NGN Interconnection: Charging Principles and Economic Efficiency

Richard Cadman

12th July 2007

NGNuk
3rd Floor, Riverside House
2a Southwark Bridge Road
London
SE1 9HA

Telephone +44(0)207 7834688
Fax +44(0)2077834700
Email peter.ryde@ngnuk.org.uk
Website www.ngnuk.org.uk

This paper “NGN Interconnection: Charging Principles and Economic Efficiency” was written by Richard Cadman (SPC Network) with the support and assistance of the NGNuk membership.

Richard Cadman is Director of Strategy and Policy Consultants Network Ltd (SPC Network), an economics and public policy consultancy active in the electronic communications market. He holds a Master Degree in Competition and Regulation Policy and Bachelors Degree in Politics and Economics. SPC Network’s clients include fixed and mobile operators in Europe and the Middle East

Strategy and Policy Consultants Network Ltd
Chapel House
Booton
Norwich
NR10 4PE
United Kingdom

+44 1603 871162
info@spcnetwork.eu
www.spcnetwork.eu

Abstract[*]

Telecommunications networks are being upgraded from current generation circuit switched technology to Next Generation all-IP networks. These new networks will have lower operating costs and offer opportunities for new services. With some exceptions, voice and messaging services on current generation networks are paid for by the calling party's network, whilst the Internet is largely characterised by Bill and Keep, whereby networks are responsible for their own costs. This paper reviews whether current interconnect charging principles or Bill and Keep will be more likely to promote dynamic and static efficiency gains when applied to voice and messaging services on NGNs. We find that efficient investment, an objective of both EU and UK law, is most likely to be supported by continuing with a system whereby the network of the party most likely to benefit from the transfer of a call or message continues to pay for the call. In this way networks are most likely to recover investments from calling or called parties who gain most. This basic economic principle is equally applicable to NGNs as it is to the current generation of networks.

[*] The author thanks NGNuk members for their support in preparing this paper. All mistakes are my own.

1. Introduction

Traditional electronic communications networks used for voice and related services employ circuit-switched technology and guarantee end-to-end quality of service. By contrast, the Internet employs packet-switched technology which, although less expensive to operate, does so on a "best efforts" basis with no guarantee of quality. To capture the efficiency benefits of the Internet, but also offer the quality benefits of traditional networks, the communications industry is developing Next Generation Networks (NGNs) capable of carrying voice and data to acceptable levels of quality depending on the consumer service.

With some exceptions, voice and messaging services have been offered on a "Calling Party Pays" (CPP) principle, i.e. at both the wholesale and retail levels, the calling party, or his/her network, is responsible for meeting the costs of transmitting the call to the called party. The exceptions will be explored in detail later in this paper, but one obvious exception is 0800 (toll free) numbers where the Receiving Party Pays (RPP) the full cost of transmission from the calling party.

This paper explores whether the current wholesale charging principles should apply in the NGN world or whether an alternative principle would be more appropriate. In particular, I examine which wholesale charging principle is more likely to offer incentives to operators to seek either productive efficiency or dynamic efficiency gains delivering lower prices and/or new services to consumers. The paper is focussed mainly on the UK and draws on UK examples, though the argument may be relevant in other countries in particular within the European Union.

The paper is set out as follows: Section 2 defines various terms used in this paper and describes the voice and messaging services discussed. Section 3 reviews the literature on NGN interconnection and on the economic efficiency of charging principles. Section 4 develops an analytical framework for assessing the efficiency gains likely to arise from maintaining the current approach or adopting new principles. Section 5 analyses the possible economic effects of adopting different charging principles. Section 6 examines the implications of NGNs for the analysis in Section 5. Section 7 explores some practical issues in relation to the implementation of NGN charging principles. Section 8 concludes and sets out policy recommendations.

2. Definitions

This section defines various terms used throughout this paper. In some cases these definitions may differ slightly from those adopted by other organisations. Where this is the case, such differences will be explained a discussed. The definitions are set out in alphabetical order.

Access The "first mile" connection from the customer's premises to the exchange, or equivalent.

Bill and Keep (B&K) The calling party's and receiving party's network are each responsible for meeting their own costs through charges made to their own retail customers. No payment is made or received for the exchange of traffic between interconnected networks. That is, one bills the retail customer and keeps all the revenue, rather than billing the customer and passing some of the revenue on to the relevant interconnected network.

For B&K to be commercially viable there needs to be a rough balance of traffic between the networks. Where traffic is not balanced, or where physical interconnection between two networks is not economic, for example a large distance separates the two, a transit arrangement would be necessary. Under transit there is a very clear customer-provider relationship whereby the transit network gets paid.

The B&K system is widely applied in the Internet world for Internet Peering.

Both Parties Pay (BPP) Where the cost of the call at the retail level is shared between the called and the calling party, we refer to this as BPP. Whilst there may not be a direct link between a wholesale charging structure and that applied at retail level, BPP is most likely to occur when

B&K is used at wholesale level. BPP also describes the retail charging principle in the USA mobile market in which, although commonly referred to as RPP, both parties pay for their portion of the call.

Calling Party The calling party is the person who initiates a call or the sending of a message.

Calling Party's Network Pays (CPNP) The Communications Provider (CP) to whom the calling party subscribes is responsible for meeting the costs of onward transmission of a call or message to the receiving party. These costs are met in the form of an interconnection fee, normally on a per minute basis, which is passed on to the receiving party's network to cover the costs of call termination. If a transit network is involved, the receiving party's network receives payment via the transit network.

Calling Party Pays (CPP) The retail charging principle under which the calling party pays the full costs of the call (message transfer) to the receiving party. This is generally the retail equivalent of CPNP.

Carrier Selection/Pre-Selection (CS/CPS) A services by which a calling party selects a CP other than that which he/she uses for access to transit their calls (messages) to receiving parties. This is also referred to as **Indirect Access**.

Communications Provider (CP) A firm providing a communications service either to wholesale or retail customers, or both.

Origination The delivery of a call (message) from the calling party to the point of interconnection with the termination network.

Receiving Party The person receiving the call or message.

Receiving Party's Network Pays (RPNP) The receiving party's network is responsible for its own costs together with those of the transit operator, if involved, and the calling party's network. This is typically the case in an 0800 call. However, RPNP also applies to indirect access calls where the CPS operator receives a call from the access network and pays for the receipt of that call. In this definition we differ from that adopted by the European Regulators Group (ERG 2007) which defines RPP as "a mixed system where the called and the calling party share the cost of call". The ERG also apply RPP to the retail level only. We argue that RPNP is appropriate to apply at the wholesale level as it underpins the 0800 and CS/CPS services.

Receiving Party Pays The retail charging principle under which the receiving party pays the entire cost of the call (message transfer) from the point of origination.

Termination The delivery of a call (message) from the point of interconnection with the receiving party's network to the receiving party.

Transit A paid for service for the carriage of a call (message) between the originating and the termination network.

This paper considers the application of NGN to current voice and messaging communications products. Each of the call types are briefly defined below together with the charging principle used.

Direct Voice Origination Services: These are voice calls placed by the consumer using his/her access service provider. These calls are generally offered on a CPNP basis. The originating network is responsible for the onward cost of transmission to the terminating network and for termination.

Indirect Voice Origination Services: These are typically CS/CPS calls where the calling party uses a different CP to carry the call to the called party than the CP he/she uses for access. At the wholesale level these calls are generally a combination of RPP and CPP, though are CPP at the retail level. The call can be divided into two components: origination and termination.

- It originates on the access network and is passed to the CS/CPS provider which pays the access provider for the delivery of the call. The CS/CPS provider can be said to receive the call from the access network and so pays on an RPP basis.
- The CS/CPS provider then passes the call on to the terminating CP for delivery to the called party and pays the terminating network for the final delivery. This leg of the call is charged for on a RPNP principle.

