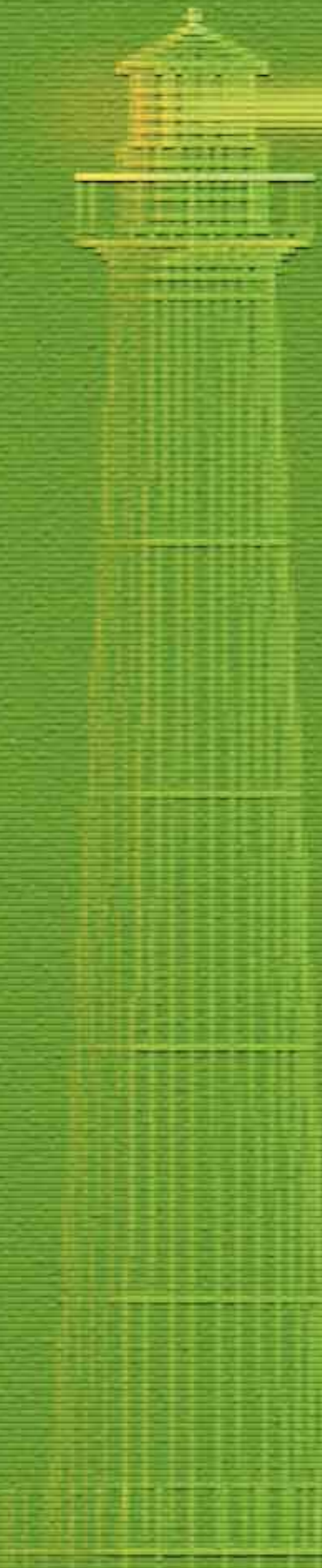
International Telecommunication Union



GSR 2007

DISCUSSION PAPER

NGN Interconnection and Access

Comments are welcome and should be sent by 1 March 2007 to GSR07@itu.int

International Telecommunication Union

GLOBAL SYMPOSIUM FOR REGULATORS

Dubai World Trade Center
Dubai, United Arab Emirates

5-7 February 2007

Work in progress, for discussion purposes

INTERCONNECTION ON AN IP-BASED NGN ENVIRONMENT

PREPARED BY J. SCOTT MARCUS
WIK-CONSULT GMBH

COMMENTS ARE WELCOME AND SHOULD BE SENT BY 1 MARCH 2007 TO
gsr07@itu.int

TABLE OF CONTENTS

GSR DISCUSSION PAPER

INTERCONNECTION IN AN IP-BASED NGN ENVIRONMENT

*This paper has been prepared by **J. Scott Marcus**, WIK-CONSULT GmbH, as an input document for the 2007 Global Symposium for Regulators (GSR), organized by the Telecommunication Development Bureau (BDT). The views expressed in this paper are those of the author and do not necessarily reflect the opinions of the ITU or its membership. Comments are welcome and should be sent to gsr07@itu.int by 1 March 2007.*

1. Interconnection in an IP-based NGN environment

This chapter addresses interconnection and access in an IP-based NGN environment, from an economic, technical and regulatory perspective.

It is worth noting at the outset that this chapter deals with both *interconnection* and with *access*. These topics are so familiar to regulators that we often lump them together without clearly distinguishing them; moreover, the conventional definitions tend to be so turgid and technical as to shed little light on what is really meant.[1] Access and interconnection are related, but they are not the same thing. For our purposes, *interconnection* enables an operator to establish communications with the customers of another operator, while *access* enables an operator to utilize the facilities of another operator in the furtherance of its own business and in the service of its own customers.[2]

Regulation is frequently needed in support of access and of interconnection. Whether regulation is appropriate in a particular case depends on the specifics of the market in question, especially on the degree to which that market is competitive, and also on the ease with which the interconnection or the facilities in question can be bypassed or replicated.

A key question that this chapter considers is the degree to which regulation of access and interconnection will be necessary in the emerging world of the NGN. How does the emergence of NGN alter market power, and the ease of bypass and replication?

The chapter seeks to apply economic reasoning, drawing on the substantial economic analysis that has been done to date of the Public Switched Telephone Network (PSTN) and the Internet, and also on substantial practical experience with both systems. There is no corresponding base of theory or practice as yet for IP-based NGNs. To a first order, it is reasonable to assume that the economic forces driving NGN market players will be similar to those that today motivate their counterparts who run the networks that most resemble IP-based NGNs.

The chapter also considers the relative merits of different wholesale arrangements. Most of the world uses a system known as Calling Party's Network Pays (CPNP); however, a less widely used system known as Bill and Keep offers a number of advantages, particularly for countries where adoption of ICT services is already well advanced. The time of migration to an NGN is a natural time for a country to consider whether its wholesale interconnection arrangements could profitably evolve to Bill and Keep.

Section 5.1 briefly discusses the objectives and rationale for regulation. Section 5.2 considers technical constraints, while Section 5.3 explores specific challenges. Section 5.4 provides general background on the economic theory of interconnection, at wholesale and retail level, both for the Public Switched Telephone Network (PSTN) and for the Internet. Section 5.5 considers whether regulators will need to set prices in the world of the NGN, and at what level if so. Section 5.6 reviews studies and proceedings conducted by a number of regulatory bodies in developed and developing countries. Section 5.7 broadens the discussion to compare fixed access concerns to those of mobile (with particular consideration of the implications for developing countries). Section 5.8 briefly considers last mile access issues. Section 5.9 provides concluding observations.

1.1 Objectives of regulation

The broad societal objectives that the regulator seeks through interconnection arrangements are largely those that telecommunications regulators everywhere seek through all of their actions[3]: to make electronic

communication services available to all of their residents at reasonable cost. The regulator would be well advised to step back and, wherever possible, rely on competitive market mechanisms to do so; however, for a variety of reasons, certain regulatory interventions tend to be necessary in most if not all countries. This is just as true for the NGN as it is for traditional networks today.

1.1.1. Addressing potential barriers to competition

The most noteworthy rationale for intervention is to deal with barriers to competitive entry. Wired telecommunications were long thought to be a natural monopoly. Initially, wired telecommunications services were generally provided either by the government or by some government-sanctioned private monopoly provider. Today, most countries encourage competitors to enter the market. The historic incumbent provider has every incentive to seek to hinder the entry of these new competitors into the market.

Competitive entrants cannot hope to successfully offer a mass market service unless they can connect their own customers to the historic incumbent's customers. The incumbent is likely to be motivated to delay or deny this access, or to make access as expensive as possible in order to raise the new entrant's costs, and thus render the new entrant a less effective competitor. Indeed, attempts to withhold interconnection are among the most common techniques used by incumbents seeking to maintain their market power.

These actions could be viewed as an anticompetitive barrier to competitive entry; unfortunately, experience strongly suggests that the application of competition law alone is not enough to enable competitive entry.[4] Where market power is entrenched, it is necessary to impose regulations in advance (*ex ante*) in order to ensure that efficient competitors can successfully achieve market entry.

1.1.2. Consumer benefits

A regulator would seek to encourage, or at least not to impede, many other societal goals through interconnection policy, including:

- Price/performance of services
- Availability of useful and innovative services
- Prospects for informed consumer choice

Access and interconnection have often implied wholesale payments between operators at regulated rates. Regulatory imposition of payments always implies a risk of creating economic distortions that could impede investment. Consequently, regulators need to balance carefully their interventions, and avoid needless meddling.

1.2 Universal service / universal access

All countries seek to ensure that some minimum, crucial set of electronic communication services are available to all residents at reasonable cost. Different countries fund universal access in different ways. International and domestic interconnection charges (call termination fees) have played an important role in financing universal service in a number of countries. The migration to NGN is putting downward pressure on call termination fees, and may ultimately make current call termination arrangements unsustainable. Should that prove to be the case, how are such countries – especially developing countries – to finance universal service? To what extent, if any, is it necessary to subsidize universal access or universal service in the world of the NGN?

1.3 Communications as an enabler to overall growth and prosperity

ICTs have long been recognized as a key enabler of economic growth. The migration to Next Generation Networks is expected to make a wealth of interrelated services available to the public. The significance of ICTs to societal growth is likely to be even greater with tomorrow's networks than it is with today's.

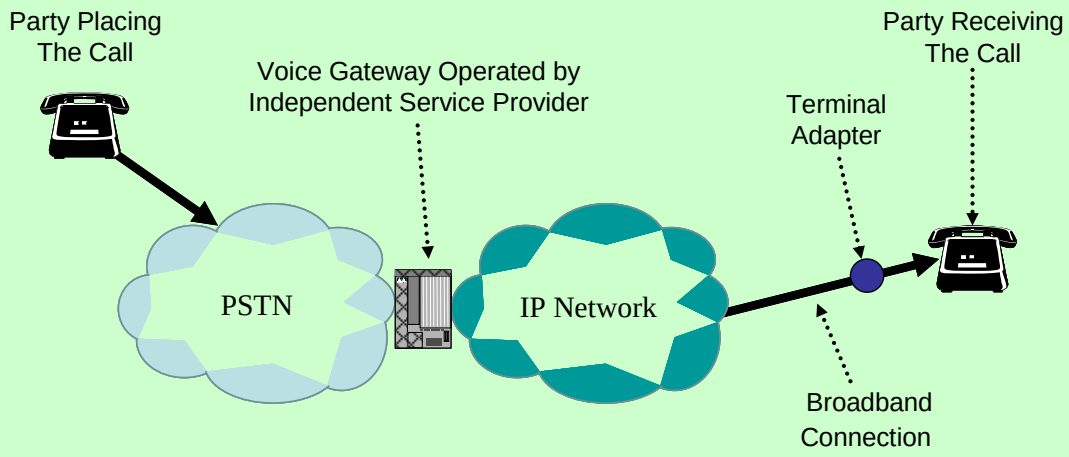
2 Technical constraints in the NGN world

A distinguishing characteristic of Next Generation Networks is the "…[d]ecoupling of service provision from network, and [the] provision of open interfaces."[5] This decoupling of *network* and *service* has profound implications for regulatory policy. Regulation of interconnection has implicitly depended on a close relationship between the network and the service. The standard mechanism for intercarrier compensation[6] depends on wholesale payments from the *service* provider originating a call to the *service* provider terminating the call in order to compensate (primarily) for the use of the *network* used to terminate the call. If the terminating service provider happens coincidentally to be the same as the terminating network provider, then the system could conceivably function much as it does today.

If, however, these are different corporate entities (as is explicitly envisioned in the definition of an NGN), then it is difficult to see how a system based solely on present paradigms could possibly function going forward. For example, the customer might have a contractual relationship with an NGN network operator to obtain broadband Internet network access, and a separate contractual relationship with a third party VoIP service provider (that does not operate a network of its own), but these two providers will not necessarily have a contractual relationship with one another.

Suppose, for example, that a caller on the traditional PSTN places a call to a customer of an independent VoIP service provider (such as Vonage), as shown in Figure 1 below. The party that receives the call typically has two separate contractual relationships (one with a broadband Internet access provider, the other with the VoIP service provider), and compensates each of these providers.[7] In addition, the VoIP service provider[8] could potentially receive a termination fee from the PSTN operator that originates the call.

Figure 1: A call from the PSTN to a customer of an independent VoIP service provider



The first challenge is that the independent VoIP provider has no network costs to speak of. More precisely, it has network costs that are very different from those of a normal fixed operator. Much of the cost of a traditional operator is associated with last mile facilities, especially the local loop. That cost is absent in this case – the broadband access plays a somewhat equivalent role, but the customer is already paying the broadband network provider for that access. The independent VoIP service provider is arranging for gateway services to translate the PSTN call to a VoIP call, and may possibly be providing the user with a terminal adapter to enable connection of a normal telephone to the Internet, but does not incur significant network costs.

In other words, an independent service provider that does not operate a network of its own is perhaps not the appropriate entity to receive compensation for network costs, and typically is under no obligation to transfer

termination fees to the network operator (in legal terms, this is an issue *privity of contract*[9] -- again, the service provider and the network provider do not necessarily have a contract with one another).

The second challenge is that no mechanism exists for the service provider to transfer those payments to the network operator, even if it chose to do so.

The third challenge has to do with the limited visibility that the service provider tends to have into the network operator's business, and vice versa. The TCP/IP protocol suite, upon which both the Internet and the NGN are based, intentionally *layers* protocols and hides information in order to simplify network design and evolution. In consequence, an independent VoIP provider is ill-equipped to measure network usage, and in the normal course of its business need not even be aware of the broadband service provider with which the customer has contracted. Conversely, the network provider can view the traffic that it carries as undifferentiated bits – it need not be aware of the nature of the traffic that it is carrying, unless that traffic requests special handling (that is, differentiated Quality of Service [QoS], a point addressed later in this chapter). Thus, the network operator is ill-equipped to account for the services that run over its network, and the service provider is ill-equipped to account for usage of the underlying networks.

The obvious way for the regulator to resolve these issues would be to determine that integrated entities that provide both the network and the voice telephony service should receive termination fees, while independent service providers should not. To do so would, however, inevitably beg the question: If termination fees are unnecessary in the case of an independent service provider, why should they be required for an integrated service provider (that is performing exactly the same function, and incurring costs that are no greater than those of the independent service provider)?

2.1 What can be measured, and what cannot?

In this discussion, we focus particularly on telephony arrangements since these are central to current interconnection payment arrangements. We assume that voice telephony in the NGN, like all traffic, will be carried over the IP protocol (i.e. it will be VoIP).

Billing in general, and call termination fees in particular, have usually depended on a few variables that in the past were relatively easy to determine:

- The duration of a call;
- The time of day and day of the week at which the call was placed;
- The physical location from which the call was placed, and the physical location of the party receiving the call;
- The identity of the network operator to which the party receiving the call is subscribed; and finally
- Where a mobile customer is placing or receiving a call at a location not served by his or her normal network operator, especially where the customer is roaming in a different country, then additional rules come into play.

The migration to NGN poses challenges in all of these areas, not only for wholesale termination payments between operators, but also for retail payments to the VoIP service provider. Some of these become difficult to determine, at least in certain instances; others no longer have a clear correlation with underlying costs.

2.1.1 Call duration and time of day

VoIP services in the world of the Internet or the NGN will typically be implemented using the *Session Initiation Protocol (SIP*). The SIP server initiating the call will unambiguously be aware of the time at which the voice session was initiated, and will in general also know the time at which the voice session ended. The VoIP service provider (which is not necessarily the same entity that is operating the network) will generally be the party operating the SIP server.

Similar considerations relate to determining the time of day, and the day of the week on which the call was initiated.

2.1.2 Source and destination location

Telephone numbers based on the ITU E.164 standard are associated with a country and, in the case of the fixed network, a permanent location. To the extent that the call is placed using telephone numbers, the source and destination could be viewed as being known to the VoIP service providers, which again are not necessarily the same as the corresponding network providers.

Even for the VoIP providers, the correspondence of telephone number to location does not mean what it once did. Nominally "fixed" VoIP services are often in practice *nomadic* – when a user changes location, the service can move with the user. This uncertainty as regards location poses a problem not only for billing, but also for identifying the location of a caller who seeks access to emergency services.[10]

Where the source or destination location are not associated with a conventional E.164 telephone number, additional challenges might be relevant. Any IP interaction will be associated with a source and desination IP address; however, these IP addresses are linked to the network's topology, not to the geographic location of the user.[11]

2.1.3 Origination or termination

At the retail level, it is common (but not universal) to bill the party that initiates the call, but not the party that receives the call (Calling Party Pays, or CPP). Analogously, it is common but not universal to oblige the network of the originating party to make a wholesale payment to the network of the party receiving the call, i.e. the terminating network (Calling Party's Network Pays, or CPNP). See "Economic background" later in this chapter.

In the PSTN, and under VoIP services that function similarly to the PSTN, the VoIP service provider will in general continue to be able to identify the party originating a call. It is, however, entirely possible that new VoIP services will emerge where it is not so clear which party has originated the call.

As a trivial example, consider the automated dialers that are available with many VoIP services. A program on the customer's PC completes the call by causing first the caller's phone to ring, then that of the recipient.[12] To the network, both parties are terminating a call, and neither is originating. It is often the case today that the party whose phone rings first could be viewed as the originator, but there is no inherent reason why this must invariably be so.

If there were an economic incentive to do so, it would be absolutely trivial for a VoIP service provider to reverse the apparent direction of a call. This exposure is reminiscent of the *refile* schemes that were popular a few years ago, when, for example, calls from Europe to America were far more expensive than calls from America to Europe. Predictably, whenever economic distortions are large enough to make bypass profitable, and where it is feasible and not unlawful (or where prohibitions are unenforceable), bypass will happen.

2.1.4 Resource consumption

As previously noted, the network operator is ill-equipped to account for activities and usage at the application level, e.g. at the VoIP level, except perhaps in cases where the same entity is in both roles; conversely, an independent VoIP provider is ill-equipped to account for network usage, and will not necessarily be able to even identify the networks that have been used to carry the VoIP provider's traffic.

3 Practical challenges

As the previous section hinted, a plethora of problems stand in the way of implementing a robust interconnection framework for IP-based NGNs, and of successfully operating such a framework were it to emerge.

3.1 Transaction costs

Establishing and maintaining an interconnection arrangement with another firm takes work. The technical effort can sometimes be substantial, depending on circumstances. What is often overlooked are the costs of administratively and contractually establishing IP interconnection arrangements.

Suppose there were 10,000 IP-based service providers, either Internet service providers (ISPs) or Next Generation Networks (NGNs). If each of them needed direct interconnection to all of the others, the sheer administrative overhead would be intractable. Each would need 9,999 interconnection agreements. In total, there nearly 50 million agreements.[13] Obviously, this is not what is done today.

In the Internet, it is common for an ISP to maintain not more than fifty *peering* interconnection arrangements, and to use between one and three *transit* relationships to reach all of the rest.[14] This number of interconnection agreements is burdensome but manageable.

Traditional PSTN telephone service providers that have not previously operated IP-based networks will need to create new IP-based interconnection agreements. Firms with existing IP-based interconnection agreements may wish to revise them to explicitly address the ability to carry traffic at a committed level of quality of service (QoS) superior to today's typical best-efforts basis. In both cases, transaction costs represent a barrier to enhancing the global network. In the case of QoS, transaction costs have historically played a large role on causing adoption to stall.

It is possible that transaction costs could be reduced, and the quality of agreements between network operators improved, if some standard agreement template were to be agreed. The GSM Association's *Standard Terms for International Roaming Agreements (STIRA)* illustrates both the benefits and the possible costs of such an approach. On the one hand, the STIRA has probably played a significant role in achieving ubiquitous global availability of roaming services (the ability to place and receive calls from one's mobile phone while traveling in an area that one's own mobile operator does not cover, such as a foreign country); on the other hand, aspects of the STIRA have arguably served to hinder competitive entry and to maintain high prices for global mobile roaming.[15]

3.2 Sensitivity of data

When an IP-based network is under heavy load, it responds by delaying or discarding traffic, rather than by blocking services altogether. For most purposes, this graceful degradation represents a strength of the IP design; however, it can be problematic for delay-sensitive traffic, and especially for real-time voice. IP-based operators who seek to provide their customers with high quality voice services to customers of other networks over an IP-based interconnection may in consequence wish to ensure that their interconnection partners are adhering to mutually agreed levels of Quality of Service (QoS). The desire to verify QoS compliance implies a need for measurement of IP-based interconnection that was not present in the PSTN.

