Prepared For:

GSM Association

71 High Holborn

London WC1V E6A

United Kingdom

Economic study on IP interworking:

White Paper

Prepared By:

Paul Reynolds, Bridger Mitchell, Paul Paterson, Moya Dodd, Astrid Jung of CRA International

Peter Waters, Rob Nicholls, Elise Ball of Gilbert +Tobin

Date: February, 2007

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The imminent migration to all-IP networks creates the potential for better, lower-cost delivery of existing services, plus the development of a wide range of exciting new services. This promises very large gains in efficiency and welfare, as supply can be more closely matched to customer demand and resources better allocated to their beneficial use.

IP interconnect is a critical enabler to capture these gains. As new and traditional services are offered via IP-based networks, and differential Quality of Service (QoS) is deployed, IP interconnection will have to become more sophisticated, and perhaps costly, in order to support these improvements. The inefficient and sometimes crude arrangements of today's IP interconnect regimes have their origin in the public Internet will not be adequate for interconnection between Next Generation Networks (NGNs) which will coexist alongside the "best-efforts" public Internet.

Which IP interconnect charging arrangements will be most efficient depends upon the efficient retail pricing of the end-user service, as well as the distribution of network costs. The efficient retail pricing model, in turn, is based on whether the benefits experienced are larger for the sender, or for the receiver of a message. The enhanced capabilities of NGNs enable retail and wholesale pricing to be better linked, in a way that current IP technology cannot.

There is no "one-size-fits-all" IP interconnect charging model that delivers superior efficiency in all situations. Initiating Party Network Pays (IPNP) is likely to be optimal in many cases. But in some circumstances, Receiving Party Network Pays (RPNP) can maximise efficiency. Bill-and-keep (BAK) is superior only in very limited circumstances particularly where traffic and costs are balanced and where there is no scope for strategic behaviour to alter that balance. BAK cannot respond to market dynamics because it effectively fixes the interconnect price at zero. Because NGNs will carry high traffic volumes bringing together a diverse range of services – including telephony, pay TV and other services with well-accepted retail charging paradigms – it would jeopardise efficiency and innovation to limit the kinds of wholesale arrangements that will underlie retail pricing. These risks are greater in an NGN environment than for traditional networks, due to the greater variety of services and greater variety of interconnection operators.

It therefore appears that, going forward, operators will need more freedom to negotiate interconnection charges that appropriately reflect their situation, rather than less, as would be implied by a mandated BAK model.

Regulators should therefore proceed cautiously in recommending or imposing any particular IP interconnect model. Existing regulatory frameworks based on objective assessment of market power are likely to prove suitable to remedy market failure where it has been identified. The instances where such market failure may arise are likely to be fewer than in traditional telephony because the very nature of IP interconnection, and the

services which it underpins, raises the potential for the traditional originating and terminating bottlenecks perceived in a legacy telephony world to be overcome.

Rather than prescribe solutions and risk regulatory errors with potentially profound negative consequences for efficiency and welfare, regulators should:

- intervene only in the event of demonstrable market failure (and if intervention can be expected to result in benefits which exceed the cost of regulation);
- intervene only to an extent that is necessary to remedy the market failure; and
- tailor the solution to the specific market circumstances, rather than applying a standard ”fall-back” option.

For the purpose of analysing which solution is best suited to fulfil the objective of the intervention – were it to occur – regulators should issue explicit assessment criteria, based on whether and how efficient market outcomes would be advanced, to guide parties as to how they will approach issues in dispute.

1. INTRODUCTION

Telecommunications networks are on the verge of profound generational change. Century-old circuit-based networks are being replaced by packet-switched "next-generation networks" (NGNs) using Internet Protocols (IP). This creates the potential for better, lower-cost delivery of existing services plus the provision of a plethora of new services that are not available on either circuit-switched networks (which will be replaced) or the public Internet (which will continue to exist alongside – and for many services in competition with – NGN). The adoption of NGN also provides an opportunity for improvements in economic efficiency and customer welfare, with supply more closely matched to customer demand and the application of society's limited resources to their most beneficial use.

Capturing these potential gains is immensely valuable. But this can only be achieved if the wholesale arrangements that underlie NGN retail services are aligned with economic efficiency considerations. Many networks make up the public Internet, underpinned by a proliferating number of network-to-network interconnect deals. If these wholesale IP interconnect arrangements distort efficient retail prices or fail to cover costs, then economic efficiency cannot be achieved, and much of the very large potential gain will simply be "left on the table".

Efficient IP interconnection is therefore fundamentally important in enabling NGNs to rise and prosper. It represents a critical leverage point for future gains – and just as importantly - for the making of investment decisions that will make those gains possible. Without them, the rationale to support such large investments may well founder, since efficiency gains deliver a large part of the value that justifies the investment.