Voice Call Termination: Refers to the final delivery of a call from the point of interconnection between the receiving party's and calling party's network and the network termination point. With the exception of 0800 calls, this leg of a call is paid for on a CPNP principle.

0800: These are calls to a non-geographic number in which the call is paid for by the receiver. They are typically used for sales and customer support lines. These calls are paid for on an RPNP principle at the wholesale level and RPP at retail level.

Other Non-Geographic Numbers: These are "Special Service" numbers in the National Numbering Plan generally in the 08XX – 09XX range, excluding 0800. At the retail level these calls are paid for on a CPP basis and where the call is originated on a network other than BT are charged on CPNP at wholesale level. However, where the call originates on BT, BT bills its retail customers and retains a proportion of the call fee passing the remainder to the terminating CP. The amount retained by BT depends on where in the BT network interconnection takes place. If a third party transit network is involved, the terminating CP pays the transit operator. So at a wholesale level these call are charged on a RPNP basis when originating on BT.

Short Messaging Service (SMS): Text messages sent between mobile networks are charged for on a CPNP principle.

As can be seen from the above, both CPNP and RPNP are currently used for traditional voice and messaging services today.

3. Literature Review

Unsurprisingly, there is little research on the charging principles most appropriate to NGN, though there has been some debate on the merits of B&K regardless of the underlying technology.

One of the first papers to study the economics of interconnection of all-IP networks is Yoon (2006). Yoon points out that in a circuit switching world, if Amy calls Bob, but Bob does all the talking, Amy still pays for the call as it is she who requests the circuit to connect to Bob. In an IP world, however, in which networks are packet switched, all the packets containing speech would be generated by Bob[1]. Therefore, the initiator of the call may be different to the sender of the packets. Traditional telecoms networks treat the initiator as the sender and so charge Amy for using the network. Whether such practice is desirable and/or feasible in both a technical and economic sense for IP networks is, according to Yoon, not settled yet.

Yoon describes the Internet as having two distinct groups of subscribers, consumers and websites, who each gain value from more of the other side being present. Internet surfers gain more value if there are more websites of interest and websites gain value if there are more surfers. If a new surfer or website joins the network, there would be a positive benefit to those already connected to the network, but this benefit is not factored into the buying decision of the individual website or surfer. Therefore, in order to achieve the highest level of welfare across both sides of the market, one needs a pricing structure that encourages the highest number of users from both sides, which may be quite different for the two sides and need not necessarily relate to the costs caused by each party.

[1] The extent to which this is true in fact depends on the voice codec used. Not all are capable of detecting silence and so will send packets of silence meaning both parties still cause a cost on the networks,

Based on this analysis, Yoon claims that:

> *…the theory of two-sided markets implies that the retail prices charged to consumers and websites, whether in the form of fixed fees or variable fees, need not reflect the benefits or costs of either side by itself. Rather, the subscription fees or usage fees depend on various factors: elasticities of demand, cross externalities, desire for variety, the pricing practice and the market power of content providers in the Internet, and so on.*

Yoon's overall conclusion is that a more flexible approach than simply charging at cost is required in all-IP network environments since it is extremely hard to calculate costs when multiple services are provided within a common network.

Yoon's two-sided market's analysis is more relevant to the content provider – content consumer communications of the Internet than to voice or message communications where the two sides are often very similar. Although for any given call or message the receiving and calling parties have different functions, in general a called party will on occasions be the calling party and vice versa. Yoon's two-sided analysis supports Wright's analysis below: CPNP means a very different price structure for calling and receiving party – one pays and the other doesn't, and in this way the number of completed calls is maximised (i.e. number of people participating in the network).

DeGraba (2000), in a paper produced for the Federal Communications Commission (FCC), proposed a B&K regime he termed Central Office Bill and Keep (COBAK). Central Office is the American equivalent of a local exchange. In DeGraba's paper he proposes two rules. First, the receiving party's carrier cannot charge an interconnecting carrier to *terminate* a call. Secondly, the calling party's carrier is responsible for the cost of transporting a call to the called party's central office. COBAK is proposed as a default rule for interconnection if carriers cannot agree on alternative terms in commercial negotiations.

The COBAK proposal is premised on three observations. First, that both parties generally benefit from a call, secondly that competition is more effective when carriers recover costs from their own customers and thirdly, that an arbitrage opportunity exists when regulation results in different charges being assessed for the same facility.

The principle current benefit of COBAK claimed by DeGraba is that it "significantly reduces" the terminating monopoly problem. Other benefits are that it will lead to more efficient pricing, and therefore more efficient usage and that it reduces the need for regulatory intervention.

Responding to DeGraba, Wright (2002) sets out two problems. First, that COBAK fails to internalise network externalities between calling parties and secondly its failure to apply Ramsey principles. On the first objection, Wright argues that the calling party receives a direct benefit as a result of the called party being willing to accept the call and that this benefit is likely to be larger than that flowing in the opposite direction. Having the calling party pay for the costs of the called party to receive the call, results in an efficient transfer between the two types of callers. By imposing B&K, this transfer will be eliminated.

The application of Ramsey principles would allocate prices between the called and the calling party based on willingness to pay. DeGraba's premise is that as both parties benefit from a call, both have some willingness to pay. Wright's second objection is that the COBAK proposal does not recognise that the called party may often have a much lower willingness to pay and a better assumption is that the called party may have no willingness to pay in many situations. Wright states that, if his assumption is correct, Ramsey principles dictate that the calling party should bear the costs of termination.

DeGraba responds to Wright's paper (DeGraba 2002) rejecting his criticisms of the COBAK proposal. The first criticism is rejected on the basis that artificially increasing origination charges by having the caller pay for termination may cause inefficient substitution of low cost wireline call by higher cost wireless calls. The second criticism is rejected on the grounds that the COBAK proposal will still impose a higher cost on the calling party even if the benefits of the call are shared equally by both parties.

Employing a somewhat more formal analysis and referring to messages rather than calls, Loder et al (2006) implicitly agree with Wright regarding the willingness to pay of the receiving party. They assume that a receiver has a private value, r, from a message at a per-message cost, c_r. The important point that Loder et al make is that the recipient cannot know her own value r until she has read the message, and therefore incurred the cost. If a message is blocked, the recipient avoids the costs but does not realise the value, if any.

Hermalin and Katz (2004) consider the economic welfare effects of the sender and receiver paying differing amounts for a message, examples of which they give as a telephone call, and SMS message, a data file, a web page or an email. They accept that the cost of some messages may be so small as to be meaningless but others, for example, large data files, such as video films, could have significant cost. They also point out that even when the cost of transmitting a message is small, the opportunity cost to the sender and/or receiver might be substantial, giving the example of taking a telesales call during dinner.

Rather like Loder et al, Hermalin and Katz point out that the message value could be unknown to both the sender and receiver at the time they make their decisions. Under Hermalin and Katz's model, total surplus is maximised if all messages, for which the combined value to the sender and recipient is greater than the costs, including opportunity costs, of transmitting the message, are sent. However, they claim that as the total value may be realised only after the message has been sent and received, this first best outcome is often unobtainable. Therefore, they say, a more realistic welfare standard is the second-best outcome which they term "information-constrained". A message is exchanged if and only if the expected value, conditional upon what the parties know, exceeds the cost.