Networks often provide *Service Level Agreements (SLAs)* to their customers. The service provider commits to deliver IP traffic at defined levels of delay and loss, and may offer financial compensation if it fails to meet its objectives. In practice, any financial guarantees typically have little relationship to the actual costs that poor performance might impose on the customer; however, it increases the customer's confidence that the service provider will meet the desired standard.

The service provider generally takes responsibility for any measurements of adherence to the SLA. More often than not, the customer does not have the tools or the knowledge to verify the service provider's measurements. Note, too, that there is no need to measure performance in the customer's network, since the customer is not making a QoS commitment to the service provider.

Where two interconnected networks are not customers of one another, measurement arrangements become much more complicated. Any attempt to measure rigorously whether two networks have adhered to their respective commitments to carry one another's traffic at agreed levels of QoS will need to somehow measure key performance metrics (such as average delay, variability of delay, and packet loss) across *both* networks. This tends to imply a need for instrumentation, either at the end user's premises or within both networks (or perhaps within *all* networks).

IP-based networks experience performance problems from time to time. Any network operator will be uncomfortable with permitting a competitor to place instruments within its network, since this makes those problems visible to the competitor and possible to the prospective customers for whose business both operators are competing. At the same time, neither network operator is likely to have full confidence in the other's measurements – neither will wish to make a payment to the other based solely on measurements that it cannot independently verify.

These problems have been discussed within industry from time to time, but never resolved. One solution that might possibly be workable would be for each network operator to deploy servers to a set of agreed-on cities for the sole purpose of responding to performance probes from the other (for example, various kinds of echo requests). This could perhaps be done on an experimental basis at first, in order to enable network operators to determine the degree to which measurements were stable and repeatable – essentially a confidence-building exercise.

An alternative would be for trusted third parties to operate measurement infrastructure in multiple networks. This would seem in principle to be an attractive solution; however, it is not at all clear who could serve as the trusted third party. For an organization to be accepted by service providers in this role, it would need to have both technical sophistication and impeccable impartiality and integrity.

Today, this entire constellation of issues could be viewed as an unsolved problem. It may be a contributory factor in the glacial deployment of QoS-capable IP-based interconnection among providers (see "Why so slow to emerge?" later in this chapter); on the other hand, in the absence of deployment of QoS-capable interconnection, there has been no impetus to solve it.

3.3 Minimizing fraud

If service providers were to implement usage-based charging, there might be a temptation on the part of one provider or both to synthetically generate traffic in order to be compensated for it.

In today's network, if a service provider were allowed to collect wholesale termination fees well in excess of cost, that service provider might well be tempted to offer favorable retail prices to free-of-charge call centers so as to stimulate calls to its network. Soliciting business that maximizes termination fees collected would in general tend to be viewed as a legitimate and permissible practice.

If the service provider were somehow to place a large number of calls to itself through a competitor's network, with the intent of collecting more wholesale termination revenue, that could reasonably be viewed as a fraudulent practice. Fortunately, this has not been a problem in the traditional network, possibly because it is not easy to do.

In an IP-based environment, it would be quite trivial to synthetically generate large volumes of traffic to or from a competitor's network. Some scenarios would be easy to detect, but others might be difficult. In practice, it might also be difficult to draw a bright line between appropriate and inappropriate practices. It is likely that complex judgment calls would be needed to distinguish between legitimate business practices that alter the traffic balance, versus improper attempts to defraud a competitor.

4 Economic background

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. It also serves to introduce the economics vocabulary that will be used throughout the balance of the paper.

For the reasons already noted, the traditional models of interconnection practiced in most of the world cannot be effective in their present form in an IP-based NGN environment. Adaptation and evolution will be necessary. It is impossible to predict the exact shape of future arrangements; however, understanding the strengths and weaknesses of alternatives already being practiced, both in the Internet and in telephony in North America, is essential to a comprehensive understanding of the likely evolution of interconnection in an NGN world.

4.1 PSTN arrangements

The interconnection of traditional telecommunications networks has been extensively studied in the literature.[16] This section seeks to provide non-specialists with a non-technical but thorough grounding in the theory and the literature.[17]

Section 5.4.1.1 deals with arrangements at the retail level, while section 5.4.1.2 deals with arrangements at the wholesale level. Retail and wholesale arrangements are interrelated, but they are not the same thing. Retail and wholesale arrangements have implications for the speed with which consumers adopt the service, the prices that they pay, and their propensity to use the service once they have it. These implications are covered in section 5.4.1.3.

4.1.1 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.

In the following sections, we discuss the two main models in use today, the *Calling Party Pays (CPP*) system and the *flat rate* (or "*buckets of minutes*") system. Each of these systems has its advantages and its disadvantages, and each has its adherents and detractors. Both systems are in need of a major re-thinking as the world evolves to IP-based NGN arrangements.

4.1.1.1 Calling Party Pays (CPP)

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).

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. These revisionist economists argue 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."[18] In this view, there is an inherent mirror-image relationship between calling and called party. There is no qualitative difference between placing a call and receiving a call, inasmuch as "it takes two to tango." If the call were of insufficient value, the party that receives the call could simply hang up – a principle referred to as *receiver sovereignty*.[19]

These observations have important implications going forward. They imply that prevailing CPP retail arrangements – which place the entire burden of cost on the calling party, and none on the receiving party – are economically inefficient to the extent that they represent a flawed mirror of the value of the call to the customer.

4.1.1.2 Flat rate / buckets of minutes

A few countries – notably, the United States and Canada – use different arrangements. For calls in the fixed telephone network, they historically employed CPP, but these days they primarily implement either flat rate plans or else the nearly equivalent "buckets of minutes" plans for both fixed and mobile telephones.[20]

With a flat rate plan, the subscriber pays a fixed fee per unit time for use of the telephone. Typically, there are no usage-based fees for normal domestic calls, but there tend to be additional per-minute charges for any calls for which the operator pays a significant charge at wholesale. Thus, flat rate plans for fixed telephones generally include per-minute charges for international calls.

Most mobile plans in the United States are in reality *banded* flat rate plans. Each band is flat rate, as long as the consumer places or receives fewer minutes per month than some maximum. Such plans typically include nominal per-minute charges for calls that exceed the agreed-on number, but in analyzing these plans it is important to remember that these per-minute charges are infrequently invoked. They tend to be high to the

point of being punitive. These high charges exist, not with the expectation that many customers will pay them, but rather in order to force consumers to upgrade to a higher band or tier of flat rate service (with a greater number of minutes of use allowed per month) when the time comes.

"Buckets of minutes" mobile plans in the United States generally include per-minute charges for placing[21] international calls, and also for international roaming (placing or receiving calls when traveling outside the United States). Some also include per-minute charges for domestic roaming (placing or receiving calls when traveling in parts of the United States where one's mobile operator does not have its own coverage). Mobile operators set these per-minute prices so as to provide a reasonable return in excess of their wholesale per-minute costs (including the termination fees that they pay for international calls).

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 high utilization 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 should thus be viewed as a gain in consumer welfare.

Some economists have 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 fine grained 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."[22] Experience in the United States strongly bears out the consumer preference for flat rate services.

In the absence of high wholesale per-minute costs, operators will tend to prefer flat rate plans as well, because flat rates provide greater revenue predictability and better reflects their real costs. Flat rate plans may also simplify customer care somewhat, to the extent that they reduce the frequency of billing disputes.

For example, the Digital One Rate service that AT&T Wireless introduced in 1998 provided a "bucket of minutes" 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.[23]

Digital One Rate proved to be immensely popular. The success of Digital One Rate effectively forced AT&T Wireless's 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 with a large "bucket of minutes" for a flat monthly fee.

Today, flat rate plans are becoming increasingly prevalent in the United States for all forms of telephony.[24] 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.[25]

Analogously, when America Online introduced flat rate pricing of USD19.95 per month 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 United States and the United Kingdom have implemented measures to enable consumers to avoid per-minute charges when using dial-up to access an ISP.[26] 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.

At this point, it is necessary to correct one of the common misconceptions about these payment arrangements. Years ago, many analysts assumed that U.S.-style retail arrangements would lead consumers to turn off their mobile phones for fear of having to pay for unwanted calls. Whatever merit that analysis might have had at that time, it is totally irrelevant to today's flat rate or "buckets of minutes" retail environment. The consumer perceives the marginal cost of placing or receiving a call as zero, and therefore has no incentive whatsoever to turn off his or her mobile telephone.

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. High wholesale per-minute costs tend to preclude flat rates. It is for this reason that flat rate plans in Europe usually exclude calls to mobile phones[27] – the high termination fees to these phones represent a wholesale cost that is too great to be ignored.

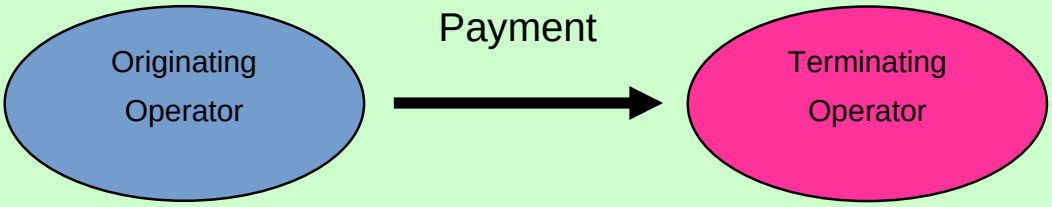
4.1.2 Wholesale level

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

4.1.2.1 Calling Party's Network Pays (CPNP)

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, as shown in Figure 2 below. The call receiver's operator pays nothing.[28]

Figure 2: Calling Party's Network Pays (CPNP) wholesale arrangements



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 by means of a payment at the wholesale level. As networks evolve to NGNs, this underlying assumption is ripe for re-thinking for two primary reasons: (1) in an NGN world, it is increasingly easy to alter or manipulate the direction of call origination, as noted in the section on "Source and destination location"; and (2) economists have come to recognize that both parties, the originator and the receiver, benefit from the call, as explained in the section on "Calling Party Pays (CPP)" at the retail level.

4.1.2.2 "Bill and keep"

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 calls to fixed incumbent operators,[29] but is effectively Bill and Keep for mobile-to-mobile calls and for calls from one non-incumbent fixed provider to another (or to a mobile operator).[30]

France used Bill and Keep for mobile-to-mobile calls until 2004, generally with satisfactory results.[31]

Singapore is particularly noteworthy. Singapore uses a U.S.-like system, with Bill and Keep for calls terminating on the mobile network, but CPNP for calls terminating on the fixed network. The termination fees for calls to the fixed incumbent are limited to 0.005 USD at peak hours, and to 0.003 USD at off-peak hours; termination fees between non-dominant operators are set by mutual agreement.

These wholesale arrangements have led to a retail market with high penetration, low retail prices, and very high utilization. Singapore has a competitive mobile market, with three operators offering 2G and 3G services. Retail arrangements for mobile services reflect per-minute prices (for calls placed or received, a system referred to as *Receiving Party Pays [RPP]*) of from 0.03 USD to 0.13 USD. Mobile penetration is 98%. Revenue per minute (for calls placed or received) for mobile operators is 0.08 USD, while the minutes of mobile use per month per subscriber is 313.[32] These results are among the best in the world, and are arguably superior, in terms of consumer welfare, to those of any European country.

There is some tendency in the literature to use the terms CPP and CPNP interchangeably, but this can lead to confusion. CPNP is a system 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. Bill and Keep wholesale arrangements make flat rate (or buckets) retail plans possible, but they do not preclude other arrangements at the retail level.

As has been previously noted, a very extensive literature exists on wholesale call termination arrangements in general.[33] A number of papers specifically address the relative merits of CPNP wholesale arrangements in comparison with Bill and Keep.[34]

4.1.3 Implications

This section considers that flow from these wholesale and retail arrangements. Of particular interest are:

- The termination monopoly – the tendency for operators (under CPNP, and in the absence of regulatory constraints) to set inefficient termination fees that are well in excess of cost;
- The impact that above-cost termination fees have on retail prices;
- The impact of retail price on consumer adoption and use of these services.

The following sub-sections begin with a brief summary of key findings, and then take up those issues in detail, in that order.

4.1.3.1 Key Findings

As a general rule, countries with CPNP systems tend to have higher retail prices and lower use of mobile service than those with Bill and Keep. Moreover, CPNP tends to subsidize the mobile network at the expense of the fixed network, at some risk of impacting or distorting the evolution of the fixed network (and thus of associated broadband services). For a country with mature ICT markets, Bill and Keep offers distinct advantages over CPNP.

For developing countries, the trade-offs between these two systems are more complex. Mobile operators respond to the higher profitability of their services by deploying rapidly and by offering incentives to new users. In effect, fixed telephone customers subsidize the mobile service. CPNP countries consequently tend to experience faster adoption of mobile services; however, they do not necessarily experience greater adoption of mobile services in the long term than in Bill and Keep countries.

This would seem to suggest that CPNP is appropriate for use in developing countries, but that CPNP should be withdrawn in favor of Bill and Keep once services are fully deployed.

This, however, implies a public policy quandary: once the subsidies associated with CPNP are in place, it is difficult or impossible to remove them. Several countries have transitioned from Bill and Keep to CPNP; no country has transitioned from CPNP to Bill and Keep.

The transition from PSTN to NGN represents a natural "watershed" event. Interconnection arrangements require massive re-thinking at that time in any event. For a country where ICT services are already close to full deployment and adoption, the point of transition to NGN would seem to be the natural point at which to consider whether a transition to Bill and Keep might be appropriate.

The following sections expand on these ideas.

4.1.3.2 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, only 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. Economists 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 mark-ups above the marginal cost of terminating calls."[35]

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.[36] In a CPNP environment, regulation of termination prices appears to represent the only viable alternative to inflated wholesale and retail prices.

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 also led to 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.[37]

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. Fixed-to-mobile termination rates in Europe averaged 0.156 USD as of October 2005,[38] somewhat higher than the global average of roughly 0.128 USD in March 2006.[39]

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 mobile operators in France prior to 2004,[40] 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.[41]

4.1.3.3 Linkage of CPNP to high retail prices

If traffic were 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.

Economists 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."[42]

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, because they do not incur termination charges.[43] 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, even where termination payments are likely to net to zero.[44] 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.

To summarize, *high termination fees tend to lead to high retail prices for originating calls.* (Under CPP retail arrangements, there is generally 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.

An informal white paper of the United States FCC described these relationships succinctly:[45]

> *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.*

The paper goes on to note 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."

4.1.3.4 The linkage between retail price and usage

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. The investment firm Merrill-Lynch provides an annual analysis of the mobile sector in a number of countries, and the U.S. FCC routinely quotes these figures in their annual reports on competition in the U.S. mobile industry.[46] Economists find it convenient to quote these figures, in part because they are readily available. This data is shown in Table 1. For this purpose, we can take the revenue per minute for all mobile operators in a country as being a reasonable proxy for mobile 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 (in USD) include minutes of both origination and termination, whether charged or

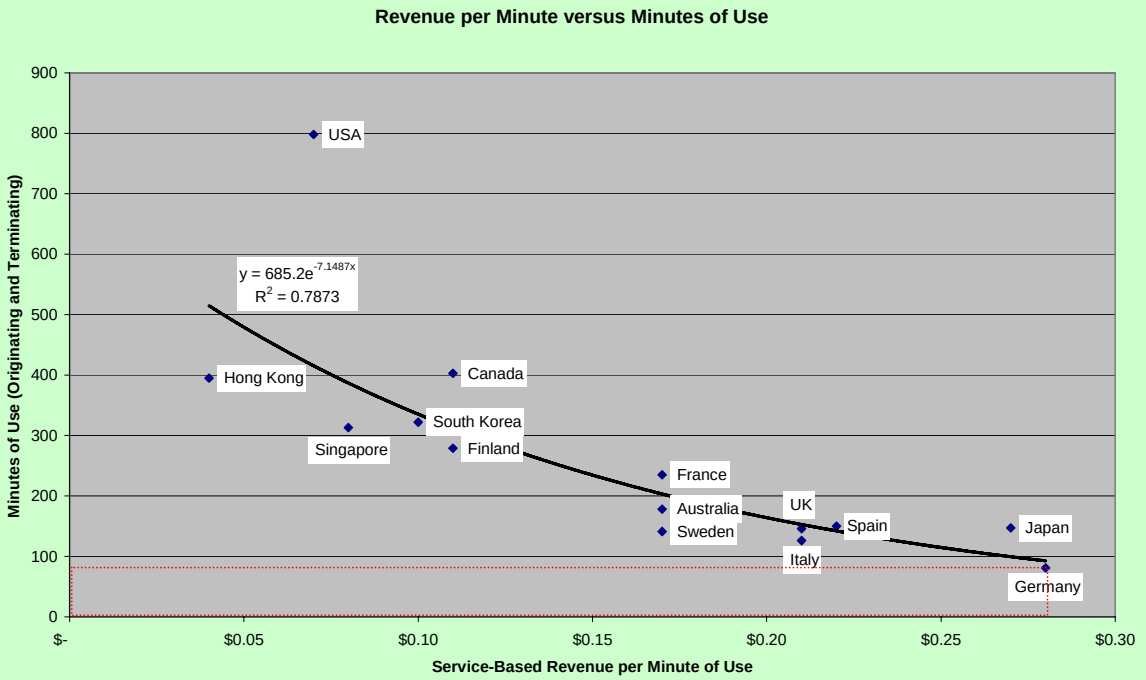
not. Based on this data, Figure 3 below depicts the relationship between service-based revenue per minute, minutes of use, and *Average Revenue per User (ARPU)* for a number of countries.

Table 1: Revenue per minute versus monthly minutes of use.

Country	Revenue per Minute	MOUs	ARPU
USA	00.07 USD	798	55.86 USD
Canada	00.11 USD	403	44.33 USD
Hong Kong	00.04 USD	395	15.80 USD
Singapore	00.08 USD	313	25.04 USD
UK	00.21 USD	146	30.66 USD
Germany	00.28 USD	81	22.68 USD
Italy	00.21 USD	126	26.46 USD
Sweden	00.17 USD	141	23.97 USD
France	00.17 USD	235	39.95 USD
Spain	00.22 USD	150	33.00 USD
Finland	00.11 USD	279	30.69 USD
Japan	00.27 USD	147	39.69 USD
South Korea	00.10 USD	322	32.20 USD
Australia	00.17 USD	178	30.26 USD

Note: Prices are expressed in USD. Revenues and ARPU are solely service-based.
Data Source: FCC, *Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 11th Report (11th CMRS Competition Report*), September 2006, Table 12, based on *Interactive Global Wireless Matrix 4Q05*, Merrill Lynch, Telecom Services Research.

Figure 3: Minutes of use versus revenue per minute.



Data Source: FCC, *Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 11th Report (11th CMRS Competition Report*), September 2006, Table 12, based on *Interactive Global Wireless Matrix 4Q05*, Merrill Lynch, Telecom Services Research.

