Distance gradients

Assessing the impact of NGNs on interconnection tariffs' distance gradients

A report for Ofcom

March 2006

Contents

Tables & Figures

We have been asked by Ofcom to consider:

- how next generation networks (NGNs) affect the relationship between cost and the distance over which traffic is conveyed (the so-called “distance gradient”); and
- the potential consequence of any changes in cost for the structure of regulated interconnection charges.

In this report, we present a qualitative assessment, focussing on the broad structure of costs. We consider in turn:

- the principles determining the relationship between distance and cost in telecoms networks;
- how this relationship affects incentives for infrastructure competition;
- the differences between NGNs and traditional networks; and
- the likely cost implications of the move to NGN.

We find that:

(i) Moving traffic onto a common platform brings unit cost savings by exploiting economies of scale and scope;

(ii) The decrease in unit costs is greater at the periphery of the network than at the core, where scale and scope economies have *already* been exhausted to a greater extent that at the periphery. Therefore, the move to an NGN makes unit costs at different levels of the network hierarchy more similar than previously;

(iii) Although NGNs are sometimes described as a move to an Internet-style network in that many services are carried on a common platform, there is no “death of distance”. Although an NGN allows traffic to traverse distance more cheaply, costs do not become independent of distance. A flat-rate charging model is inappropriate for regulated interconnection, as it would not reflect costs and would eliminate incentives for efficient bypass of the incumbent;

(iv) The current element-based charging (EBC) methodology for interconnection pricing can remain, but cost differences across the various levels of network hierarchy become smaller. Potentially, if the trend to more similar unit costs across the network were sufficiently pronounced, charging by distance rather than by level of network hierarchy might become a better proxy for the network costs of conveying traffic;

(v) There are no strong reasons to expect the potential for competition at the core to be affected by the move to NGN, as entrants in this segment can aggregate traffic significantly and achieve the minimum efficient scale;

(vi) However, other factors equal, the move to NGN may reduce incentives for infrastructure competition at the periphery, as incumbent’s costs in this segment are likely to fall significantly due to increased aggregation of traffic. However, where alternative networks can aggregate traffic in a similar way to the incumbent and have access to similar technologies, the differences in unit costs between entrants and incumbent should narrow. Therefore, infrastructure-based competition, where feasible, is likely to involve competition for the entire bundle of services offered to customers, rather than for individual services.

1 Introduction

BT is planning to undertake fundamental changes to their network, moving it to a Next Generation Network (NGN). NGNs are based on integrating a wide range of communications services on a single multi-use network, using IP transmission on all shared network segments.

The move to an NGN substantially affects the network costs faced by BT and the relationship between the cost of conveying traffic and distance over which traffic is carried, which is our particular concern in this report. Throughout this report, we use the term “distance gradient” to refer to this relationship.

Underlying network costs determine the structure of interconnection charges, which in turn set the distance gradient faced by operators using the incumbent’s wholesale products in order to provide their services. Thus, the interconnection charge distance gradient provides the framework upon which entrants base alternative infrastructure investment decisions. From a regulatory standpoint, the structure of interconnection charges is crucial in providing appropriate incentives to encourage socially efficient infrastructure build-out and platform-based competition.

The similarities between NGNs and the Internet have raised the question of whether the move to NGN will bring the “death of distance” in interconnection charges. Underpinning this argument is the observation that Internet charges are typically independent of the distance over which data is conveyed. However, Internet charges are not regulated and do not need to reflect underlying detailed structure of costs. In order for the distance gradient to disappear entirely from cost-reflective interconnection charges, it would be necessary for costs to become unrelated to distance. Although distance-related costs may become smaller with the move to an NGN, networks costs still increase with the distance covered.

In this report we consider how the move to NGN may affect the distance gradient and the implications for regulation of interconnection charges and the incentives for infrastructure competition. We start by looking at the basic concepts and definitions underlying distance gradients in Section 2.

Throughout the report, we use the terms “core” and “periphery” loosely to contrast the various layers of hierarchy within a network. In the “core”, there is considerable aggregation of traffic achieved on network links, with traffic having passed through one or more nodes at which aggregation occurs. Conversely, in the “periphery”, much less aggregation is achieved on network links. We will typically identify the local loop as a further level of network hierarchy, which has the distinguishing feature that links are dedicated to individual customers. When we refer to “core” as opposed to “periphery”, we do not necessarily use the term “core” in quite the same way as BT’s published plans for an NGN.

Due to scale economies, the unit cost of carrying communications over a given distance is smaller the greater the degree of traffic aggregation on links within the network. This typically implies a steep distance gradient at the edges of the network where links are dedicated to each user; the distance gradient becomes flatter as aggregation of traffic across customers allows for costs to be shared between different users, and becomes flattest at the core of the network, where distance can be traversed most cheaply. This logic is confirmed when looking at the current interconnection charges set by BT and their underlying costs.

The distance gradient in interconnection charges provides an instrument by which variation in costs related the distance over which communications are carried can be reflected in interconnection charges. The extent to which the distance gradient accurately reflects underlying costs impacts on the incentives for efficient alternative network infrastructure investment.

In order to assess the potential impact of upgrading the network to an NGN structure, we need to identify the main differences between NGNs and traditional telecommunications networks. In Section 3, we briefly review BT's proposed plans to upgrade their network to NGN, and identify the changes in network structure that can typically be expected from NGNs. The move to an NGN allows for greater sharing of links, and decreases the extent to which traffic is differentiated both across services and in relation to where it is within the hierarchy of the network.

In Section 4, we assess the likely impact of the changes in network structure from moving to an NGN on the distance gradient. NGNs allow for greater exploitation of scale economies through a number of different mechanisms. This results in lower costs overall, but can be expected to have a stronger impact on the edges than at the core of the network. This decreases differences in the distance gradient between different network segments. We finish our assessment of the likely impact of NGN by looking at how competition may affect or be affected by these changes in the distance gradient.

We finish by drawing our conclusions in Section 5. These conclusions are qualitative and are intended to identify the general issues arising from the move to NGNs. Therefore, the analysis provides a starting point for quantitative assessment, but is not intended to provide quantification by itself.

2 Distance gradients – general considerations

In this section, we set out some general principles and definitions concerning distance gradients, focusing mainly on how 'cost' varies with 'distance'. We set out terminology for describing the structure of distance-related costs that we use throughout the document.

For access seekers, distance gradients in interconnection charges are a major factor determining the number and location of points at which they wish to interconnect with the incumbent's network and thus the extent to which they will need to build out their own infrastructure. In order to provide the efficient incentives for infrastructure investment, the cost impact of the distance over which traffic is carried should be reflected in interconnection charges.

Both the absolute level of interconnection costs and the structure of these costs affect investment decisions by entrants. In order to illustrate this, we provide some examples of how distance gradients affect investment decisions of alternative providers, and set out how we expect distance gradients to affect the nature of competition in general.

2.1 Distance related costs

Network costs increase with the distance over which communications are carried. However, although it is intuitively plausible that 'more' network infrastructure is needed in order to carry traffic over a greater distance, and that therefore greater distance implies greater costs, the precise nature of the relationship between cost and distance is far from straightforward, as it depends on network topology.

2.1.1 Network costs specifically related to distance

Consider the simple case of a communication link between two points, providing a certain capacity (measured, for example, in busy hour Erlangs or available bandwidth). The costs incurred in providing this link are likely to be higher the further apart the two points, for example, for the following reasons:

- right-of-way costs are likely to be higher the larger is the distance;
- the cost of civil engineering (trenching, ducting, erecting poles etc.) increase with the distance that has to be covered;[1]

[1] The impact of distance on costs, however, is not uniform but depends on factors such as topography, terrain characteristics or population density, although in practice it may be necessary to establish cost measures that average across differences in these factors. In addition to affecting costs directly, such factors may also impact on the optimal network topology, and therefore they are often considered in bottom-up engineering models of network costs. For example, the Loop Design Module of the FCC's Hybrid Cost Proxy Model for the local access network requires the user to specify terrain data consisting of bedrock depth, rock hardness, soil type, depth of water table, minimum slope, and maximum slope.

- the cost of cabling increases with cable length;
- where wireless links are used, a greater number of transmitters and receivers are required in order to bridge a greater distance;
- links between two points that are further apart tend to involve traffic passing through a greater number of switches.

Assuming a positive relationship between investment costs and operating expenditure, these effects imply that both capital costs and operating costs increase with the distance between the two points connected by the communications link.

Obviously, the impact of distance on unit cost varies across the different categories of distance-related costs. Costs of cabling, for example, are likely to vary with capacity (which is related to actual and forecast traffic volumes) and with distance. Civil engineering costs, by contrast, are likely to vary only with distance, but not necessarily with capacity. Switching costs may vary with distance to the extent that more switches are passed by traffic travelling further, with the costs of switches depending in turn on the number of links at that node and the switching capacity they provide.

The link between distance and any reasonable measure of 'unit cost'[2] is complicated by the fact that:

- network investment is often lumpy rather than varying smoothly with the volume of traffic (or available capacity);
- many costs do not vary at all with traffic volumes or capacity;
- network topology plays an important role, with routing and sharing of link capacity at certain points in the network; and
- network investments are often sunk.

We consider complications related to investments being sunk in Section 2.2. However, we first consider the issue of recovery of costs that do not vary with distance.

2.1.2 Sharing and recovery of fixed and common costs

Broadly speaking, efficiency is promoted where access charges reflect forward-looking incremental costs, as then alternative providers have incentives to bypass regulated networks only where they can provide the service more cheaply. This would suggest that distance gradients should reflect the change in incremental costs that are related to distance, e.g. the higher cost of cabling and

[2] Even though it might be possible, in theory, to measure costs for each particular network element (links and nodes) that is actually used in the provision of a specific interconnect service, and then set a price in line with the actual cost incurred in carrying, say, a voice call between two specified points, such an approach is unlikely to be feasible in practice. Realistically, it is necessary to define a number of standardised products, and calculate an appropriate measure of unit cost by averaging across the different costs that might be incurred for different products in a particular product category (which results in interconnection products reflecting the average distance in each category rather than the actual distance for each individual transmission).

switching (which are directly driven by distance and capacity) and a contribution towards costs that are fixed with regard to capacity, but are related to distance (e.g. civil engineering costs averaged over the total capacity available on a particular link). However, in practice, it is typically necessary to recover some costs that are common across services and fixed with respect to both distance and capacity.

In order to allow for the recovery of common network costs, interconnection charges for individual services may need to include some mark-up above incremental cost. This leads to an efficiency loss, as there may be inefficient bypass with some alternative networks choosing to provide a service even though they are less efficient than the regulated network, and because demand is being priced off even though willingness to pay exceeds the cost of serving such demand.

However, this situation may be unavoidable if the regulated network is to recover common costs. In this case, it is appropriate to consider a second-best notion of efficiency, with interconnection charges being set at a mark-up on LRIC to allow recovery of these common costs in such a way as to minimise the resultant inefficiencies. In any case, including such a mark-up can only strengthen the incentives for alternative infrastructure provision.

With regard to distance gradients, it is helpful to distinguish between:

- costs which are common across products and services that are neither related to distance, nor to capacity or traffic volumes; and
- costs which are common across a range of products, or fixed with regard to capacity, but vary with distance.

If common costs unrelated to distance were recovered on the basis of distance, doing so would distort prices by making traffic carried over large distances unduly expensive relative to that travelling a shorter distance. This would amplify the impact of distance on cost, and tend to make distance gradients steeper. As a result, replication of longer links would become relatively more attractive.

With regard to the second class of common costs, although it would seem intuitively plausible that link-specific fixed costs should be recovered from charges for services carried over this link, and that averaging these costs over the traffic carried on this link might be an obvious solution, complications arise if multiple different services are carried over the link (as will be the case in full IP networks). In this case, the argument for simple averaging is reduced to one of mere practicality without any other justification.

Differential mark-ups over distance-related variable costs (or incremental costs) may be appropriate where demand for different services has significantly different price responsiveness and where simple averaging would lead to inefficiencies.[3] A FAC model may avoid some of these problems, but only provided the cost allocation key used in some way proxies for relative demand responsiveness.

A common regulatory practice is to use equi-proportionate mark-ups (EPMU) due to the strenuous information requirements of structuring optimal mark-ups. Although EPMU is widely used due to its simplicity, it may provide an inefficient structure of interconnection charges. Where there is different demand responsiveness for different interconnection products, it is efficient to set different mark-ups in order to minimise demand distortions.

Similar incentives to set differing mark-ups may be present for operators in competitive markets, as they may face differing demand responsiveness for different products. This might lead to differences between the distance gradient of interconnection charges and the distance gradient of LRIC where prices are not regulated with reference to underlying costs. This needs to borne in mind when examining interconnection pricing in non-regulated context (see Annex A).

2.2 Cost measurement issues

There are a variety of issues regarding how costs are measured, which have a material effect on investment incentives. Specifically, the valuation of network costs can differ substantially depending on whether costs are measured using actual costs incurred or using a forward-looking approach.

In particular, uncertainty about future network requirements and uses, combined with the existence of legacy network infrastructure, makes it impossible to design a network that would be optimal if built anew at any point in time. This can result in additional network costs due to inefficient network dimensioning and/or topology inevitable from outturns being different to forecasts.

2.2.1 Network dimensioning

The capacity to which the network is dimensioned has an important role in determining total costs. Additionally, network operators often are able to lower the cost of adding future capacity, as and when required, by incurring a greater proportion of fixed costs at the point of building a network – although these fixed costs are sunk and cannot be recovered in the case that no further capacity is needed in the future. In effect, this is building ahead for risk management reasons. For example, consider the following three cases:

- Variant A: Excess capacity has been installed when the link was constructed (e.g. in the form of dark fibre). Such excess capacity can be brought to live with little additional costs, which may not at all depend on distance. In this case, the initial distance-related costs may have been

[3] Economic theory show that the loss of efficiency from recovering costs from customers is minimised when customers are charged according to their demand elasticity. By recovering a higher proportion from customers more insensitive to the increase in price (Ramsey pricing), the impact of price on consumption is minimised, and the outcome is closer to the efficient level of consumption.

relatively high (e.g. the cost of cabling would have been higher given the redundant capacity), but as additional capacity causes little or no additional cost, the impact of distance on the cost of capacity expansion is negligible.

- Variant B: The link was designed with future expansion in mind. For example, rather than simply burying the cable, ducts were installed, allowing the addition of future capacity at relatively little additional cost. In this case, the initial distance-related costs would have been higher than strictly necessary in the short term, but the impact of distance on the cost of adding capacity is relatively modest.
- Variant C: The link was built by simply burying cable. Unless methods of extending capacity such as multiplexing different communications over the same link can be used, installing additional capacity will require almost complete replication of the initial build. Minimising distance-related costs for the initial build implies that additional capacity can only be added at relatively high cost.

In variant A, incremental costs of adding capacity are small and do not vary much, if at all, with distance – though the cost of installing spare capacity has to be written off in case that an anticipated increase in demand is not being realised. Although most costs of the are fixed with respect to capacity they are still distance related. Variant B has higher distance-related incremental costs of adding capacity to existing links, as distance will have an impact on the cost of links added to the ducts. Variant C has the lowest unit costs for the initial chunk of capacity, but the operator incurs significant cost when adding capacity becomes necessary, with the impact of distance on incremental costs if adding capacity similar to the initial build stage.

The fact that distance-related costs and capacity-related costs may be traded off against each other by installing excess capacity of ducts rather than simply burying cables raises the question of what particular build scenario should be used in order to determine the appropriate level of costs, and what capacity measures should be used in order to determine appropriate levels of costs. A regulator would want to take a sufficiently long-term view of costs in order to give appropriate incentives for investment.

2.2.2 Sunk investment and inefficiencies in network topology

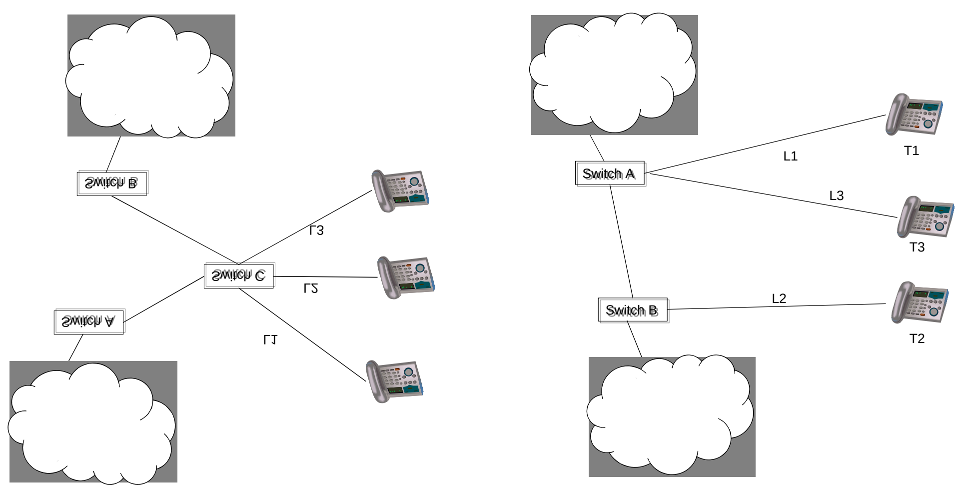
In practice, the presence of sunk costs implies that networks are not necessarily optimally configured. Rather, because of sunk costs, historical factors are important, as we cannot assume that network topology is optimised at each and every point in time. Network topology and the capacity of key network infrastructure (links and switches) are determined at least in part by expectations about future traffic. These expectations will generally prove to be incorrect, in which case network design will not be optimal ex post given outcome network traffic. The actual network topology is likely to diverge from the optimal design that would minimise overall network costs given current traffic volumes because the location of nodes is largely fixed, and it is often cheaper to upgrade capacity on existing links than to build new links and nodes.