In a discussion on mobile termination charges, Littlechild (2006) writes:

> *Other things being equal, a charge for receiving calls under RPP may discourage the receipt of some calls and in other circumstances may lead subscribers to turn off their phones or not to join in the first place. This could reduce both calling rates and mobile penetration. On the other hand, if there is no charge for receiving calls under CPP but a higher charge for making calls, this will presumably discourage the making of some calls and could again lead some subscribers not to join…The net effect of CPP and RPP will thus depend, amongst other things, on the prices charged and on the relative levels and elasticities of demand for making and receiving calls.*

4. Analytical Framework

4.1 Legal Framework

Regulation of the European electronic communications sector is underpinned by the New Regulatory Framework (NRF) introduced by the European Commission in 2002. A central objective of the NRF is the promotion of efficient markets delivering choice and quality at low cost to consumers. This requirement is embodied in particular in Article 8.2 of the Framework Directive which states:

> *The national regulatory authorities shall promote competition in the provision of electronic communications networks, electronic communications services and associated facilities and services by inter alia:*
>
> *(a) ensuring that users, including disabled users, derive maximum benefit in terms of choice, price, and quality;*
> *(b) ensuring that there is no distortion or restriction of competition in the electronic communications sector;*
> *(c) encouraging efficient investment in infrastructure, and promoting innovation.*

Each Member State of the EU is required to transpose the NRF into national law. In 2003, the UK Parliament passed the Communications Act 2003 (the Act) which enacted the NRF Directives into UK law. The Article 8 objectives were restated in Section 3(1) as:

It shall be the principal duty of OFCOM [the UK regulator]*, in carrying out their functions-*

(a) to further the interests of citizens in relation to communications matters; and

(b) to further the interests of consumers in relevant markets, where appropriate by promoting competition.

Section 3(4)(d) of the Act requires Ofcom to have regard to "the desirability of encouraging investment and innovation in relevant markets" in carrying out its duties. Unlike the Framework Directive, the Act does not qualify investment by "efficient".

Section 3(5) of the Act defines the interests of consumers in the same language as the Framework Directive: choice, price and quality.

In assessing the interconnection charging principles, we will take particular notice of the objectives in Article 8.2 and Section 3(1) and consider how different charging principles affect consumers and promote efficient investment.

4.2 Economic Framework

The starting point for our analysis is the consumer, for two reasons. First they are the people whose interests have to be furthered by Ofcom and secondly, but just as importantly, they are the people who are served by communications providers. Clearly any change which would cause harm to consumers would be against the objectives of sector policy at both EU and UK level. Our first objective, therefore, is to ensure that any change in charging principles arising from the introduction of NGNs does no harm to consumers.

Our second concern is whether a change in charging principles would promote *efficient* investment, in line with the Framework Directive's objectives. The term efficient in economics has three meanings:

- *Productive Efficiency* refers to producing current goods and services at the lowest possible costs. Productive efficiency gains can be passed on to consumers in the form of lower prices. The most productively efficient firm in a market can earn additional profits by setting its price just below the price of the next most efficient firm, but above its own costs. This additional profit (referred to as Ricardian rents) is the incentive for a firm to become more efficient. However, as rivals also become more efficient the firm's ability to continue earning Ricardian rents is competed away unless it continues to find new ways of being the most efficient provider.

- *Allocative Efficiency* refers to resources being allocated to the production of goods and services which are most valued by consumers and is necessary to send correct signals to existing market players and entrants for investment decisions. This requires products to be sold at a cost based price including a fair return on capital. In the event that competition does not reduce prices to the level of cost there may be a case for regulatory intervention, provided that such intervention does not damage investment incentives.

- *Dynamic Efficiency* refers to the incentive for firms to invest in new products/services and bring innovative products to the market, which consumers value and are prepared to pay a price for which allows the investor to earn an economic return on its investment. A dynamically efficient firm may gain temporary market power through a "first mover advantage", though this may soon be competed away by rivals' innovations. The additional profits (Schumpeterian rents) earned by dynamically efficient firms provides the incentive to invest. As rivals invest in their own new services, the first mover advantage is weakened removing its market power.

In assessing the impact of charging principles in NGN interconnection, we will examine how each principle (CPNP, RPNP and B&K) affects incentives for both productive and dynamic efficiency gains by firms for each of the communications products described.

"Efficient investment" has been described as "the right amount of investment at the right time". Sadly, whether this laudable aim is met cannot be known in advance of an investment being made and the market responding to the new or lower cost services the investment creates by either buying sufficient volume to earn the investor a return, or finding no value in the new service. Our interest is in whether the current charging principles maximise the opportunity for investors to earn a fair return, subject to their services being attractive to the market, or whether an alternative structure would encourage investment without harming consumers.

5. Analysis

In this section, I consider the economic benefits or otherwise of changing from the current mix of charging models to an alternative. I assume that any change from current models can only be justified on the basis of promoting efficient investment in NGN whilst doing no harm to consumers. Any such harm would be measured in decreased demand for services and thus lower consumer welfare.

5.1 Who should pay for an electronic communication?

At least some of the discussion amongst researchers on interconnection economics revolves around the normative question of who *should* pay for the costs of transporting a telephone call or an electronic message from one party to another. There are two ways of answering this question: who causes the most cost and who benefits most. I look first at the question of cost causation before addressing the question of who benefits the most.

During the UK Competition Commission (UKCC) enquiry into Number Portability, the Director General of the then telecoms regulator, Oftel, referred to six principles which are relevant to determining how the burden of cost recovery should fall upon communications providers and their customers (UKCC 1995, p128-129). The first of these principles was cost causation: whose actions caused the costs to be incurred at the margin.

Referring to Yoon's example, (Yoon 2006), in a circuit switched network Amy, the caller, causes the circuit to be set up through to Bob, the receiver, even if Bob does all the talking. Therefore Amy can be said to be causing the cost on both her and Bob's network. However, in a packet switched network it may be possible continuously to assess which party causes costs by monitoring the flow of packets. If Amy remains silent then, dependent on the voice codec used, she causes far fewer packets to be sent through the network than Bob. On the fairly safe assumption that traffic is the most significant driver of network costs (over the long run), Amy causes far less cost than Bob despite initiating the call. Similarly, if Amy contacts a website, rather than Bob, and requests a page or file to be downloaded, it is the website that could be seen as causing the cost by sending large quantities of data across the network, even though it was Amy who initiated the download.

However, it is also arguable that, despite not talking, Amy still causes the cost of the entire call by initiating it in the first place. She may well know in advance that Bob will dominate the conversation but still wishes to call him to hear what he has to say. Her choice to call Bob is therefore the major factor that causes costs to be incurred by both networks.

Yoon considers that if cost causation is continuously assessed it would be feasible to charge the sender for the transmission of the packets rather than initiator. In our view, at least with regard to voice calls, cost causation simply based on who sends the most packets is a misleading way to address the question. Of more significance is who benefits from the transmission of the data, and as I discuss below, who expects to benefit at the time a call is initiated.

Another of Oftel's six principles was the distribution of benefits. Although the principles were put forward in the context of number portability, in essence this principle suggests that the party who benefits most from the communication should meet most of the costs. Underlying DeGraba's argument for a variant of B&K is that benefits are broadly equal between the parties to a call and therefore the costs should be shared. I argue below that such an assumption is incorrect and that the distribution of benefits between parties may be entirely in favour of either the initiator, the receiver or sufficiently in favour of the receiver that he is willing to subsidise some of the calling party's costs. Provided that the

called party is able to signal to the caller the level of his willingness to pay for calls, revenue will be maximised to recover the costs associated with investment in NGN.

The primary focus of this paper is existing voice and SMS services running on NGNs. I first discuss voice calls and then SMS text messages.