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 service-based revenues of just USD 0.07 per minute, but with a marginal price that many users perceive (somewhat inexactly) as zero[47] – experiences about ten times as much consumption, expressed in terms of minutes used per month (both originating and terminating), as a country like Germany, where average revenue per minute is about USD 0.28.

Bill and Keep arrangements tend to correlate with low service-based revenues per minute, and CPNP systems with high revenues per minute, but there are exceptions in both directions. Note, too, that low revenue per minute does not necessarily imply low gross revenues – monthly *Average Revenue per User (ARPU)* in the United States and Canada are the highest in this group (USD55.86 for the United States, USD 44.33 for Canada), despite low revenue per minute. Customers are willing to pay substantial monthly subscription fees for cost-effective "buckets of minutes" services.

Conversely, high revenue per minute does not necessarily correspond to high ARPU – in a country like Germany, with high service-based revenues of USD 0.28 per minute, monthly ARPU is low (USD 22.68, compared to USD 26.46 in Italy and USD 39.95 in France) because customers are reluctant to place calls at such high prices. In the graph above, ARPU is the area under the rectangle from the origin (the 0,0 point) to a point associated with a particular country. Germany is associated with a long, low rectangle which encloses very little space – hence, not much ARPU. Paradoxically, the graph above suggests that German operators could probably *increase* ARPU by *lowering* their prices. The resultant increase in usage would overwhelm the reduction in price per minute.

Results in India are particularly interesting. Termination fees for both fixed and mobile are limited to roughly 0.007 USD. This has led to some of the lowest retail rates in the world, roughly 0.02 USD of service-based revenue per minute. These low retail prices have in turn driven high usage of 350 minutes per month.[48] India has achieved this strong usage while simultaneously increasing mobile penetration enormously. India has

apparently found a "sweet spot" where both usage and the rate of penetration are experiencing healthy growth.

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.

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, as discussed in the next sub-section of this chapter. The more common CPP/CPNP arrangements generate effective subsidies to mobile operators. Portions of these subsidies are returned to consumers 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.

4.1.3.5 Linkage between CPNP arrangements and penetration

The low fixed fees and low monthly price associated with CPNP arrangements 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.

Similarly, mobile operators under CPNP are highly motivated to offer pre-paid service with no monthly fee, once again in the hope of receiving termination fees well in excess of marginal cost.

These low monthly fees are usually accompanied by handset subsidies. Mobile operators provide handsets at prices well below cost, or else give them away outright. The low, subsidized initial price is a clear case of "giving away a razor in order to sell the blades".[49] These subsidies may, however, encourage subscribers to replace perfectly good handsets long before they become obsolete.

The combined effect is to encourage consumers in CPNP countries to initially adopt mobile service.[50] The disadvantage, however, is that per-minute usage prices well in excess of marginal cost discourage users from placing calls once they have the service.

Conversely, in Bill and Keep countries, prices track more closely to real costs. Handset subsidies are smaller. Customers have less incentive to initially acquire the service, but much greater incentive to use the service once they have acquired it.

In Europe, there is a growing sense that it is no longer necessary to subsidize the adoption of mobile services.[51] The European Union as a whole claims a 91% penetration of mobile phones, and a number of European countries claim penetration rates in excess of 100%.[52] One needs, however, to be cautious in interpreting these penetration numbers: Penetration rates are computed by dividing the number of subscriptions by the total population. Penetration rates in excess of 100% may reflect consumers who acquire multiple mobile services (multiple SIMs) simultaneously, or possibly consumers who stop using a pre-paid service but have no incentive to affirmatively terminate it. The need for multiple mobile services flows in part from high termination rates, and the desire to obtain cheaper on-net prices to other consumers who subscribe to different mobile networks. Many European countries report penetration well in excess of 100%, but in 2006, 20% of European households did not have a mobile phone at all, and in only three of the then 25 Member States of the European Union did more than 90% of households have a mobile phone.[53] Nonetheless, the bottom line is that mobile phone penetration in Europe is high to the point where there is no public policy basis for subsidizing further mobile penetration from the revenues of the fixed network.

One must be cautious in comparing penetration in CPNP countries versus that in Bill and Keep countries. Customary measurements tend to overstate mobile penetration rates for CPNP countries in comparison to those of Bill and Keep countries. Under CPNP, individual customers often subscribe to more than one service in order to get favorable on-net rates on more than one mobile operator's network. In the absence of monthly subscription fees, there is no disincentive to multiple subscriptions. These duplicate subscriptions provide little benefit to social welfare. In some senses, a more meaningful comparison of penetration would

compare the number of individuals with no mobile service at all. By this measure, the disparity in mobile penetration rates is much less than has been assumed in the literature to date.

Most experts have assumed that CPNP leads to greater adoption of mobile services than Bill and Keep. It may in fact lead to *faster* adoption, but it does not necessarily lead to *greater adoption in the long term*. Singapore, a Bill and Keep country, enjoys mobile penetration of 98%. Mobile phone penetration in the United States (currently at some 71%, and growing by about six points per year)[54] is probably only slightly behind the effective 80% penetration of Europe, once double-counting is taken into account, and is within a small number of years of equalling European levels. Canada, another Bill and Keep country, is following the same pattern but trailing by a few years.[55] Thus, countries that have buckets of minutes arrangements, based on Bill and Keep wholesale arrangements, may tend to experience slower take-up, but can in time achieve comparably high adoption rates to those of CPNP countries.

India's experience is particularly intriguing. In 2003, India introduced a CPNP regime but implemented unusually low fixed and mobile termination rates of just 0.007 USD per minute. The number of subscribers went from some 13 million at the beginning of 2003, to more than 100 million subscribers by the middle of 2006. This dramatic surge in penetration was not at the expense of usage, which nearly doubled over the same period.[56]

The cross subsidies from fixed to mobile that are inherent in high mobile termination rates may also have a tendency to slow the adoption of fixed telephone service.[57] To the extent that these CPNP cross subsidies slow deployment of the fixed network, they might have a negative impact on the deployment of wired broadband access services.[58] Whether this might be a concern will vary greatly depending on national circumstances.

4.1.3.6 CPNP versus Bill and Keep in developing countries?

In summary, what appears to be known is:

- Bill and Keep wholesale arrangements enable low or zero retail per-minute usage fees, but tend to lead to higher initial and fixed per-month retail charges;
- CPNP wholesale arrangements (coupled with high termination fees) 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 retail charges;
- Countries with flat rate or banded "buckets of minutes" retail arrangements tend to experience high and efficient utilization, but may experience slower adoption of mobile services;
- Countries with conventional CPNP/CPP arrangements tend to experience lower utilization, but faster adoption of mobile services; and
- On the whole, CPNP arrangements seem to lead to larger economic distortions than Bill and Keep.

An obvious implication is that CPNP countries in which the market for mobile services is already mature or saturated should consider migrating to Bill and Keep arrangements. As a practical matter, however, it is very difficult to abandon the subsidies implicit in a CPNP environment once they are in place.

The migration to NGN represents a natural transition point at which interconnection arrangements must necessarily change in any case, and may represent a rare point in time at which a migration to Bill and Keep is worth considering. In Europe, it is the migration to NGNs that is prompting national regulators to reconsider the kind of interconnection arrangements that they might want going forward.

Conversely, developing countries seeking to foster the widespread initial adoption of mobile services might possibly prefer CPP/CPNP, even though they tend to embody distortions. CPNP arrangements may have additional advantages for developing countries:

- CPNP for international calls will tend to generate net monetary transfers in the direction of the less-developed country due to asymmetries in the number of calls placed, probably as a result in differences in the level of disposable income; and

- CPNP for domestic and international calls, implemented so as to reflect the higher costs that rural operators incur, may provide a means of supporting universal access or universal service.

Seen in this light, a key question remains: Will CPNP arrangements remain viable in light of the transition to NGN?

4.2 Internet arrangements

As with the analysis of the PSTN, it is helpful to distinguish between retail and wholesale arrangements. Section 5.4.2.1 discusses the retail level; section 5.4.2.2 discusses the wholesale level; and section 5.4.2.3 considers the implications for NGNs going forward.

The chapter focuses on experience with IP-based interconnection in the Internet rather than in NGNs, for the simple reason that there is no operational experience with interconnection of NGNs. At the same time, NGN operators will be subject to the same economic forces as their Internet counterparts, and can reasonably be expected to respond similarly.

4.2.1 Retail level

At a retail level, large enterprises purchase access on a somewhat different basis than do individual consumers (including small offices and home offices). In addition, pricing characteristics for individual consumers tend to be distinct for dial-up Internet access versus broadband.

4.2.1.1 Enterprise pricing

Internet access is most often sold to large enterprises on a flat rate basis, with a fixed monthly fee and no usage-based charges. The maximum traffic that the ISP will carry is nominally limited only by the size of the pipe to which the enterprise customer has subscribed, but may additionally be limited by the amount of capacity available in the ISP's overall network.

Enterprises have occasionally acquired Internet access on a usage basis, most often based on some approximation of their traffic during the busiest hour of the day. For example, some U.S. ISPs historically offered access based on the 95th percentile of traffic measurements taken every 15 minutes. For an ISP, a measure of near-peak traffic probably equates reasonably well to cost causation.

Alternatively, charges (especially for web hosting traffic) might reflect the total number of bytes of data transferred in the billing period.

If the customer's traffic increases in a pure flat rate arrangement, the ISP hopes to benefit in the longer term when the customer is obliged to procure more capacity.

4.2.1.2 Consumer dial-up Internet access

Dial-up Internet access should not be viewed as an NGN service; however, the operation of that side of the present market helps shed light on consumer preferences.

Historically, many dial-up ISPs charged consumers based on the number of hours for which the consumer was connected to the service. In many countries, the consumer also pays by the minute for the access over the telephone.

In the United States (and in a number of other countries, including the UK and Italy), telephone access to the ISP is typically available without per-minute charges. In addition, flat rate has been the norm for dial-up Internet in the U.S. since America Online first introduced its "all you can eat" plan in 1995. Countries where neither telephone access nor dial-up access to the ISP incur per-minute charges tend to experience much higher Internet usage, and customers clearly favor these flat rate arrangements.

4.2.1.3 Consumer broadband Internet access

Broadband Internet access is generally offered on a flat rate basis. The maximum amount of data that the user can send or receive is usually limited either by the physical capacity of the pipe to the consumer's home, or else by administrative controls.

The capacity upstream (from the consumer to the network) is often less than the capacity downstream (from the network to the consumer).This difference may reflect technical limitations, or the desire of the broadband ISP to charge a premium to customers who use the service to perform upstream-intensive operations such as web hosting, or both.

As with enterprise pricing, the consumer generally does pay a premium for slight increases in usage; however, the ISP hopes to benefit if the consumer's usage increases to the point where he or she needs a larger pipe.

4.2.2 Wholesale level

In this section, we consider the nature of interconnection arrangements among ISPs; the degree to which they are motivated to interconnect; and the similarities and differences, at an economic level, between PSTN interconnection and Internet interconnection.

4.2.2.1 Peering and Transit

The two most prevalent forms of interconnection at an Internet Protocol level between *Internet Service Providers (ISPs*) are *peering* and *transit*. For a definition of these terms, we turn to a publication of the Network Reliability and Interoperability Council (NRIC)[59], 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.

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. Some years ago, about 10% of the peering relationships of GTE Internetworking (at the time one of the five largest Internet backbones in the world) 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 the Internet backbone service provider more, than the traffic sent to the other network.

4.2.2.2 To peer, or not to peer?

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

How many ISPs are there? It is difficult to say. A few years ago, *Boardwatch Magazine* listed more than 7,000 ISPs in the United States alone. There are no reliable statistics today, but it is possible to estimate an upper limit. Any network that participates in the Internet's global routing system requires an Autonomous System Numbers (ASNs).[60] A tiny ISP might not need an ASN, but any ISP of any size will have multiple upstream service providers and will therefore require an ASN.[61] There are not more than 40,000 ASNs

currently assigned,[62] so there are not more than 40,000 ISPs of any size. In reality, there are probably far fewer.

Analogously, there is no unambiguous answer to the upper limit on the number of ISPs with which a peering relationship is technically feasible. It is reasonable to assume that technical constraints limit to a few hundred peering relationships at the maximum.

Technology poses constraints, but they are by no means the only constraints. The number of peering relationships is also limited by:

- The infrastructure 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 less than the number of independent IP-based networks in the world by at least a factor of one hundred. These hierarchical arrangements contribute to efficiency. "[E]conomic 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."[63]

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.[64]

4.2.2.3 Incentives to interconnect

A body of economic theory that first appeared twenty years ago[65] analyzed incentives of firms to conform to standards when participating in markets characterized by strong network externalities. 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 interchangeability 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 has come up in the context of a number of major mergers, and has been analyzed at length by a number of economists.[66] 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. "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."[67]

The thresholds at which these potential anticompetitive effects might dominate have not been rigorously determined. 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 commercial Internet is operating well below the thresholds where this form of anticompetitive effects would predominate.

4.2.2.4 Linkages between PSTN interconnection theory and Internet interconnection theory

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.[68] A key linkage between the economics of Internet backbone peering and the economics of PSTN interconnection has been identified:

> 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 operators' optimal usage price reflects their perceived marginal cost. … [T]he missing payment affects the backbones' perceived costs, and it reallocates costs between origination and reception. [69]

In other words, in PSTN interconnection under a CPNP regime, the "missing price" is the price that the call recipient would pay at retail. In an RPP system, there is no missing price, because the customer pays for both placing and receiving calls. Likewise, in a flat rate system, there is no issue of a missing price, because all of the usage charges (both for placing and receiving calls) are subsumed in the flat rate (e.g. monthly charge).

4.2.3 Implications

The interplay between peering and transit in IP-based networks yields a rich tapestry of interrelationships between and among Internet Service Providers (ISPs).

It is important to remember that peering and transit are not the same service. With peering, two ISPs agree to exchange traffic for their respective customers, but are not obliged to carry traffic for one another to third parties. With transit, one ISP agrees to carry traffic, usually to all destinations and usually for a fee, for another.

The complex dynamics of Internet play out quite differently than those of the traditional world of switched telephony. Notably, the implications of a refusal to peer in the Internet can be quite different from those of a refusal to interconnect in the PSTN. These differences have potentially important implications for future regulation, or lack of regulation, of interconnection among IP-based networks.

Like the rest of the chapter, this section draws experience from the Internet in order to reason about the future world of IP-based NGNs. There is no base of experience with IP-based interconnection of NGNs; however, it is reasonable to assume that the economics of IP-based interconnection that is visible in the Internet today will influence the NGN decision makers of tomorrow.

4.2.3.1 Relative desirability of peering versus transit

It is sometimes erroneously assumed that peering is inherently preferable to transit, because peering is often (but not always) "free". The reality is much more complex.

A handful of the largest and best-connected backbone ISPs have no need of a transit provider.[70] They are richly connected to one another, and can reach all destinations over their peering connections. To be in this position, an ISP needs to have dozens of peering relationships. If there are any gaps (that matter) in the ISP's peering coverage, the ISP will generally need to purchase transit from some other ISP that has coverage to those destinations. As a result, nearly all ISPs find it necessary to purchase some transit.

Suppose that two ISPs both have transit providers, and can therefore reach all destinations. When will they choose to peer? The answer is that they will view peering as an economic optimization – each will be motivated to peer if the cost of peering (considering capital, operating and transaction costs) is less than the cost reduction it obtains by avoiding sending this peering traffic to its transit provider. If both ISPs perceive a net reduction in cost, they will be motivated to implement the peering connection.

Nearly all ISPs find it necessary to purchase transit, and most find it cost-effective to supplement the transit with at least some peering.

If an ISP were to expand its peering relationships to the point where it no longer needed a transit relationship, would it save money by doing so? Not necessarily. In the few concrete cases where this has been accomplished, it has sometimes caused costs to *increase*, not to decrease. A transit customer benefits in many ways from the transit relationship. The transit customer has no obligation to carry traffic for the transit provider; moreover, the transit customer can expect the provider to deliver traffic over whichever of its multiple connections the customer prefers. If the transit customer "upgrades" to become a peer, it loses these benefits. The former transit customer will generally find it necessary to invest in circuits and equipment to

carry more traffic. In concrete instances, the increased costs of infrastructure have sometimes overwhelmed the savings from eliminating payments for transit service.

4.2.3.2 What happens when peering is denied?

Suppose an ISP (call it ISP A) refuses to peer with another (ISP B). What are the implications for both in terms of the likely arrangements that each will pursue, and the costs associated with each?

In most cases, the refusal to peer does *not* result in a connectivity breakdown. In declining to peer, ISP A typically reasons that ISP B's customers need to reach those of ISP A, and that ISP B will therefore find it necessary to make other arrangements.

The refusal to peer does not necessarily lead to more revenue for ISP A, nor does it necessarily reduce direct costs for ISP A. ISP B is by no means forced to become a customer of ISP A; it could instead use its arrangements as a transit customer *of any ISP* to reach ISP A (assuming that ISP A is generally reachable), and that is the more likely outcome.

Should ISP B choose to use its transit arrangements with some other ISP to route its traffic to ISP A rather than being able to do so over a peering connection to ISP A, doing so does not change the volume of traffic entering and leaving either network. To a first order, the volume of traffic is a function of customer demand, not of the way in which the networks are interconnected.[71] It does, however, change the location at which traffic enters and leaves each network, and to that extent potentially changes costs for both ISPs. The change in costs could be small, or they could be quite substantial.

5 Should interconnection prices be set in an NGN, and at what rates if so?

An important difference between PSTN interconnection and Internet interconnection is that the former had generally been subject to regulation, while the latter has generally not. Both systems seem to have led to generally satisfactory outcomes. What can we say about the differences between the two systems? Going forward, which system should be preferred?

In this section, we consider first the rationale for a regulated solution with interconnection obligations and established rates; then the arguments against interconnection obligations; and end with a quick summary.

The discussion that follows is largely in terms of termination fees based on minutes of use, because that is the way the debate has traditionally been framed, and is also the basis for most of the existing economic analysis; however, the analysis would not be very different if some other measure of traffic exchange were used.

5.1 Rationale *for* interconnection obligations

Interconnection obligations exist in the PSTN primarily to deal with market power imperfections that might otherwise motivate powerful, entrenched incumbents to refuse to interconnect on reasonable terms, thus inhibiting the entry of competitors. Interconnection has historically been a key locus for the exploitation of market power.

They exist secondarily as a means of limiting the degree to which CPNP wholesale payment arrangements might be set at levels well in excess of cost, thus harming consumer welfare.

5.1.1 Market power of incumbents

The most important reason for mandating interconnection in the PSTN is that, in the absence of regulation, large established incumbents would not be inclined to grant interconnection on reasonable terms. The discussion of "Incentives to interconnect" earlier in this chapter explains why: perfect interconnection would limit the ability of the incumbent, which has market power by virtue of the network effects associated with its large customer base, to exploit its market power. To the extent that a powerful incumbent were to provide perfect interconnection, it would undermine its competitive advantage.