This implies that costs in any real-life network are likely to be higher than those that would be obtained from a bottom-up engineering model. Moreover, it is likely that distance gradients are steeper than would be the case in an optimally configured network because, owing to the fixed location of nodes, some of the links may be longer than would be the case in an optimally configured network, and some of the routing may be inefficient (i.e. pass through more switches, or be carried over longer distances) relative to that in an optimally configured network.

For example, consider the impact of adding a further terminal point T3 to an existing network with Terminals T1 and T2 connected to Switch A and Switch B respectively (see Figure 2-1). Assume that the optimal network configuration for such a network would be Network I, but that given the existing links L1 and L2 it is more cost-effective to connect the new terminal to Switch B, thus producing Network II. This means that the cost of routing traffic to and from the new terminal is more costly than it would be in the optimal network configuration, and that the distance between T3 and the rest of the network has a greater impact on costs than would be the case in an optimally configured network.[4]

Figure 2-1: Inefficient networks

Network I | Network II



[4] Note that only if, owing to the distance between T3 and the nearest existing switch it would be more costly to build L3 rather than reconfigure the entire network would it be appropriate to do so.

2.2.3 Historic versus forward-looking costs

In practice, interconnection charges are often set by reference not to the costs actually incurred by the provider, but those suggested by an engineering model of an efficient network. However, costs may be higher than a model of network planning under perfect foresight might suggest as result of both historical inefficiencies caused by unforeseen events and the need to build in spare capacity as insurance against higher than expected demand. This raises a variety of complex issues about how risk and uncertainty should be dealt with, in that efficient building of real-world networks may need to make use of legacy assets and to take account of likely future demand.

At the same time interconnection charges cannot necessarily be set with regard to actual costs incurred, as there is no guarantee that these costs have been efficiently incurred, even taking account of the problems of sunk investments, stranded assets and future uncertainty. As a result the informational requirements for setting efficient access prices are severe.

2.3 The relationship between distance and cost

The relationship between distance and costs is strongly affected by network topology. Network design is affected by trading off various types of costs[5], and is conditioned to a great extent by existing sunk investment and capacity available on the existing network (discussed above).

In an optimally configured network, costs of switching and costs of conveyance are traded off against each other with the aim of minimising the total costs of traffic handling (albeit mixed with other objectives, such as achieving a certain level of resilience and considerations of service quality).

2.3.1 Sharing of costs

Given the presence of costs that vary with distance, but not with capacity (such as civil engineering costs), one should expect distance gradients in an optimally configured network to be shallower than they would be in a fully connected network (i.e. a network where there are links between each terminal).

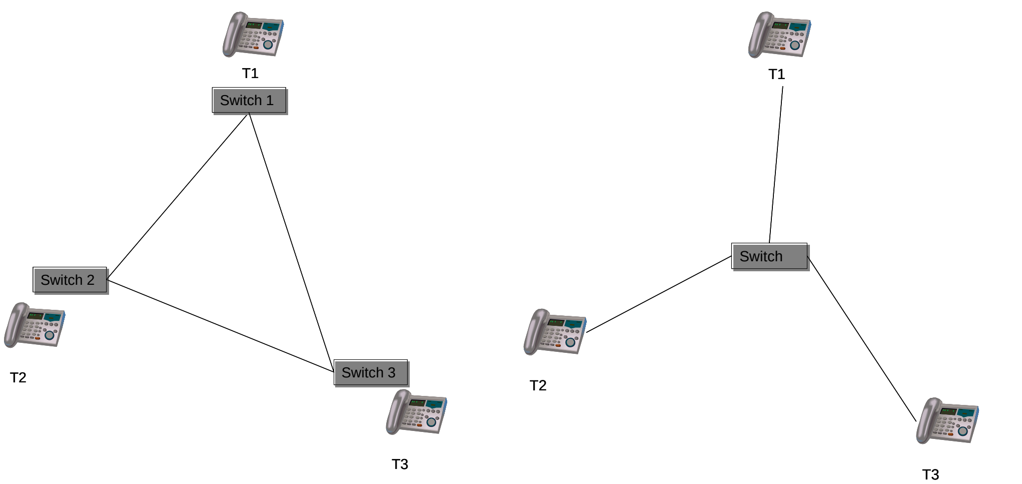
Although aggregating traffic on links results in longer physical paths for traffic, it may result in lower costs overall and lower unit costs for traffic conveyance. For example, one should expect the cost of Network I in Figure 2-2 to be greater than the cost of Network II, where the number of links and switches has been rationalised. In addition, as the cost of links in Network II can be shared over a

[5] For example, the optimal size of service areas in the FCC HCPM model is determined by considering the impact of reducing the number of wire centres on the average length of connections with end users. The FCC HCPM includes a 'clustering algorithm' which considers that reducing the number of serving areas (customers connected to a wire centre) implies that the average distance of customers from the wire centre increases, thus increasing the cost of cabling and associated structures.

greater volume of traffic (all traffic coming and going from each terminal), one can expect the impact of distance on the cost of the call to be smaller.

Figure 2-2: Sharing of periphery links

Network I	Network II

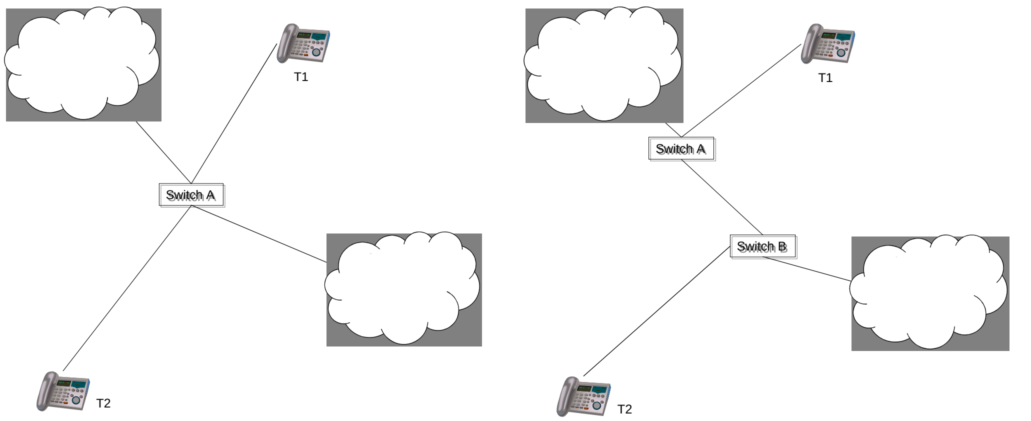


Broadly, one would expect distance gradients to decrease with the extent to which capacity on particular links is being shared across different users and uses, which in turn depends on the degree of aggregation at nodes and the location of this nodes. For this reason, adding nodes in a hierarchic fashion allows a greater degree of sharing of links at the core of the network, as traffic from shared links is further aggregated when going up a hierarchical level. Thus, one should expect the distance between the two Terminals T1 and T2 to have a greater impact on cost in Network III in Figure 2-3 than in Network IV, where the capacity (and cost) of the link connecting Switch A to Switch B is shared between all the terminals connected to the network, exploiting economies of scale on this link (and economies of scope across different connections within the network).[6]

[6] Note that the cost of additional switches in Network IV (each of which may have less capacity than the switch in Network III) must be lower than the cost savings obtained from sharing the link between switch A and switch B across a larger volume of traffic provided that the network is optimally configured.

Figure 2-3: Sharing of core links

Network III	Network IV

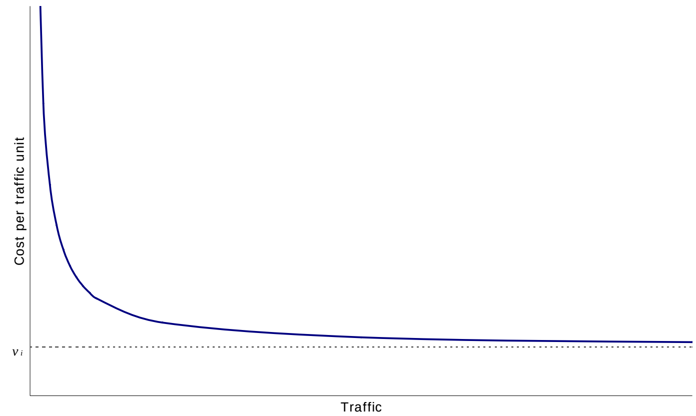


Scale economies from aggregation of traffic result from the fact that a number of cost elements can be shared across all traffic using a particular link. Therefore, the additional cost savings from sharing a given link over greater traffic volumes can be expected to fall with traffic already supported on the link. We can expect traffic unit costs to be related to the volume of traffic on a network segment to be as shown in Figure 2-4. For large traffic volumes, the share of common costs allocated to each traffic unit is negligible, and thus the unit cost of traffic becomes constant with respect to additional traffic over the link.

It is useful to define the *unit cost volume elasticity*. This is the proportionate amount by which unit cost falls for a given proportionate increase in traffic. For small traffic volumes, this elasticity is large in magnitude, as increases in traffic cause unit costs to fall rapidly. However, the elasticity becomes increasingly close to zero as traffic volumes increase, reflecting the fact that unit costs become increasingly independent of traffic volumes as divisibility problems and the importance of fixed costs diminish.[7]

[7] For example, suppose that there is a fixed cost F and a variable cost v. Total costs are $F+vQ$ for traffic Q. Therefore, unit costs are $v+F/Q$, which fall as traffic increases. It can be shown that the unit cost volume elasticity is given by $-F/(F+vQ)$. For small traffic volumes, this is approximately -1. As Q tends to infinity, this elasticity tends to zero.

Figure 2-4: Per traffic unit cost



Within switched or routed networks, greater aggregation of traffic at the core of a network than at the periphery means that distance can typically be covered more cheaply. However, due to the nature of scale economies from aggregating traffic over shared links, this also means that the scope for reaping cost savings further from accommodating additional traffic is more limited at the core than at the periphery.

2.3.2 Scale economies in traffic conveyance and structure of cost-distance relationship

We can explore the relationship between cost and distance through the following thought experiment. Consider communications traffic between two given points. Consider the case of a voice call, in which case we can think of these as originating and terminating customers' terminals.[8] Initially, consider the traffic is being carried over the incumbent's network from origin to end. However, we can also imagine the traffic being passed off the incumbent's network to an alternative network at some intermediate point (X) between origin and end. Ignoring any costs of establishing a point of interconnection, we can ask how much of the cost of carrying traffic between origin and end has been incurred in carrying it to the intermediate point of interconnection X. By varying the distance between the origin and the point of interconnection X, we can examine how the costs of carrying the traffic between origin and end divide across the route.

In the case of a simple point-to-point link between two terminals, there are some costs associated with setting the link up (regardless of distance) and other costs associated with the distance of the link. In the case of a wired link, these distance-related costs are significant and can be expected to scale approximately

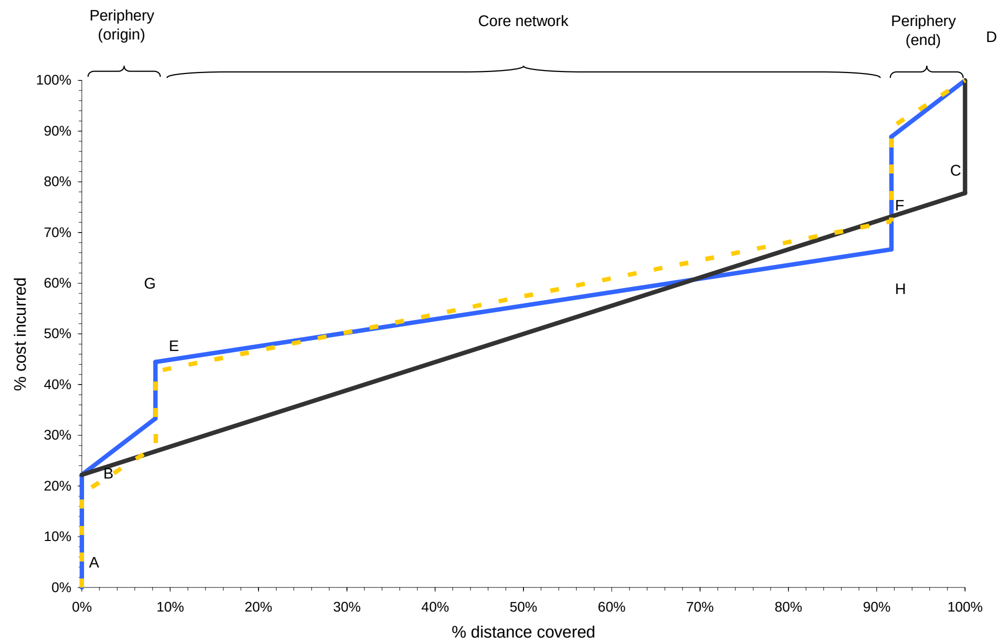
[8] In the case of broadband data, these might be the customer's terminal and the point of interconnection with the Internet backbone.

linearly with distance. Therefore, we would expect to see a cost-distance relationship such as the line ABCD in Figure 2-5, which provides the proportion of the total cost of conveying traffic between the originating terminal A and the end terminal D as we move the point of interconnection X further away from the origin.

Within switched or routed networks exploiting scale economies on links, there is a characteristic “inverse-S” relationship between the cumulative proportion of cost incurred so far on a route and the distance travelled.[9] This occurs because incremental costs per km are lower in the core network than in the periphery as there is greater aggregation of traffic on links. However, to achieve this aggregation it is necessary to incur additional fixed costs along the path, as switches are required to combine traffic on to higher capacity links. Therefore, we would expect to see a cost-distance relationship such as the curve ABEFD in Figure 2-5, where the slope of the cumulative cost with respect to distance changes as the transmission goes through different network nodes; in particular the slope is shallower closer to the core of the network. In the figure we show the proportion of cost incurred in relation to the proportion of distance covered, and therefore the figure does not reflect difference in the total cost of conveying a communication; instead, the figure reflects the impact of network topology on the distribution of cost across different parts of the network.

[9] In the case of broadband data, where users carry their own costs for getting to the network, only the first half of the relationship would be relevant.

Figure 2-5: Cost - distance relationships



The inverse S-shape of the distance gradient is very typical of a network: the slope is shallowest in the centre and steepest at the ends. This is a reflection of exploitation of scale economies in the core network being greater than at the periphery, where less traffic can be aggregated over links. An efficient network uses a topology where traffic at the core can be concentrated on high capacity links and then distributed on using switches. Where traffic is dense on a route, the unit per-kilometer cost is relatively low. Conversely, closer to customers, it is not possible to exploit scale economies; in the extreme case, the local loop is dedicated to just one customer. Thus, costs per kilometer to be higher at the periphery than at the core.[10] This loss of scale is typically progressive as we get closer to the customer and so we would expect, for example, traffic to be less dense on local-tandem links than on inter-tandem links.

2.3.3 Distance gradient: terminology used in this report

Throughout, we adopt the approach of ignoring the local access segment and concentrate on those segments of the network that are shared by different users. We can make this assumption because the move to an NGN does not

[10] It would be possible for the distance gradient at the core to be steeper than at the periphery. In this case, the additional level of aggregation provided by the core network might still reduce overall costs, as the distance gradient present at the periphery would increase if aggregation at core nodes were suppressed and substituted by a number of individual links (this is the effect shown in Figure 2-2).

affect the local loop in the sense that it remains dedicated to particular customers. This approach greatly eases the presentation.

We can distinguish two different aspects of the relation between distance and costs:

- We can consider the average cost per km over the entirety of the network, including links between terminals both near and far. This is the total cost divided by the average distance traversed. Any reductions in network costs clearly reduce this average per km cost. We call this the ***average distance gradient***.
- On the other hand, we can also consider the relative costs of carrying traffic over different segments of the network. To the extent that *incremental* costs of covering distance are different for different parts of the network, the rate at which costs are incurred per km travelled varies and so the cost-distance relationship deviates from linear. Therefore, it is useful to consider the ***uniformity*** of the distance gradient across different network segments, i.e. how closely the cost-distance relationship tracks a straight line. As discussed, with a point-to-point network, the distance gradient would be uniform.

We use the term “distance gradient” broadly to encompass the entire relationship between distance and cost. However, this gradient varies according across different points in the network structure. The structure of costs over different segment of the network, and how it is affected by technological change, plays a major role in determining the incentives for infrastructure-based competition.

A curve such as ABEFD in Figure 2-5 represents how distance-related costs are incurred in a typical network with aggregation of traffic at its core. Consider a change in technology that leads to greater potential for scale economies in link capacity. If scale economies in the core network become *relatively* stronger than at the periphery, then we would expect the middle part of the curve to become relatively flatter (as in the relationship ABGHD), as the difference in incremental costs per km between periphery and core becomes greater; the cost relationship deviates further from the linear relationship ABCD, and there is a flatter ‘plateau’ in the cost-distance relationship related to the core network. A greater proportion of costs are related to getting to the core network in the first place and a smaller proportion related to traversing the core network once on it. The location of aggregation nodes also affects this relationship, as it will determine the extent of each network segment.