A voice call can be considered to require two actions to be completed: first Amy must decide to initiate a call and Bob must decide to accept it (or have a machine answer on his[2] behalf). Once the call is accepted by Bob either party can decide to finish the call at any time. Completed calls are the relevant unit of analysis in this paper as they generate revenues for the network operators and, if the value of a call to both parties is greater than the cost, they deliver welfare to consumers. Following on from the Article 8.2 objectives, we are interested in creating economic conditions which will encourage investment. This will happen if firms have an expectation that they will receive additional revenues from their investments and/or the same revenues at lower costs. As it is the completion of a call, i.e. the decision by Bob to accept the call, that generates revenue, we are primarily interested in completed calls.

I therefore define the relevant market as being that for completed voice calls which consists of two parts: origination and termination, except when two networks are connected via a third transit network. Once the call is completed, this paper makes no further assumptions about the content of the call.

In the analysis below, the price (P) of a call is taken the be expected price of the entire call, regardless of duration.

Let us assume our two consumers Amy (the initiator) and Bob (the recipient) are connected to networks α and β respectively, and that the two networks are directly interconnected. Each consumer has a willingness to pay, or utility (U_A and U_B respectively) for each call. The level of utility is based on the value each will receive from the call and, for simplicity is not constrained by income. We scale the utility of each party between 0 and 1. Each consumer is also faced with a cost comprising two parts: the price charged by the network for making or receiving call as appropriate and the opportunity cost of spending his/her time on the call rather than on some other activity. It is likely that the opportunity cost is substantially greater than the financial cost.

We are interested in maximising completed calls, i.e. calls made by Amy and accepted by Bob. As Amy is the initiator, for her to make a call we can expect Cost < $U_A \leq 1$. Provided that the cost of the call is less than her expected utility she will initiate the call to Bob. To understand Bob's expected utility at the time he receives the call we can assess three scenarios:

i) Bob has the same level of knowledge as Amy (e.g. through caller display and prior knowledge of call, or pre-arranged call, etc). In this case, A and B are in exactly the same situation and have the same expected utility function. Again, we are only interested in completed calls, so the relevant utility curve for B is: Cost < $U_B \leq 1$.
ii) Bob has less knowledge than Amy (e.g. an unrecognised or hidden caller ID). Again we are only interested in completed calls, but this time Bob cannot rule out the call being less value than the cost, i.e. utility curve is $0 \leq U_B \leq 1$.
iii) Bob has a greater degree of knowledge than Amy (e.g. Bob wants to receive Amy's call to sell a service). Given that not all calls result in a successful sale the utility curve will be the same as scenario (i): Cost < $U_B \leq 1$.

The average utility for each party ($\bar{U}_A$ and $\bar{U}_B$) depends on the relative frequency of the three scenarios. Let us first consider the case when scenario (ii) is the most frequent such that $\bar{U}_A > \bar{U}_B$.

In a CPNP environment, with CPP at the retail level, network β will pass the costs of termination back to network α who will recover both its own costs (C_α) and the termination charge raised by β (C_β) from A through its retail price. I make no assumption here about the price setting ability of α, nor about its objectives (profit or revenue maximisation). Therefore I assume that the retail price charged to A is

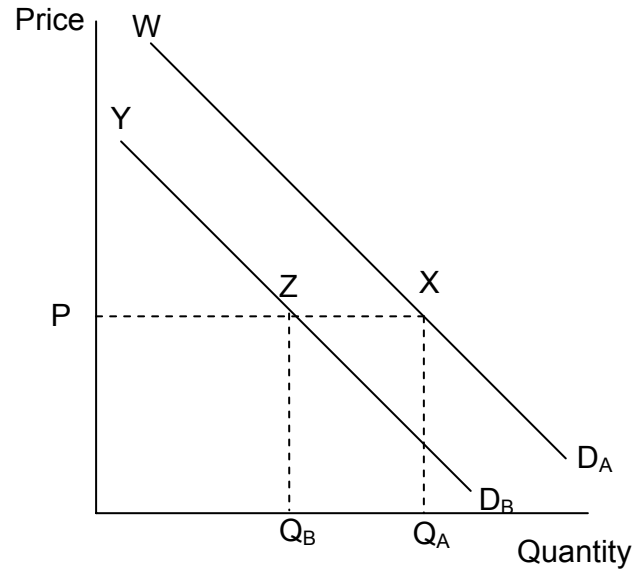
[2] Throughout this discussion I follow Yoon and refer to the caller in the feminine and the receiver in the masculine.

simply $C_\alpha + C_\beta$. As all the costs are met by A, the quantity of completed calls is set by the demand of all initiators - Q_A in Figure 1 – where D_A is the demand curve for call initiators and D_B is the demand curve for call receivers where $U_A > U_B$. Consumer welfare is depicted by the triangle PWX and total revenue is the product of price and the quantity demanded (PQ_A).

Let us now assume that the charging mechanism was changed to an RPNP environment whereby network α passes its costs to network β who must then recover those costs from Bob (and all other receivers). The costs faced by Bob are the same as Amy, thus its price (P) is again $C_\alpha + C_\beta$. However, as $\bar{U}_A > \bar{U}_B$ then Bob will demand a smaller quantity at the same price and so the demand curve for received calls (D_B) is further to the left. The consumer welfare obtained under a RPNP regime for normal call is depicted by the triangle PYZ.

It can easily be seen by the smaller size of triangle PYZ that consumer welfare is reduced as is total quantity demanded on the basis that $P_AQ_A < P_BQ_B$. Thus for "normal" calls, it is obvious that a CPNP environment is better for both consumers and networks.

Figure 1



In scenario (iii) receiving parties wish to generate inbound calls for the purposes of, say, sales enquiries, bookings or orders. In these cases the utility to the receiver may be higher on average than to the caller. We would therefore have a situation where $\bar{U}_B > \bar{U}_A$. The demand curves in Figure 1 would therefore be reversed such that D_B is further to the right than D_A and welfare and revenues are maximised in an RPNP environment.

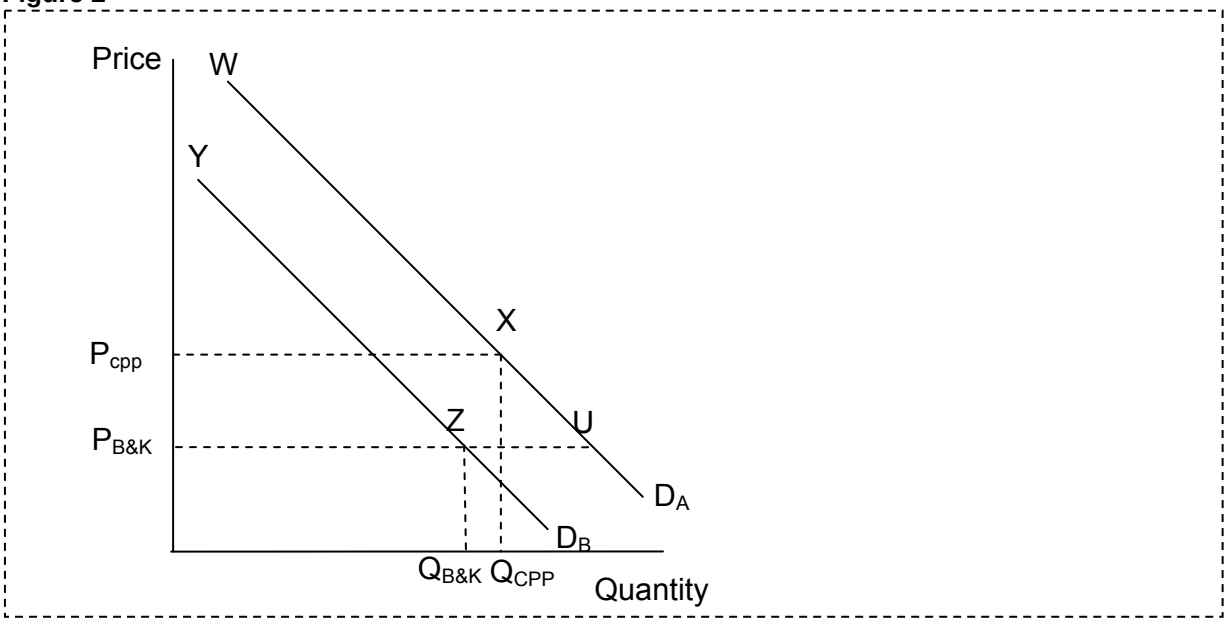
Some researchers are interested in promoting a B&K environment where each network is responsible for recovering its own costs from its own consumers. This type of environment exits today with Internet peering, where two networks of broadly equal size do not charge each other for traffic which flows between them. It also exists in the US wireless market where the receiving party pays for the mobile termination charges.