This paper is a summary version of the long-form report *Economic Study of IP Interworking* by CRA International and Gilbert + Tobin, dated February 2007. In the following sections, we summarise:

1. *Why large efficiency and welfare gains beckon*: how IP interconnection works today, and why it can work much more efficiently in future.

2. *How IP interconnect can help these gains be captured*: why welfare and efficiency depends on the underlying interconnection arrangements, and how to assess what the right interconnection model might be.

3. *What the policy implications are*: what regulators and policy-makers need to do – and not do – to ensure that IP interconnect will support efficient outcomes.

2. LARGE EFFICIENCY AND WELFARE GAINS BECKON

IP interconnection is not a new phenomenon – it underpins the public Internet today. However, reflecting the technological constraints of the past and current IP environment, IP interconnection today suffers from many limitations that constrain its efficiency. Many

of these constraints will be resolved in the NGN environment, creating a substantial opportunity for welfare gains.

2.1. IP INTERCONNECT TODAY

2.1.1. Technical background

Today's public IP-based networks (the public Internet) differ from traditional circuit-switched networks in several important ways that limit the kind of interconnection arrangements that can be used.

Circuit-based networks were developed historically in the context of a limited number of operators and a high degree of centralised control. They establish a single physical path for the duration of a call or session, via a signalling network that provides end-to-end traffic management and billing information. This requires that all the networks used for a call be known to the originating and terminating networks, and have a commercial agreement to interconnect.

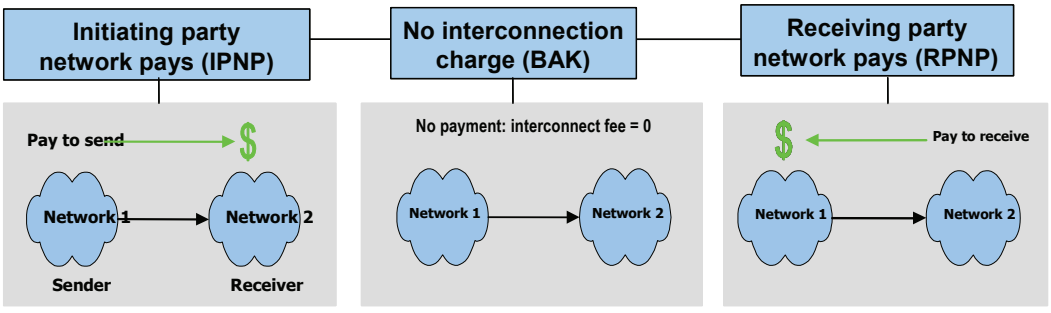
By contrast, current IP-based networks operate without centralised control and have proliferated in a far more liberal regulatory environment. They send the message as a series of packets, each bearing the destination address. These packets can take multiple, independent paths and are re-compiled at their destination into a coherent message. This is a connectionless system with no central control or central collection of billing information. Network partners are not necessarily known, other than the possible next network along a packet's pathway. As a result, services are limited to best-efforts quality without regard to the content of the packets, as the message is carried in numerous packets by an indeterminate set of operators along multiple unknown routes. Traffic is generally only measurable at the handoff points between each successive pair of networks along a packet's pathway.

These differences are important to understanding the varied – and sometimes crude – nature of current IP interconnect arrangements.

2.1.2. Charging models

The question of who pays whom for interconnect has three broad answers along a continuum of possibilities, as illustrated below.

Figure 1: Continuum illustrating who pays whom in current IP interconnect



In traditional telephony, IPNP arrangements most commonly apply. The caller is charged a retail price, a part of which is passed on to transit and/or terminating networks in order to complete the call. For most traffic, the called party does not pay to receive a call. There are of course exceptions – for example, RPNP applies to reverse-charged traffic e.g. 800-number calls, and calls to mobiles in some countries.

Importantly, these arrangements demonstrate a clear relationship between retail and wholesale models. Interconnect arrangements are struck with the retail model clearly in mind, since those retail arrangements generally have a high degree of historic consumer acceptance.

IP interconnect on the public Internet today, on the other hand, looks very different to telephony interconnect. Pricing arrangements apply at each handoff point (where measurement occurs) and are generally struck in isolation from other similar arrangements along the route. The lack of central control also means that IP interconnection deals are struck largely in ignorance of retail pricing arrangements – in other words, there is no real link between wholesale and retail charging models.

Presently, the most commonly applied IP interconnect principle is that a network receiving a packet should pay to do so (RPNP). Again, this is not without exception, and many different arrangements are struck on a bilateral basis between interconnecting providers. Some networks pay to send, others receive a payment for sending, yet others offset packets they send against the packets they receive. Where traffic is balanced, sometimes there is no payment at all – networks simply exchange traffic at an interconnect price of zero.