Throughout, we use the terminology that the cost-distance relationship is ***less uniform*** when the difference between incremental cost with respect to distance at the core and at the periphery are substantially different. Conversely, if we say that the cost-distance relationship is ***more uniform***, there is more similarity in costs per km across the entire network, in which case the ‘plateau’ in the cost-distance relationship at the network core is relatively steeper. Referring to Figure 2-5, the cost-distance relationship curve ABGHD is less uniform than ABEFD, as we move further from a straight-line relationship between cost and distance.

Using this terminology, we can see that distance might become less important in absolute terms where distance-related costs fall. However, there is also a separate question about whether even given a lower overall level of distance-related costs, we have a less uniform cost-distance relationship, so that once traffic has penetrated sufficiently into the core of the network there is little further additional cost of carrying a little further.

2.4 Current distance gradients for interconnection charges

Looking at current interconnection charges for different types of services provides useful validation of our conclusion about relative distance gradients in the core and periphery.

2.4.1 Distance gradients in BT's current interconnection products

We have derived the implied relationship between cost and distance for hypothetical communications by decomposing the costs incurred in carrying traffic according to which of BT's interconnection products would be required to carry it a certain distance. Details of this exercise can be found in Annex A.

Current (regulated) narrowband interconnection charges and their underlying costs demonstrate the characteristic inverse-S relationship, where the incremental cost of additional distance is smaller at the centre, particularly when the local access segment is taken into account. This is consistent with our expectations about the cost-distance relationship within switched networks.

Unlike narrowband, broadband interconnection charges do not have the same regulatory requirement to be cost reflective, nor to publish cost data. This limits the extent to which interconnection charges can be used in order to assess how costs change with distance. Nevertheless, the relationship between interconnection charges and distance for Datastream products (which use virtual paths) is less uniform than for narrowband PSTN products. Assuming that the underlying costs follow a similar relationship, this would then suggest that the difference in costs between core and periphery (resulting from greater scale economies from traffic aggregation at the core relative to the periphery) is more significant for broadband communications than for narrowband. This is intuitively reasonable given the greater ability of broadband traffic to use shared capacity rather than requiring dedicated or virtual circuits.

2.4.2 Distance gradients in commercial tariffs of other services

It is difficult to find direct analogies to distance gradients for PSTN and broadband services that have been determined entirely by competition rather than regulation. Nevertheless, Internet peering arrangements and mobile interconnection provide some weak analogies.

In both cases, there is typically just one charge for interconnection. In the case of Internet interconnection, this has often been zero (traditionally through peering arrangements, though there are now other models in use). In the case of mobile networks, there is a call termination charge regardless of the distance that a call must travel on the mobile network. Therefore, in both cases there is an incentive for an interconnecting operator to hand over traffic as early as possible.

Therefore, commercial analogies appear to demonstrate more averaging across network elements and less uniform distance gradients (i.e. single fee regardless of distance) that in the case of regulated interconnection charges. This is unsurprising as such networks are not subject to unbundling obligations and compete end-to-end. Therefore, they are likely to want to benefit for scope economies across provision across different network elements and have little incentive to offer unbundled network elements.

2.5 Distance gradients, investment incentives and potential inefficiencies

The distance gradient affects the attractiveness of different entry strategies and, in particular, infrastructure-based entry. Incentives for alternative network entry depend on both the level and structure of access charges. Efficient entry depends on the relative costs of the incumbent and entrant for each network segment, which provide the potential for new entrants to save costs by extending their own network infrastructure into that segment.

It is a generally acknowledged principle of regulatory policy that regulated prices, such as interconnection charges, should be cost reflective. Aligning prices with the costs of providing a particular service ensures correct incentives for efficient use of existing network infrastructure, and appropriate investment signals for the construction of new networks (both by the regulated firm, and those making use of the incumbent's infrastructure): appropriately set interconnection charges should provide incentives for provision of alternative infrastructure where it is efficient.

However, it is seldom the case that there is an objective and uncontroversial measure of incremental cost. Given this, it is appropriate to judge the efficiency of access prices at least partially by regard to the outcomes they achieve.

There are three main problems that may occur with the pricing of interconnection that may lead to being insufficient incentives for alternative networks:

- the general level of interconnection prices may be too low, in which case it may be cheaper for competitors to use the incumbent's network even if actual costs could be lowered by building own network infrastructure;
- the distance gradient could be too shallow, distorting choices about the extent of alternative networks and providing insufficient incentive to extend from the core network to the periphery;
- the structure of averaging could lead to inefficient entry in some geographical areas and insufficient entry elsewhere.

On the other hand, interconnection charges could be too high. This would lead to excessive incentives for alternative infrastructure investment in situations where it were not socially efficient.

The *level* of interconnection charges is not the focus of this study. Ofcom has already undertaken extensive processes to determine interconnection charges for both narrowband and broadband services. Rather our concern is primarily with the *structure* of interconnection charges, especially where there are a number of related interconnection products.

2.5.1 Structure of interconnection charges and entry strategies

Where there are a number of interconnection products available, alternative providers will take into account the relative prices of these products when deciding whether to purchase network services from the incumbent or else self-provide. Therefore, interconnection services (at least where these include common network elements) may be substitutes; lowering the price of one may lead to switching of wholesale demand from other interconnection products, as entrants will chose between various possible modes of entry.

We are interested in the cost to alternative operators of interconnecting with the incumbent's network in order to provide communication services to their customers and how this may affect their infrastructure investment decisions. It is useful to distinguish between two different situations in which alternative providers may require the incumbent's interconnection products:

a) when providing core services between incumbent's (or other competitors') subscribers, where the entrant does not provide local access services;

b) providing communication services between entrant's subscribers and incumbent's subscribers, where the entrant provides local access services to some end users;

Assuming entrants have access to the same (or better technology) than that used by the incumbent, any cost advantages from the incumbent operator source from the greater traffic volumes carried on the network, which allow for further exploitation of scale economies. For this reason, toe-hold entry starts at the core, where minimum efficient scale is smaller relative to traffic than at the periphery. This is the 'ladder model' of investment, according to which entrants first build their alternative networks at the core, and then extend them towards the periphery as they gain sufficient market share.

Some simple examples show how distance gradients can affect investment incentives. We use two examples that are already well documented, although with the result that the data discussed below is a few years old and may not necessary be a good description of current market outcomes. We assume here that the total end-to-end cost of conveying traffic is known, and concentrate on how the impact of distance on this cost is reflected in interconnection charges.

First, the use of UNE-P in the US provides a simple example of substitution between various access-based entry strategies where incentives for alternative provision of core network may arguably have been adversely affected. UNE-P consists of a bundle of network services to provide end-to-end access, including access to the local loop and onward conveyance and switching. It has grown to be the most important access product for alternative providers seeking to access the local loop. It is cheaper to purchase UNE-P than to purchase the individual network elements it implicitly comprises. Although the UNE-P problem arises because of bundling of access to the core and periphery network, it corresponds to a flattening of the distance gradient.

Alternative providers require access to local loops as this is the least contestable aspect of the network. The extent of alternative infrastructure provision in the core network is then determined by the relative cost of taking an interconnect product that provides access to both core and periphery (i.e. UNE-P) or using access to the periphery and self-building the core. Given the price of UNE-P relative to the charges for individual network elements, there is an incentive for alternative providers to purchase UNE-P to benefit from the implicit bundling discount relative to buying its component network elements separately, even though it might be that they only need local loop access and could more efficiently provide switching and transit themselves. In effect, UNE-P has not only replaced resale, but crowded out other access products, and has thus had a detrimental impact on investment in alternative infrastructure which might otherwise have occurred.

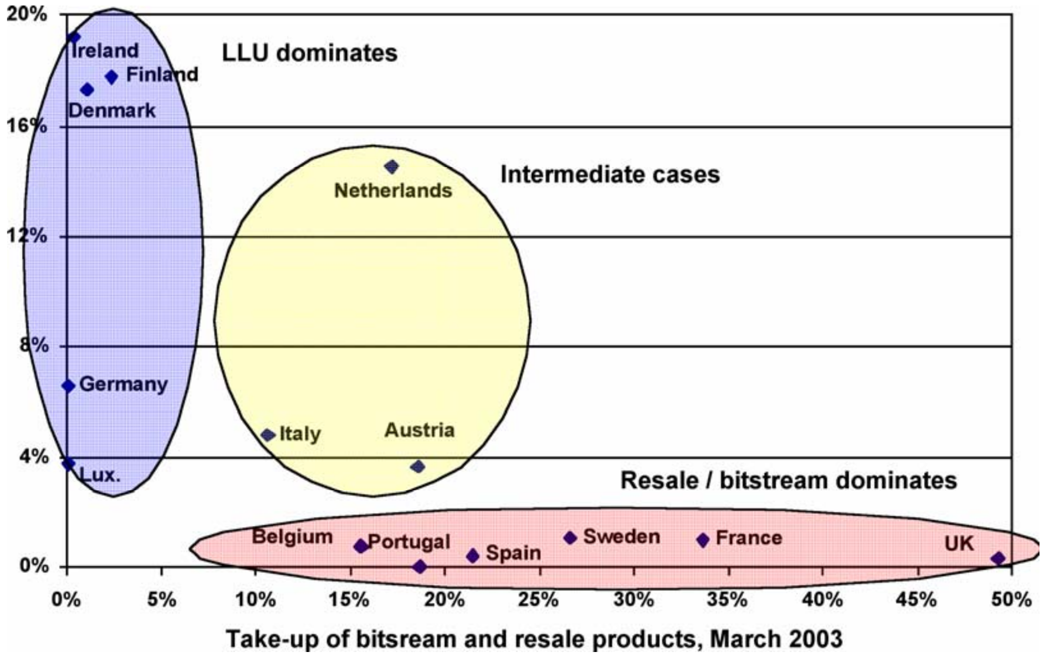
In this case, insufficient incentive to invest in the core network occurs because the cost of access to both core and periphery is too cheap relative to the cost of access to just the periphery. In general, incentives to build alternative core network infrastructure are depressed if the distance gradient become less uniform.

However, the effect is not limited to core network investments. The more conventional worry is that access regulation might lead to insufficient incentive for provision of alternative local loop infrastructure. According to this view, alternative providers are likely to enter with core networks, as this is the most contestable aspect of the network, but have a choice as to how far these alternative core networks extent.

In this hypothetical example, if there is already some degree of competition at the core network level, then alternative providers then have a choice whether to extend self-provision closer to the end customer, ultimately replicating local loops. Clearly this choice depends primarily on the costs of access to the local loop. Raising this price, so flattening the distance gradient, increases this incentive. Therefore, incentives to extend the network from core to periphery are smaller the more uniform the distance gradient.

A second example for the potential for such effects is provided by looking at the use of bitstream access for broadband as opposed to using unbundled local loops, which we might interpret as extending the network closer to the customer. Looking across the EU, there is some evidence of polarisation of the strategies adopted. In some cases this may be due to particular countries pursuing one model of access more vigorously, though in most cases both forms of access have been available for some time. Therefore, it is reasonable to suppose that this polarisation is due at least in part to differences in relative pricing of the two access products and alternative providers choosing between them.

Figure 2-6: Comparative take-up of LLU and bitstream access



Source: Figure 18 of Maldoom at el (2005) "Broadband in Europe", Springer, New York.

These two illustrative examples demonstrate an important general point: although it is clear that a flatter distance gradient discourages investment – which is unsurprising given that it reflects stronger scale economies in some parts of the network – it is not necessarily obvious where investment is discouraged. In particular, what differentiates the two examples is that in the case of UNE-P, the relatively low interconnection charge for core interconnection services might discourage entrants from building their own core network infrastructure, as if access services need to be purchased from the incumbent anyway, then the *additional cost* of also acquiring bundled core services may be small relative to the cost of building own infrastructure. This becomes a problem when competitors face a decision on whether to build their own core network infrastructure.

In contrast, in the broadband example, the relatively low cost of periphery interconnection services may prevent infrastructure-based competition in the core network from extending to the periphery. Operators who already own core network infrastructure may be unable to gain the traffic volume at the periphery that would make building their own periphery infrastructure cheaper than acquiring periphery interconnection products from the incumbent.

As these examples show, when the cost of core interconnection products is low relative to the cost of periphery interconnection products, there may be little incentives for entrants to build (or use) their own core infrastructure. However, provided entrants *already* have core infrastructure, there are incentives to extend infrastructure to the periphery in order to avoid the relatively high periphery interconnection charges. On the other hand, more uniform distance gradients (other factors equal) tend to increase the potential for entrants to benefit from building their own core network infrastructure. However, this situation might prevent from infrastructure-based competition extending to the periphery.

2.5.2 Impact of EBC on infrastructure investment incentives

At present, BT narrowband interconnection charges are set in reference to the number of network elements used, following the Element-Based Charging (EBC) approach. Under this approach, interconnection costs are averaged within each network segment, and distance is only explicitly considered in some instances by using banding within network segments (e.g. short, medium and long inter-tandem conveyance). The reason for this is that differences in cost between different network segments are currently more important than distance-related differences.

In the context of EBC, the distance gradient measures the magnitude of the increase in cost to alternate telecommunications providers as they interconnect at higher tiers of the BT network, i.e. when they make use of the more substantial BT wholesale products. In this respect, a uniform distance gradient indicates that cumulative price increases substantially as providers utilise higher tiers of the network in generating a connection for their customers, thus it results more expensive to use the incumbent's network for longer routes. Conversely, in the case of less uniform distance gradients, the cumulative price to alternate providers for interconnecting at different tiers of the network does not vary significantly, and thus distance has a small impact on interconnection charge paid.

The EBC approach results, on average, in a positive relationship between distance and interconnection tariffs, as typically more network elements are required for covering greater distances. However, there are some limitations to this, resulting from averaging of prices within network tiers.

Averaging prices across categories has the advantage of providing a much simpler tariff scheme, which facilitates interconnection deals and regulation. However, this somewhat limits the extent to which the tariff can reflect the true cost of the interconnection service, as communications over links which are longer than the average will thus be subsidised while communications over links shorter than the average are over-charged.

Due to geographical averaging of links, there might be insufficient infrastructure investment in areas where links are longer than average, as alternative networks can access the interconnection product at a price which is lower than the cost of building that segment. Similarly, in areas with links shorter than the average, where interconnection tariffs are greater in relation to their true costs, there might be an incentive to build alternative infrastructure even when it might be inefficient to do so; for example, even if the cost of the service using an alternative network were greater than the cost of the service using the PSTN, this could still be lower than the interconnection tariffs due to the surcharge resulting from tariff averaging. Note that the same reasoning can be followed if geographical averaging limits the extent to which area-specific factors are reflected in the cost structure, for example due to lower use if there are fewer or lighter users than average. However these are out of the scope of these study.

2.6 Concluding remarks

The distance gradient in the cost of conveying traffic depends on the degree of sharing of link costs. Due to greater aggregation at the core of the network relative to the periphery, the cost of conveying traffic an additional distance is typically cheaper at the core than at the periphery. Current distance gradients for regulated interconnection charges reflect this cost structure.

The distance gradient of interconnection prices affects the attractiveness of different entry strategies. When interconnection prices are cost-reflective, there are efficient incentives build alternative network infrastructure. If interconnection charges are set above or below their cost, investment of alternative infrastructure may take place where it is inefficient.

As the magnitude of the unit cost-volume elasticity decreases with the traffic volume already aggregated, cost differences due to differences in traffic volumes can be expected to be smaller at the core than at the periphery. Therefore, one can expect the differences in costs between the incumbent and entrants using similar technologies to be smaller at the core. This yields the traditional 'ladder theory' of investment, where toe-hold entry starts at the core (where minimum efficient scale is smaller relative to the periphery and entrant disadvantages from smaller scale relatively less severe) and may be later extended to the periphery. Given this:

- other factors equal, a more uniform distance gradient across network segments increases the incentives for entry of alternative core networks, as the incumbent's charge per kilometer at the core is relatively high given the potential cost savings that could be gained from aggregation of traffic at the core;
- other factors equal, a less uniform distance gradient provides a cost structure where the *incremental* cost for an alternative core network of using the incumbent's access services at the periphery is *relatively* high, and so provides greater incentives for vertical integration (i.e. extending own network to the periphery) where entry at the core exists.

Therefore, there is a tension: a more uniform distance gradient encourages core network competition; a less uniform distance gradient encourages the extension of core competition to the periphery, but at the cost of less incentive to enter at the core in the first place.

3 Differences between an NGN and traditional networks

In this section, we review the NGN plans proposed by BT, and then identify the main implications of moving to an NGN structure. In summary, moving to an NGN entails the following general effects:

- A common protocol means less differentiation between traffic generated by different services and greater ability to launch new services. The use of an end-to-end packet-based network allows different services and users to share capacity dynamically. This increases the efficiency with capacity is used and deferring the need to build new capacity;
- Simply bringing separate networks together may lead to scale economies even without any topology changes. For example, there may be savings in maintenance costs, network planning and equipment procurement;
- Further cost savings may be achieved by changes in network topology. Aggregation nodes can move closer towards end-users. Greater aggregation of traffic sooner in the network structure allows more exploitation of scale economies on links within the periphery of the network. In effect, the periphery becomes more like the core;

There may also be a centralisation of network intelligence, resulting in cheaper local nodes but less direct routing

3.1 BT's plans

BT set out plans for its intended NGN, 21st Century Network (21CN), for consultation at the beginning of 2005. BT's 21CN project consists in merging all of BT's existing networks into a single multi-service network. BT estimated[11] that the investment required for this would be around £10bn over five years, and would reduce operating costs by £1 billion per annum by 2008/09 through increased economies of scale from the converged network.