As the costs are shared in a B&K environment, we can assume that $P_A = C_\alpha$ and will thus be lower than in an CPNP model. We can also assume that $P_B = C_\beta$ whereas $P_B = 0$ in a CPNP environment. Applying this structure to geographic calls, the lower origination price will stimulate more calls but, as $\bar{U}_A > \bar{U}_B$ and B now has to pay to receive calls, this is likely to suppress demand for completed calls. The introduction of B&K in comparison to CPNP is shown graphically in Figure 2.

As with Figure 1, the demand curve for originators is D_A. Under CPP price is set at P_{CPP} and the quantity demanded is Q_{CPP}. If pricing is now changed to a B&K environment, then receivers' sets the demand for completed calls, D_B. Assuming that costs are equal for the originating and terminating network then $P_{B\&K}$ will be half P_{CPP}, so demand will no be $Q_{B\&K}$, i.e. where price intersects D_B.

Figure 2 shows that the total number of completed calls would be less under B&K than CPNP, reducing consumer surplus and industry revenues which may lead to less investment. However, this is a function of the illustration. Consumer welfare under B&K is represented by the triangle $P_{B\&K}$ Z Y and consumer welfare under CPP is P_{CPP} X W. To determine the total welfare and industry revenue effect would require knowledge of: the price elasticity of demand for origination and termination by A and B respectively, the difference in $\bar{U}$, and the price difference between CPNP and B&K. The less the difference in $\bar{U}$, the less the price elasticity of demand and closer P_α is to P_β, the more B&K is likely to generate more revenue than either CPNP or RPNP. The calculation required is shown formally in Annex A.

Figure 2



The above analysis shows that as the expected utility of each party relative to price converges, so the more efficient it would be to introduce a B&K system at wholesale level which would probably be reflected in retail pricing. The practical problem lies in the level of information available to the receiving party on the origin of the message and therefore his expected value. Under Article 8 of the Data Protection Directive[3], service providers must offer calling parties the ability to hide calling line identification information free of charge. The called party must also have the ability to hide such information, again free of charge. Operators can therefore never be certain that the called party will have the necessary information required to accept or reject a call or message.

Figure 2 also illustrates that under B&K some consumer welfare would not be realised. A $P_{B\&K}$ originators are willing to make more calls than receivers are willing to receive. Initiators therefore lose welfare from calls not being accepted, represented by the quadrangle UWYZ.

We recognise that there are a number of variations to the general theme set out above. For example, Amy may call Bob on his landline and hear a message informing her that Bob is out and to call him on his mobile. As Bob has given this message, we can assume that he expects to receive some utility from calls he receive whilst away from his normal location. In these circumstances it might be arguable

[3] DIRECTIVE 2002/58/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)

that Bob should pay at least some of the termination charge as he is expecting some value. The counter to such an argument would be that if Amy has made the decision to call Bob on his mobile then the condition Cost < $U_A \leq 1$ has been satisfied. Bob's utility function would still be dependent on his level of knowledge when the call is received still applies and so it is still economically efficient for Amy to pay for the call. This and other variations are not explicitly explored in this paper.

I now turn briefly to SMS text messages. Following the principle of cost causation, the sending party imposes the costs on both her own and the receiver's network by initiating the transmission of the message. The sending of a return message is a separate communication so if Bob replies to a text sent by Amy, he is the one now imposing costs. Unlike with voice calls, costs cannot be assessed constantly through the communication.

Turning the principle of distribution of benefits, a fundamental difference between voice calls and SMS messages is that the receiver has no option but to receive the SMS. If Amy sends Bob a text, he cannot reject the message or perform the SMS equivalent of hanging-up part way through receipt if the message is of no value to him. As with a voice call, we can assume that Amy will derive some utility from sending a text, otherwise she would not invest her time in preparing the message, and that Bob's utility is unknown when the message is received. So it is likely that the condition $\bar{U}_A > \bar{U}_B$ will hold true and therefore the initiator should pay for the cost of both origination and termination.

If the charging principle for SMS was changed to RPNP then Amy would impose a cost on Bob which Bob would have no option to reject. As the cost to Amy would be reduced to her opportunity cost only, she may be expected to send more messages to Bob than under CPP, thus creating a cost burden on Bob from which he may gain no utility. It would therefore seem inappropriate to change the charging principle applied to SMS away from CPNP/CPP.

5.2 Network Externality and Termination Subsidy

Externalities are defined as consequences for welfare not fully accounted for in the price of a good. In electronic communications there are two forms of externality. The first is the network effect of subscribers: as each additional subscriber joins a network so the value of the network increases for all existing subscribers. However, this increase in value is not captured in the price paid by subscribers. The second form of externality is call externalities. The initiator of a call derives value from the call being answered by the called party. If the price to the caller was set only at the direct cost of call origination then this externality would not be captured (internalised).

We have seen above that the benefits of a call are likely to be asymmetrically distributed between the called and calling party. In most cases we can assume that *at the time the call is made and received* the calling party expects to benefit and the called party may or may not expect to benefit dependent on his knowledge about the inbound call. In the worst case, the receiver will be sufficiently pessimistic about his level of utility that he will keep his phone switched off and only use it for outbound calls. There is some evidence for such behaviour. Samarajiva and Melody (2000, p4) report that *"Subscribers in RPP countries are much more likely to turn their phone off, or refuse to answer calls, in order to avoid paying for them."* They also show a near trebling of mobile subscribers in Mexico after it converted from RPP to CPP.

However, Littlechild (2006) argues that there is no empirical evidence to support the claim that phones will be kept switched off if the receiver has to pay for calls, suggesting that all such evidence is either anecdotal or out of date and claims that the lack of quantification of such behaviour is an "unsatisfactory basis for responsible policy making".

The examination of the benefits of calls discussed in 5.1 and the model developed in Annex A suggests that the difference in average utility between the calling and called parties, and the difference in price each pay will be the determining factors.

I now go on to consider the call-by-call externality. We can describe the called party as subject to two kinds of error when making his decision about whether to accept a call. Either he might reject a call from which he would gain utility (type I error) and he accepts a call from which he has zero or negative

utility (type II error). Type II errors may be less serious from an economic perspective as the recipient can cut short the call if he quickly ascertains there is no benefit to him in continuing with the call. For example, he could tell the double-glazing salesperson who has interrupted his meal that he is happy with his windows as they are. Arguably the caller would also gain from a wasted sales call being cut short as she can call a potentially more responsive prospective customer.

Rejecting calls from which both parties would gain utility is more of a problem. In this case neither the caller's nor the receiver's utility is realised and so there is a consumer welfare loss. However, the value of the call to the calling party may be sufficiently high that she is willing to subsidise the called party to receive the call. This is the positive externality enjoyed by the calling party of the called party being willing to accept the call and arises from the basic condition that $\tilde{U}_A > \tilde{U}_B$.