Several typical arrangements can be observed, based on the level in the Internet hierarchy of the interconnecting networks, for example:

- Internet backbone operators (so-called Tier 1 Internet providers) typically agree not to make any payments to each other (BAK), usually on the proviso that traffic is roughly balanced in each direction[1]. If traffic is imbalanced, the receiving network pays (RPNP);

- As between backbone operators and so-called Tier 2 Internet providers[2], the Tier 2 operator pays to receive (RPNP) but offsets the packets it uploads to the backbone operator (settlement-based interconnection or SBI);

- As between Tier 2 providers and pure resellers (so-called Tier 3 providers), the reseller always pays for both downloads (RPNP) and uploads (IPNP);

---

1 Strictly, if there is a condition of traffic balance then the arrangement is not pure BAK, but a settlement-based IPNP or RPNP arrangement where the traffic nets to zero. If the balance changes, then BAK no longer applies.

2 Tier 2 providers host some of their own content and peer at their own level, but still rely substantially on transit.

- For transit, the sending network pays (IPNP).

Because messages on IP-based networks consist of packets which travel via multiple routes over numerous networks, a single Internet session will result in several different charging models applying at different points between origin and destination.

In short, traditional telephony mainly operates with IPNP, under arrangements that reflect a strong link between retail and wholesale charging models, while IP-based services apply several charging principles, with wholesale pricing set largely in isolation from retail. While in traditional circuit-switched networks many interconnection services are subject to regulation, in the public Internet, each operator decides on a commercial basis whether it wants to peer with a second operator (BAK) or rather enter into a ”customer-provider” relationship (IPNP, RPNP).

One explanation for these differences may be that telephony typically serves users who are engaged in unique 1:1 interactions. While both parties usually benefit, the party triggering the message exchange bore the cost, although this asymmetry was often addressed over time with repeated and returned calls. By contrast, the Internet very rapidly became an enormous repository of publicly hosted content from which users could download at their request. Users came to pay to download (just as they pay to access other content such as books and pay TV).

The differences also reflect the lack of centralised control over the public Internet, with no end-to-end management of message routing or billing. These technical limitations constrain the extent to which wholesale and retail pricing can be linked.

But whatever the source and rationale for the differences, it is clear that an important question is raised as to how IP interconnection might be applied to a much wider range of services in the future. This is especially so for traditional services (such as telephony and pay TV) as they evolve towards IP. These services have retail pricing paradigms which are widely accepted and understood by consumers, but which would be undermined if today’s somewhat crude IP interconnect models were simply transposed onto them as they move to IP.

2.2. FUTURE IP INTERCONNECT AND NGNS

NGN rollout, first in the core networks and later in customer access networks, is a key enabler for the development of new services. Over time, IP transmission and switching will carry messages end-to-end, and interconnection will occur between NGNs.

2.2.1. Technical advances

In addition to dramatic increases in bandwidth, NGNs will also bring enhanced architecture, including control and service planes that offer something analogous to a signalling system, and enable more intelligent services to be deployed. This architecture will support quality of service (QoS) parameters, involving:

- Labelling of each packet by quality or priority;

- Creation of virtual QoS "paths", by which the necessary packet carriage priority is established for the message transfer (in order to deliver the required QoS, packets belonging to the same message will be more likely to travel the same route rather than multiple routes as now occurs);

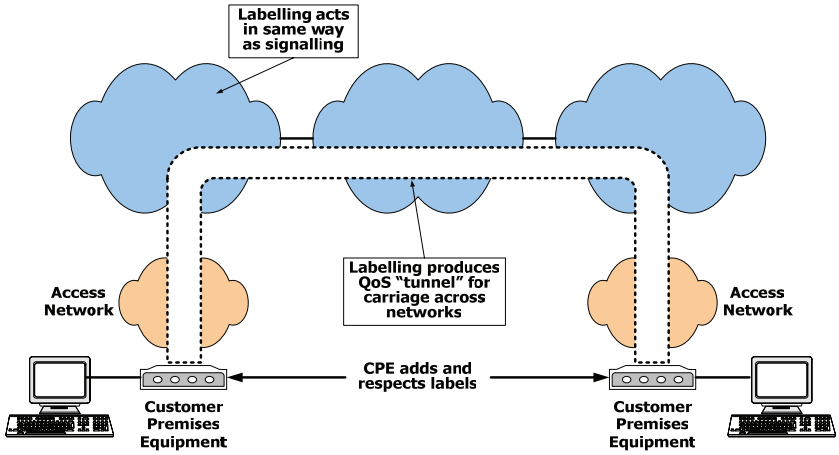
- Customer premises equipment that allow customers to select QoS demands on a message-by-message basis; and

- Billing differentially for QoS at both retail and wholesale levels.

As a result, QoS will substantially "raise the bar" for IP interconnection because all of the above parameters must be supported in the handover of packets from one network to another. Transport must be provisioned at each quality level and QoS paths sustained across multiple networks for the right duration. Customer selections of the desired QoS level must be fulfilled. Importantly for the development of pricing models, IP interconnect must also support the tracking and valuing of packets of various qualities and apply this to inter-operator billing.