BT's initial 21CN design would fundamentally change the capability of nodes at different levels of the network hierarchy, limiting the availability of interconnection points between alternative providers and BT. In addition, BT's plans contemplate changing the number and location of many of the nodes in its current networks.

3.1.1 BT's current networks

BT currently offers a number of products for voice telephony (including data over voice lines from dial-up connections) and broadband data. These services are carried on a number of different platforms, involving separate networks and a number of different communication protocols. In this setup, network sharing across services is limited to the core for those services that use compatible protocols.

[11] Press release, June 2004.

In the case of residential services, i.e. voice telephony and broadband, traffic from both services initially shares the local segment of the network up to the first node (local exchange), where traffic from different services is split and routed onto different networks. Broadband traffic is routed onto the ATM core network and typically to the destination ISP, while voice traffic is routed onto the PSTN, where it is routed to its final destination.[12]

A feature of the current PSTN network is that traffic routing occurs both at local nodes and core nodes. This allows for interconnection between BT and alternative service providers both at local exchanges and tandem switch nodes.

3.1.2 The proposed 21CN

With the proposed NGN setup, traffic from the customer would be carried over to the MSAN (multi-service access node, co-located with the Main Distribution Frame in BT's 21CN plans) and from there directed to a core network node (metro node). Thus, under the proposed 21CN setup, traffic originated by customers would not only share the local network segment, but would be *aggregated* at the local node using a common transmission protocol and from there routed towards a common multi-service network.

The advantage of the 21CN MSAN relative to local nodes in current existing networks is that it supports many different access technologies and allows services using different technologies to be carried on the same lines. Therefore, sharing takes place not only in parts but across the whole network (including the network segment covering from the local nodes to the core network). The local node of 21CN is substantially simpler relative to local nodes in current networks and its simplicity represents a major cost saving.

The 21CN plans also involve changes in topology, with different connectivity and functionality of nodes at the different hierarchical levels. The new topology is to consist of three tiers:

- the first tier will consist of around 6000 sites containing MDFs and MSANs (local network);
- the second tier will be made up of approximately 120 metro nodes (periphery network); and
- the third tier will be made up of about 10 core nodes (core network).

The Metro Node defines the core edge of the proposed new network. It directs traffic to the core or performs a turnaround of local traffic. Each Metro Node is connected via diversely routed fibre to at least two Core Nodes.

[12] An overview of BT's current PSTN and broadband network can be found in Annex B.

The Core Node differs from a Metro Node by virtue of increased functionality and connectivity. There are two types of nodes at the core, Inner and Outer Core Nodes. They have identical functionality with the Inner having improved connectivity; each Outer Core Node is connected to at least three Inner Core Nodes while all Inner Core Nodes are fully interconnected. This contrasts with current network topology, where all main tandem switch nodes are interlinked.

Node sites in the core network (i.e. Metro and Core Nodes) may also contain iNodes. These contain voice related equipment such as call servers.

Based on their proposed network topology, BT plans to date are to offer physical unbundling (LLU) at the MDF/MSAN nodes and interconnection at metro nodes only, although different options are still being discussed.

3.2 General network implications of moving towards NGNs

BT's 21CN plans involve a number of differences between their proposed NGN and current traditional networks. From these, we identify a number of potential differences between NGNs and traditional networks likely to be relevant. The main features of NGN relative to traditional networks appear to be:

- the use of a common protocol across services, which in turn implies a lower degree of differentiation between communications with respect to service type;
- the greater aggregation of traffic, resulting in more intensive use of network and greater scale economies;
- the move of aggregation nodes towards the customer (or rather, the absence of a split of the signal of different services at local nodes);
- potential efficient geographical re-location of nodes (for example moving core aggregation nodes closer to customers if higher traffic on the multi-service core network can be achieved earlier); and
- potential centralisation of routing capabilities to the core of the network (where before there was intelligence both at core and local nodes), in turn resulting in:
  - cheaper local nodes and cheaper access to the core network (as costs of routing equipment at local nodes can be suppressed); but
  - less direct routing, with greater distances required for local communications (local calls cannot be re-routed at local nodes but need to be directed to core nodes before being routed back).

  Overall, centralisation of routing leads to a lower degree of differentiation between communications with respect to distance as the path taken by traffic is less determined by where it originates and terminates.

We can distinguish two categories of changes:

- those involving a greater degree of network and capacity sharing; and
- those involving changes in network topology.

We consider these two categories in turn.

3.2.1 Network and capacity sharing

The move towards a common packet-based data transmission protocol could (at least in principle) be achieved through upgrades of the network components at nodes, without necessarily affecting the location of transmission links or network topology. However, even without any changes in network topology, NGNs can make a more capacity-intensive use of existing links and lower unit costs. This occurs through two mechanisms:

- the use of a packet-based data transmission protocol that allows for capacity sharing through statistical multiplexing[13] (where previously transmissions may have required exclusive use of a dedicated line or a virtual circuit) which allows accommodation of greater traffic volumes without requiring additional capacity; and
- the combination of different services (previously delivered through different networks) onto a single platform that allows for greater exploitation of economies of scale so reducing unit costs.

Operators are already taking advantage of statistical capacity sharing in some parts of their networks (mostly at the network core). The move to an NGN would provide further opportunities for capacity sharing across different users and services resulting from end-to-end packet-based transmission across all parts of network except the local loop. Capacity sharing (across both users and services) can smooth usage peaks that drive requirements for capacity.

The greater sharing of infrastructure allows for greater exploitation of scale economies. For instance, bring two separate platforms together into one can have an impact on common network costs by rationalising fixed network expenses (e.g. network maintenance).

Providing that traffic from different services is sufficiently homogenous (rather than requiring different quality of service), capacity can be allocated to each service dynamically. For this reason, there may be little need to attribute network costs to individual services, but rather similar charging models can apply to all types of traffic. Therefore, the use of common transmission protocol results in smaller scope for differentiation of traffic (and interconnection charges) depending on the type of service that generates it.[14]

Although sharing capacity creates economies, in a multi-service network there may be increased dangers of unforeseen network congestion to some extent and

[13] In any link, the capacity required depends on the expected usage, and more importantly, the expected peak usage. As the number of users and services sharing this link increases, peak demand becomes more predictable provided individual usage demands are at least in part uncorrelated.

[14] Given that traffic from different services uses the same transmission technology, differences in the price for different services could lead to arbitrage from operators purchasing capacity from cheaper services and using it for transmitting data from more expensive services. Therefore, a common underlying protocol leads naturally to a common notion of capacity for all service types and common pricing according to the capacity demands that services make.

so a greater need to build ahead. Although the variance of traffic becomes smaller due to the law of large numbers, rare events may still cause traffic peaks and these may be large in size even if more improbable. For example, a new application (say a peer-to-peer file sharing craze) could unexpectedly and suddenly increase broadband traffic. This could congest all services on the multi-service network.

This suggests a greater need for pricing of services in relation to the stochastic demands they make on common network capacity (i.e. peak-load pricing). Further, this may require schemes for managing service priority given the different needs of different services (e.g. video/audio streams requiring low latency vs. email). We do not consider quality of services charging models in this paper.

3.2.2 Network topology

At least in principle, there is no automatic need for changes in the geographical layout of the network, as MSANs could be installed at current local node sites and Metro nodes at current tandem switch sites. Therefore, we can at least hypothetically consider a move to an NGN were network topology remains broadly similar to the existing one. Transmission links between nodes are to be largely reused, and thus the move to NGN would appear to mainly affect equipment at nodes.

For this reason, geographical changes in network topology should only be expected to take place if they are expected to result in net cost savings. For example, there could be a substantial rationalisation of nodes and links, which can be expected to result in lower total costs.[15] In addition, it may be efficient to re-locate aggregation nodes, as the current network layout may not be optimal for a number of reasons, for example:

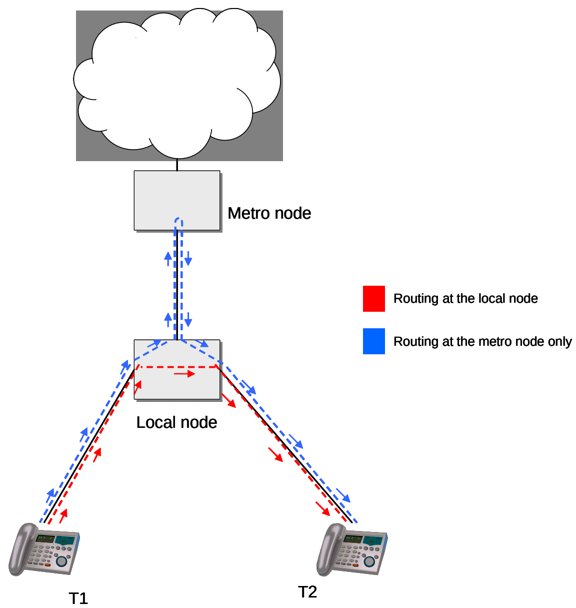
- due to the greater aggregation of traffic resulting from integrating traffic from different services one can expect the nodes in an optimally designed network to be closer to the end users (aggregating traffic from a smaller number of users but increasing the proportion of shared links overall);
- due to the sunk nature of network costs, the location of nodes may result from historical legacy, rather than an optimal location given current traffic patterns and costs (for example, it may be cost effective to move nodes further away from urban centres due differences in the land costs). Although such changes in topology may have not been cost-effective in previous years when current networks were already fully functional, it may be efficient to undertake some of these network upgrades in conjunction with the NGN upgrade.

[15] If the rationalisation of nodes is optimal, cost savings from not upgrading and using suppressed nodes are larger than any additional costs for extra transmissions between nodes and potentially longer routes. In addition, if the routing capability of local nodes is suppressed (as we discuss below), then a number of local nodes historically used only or mainly for routing local calls may become redundant.

On the other hand, due to changes in the functionality of the equipment at nodes, there could be changes in hierarchical tiers within the network and the level at which routing of transmissions takes place, which would result in the number of interconnection products decreasing (from three-tier to two-tier, a fusion of call origination with local to tandem conveyance).

The changes in network hierarchy levels and in particular of moving the routing of traffic further in towards the core of the network has the effect of increasing the distance over which local transmissions need to be carried. This is shown in Figure 3-1, which represents the path over which a telephone call is carried in order to reach a terminal connected to the same local node as the originating terminal both in the case that the local node had routing capability and in the case where routing takes place at the Metro node.

Figure 3-1: Impact of moving routing towards the core of the network



Overall, centralisation of routing capabilities result in longer routes for some calls and results in communications being carried over greater average distances. This reduces the difference in traversed distance over which calls with different destinations have to be carried over. For example, the average distance over which traffic between two terminals connected to the same local node is carried is in the new setup equal to the average distance over which traffic between two terminals connected at different local nodes if the two local nodes share the same primary metro node. Similarly, there will now be less differences between the distance over which the traffic is carried in all communications, as a greater part of the transmission path (the link between local and metro nodes) is common to all communications.

However, the lack of efficiency in routing is traded off against cost savings in the equipment at network nodes, which may offset additional costs from increasing the distance travelled by communications. Thus, as we discuss below, centralisation of routing could result from changes in the distance gradient of the cost of communications becoming flatter for the additional segment over which communications are to be carried, in this case for the MSAN to Metro link segment.

4 Distance gradients for NGNs

In this section we investigate how moving to an NGN may affect network costs and, in particular the link between distance and interconnection costs. In order to do so, we consider each change identified in the previous section one at a time. Once the individual partial effect of each change is addressed separately, we consider what the overall effect is likely to be. We conclude that overall the distance gradient tends to be smaller in absolute terms, and more uniform across shared network segments.

The decision to move towards NGN is taken in a partially competitive environment, and thus we consider to what extent competitive interaction may affect this conclusions. We conclude this section with a brief consideration of the likely impact of the move to NGN on competition of infrastructure and other implementation issues.

Throughout this section we distinguish three network segments: the local access loop (the link going from the end user to the local node), the periphery (including the local node and the link from the local node to the core node) and the core of the network (including core nodes and inter-core links).

This Section provides a *qualitative* assessment of the likely impact of the move to NGN on costs and their relationship with distance. All the figures in this Section are provided as a visual aid, and based on examples rather than on actual data.

4.1 Impact of NGN plans on costs

The move towards NGN is expected to bring substantial cost savings for operators, by increasing usage and providing the opportunity to exploit scale economies further. This can be expected to flatten the distance gradient: total network costs become smaller relative to traffic volumes and thus average network costs associated with each traffic unit decrease. This is the result of a combination of effects.

In order to assess the potential impact of structural NGN changes to the network, we analyse the potential cost implications of the effects of the move towards NGN identified in Section 3.2, namely of:

- increasing network and capacity sharing between users and services; and
- potential changes in network topology.

We then address the potential impact of changes in the unit cost of network elements, and to what extent can this affect the relative per traffic unit cost of different network segments, and thus the distance gradient.

4.1.1 The initial cost-distance relationship

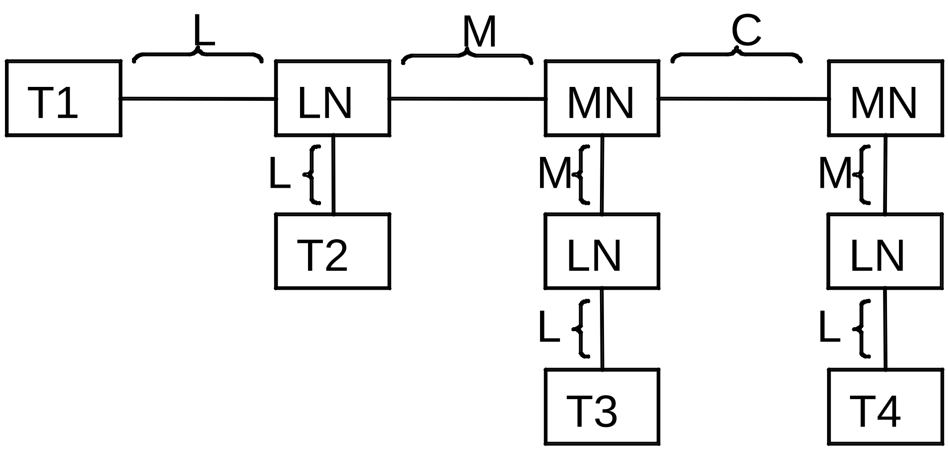
Following a simple network structure with two levels of aggregation, we identify the different elements of the network for which the cost of traffic can be expected to differ:

- local access link (L);
- local node (LN);

- local to core link (M);
- core node (MN); and
- core link.

These are represented in Figure 4-1. Here T1 to T4 represent terminals.

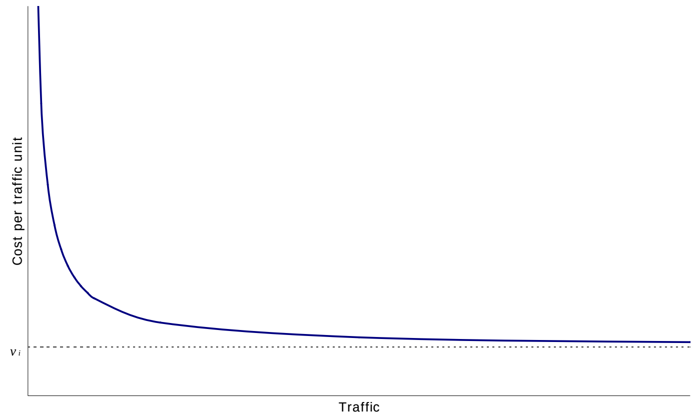
Figure 4-1: Network elements



The local loop is largely unaffected by NGN, except for overall traffic increases that may reduce unit costs over this segment but leave total costs of the link to the user unchanged. For this reason, we concentrate our analysis on changes in the other elements of the network.

Network elements typically have both fixed costs of infrastructure and constant costs per traffic unit utilising that element. In this case, the cost per traffic unit can be expected to decrease with traffic, asymptotically tending towards the variable cost of traffic (v_i), as explained in Section 2.3.1. This is shown in Figure 4-2, which reproduces the relationship between total traffic volumes and unit costs already presented in Section 2.3.1. It is helpful to bear this figure in mind throughout this Section.