Wright (2002) says that the externality can be funded through the caller paying the termination charge of the receiver and that this could be an efficient transfer between the two parties. Restricting the analysis to "normal" phone calls, i.e. non-0800 calls, type I errors can be avoided by the calling party subsidising the called party for the cost of termination. If we extend the analysis so that it is more general, then the party who expects to receive the most benefit at the time the call is placed (message sent) can efficiently subsidise the other party's direct costs. Therefore, it is efficient for a firm wishing to receive calls to offer an 0800 number and pay the origination charge of the caller.

5.3 SMP in Termination

One of the principle problems associated with CPNP is that it creates a termination monopoly. In brief, this problem exists because the calling party has no control over which network the called party is connected to and therefore, under a CPNP arrangement, the caller must pay whatever charge the terminating network sets for termination. This may result in call originators paying an excessively high price, though this is disputed by, in particular, the mobile networks[4]. To get over this problem regulators have generally stepped in to force termination charges down to cost based prices. Such an action requires regulatory intervention which is generally regarded as costly and second best to a properly functioning market, not least because it is difficult for the regulator fairly to calculate the cost of termination.

DeGraba (2000) points out that this termination problem presents regulators with the unattractive choice of allowing operators to exercise their terminating market power, which could raise retail prices and reduce network usage, or regulating the terminating access rates of all carriers.

B&K is often put forward as a means of overcoming this termination monopoly problem. Advocates of B&K claim that, in the presence of competition, prices for receiving a call or message would be competed away to cost. Communications providers will compete for both origination and termination traffic and consumers will be sensitive to both prices. Some advocates claim that termination will be offered free of charge, subsidised through higher prices for origination, and others that a bundle of minutes will be included with monthly rental which will be sufficiently large to make users indifferent as to whether they are used for inbound or outbound calling. Assuming this will be the case, then regulators could step back from their involvement in setting call termination prices and so reduce the regulatory burden on operators. This would undoubtedly be a benefit to consumers and communications providers as the cost of regulation could be reduced. However, given that payment for voice call termination would raise the costs of receiving a call from the current level of zero and so result in some calls not being accepted, would the benefit of loss of SMP in termination be sufficient to off-set the potential welfare loss from call not being accepted?

Again referring to our illustration in Section 5.1 and the formal explanation in Annex A, we can see that the burden of regulation is justified if a receiver's utility is relatively low, price elasticity is high and the price differential is large.

However, B&K follows the same principle as Internet peering, that is for it to be commercially attractive to both parties the cost ratio, i.e. the relationship in the costs the two networks cause on each other,

[4] For a full discussion, see UKCC 2003

needs to be close to one. An Internet Backbone Provider (IBP) will not enter into a peering agreement unless it is satisfied with the cost, as well as the technical and operational characteristics of the prospective peer network (Laffont et al 2001). If costs are not in balance, then the IBP may refuse an peering agreement and require transit charges.

The following example assumes that distance is the primary driver of cost, but the analysis is equally valid for any other significant difference in costs between the two networks. Suppose that network α is geographically limited such that the average distance a call or message travels from the calling party to the Point of Interconnection is 10 kilometres whilst on network β the average distance is 100 kilometres. For each interconnected call placed by network α costs are caused in the ratio 1:10, assuming that distance is the primary driver of cost. As calls passed from β to α travel the same average distance on each network, the same ratio applies. Network β incurs ten times the costs of network α for any call or message regardless of the direction of the message.

Since the two networks compete at a retail level, a standard B&K arrangement is clearly inefficient in this context as either network β will have to charge below price to be competitive with network α or it will have to sell at an uncompetitive price and so lose market share.

DeGraba's COBAK solution is supposed to get around this problem by requiring the calling party's network to meet the costs of transit all the way to the called party's Central Office (local exchange). Whilst the proposal may remove SMP in termination, it does not remove the problem altogether as it simply shift the SMP to transit, in the absence of perfect competition.

When Amy wants to call Bob she has no control over the network to which Bob is connected. Under COBAK, network β charges transit from the Point of Interconnection (PoI) to the local exchange. In a two network market, where network β is the only other network, it remains in a monopoly position to transit the call the Bob's local exchange. Network β's SMP has simply shifted from the termination leg (from the local exchange to Bob's premises) to transit (from the PoI to local exchange).

There are two potential regulatory responses to this problem. One response is for the regulator to estimate transit costs and impose a cost oriented transit price on network β. This has no advantage over the regulator estimating the cost of termination. The alternative is for the regulator to determine that B&K, even COBAK, is only available to networks with a minimum number of PoI's. This is an even more unpalatable decision for the regulator as it is then effectively determining the design of a network and is also left with determining a regulated price for termination for all networks which do not have the requisite number of PoIs.

The problem of SMP in part of the network does not completely disappear even if the transit market is effectively competitive. Suppose there is now a third network, γ, which offers transit services in competition with network β. Unless network γ perfectly matches network β, at some point it will need to hand over traffic for onward transit to the local exchange leaving some degree of bottleneck which will need to be regulated to prevent abuse of a monopoly position. The reality in the UK is that even the largest alternative network to BT, Cable & Wireless, does not perfectly match the BT network and so would have to hand some traffic on to the BT network.

This concern may be ameliorated in an NGN environment. In BT's 21CN, there will be 20 – 30 PoIs which will be the deepest level of the network for interconnection. If the Central Office is equivalent to the PoI, rather than the local exchange, then either the alternative CP would be connected to all PoIs or it could buy transit from BT or an alternative carrier in a market where there is no SMP. This model would see the terminating operator responsible for the cost of termination from the PoI to the called party[5].

However, if the receiving party's network remained responsible only for the cost from the local exchange to the called party, then a problem of SMP would arise between the PoI and the receiving party's exchange.

[5] The problem of SMP is not dependent on charging principles. It will exist under all charging models considered.

5.4 The "Hot Potato" Problem

An argument often used against B&K is the "hot potato" problem which arises because communications providers have an incentive to hand over traffic to another network for termination as close to the point of origin as possible, thereby reducing their own costs and maximising the costs of the terminating network. If the terminating network is not able to recover these costs then, it is claimed, the terminating network will under invest (Gilert + Tobin and CRA International 2006).

The ERG suggests that the problem could be overcome by requiring operators to have a reasonable minimum number of interconnection points for B&K to be applicable to that operator. As discussed above. this would involve the regulator in determining the topology of points of interconnection. The ERG then points out that if operators had to increase their network size to be B&K partners for other networks, the investment involved could be inefficient if infrastructures are unnecessarily duplicated (ERG 2007).

DeGraba (2000) however neatly removes the hot potato problem by making the originating network responsible for the costs of transport all the way to the terminating network's central office, including the cost of transit if involved. This proposal maintains the incentive for the originating network to build an efficiently sized network and also maintains the customer-supplier relationship between the originating and terminating network and thus the incentive for investment as costs can be recovered from the originating network.

5.5 SPAM and SPIT

In 5.1 we point out that some calls or received messages result in negative utility for the called party, such as when the receiver is interrupted during some more valuable activity to receive a "junk" message (SPAM) or call (SPIT). Under the principle of doing no harm to consumers, a change from the existing wholesale billing regime which reduced the disincentives to sending SPAM and SPIT would clearly be a wrong.

We can presume that at least some SPAM and SPIT will have a utility for the receiver. For example, the double glazing salesperson must occasionally find a willing buyer or presumably the company would not find this a sensible way to market its products. Similarly a firm which uses large postal mailshots must also derive some value from this activity. However, across all recipients of messages we can assume that average utility ($\bar{U}$) is low. In the presence of sufficient information to the recipient, a large number of the calls would not be completed. However, with insufficient information on which to reject calls, a number of type II errors will occur from which neither party will benefit.