Perhaps the first and best known example of this phenomenon was the refusal of the Bell System in the United States to interconnect with its long distance network with rivals. In the absence of interconnection to

the best and most richly connected long distance network, its competitors were not viable. Disputes with competitive entrants led to shareholder lawsuits and threats of a government antitrust investigation. The issue was ultimately resolved by limited commitments on AT&T's part to interconnect (the "Kingsbury Commitments").[72]

The dispute between Telecom New Zealand (TCNZ) and Clear represents a well known example in modern times. Reformist New Zealand attempted to avoid regulating interconnection in advance (*ex ante*), hoping that lightweight remedies based on after the fact (*ex post*) competition law would be sufficient to resolve problems. Unfortunately, it took six years to resolve the dispute on this basis, during which time New Zealand consumers were denied the potential benefits of competition. General dissatisfaction with this delay led New Zealand to implement a regulatory framework capable of dealing with interconnection issues on an *ex ante* basis.[73]

The migration to IP-based NGNs has the potential to open portions of the network up to competition, thus ameliorating these concerns. Unfortunately, that is not the end of the discussion. First, there is a strong likelihood that some forms of market power will remain, especially in regard to last-mile access (where it is unusual to find more than two significant facilities-based operators in any geographic area).[74] Second there is the risk that NGN might create opportunities for the exploitation of new bottlenecks. We return to these points shortly, in the "Changes in market power effects" section.

5.1.2 Termination monopoly

As previously noted, a CPNP system encourages operators to set termination prices at very high levels – levels that are in excess in many cases even of monopoly prices.

This is a problem that has been *created by regulation* – without a regulatory obligation to pay termination fees, the operators would tend either to mutually agree to much lower levels, or more likely to avoid them altogether. But it is also a problem that can be ameliorated by regulation. In Europe, concerted regulatory action has been effective in systematically moving both mobile and fixed termination fees to levels that are at least closer to usage-based marginal cost.[75]

5.1.3 Rate-setting

It is generally accepted that call termination fees should be set at levels that correspond as closely as possible to the forward-looking long run incremental costs of the operator that terminates the call.

This principle is simple enough to express, but actual computation poses a great many practical problems. We will provide only a brief sketch here, since these issues are well beyond the scope of this chapter and are in any case familiar to many readers:

- The operators understand their costs far better than does the regulator, and may be motivated to slant their reporting to the regulator.
- In computing a forward-looking cost, it is unclear whether the regulator should use a perfect, idealized network (which is perhaps unachievable in practice), the current network, or some blend of the two. To the extent that the rate reflects the current network, it may tend to reward bad design decisions.
- If termination fees are set too high, it might be difficult for competitors to achieve market entry. (The incumbent is not constrained by these rates in setting prices for on-net calls among its own subscribers.)

5.2 Rationale for *not* mandating interconnection

Bilateral negotiations for Internet interconnection have frequently led to satisfactory arrangements for all parties concerned.[76] In the United States, where mobile operators and non-dominant fixed operators are free to negotiate any rate they choose as long as it is symmetric, negotiations have led to very satisfactory arrangements for all parties concerned (and in most cases to an agreement to set termination fees to zero);[77] however, this has not always been the case in developing countries, where the historic incumbent often benefits from a huge size disparity in comparison with its smaller competitors.

These outcomes are best understood in terms of (1) the Coase Theorem, and (2) issues of market power.

IP-based NGNs differ somewhat from traditional networks in their technology, and in the associated value chain, as noted at the beginning of this chapter. A number of these differences tend either to undermine current interconnection arrangements, or to make them less relevant than they once were.

All things considered, it appears that:

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

5.2.1 Coase Theorem

The Nobel-prize-winning economist Ronald H. Coase has argued, most notably in a famous 1959 paper[78], 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 key intuition here is that the operators themselves are in a better position to understand their costs, and the respective value to one another of interconnection, than are the regulators. The generally positive experience with Internet peering and with mobile operators in North America appears to bear this out.

In the United States, mobile operators (and non-dominant fixed operators) have generally been under no regulatory obligation to interconnect with one another; nonetheless, privately negotiated Coasian wholesale interconnection arrangements have worked well.[79] The sector has tended in practice to operate on a Bill and Keep basis.[80]

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.

There are also important differences. In the United States, the regulator does not set the termination fee (except for dominant wired incumbents), but fees must be symmetric.[81] Also, the FCC found it necessary to intervene to prevent local competitive operators (in American parlance, these are *Competitive Local Exchange Carriers*,or *CLECs*) from setting unreasonably high charges for completing calls from long distance providers. The problem was in part a consequence of the fact that the long distance operator was under a regulatory obligation to complete the call, but had no right to compensation from the local operator, while the CLEC had no significant constraints on the price that it could charge for completing the call. Additional regulation was needed to correct a regulatory asymmetry: the FCC imposed a rule preventing CLECs from charging a rate in excess of the (generally regulated) rate charged by the fixed incumbent (the *Incumbent Local Exchange Carrier*, or *ILEC*) in the same area.[82]

5.2.2 Changes in market power effects

If one party to a bilateral negotiation had significant market power, and the other lacked countervailing power, then one might expect that a 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[83] 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, this form of 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 last mile market power.

5.2.2.1 Internet interconnection versus PSTN interconnection

Market power over interconnection manifests significantly differently for Internet service providers than for traditional PSTN-based operators. Institutional arrangements in the United States help shed light on the differences. In the United States, as in many countries, wired incumbents are subject to a range of regulatory obligations. Limited regulation of interconnection has also been necessary for mobile operators and CLECs in the United States, but not for ISPs. Why not?

Part of the answer is that these limited rules were necessary to correct for unforeseen consequences flowing from other regulations, including the interconnection obligation on long distance carriers. The other part is that it is the fundamental structure of interconnection in the Internet, with a mix of peering and transit, that largely obviates the need for regulation. This second point is relevant globally.

If a traditional incumbent PSTN carrier denies interconnection, its actions generally prevent a competitor from reaching the incumbent's customers. In the Internet, however, any provider can purchase transit in order to reach all Internet destinations. Therefore, when an ISP (call it ISP A) refuses to peer with a second ISP (ISP B), and in the absence of other bottlenecks, the second ISP can in general still access the customers of ISP A as long as one or both providers have a transit provider (see the section "What happens when peering is denied?" earlier in this chapter).

As long as the ISP that refuses peering (ISP A) offers peering to at least one or two ISPs who can access end-user customers at costs not very different from those of ISP A, the peer ISPs should be able to offer transit arrangements that include access to ISP A's customers at prices not very different from those of ISP A. From ISP B's perspective, this should be sufficient to enable it to compete effectively with ISP A, other things being equal.

These same considerations also limit the value to ISP A of denying peering. ISP A's denial of peering does not force ISP B to become a transit customer of ISP A; *ISP B can be or can become a transit customer of any ISP*, as long as ISP A is generally reachable at reasonable cost. Nor does ISP A's denial of peering necessarily reduce ISP A's network infrastructure expenses, because ISP B's traffic demands (and the demands of ISP A's customers to reach customers of ISP B) have not gone away – they simply enter ISP A's network at some other point, typically over some other peering interface.

This means that, as long as these markets remain reasonably competitive, ISP A should be making its decision on a rational business basis. It should be observing potential infrastructure savings from implementing peering with ISP B, considering its infrastructure costs and administrative costs to implement and maintain the peering, and agreeing to peer if the benefits exceed the costs, all things considered.

What if underlying markets are not competitive? What special challenges do regulators face in the event that a large ISP refuses to peer at reasonable cost with any of its domestic competitors? Section 5.5.3, "Market power of the incumbent", deals with this concern.

5.2.2.2 Independent service providers

The presence of independent third party providers of services such as Voice over IP (VoIP) may have some tendency to moderate market power, or to reduce the ability of incumbents with last mile market power to leverage that market power into upstream services markets.

Incumbents may attempt to restrict the activities in these party providers – by impacting, for example, the quality of service that their services receive. From a public policy perspective, it would be unfortunate were they to succeed. Third party services clearly enhance competition and consumer choice.

At present, it is not clear whether a degradation strategy would be effective. For DSL-based broadband services, all indications are that normal best-efforts services are good enough to enable competitive third party services; thus, incumbents would have only limited ability to discourage use of the service unless they were to intentionally cripple its performance, which could perhaps be viewed as an anticompetitive act. For cable-television-based broadband services, however, degradation might be a more significant concern. Many cable users share a common transmission link – significant delays are possible or likely, at least on occasion.

5.2.2.3 New loci of market power

As networks evolve to NGNs, there is a risk that new forms of market power might emerge. Historically, market power was often associated with physical bottlenecks, especially in connection with last mile access. In the future world of the NGN, it is entirely possible that new choke points will emerge in higher layers of the network, associated with logical control functions.

A study conducted for the European Commission[84] raised concerns over a variety of potential bottlenecks going forward; at the same time, it cautioned against an overly hasty attempt to apply regulation before problems have emerged. Their concerns included a range of network capabilities, call set-up capabilities, application program interfaces (APIs), various "walled garden" restrictions on access to content, and user identity and location information. For now, these should be viewed as potential threats, not as immediate problems requiring action today.

5.2.3 Changes in the value chain

Many IP-based services are available from traditional incumbents and also from independent third parties. Voice services, in particular, are available from third party service providers including Vonage, Skype, and SIPgate. IPTV providers are emerging. These independent services have already gained a following, and are likely to remain popular. Their presence in the marketplace provides clear procompetitive benefits.

Regulators will tend to view these changes as disruptive, and may be tempted to try to prevent their markets from developing in these directions. These independent services provide benefits, and should be embraced, not thwarted. Rather than attempting to suppress independent service providers, a regulator would be well advised to try to anticipate these changes, to support them, and to apply regulation only to the limited extent necessary to address likely competitive harms.

These independent service providers do not have networks of their own, and are ill-equipped to measure network usage. Conversely, network operators are ill-equipped to measure minutes of use from these third party providers.

Minutes of use are only weakly correlated with cost in an IP-based NGN. In addition, they will be difficult or impossible to measure consistently. In the not-too-distant future, it most likely will be impractical to continue to base termination fees on the duration of the call in minutes.

Does this mean that the entire system of termination fees will disappear? It very well might. But an alternative that is not absolutely precluded is that a system of access charges might emerge that reflects measurable and quantifiable aspects of the network provider's service that correlate better with the network operator's cost drivers – for example, some measure of average traffic volume and of traffic variability, possibly with a premium added on for traffic carried at better-than-best efforts quality of service.

5.2.4 Changes in retail price structure

The current CPNP system of call termination fees is largely a response to Calling Party Pays (CPP) retail pricing arrangements. Since the receiving party does not pay for the call, the CPNP termination fee compensates the terminating network operator for the use of its network.

With the disintegration of the traditional value chain, as discussed in the previous section, traditional providers will be subject to competition from independent providers of VoIP services (and other services as well, such as IPTV video). These new providers will have a very different cost structure from that of traditional service providers. All indications are that their usage-based costs are very low, and are roughly linear in the number of customers.[85] This will effectively set a ceiling on the prices that traditional providers

can charge for voice services – if they price substantially in excess of the costs of the independent providers, they will simply lose market share to them.

The migration to IP-based NGNs should alter the costs of the traditional operators such that their usage-based marginal costs are similar to those of their independent third party competitors. Large incumbents benefit from economies of scale and scope, but it is not clear that these benefits are compelling in regard to services such as VoIP. The most likely outcome would seem to be that the pricing structures offered by all market players will be generally similar, and will lead to retail pricing structures for voice services that are also similar. To the extent that traditional providers are viewed as offering superior service, they may be able to command a modest price premium commensurate with the perceived value of the difference.

Independent service providers incur minimal network expense.[86] Typically, their customers pay for their own broadband Internet access on a flat rate basis. The independent providers have one noteworthy per-call per-minute expense: the termination fees that they pay to traditional telephony service providers for terminating calls originated on the VoIP service. The pricing plans of providers such as Skype and Vonage are exactly what one would expect under those circumstances: They are flat rate and inexpensive for calls where the termination fees are low enough to ignore, but reflect per-minute fees somewhat in excess of the termination fee where the termination fee is larger (greater than roughly 0.02 USD per minute).

Given the competitive nature of the marketplace, it seems likely that traditional telecommunications firms will necessarily evolve to roughly the same flat rate pricing structure as their third party competitors as they migrate to IP-based NGNs, but perhaps with a slight premium for quality. The traditional providers will find it difficult to avoid tracking the contours of the cost structure of their competitors, particularly when that cost structure is roughly the same as the cost structure to which they themselves are subject.

Should this prove to be the case, then a key rationale for charging termination fees disappears. In a flat rate retail system, there is no pricing difference betweens calls placed and those received, nor between traffic sent and traffic received. Since there is no "missing charge" for calls that are received but not charged for, there is no need to correct for the missing charge with a termination fee.

5.2.5 Practical difficulties in allocating costs

As previously noted, it is difficult to determine the correct level for termination fees. The analysis is challenging and time-consuming, it is difficult to know how to properly interpret a forward-looking cost for a purely hypothetical network, and there are adverse consequences for errors.

One view from the United States is that true usage-based forward-looking marginal costs are closer to zero than they are to the levels at which call termination fees have traditionally been set. In this school of thought, setting fees to zero reduces errors at the same time that it simplifies the billing process between the operators.[87]

5.3 Special challenges for developing countries

All countries will tend to face significant challenges in dealing with entrenched incumbents, and in providing universal access or universal service to their citizens. Developing countries are likely to experience particular difficulties. The following sub-sections deal with market power and with universal access with a particular emphasis on developing countries.

5.3.1 Market power of the incumbent

An earlier study (prepared under ITU sponsorship, but not necessarily reflecting the views of the ITU) evaluated interconnection among IP-based NGNs in the context of developed countries, with an eye toward Europe.[88] That study concluded that IP-based NGNs were likely to interconnect even in the absence of an explicit interconnection obligation. The most likely form of interconnection would be peering, with no money changing hands (i.e. Bill and Keep). If that were indeed the case, then multiple service providers would be able to offer connectivity to one another's customers at reasonable prices through some combination of peering and transit. Under those assumptions, regulators would be well advised to pay primary attention to ensuring that markets for underlying components that competitive entrants would need in order to offer fully competitive services remain competitive – especially broadband and leased lines.

This is consistent with the notion that, wherever possible, the regulator should prefer the working of the competitive market. Ideally, the regulator should intervene only to the extent necessary to correct for market failures and deficiencies. Deferring to the competitive market eases the task of the regulator, and generally leads to more efficient outcomes for all.

The assumption that large IP-based NGNs operated by former national incumbents will voluntarily interconnect with other IP-based operators seems natural based on experience with the Internet, but it is by no means guaranteed. Historically, large Internet backbone ISPs have agreed to peer with some of their largest competitors, but that practice may in part reflect the decision of antitrust authorities to block worrisome aspects of mergers[89] that might have enabled a backbone to amass enough market power to find it profitable to have less-than-perfect interconnection with some competitors. It may also reflect the historical reality that the largest Internet backbones mostly derived from the U.S. long distance industry, a competitive segment, during a period when the local incumbents (the market segment that arguably possessed market power) were effectively precluded from offering Internet access services – it is by no means assured that the same outcome would emerge today.

There is also a notable counter-example – the Australian incumbent declined to peer with any of its domestic competitors. The Australian Competition and Consumer Commission (ACCC) ultimately forced Telstra to peer with domestic competitors in 1998.

Historically, regulators have been reluctant to regulate IP-based interconnection, and appropriately so. Most regulators lack authority to regulate IP-based interconnection. Going forward, as networks evolve to IP-based NGNs, at some point in time the only meaningful interconnection to the incumbent's network will be an IP-based interconnection. If the incumbent refuses to interconnect on reasonable terms, it seems clear that the regulator will need to have the tools and authority to enable it to respond effectively.

Circumstances may vary greatly from one country to the next, but it is quite possible that developing countries will face greater challenges than industrialized countries. The incumbent operator is more likely to be solidly entrenched. Moreover, the size disparity between the incumbent and its competitors is sometimes enormous, putting the latter at a great disadvantage.

Compounding these concerns, the legal and regulatory institutions associated with network interconnection in developing countries will in many cases be less mature than those in developed countries – not only for IP-based interconnection, but in many cases for traditional interconnection as well.

It is also worth noting that the prediction of the previous ITU-sponsored study was that *interconnection* would in many cases be implemented voluntarily. *Access* – to unbundled local loops, for example – is quite another matter. An established incumbent is unlikely to grant access to its facilities at favorable prices in the absence of regulatory pressure.

5.3.2 Universal access / universal service

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. *Universal service* refers to services delivered to the home; *universal access* refers instead to the closely related problem of delivering these services to public, shared facilities such as schools, libraries, post offices, or telecenters. For many developing nations, universal access represents a more appropriate and more readily achievable goal.

The transformation associated with migration to NGN places great stress on universal access and universal service institutions. The scope of universal service, the funding, and the very viability face enormous challenges going forward.[90]

Section 5.5.3.2.1 explains the rationale for universal service, in terms of network externalities, economic distortions, and consumer welfare. Section 5.5.3.2.2 explains the use of implicit interconnection-based subsidies within a developing country, while Section 5.5.3.2.3 explores subsidization mechanisms employed internationally. Section 5.5.3.2.4 expands on the implications for policy.

5.3.2.1 Network externalities, economic distortions, and consumer welfare

Markets characterized by strong *network externalities* (that is, markets where the benefits that a customer derives are strongly influenced by the number of other people using the same service, as is the case with a network) have a tendency to reach stable equilibrium at levels of service adoption that are much lower than those that are societally optimal.[91] 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 somehow extract enough profit from other customers to cover the cost. Still others provide an explicit 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.

5.3.2.2 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.

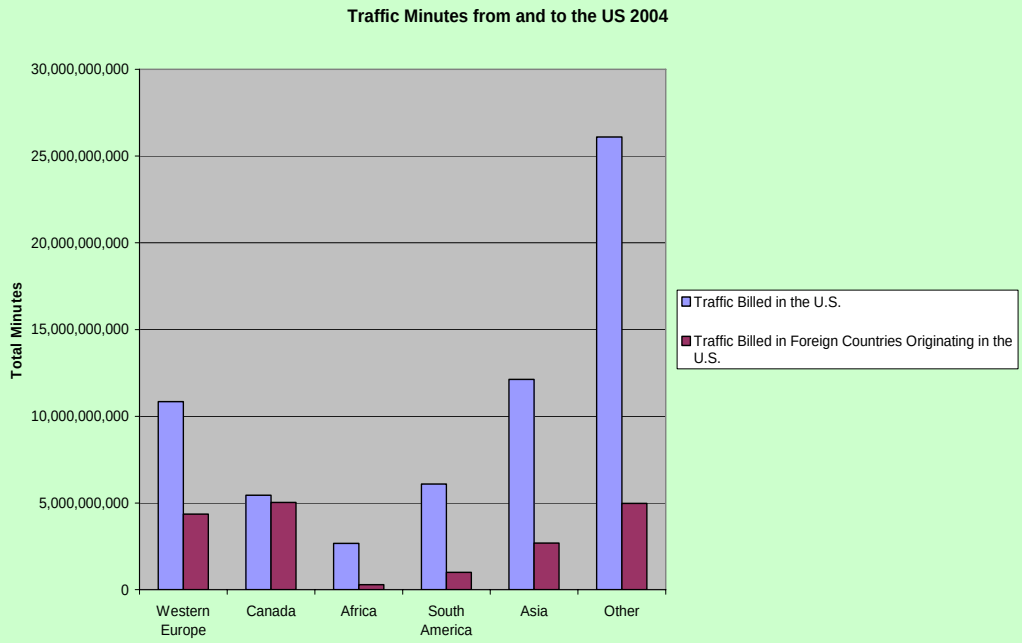
A number of concerns must be raised in connection with these subsidies. They represent an economic distortion. They are subtle, and not likely to be understood by the public – there can thus be a notable lack of transparency. And they can easily turn into "slush funds". It is for this reason that there has been a move away from the use of such subsidies towards other mechanisms. There are explored in the GSR Paper on Universal Access.