As illustrated in Figure 2, mechanisms used to create QoS enabled transport paths within an NGN will be needed between interconnected NGNs.

Figure 2: QoS based IP interconnect



As a result, a packet's pathway across interconnected networks will be set up before it is despatched and the packets comprising a single message will flow along a single pathway.

2.2.2. Future charging models

QoS will enable major commercial developments at the retail and interconnection levels.

Retail services that require particular QoS standards, such as voice and television, can migrate onto IP-based networks (VOIP and IPTV). These can be supported by differential charging according to QoS to best match service requirements, thereby enabling a range of retail charging models that are not presently available. For example, VOIP services do not tolerate jitter (i.e. delay in the arrival of some packets relative to other packets) but do

not require high bandwidth. Web browsing, on the other hand, can tolerate jitter, but high bandwidth is important. QoS-based charging would be able to differentiate the pricing of these variations in a way that efficiently matches demand with supply.

As we explain in more detail below, if these new retail charging models are to be efficient, they must be supported by efficient wholesale (IP interconnect) pricing. Future IP interconnection will both require and support more sophisticated commercial interconnection relationships. Unlike in today's IP interconnect environment, the creation of virtual QoS paths across networks means that the networks providing interconnection will be arranged on an end-to-end basis and billing information can be gathered and passed along the chain.

Creating a QoS path does not determine the direction of charging for wholesale or retail services. QoS enabled transport services can support retail calling party pays and receiving party pays approaches. But importantly at the interconnection level, interconnection between networks along the virtual pathway can be consistently configured to a particular interconnection model, such as IPNP or RPNP, which is not possible with current IP interconnection.

3. CAPTURING GAINS BY EFFICIENT IP INTERCONNECT

In this section, we explain why interconnection charging is so important to achieving efficiency and consider how to work out what kind of interconnect charging will be efficient. We then briefly discuss the main IP interconnect charging models in terms of their effect on economic efficiency.

3.1. WHY INTERCONNECT IS A LEVER FOR EFFICIENCY

IP interconnection is a critical lever for economic efficiency. Interconnection charges represent the underlying wholesale costs that must be borne by retail services. They impact network cost recovery (and hence investment and innovation incentives) for all interconnecting networks and they impact retail prices (and hence consumer demand).

To be efficient, IP interconnect charges must ensure that:

- the costs of each interconnecting network are covered; and
- efficient retail pricing is supported.

In an NGN environment, it is likely to be even more important to optimise IP interconnection. Not only is a much larger range of services delivered via IP, but the additional costs of supporting QoS raise the risk involved. If, for example, interconnect arrangements did not allow QoS costs to be fully recovered by all networks, then QoS may not be developed and deployed widely. This would delay the migration of QoS-dependent services, thereby requiring costly and outdated circuit-switched networks to be maintained in parallel with NGNs for a longer period.

3.2. EXTERNALITIES AND EFFICIENT RETAIL PRICING

The need to support efficient retail pricing raises some issues that are specific to communications, because messages are *jointly consumed* by both sending and receiving parties. When a party other than the paying party receives a benefit, an economic "externality" is the result. For example, if one person calls another to make a mutually beneficial arrangement, and the calling party pays for the call, then the called party gains a benefit (or positive externality) for which they have not paid. Externalities can also be negative, as occur with nuisance calls or spam.

Because of this joint consumption, retail pricing for communications faces special challenges in achieving efficiency. Ideally, pricing should be such that it encourages only the messages that would pass a cost-benefit test. That is, the only messages sent should be those where the combined benefit (to both parties) exceeds the total costs to all networks involved in the delivery. Any messages where this is not so – for example, if the message actually yields a total net benefit (taking into account the benefit to the sender and any cost to the receiver) *lower* than the cost of delivering the message – then ideal pricing would exclude that message from being sent (say, by pricing it higher than the sender's willingness to pay based on the benefit it alone derives). Unsolicited messages (spam) fits within this example.

In an ideal world, efficient retail pricing would follow the allocation of benefit. The sender alone would pay where it alone benefits and the receiver alone would pay where it alone benefits. Where the benefit is shared, the retail payments should also be shared.

But precise measurement of the benefit allocation is difficult, and billing additional parties imposes transaction costs, so practical considerations often dictate that only one party pays even though benefits are likely shared. The most efficient party to pay is the one for whom there is sufficient benefit available to induce them to send all (or most) of the socially desirable messages that they might initiate.

In many cases of two-way communication (including most telephony), it is efficient for the calling party to pay, because this regime generally leads to most socially undesirable messages being stopped and the most desirable messages being sent. In particular, calling party pays will be suitable to most cases of one-off calls or messages where it is likely to be the person initiating the call who obtains most of the benefit (e.g. where a caller is seeking information). In repeat calling arrangements where parties take turns to call each other, the benefits may be more evenly shared. However, in these cases, the retail charging model is less important, precisely because both parties are prepared to call each other over time. Where one party compensates the other party, such as a parent paying their children's mobile bill, then the retail charging model will also be less significant for ensuring efficient outcomes.

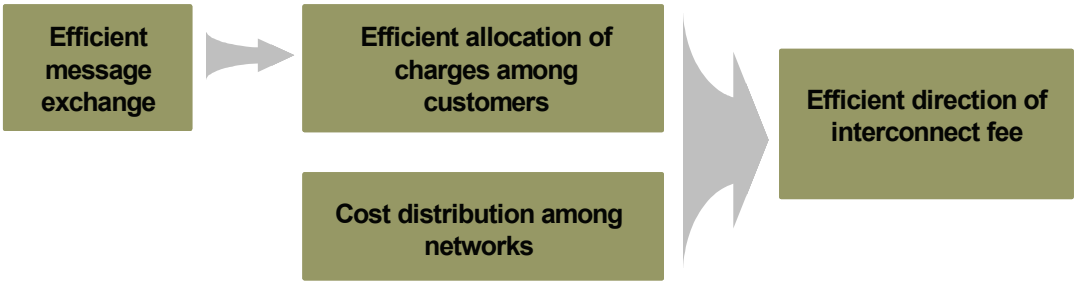
3.3. DEDUCING EFFICIENT INTERCONNECT CHARGES

3.3.1. From the retail model and network costs

Once the efficient retail model is understood (i.e. which of the interconnected networks should be the retail provider), the efficient interconnect charge must be aligned in order to provide the right incentives to the retailer. As noted above, interconnect charges must also ensure that network costs are recovered.

These determinants are shown in Figure 3.

Figure 3: Determinants of efficient interconnection



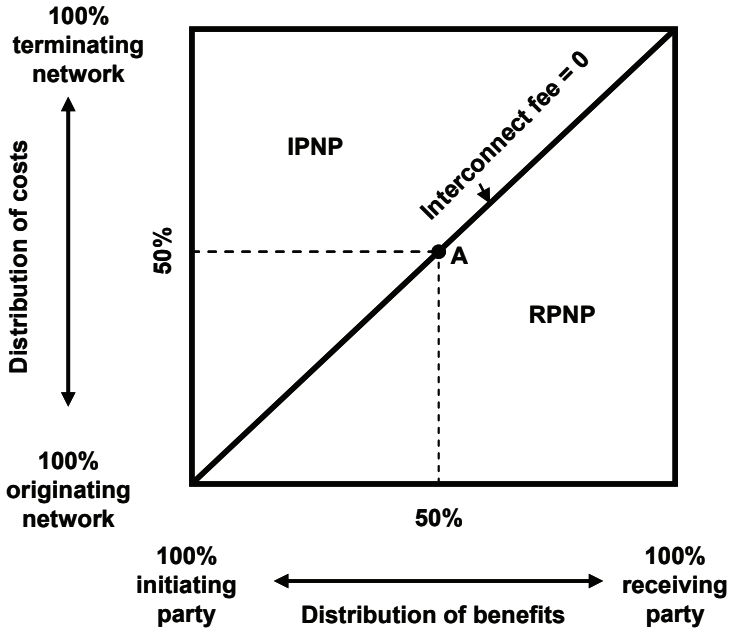
The relationship between the determining factors and the efficient outcome can also be shown in graphic form (see Figure 4 below). This figure shows how both the distribution of costs (as between originating and terminating networks) and the distribution of benefits (as between sending and receiving parties, which defines the efficient retail pricing model) determine the most efficient model for the underlying IP interconnection.

Figure 4 below shows how an economically efficient interconnect model can be derived if the distribution of benefits and costs for a specific message are known[3]. The area to the left of the diagonal line represents situations where IPNP is efficient; and the area to the right represents those situations where RPNP is efficient. If the plotted position falls exactly on the diagonal line, then it is efficient to pay no interconnect fee at all (BAK) – a situation where the distribution of costs happens to align exactly with the distribution of benefits.

For example, at point A on the line, exactly half of each of the costs and benefits lie with each party, and BAK would work efficiently as the interconnect model. If, in contrast, both retail parties benefit equally, but the terminating network incurs more than half of the costs, then payment of a termination fee would be efficient (points above point A in the graph). Similarly, if both retail parties benefit equally, but the originating network incurs more than half of the costs, then a fee for origination would be efficient (points below point A in the graph).

3 This model assumes that the total benefit of the message is equal to the total cost of providing it.

Figure 4: Deducing the efficient interconnection model from retail benefits and costs, where total costs = total benefits



If only the initiating party benefits, but some costs are borne by the terminating network, then IPNP is efficient regardless of the exact distribution of costs. Similarly, if only the receiving party benefits, but some costs are incurred by the originating network, then RPNP is always efficient.

3.3.2. When traffic is balanced between peers

In very specific circumstances where traffic is balanced between peers (that is, networks that have the same cost structure and customer profile) finding an efficient interconnection fee is much simpler because net interconnection payments are equal to zero *regardless* of the size of the interconnection fee and its direction. In other words, BAK, IPNP and RPNP would all yield the same result. However, BAK would be the preferred model, because it avoids transaction costs (e.g. measurement, billing).

For BAK to both *be* efficient and *stay* efficient, the traffic must remain balanced. If operators can strategically alter the traffic balance, then BAK would no longer be the most efficient model.

3.4. THERE IS NO "ONE-SIZE-FITS-ALL" INTERCONNECT MODEL

Because retail pricing models vary (especially as different types of services make their way onto IP-based networks) and cost conditions vary across markets and networks, there will be no single "one-size-fits-all" interconnection model that maximises efficiency in all situations. Instead, a variety of models employed across different circumstances and networks is likely to best promote efficient market outcomes overall.

Moreover, dynamic effects must be taken into account. Setting the right price is one matter; but it can quickly become the wrong price if circumstances change. Any model that "locks in" a static price (as does BAK, with a zero price) risks becoming inefficient, even if it is efficient to begin with. This is particularly important when market circumstances change because of strategic action by network operators (e.g. the "hot potato" problem, where networks using BAK hand over traffic at the earliest possible point in order to minimise their own tasks). An interconnect model that is able to evolve to address any such distortion is critical to preventing inefficient strategic behaviour.

3.5. OBSERVATIONS ON EXISTING MODELS

No charging model is universally superior and each has strengths and weaknesses in different situations. Here we note some characteristics of each.

3.5.1. BAK (Bill and Keep)

Disadvantages of BAK derive from the fact that there are only limited conditions under which it yields efficient market results:

- BAK is superior only under very limited conditions: balanced traffic between peers; and where the distribution of costs among networks happens to align exactly with the distribution of benefits among retail customers.
- In most cases, BAK leads to market distortions and damages efficiency. With zero interconnect revenues, networks must recoup all costs from their own customers (e.g. in the case of VOIP calls, both the calling and the called party would have to pay a retail charge) and this usually leads to inefficient retail pricing.
- Because BAK is inflexible, it can lead to the "hot-potato" problem. The result is network structure bias: costs are pushed onto other networks; but despite their increased costs, no adjustment is made to the zero interconnect fee. If costs are under-recovered, networks will under-invest.
- Applied to transit, BAK would discourage the provision of transit services entirely since transit networks have no direct customers from whom to recover their costs.
- These inefficiencies are likely to be amplified in a QoS world, where network costs are greater (so the unrecovered costs would be larger).
- Applying BAK to services like telephony – where IPNP is the historic model – would lead to upheaval in retail pricing models and major transitional issues for customers.

Against these, a potential advantage is that BAK avoids transaction costs between operators in case of symmetric traffic between peers. However, this benefit is offset if strategic behaviour and traffic balance needs to be monitored to check whether the conditions in which BAK is efficient still hold.

3.5.2. IPNP (Initiating Party Network Pays)

IPNP has a number of advantages that avoid the problems of BAK:

- IPNP will outperform BAK in most situations, so long as there is flexibility for the price to adjust to changing retail and cost conditions.
- It represents the model most commonly applied in traditional telephony interconnection. Customers may be resistant to alternative models that require charges for activities such as receiving calls that they do not currently pay for.
- IPNP also helps to discourage unsolicited messages, e.g. spam.
- It works well when the calling party gains most benefit. It may discourage some messages beneficial to recipients, as the retail price to the sender may outweigh the sender's benefit. However, in many situations message benefits may be able to be re-balanced through repeated or returned calling (e.g. taking turns to call) or offline relationships between the sender and receiver.
- For transit, it is superior to BAK in that it provides a revenue source to cover otherwise stranded transit costs.

Against these, a disadvantage of IPNP is that regulators in some cases have been concerned that the level of the termination charge may not be competitively constrained to the efficient level. However, such concerns are likely to be less relevant in future IP networks.

3.5.3. RPNP (Receiving Party Network Pays)

Advantages associated with RPNP are:

- RPNP has the potential to outperform BAK, so long as there is flexibility for the price to adjust to changing retail and cost conditions.
- Messages that primarily benefit the receiving party may be sent when they would not otherwise (for example, a call to report that a company's lost property has been found).
- For transit, unlike BAK, RPNP provides a revenue source to cover otherwise stranded transit costs.

The main disadvantage of applying RPNP generally would be:

- RPNP applied generally would risk a massive growth in unsolicited messages (spam) to the detriment of recipients, as the sending network (and possibly, the sender) will face zero cost and often cannot be "punished" by the recipient.