Further, in a B&K regime, the price to the caller will decrease and so, given a negative price elasticity, we can presume that demand for such calls will increase. Receivers are therefore likely to suffer from an increase in unsolicited calls from which consumer welfare is negative, damaging consumers' interests. In response to this problem, Loder et al (2005) set out a number of solutions from law, technology and regulation and propose an "Attention Bond Mechanism" (ABM) designed to signal preferences whilst screening low value communication. Perhaps simply maintaining the cost to the caller would be simpler.

6. Implications for NGN Interconnection

The above analysis has not specifically referred to NGN interconnection and could apply equally in the current generation network world as for NGN. So the question arises as to whether there is any unique quality of NGN that would change this analysis.

There are two characteristics of NGNs which are relevant in such a discussion. First, NGNs are packet based and the packet flow in each direction may not be equal. By contrast, circuit switched, current generation networks open a circuit between the calling parties and so the effective traffic flow in each direction is the same. Secondly, NGN architecture consists of a common transport layer above which sit different services (e.g. voice, data, multimedia, etc.) each of which may have different quality of service parameters. Each service may use different elements of the transport layer which may or may

not be shared with other services. I explore the potential implications of these two characteristics below.

6.1 Packet Based

The packet based network means that the flow of packets between each party can be counted. In a voice call it could be assumed that the person who speaks most generates a larger volume of data and therefore causes the most cost to the two networks. Voice Quality of Service (QoS) requires that packets containing voice have high priority over the network and are not subject to delay or jitter, and therefore tend to impose a higher cost than non-real time services.

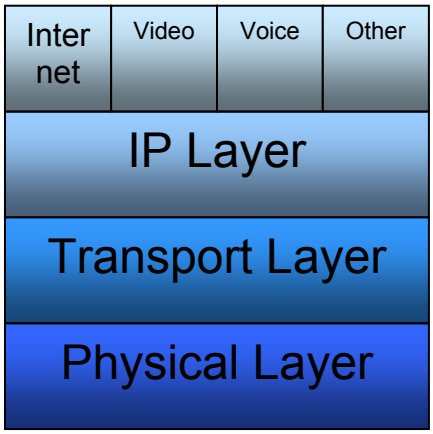
However, to assume that the party who creates most packets gets the most utility from the call would be naïve. It is quite feasible that the called party asks short questions to which the calling party responds with long answers. This does not mean that the calling party does not benefit most from the call, or even receives the most benefit. Imagine, for example, a market research call in which the researcher asks short questions but elicits long answers. Alternatively, imagine a call which combines voice and data transmission in which the initiator requests a large file from the receiver. Again the receiver may cause the most cost, by sending the file, but it is the initiator who benefits.

Given the high cost burden of voice QoS in conjunction with the uneven distributions of costs due to the nature of a packet based network, the relationship between cost causation and distribution of benefits is probably less clear cut in a NGN world than in current generation networks.

6.2 NGN Architecture

The architecture of an NGN consists of a number of common layers, from the physical layer (cables, etc) to the IP layer which support end-user services (Figure 3). As Yoon pointed out, interconnection can happen at either the transport or the IP layer or retailers could buy wholesale services. Each service may use network elements from lower levels in different combinations. Thus there may be some elements used which are common to two or more services and some which are unique to a single service.

Figure 3



It may also be that usage of network resources may be counted in different ways depending on the service provided. For example, voice services may be best counted in minutes on the softswitch[6] whilst downloading of video file may be best counted in megabytes of network traffic. Furthermore, different services may have differentiated QoS requirements at the service layer whilst using common network elements at lower layers.

The question that arise from this understanding of the architecture of NGNs is whether network elements which are used in the same way, should be charged for using the same principle? Should the costs of transport layer, for example, be always recovered on a B&K basis regardless of what service is being carrier over it, whilst the costs of elements unique to voice are recovered from the calling party's network?

Our response to such a question is again phrased in the same framework as above. Would such a scheme harm consumers and, if not, could it benefit consumers by lowering costs or stimulating investment in innovative services? To answer this we have again to refer to which party to a call

[6] A softswitch is a central device in a telephone network which connects calls from one phone line to another, entirely by means of software running on a computer system. This work was formerly carried out by hardware, with physical switchboards to route the calls (Wikipedia).

expects to gain the most utility and will therefore have a greater willingness to pay, at the time the call is placed (or message sent).

At the wholesale level, the primary beneficiary will be the CP which has the primary retail beneficiary as the client. Who this is depends on the call. For "normal" call it is most likely to be the initiator, whilst for inbound sales and marketing calls it may well be the receiver. This fundamental assessment of benefits seems to us to exist independently of the technology used to transfer the call or message and so is equally applicable to next generation as to current generation networks.

7. Practical Issues

Until now I have concentrated on the economic issues related to interconnection charges on NGNs, which point to continuing with the current mix of charging principles to ensure that the party, and therefore the party's network, which most values the communication should pay for it. I now turn to some practical issues. These issues were identified by a survey of NGNuk members in early 2007.

NGNuk established fifteen principles of interconnect charging. Below I have grouped these principles according to three economic objectives: doing no harm to consumers; promoting productive efficiency; and promoting dynamic efficiency. I then apply them to each of three types of communication: voice/SMS origination; 0800; and voice termination. In this application I particularly focus on practical issues.

Figure 4 shows the fifteen principles grouped according to the three economic objectives.

Figure 4: Objectives and Principles of Interconnection

Do no harm	• Protect/enhance user experience • Minimum user/operator disruption • Be cost effective to implement • Enable customer protection/security • Recovery of efficiently incurred costs for SMP operators • Promote interoperability • Charging model coexistence • Enable end-to-end services to agreed quality levels
Productive Efficiency	• Efficient investment in NGNs • Efficient routing
Dynamic Efficiency	• Promotes retail innovation and investment • Promotes wholesale innovation and competition • Efficient investment in NGNs • Promotes efficient and timely migration to NGNs • Voluntary adoption of charging model

7.1 Voice/SMS Origination

Consumers of voice and SMS message origination are used to paying for calls and the sending of text messages. The maintenance of a CPP system, at retail level, would therefore introduce no new harm to consumers and so the user experience would be protected.

It follows from the above that there would be minimum end-user disruption by maintaining the CPP principle for voice/SMS origination. Operator disruption would be minimised by maintaining RPNP for indirect access calls with CPNP for direct access calls. A change to any other regime would cause disruption with attendant costs and confusion for both consumers and CPs.

It also follows that, as no change is required, maintenance of the current mix of payment principles would be cost effective to implement as no additional costs are incurred by Communications Providers (CPs).

The cost of sending SPAM/SPIT remains unchanged and continues to ensure that the receiver of such communications does not pay for them, again protecting consumer experience.

The economic analysis above demonstrates that CPs are able to recover efficiently incurred costs from the user who most value the service and that in the case of voice/SMS origination it is likely to be the initiator of the communication who places most value on it.

The promotion of interoperability may not be affected by the charging model but rather by electronic communications policy promoting any-to-any communications. Whichever charging principle is employed, some regulatory requirement to ensure any-to-any communications will remain essential.

As demonstrated above, all charging models co-exist today and are applied to parties and networks which gain the most benefit. As, on average, we expect calling parties to benefit most from a communication then the continuation of CPP to voice/SMS origination remains appropriate, at the retail level. The current mix of CPNP and RPNP at the wholesale level would continue to support CPP.

The calling party's network will want to ensure end-to-end quality and, as it is the customer of the terminating network, it has the bargaining power to require equivalent quality on the terminating segment of the communication to that on the originating segment.