5.3.2.3 Traffic imbalance – the "Robin Hood" effect

As previously noted, PSTN interconnection fees for switched telephone calls in most countries are 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.

Consider, for example, the number of messages and of minutes of use between the United States and (1) Western Europe, (2) Canada, (3) Africa, (4) South America, (5) Asia, and (6) everywhere else. Data from the U.S. FCC show a roughly balanced calling pattern with Canada, but a huge preponderance of calls placed from the United States to the other foreign countries.

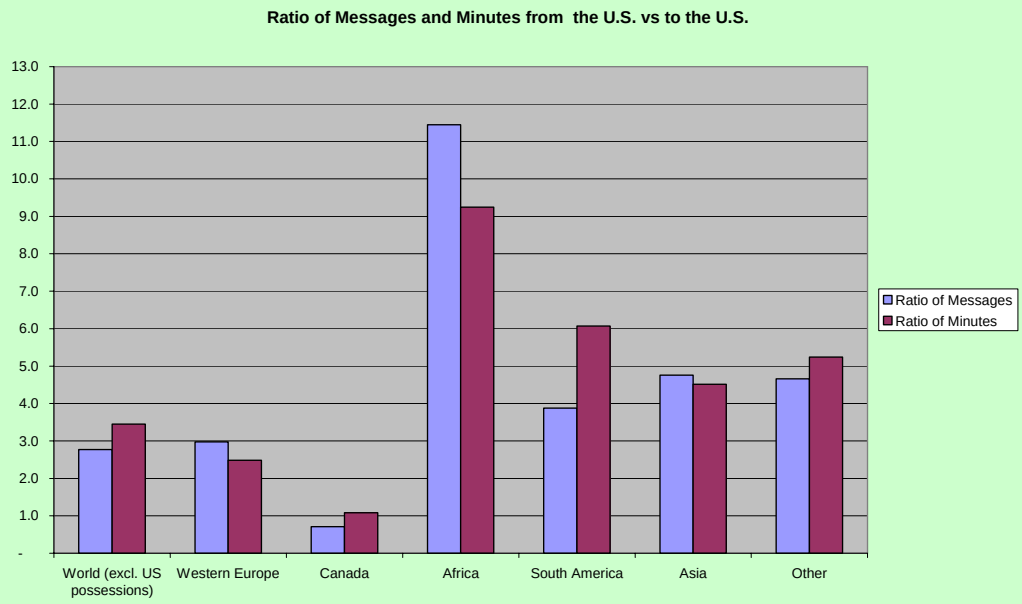
Figure 4: Traffic minutes between the U.S. and selected regions of the world.



Note: Author's figure prepared from 2004 FCC data

The next figure shows the ratio between traffic originated in the United States and traffic originated in a foreign country. This ratio is 1.1 for Canada, 2.5 for Western Europe, but 6.1 for South America and 9.2 for Africa. The disparity is thus far greater with developing countries. Most experts interpret these patterns as reflecting the tendency for residents of industrialized nations to have far more disposable income than those of developing nations.

Figure5:. The ratio between international traffic originated in the U.S. and traffic to the U.S. originated overseas (messages and minutes of use)



Source: FCC 2004

This asymmetry 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 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]

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.

5.4 QoS

The ability to offer different levels of Quality of Service (QoS) has been seen as a key factor differentiating IP-based NGNs from the public Internet. It has long been recognized that tighter limits on network delay, and on variability of delay, would enable IP-based networks to deliver real-time bidirectional voice and video services more predictably and perhaps more reliably than the current best-efforts arrangements.

At the same time, the technical capabilities to deliver enhanced QoS have existed in the "public" Internet for at least a decade. Enhanced QoS is widely deployed within IP-based networks, but not *between* them. The reasons for this failure to deploy have little or nothing to do with technology, but rather flow from economic considerations. Given that the migration to IP-based NGNs does not intrinsically alter these economics, it is natural to ask whether deployment of differentiated QoS into IP-based NGNs will in fact be taken up without "help", and if not to wonder whether public policy initiatives should be considered. What sort of "help" might be beneficial?

In this section, we consider the economics of QoS; the technology of QoS; the reasons for its slow emergence between networks in the public Internet; the levels at which interconnection charges might be set; the implications of the network neutrality debate that has emerged in the United States; and finally whether it might be preferable to leave the economic arrangements to private arrangements (a "Coasian" approach) rather than attempting to regulate economic aspects of interconnection of IP-based NGNs.

5.4.1 General economic interpretation

It has long been recognized that providers of goods or services could potentially achieve some pricing power and profitability by distinguishing their goods and services, and by offering different qualities at different prices to different groups of customers. We experience this service and price differentiation every day. We drive into a gas station, and choose to purchase regular gasoline or premium. When we buy a ticket for a train or an airplane, we take it for granted that we may be offered first class and second class tickets, with a higher price for the former. To the extent that the amenities offered in first class have value to us, they increase our *surplus* (the difference between perceived benefits and cost), which in turn increases the price that we are willing to pay. The airline charges a higher price because it recognizes that those customers that value the amenities are willing to pay the higher price.

The basic notion of service differentiation is not new,[95] and the underlying economics have been well understood for many years.[96] Service differentiation recognizes that different consumers may have different needs and preferences, which translate in economic terms into a different *surplus* 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.[97] In practice, they generally target their distinct offers at different *groups* of consumers (second order price discrimination) rather than targeting different *individual* consumers (first order price discrimination).

In some cases, price discrimination may be linked solely to the willingness of the customer to pay, and largely unrelated to underlying costs. When an airline offers cheaper tickets to passengers who are willing to stay overnight on Saturday, the lower price has nothing to do with the airline's costs; rather, it reflects the greater willingness to pay (lower elasticity of demand) of business travelers. Business travelers are able to pay more, but are in most cases unwilling to stay overnight outside of the Monday to Friday time frame.

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.[98] 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."[99] 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.

Price discrimination is not invariably effective. It has been suggested that the propensity for Internet access to be priced at a flat rate is practically irresistible; however, this argument may be more persuasive in regard to prices paid by residential consumers than it is for prices paid at retail by large enterprises. "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. … Constraining architectures and pricing structures work against increased usage. … [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. … Now the Internet already pervades society, and will be even more ubiquitous in the future, used round the clock in a variety of applications. Simplicity is likely to be key to acceptance. … Hence telecom service providers are likely to discover that the elaborate architectures they are dreaming of will work against their interests."[100]

In summary, price discrimination tends to enhance consumer welfare (to the extent that it works effectively, and in the absence of market power). An open question is whether markets in some countries are sufficiently competitive for this to be the case. We return to this point shortly, in the sub-section "QoS and network neutrality".

5.4.2 Technical background

The following sub-sections discuss the nature of application requirements, the basic technical characteristics of packet transmission in the Internet as they relate to variable delay, and the implications of differentiated QoS for billing, accounting, and Operational Support systems.

5.4.2.1 Applications delay requirements

What sort of performance does a user need from the network? This is a function of what the user is attempting to do.

For typical data applications – email, for example – the user may need the ability to send a fair amount of data, but a great deal of packet delay is acceptable. Even if packets were to take several seconds to get through the network – which is an extremely long time by the standards of today's networks – the user would receive his or her email and would likely be satisfied with the service.

For streaming video – one way transmission akin to watching television – the network needs to transmit a great deal of information, but can still tolerate some variation in delay, as long as the user will accept a second or two of delay when the transmission begins. This is so because the receiving system can *buffer* the video data. As long as individual packets are not delayed more than the original start-up delay, the buffer corrects for occasional slow-downs. Even occasional loss of data will not necessarily result in performance unacceptable to the user, as long as the receiving system is designed to try to smooth over the gap.

Two way real time voice is, however, much more demanding. Many tests over the years have found that, where delay exceeds about 150 milliseconds, people on both sides of the conversation are likely to start speaking at once (because neither knows that the other has already begun to speak). This phenomenon is familiar to those of us who have used satellite circuits for international telephone calls. One can still conduct a conversation, to be sure, but nobody would prefer such a conversation to one conducted with low delay.

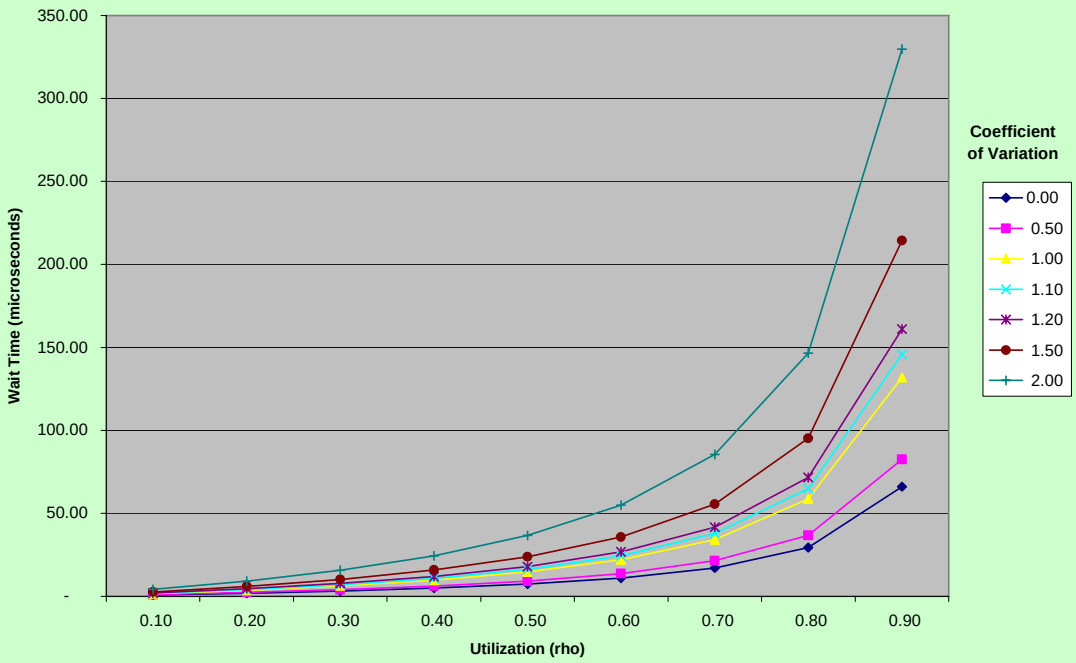
5.4.2.2 Delay in packet networks

It is natural to begin by asking the degree to which normal Internet traffic would meet demands for delay-sensitive traffic. The answer is, quite simply, that normal Internet traffic performs well enough nearly all of the time. This is the reason why services such as Skype and Vonage have customers – they work well enough today over the public Internet, with no special provisions taken to ensure quality of service.

This follows from basic queuing theory. Queuing theory is the branch of mathematics that deals with waiting lines – in this case, the waiting line to place a packet onto a high speed link between two routers in the core of the Internet.

The standard formulae for variable delay depend on how busy the transmission link is, how fast it is, how big the packets are and how variable in size. The standard analysis (which corresponds to a so-called M/G/1 queuing system) results in a family of curves, as shown in the figure below, where delay increases as the link becomes increasingly busy (moving toward the right of the graph). These particular curves reflect a 155 Mbps link, which is the *slowest* link that one is likely to find in the core of a high speed IP-based NGN. Variability in the length of the packets is reflected in the *coefficient of variation* of the packet length (and thus the service time), where a coefficient of variation of 0.0 denotes no variability at all, while a coefficient of variation of 1.0 reflects a nominal degree of variability corresponding to exponentially distributed packet lengths. A few years ago, a coefficient of variation of 1.2 was typical for Internet traffic.[101]

It is immediately obvious from the graphs that, even at exceedingly high loads of 90%, expected variable delay is less than 150 *microseconds* (where a microsecond is one millionth of a second). Given that our "budget" is in the range of 150 *milliseconds* (where a millisecond is one one-thousandth of a second) it is no surprise that IP traffic performs adequately most of the time – we could afford about 1,000 router hops, even under heavy load (as long as no link is truly saturated with traffic).



This does not mean that measures to better manage traffic are irrelevant. First, there is the risk that some link is completely saturated with traffic – a risk that cannot be completely avoided in light of the bursty nature of Internet traffic.[102] Second, there are the slower circuits at the edge of the network, especially shared media such as broadband based on cable television. And third, there is the matter of operation when the network is operation in a partially degraded mode due to the failure of some but not all components.

So the ability to provide QoS in a network has its place in a modern IP-based network. Still, the willingness of customers to pay a premium for a service that, most of the time and under most circumstances, provides no customer-visible benefits will surely be limited. This limited willingness to pay on the part of the customer naturally leads to limited willingness on the part of network operators to invest in making the capabilities available.

5.4.2.3 Technical solutions

In the early Nineties, the Internet Engineering Task Force (IETF) was active in evolving a series of relatively complex solutions under the rubric of the *Integrated Services Architecture*, as exemplified by the RSVP protocol. The common wisdom has been that these protocols were hopelessly complex.

In fact, at least one firm (BBN) had a working network based on RSVP. It was delivering services to real customers. It was a technical success but a commercial failure. It was eventually shut down, not due to technical problems, but rather because the company never found enough customers who were willing to pay much of a premium to use the RSVP-capable network.

Be that as it may, the IETF subsequently evolved a much simpler set of communications protocols in conjunction with *DiffServ* (for *Differentiated Services*). DiffServ enables hop-by-hop traffic management, where selected packets are marked as having requirements other than best efforts. It is up to each router, then, to do what it can to implement the desired transmission characteristics (or to decline to do so). Various techniques can then be used to ensure hop-by-hop performance, with *Multi-Protocol Label Switching (MPLS*) being perhaps the most common.

DiffServ provides a more limited service than RSVP in the sense that it assures performance only on a hop-by-hop basis, rather than end-to-end. Still, it can provide adequate overall assurance at a statistical level.

DiffServ and MPLS are trivial to implement within a network, and are in use in many large networks today. Nonetheless, there is no significant use *between* networks. The lack of deployment reflects economic and business factors rather than technical ones. We return to this point in a later section: “Why so slow to emerge?”

5.4.2.4 Implications for billing and accounting systems

The implications for Operational Support Systems (OSS) in support of differentiated QoS tend to be overlooked in most discussions. Technologists tend to focus more on the problem of getting the bits to move as they are supposed to move, and less on the problem of how to ensure that someone pays for those movements.

It has generally been assumed that a network operator would be willing to provide better-than-best-efforts quality only to the extent that either the end user or another network operator was willing to pay them a premium to do so. To the extent that this implies the need to account for QoS-capable traffic, it implies surprising complexity.

First, a pair of network operators would need to agree on how much QoS-relevant traffic each delivered to the other. Second, they would need to verify that each actually delivered the quality that it had committed to the other. Finally, each would need some tools to deter fraudulent use or “gaming” of the system. The first is trivial, the second and third are difficult. Finally, there would be the need to reconcile statistics, and to deal with discrepancies between the measurements of the parties.

Measuring traffic across a link would seem to be straightforward, and distinguishing among different marked classes of traffic is no harder. Capturing first-order statistics on traffic sent between the parties is straightforward. Even here, some prior agreements will be needed as regards what is being measured, and when – otherwise, there is the risk that network A has a slightly different view of the traffic delivered on the link from A to B than does network B, even though both are measuring (different ends of) the same link using substantially similar tools. And sampling intervals need to be mutually agreed, otherwise any measures of variability (quantiles, standard deviation) are likely to reach different conclusions due to the perverse effects of the Central Limit Theorem (if two sensors sample the same distribution, the one that is sampling at more frequent intervals will tend to see an apparently more lumpy distribution).

Reconciling data would be challenging. There is an old Dutch proverb: “Never go to sea with two compasses. Take one or three.” If the providers do not agree, whose statistics should govern? Is there scope here for a trusted intermediary, and if so who might that trusted third party be?

The challenges in verifying that the service was actually delivered are much more profound. In this case, network A needs to ensure that network B delivered the committed performance, and neither will want to rely on measurements provided by their respective end users. Network A thus needs performance statistics

about network B's network. Yet these networks are likely to be direct competitors for the same end users – network B is not about to let network A place sensors in its network. Both networks are likely to uncomfortable with providing internal performance data to one another.

It might be far simpler to bill, not for the use of the network, but rather for the services that benefit from differentiated QoS. Here, too, there are challenges – in an IP-based NGN, the service provider might not be the network provider. Moreover, it is quite possible, for reasons noted earlier, that services without QoS will compete successfully with services that are supported by QoS. It is not clear that network operators would be able to extract enough revenue from independent service providers to enable them to fund the differentiated services.

5.4.3 Why so slow to emerge?

Given that the technology of differentiated QoS is not particularly challenging, and given its widespread use *within* networks, why has it been so slow to achieve deployment *between and among* networks?

From an economic perspective, the basic answer is obvious: Had the benefits of deployment clearly exceeded the costs, it would have deployed. Thus, one might infer that either the perceived costs are too high, or the perceived benefits too low, or perhaps both.

Given that most users will be unable to notice the difference most of the time, there are indeed questions as to whether the perceived benefits are too low. In addition, a series of challenges related to *network externalities* and to *transaction costs* have inhibited deployment.

Many industries experience network externalities. A service may be most useful when a great many people use it (and not just because of economies of scale). This is true of telephone service, and also of the Internet. My telephone is worth more if there are a great many people whom I can call. My Internet connection is worth more when there are a great many people to whom I can send an email, and a great many websites to which I can connect.

Getting a new service launched in a sector dominated by network externalities can be challenging. In effect, the externalities of the old service keep pulling the industry back. It is difficult to get past the *initial adoption hump* in order to achieve critical mass.[103] The economist Geoff Rohlfs[104] has explained that different services got past the initial adoption hump in different ways. VCRs were initially purchased for time-shifting of television programs; only when enough consumers had purchased VCRs did a rental business emerge. CDs were successful because Matsushita and Phillips had commercial interests in both CD players and studios, and were thus motivated to ensure that both players and content were available.

Differentiated QoS *between and among networks* is subject to these network effects. The service has some value within a network. It might have great value if it were available to every destination on the Internet. If it were available to only two or three networks, then it is of limited value. Thus, the value of deployment might be significant to those networks that implement it later, but the initial benefit to the first two or three networks to deploy it is minimal.

At the same time, extending differentiated QoS to each additional network implies *transaction costs*. Agreements, monitoring tools, and coordination in general would need to be put in place. These costs might be roughly linear in the number of networks with which one network has agreements in place.

Thus, it is hard to get the process started, and it would be hard to get it to completion once it had been launched.