3.5.4. Summary

The following table summarises the key performance characteristics of each interconnect model and its implications:

Model	Advantages	Disadvantages	Implications
BAK	- In case of balance of traffic, this appears simple and low-cost. However, efficient market results could require monitoring of retail market conditions, operator costs and/or traffic balance	- Even in a static market, BAK is superior only under very limited conditions. It leads to market distortions in most cases, which are amplified in QoS environment and when BAK is applied to transit - Inflexibility of fee to adjust to evolving market conditions and cost-avoidance (e.g. "hot-potato routing") creates further distortions	- Suitable only in limited situations (e.g. sustained traffic balance between peers) - Lacks the flexibility required to maintain efficiency when circumstances change
IPNP	- IPNP avoids the problems of BAK because it does not set a specific fee and prices can adjust dynamically as conditions change - Likely to perform well in many situations because it discourages spam, whilst not significantly impeding messages that benefit mostly the receiving party	- Regulators have been concerned in some cases that the level of termination charges may not be effectively constrained. Such concerns are likely to be less relevant in an all-IP world.	- Likely to be the best performing model in most common situations
RPNP	- RPNP avoids the problems of BAK because it does not set a specific fee, and prices can adjust dynamically as conditions change	- Encourages spam	- May be suitable in some situations, although dominated by IPNP in most cases

3.6. CONSEQUENCES OF INEFFICIENCT INTERCONNECT MODELS

If an inefficient interconnection model is imposed, consumer welfare is harmed. Services may not be provided to their fully optimal extent, thereby retarding the market and any dependent markets, and costs may not be covered. This can damage investment incentives and stifle service development. It can also bias network design (e.g. the "hot potato" problem of operators facing incentives to hand off traffic as soon to minimise their costs even if total network costs are increased) or force networks into engaging in other measures to try to recover costs from a less efficient source.

Should a single IP interconnect model be imposed, retail innovation will also be stifled because variety in retail pricing models must be supported by a range of appropriate wholesale pricing models. If, for example, the close link between retail and wholesale pricing of telephony is lost in the migration to IP-based networks (say, because BAK was mandated) then efficiency losses are likely to result.

In short, inefficient IP interconnect could strand many of the anticipated benefits of an NGN and thus fail to capture the opportunity for a step-change in efficiency levels.

4. POLICY CONSIDERATIONS

4.1. ROLE FOR REGULATION

Regulatory intervention is always risky. If the regulatory settings are wrong, investment can be chilled, competition damaged and service development hindered.

Moreover, there is no certainty that market failure will occur as a result of IP interconnection as more services migrate onto IP-based networks; and it is even less clear what the right regulatory response might be. In particular, there is no clear view of where future potential bottlenecks might lie. Any-to-any connectivity requirements, for example, may well become less important as interconnection cannot effectively be blocked in a fully IP-based world.

What is clear is that many different kinds of services, with a great variety of retail charging models, will be carried by NGNs. Some of these models are very well-established (for example: calling-party pays in fixed telephony; or time-based rather than data-based charges for mobile access). These retail models must be translated into efficient wholesale charging models. Simply transposing the limited precedents of today's IP interconnect world would not only require a dramatic shift in consumer preferences (and probably generate considerable resistance), but would also be likely to stifle more efficient, market-driven solutions.

Regulators are therefore left with the quandary of wanting to provide regulatory certainty, but without incurring the risks of intervention and error.

Rather than prescribe specific interconnection solutions and risk regulatory errors with potentially profound negative consequences for efficiency and welfare, regulators should:

- intervene only in the event of market failure (and if intervention can be expected to result in benefits which exceed the cost of regulation);
- intervene only to an extent that is necessary to remedy the market failure; and
- tailor the solution to the specific circumstances, rather than applying a standard "fall-back" option.

The regulatory frameworks that exist in most countries are sufficient to address market failure problems where and when they arise – there is no need for regulators to define a special framework for IP interconnection.

For the purpose of analysing which solution is best suited to fulfil the objective of the intervention – were it to occur – regulators should issue explicit assessment criteria, based on whether and how efficient market outcomes would be advanced, to guide parties as to how they will approach issues in dispute.

This approach would provide regulatory certainty without losing the flexibility to adopt a solution that will actually solve the problem that has been identified (without creating new ones).

Below we outline a recommended set of criteria for assessing different IP interconnect charging models in various circumstances, based on maximising efficiency and welfare.

4.2. ASSESSMENT CRITERIA

Economic efficiency is defined as the best use of resources (allocative efficiency), least cost production (productive efficiency) and incentives for innovation and investment (dynamic efficiency). Efficiency is a precondition to maximising welfare and in most practical circumstances consumer welfare is enhanced by increasing efficiency. With efficiency gains, prices fall, quality improves (to the extent consumers are willing to pay for it), costs are recovered (so investment incentives are preserved) and all messages carried have a value that is at least as high as the cost of delivering them.

The following checklist encompasses the practical market outcomes of economic efficiency applied to the context of interconnection. It can be used as a guide to test whether any proposed IP charging model is likely to provide any real efficiency benefits. Any proposed regulatory intervention in IP interconnection charging should improve overall outcomes with respect to the checklist below, above and beyond what the market could achieve without intervention.

	Type of impact	Efficiency outcome
1.	Consumer benefits	• All customers are served for whom the total benefit of having them on the network is greater than the cost • Full range of services demanded by customers is provided, including innovative new services • Differential QoS is available, to match customer demand • Individual messages are sent if and only if the total benefits to the initiating and receiving customers are equal to or exceed the incremental cost of the messages • Low prices, provided that prices cover the long-term costs of providing services efficiently
2.	Network operator impacts	• Efficiently-incurred operating costs are recovered • Operators have the incentive to undertake efficient investment and innovation • Interconnection arrangements are available which allow services to be provided in line with consumer demand (e.g. end-to-end QoS)
3.	Market operations benefits	• Efficient competition is promoted and inefficient arbitrage is avoided • Costs are minimized by efficient network usage and call routing, including packets being handed off at economically efficient points • Changes in interconnection charging models are made if and only if the benefits exceed the transition costs
4.	Regulatory impacts	• If regulation is applied, regulatory administration and operator compliance costs are minimized

In some cases, there may be a need to trade-off particular criteria so as to determine the optimal charging model. This can be done by assessing the quantum of the competing impacts.

4.3. POLICY RECOMMENDATIONS

There are some clear implications for policy-makers, based on the preceding analysis:

- *Proceed cautiously*: The above analysis of interconnect models, and the absence of evidence of market failure, imply that there is no justification for regulatory intervention at this stage. It is too early to tell what model or models will prevail commercially, particularly as many of the new services are still being developed. Regulatory intervention to prescribe a particular model, such as BAK, would be pre-emptive and risky. There is no evidence that the industry will not be able to work out appropriate IP interconnection models without *ex ante* regulatory intervention (for example, global connectivity for the current Internet was achieved without any regulatory intervention). Mandating particular interconnection charging arrangements may well inhibit the development of inherently more effective and efficient IP operating models.

- *Don't mandate a single charging model.* Our analysis also suggests that, even if a particular charging model develops commercial currency, it is not necessarily appropriate for regulators to mandate this model. A single wholesale model will constrain the variety of retail models that are necessary for efficiency. With the multitude of products being developed in the IP environment it would appear that operators will need more freedom to negotiate interconnection charges tailored to their situation, rather than less, as would be prescribed by a mandated charging model.

- *Don't assume bottlenecks will be replicated.* Our discussion of the role of regulation shows that the deployment of NGNs has the potential to change the way many services are delivered – for example, by leading to more "multi-homing" of content which undermines the potential for bottlenecks. As a consequence, a statutory any-to-any connectivity condition is unlikely to be needed to ensure competitive outcomes in a fully IP world. Moreover, even where regulators intervene to ensure any-to-any connectivity, regulators will still need to determine the most efficient charging arrangement as multiple charging models can be consistent with achieving any-to-any connectivity.

- *Use existing regulatory frameworks.* Existing regulatory frameworks are likely to be sufficient to resolve problems should they arise. Current sector-specific and competition powers generally permit regulators to intervene if bottlenecks emerge in IP interconnection.

- *Employ consumer welfare analysis.* However, in circumstances where regulators identify market failure or are requested to resolve disputes, the resolution should be tailored to the specific circumstances and be applied only as broadly as necessary to solve the problem. Regulators should not define a single charging model that would be the "fall-back" option, but rather should employ a clearly defined assessment framework that appropriately reflects the drivers of consumer welfare and broader economic efficiency. We have suggested such a framework in section 4.2 above.

5. CONCLUSION

The large welfare and efficiency gains made possible by NGNs critically depend upon the efficiency of future IP interconnect arrangements.

NGNs will carry a wide range of services (including telephony and pay TV) with diverse retail pricing models. Wholesale (i.e. IP interconnection) pricing must support that diversity if it is to sustain efficiency and innovation in retail markets.

The efficient wholesale pricing model for a service can generally be derived from two factors – the efficient retail price, and the distribution of network costs among the transporting networks – but the answer is highly situation-dependent and may change over time.

Consequently, no single IP interconnection model is superior in all circumstances. IPNP is likely to be superior in most cases, but RPNP and BAK can also be optimal (although, in the case of BAK, this is only in very limited circumstances).

Regulators should therefore be cautious in imposing any particular IP interconnection solution. Prescribing a specific model is fraught with risk, especially when it is considered that IP interconnection must in future support telephony, which has a well-accepted retail charging model. For example, if BAK were mandated, telephony customers may be charged to receive telephone calls.

Rather than prescribe solutions or transpose them from today's (inefficient) IP interconnect environment, regulators should instead set out the assessment criteria against which they would test regulatory proposals. In essence, they should ask whether the proposal would advance efficient long-term outcomes for consumers beyond what would occur in the absence of regulatory intervention.