Continuing with a CPP principle at the retail level for origination allows the originator's CP to recover costs from the party which most values the communication and therefore has a higher willingness to pay. This will maximise the opportunity for recovering investments and so do the most to promote investment.

CPNP maintains the current incentives for operators to keep traffic on their own networks until it becomes inefficient to do so and so promotes far-end handover, where the call is handed on to the terminating network as near to the receiving party as possible. The incentive arises from the calling party's network desire to maximise its own revenues and minimise outpayments to other operators. To minimise costs whilst traffic is on its network, both operators have a strong incentive to route traffic as efficiently as possible.

All five of the principles linked to dynamic efficiency gains are supported by implementing a charging mechanism that allows firms to recover costs from those that value the service most.

7.2 0800

The argument for maintaining the RPP/RPNP charging principle for 0800 appears self evident. The purpose of 0800 is to allow the receiver to pay for the cost of the call and to signal to callers that it is willing to do so. To mandate an alternative charging principle for 0800 would remove a valuable product option.

Many of the practical comments above apply equally to 0800, though self-evidently in reverse.

7.3 Voice Termination

Other than for 0800 calls the receiving party currently does not pay for termination in European countries. Changing the retail charging principle to either RPP would result in consumers paying for a

service they may not value and so would be detrimental to consumers on average. Such a change would cause user and operator disruption and it is unlikely that it would be cost effective to implement.

If the cost of calling is reduced, then the amount of SPAM/SPIT is likely to increase and so consumers would lose some of the protection they currently receive from the costs all falling on the calling party. If consumers were required to pay to receive SPAM/SPIT this would damage the interests of consumers.

In Section 5.2 above on network externalities, I discussed how the calling party subsidising the termination costs of the called party is an efficient transfer as the calling party benefits from her call being accepted by the called party. The reverse is the case in relation to 0800 calls. As all operators have SMP in termination then such an efficient transfer allows the SMP operator to recover the efficiently incurred costs of termination.

Both the called and the calling party have an interest in end-to-end services being provided to an agreed and acceptable level of quality. Provided that competition exists, the relevant networks therefore have an interest in ensuring such quality to capture maximum market share.

8. Conclusion and Recommendations

Efficient and timely investment in NGN will happen if firms making the investments expect to generate a return on their investment either by being more efficient than their competitors or by being able to deliver new and innovative services which consumer are willing to pay for. This requires that CPs are able to recover costs from the party to the call or message transfer which gains most value from, and so has a greater willingness to pay for, the call or message. We make the assumption that the initiator of a call (or the sender of a message) always expects to receive some positive utility from the call, if it is completed (accepted) by the called party. However, at the time the call or message is received the called party has imperfect knowledge of his expected utility, which may be negative if the opportunity cost is too high. This is not always the case, though, as inbound call centres may wish to generate sales enquiries.

A flexible wholesale payment system, which aligns with the product being sold to the end consumer, has the best chance of allowing CPs to recover investments from those who most value the results of the investment. Prescribing any one wholesale payment system may therefore damage the interests of the industry and of consumers. Imagine, for example, if CPNP was prescribed as the only payment principle for NGN interconnection. Two products of significant value to consumers would become unviable: 0800 and CS/CPS.

The economic fundamentals of who benefits from the exchange of a voice call or message seem to us to be unaffected by the introduction of NGNs:

i) On average, calling parties will always expect some utility from a call which, with some exceptions, will be more than the utility expected by the receiving party at the time the call is placed.
ii) The call-by-call externality, i.e. the value to the calling party over and above the cost of the call, is best internalised by the initiator subsidising the called party for the direct costs of receiving the call.

Our view therefore is that the current mixture of CPNP and RPNP should remain in place for today's services on NGNs. The charging principles applied to any new service should be decided upon taking into account which party is likely to derive most benefit from a call and may include CPNP, RPNP and B&K. For example, the costs of billing for Presence Information might outweigh the economic benefits and so it may be appropriate to use B&K as the charging principle. The numbering system, perhaps suitably adapted to ENUM, can continue to be used for receiving parties to signal to calling parties how much of the call charges they are willing to pay for.

References

DeGraba, Patrick (2000) *Bill and Keep at the Central Office as the Efficient Interconnection Regime* Federal Communications Commission, OPP Working Paper Series No. 33

DeGraba, Patrick (2002) *Bill and Keep as the Efficient Interconnection Regime?: A Reply* Review of Network Economics, Vol. 1 Issue 1 – March 2002 pp61 - 65

ERG (2007) *Final Report on IP Interconnection*

Gilbert + Tobin and CRA International (2006) *Economic Study of IP Internetworking: White Paper* Prepared for GSM Association.

Hermalin, Benjamin E., and Katz, Michael L., (2004) *Sender or Receiver: Who Should Pay to Exchange an Electronic Message* Rand Journal of Economics, Autumn 2004, pp423 - 447

Laffont, Jean-Jacques, Marcus, Scott, Rey, Patrick and Tirole, Jean (2001) *The American Economic Review*, Vol. 91, No. 2, Papers and Proceedings of the Hundred Thirteenth Annual Meeting of the American Economic Association (May, 2001), pp. 287-291

Littlechild, S.C., (2006) *Mobile termination charges: Calling Party Pays versus Receiving party Pays* Telecommunications Policy 30, pp242 - 277

Loder, T., Van Alstyne, M., and Wash, R., (2006) *An Economic Response to Unsolicited Communication Advances in Economic Analysis & Policy*: Vol. 6:Issue 1, Article 2. Available at: http://www.bepress.com/bejeap/advances/vol6/iss1/art2

Oftel(2003) *Review of fixed geographic call termination markets: Final explanatory statement and notification* 28th November 2003

Samarajiva, Rohan and Melody, William H. (2000) *Briefing Paper in Fixed Mobile Interconnection Workshop* ITU New Initiative Programme, Geneva, 20 – 22 September. Available at www.itu.int/osg/spu/ni/fmi/workshop/

UKCC (1995) *Telephone number portability: A report on a reference under section 13 of the Telecommunications Act 1984*

UKCC (2003) *Vodafone, O2, Orange and T-Mobile: Reports on references under section 13 of the Telecommunications Act 1984 on the charges made by Vodafone, O2, Orange and T-Mobile for terminating calls from fixed and mobile networks*

Wright, Julian (2002) *Bill and Keep as the Efficient Interconnection Regime?*

Yoon, Kiho (2006) *Interconnection Economics of All-IP Networks* Review of Network Economics Vol. 5 Issue 3, September 2006

Annex A

Let

$$q_a^d = a_0 - a_1(P)$$

be A's general demand function, and

$$q_b^d = b_0 - b_1(P)$$

be B's general demand function. Where a_0 and b_0 represent the sum of all variables affecting demand other than price and describe the point at which the demand curve intercepts the Y axis. The difference between a_0 and b_0 therefore reflects the difference in Ū of A and B.

Under CPNP, let the price charged be P* which comprises C_α + C_β such that:

$$q_a^d = a_0 - a_1(P^*)$$

and under B&K the price charged is P*/x where */x is the ratio of the retail price to A under CPNP to the retail price to B under B&K such

$$q_b^d = b_0 - b_1(P^*/x)$$

For there to be more completed calls under CPNP than B&K we have an inequality condition between the two demand functions where

$$a_0 - a_1(P^*) > b_0 - b_1(P^*/x)$$

Assuming $a_1 = b_1$ Solving the above to find the relationship at which demand under CPNP is greater than that under B&K we have:

$$a_o - b_0 > a_1(P^*) - a_1(P^*/x)$$

Then

$$a_0 - b_0 > a_1 P^*(1 - \frac{1}{x})$$