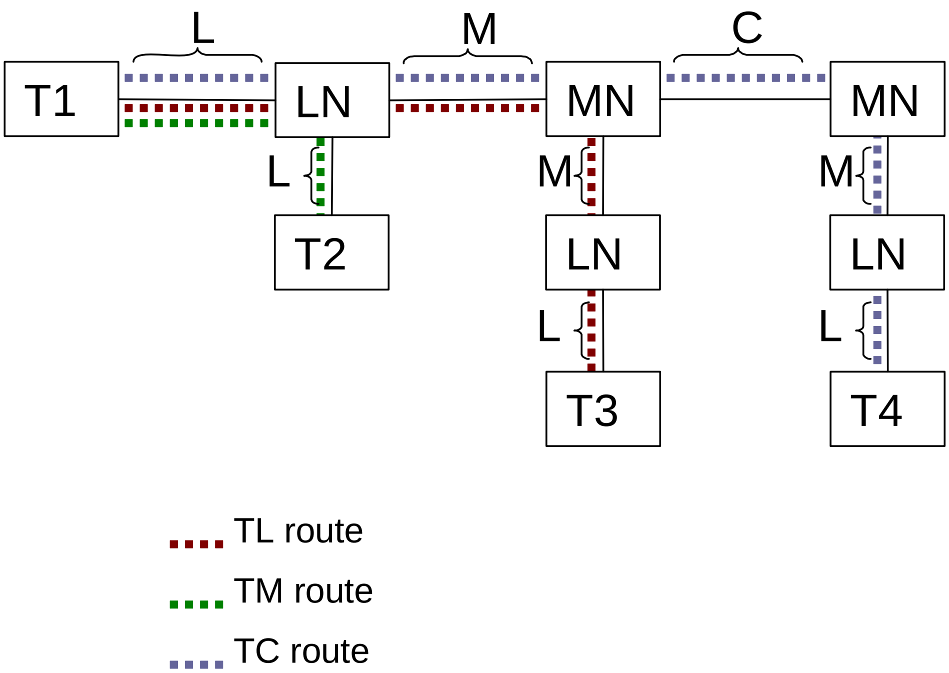
Figure 4-2: Per traffic unit cost



Depending on the type of communication that generates traffic, not all the network elements may be used. This is represented in Figure 4-3:

- calls between terminals connected to the same local node (e.g. T1 and T2 in the figure) could in principle (assuming there is routing capability at local nodes) be provisioned without need to go through any of the M, MN and C elements – we denote this type of traffic TL;
- calls between terminals connected to different local nodes which are both connected to a common core node (e.g. T1 and T3 in the figure) could in principle be provisioned without need to go through the C element – we denote this type of traffic TM;
- last, calls between terminals connected to different local nodes which are not connected to the same core node (e.g. T1 and T4 in the figure) and need to be routed though the core network, and thus use all network elements – we denote this type of traffic TC.

Figure 4-3: Types of traffic and their use of network elements

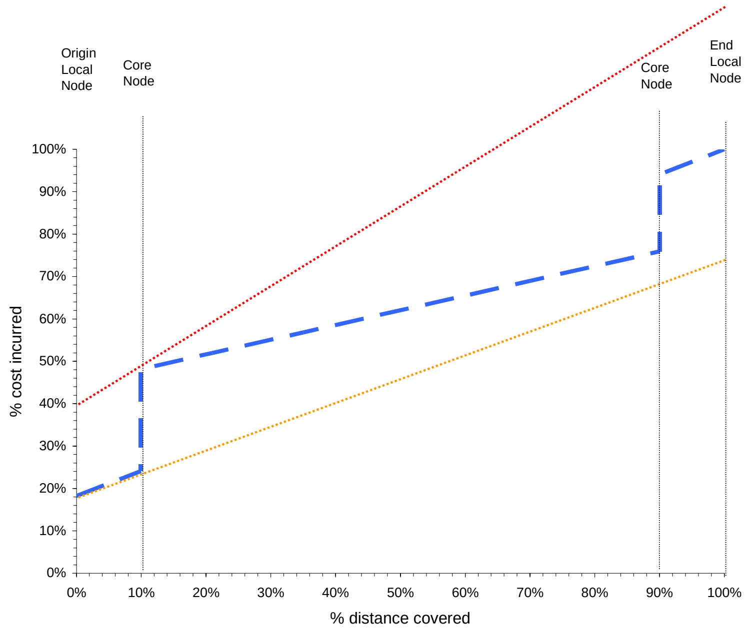


The decrease in overall per unit costs is expected to offset any increase in the cost of components involved in aggregating to the next level of network hierarchy, as otherwise aggregation to a higher level would not be cost effective. This is not directly apparent from the cost-distance relationship figures. For example, Figure 4-4 shows the cumulative LRIC over distance for a long distance call over the average long double-tandem conveyance distance.[16] One could wrongly conclude that overall network costs can be reduced by extending M up to the end LN, bypassing the core network altogether (MN and C), as shown by the dotted orange line in Figure 4-4. However, this would only be possible if the distance gradient for M were unaffected when bypassing the core. In reality, bypassing the core network would mean that now traffic could not be routed to other end nodes, and therefore link M would become dedicated only for traffic going to each particular end node (LN), which would result in greater per unit costs due to costs being shared over a smaller traffic volume (as for example the dotted red line); this would need to be accompanied by building additional M

[16] In the figures we show the cost-distance relationship for communications between terminals connected to the same network. However, the graph can be interpreted for all types of communications, the only difference being where the call would end its journey on this network. For example, for broadband internet data communications, the final destination would usually be located within the core segment.

links to each destination node. Therefore, we cannot necessarily consider one particular route for traffic without considering other traffic.

Figure 4-4: Cumulative cost over distance (excluding local loop)



4.1.2 Network and capacity sharing

Using IP technology allows for a more intensive use of links and for different services to be run over the same network, where before different core networks used different transmission protocols. Therefore, the move towards a common IP technology allows:

- increased usage on shared links, and
- increasing the proportion of shared links (thus reducing dedicated links).

The benefits of statistical capacity sharing are greater as the sources of traffic sharing the network become more various, which has the effect of reducing the variance of the total demand for capacity. In addition, sharing of network assets across services is expected to reduce both overall fixed network costs and the per traffic unit cost of equipment at network nodes, and provides flexibility in using capacity for different services, as less spare capacity required to meet potential demand shocks for individual services.

In addition, consolidation of networks can substantially reduce costs. In absence of any cost savings from network consolidation, i.e. if the two networks were run in parallel, one could expect unit costs to be equal to the weighted average of the standalone costs of each network. However, integrating different services onto a single network may result in an overall decrease in cost due to:

- there may be scope for reducing some common costs that may scale up with the number of independent networks used rather than with the utilisation of the network (e.g. maintenance costs, duplicate links);

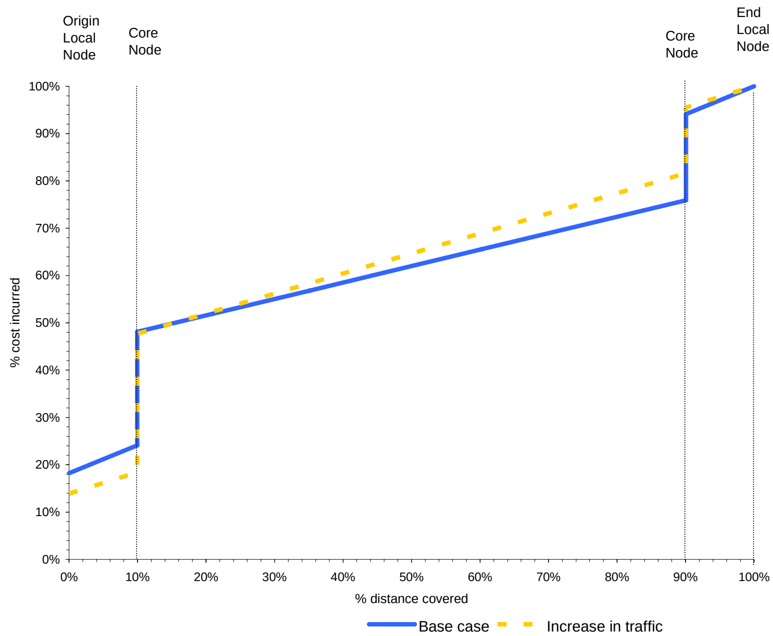
- if different services were using different technologies, one can expect traffic to be integrated using the most efficient or cheaper technology.

Overall, the effects discussed above result in an increase in capacity and usage across all segments, thus shifting the cumulative cost curve down. However, the effect of combining different networks into a single multi-service network depends on the potential cost savings that network consolidation can provide to each segment. One would expect the benefits from consolidation to be felt more strongly in those parts of the network where economies of scale were only mildly exploited to date and thus in segments closer to the edge of the network, rather than those segments where economies of scale are already strongly exploited even before the move to an NGN. Thus, relative to the costs of broadband traffic (which have a structure closer to NGN than PSTN), the reduction in cost can be expected to impact significantly more at the periphery than at the core. This is because of two reasons:

- Given an even increase in traffic throughout the network, ***there may be a greater scope to benefit from further scale economies at the periphery than at the core***. If we imagine services being carried on two separate networks with similar topology being combined onto a single network, there will be benefits from the greater exploitation of scale economies. This is true even without changes in network topology. However, reductions in unit cost could be more significant at the periphery of the networks where unit cost-volume elasticities are likely to be greater (in absolute value) as traffic is less aggregated. Conversely, aggregation of traffic at the network core is already significant even before the networks are combined. In other words, unit costs in the periphery are likely to be higher up in its cost-traffic relationship (shown in Figure 4-2) than unit costs at the core; therefore, the impact of an increase in traffic on cost is likely to be greater at the periphery than at the core. Therefore, there is a relatively smaller decrease in unit costs from further aggregation than at the periphery.
- In addition, ***the increase in traffic due to aggregation is likely to be greater at the periphery than at the core***. First, unlike broadband internet traffic, a large proportion of voice traffic (local calls) can typically be routed without requiring an extensive use of the core network (either by routing it at local nodes or by routing it at core nodes without requiring the use of inter-core links). This increase in traffic would thus impact on the periphery but not necessarily on the core. Second, IP technology provides a more flexible routing for communications and may spread traffic more evenly throughout the network.

Therefore, periphery costs may fall relative to core costs, which would lead to greater uniformity of the cost-distance curve. This effect is represented in Figure 4-5, where the proportion of cost incurred relative to the proportion of distance covered is represented, assuming that the reduction of costs is reduced proportionally greater at the periphery than at the core. (The figure does not capture the effect on the average distance gradient because it is scale-free and does not reflect overall costs.)

Figure 4-5: Impact of an increase in traffic on the cumulative cost over distance (excluding local loop)



Another consequence of the adoption of a common data transmission technology is that the differentiation in charging models between different services (e.g. differentiation between voice and bitstream transmissions) may become unsustainable. This is because as different types of communication use the same underlying transport layer, interconnection products designed for different services may be substituted and thus price differentiation would lead to arbitrage between interconnection products[17].

4.1.3 Topology changes

In this section, we assume that operators will only change topology if network topology changes to result in cost savings for the operator undertaking the network upgrades. Competitive interaction and the potential for topology changes being aimed at increasing the cost of competitors (rather than reducing the cost of the incumbent provider) are addressed below in Section 4.2.

As discussed above, changes in network topology are not a necessary requirement for moving towards NGN setup, but rather arise from the possibilities that the new structure provides for increasing efficiency of the network and for reducing overall costs. We distinguish here between three potential changes in network topology:

[17] There may be potential limitations to this if different services require different levels of quality of service or if there is policing of the nature of traffic carried.

- rationalisation of nodes, where the benefits from combining different networks might allow to override a number of nodes and links;
- geographical relocation of nodes; and
- changes in the routing capabilities of different nodes, affecting the potential routes for traffic.

The first type of change stems from the combination of services onto a single network, and is one of the factors that would contribute towards the cost reduction – the greater the proportion of nodes that can be suppressed, the greater the cost reduction. Thus, the effect of these changes will be similar to that described above, depending mainly on the relative impact across the network hierarchy of combining network traffic previously on separate networks. It can be expected that the scope from rationalising network nodes will be greater where capacity is not fully used. There is less exploitation of economies of scale. Therefore, the effect of network rationalisation is consistent with our view that cost savings are likely to be more significant in the periphery than at the core.

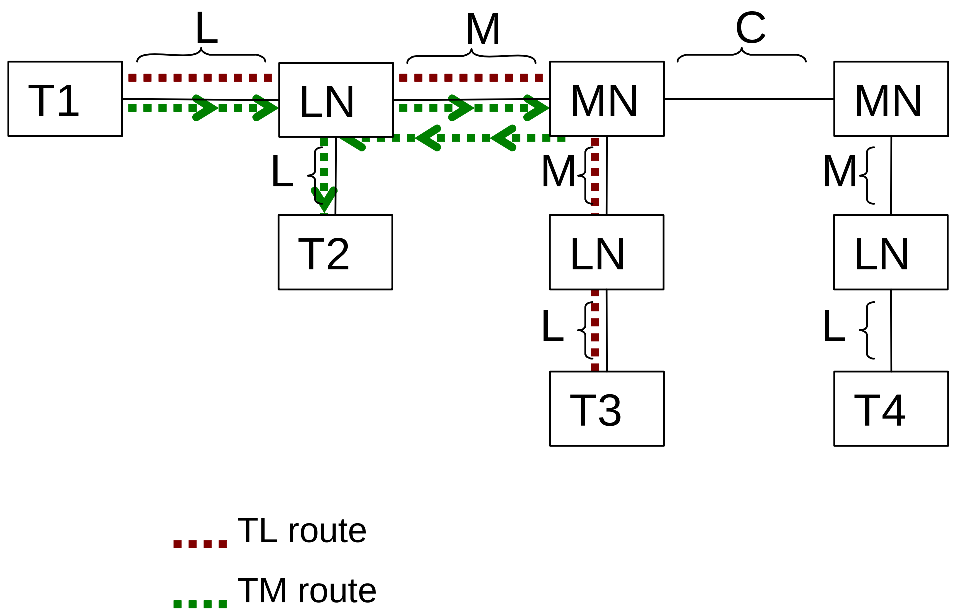
We now consider the second potential change, where aggregation nodes may be moved closer to end users. As overall traffic on links increases from potential new services or greater sharing across users and/or services, it may be possible to achieve the critical traffic mass for which it is beneficial to build an aggregation node closer to the end user. In this eventuality, it would be efficient to relocate aggregation nodes closer to the edge of the network. We can see that in general combining a number of networks into a single multi-service network should make it desirable to locate aggregation nodes closer to end users. However, in practice, existing sunk investments mean that such changes will only occur if the cost savings are sufficiently large.

Relocation of aggregation nodes closer to end users would imply a smaller concentration ratio at relocated nodes (thus potentially reducing scale economies at that level), but might allow for substantial cost savings from reducing the average length of lines with a smaller degree of aggregation. At the same time, the relocation of nodes is likely to increase the total distance over which the transmission are carried without any real increase in the distance between terminals, because the path over which the communication is routed typically becomes less direct the greater the proportion of shared links. This might to some extent counter the cost reduction moving aggregation closer to the edges.

Last, and the most relevant change in network topology proposed by BT in its 21CN plans, we consider suppressing routing capability at the local nodes. The main reason for this change is that expected equipment cost savings at the local node are expected to offset any additional costs from TL having to be routed to the core node and back to the originating local node before reaching its final destination.

In fact, now TL becomes equivalent to TM with respect to usage of network segments, as shown in Figure 4-6.

Figure 4-6: Equivalence between TL and TM when routing is centralised at MN



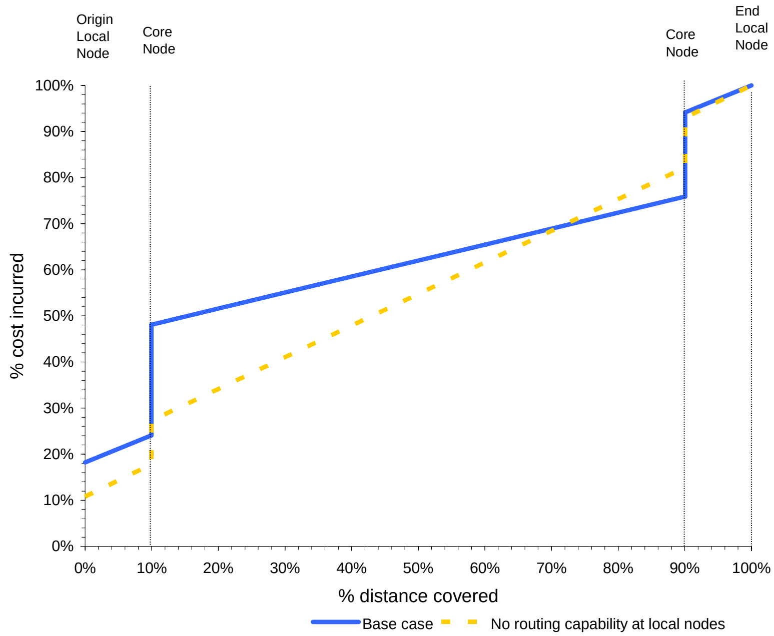
This has an impact on the traffic of all segments except for local access links and core links:

- traffic at LN increases as TL traffic now goes through this node twice;
- traffic at M increases as TL, that previously did not require using this segment, goes back and forth on these links;
- last, traffic at MN increases from TL traffic reaching this node in order to be routed back to the originating LN.

Therefore, both types of traffic now become equivalent and require use of the same network elements to the same extent.

With centralisation of routing, access to the core becomes cheaper. The suppression of routing at local nodes involves lower costs for the equipment at this node, reducing the cost per traffic unit for using this element. In addition, the increase in traffic in the periphery (due to TL traffic making more extensive use of the network) would result in lower costs per traffic unit in the periphery (as each traffic unit can be allocated a smaller share of fixed costs). Therefore, the distance gradient at the periphery falls both in absolute terms and relative to the distance gradient at the core. The effect on the proportional cost-distance relation of lower cost of local nodes and of increased traffic on the periphery is shown in Figure 4-7.

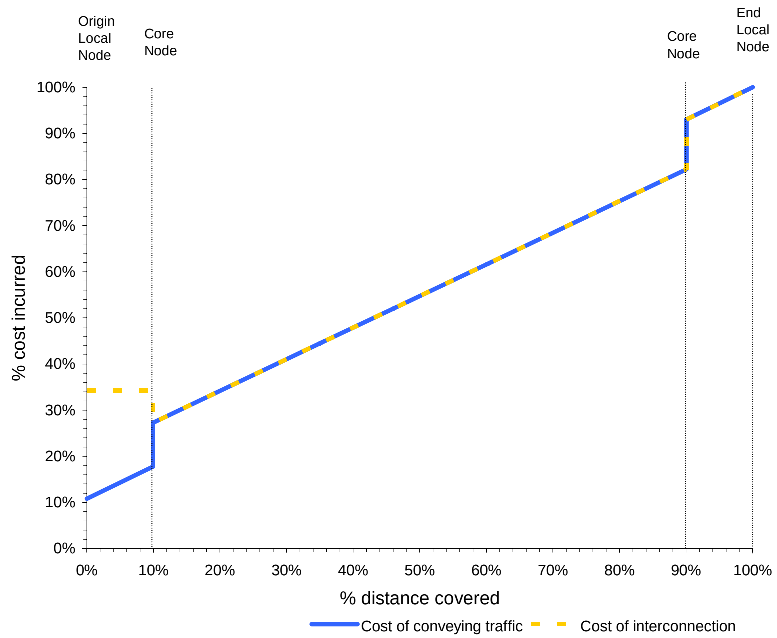
Figure 4-7: Effect of removing routing capability at local nodes on cumulative cost over distance (excluding local loop)



If interconnection at local nodes were offered, such interconnection would entail greater costs than interconnection at the core node (assuming it is not possible to route traffic directly at the local node), as this would involve costs of routing the call back to the initial local node. In this case it would be cheaper to interconnect at the core node than at the originating local node. This is shown in Figure 4-8: although the cost-distance relationship has the inverse S-shape exhibited by the blue line, the relationship between the cost of interconnection and distance would be the one exhibited by the yellow dotted line.

If routing at local nodes were not offered, then the relevant part of the cumulative cost structure becomes more uniform, as the local to core segment would not be available and there would be only one segment (inter-core conveyance) for which the distance gradient would apply.

Figure 4-8: Effect of offering interconnection at local nodes on cumulative cost of interconnection over distance (excluding local loop)



As discussed in Section 3.2.2, cost savings from suppressing routing functions are achieved at the expense of an increase in the distance over which TL is carried, potentially resulting in an increase in the cost for this type of traffic. This, combined with the effect of traffic which would anyway require access to the core (TM and TC) becoming cheaper, results in a lesser degree of differentiation between transmissions over the distance between originating and end terminals.

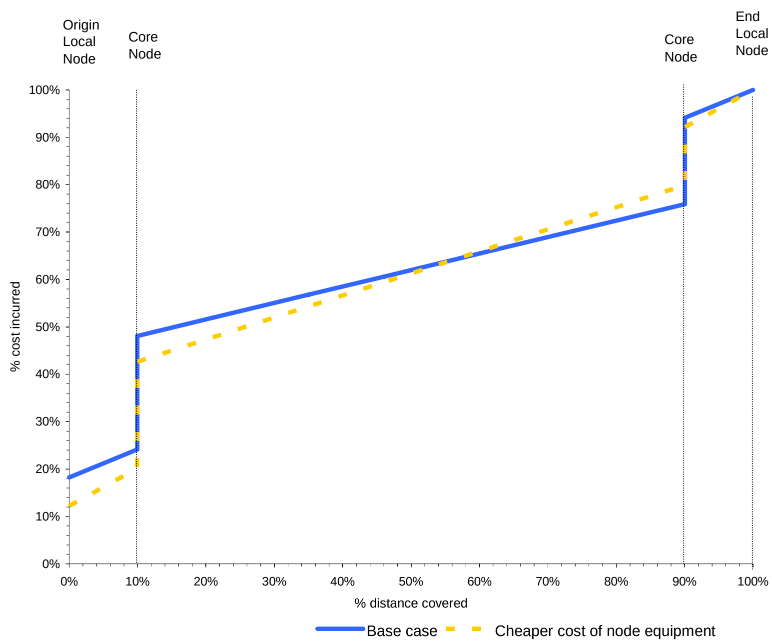
4.1.4 Impact of changes in the cost of network elements

Changes in the cost of network elements also affect the distance gradient. However, existing physical network links (e.g. fibre) are likely to be reused where possible, even if new links were to be built. Therefore, changes in the cost of equipment are likely primarily to affect the fixed costs at each node related to switching and routing, rather than the costs of links. Indeed, if we imagine a move to an NGN without significant topology changes, it would primarily be changes in equipment at nodes that would drive cost changes.

The relevant issue with regard to changes in the cost of equipment is the extent to which the *relative* cost of the equipment at nodes in different hierarchical levels of the network may change. However, there is no particular reason to believe that the unit cost of equipment at nodes closer to the edge would become *relatively* more expensive and accentuate differences in the distance gradient of different network levels. Indeed, changes in function are more significant in the periphery of the network and there are likely to be more legacy assets stripped away than in the core. Network intelligence may move in from the periphery to the core. Therefore, it would seem plausible that unit cost savings at nodes might well be greater closer to the edges of the network.

The effect of such a change would be similar to that of combining different networks onto a single multi-service network, discussed in Section 4.1.2. An example of this effect is shown in Figure 4-9.

Figure 4-9: Impact of changes in the cost of node equipment on the cumulative cost over distance (excluding local loop)



4.1.5 Summary

As we have seen above, the move towards NGN may involve a number of different mechanisms by which the cost structure is affected. All these effects accentuate scale economies and extend them towards the periphery of the network. Therefore, we expect the move to an NGN to:

- lower the average distance gradient in general; and
- make it more uniform between periphery and core.

The overall effect of flattening the cost on shared network segments, is to tend towards a cost structure where once the aggregation node is reached, the distance over which the communication is carried makes less difference to the overall cost of the communication. Thus, variable costs of traffic conveyance become more important relative to distance-related costs.

The more uniform cost structure reduces the need for differentiating interconnection charges across different network segments. This is because, as costs at the periphery are more similar to costs at the core, the unit cost impact of moving up a hierarchical level in the network is smaller.

4.2 Competitive interaction and distance gradients

Changes in the distance gradient resulting from upgrading networks to an NGN structure may affect interconnection charges. In turn, this may affect competition, and possibly feed back to affect the provider's costs and in turn the distance gradient.

We are concerned in the extent to which network changes may affect the options available to competitors. In particular:

- the move to an NGN may reduce interconnection options for alternative networks and the balance of incumbent's and alternative networks;
- changes in the distance gradient and its degree of uniformity may affect the nature of competition, as discussed in Section 2.4.

We first address the implications of changes in interconnection possibilities on competitors. Then we assess the likely overall impact of the move to NGN and the effect that this is likely to have on competition and alternative infrastructure investment.

4.2.1 Implications of changes in interconnection possibilities

The 21CN plans proposed by BT are likely to affect interconnection with alternative networks, both due to:

- relocation of nodes; and
- potential absence of routing capabilities at local nodes.

If there is geographical relocation of nodes (as opposite to reconfiguration of interconnectivity at current node locations), part of the current interconnection infrastructure of competitors may become unusable. In this case, competitors desiring the same degree of interconnection possibilities as before the move to NGN might need to undertake further infrastructure investment. Conversely, if competitors decided not to roll-out additional infrastructure, this would result in greater traffic onto the incumbent's network, potentially increasing scale economies and thus further flattening the distance gradient. This is unlikely to be a material consideration unless relocation of nodes affected a large proportion of the network.

The loss of routing capabilities at the local nodes impedes interconnection directly at the origin local node before being transferred to the core network. If communications cannot be routed at the local node, they are systematically conveyed to the core node before the destination of the transmission has an impact on the path it will follow. Although handing over the communication at the origin local node may still possible, this would require routing the communication back from the core to the origin local node, and would therefore be more costly than handing over the communication at the core node. Therefore, absence of routing capability at the local node may have a substantial impact on competitors' interconnection possibilities or costs.

No interconnection at local nodes

The points of interconnection available for handing over communications between networks of different operators affect the "divisibility" of network investment decisions, and thus impact on alternative network build-out.

Assuming interconnection were not offered at the local node, this would imply a smaller degree of divisibility.

Reducing the number of potential interconnection points along the distance over which communications are carried reduces the potential for using different networks for different segments. Therefore, the less interconnection points available the more constraints when selecting the path of traffic, as the range of alternative paths – in particular paths involving the use of different networks for different segments – decreases.

In order to illustrate this, we can consider a situation with interconnection products defined in a similar way as current BT's PSTN interconnection products, including local access. Suppose that an alternative operator were able to provide cheaper conveyance between local and core nodes, but more expensive for all other path segments, as shown in Figure 4-10. In addition, assume that the cost savings on conveyance between local and core nodes when using the alternative provider were smaller than the additional costs at the local access segments, resulting in the alternative provider having a higher cost for the aggregate path of local access plus conveyance between local and core nodes, as shown in Figure 4-11.

Figure 4-10: Example of build/buy choices, full interconnection possibilities

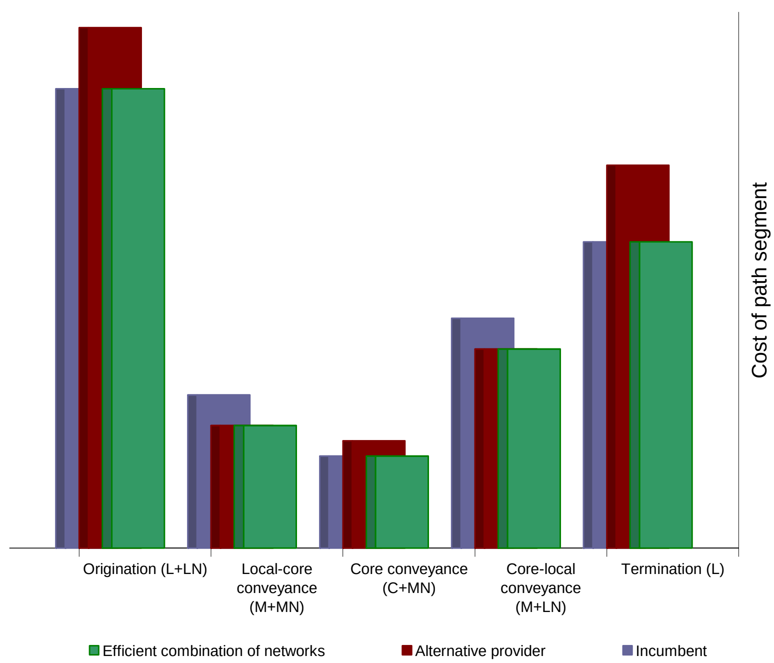
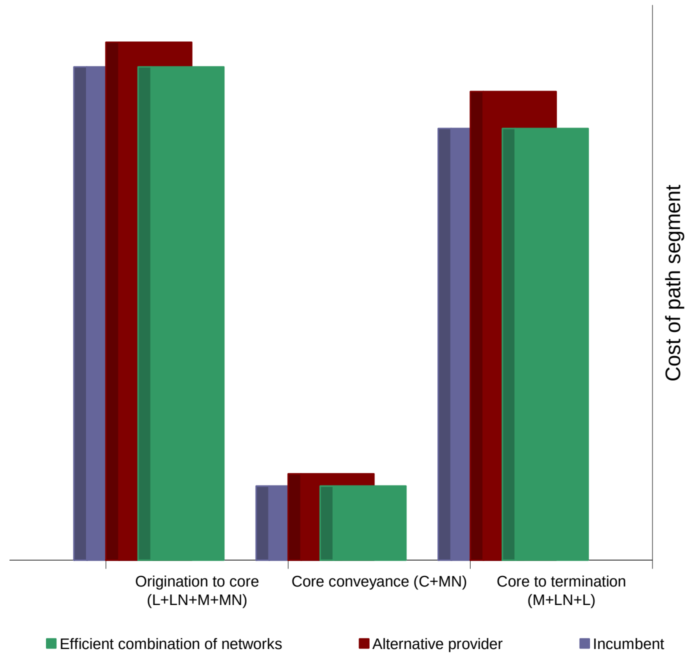
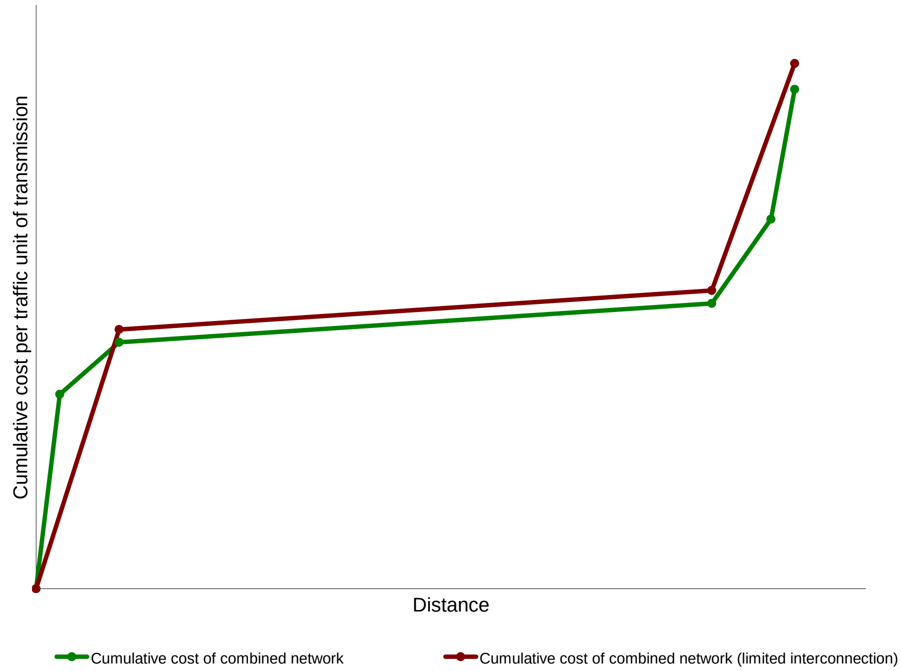


Figure 4-11: Example of build/buy choices, limited interconnection possibilities



Were interconnection available at the local nodes, the choice of building versus buying would take advantage of the cheaper option for each segment. However, if interconnection at the local nodes were not available, then the potential cost savings from using the alternative provider for conveyance between local and core nodes would not be materialised, resulting in a higher overall cost of transmissions relative to the case of full interconnection. This is shown in Figure 4-12.

Figure 4-12: Cumulative cost of combined network



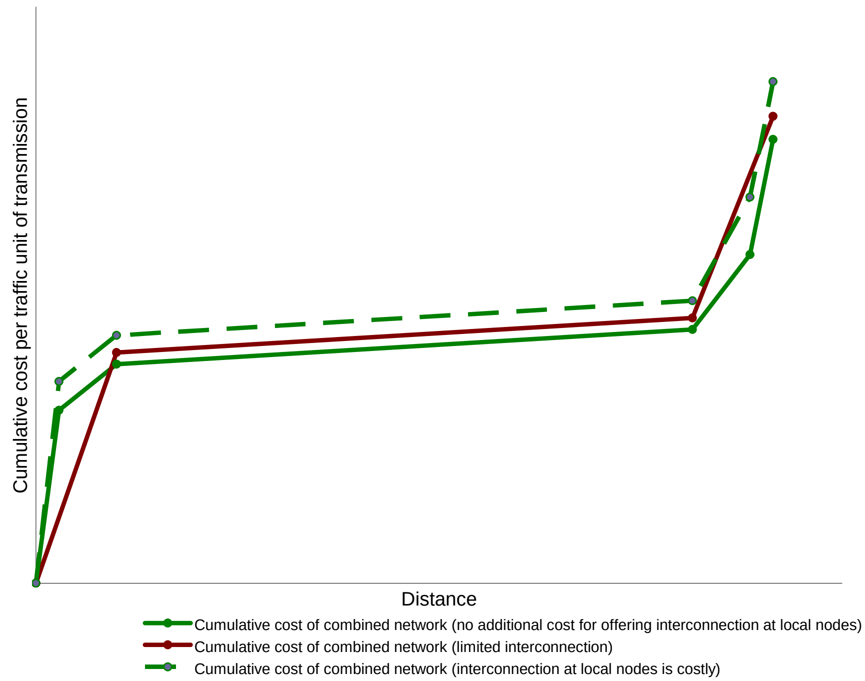
Thus, due to the absence of interconnection at local nodes, the mode of competition would tend towards competition for the whole length of the communication, with operators competing for carrying the communication from end-to-end rather than carrying it over specific segments.

However, the relative advantages of operators for *particular* segments are likely to become less relevant under the NGN structure. As the incumbent upgrades its network it becomes unlikely that entrants can benefit from potential technological advantages derived from using newer technology, and thus any cost differences are likely to derive mainly from differences in traffic volumes and operating efficiency differences. In this context it is less likely that we have a situation with different a cheapest provider for different segments (local loop excepted). Thus, the advantages of divisibility in the path of a transmission become less relevant under NGN, and the limitations from a having less interconnection points can expected to be moderate.

Costly interconnection at the local node

Providing interconnection at local nodes in order to have a greater degree of divisibility may result in higher cost of network components, as it would require routing capability at the local node. Therefore, the benefits of enabling greater flexibility on the point at which transmissions can be handed over between networks need to be traded off against the cost savings in the absence of additional interconnection points, as the increase in cost could offset benefits from greater flexibility. This is shown in Figure 4-13, where the provision of routing and interconnection at the local node is assumed to increase the cost of node equipment to the extent that benefits from using the alternative provider at the local node are forgone.

Figure 4-13: Cumulative cost of combined network when offering interconnection at local nodes is costly



This example shows only a static comparison of costs. When looking at the problem from a dynamic standpoint, one needs to take account of the potential impact on incentives to build alternative infrastructure and the extent to which lack of incentives might reduce future benefits from infrastructure competition. In our example, removing interconnection at the local nodes results in the alternative provider being more expensive in all segments, which might remove the incentive for the alternative provider to maintain and use its own network.

4.2.2 Impact of changes in topology on rivals costs

The fact that the move towards an NGN may affect the cost of competitors raises the concern that there might be incentives for incumbent operators to undertake some network upgrades and changes in order to raise rival's costs rather than reduce the provider's costs. For example, competitors may need to build additional infrastructure in order to accommodate to the new interconnection possibilities. This creates a risk that certain changes in the network may not be motivated by cost savings and be socially efficient, but rather be aimed at distorting competition.

An assessment of the potential for such distortions of competition is out of the scope of this study. The fact that most entrants have built their networks more recently than the incumbent operators suggests that most operators might be able to accommodate new transmission standards relatively easy. Difficulties would seem to involve primary sunk investments in alternative networks that might be stranded by certain changes occurring in the move to an NGN. In practice, regulatory action can be expected to mitigate such problems, for example through provision of legacy wholesale services during some transitional period.

Nevertheless, even if distorting effects were present, these would not necessarily affect our conclusions about the effect of the move to an NGN on the distance gradient, unless the potential damage that the incumbent could inflict upon competitors were sufficiently large as to motivate inefficient network upgrades that resulted in higher costs. Instead, actions aimed at raising rivals costs would tend to increase traffic on the incumbent's network and allow the incumbent to exploit scale economies in its network further, thus reinforcing the effects discussed above. Therefore, we do not see this issue as affective our overall conclusions.

4.2.3 Alternative infrastructure and bypass

The overall absolute fall in the distance gradient resulting from scale economies being exploited further means that distance loses importance as a driver of cost. If this is reflected in the structure of interconnection charges, once the fixed part of interconnection costs is paid for there is little additional charge for having the communications carried elsewhere in the network. Thus, relative to the current situation, alternative service providers would face less difference in the cost of using the interconnection services on the incumbent network for short or long distances. Therefore, the smaller average distance gradient reduces scope for competitors to reduce their costs by building their own infrastructure.

However, incentives for alternative infrastructure investment do not depend solely on the incumbent network's distance gradient. Rather, these depend on the *difference* between costs to competitors of using the incumbent network relative to costs of building out their own network.

We can consider the case of a competitor with costs similar to the incumbent's (i.e., similar cost-volume relationships, patterns of aggregation and topology) but lower overall traffic volume. Such competitor would face higher unit costs due to lower volumes. Supposing traffic volumes on the competitor's network remained constant, the cost reduction resulting from the increase in traffic on BT's network following the move to NGN would increase the difference in costs between the incumbent and the entrant.

In this case, assuming that there is more traffic aggregation and a smaller magnitude of the unit cost volume elasticity at the core than at the periphery, the impact of traffic volume differences between the incumbent and alternative networks on costs may be smaller at the core than at the periphery. Hence, the disadvantage of the entrant created by a move to an NGN would be relatively greater at the periphery than at the core, and we would expect relatively greater incentives to enter at the core than at the periphery as a result. This confirms that the impact of distance gradients becoming more uniform would be to reduce the incentives to invest at the periphery relative to incentives to invest at the core, as anticipated in Section 2.4.

However, assuming that entrants followed suit, launched multiservice offers similar to BT's and managed to increase traffic in a manner similar to BT, then both the entrant and the incumbent would benefit from greater exploitation of scale economies relative to the current situations. This would tend to narrow cost differences between the incumbent and the entrant. Both operators would lower their unit costs, but also their unit cost-volume elasticities (moving to the left side of Figure 4-2). This means that the difference in volumes between the entrant in the incumbent would become less important. Therefore, in order not to be left behind by the move to an NGN, alternative networks need to find traffic aggregation possibilities.

Last, changes in interconnection possibilities impact on the parts of the network that are relevant to competitors when making build-out decisions. In particular, reducing the number of points of interconnection reduces the divisibility of wholesale services, and the possibilities for using the most efficient network provider for individual network segments. This may limit the benefits from having access to different networks and reduce the incentives for alternative infrastructure investment in segments where it may have been efficient.

As changing the structure of interconnection charges has a different impact on the degree of competition in different network segments, it may affect the nature of competition. In the case at hand, suppression of interconnection at the local node would be equivalent to bundling the local access loop with the periphery segment, impeding the use of different networks for each one of these segments.

The effect of this could be substantial. However, cost advantages in different segments, and in particular between the local access link and the first aggregation layer of the network may be achieved through LLU. Indeed, if interconnection at the local nodes is suppressed, LLU access at the local nodes might become crucial feature for enabling the benefits of competition to materialise and be effective in the whole network rather than just the core.

The main difference between interconnection at the local nodes and LLU would appear to be the ability to route different types of communications onto different networks. However, traffic becomes more homogeneous due to data from different services becoming more similar under a single transmission protocol and differentiation of the cost of communications according to destination disappearing with flatter distance gradients and centralisation of routing. It is not unreasonable to suppose that this may suppress the relative advantages of different operators handling different types of data, and thus the advantages of directing different types of communication onto different operators. Therefore, interconnection at the local node might offer little advantages over LLU in the NGN context.

4.3 Risk and uncertainty

As we have explained above, a common IP-based platform for conveying traffic related to many different services has the advantage allowing operators to use installed capacity more efficiently. For this reason, when dimensioning the network for *existing capacity requirements*, there is less need for spare capacity as demands should be more predictable.

However, a clear benefit of using a single standardised transmission protocol for all services is that the development and deployment of new services should become easier. This may increase the uncertainty with regard to *future capacity requirements*. Services are using a common platform and terminal equipment may be very flexible with regard to the services it can provide, as this may be more a matter of software than hardware. Therefore, there is the possibility that new services may become widespread rapidly, causing a sudden increase in demand for network capacity that may be difficult to satisfy sufficiently quickly. For example, new traffic from broadband-connected computers might be driven by innovative content. Consumers may not need any new equipment or an upgraded network connection to use such new applications. Given this, take-up of such new applications might diffuse rapidly and suddenly stress the network.

Therefore, the potential for risks from new services (and particularly demand shocks from consumers using existing services in new ways) suggest that there might be a greater need to build ahead for risk management reasons. This raises the question of whether costs related to spare capacity held for these reasons should be recovered from existing services. This is a very difficult problem.

Some need to hold spare capacity may be related to risks arising from existing services. For instance, the example of a new application for broadband-connected computers (discussed above) would fall into this category. Clearly costs of holding spare capacity for such reasons can be attributed to providing existing services and so should be recovered from them.

Where capacity is held spare for possible new services, the case for recovering these costs from existing services is much weaker. It is particularly problematic as a network operator with market power might even use spare capacity as an entry deterring mechanism, making it unattractive for entry of alternative networks as it could threaten to cut prices and fill up this capacity. In particular, it might use spare capacity as a means of discouraging innovation in new services by alternative network providers.

4.4 Implementation considerations

A decrease in the distance gradient reduces the absolute importance of physical distance in determining the cost of communicating between two points. The additional cost of traversing each extra kilometre is on average lower with an NGN. However, this tells us nothing about whether distance-related charging is an appropriate model or not for setting regulated interconnection charges.

The move to an NGN also reduces the importance of other factors affecting cost, such as the levels traversed within the network hierarchy; unit costs within core and periphery become more similar. Given this, it may well be that distance becomes a better proxy for measuring cost in an NGN as compared with the alternatives. For instance, rather than setting interconnection charges using an EBC regime based on a small number of levels of network hierarchy with few distance-based subdivisions (for instance as at present with inter-tandem conveyance), simply charging by distance without regard to the level of network hierarchy might at some point become more cost reflective if unit cost differences at different levels of the network became sufficiently similar.

Although distance might lose absolute importance, distance still matters. Other cost drivers may become unimportant faster than distance does. In addition, many of the drivers of cost than may be practically difficult to measure remain closely correlated with distance.

For example, there may be other metrics than become relatively better as measures of cost with an NGN, such as the number of nodes traversed by traffic. For instance, short run capacity constraints may become more relevant in an NGN than at present, with the physical distance over which the traffic is carried having a limited impact on short-run costs. Because of traffic congestion at nodes and short run capacity constraints it is often optimal not to route traffic through the most direct route but rather to deviate it towards nodes with a lower traffic load.

However, just because factors other than distance may be revealed to be important by the move to an NGN, this is not to say that distance stops being a good proxy for measuring cost. For instance, pricing interconnection products according to factors such as the number of nodes traversed, or the congestion at these nodes, would provide the wrong incentives for traffic routing and network dimensioning by the incumbent. For this reason it is more appropriate to define prices as a function of a variable over which the incumbent has no control. This may make distance a good determinant of cost for efficiently routed traffic within an efficiently dimensioned network.

In particular, in a similar way in which the EBC approach currently provides a proxy for distance based on the network hierarchical levels used, distance may provide a reasonable proxy for the number of nodes used or the congestion generated within an efficient network. Distance-related costs still exist (even if small) even with an NGN. This suggests that rather than the “death of distance”, the move to NGN may increase the potential of the actual distance between to terminals to act as a proxy for measuring cost.

5 Conclusions

The cumulative cost of conveying communication traffic between two terminals takes the form of an inverse S-shape curve, with a steeper distance gradient at the edges and a flatter gradient at the centre. This results from the benefits of aggregating traffic from different users and services over shared links and exploiting scale economies on links. Despite the fact that aggregation of traffic increases the physical distance over which traffic is carried, benefits from sharing the cost of links over more traffic offset additional costs from the extra distance travelled, and the net result is that unit cost falls.

The main implications of upgrading the incumbent network to an NGN structure are:

- greater aggregation of traffic, both across users and services, resulting in more intensive use of network and greater exploitation of scale economies;
- smaller degree of differentiation between communications with respect to service type, resulting from the use of a common transmission protocol; and
- potential changes in network topology, including:
    - the potential re-location of nodes; and
    - the potential centralisation of routing capabilities to the core of the network (where before there was intelligence both at core and local nodes) which would result in a smaller degree of differentiation of the cost of communications with respect to distance.

The move to an NGN may exploit scale economies and achieve a reduction in unit cost particularly in the periphery of the network. This has the effect of lowering the average distance gradient in general.

The impact of additional traffic can be expected to be larger in the periphery. This is because existing traffic volumes on links in the core leave less scope for exploiting *further* scale economies than at the periphery. The increase in traffic and capacity resulting from the move to NGN therefore makes the distance gradient more uniform between periphery and core.

The more uniform cost structure reduces the need for differentiation of interconnection charges according to the network segment used. As costs at the periphery are more similar to costs at the core, the cost impact of moving between hierarchical levels within the network is smaller.

Assuming the costs of competitors remained unchanged by the incumbent moving to an NGN, the lower average distance gradient reduces the scope for competitors to cut costs from interconnection out-payments by building their own infrastructure. In particular, the incumbent’s ability to exploit greater scale economies at the periphery following a switch to an NGN may limit competition in this segment.

However, the relative advantages of operators for *particular* segments are likely to become less relevant under the NGN structure. As the incumbent upgrades its network it becomes unlikely that entrants can benefit from potential technological advantages derived from using newer technology, and thus any cost differences are likely to derive mainly from differences in traffic volumes and operating efficiency differences.

Consider the case of a competitor with a cost *structure* similar to the incumbent's (i.e. an NGN with similar cost-volume and topology) but lower overall traffic volume. Such competitor would face higher unit costs due to lower volumes, but the disadvantage would be proportionally greater at the periphery than at the core. However, greater aggregation at the core reduces the unit cost impact of volume differences between the incumbent and entrants. Therefore, we can expect the impact of differences in the size of the operator to be greater at the periphery than at the core.

This implies that the move to an NGN could suppress incentives to extend alternative networks from the core to the periphery. However, clearly in this situation this would be an efficient outcome reflecting the greater potential ability to exploit scale and scope economies in the periphery. Nevertheless, if competing networks are able to aggregate their traffic in a similar manner to the incumbent (e.g. by offering a full range of services) then entrants may also be able to benefit from exploiting economies further, thus reducing the difference between their and the incumbent's costs.

Annex A

Current implementation of distance gradients

In this section we provide a brief overview of the distance gradients that apply to current PSTN and broadband interconnection tariffs, and their relation to underlying network costs of the services provided. This exercise also allow us to investigate whether there are any differences between narrowband and broadband services. However, due to the different regulatory obligations on different services, the interconnection charge structure of some of these services does not need to reflect the structure of the underlying costs. Therefore, it is difficult to infer from these data how the different nature of services or their networks may affect the cost-distance relationship.

Narrowband

At present, narrowband interconnection charges are set in reference to the number of network elements used, following the Element-Based Charging (EBC) approach. Under this approach, interconnection costs are averaged within each network segment, and distance is only explicitly considered in some instances by using banding within network segments (e.g. short, medium and long inter-tandem conveyance). The reason for this is that differences in cost between different network segments are more important than distance-related differences.

Under the EBC approach, the level at which the call is handed over from one network to the other determines the interconnection services required, which include:

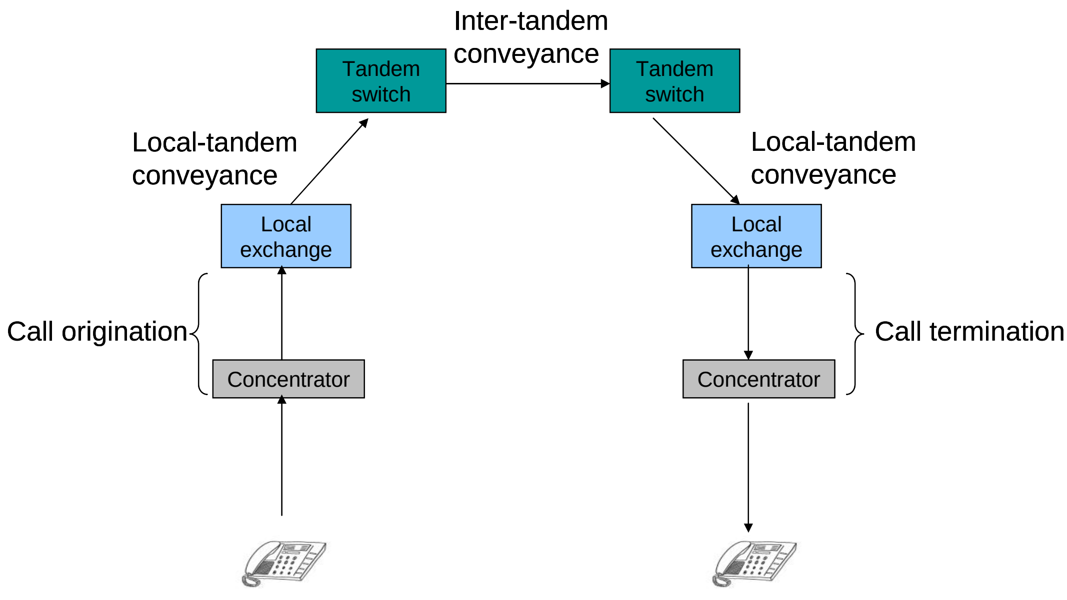
- call origination;
- local tandem conveyance/transit[18];
- inter-tandem conveyance/transit;
- call termination.

These interconnection services, and the extent to which they use PSTN elements, are represented in Figure A-1.[19]

[18] Transit services are those which originate and terminate on a network other than that of the transit provider.

[19] A brief overview of BT's PSTN structure is provided in Annex B.

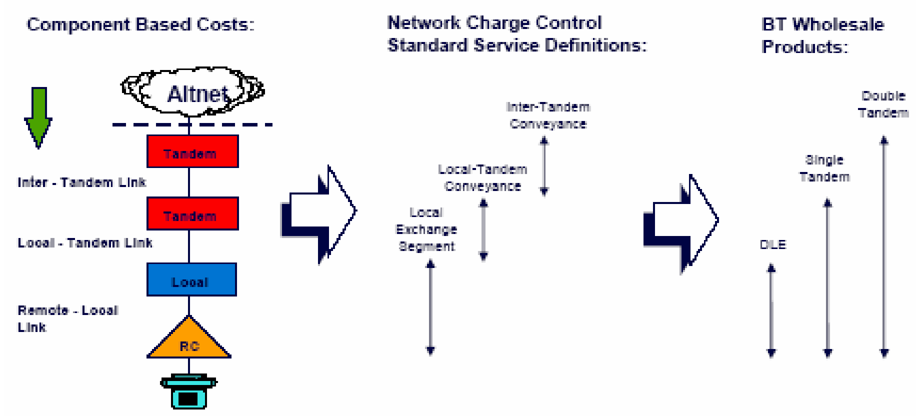
Figure A-1: Narrowband interconnection services over PSTN



Charges are determined by reference to the levels of network hierarchy used and, in the case of inter-tandem conveyance, priced in three bands according to distance: short (<100km), medium (100-200km) and long (>200km). Nevertheless, there is still significant averaging of costs for each network element across traffic travelling different distances.

The service including call origination and local-tandem conveyance is known as single tandem conveyance/transit, while the service including call origination, local tandem conveyance and inter-tandem conveyance is known as double tandem conveyance/transit (and has thus different prices depending on whether it uses short, medium or long inter-tandem conveyance). This is shown in Figure A-2.

Figure A-2: Distance and network hierarchy



Source: Extract from BT presentation, NCC review 2004

Within its regulatory account statements, BT is required to report the cost of regulated interconnection services, along with an estimated floor (forward-looking LRIC) and ceiling (stand-alone costs) for each one of these services. Costs are measured on a current cost basis, based on the network components required for providing each service following the Element Based Charging (EBC) model. The network elements considered include the links through which the communication is carried and the nodes and network switches required for routing the communication from the origin to its final destination, and a cost specifically associated to the length of the link. Using these information, we can investigate the relationship between PSTN interconnection charges and their underlying costs. BT's costs and average interconnection charges for the interconnection products specified above are shown in Table 1**.**

Table 1: PSTN interconnection costs and average charge for the year (pence per minute), for year ended 31.03.04

Service	LRIC (Floor)	CCA FAC	SAC (Ceiling)	Average charge for the year
Call origination	0.121	0.162	0.294	0.189
Local tandem conveyance	0.116	0.154	0.276	0.187
Inter-tandem conveyance – short (<100km)	0.06	0.083	0.154	0.084
Inter-tandem conveyance – medium (100-200km)	0.074	0.099	0.167	0.18
Inter-tandem conveyance – long (>200km)	0.097	0.133	0.265	0.303
Call termination	0.138	0.193	0.437	0.48

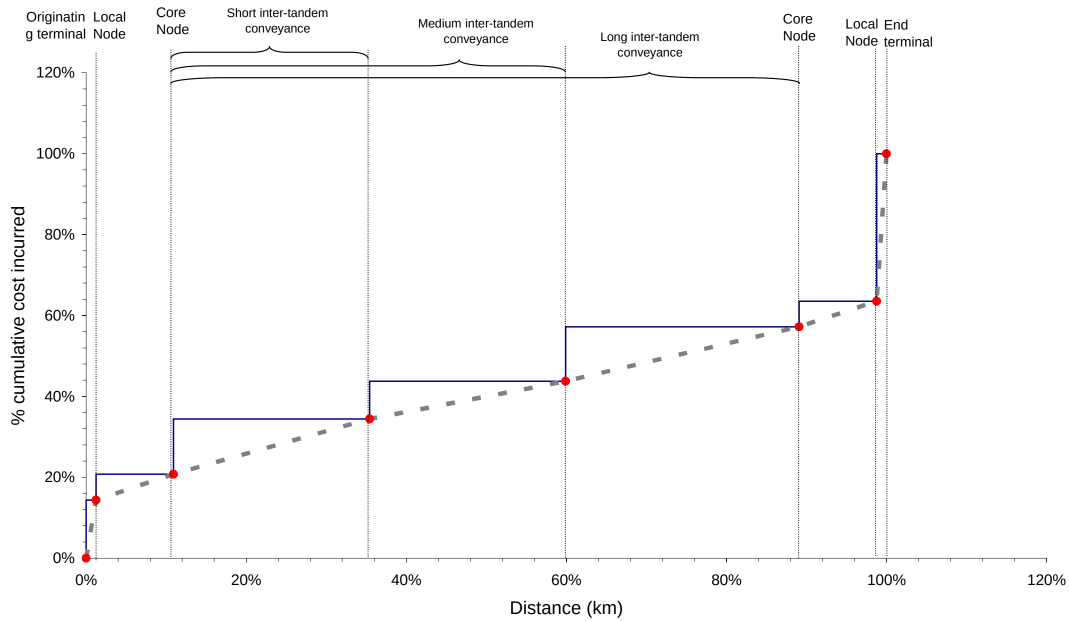
Source: BT, Current Cost Financial Statements for 2005

We can plot a cumulative price-distance relationship for a hypothetical PSTN communication, decomposing it according to the BT's interconnection products required to carry it a certain distance. Due to the EBC structure, this is relationship is a step function. However, if we spread[20] the additional costs over the distance covered we can clearly see how the distance gradients for interconnection to BT's network demonstrate the characteristic inverse-S relationship, where the incremental cost of additional distance is smaller at the centre, particularly when the local access segment is taken into account. This is shown in Figure A-3, where we plot the cumulative cost function over the distance covered for a call requiring long inter-tandem conveyance[21], including the local access segment.

[20] We simply average the increase in charge over the additional distance provided by the interconnection product.

[21] The assumed relative distances for this call are: 2km local access segment, 40km local-tandem link, 319km inter-tandem link. The latter figures have been estimated using figures from BT's Current Cost Financial Statements for 2005, dividing the length usage factor by the link usage factor for each product.

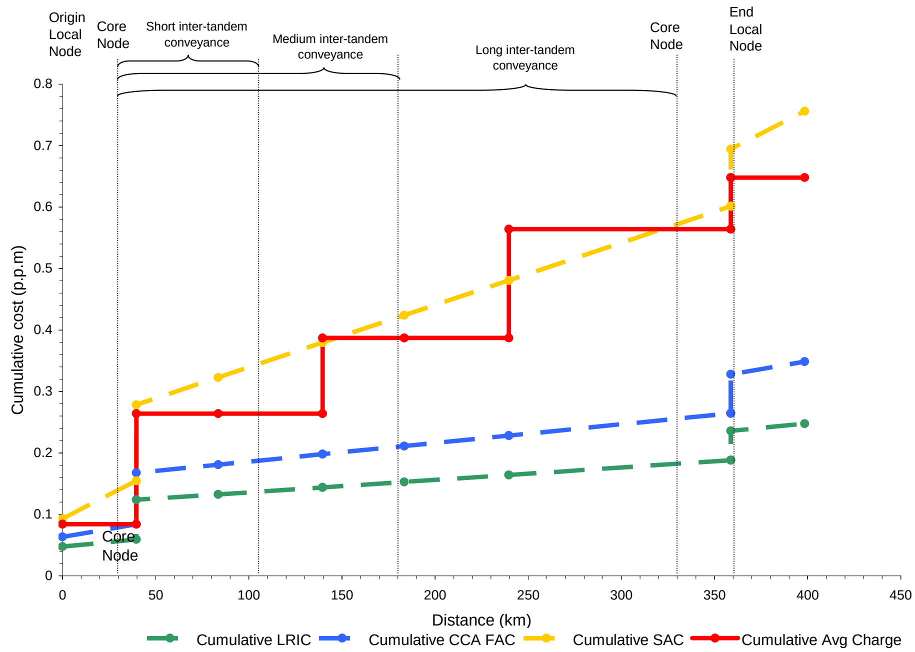
***Figure A-3: End to end cumulative cost function for a long distance call*[22]**



This is consistent with our expectations about the cost-distance relationship. The inverse S-shape of the cost-distance relationship is further confirmed by plotting interconnection charges against underlying costs, shown in Figure A-4 (excluding the local access segment).

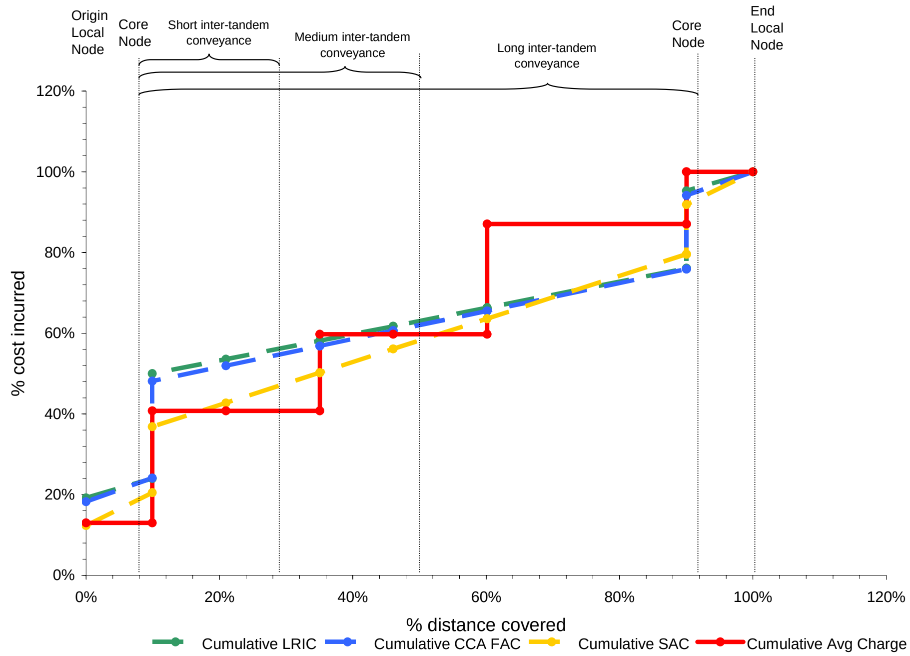
[22] The cost of increasing distance for inter-tandem conveyance (i.e. from short to medium, and from medium to long) is computed as the difference between interconnection charges (i.e. medium inter-tandem conveyance minus short inter-tandem conveyance, and long inter-tandem conveyance and medium inter-tandem conveyance respectively).

Figure A-4: Cost-distance relationship for BT's PSTN interconnection charges and underlying costs (excluding local access segment)



The structure of interconnection prices broadly reflects the relative costs of various network elements. However, the cumulative cost curve for wholesale prices is more similar to that for standalone costs (SAC) than it is for long run incremental costs (LRIC). The cumulative interconnection charge curve is steeper than the cumulative LRIC; it is also more uniform (as can be seen in Figure A-5), with a greater mark-up over LRIC at the core than at the periphery.

Figure A-5: Proportional cost-distance relationship for BT's PSTN interconnection charges and underlying costs (excluding local access segment)



Given that competition at the core already exists, the current relatively uniform distance gradient of interconnection charges relative to the LRIC distance gradient arguably diminishes to some extent the incentives of alternative networks to extend towards the periphery (relative to the case where EPMU over LRIC were used), as charges at the periphery are relatively cheap.

Broadband interconnection

There are no regulatory requirements to publish cost figures for broadband interconnection products. Therefore, we can only use interconnection charges as a proxy for the underlying cost structure. However, broadband interconnection charges are regulated using a retail minus approach and thus need not be cost reflective in the same way as PSTN charges. This limits the extent to which interconnection charges can be used in order to assess how costs change with distance.

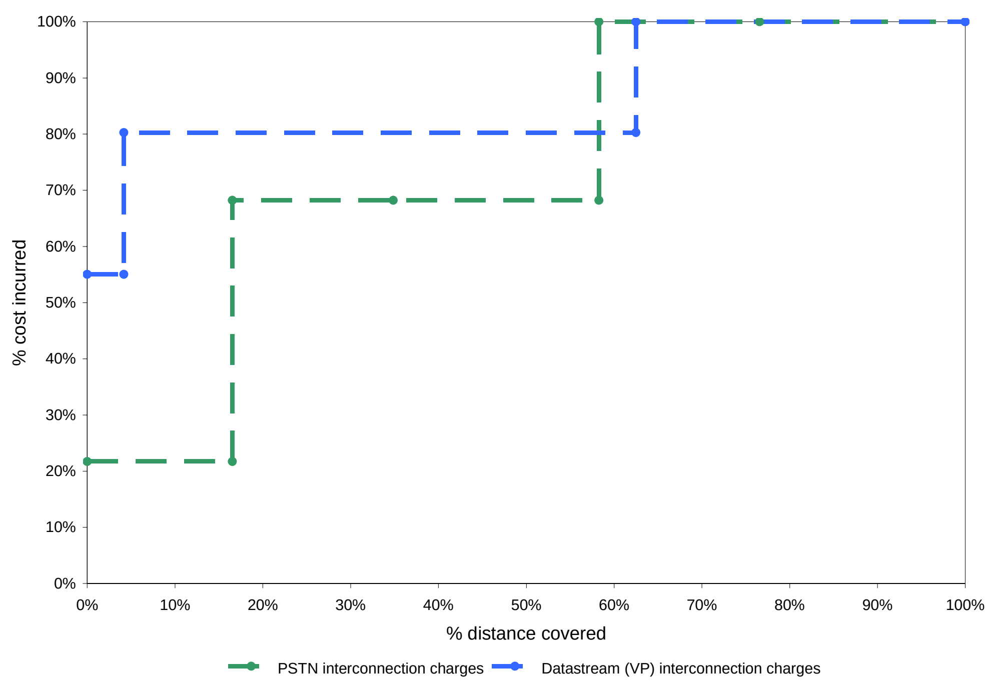
The structure of the core network over which broadband conveyance is provided tends to be flat, without tiered levels of aggregation.[23] Charges for broadband conveyance include a more explicit distance component.

[23] A brief overview of BT's broadband ATM network structure is provided in Annex B.

The structure of interconnection charges for broadband products is different for Datastream (which uses virtual paths) and IPstream products (which are IP based). Datastream products are roughly comparable to the first half of PSTN interconnection products, in the sense that they reserve a virtual dedicated line from the terminal to the core of the network for every communication established. Comparing the proportional cost-distance relationship of interconnection charges for PSTN and Datastream products, shown in Figure A-6, we observe that the cumulative cost curve is to some extent more uniform for narrowband PSTN than Datastream broadband.

This indicates that, as compared with PSTN, Datastream broadband products exhibit a relatively greater cost for access to the core network, and a relatively more modest cost for traversing the core network. The somewhat less uniform cost structure of broadband suggests greater scale economies from traffic aggregation at the core, which lead to greater unit cost differences between the core and periphery. This would indicate that the benefits from traffic aggregation at the core can be felt more strongly in the broadband network. For example, all broadband communications are typically routed to the core network, while a significant proportion of voice calls are local and may be routed without going through the core network. This means that the difference between traffic at the periphery and core may be smaller in the PSTN network, and thus the degree of aggregation at the broadband network core more substantial.

Figure A-6: Cost-distance relationship for PSTN (to the core) and Virtual Path interconnection charges, (excluding local access segment, total distance of 240km)



Interconnection charges for IPstream products are unrelated to the distance traversed on BT's network. The only distance-related component consisting in charges for conveying traffic from BT's network to the alternative provider's Point of Presence (PoP) if it is beyond 40km; this charge is linear with respect to distance and unrelated to bandwidth, as it requires a dedicated link for the customer.

As interconnection charges for IPstream products do not depend on the distance over which traffic is carried on BT's network, the appropriate comparison when making alternative infrastructure decisions would be the cost of using a combination of Datastream products and own network instead. Whether alternative networks are able to reduce their costs by renting a Datastream virtual path and building own network infrastructure will depend on the extent to scale economies on these links can be exploited, and thus on the traffic they can aggregate on these links.

Annex B

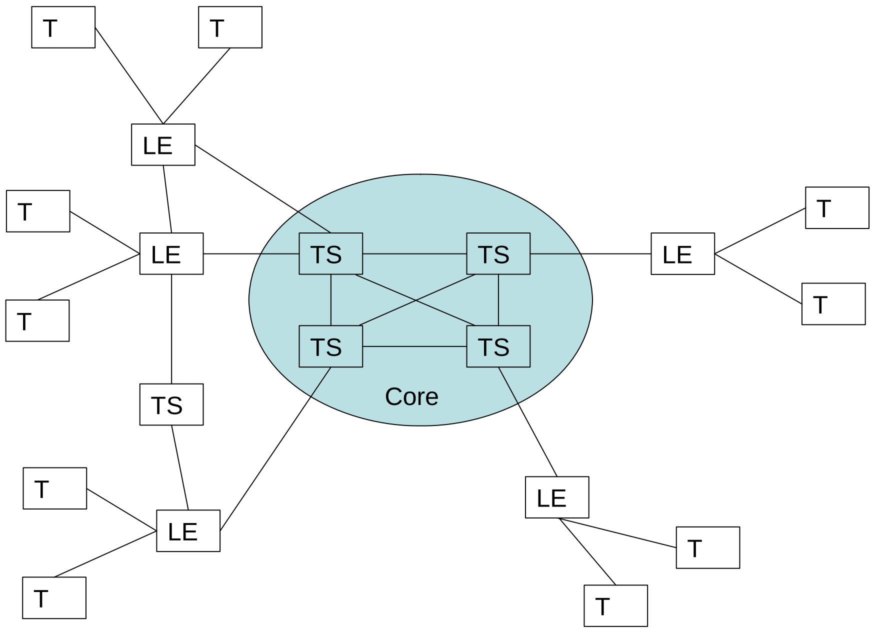
BT's PSTN structure

Figure B-1 shows the structure of BT's current PSTN. