-based cost per packet is very, very low.

Yet another challenge relates to cost causation. Historically, it has been assumed that party that originates the call is the sole cost causer. This assumption has always been questionable. Going forward, it will be difficult if not impossible to ascribe cost to one or another party to a communication.

7.6 Implications for differentiated Quality of Service

Within individual IP-based networks, differentiated QoS has existed for many years.

If BigCo prices Internet transit competitively, many competitive operators are likely to choose to procure transit service from BigCo. This positions BigCo to offer QoS-capable access to its competitors, not only to BigCo's own customers, but also to the customers of most domestic competitors.

For reasons noted in section 3 of this paper, inter-provider QoS has been slow to deploy in connection with peering interconnection. Paradoxically, offering it in connection with transit service could be less problematic, provided that it is offered at a price that is not disproportionate to the benefits that it provides. In this scenario, the network externalities advantage that BigCo enjoys by virtue of its large customer base positions it to provide QoS capable transit to most or all competitors on the national market.

This is not a model that a regulator will hasten to embrace, since it implies a unique role for the country's historic incumbent provider. Given the limited benefits that differentiated QoS confers, however, it might represent a quite reasonable trade-off. Whatever market power these arrangements confer on BigCo in regard to QoS would appear to be of limited value.

At the same time, these arrangements do not necessarily lead to a global NGN with ubiquitous support for differentiated QoS. Transaction costs are likely to continue to inhibit implementation of diffentiated IP QoS at the level of peering relationships; consequently, differentiated QoS at the international level is likely to have at best a spotty availability for an extended period of time, even in the event that most service providers ultimately migrate to NGN and to IP-based NGN interconnection.

7.7 Policy implications

With all of this in mind, my view is that interconnection arrangements in an NGN world are likely to be most rational and sustainable to the extent that they adhere to a few guiding principles:

(1) Wherever competitive conditions warrant, a Coasian solution reflecting market-based negotiations between the NGN operators is likely to lead to more efficient solutions than a regulatory rate-setting.

(2) National regulatory authorities might therefore be well advised to focus their attention primarily on ensuring adequate competition for wholesale Internet transit services, and for consumer broadband Internet access.

Where a Coasian resolution is not feasible, the following considerations follow from the previous discussion:

(3) The wholesale charge assessed should either be zero (i.e. Bill and Keep), or should be no higher than the forward-looking marginal usage-based cost associated with carrying the incremental traffic.

(4) As a corollary, incremental charges are appropriate only to the extent that they are associated with incremental costs.

(5) Charging should reflect only things that can be measured in a straightforward and documentable way by the party that assesses the charges.

(6) Charges could reasonably consider the volume of traffic exchanged at each level of QoS requested (and delivered), but should otherwise be independent of the nature of the application employed by the user.

ENDNOTES

[1] Charles Dickens, A Christmas Carol.

[2] See http://www.btglobalservices.com/business/global/en/business/business_innovations/issue_02/century_network.html.

[3] See Haucap, J., and Marcus, J.S., "Why Regulate? Lessons from New Zealand", IEEE Communications Magazine, November 2005, available at: http://www.comsoc.org/ci1/Public/2005/nov/ (click on "Regulatory and Policy").

[4] See Ofcom's Final statements on the Strategic Review of Telecommunications, and undertakings in lieu of a reference under the Enterprise Act 2002, September 22, 2005.

[5] "ECTA comments on NGN public policy", November 2005.

[6] Section 7 benefits from recent developments and from regulatory proceedings in the UK, but the scenario is not patterned after the proposed BT evolution, nor after specific developments in any particular country.

[7] Jean-Jacques Laffont, Patrick Rey and Jean Tirole, "Network Competition: I. Overview and Nondiscriminatory Pricing" (1998a), Rand Journal of Economics, 29:1-37; and "Network Competition: II. Price Discrimination" (1998b), Rand Journal of Economics, 29:38-56.

[8] Armstrong, M. "Network Interconnection in Telecommunications." Economic Journal, Vol. 108 (1998), pp. 545–564.

[9] Jean-Jacques Laffont and Jean Tirole, Competition in Telecommunications, MIT Press, 2000.

[10] I should hasten to add that I myself am not formally trained as an economist.

[11] See Federal Communications Commission (FCC) Office of Strategic Planning and Policy Analysis (OSP) Working Paper 33: Patrick deGraba, "Bill and Keep at the Central Office As the Efficient Interconnection Regime", December 2000, available at http://www.fcc.gov/osp/workingp.html.

[12] "On the receiver pays principle", RAND Journal of Economics, 2004.

[13] Andrew Odlyzko, "The evolution of price discrimination in transportation and its implications for the Internet", Review of Network Economics, vol. 3, no. 3, September 2004, pp. 323-346, available at http://www.rnejournal.com/articles/odlyzko_RNE_sept_2004.pdf.

[14] Cf. FCC, 8th CMRS Competition Report, §94: "AT&T Wireless's Digital One Rate ("DOR") plan, introduced in May 1998, is one notable example of an independent pricing action that altered the market and benefited consumers. Today all of the nationwide operators offer some version of DOR pricing plan which customers can purchase a bucket of MOUs to use on a nationwide or nearly nationwide network without incurring roaming or long distance charges." Several mobile operators offer a variant of this plan where there are no roaming charges as long as the customer is using that operator's facilities.

[15] These flat rate plans are truly flat rate, whereas the mobile plans are generally two part tariffs. The usage charges of the mobile plans are usually set to very high levels (as much as $0.40 per Minute of Use). They are not so much intended to be used, as to punish consumers who purchase bundles that are too small. The common feature between the mobile plans and the newer truly flat rate plans is a movement away from meaningful usage charges.

[16] For example, Vonage offers unlimited calls to or from the U.S. and Canada for just $24.99 a month. See www.vonage.com.

[17] In the United States, by means of the Enhanced Service Provider (ESP) exemption; in the UK, by means of FRIACO.

[18] This definition is adapted from Laffont and Tirole (2001), page 182.

[19] In the interest of simplicity, we will gloss over the historically important distinction between access charges and reciprocal compensation in the United States. As the industry consolidates (with the disappearance of AT&T and MCI as independent long distance carriers), this distinction is somewhat less relevant than it once was. For a more detailed treatment of arrangements in the U.S., see Marcus, "Call Termination Fees: The U.S. in global perspective", presented at the 4th ZEW Conference on the Economics of Information and Communication Technologies, Mannheim, Germany, July 2004. Available at:

ftp://ftp.zew.de/pub/zew-docs/div/IKT04/Paper_Marcus_Parallel_Session.pdf.

[20] In 2001, the FCC signaled its intent to migrate to a much broader implementation of Bill and Keep; however, this regulatory policy change has been stalled for years. See FCC, In the Matter of developing a Unified Intercarrier Compensation Regime, CC Docket 01-92, released April 27, 2001.

[21] See Laffont, Rey and Tirole (1998a) and (1998b); Armstrong (1998); Laffont and Tirole (2001), all op. cit. See also Cave et. al. (2004); de Bijl et. al. (2004); and Haucap and Dewenter (2004).

[22] See FCC OSP Working Paper 33: Patrick DeGraba, "Bill and Keep at the Central Office As the Efficient Interconnection Regime", and Working Paper 34: Jay M. Atkinson, Christopher C. Barnekov, "A Competitively Neutral Approach to Network Interconnection", both December 2000, both available at http://www.fcc.gov/osp/workingp.html; Stephen C. Littlechild, "Mobile Termination Charges: Calling Party Pays vs Receiving Party Pays", forthcoming, available at http://www.econ.cam.ac.uk/dae/repec/cam/pdf/cwpe0426.pdf; Robert W. Crandall and J. Gregory Sidak, "Should Regulators Set Rates to Terminate Calls on Mobile Networks?", Yale Journal on Regulation, 2004; and Marcus (2004), op. cit.

[23] Laffont and Tirole, Competition in Telecommunications (2001), page 186. The italics are theirs. See also Haucap and Dewenter (2005).

[24] There are, of course, numerous exceptions and caveats to this statement. See chapter 5 of Laffont and Tirole (2001).

[25] See Martin Cave, Olivier Bomsel, Gilles Le Blanc, and Karl-Heinz Neumann, How mobile termination charges shape the dynamics of the telecom sector, July 9, 2003; Paul W.J. de Bijl, Gert Brunekreeft, Eric E.C. van Damme, Pierre Larouche, Natalya Shelkoplyas, Valter Sorana, Interconnected networks, December 2004; Littlechild (2006); and Marcus (2004).

[26] See European Commission, 10th Implementation Report (December 2004); and Marcus (2004)

[27] Laffont and Tirole (2001), page 190.

[28] Milgrom et. al. suggest that this is the economically predicted result for Internet backbones. See Paul Milgrom, Bridger Mitchell and Padmanabhan Srinagesh, "Competitive Effects of Internet Peering Policies", in The Internet Upheaval, Ingo Vogelsang and Benjamin Compaine (eds), Cambridge: MIT Press (2000): 175-195. At: http://www.stanford.edu/~milgrom/publishedarticles/TPRC%201999.internet%20peering.pdf.

[29] To understand the motivation for this, see Laffont and Tirole (2001) pages 201-202.

[30] An operator might choose to ignore a termination fee that constitutes only a small fraction of the total cost of the call. Termination fees set in the absence of regulation often represent the preponderance of the total cost of the call.

[31] In support of this interpretation, it is worth noting that flat rate mobile plans are common in Europe, but generally only for on-net calls and for calls to the fixed network. The calls that are excluded are precisely the calls to off-net mobile subscribers – the calls where termination fees would tend to be high.

[32] See Federal Communications Commission (FCC) Office of Strategic Planning and Policy Analysis (OSP) Working Paper 33: Patrick DeGraba, "Bill and Keep at the Central Office As the Efficient Interconnection Regime", December 2000, at 95, available at http://www.fcc.gov/osp/workingp.html.

[33] See footnote 69 on page 28.

[34] Mobile termination fees in the European Union are increasingly subject to regulation, but this is still in the process of being phased in. The fees today continue to reflect to a significant degree the previous unregulated arrangements.

[35] See FCC, Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 10th Report (10th CMRS Competition Report), July 2005, Table 10, based on Glen Campbell et al., Global Wireless Matrix 4Q04, Global Securities Research & Economics Group, Merrill Lynch, Apr. 13, 2005.

[36] Cf. Crandall and Sidak (2004).

[37] The per-minute fees in US "bucket of minutes" plans are probably not exercised much in practice. They are so high, in comparison with the bucket of minutes arrangements, as to serve primarily as a punitive measure to force users to upgrade to a larger bucket. At the same time, this implies that, for a given user over time, consuming more minutes will equate to higher charges.

[38] In analyzing European experience, Cave et. al. (2003) find that only a small portion of the subsidy is returned to the consumer.

[39] One might well expect a corresponding tendency for CPP/CPNP arrangements to slow the ongoing adoption of the services from which the subsidies are generated, that is, fixed services. Eastern European experience might possibly support this view – for example, mobile phone penetration in Hungary is more than 70% and growing, while fixed phone penetration is about 40% and declining. I am not aware of any rigorous analysis on this question.

[40] See, for example, Cave et. al. (2003); Littlechild (2006); and Crandall and Sidak (2004), all op. cit.

[41] The penetration is usually computed by dividing the number of subscriptions by the population. For penetration to exceed 100% implies that some consumers have more than one subscription at the same time. This probably reflects either (1) pre-paid cards that are nominally active but no longer being used, or (2) consumers who find it cost-effective to place on-net calls on more than one network.

[42] Crandall and Sidak (2004), op. cit.

[43] Report of the NRIC V Interoperability Focus Group, "Service Provider Interconnection for Internet Protocol Best Effort Service", page 7, available at http://www.nric.org/fg/fg4/ISP_Interconnection.doc.

[44] Ibid., pages 4-6. See also Marcus, Designing Wide Area Networks and Internetworks: A Practical Guide, Addison Wesley, 1999, Chapter 14.

[45] The current number is probably far less.

[46] See http://www.apnic.net/services/asn_guide.html.

[47] A very innovative paper by Prof. Gao of the University of Amherst confirms this structure. See Lixin Gao, "On inferring autonomous system relationships in the Internet," in Proc. IEEE Global Internet Symposium, November 2000. The Internet is probably more richly interconnected today than was the case in 2000, but there is no reason to believe that these basic aspects have changed very much.

[48] Paul Milgrom, Bridger Mitchell and Padmanabhan Srinagesh, "Competitive Effects of Internet Peering Policies", in The Internet Upheaval, Ingo Vogelsang and Benjamin Compaine (eds), Cambridge: MIT Press (2000): 175-195. At: http://www.stanford.edu/~milgrom/publishedarticles/TPRC%201999.internet%20peering.pdf.

[49] See M. Katz and C. Shapiro (1985), "Network externalities, competition, and compatibility", American Economic Review 75, 424-440.; and J. Farrell and G. Saloner (1985), `Standardization, compatibility and innovation', Rand Journal of Economics 16, 70-83.

[50] Jacques Cremer, Patrick Rey, and Jean Tirole, Connectivity in the Commercial Internet, May 1999.

[51] Milgrom et. al., "Competitive Effects of Internet Peering Policies" (2000).

[52] Private communication, Marius Schwarz, Georgetown University.

[53] Armstrong, M. "Network Interconnection in Telecommunications." Economic Journal, Vol. 108 (1998), pp. 545–564.

[54] Laffont, J.-J., Rey, P., And Tirole, J. "Network Competition: I. Overview and Nondiscriminatory Pricing." RAND Journal of Economics, Vol. 29 (1998a), pp. 1–37.

[55] Laffont, J.-J., Marcus, J.S., Rey, P., And Tirole, J., "Internet interconnection and the off-net-cost pricing principle", RAND Journal of Economics, Vol. 34, No. 2, Summer 2003, available at http://www.rje.org/abstracts/abstracts/2003/rje.sum03.Laffont.pdf. A shorter version of the paper appeared as "Internet Peering", American Economics Review, Volume 91, Number 2, May 2001.

[56] This is not to suggest that all parties have been satisfied with the results. An ongoing dispute over International Charging Arrangements for Internet Service (ICAIS) has been simmering for some years now.

[57] Ronald H. Coase, "The Federal Communications Commission", Journal of Law and Economics 2 1-40, 1959.

[58] Industry consolidation is another noteworthy contributory factor.

[59] FCC, In the Matter of developing a Unified Intercarrier Compensation Regime, CC Docket 01-92, released April 27, 2001, section 95. See also Marcus, "Call Termination Fees: The U.S. in global perspective", July 2004, available at: ftp://ftp.zew.de/pub/zew-docs/div/IKT04/Paper_Marcus_Parallel_Session.pdf.

[60] Andrew Odlyzko has written a number of insightful papers exploring the historical roots of price discrimination, and the relevance to the Internet. See Andrew Odlyzko, "The evolution of price discrimination in transportation and its implications for the Internet", Review of Network Economics, vol. 3, no. 3, September 2004, pp. 323-346, available at http://www.rnejournal.com/articles/odlyzko_RNE_sept_2004.pdf.

[61] See the classic paper by the Stanford University mathematician Harold Hotelling, "Stability in Competition", The Economic Journal, March 1929, pages 41-57.

[62] The Hotelling paper argues, in fact, the providers will tend to prefer to provide products very much like those of their competitors, even at the cost of leaving some demand only imperfectly satisfied.

[63] See, for example, Joskow, P., "Regulation and Deregulation after 25 Years: Lessons Learned for Research in Industrial Organization", 2004, pages 26-27, available at:

http://econ-www.mit.edu/faculty/download_pdf.php?id=1005.

[64] Alfred E. Kahn, "Whom the Gods would Destroy, or How not to Deregulate", available at http://www.aei.brookings.edu/admin/authorpdfs/page.php?id=112.

[65] Jean-Jacques Laffont, J. Scott Marcus, Patrick Rey, and Jean Tirole, "Internet interconnection and the off-net-cost pricing principle", RAND Journal of Economics, Vol. 34, No. 2, Summer 2003, available at http://www.rje.org/abstracts/abstracts/2003/rje.sum03.Laffont.pdf.

[66] This is not altogether true. My former firm, BBN, operated a commercial RSVP-based network for many years. It was a commercial failure, but not a technical failure.

[67] Those of us who remember international telephone calls routed over satellites are familiar with this phenomenon.

[68] For an introduction to the use of queuing theory in this context, see Chapter 16 of my textbook, Designing Wide Area Networks and Internetworks: A Practical Guide, Addison Wesley, 1999.

[69] The graph was computed using the Pollaczek-Khinchine formula for an M/G/1 queuing model. This implies a Markovian arrival pattern; however, the so-called operational analysis school of queueing theory has demonstrated that the formula can also be derived under greatly relaxed assumptions. A mean packet length of 284 octets is assumed, consistent with observational experience around 2001.

[70] This was, of course, the key root problem in BBN's inability to successfully commercialize its RSVP-based commercial QoS-capable network.

[71] In a classic joke, a child looks for a lost coin under a lamp post, not because he lost it there, but rather because that is where the light is best.

[72] Jeffrey H. Rohlfs, Bandwagon Effects In High-Technology Industries 3 (2001). Much of the discussion in this section derives from Rohlfs's excellent book.

[73] I make this case at much greater length in "Evolving Core Capabilities of the Internet", Journal on Telecommunications and High Technology Law, 2004.

[74] If each of n interconnected networks need to reach agreement with every other network, this implies a need for n(n-1)/2 interconnection agreements. The number of agreements goes up as the square of the number of networks.

[75] The thought here is to provide examples of contractual arrangements that seem to work, but emphatically not to intrude on the ability of commercial service providers to conclude whatever arrangements they might choose.

[76] Including slow rolling, cost-price squeezes, and strategic litigation.

[77] In the absence of regulation, these behaviors can arise quickly and spontaneously. In the United States in the early 1900's, it was a refusal of AT&T to interconnect with competitors that led to the Kingbury Commitment of 1912, and ultimately to the regulation of telecommunications.

[78] Justus Haucap and J. Scott Marcus, "Why Regulate? Lessons from New Zealand", IEEE Communications Magazine, November 2005, available at: http://www.comsoc.org/ci1/Public/2005/nov/ (click on "Regulatory and Policy").

[79] See M. Katz and C. Shapiro (1985), "Network externalities, competition, and compatibility", American Economic Review 75, 424-440.; and J. Farrell and G. Saloner (1985), "Standardization, compatibility and innovation", Rand Journal of Economics 16, 70-

83. Two threads of economic research, one related to standards, the other to interconnection, proceeded in parallel for many years. Only later did the economists realize that the underlying economics were nearly identical.

[80] Jacques Cremer, Patrick Rey, and Jean Tirole, Connectivity in the Commercial Internet, May 1999. Note that the author was a prominent intervener in both cases.

[81] The emergence of "network neutrality" as a hot issue in the United States may reflect a recognition or belief, perhaps not fully understood or articulated, that broadband Internet providers in the U.S. might be approaching this threshold. Whether this is really so remains unclear.

[82] It must, however, be noted that a framework of this type requires some sophistication as regards economics. Moreover, the effectiveness of implementation depends on institutional arrangements that enable economic tests to be applied impartially and transparently.

[83] Nicholas Garnham, "Contradiction, Confusion and Hubris: A Critical Review of European Information Society Policy", available at http://www.encip.org/document/garnham.pdf.

[84] See Haucap, J., and Marcus, J.S., "Why Regulate? Lessons from New Zealand", IEEE Communications Magazine, November 2005, available at: http://www.comsoc.org/ci1/Public/2005/nov/ (click on "Regulatory and Policy").

[85] Ofcom's approach to risk in the assessment of the cost of capital: Final statement, August 18, 2005

[86] Ofcom defines a real option as "… the term given to a possibility to modify a project at a future point." It relates to "… the option for a firm that faces significant demand uncertainty to 'wait and see' how the demand or technology for a new product will evolve before making an investment."

[87] Content in this context should be construed broadly. It is any information that one might possibly access using the Internet. It could be a website, or a movie, or an audio recording.

[88] "The chief executive of AT&T, Edward Whitacre, told Business Week last year that his company (then called SBC Communications) wanted some way to charge major Internet concerns like Google and Vonage for the bandwidth they use. 'What they would like to do is use my pipes free, but I ain't going to let them do that because we have spent this capital and we have to have a return on it,' he said." New York Times, March 8, 2006

[89] FCC, "In the Matter of Madison River Communications, LLC and affiliated companies", available at:

http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-05-543A1.pdf.

[90] The author has no first hand knowledge of this case.

[91] Marcus, J.S., "Is the U.S. Dancing to a Different Drummer?" Communications & Strategies, no. 60, 4th quarter 2005. Available at: http://www.idate.fr/fic/revue_telech/132/CS60%20MARCUS.pdf.

[92] Referred to in this context as international settlement rates.

[93] I am not a neutral party in the matter. I have an ongoing relationship with the Jamaican regulatory authority.

[94] FCC, Modifying the Commission's Process to Avert Harm to U.S. Competition and U.S. Customers Caused by Anticompetitive Conduct, IB Docket No. 05-254, Released: August 15, 2005.

[95] See also the remarks of Thilo Salmon of SIPgate at the recent NGN and Emerging Markets workshop, Koenigswinter, Germany, December 5 2006.

[96] The IP address reflects the topological location. The telephone number implies a geographic location, but that implication will not necessarily be reliable for "nomadic" services, including VoIP.

[97] In this section in particular, I am my own primary source. When I was in industry as a Chief Technology Officer for a major Internet backbone service provider, I was very active in trying to evolve peering arrangements to accommodate QoS.

[98] Mike Muuss, "The Story of the PING Program", available at http://ftp.arl.mil/~mike/ping.html.

[99] See, for instance, http://www.ripe.net/projects/ttm/about.html.

[100] See Next Generation Networks – Future arrangements for access and interconnection, October 24, 2004; and Next Generation Networks: Further consultation, June 30, 2005.

[101] See M. Katz and C. Shapiro (1985), "Network externalities, competition, and compatibility", American Economic Review 75, 424-440.; and J. Farrell and G. Saloner (1985), "Standardization, compatibility and innovation", Rand Journal of Economics 16, 70-83.

[102] If BigCo were to refuse to peer on reasonable terms with any domestic competitors, it is possible that regulatory intervention might be appropriate. There are parallels to the circumstances that pertained in Australia a few years ago, where the government considered it necessary to impose a peering obligation on their historic incumbent.

[103] For example, circuits from a foreign provider to a commercial end user will be longer and more expensive than circuits from BigCo, in general.

[104] These "secondary" peering arrangements will tend to emerge spontaneously, without regulatory intervention. They are already evident, for example, among VoIP providers in the UK.