These concerns are not unique to differentiated QoS. A number of Internet capabilities are faced with similar economic challenges. The adoption of Internet Version 6 (IPv6, a new version of the Internet Protocol with a greatly expanded address range) and of DNSSEC (a security enhancement to the Domain Name System – the acronym alludes to *DNS SECurity*) have arguably been impacted by similar considerations.[105]

On the other hand, certain Internet capabilities have deployed effortlessly – for example, the worldwide web. In many cases, the successful capabilities benefit from the *end to end principle* – that is to say that they can be implemented independently by end-user organizations or consumers, without requiring coordination or for that matter any action at all on the part of the providers of the underlying IP-based network.

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 coordinated 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:

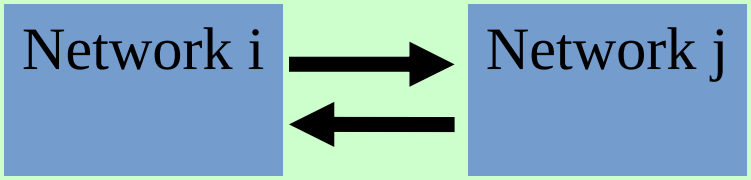
- Benefits that are in some sense insufficient: too limited, too difficult to quantify, too slow to appear, too difficult for the investing organizations to internalize.
- Limited benefits until the service is widely deployed.
- The need for coordination among a large number of organizations, leading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

5.4.4 QoS and interconnection charges

In an unregulated environment, how would service providers (be they ISPs or NGNs) choose to set access charges for carrying one another's traffic at levels of quality other than the current best-efforts level of quality? It has widely been assumed that QoS between IP-based backbone ISPs would work only if the backbones compensated each other at a higher rate (i.e. with a higher access charge) for service of better quality.

Economists have modeled a pair of interconnected backbone ISPs (networks *i* and *j* in the figure below) serving a customer base of web sites (senders of traffic) and consumers (recipients of traffic). [106]

Figure 6: Traffic flows from network *i* to network *j*



In the absence of differentiated Quality of Service, backbone ISPs in a competitive market would tend to price to usage-based marginal cost. Profits in this scenario are nil – the backbones compete away their profits.[107] If, on the other hand, the backbones have market power (as a result of having designed their respective networks to inherently provide different levels of quality of service)[108], then profits can be positive – the network providers can make a profit.

In this latter scenario, if the ISPs face identical costs but distinct demands based on the differences in their quality of service, both will be motivated to price at an equilibrium price that maximizes their profits. At that point, an increase or decrease in either ISP's price would not further increase profits.[109]

In sum, the analysis at this level is consistent with the notion that a system where IP-based networks have the ability to offer a higher level of quality of service (for traffic delivered at higher cost and associated with higher access charges) could lead to an economically rational system that would tend to increase provider profits. As we show in the next sub-section, in a generally competitive market, this would tend to enhance consumer welfare.

5.4.5 QoS and network neutrality

In the United States, a recent debate has emerged over *network neutrality*. The arguments on both sides of this complex debate have arguably been somewhat misplaced, but it is worth noting that a number of experts have implicitly objected to price discrimination and to the use of technology to support the excludability that would make price discrimination effective.

The network neutrality debate emerged in the United States due to a "perfect storm" of three simultaneous market and regulatory changes:

1. The collapse of the wholesale market for broadband Internet access;
2. A series of mergers (Cingular/AT&T Wireless, SBC/AT&T, Verizon/MCI, and now AT&T/BellSouth) with insufficient conditions imposed; and
3. The overly hasty and ill-considered withdrawal of procompetitive regulation.[110]

These concerns about price discrimination appear to reflect excessive concentration in the U.S. market – regulatory experts are objecting to many practices that, in a healthy market, would be welfare-enhancing. In the U.S. context, these concerns are real; moreover, they cannot easily be fixed through regulation. The problems are too complex. Several U.S. FCC proceedings[111] suggest that the FCC lacks the necessary expertise to distinguish between welfare-enhancing service discrimination versus harmful anticompetitive acts. In any case, once markets have been allowed to deteriorate to this degree, no regulatory fix is likely to be satisfactory. The fox is already in the chicken house, the horse has already left the barn.

These network neutrality problems will not necessarily manifest themselves in other parts of the world, or at least not in exactly the same way. In Europe, for example, the underlying broadband markets are much more competitive than in the United States, in the sense that far more options are available to the average consumer; moreover, the regulatory system in Europe is likely to ensure that they remain competitive. Even in relatively concentrated European markets such as Germany, most consumers can choose among multiple broadband service providers (many of them service-based rather than facilities-based). As long as regulators continue to ensure competitive underlying markets, offers of different quality of service at different prices are likely to enhance consumer welfare rather than to detract from it.

5.4.6 A Coasian approach to QoS

As we have seen, Coasian arrangements (i.e. commercially negotiations) work well in a wide range of settings for interconnection agreements to convey best-efforts traffic. (In this sense, *best-efforts* connotes traffic for which the network exerts its best efforts to ensure that the traffic is delivered, but provides no ironclad guarantee that all traffic will be delivered. There may, of course, be *statistical* assurances as to the probability of delay or loss of data traffic.) What are the prospects that Coasian arrangements might work for traffic some fraction of which must be delivered with quality better than best-efforts?

Reasoning by analogy with current Coasian best-efforts arrangements, we might reasonably expect that each network would seek to be paid by its own customers, rather than seeking payment from either the consumer or the content provider customer of the other network. The charge to one's own customer for carrying better-than-best-efforts traffic would presumably be higher, on a unitized basis, than the charge for carrying best efforts traffic.

The *service* providers (of VoIP, for example) would likely charge their customers, and might also agree to charge one another. These payments between the services providers would then constitute a form of *micro-payment*.[112] In a competitive market, the *service* providers would tend to set these payments in such a way as to cover their costs, including their payments to the underlying *network* providers. A key insight from Laffont et. al. is that, where such micropayments exist, the access charge typically has no impact on traffic or on economic efficiency.[113]

This possibly simplistic analysis would seem to suggest that differentiated QoS might emerge spontaneously without any policy intervention on the part of governments. The seemingly intractable problems associated with access payments between service providers need not be dealt with at all – access charges could be set to zero, or simply ignored altogether. Unfortunately, this analysis begs the question: If it were that easy, why has it not already happened?

6 A sampling of regulatory analysis of NGN interconnection to date

This section briefly reviews a number of regulatory proceedings and studies that have considered the impact of IP-based interconnection. Many countries have looked at these issues.

Section 5.6.1 considers the overall approach of the European Union, which largely uses economic analysis to address market bottlenecks in a technologically neutral and reasonably future-proof way. Sections 5.6.2 and 5.6.3 consider recent developments in the UK and in Germany, which could be viewed as reflecting forward-looking thinking in the industrialized world. Section 5.6.4 and 5.6.5 briefly summarize recent proceedings in India and Saudi Arabia by way of showing emerging views in developing countries.

6.1 The European approach

Different countries will have developed different methodologies for addressing market power as it relates to access and interconnection. 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 (insofar as practicable) 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. The determination reflects consideration of the degree of substitutability with 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.[114]

6.2 The United Kingdom (U.K.)[115]

The use of the European regulatory framework simplifies the task of the regulator in addressing the challenges of the Next Generation Network, but by no means does it provide all of the answers. Ofcom, the national regulatory authority (NRA) for the UK, has been in the forefront in dealing with NGN migration challenges, largely as a result of the commitment of British Telecom (BT) to migrate rapidly to an NGN and to phase out the existing "traditional" PSTN network. These proceedings represent cutting edge thinking, and merit careful study by regulators.[116]

Among these Ofcom proceedings are:

- *Next Generation Networks – Future arrangements for access and interconnection* (First Consultation), 24 October 2004.
- *Next Generation Networks: Further consultation* (Further Consultation), 30 June 2005.
- *Final statements on the Strategic Review of Telecommunications, and undertakings in lieu of a reference under the Enterprise Act 2002* (Strategic Review), 22 September 2005.
- *Regulatory challenges posed by next generation access networks*, 23 November 2005.
- *Ofcom's approach to risk in the assessment of the cost of capital*, 26 January 2005 (updated 2 February).
- *Ofcom's approach to risk in the assessment of the cost of capital: Second consultation in relation to BT's equity beta*, 23 June 2005.
- *Ofcom's approach to risk in the assessment of the cost of capital: Final statement* (Final Statement), 18 August 2005.

- *Review of BT's network charge controls: Explanatory Statement and Notification of decisions on BT's SMP status and charge controls in narrowband wholesale markets*, 18 August 2005.

The Ofcom proceedings have considered a range of issues, including the degree to which regulatory remedies would continue to be needed going forward, the number of points of interconnection that would be required, the appropriate way to determine regulated costs and prices in an NGN environment. Many of these proceedings have concluded that it was simply too early to make a firm regulatory determination; consequently, they have tended to focus on putting clear and consistent *process* in place rather than making firm conclusions about the regulatory *outcome*.

At the same time, Ofcom has reached understandings with BT that establish a semi-independent subsidiary to provide those wholesale products that are closely associated with the local loop to the customer's premises, where BT is presumed to still possess market power.

Ofcom concluded early on that the migration to an NGN would not in and of itself eliminate BT's market power on the local loop any time soon. The subsidiary (*OpenReach*, referred to in some Ofcom documents as the *Access Services Division (ASD)*) is obliged to provide access products at wholesale to the competitors on the same basis on which it provides them to BT itself, a principle known as *Equivalence of Input*. Through these arrangements, Ofcom hopes to redirect critical obligations to a small number of services where barriers to competition are most likely to be durable, and by so doing to obviate the need for obligations on other services (higher level services) that use those lower level services as key piece parts.

6.2.1 New access products required

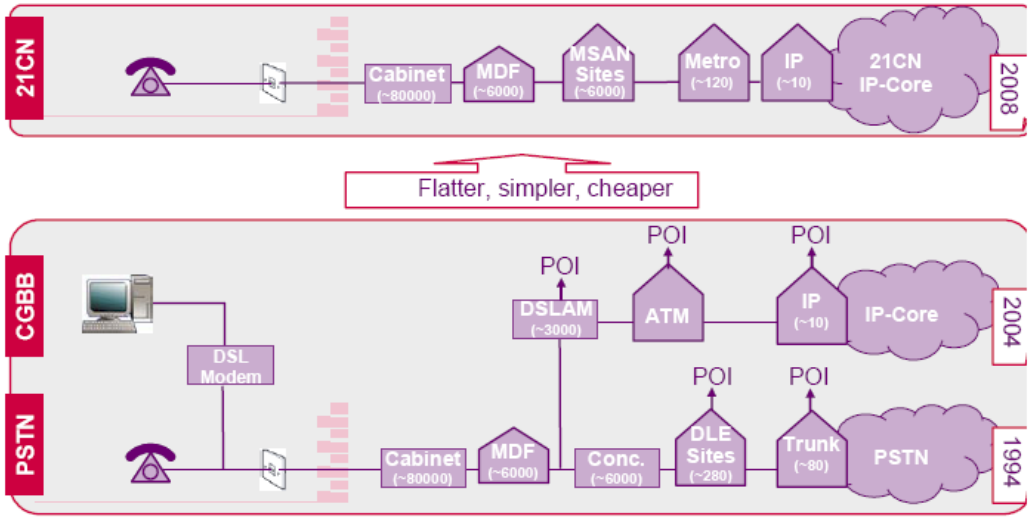
As regards existing SMP obligations, and specifically existing SMP interconnection offerings, Ofcom came to the unsurprising conclusion that those offerings would need to be maintained for some period of time. At the same time, they came to the equally unsurprising realization that new SMP interconnection offerings would be appropriate in the future. This necessarily implies some period of overlap:

> To enable business planning for alternative providers there initially needs to be continuity of existing SMP products (those products that BT is obliged to offer in markets where they have Significant Market Power), but we believe that this should only be for an interim period during which both legacy and next generation products are available. To ensure a timely move to next generation interconnect we propose that legacy products should be withdrawn once there is no longer reasonable demand or when next generation products provide an adequate replacement that providers are able to migrate to.[117]

6.2.2 Fewer access points

Ofcom recognized[118] that the new structure that BT envisioned for 21CN necessarily implied a flatter network with fewer points of interconnection. Today, BT has some 3,000 locations at which competitors can connect to the DSLAM, and some 280 Digital Loop Exchange (DLE) sites at which competitors can gain access to the voice network. In the 21CN as currently envisioned, interconnection will be possible only at the metro nodes, i.e. only at 100 – 120 sites.[119]

Figure 1. Comparison of existing BT voice and broadband networks with 21CN



Key:

PSTN	Public Switched Telephone Network	Conc	Concentrator
CGBB	Current generation broadband	DSLAM	DSL Access Module
21CN	BT's 21st Century Network	DLE	Digital Local Exchange
MDF	Main distribution frame	MSAN	Multi Service Access Node

This inevitably raised many questions. Alternative operators had invested significant sums to interconnect with BT at existing interconnect locations. Now, as a result of unilateral decisions set in motion by BT, many of those interconnect locations would no longer exist. How should the costs of these stranded assets be apportioned?

Ofcom found[120] "…that the key factors relevant to compensation arrangements for BT's 21CN migration are:

- the extent to which these changes are unilaterally decided by BT without industry agreement;
- the distribution of benefits that accrue from these changes;
- the remaining life of any legacy interconnect equipment employed at the time of the change;
- the extent to which new interconnect investments are made by communication providers after they have been made aware of forthcoming changes that would impact that investment; and
- the additional cost necessarily and directly incurred as a result of having to bring forward investment in new interconnect equipment.

6.2.3 Call termination in the traditional (narrowband) PSTN

Ofcom's analysis of *narrowband* (where narrowband refers to the existing PSTN, at bandwidths sufficient to carry a telephone call) call termination fees is particularly interesting, both for what is said and for what is not said. Ofcom must address an incumbent (BT) that is moving rapidly today, not merely to augment its network with NGN capabilities, but rather to replace its existing PSTN network altogether in just a few years with a new IP-based NGN known as the *21st Century Network (21CN)*. Driven by that need, they have developed a thoughtful analysis of how they might carry their existing narrowband arrangements forward during a transition period where the incumbent network is based partly on PSTN technology and services, and partly on those of the NGN.

Ofcom does not attempt the broader and more difficult task that is attempted in this chapter – determining whether the traditional PSTN model of interconnection will be appropriate going forward in a purely NGN world. In that sense, their analysis could be said to look back to the past, rather than forward to the future. Nonetheless, it contains valuable insights for regulators seeking to address the transition period from PSTN to NGN.

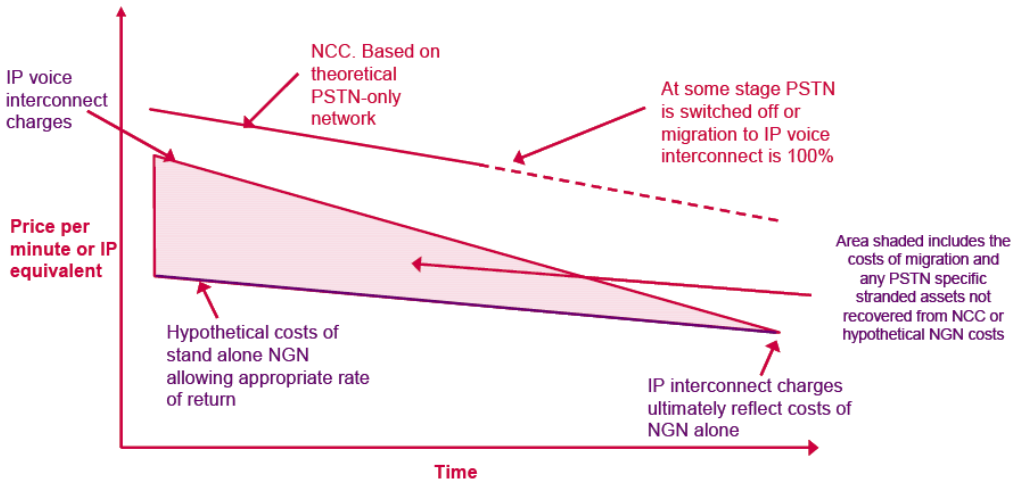
Ofcom recognized several interrelated factors of the current regulatory and market environment that could influence BT's profitability and that of its competitors, including:

- the higher risk to BT and its shareholders in implementing 21CN;
- the ultimately lower unit costs of operation for 21CN;
- the losses to competitors associated with stranded investments in interconnection facilities; and
- the cost to BT of simultaneously offering both old and new SMP offerings during the period of transition.

Ofcom has put forward the following over-arching view of the relationship between BT's risk, its initial capital expenditures in creating the 21CN, its lower overall unit costs once 21CN is fully operational and once legacy SMP offerings are no longer required in parallel, and the desired course for regulated rates for narrowband voice interconnection (and presumably to other regulated prices) as a result:

> *… IP voice interconnect charges would need to start above the costs of a hypothetical stand alone NGN, because to do otherwise would create an arbitrage opportunity where (for example) migration costs would not get recovered. However, these IP voice products could still be priced below [current] narrowband interconnect products to the extent they cost less to provide than [those current] interconnect products. Finally, at a point in the future, when all traffic is via IP voice interconnect, and all migration / PSTN costs had been recovered, IP interconnect pricing would end up reflecting the costs of the NGN allowing an appropriate rate of return.*[121]

Figure 5. 'Holistic' approach to narrowband voice interconnect cost recovery (illustrative only)



Source: Ofcom

This intriguing diagram represents a fascinating thought model, but it also raises many questions.

The upper line ("NCC. Based on theoretical PSTN-only network") reflects the expected trend for the regulated *Network Charge Control (NCC)* for BT's existing wholesale narrowband interconnect. It declines over time because BT's efficiency is presumed to improve over time. It is implicitly assumed that the efficiency of a network that is part PSTN and part NGN will improve no less quickly than BT's current PSTN network. In the event that the migration to NGN enables still greater efficiency gains, then BT reaps the benefit over the defined lifetime of these cost controls, which is 2005-2009 – the NCC level will not be revised other than in exceptional circumstances.[122]

The next line below, "IP voice interconnect charges", represents a future NCC for a new wholesale SMP product enabling interconnection to narrowband voice services. It is presumably some form of IP interconnection. Given that this interconnect offering is not yet defined, much less implemented, Ofcom has not attempted to set the level of these charges; however, the general notion is that they should be less than those of traditional voice interconnect charges, but still sufficiently in excess of incremental cost to enable BT to recover the cost of migration from the PSTN to the NGN.

The question of NCC levels during a period of coexistence between traditional interconnect and new IP-based interconnect is complex. To the extent that network costs are lower, the interconnect price for IP-based interconnect should be lower; however, maintaining different interconnection prices for the same service would tend to result in lower retail prices, and would thus tend to drive customers of the wholesale service to the new IP-based mode of interconnection. That could accelerate an already rapid transition.

6.2.4 Risk, the permissible rate of return, and the Weighted Average Cost of Capital (WACC)

Regulation must not interfere with the ability of 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) can be computed using the *Capital Asset Pricing Mechanism (CAPM)*, a widely used and theoretically well grounded methodology for reflecting risk and its impact on the returns that shareholders should expect. A key component of the WACC under CAPM is the *beta*, which is a relative measure of the risk that is relevant to the specific firm.

Ofcom's analysis of the appropriate WACC for BT[123] 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*[124].

6.2.5 The Equivalence of Input approach

Ofcom announced an agreement with BT in June 2005 that represents a significant departure from previous regulatory practice.[125] BT made legally enforceable commitments[126] to provide a range of access services to competitors on a nondiscriminatory *equivalence of input* basis. Ofcom defines *equivalence of input (EoI)* as "…a requirement for BT to make available the same [*Significant Market Power (SMP)*] products and services[127] to others as it makes available to itself, at the same price, and using the same systems and processes." EoI obligations would be applicable "… when the cost is proportionate, and in particular [to] all new wholesale SMP products, processes and systems, and therefore to all new SMP products delivered over 21CN."[128]

BT has agreed to make key wholesale offerings where it has previously been found to be dominant available to competitors *on an equivalence of input basis*. Most if not all of these offerings are, to be sure, available today in connection with *ex ante* remedies imposed in response to BT's market power. What the commitment implies is that these wholesale services must be delivered by BT's Access Services Division (Openreach) as of the committed dates using new order processes and new nondiscriminatory ordering systems.

This is an interesting and promising model for the regulatory community, to the extent that it implies an attempt to achieve regulatory aims, not by traditional regulatory means, but rather by an enforced but limited structural separation of the wired incumbent. The hope is that the incentives of the new Access Services Division will be such that day to day regulatory intervention will be unnecessary. It is too soon to say whether this approach will prove to be effective, but it merits careful study going forward.

6.3 Germany

The German BNetzA initiated a study group on the interconnection of IP-based networks, which issued its final report at the end of 2006. The report notes the importance of the separation of service and network, and the impact of the centralization of control functions.

As regards the number of interconnection points, the report concludes that it is not yet possible to predict what will happen, for both technical and business reasons. From a regulatory perspective, the number of interconnection points should insofar as possible be consistent with an efficient network architecture for the incumbent and also for competitors, and should seek to minimize stranded investments for all concerned.

IP-based networks have a clear focus on end-to-end quality. The report recognizes the value of providing different classes of service, and identifies four possible classes: real time service, streaming service, data service, and best efforts service. The questions associated with compensation for the higher costs of more demanding services are complex and interlinked. The report draws a useful distinction between VoIP services versus VoNGN (voice over Next Generation Networks), but recognizes that the marketplace should make the ultimate choice between the two.

Unit costs for NGNs will presumably be lower than for current networks; however, an immediate transition to these forward-looking costs could potentially be too disruptive to market participants. A glide path may be more appropriate, and may better reflect the transition period where the network comprises a mix of circuit-switched and NGN technologies. A single interconnection regime reflecting a "blended" cost structure may also serve to reduce the risk of arbitrage and bypass in comparison with an interconnection regime where PSTN interconnection is priced at a very different level from that of NGN interconnection. A unified interconnection regime is also likely to result in more appropriate market-based incentives for all concerned for a migration to NGN with an appropriate number of points of interconnection. The BNetzA will in any case need to undertake more detailed study of costs in an NGN environment.

The migration to NGN puts in sharp contrast the common practice of CPNP in the traditional switched network, versus the practices of Bill and Keep and of transit arrangements in the IP-based world. CPNP leads to the termination monopoly problem, which necessitates a regulatory correction; Bill and Keep eliminates the termination monopoly, and thereby reduces the need for regulation, since there is no need to set termination prices.

Bill and Keep would tend to reduce the incentive for operators to concentrate on groups of customers who receive more traffic, and may increase the tendency toward roughly comparable traffic among network providers, even when the providers are very different in size. These characteristics can enhance efficiency. At the same time, Bill and Keep might lead to the "hot potato" problem, which could result in underinvestment in network infrastructure.

A possible way to integrate these views would be to implement Bill and Keep in the concentration network, but CPNP on an *Element Based Charging (EBC[129])* basis in the core network.

The report provides a wealth of analysis, but does not reach firm conclusions as to the preferred long term interconnection regime (CPNP, Bill and Keep, or a hybrid of the two), nor as to likely time frame for transition from current to future arrangements.

6.4 India

The Indian TRAI issued a consultation paper[130] on NGNs early in 2006. The section on interconnection serves primarily to raise questions to stakeholders. The TRAI notes potential concerns that a dominant operator might refuse to offer new forms of interconnection to competitors, thereby disadvantaging them; and legacy services might be withdrawn or altered on short notice, to the detriment of competitors.

Among the questions that the TRAI poses are:

- Should independent VoIP providers who wish to terminate calls have interconnection rights?
- What interconnection services are appropriate going forward? Should NGN interconnection cover "… only [the NGN] Core, Core and Access or all three layers including Service?"
- Should interconnection charging continue to be based on time and distance, or should it immediately evolve to be based on capacity, quality, and class of service?
- What obligations, if any, should the incumbent have to continue to offer legacy interconnection products?
- How should the transition period to the NGN be handled?
- Given the complexity of interconnection issues, is a separate consultation dealing only with interconnection issues necessary?

6.5 Saudi Arabia

The Saudi Communications and Information Technology Commission (CITC) briefly considered IP-based interconnection as part of their 2006 review of the Reference Interconnection Offer of the Saudi Telecommunications Company (STC).[131] They note that "…IP traffic does not lend itself easily to per minute charging, and it is technically complex to separate one kind of traffic (e.g. voice) from another (e.g. http traffic) where many different types of traffic may be carried simultaneously across the same interconnection link. This raises issues about how service providers should charge for interconnection, and the issues are particularly complex when traffic has to be passed from a circuit-switched to an IP environment, or vice versa. The CITC believes that interconnection to NGNs will be an important issue in the Kingdom in future , and intends to begin a process of analysis and consultation to arrive at an optimal set of regulations for NGN interconnection within the Kingdom."

7 Access and Mobile Next Generation Networks

This section briefly addresses some concerns that are unique to mobile networks. As networks evolve into Next Generation Networks, fixed and mobile networks are likely to become increasingly interoperable and substitutable for one another. Nonetheless, important differences will remain, and will continue to require regulatory attention.

7.1 Is it appropriate that termination fees be higher for mobile networks?

Traditional economic theory has held that the termination fee should reflect the real marginal cost of terminating the traffic. To the extent that the mobile network entails greater cost than the fixed network, it is appropriate that the termination fee should be higher.[132]

Computing the appropriate termination fee is a complex specialist area. The operator should be able to recover not only usage-based marginal costs, but also a reasonable return on the capital invested. For a mobile service, a large part of that capital may have been invested in procuring spectrum at auction. The costs are generally based on some measure of long range forward-looking incremental cost for an idealized network. The rationale for the use of an idealized network is to avoid rewarding operators for poor design decisions.

Over time, the migration to NGNs is likely to provide competitive alternatives that will make the current interconnection arrangements unsustainable. For the moment, however, mobile operators seem to be more

tightly vertically integrated than their fixed counterparts. Independent third party providers of services such as VoIP have been less successful in cutting into the revenues of mobile operators. The current minutes-based arrangements might be sustainable somewhat longer in the mobile network than in the fixed network.

Regulators may be willing to migrate the fixed network to different interconnection arrangements (such as bill and keep) more quickly than the mobile network, in part because there is far less revenue involved. Mobile termination rates are many times as great as fixed.

Is there cause for concern if the regulators implement a mixed system, at least for some period of time? Perhaps not. Mixed systems are ungainly, but not impossible. France used Bill and Keep wholesale arrangements among mobile operators for many years, while using CPNP for calls to or from the fixed network. This system was ungainly, but in practice was unwieldy rather than unworkable.[133] Similarly, the United States has used Bill and Keep for years for mobile operators and for non-dominant fixed operators, but has used CPNP for calls to wired incumbents. The resultant system is hideously complex, but the practical results are surprisingly good. Singapore has achieved outstanding results with a U.S.-like mixed system that uses Bill and Keep for calls terminated by mobile operators, but regulated rates for calls terminated by the wired incumbent.

7.2 Mobile services versus WiFi services

One interesting competitive alternative to mobile telephony that is emerging is the use of WiFi in conjunction with VoIP. New handsets have recently emerged that operate as VoIP phones when WiFi is available, but as conventional mobile phones at all other times.

The original expectation was that people would use these phones to avoid high mobile charges when making calls from their own homes, but would use them primarily as conventional mobile phones when outside. A somewhat unexpected manifestation is that many WiFi hot spots are completely open and unsecured. A person walking down a big city street might very well be able to place a free VoIP/WiFi call by piggy-backing on a stranger's WiFi router.[134]

This introduces numerous practical and ethical questions. The person who owns the WiFi router probably will never notice that his service is being used, and suffers no significant degradation of service; nonetheless, he or she is arguably experiencing a microscopic theft of service. Should the practice be viewed as being acceptable, or objectionable?

Also, the owner of the router could easily eavesdrop on the conversation. Should this be a matter of concern?

7.3 Mobile Roaming

Mobile roaming is a complex market that is subject to many of the same defects as the mobile termination market. In the absence of regulation, prices have a tendency to be greatly in excess of real marginal cost.

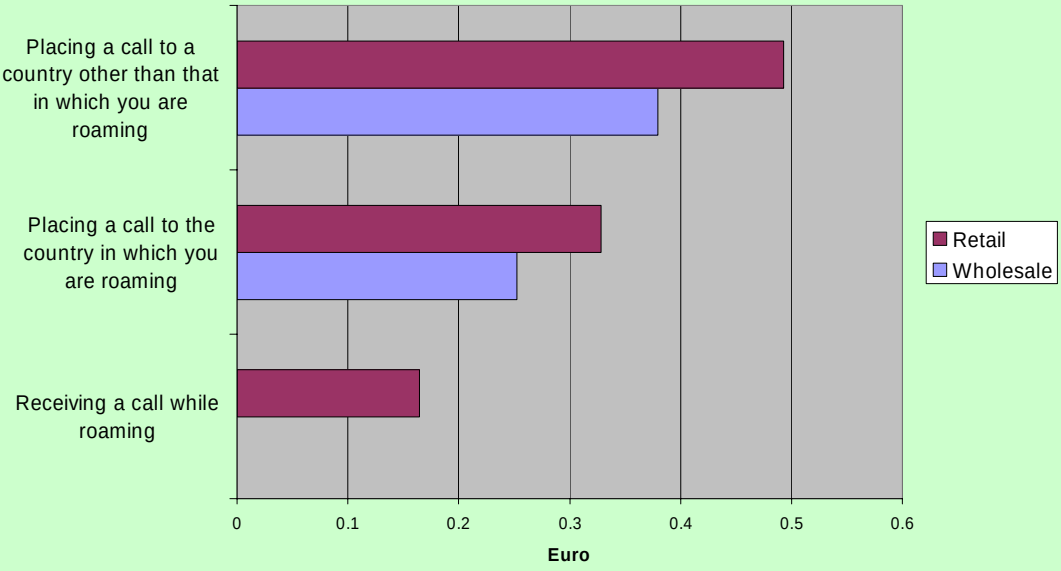
Mobile roaming consists of using one's mobile phone when one is outside of the service provider's service footprint. Roaming in a country where the service provider lacks a presence is *international mobile roaming.* Placing a call while roaming is *roaming out*; receiving a call while roaming is *roaming in*.

European regulators have become increasingly concerned over high roaming prices, both at wholesale and also at retail levels. Roaming out in Europe typically costs more than 1.00 EUR per minute, while roaming in often costs 0.50 EUR per minute. Real marginally costs are surely much less.

The European Commission has proposed a regulation to impose both wholesale and retail price caps on mobile roaming.[135] The imposition of retail price controls is a rather extreme measure, but is motivated in this case by the recognition that retail prices of mobile roaming have stubbornly refused to move very much over the past few years; moreover, since there is no wholesale price for roaming in, there is no serious prospect of addressing that aspect of the problem solely at the wholesale level.[136]

The Commission's proposed regulation would "peg" wholesale prices for mobile roaming to levels that are based on regulated mobile termination rates (MTRs). Based on the mobile termination rates that were in effect as of October 2005,[137] maximum wholesale and retail prices per minute would be as shown in Figure 7 below for wholesale and retail mobile roaming.

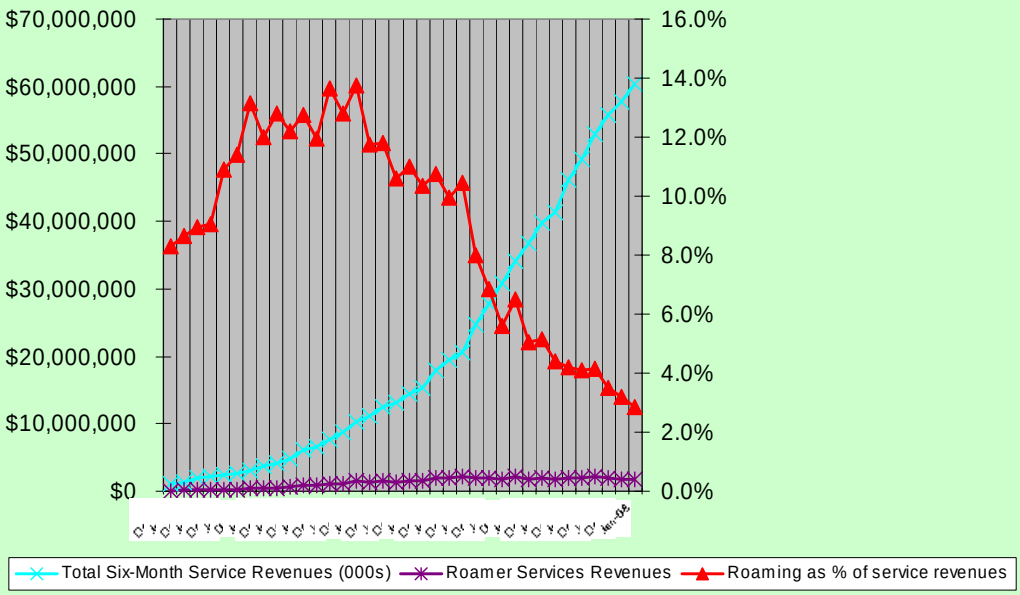
Figure 7: Maximum wholesale and retail prices for mobile roaming under the European Commission's proposed mobile roaming regulation



Source: European Commission

In the United States, domestic mobile roaming (within the U.S.) has been a problem of declining significance for many years. No regulation was necessary. Once AT&T Wireless offered a flat rate package with no domestic roaming charges, the industry moved quickly to transform itself. Mobile operators were effectively forced to become national in scope in order to compete, either through acquisitions or through partnerships. Mobile revenues in general have climbed steeply over the years, but mobile roaming revenues have been essentially flat and more recently have declined somewhat. As a percentage of total revenues, domestic roaming has declined from 13.8% in December of 1995 to 2.8% in June of 2006.[138]

Figure 8:. Domestic mobile roaming revenues in the U.S. in relation to total service revenues



Source: Cellular Telecommunications and Internet Association

Mobile roaming is proving to be a challenging problem; however, there is no particular reason to expect that the migration to IP-based NGNs will change the character of the problem very much.

8 Last Mile Access

Up to this point, this chapter has dealt primarily with network *interconnection*. The section deals primarily with *access*. For our purposes, *access* enables an operator to utilize the facilities of another operator in the furtherance of its own business and in the service of its own customers.

At the regulatory and policy level, access and interconnection have 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.[139] 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.[140]

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.[141]

9 Concluding thoughts

As today's networks evolve into Next Generation Networks of tomorrow, much will change, but some things will remain the same.

- The character of competition may change, but competitive bottlenecks will continue to be of concern to regulators.
- The close link between the *service* and the *network* will be broken, but the regulatory interest in protecting consumer rights in connection with existing services and with their successors is likely to remain.

The time of transition to NGNs represents an appropriate point at which to reconsider the entire interconnection regime. The CPNP wholesale arrangements that are globally widespread today will be difficult or impossible to maintain without change going forward, for a variety of technical and practical reasons. Interconnection arrangements must evolve in any case, This is the appropriate time at which to reexamine the workings of the entire interconnection system.

A better trade-off between *adoption* and *use* of communication services is needed. CPNP has stimulated take-up of mobile services, but at the cost of substantially depressing the usage of mobile phones, and possibly also to the detriment of the fixed network.

In the longer term, arrangements reflecting withdrawal of a regulatory mandate for CPNP, akin to regulatory practice in North America (Bill and Keep) and in the Internet, may represent the most appropriate and most sustainable economic model. These systems are economically efficient; they encourage usage (albeit possibly at some cost in the speed of mobile penetration); they ease the task of the regulator, to the extent that regulatory rate-setting is not required; and they pose no conceptual or implementation difficulties in the world of the NGN.

Based on existing experience, Bill and Keep arrangements are likely to work well without a regulatory mandate to interconnect,[142] provided that key underlying markets are competitive, and provided that otherwise dominant operators are motivated (or possibly required) to interconnect on an IP basis to at least two or three significant domestic competitors. The regulator's task might be simplified in this scenario, but the regulator must nonetheless remain vigilant.

In the nearer term, CPNP systems with much lower termination fees than those typical today might represent a promising interim step. Experience in India suggests that CPNP arrangements with mobile termination fees less than 0.01 USD per minute can be compatible with both high usage and rapid adoption. By reducing the spread between CPNP and Bill and Keep, the regulator also greatly reduces the pain associated with a subsequent transition to Bill and Keep arrangements should such a transition prove necessary.

---

[1] See, for example, Article 2 of the European Union's *Access and Interconnection Directive*: "(a) access means the making available of facilities and/or services, to another undertaking, under defined conditions, on either an exclusive or non-exclusive basis, for the purpose of providing electronic communications services. It covers inter alia: access to network elements and associated facilities, which may involve the connection of equipment, by fixed or non-fixed means (in particular this includes access to the local loop and to facilities and services necessary to provide services over the local loop), access to physical infrastructure including buildings, ducts and masts; access to relevant software systems including operational support systems, access to number translation or systems offering equivalent functionality, access to fixed and mobile networks, in particular for roaming, access to conditional access systems for digital television services; access to virtual network services; (b) 'interconnection' means the physical and logical linking of public communications networks used by the same or a different undertaking in order to allow the users of one undertaking to communicate with users of the same or another undertaking, or to access services provided by another undertaking. Services may be provided by the parties involved or other parties who have access to the network. Interconnection is a specific type of access implemented between public network operators …". *Directive 2002/19/EC of the European Parliament and of the Council of 7 March 2002 on access to, and interconnection of, electronic*

communications networks and associated facilities (Access Directive), Official Journal of the European Communities, L 108, April 24, 2002, Article 9.

[2] Access can be viewed as an inherently asymmetric relationship, in that one operator is providing access to another. Interconnection, by contrast, can be viewed as being a symmetric relationship between two operators.

[3] See, for example, Article 8 of the European Union's *Directive 2002/20/EC of the European Parliament and of the Council of 7 March 2002 on a common regulatory framework for electronic communications networks and services (Framework Directive)*, Official Journal of the European Communities, L 108, April 24, 2002.

[4] See, for instance, Haucap, Justus, and 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").

[5] ITU's working definition reads, in part: "A Next Generation Network (NGN) is a packet-based network able to provide services including Telecommunication Services and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions are independent from underlying transport-related technologies. It offers unrestricted access by users to different service providers. It supports generalized mobility which will allow consistent and ubiquitous provision of services to users. …" See http://www.itu.int/ITU-T/studygroups/com13/ngn2004/working_definition.html..

[6] These wholesale payment arrangements most often follow the *Calling Party's Network Pays (CPNP)* principle, which is described later in this chapter.

[7] There may be a monetary payment; there may be an acceptance of advertising; or the customer may compensate the providers in some other way. In general, each service provider derives some quantifiable benefit from the customer.

[8] In practice, the termination is often implemented by a more-or-less traditional competitive fixed operator on behalf of the VoIP service provider.

[9] Cf. Wikipedia contributors, "Privity of contract," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/w/index.php?title=Privity_of_contract&oldid=90675807 (accessed December 10, 2006).: "The doctrine of privity in contract law provides that a contract cannot confer rights or impose obligations arising under it on any person or agent except the parties to it."

[10] See Marcus, J. Scott, "Voice over IP (VoIP) and Access to Emergency Services: A Comparison between the U.S. and the UK", *IEEE Communications Magazine*, August 2006, available at http://www.comsoc.org/livepubs/ci1/public/2006/aug/cireg.html.

[11] The migration of the Internet Protocol from IPv4 to IPv6 is not expected to change this.

[12] See, for instance, the description of Click2Call at http://www.vonage.com/features.php?feature=click_2_call. "You simply select a number from your contact list or anywhere on your computer, press the Hotkey or click the "Place Call" button - and we dial the number for you, instantly. Your Vonage phone rings first. When you pick it up it, it rings the number you clicked to call."

[13] For *n* providers, (*n* * (*n*-1)) / 2 agreements would be needed.

[14] Peering and transit are discussed later in this chapter, in section 5.4.2.2.1 "Peering and Transit".

[15] See, for instance, INTUG's submission to the European Regulators' Group (ERG), "International mobile roaming", June2005, available at: http://erg.eu.int/doc/publications/consult_wholesale_intl_roaming/wir_intug.htm.

[16] Many economists would view the authoritative sources as being Jean-Jacques Laffont, Patrick Rey and Jean Tirole, "Network Competition: I. Overview and Nondiscriminatory Pricing" (1998a), *RAND Journal of Economics*, 29:1-37; "Network Competition: II. Price Discrimination" (1998b), *RAND Journal of Economics*, 29:38-56; Armstrong, M. "Network Interconnection in Telecommunications." *Economic Journal*, Vol. 108 (1998), pp. 545–564 ; and Jean-Jacques Laffont and Jean Tirole, *Competition in Telecommunications*, MIT Press, (2001). I choose to draw primarily on Laffont and Tirole (2001).

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

[18] Patrick DeGraba, "Bill and Keep at the Central Office As the Efficient Interconnection Regime", U.S. FCC OSP Working Paper 33.

[19] Doh-Shin Jeon, Jean-Jacques Laffont, and Jean Tirole, "On the receiver pays principle", *RAND Journal of Economics* (2004).

[20] The United States is often referred to in the literature as a country that employs *Receiving Party Pays (RPP*), but this rather misses the point. RPP was never used for fixed services. Under RPP, the mobile subscriber pays for air time, whether originating or receiving a call. RPP plans may possibly still be available from some operators in the U.S., but their role in the U.S. mobile marketplace is small and is diminishing over time. Many mobile operators no longer bother to offer RPP subscription plans.

[21] Calls received from foreign countries are typically billed for air time at the same rates as domestic calls.

[22] 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.

[23] 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.

[24] Flat rate plans for fixed service are truly flat rate, whereas the mobile plans are generally constructed as if they were two part tariffs. The usage charges of the mobile plans are usually set to levels that must be seen in the context of the U.S. as unreasonably high (as much as 0.40 USD per Minute of Use). As previously noted, 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.

[25] For example, Vonage offers unlimited calls to or from the U.S. and Canada (and also calls to fixed telephones in the UK, Ireland, France, Italy and Spain) for just $24.99 a month. See www.vonage.com.

[26] In the United States, by means of the Enhanced Service Provider (ESP) exemption; in the UK, by means of a system known as FRIACO.

[27] Mobile operators may, however, include *on-net* calls to mobile phones (calls to their own mobile customers). They do not pay a termination fee for these calls, and thus do not face a high marginal wholesale cost.

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

[29] In the interest of brevity, 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.

[30] 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.

[31] The French ARCEP was obliged to end Bill and Keep arrangements because the European regulatory framework for electronic communications implemented in 2003 effectively prohibited operators from charging a different termination fee for mobile-to-mobile calls than for fixed-to-mobile calls. The mobile-to-mobile Bill and Keep arrangements had caused some minor anomalies, notably including relatively high use of so-called "SIM boxes" (a technological means of converting fixed-to-mobile calls into mobile-to-mobile), but were generally felt to have worked well.

[32] FCC, *Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 11th Report (11th CMRS Competition Report*), September 2006, Table 12, based on *Interactive Global Wireless Matrix 4Q05*, Merrill Lynch, Telecom Services Research.

[33] 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).

[34] 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.

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

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

[37] 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).

[38] As of October 2005. European Commission, *11th Implementation Report* (December 2005). See also Marcus (2004).

[39] ITU, "Report of the meeting of Working Party 2/3 (Geneva, 19-27 June 2006)", COM3–R8–E, September 2006, page 12, paragraph 6.14.

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

[41] 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.

[42] Laffont and Tirole, *Competition in Telecommunications* (2001).

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

[44] 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, however, are generally too high to ignore – they often represent the preponderance of the total cost of the call.

[45] 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.

[46] FCC, *Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 10th Report (11th CMRS Competition Report*), September 2006, Table 12, based on *Interactive Global Wireless Matrix 4Q05*, Merrill Lynch, Telecom Services Research.

[47] As a consumer originates more calls, eventually he or she exceeds the number of minutes in the bucket, incurs exceedingly high charges, and is effectively forced to upgrade.

[48] Private communication, Indian TRAI.

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

[50] This section presents the traditional view. Some recent studies have challenged this view, including Haucap and Dewenter (2006) and Littlechild (2006).

[51] See, for example, Cave et. al.; Littlechild; Crandall and Sidak.

[52] European Commission, *11th Implementation Report*, page 43.

[53] Eurobarometer, *E-Communications Household Survey*, July 2006 (reflecting December 2005 – January 2006 survey data).

[54] FCC, *Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 10th Report (11th CMRS Competition Report*), September 2006, paragraph 5.

[55] Whether the U.S. will achieve European levels of subscriptions per inhabitant is in fact doubtful; however, the U.S. is quite likely to achieve European levels in terms of the more meaningful measure of the fraction of households that have access to at least one mobile phone. As previously noted, Europeans are probably much more likely than Americans to have multiple mobile subscriptions at once.

[56] Private communication, Indian TRAI.

[57] See Olivier Bomsel, Martin Cave, Gilles Le Blanc and Karl-Heinz Neumann, How mobile termination charges shape the dynamics of the telecom sector, 9 July 2003, section 6.4,available at: http://www.med.govt.nz/upload/39184/06-appendix.pdf. The impact is difficult to analyze because many incumbents operate simultaneously in both the fixed and the mobile markets.

[58] Consider, for example, a country like Hungary. Mobile penetration is about 90%, and growing; fixed penetration is about 35%, and declining.

[59] 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.

[60] An Autonomous System, as defined in the IETF's RFC 1930, is "… a connected group of one or more IP prefixes run by one or more network operators which has a SINGLE and CLEARLY DEFINED routing policy." A globally unique AS number must be assigned to every AS. These numbers are currently assigned from a 16 bit field, and thus constitute number from 1 to 65,535.

[61] Large organizations with more than one upstream ISP also require an ASN.

[62] A list of assigned ASNs appears at http://www.iana.org/assignments/as-numbers.

[63] 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.

[64] A very innovative paper by Prof. Lixin Gao of the University of Amherst confirms this structure. See Gao, 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.

[65] 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..

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

[67] 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.

[68] The definitive works on intercarrier compensation in the world of the PSTN are generally considered to be Armstrong, M. "Network Interconnection in Telecommunications." *Economic Journal*, Vol. 108 (1998), pp. 545–564; and 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.

[69] 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.

[70] The absence of a significant transit relationship has in some contexts been used as a definition of a *backbone* ISP.

[71] To the extent that the different interconnections have different bandwidth and delay characteristics, this could have some impact on the total volume of data transferred. A user might not make a second request to a web site, for example, until the first request is complete.

[72] See "A Brief History: The Bell System", at http://www.corp.att.com/history/history3.html. See also Wikipedia contributors, "Kingsbury Commitment," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/w/index.php?title=Kingsbury_Commitment&oldid=60945357 (accessed December 17, 2006).

[73] See 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").

[74] Ofcom has noted this problem in their many public consultations on NGN. Consider, for instance: "The core problem is in the fixed line market. 20 years of pro-competition regulation has led to some improvements for consumers. But years of intrusive regulation have not created the conditions for the sustainable competition necessary for long-term consumer benefit and which, in other countries, has spurred investment in next generation core and access networks.

Ofcom's overall approach, therefore, … has been to create a regulatory framework which seeks to encourage and incentivise sustainable, scale, infrastructure competition at the deepest extent possible. However, some assets in the network are either economically impossible or highly economically inefficient to try to replicate: the so-called enduring bottlenecks- mainly, though not exclusively, in the access part of the network. Without open and truly equivalent access to such assets, sustainable infrastructure based competition would be too risky and too easily frustrated." Available at: http://www.ofcom.org.uk/consult/condocs/telecoms_p2/statement/.

[75] European Commission, *11th Implementation Report*, COM(2006)68, 20 February 2006, Volume II, Annex 2, pages 30-36. Mobile termination fees for all operators have fallen from 0.2054 EUR per minute in July 2001 to 0.1279 EUR per minute in October 2005. Fixed interconnection fees have also fallen under pressure from regulators.

[76] 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.

[77] 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. See also Patrick DeGraba, "Bill and Keep at the Central Office As the Efficient Interconnection Regime", op. cit.

[78] Coase, Ronald H., "The Federal Communications Commission", *Journal of Law and Economics*, Vol. 2, pages 1-40, (1959).

[79] Marcus, "Call Termination Fees: The U.S. in global perspective", op. cit.

[80] Retail pricing arrangements for mobile operators are completely unregulated, but operators and consumers have increasingly chosen flat rate (buckets of minutes) plans.

[81] Strictly speaking, this is true for *reciprocal compensation*, but not for *access charges*. The distinction will not be of interest to most readers. See Marcus (2004), op. cit.

[82] Marcus (2004), op. cit.

[83] Industry consolidation is another noteworthy contributory factor.

[84] Cullen and Devoteam, *Regulatory implications of the introduction of next generation networks and other new developments in electronic communications*, 16 May 2003.

[85] Verbal remarks of Thilo Salmon of SIPgate, at the WIK/BNetzA workshop on NGN and Emerging Markets, 5 December 2005, Koenigswinter, Germany.

[86] Their costs are not zero, however. They typically need a traditional switch, or a media gateway, to terminate calls placed to their subscribers. Alternatively, they may pay a traditional operator for providing this service. It is a cost in either case, but it is a small cost in comparison with that of traditional telephony network.

[87] It does not totally eliminate administrative expense, of course. Operators would typically continue to *monitor* traffic exchange, even if they do not explicitly *charge* for it. Having been in both positions, I would say that charging represents a vastly greater level of administrative complexity and of overhead.

[88] J. Scott Marcus, "Interconnection in an NGN Environment", a background paper commissioned for the ITU New Initiatives Programme workshop on "What rules for IP-enabled Next Generation Networks?" held on 23-24 March 2006 at ITU Headquarters, Geneva. Available at: http://www.itu.int/osg/spu/ngn/documents/Papers/Marcus-060323-Fin-v2.1.pdf . Also available as WIK Discussion Paper 274 (see http://www.wik.org/content_e/diskus/274.htm).

[89] Notably Worldcom/MCI and Worldcom/Sprint.

[90] See OECD, Patrick Swinburne, "Rethinking Universal Service for a Next Generation Network Environment", DSTI/ICCP/TISP(2005)5/FINAL, 18 April 2006.

[91] See Rohlfs, op. cit.

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

[93] The author is not a neutral party in the matter. The author has 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] 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.

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

[97] 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.

[98] 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.

[99] 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.

[100] 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, page 340.

[101] J. Scott Marcus, *Designing Wide Area Networks and Internetworks: A Practical Guide*, Addison Wesley, 1999.

[102] There has been a long and tiresome debate in the literature about the degree to which Internet traffic is *self-similar*. Suffice it to say that traffic patterns are not perfectly random, and that in consequence the risk of an occasionally saturated link is a bit higher than the standard equations would lead one to expect. The network design consequently needs to leave slightly more "headroom" than the basic mathematics would otherwise imply.

[103] Rohlfs notes that many new high technology services encounter difficulty in achieving sufficient penetration to get past an initial adoption hump. A certain number of end-users might take up a product or service based solely on its intrinsic value, but that number is likely to be far less than the number of end-users who 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.

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

[105] See Marcus, "Evolving Core Capabilities of the Internet", *Journal on Telecommunications and High Technology Law*, 2004, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=921903.

[106] 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, page 371.

[107] Ibid., page 374-376.

[108] This is, in a formal sense, a form of market power, but it is not a manifestation of the kind of market dominance that regulators generally seek to control, and it does not necessarily imply a loss of consumer welfare.

[109] Ibid., page 380-381.

[110] "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. Also available in *intermedia* (the journal of the International Institute of Communications), vol. 34, no.3, July/August 2006.

[111] In the Madison River proceeding, the FCC levied a fine for conduct that arguably might have been objectionable, but without identifying any rule that was violated or any way in which Madison River could reasonably have known that they were in violation. In the "Broadband Policy Statement", they established lofty but ill-defined goals without clarifying their meaning or establishing an enforcement mechanism. Collectively, the rulings leave an ambiguous legacy.

[112] 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, op. cit., especially the discussion in section 6, "Micropayments and neutrality".

[113] Ibid.

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

[115] The discussion of Ofcom's proceedings draws heavily on an earlier paper, "Framework for Interconnection of IP-Based Networks – Accounting Systems and Interconnection Regimes in the USA and the UK", a background paper prepared for the German Federal Network Agency's study group on a Framework for Interconnection of IP-Based Networks, 27 March 2006, available at: http://www.bundesnetzagentur.de/media/archive/6201.pdf.

[116] See also the final report of the German BNetzA's project group: *Rahmenbedingungen der Zusammenschaltung IP-basierter Netze*, 15 December 2006, available at http://www.bundesnetzagentur.de/media/archive/8370.pdf. For a developing country perspective on these interconnection issues,

[117] *Further Consultation*, section 1.11.

[118] *First consultation*, Figure 1.

[119] Ibid., section 2.9. Figure 1 of the *First Consultation* shows 120 sites; however, at the December 5, 2005 "NGN and Emerging Markets" workshop, hosted by WIK on behalf of BNetzA, BT CTO Matt Bross referred to 100 metro sites.

[120] *First Consultation*, section 1.13.

[121] *Further consultation*, section 3.10.

[122] It is generally recognized that regulators should refrain from frequent or arbitrary reductions in regulated rates, due to the risk of reducing the incentive for operators to invest in efficiency improvements . Cf. Laffont and Tirole, *Competition in Telecommunications*, 2001.

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

[124] 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."

[125] See http://www.ofcom.org.uk/media/news/2005/06/nr_20050623 and http://www.ofcom.org.uk/consult/condocs/telecoms_p2/statement/main.pdf. See also Ofcom's *Final statements on the Strategic Review of Telecommunications, and undertakings in lieu of a reference under the Enterprise Act 2002* (Strategic Review), 22 September 2005.

[126] BT offered undertakings in response to an implied threat by Ofcom to refer them to UK competition authorities. The undertakings are thus pursuant to competition law, and operate in a parallel and complementary fashion to Ofcom's *ex ante* sector-specific regulation. See http://www.ofcom.org.uk/consult/condocs/sec155/sec155.pdf. BT's commitments appear as Annex A to Ofcom's *Strategic Review*.

[127] An "SMP product or service" is a product or service that BT is obliged to provide as a regulatory remedy because Ofcom has found BT to have Significant Market Power, or SMP.

[128] *Further Consultation*, op. cit., section 1.21.

[129] In an element-based charging (EBC) charging system, termination fees reflect the long range incremental costs of the underlying network elements used in service provision, plus a surcharge that provides a reasonable return on the capital invested.

[130] See section 3.7 of the Indian TRAI's Consultation Paper No: 2/2006, *Consultation Paper on Issues Pertaining to Next Generation Networks (NGN)*, 12 January 2006.

[131] Saudi CITC, *Report on CITC's Review and Analysis of Saudi Telecommunications Company's Revised Reference Interconnection Offer Filed on 21/1/1427h*.

[132] Mobile termination rates are indeed higher. In Europe, the average (peak rate) mobile termination rate for all operators in October 2005 was 0.1296 EUR per minute, while the fixed termination rates for local level, single transit

and double transit were 0.0060 EUR, 0.0094 EUR, and 0.0139 EUR, respectively. Mobile termination rates are thus ten to twenty times higher.

[133] The French experienced considerable take-up of so-called "SIM boxes" that effectively converted calls that would otherwise have been fixed-to-mobile into mobile-to-mobile calls. In doing so, they bypassed CPNP wholesale charges and capitalized on cheaper mobile-to-mobile retail prices.

[134] See Matt Richtel, "The Air is Free, and So Are the Phone Calls That Borrow It", New York Times, 27 November 2006; and Glenn Fishman, "Marrying the Cellphone to Cheap Internet Calling",*New York Times*,14 December 2006.

[135] European Commission, *Proposal for a Regulation of the European Parliament and of the Council on roaming on public mobile networks within the Community and amending Directive 2002/21/EC on a common regulatory framework for electronic communications networks and services*, {SEC(2006) 925} and {SEC(2006) 926}, COM(2006) 382 final, 2006/0133 (COD), 12 July 2006; and *Commission Staff Working Paper: Impact Assessment of Policy Options in Relation to a Commission Proposal for a Regulation of the European Parliament and of the Council on Roaming on Public Mobile Networks within the Community*, {COM(2006) 382 final} {SEC(2006) 926}, SEC(2006) 925, 12 July 2006.

[136] See Ulrich Stumpf, "International Roaming: A Way Forward", presented at IBC's 9th Annual Conference "Communications and EC Competition Law", Brussels, 14-15 October 2004; and Ulrich Stumpf, "Prospects for improving competition in mobile roaming", presented at TPRC, Alexandria, VA, USA, 27-29 October 2001.

[137] European Commission, *11th Implementation Report* (December 2005). See also Marcus (2004).

[138] Data source: Cellular Telecommunications and Internet Association, *Semi-Annual Wireless Industry Survey* (see http://www.wow-com.com/industry/stats/surveys/).

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

[140] 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.

[141] 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").

[142] The same claim cannot be made for *access*, as distinct from *interconnection*.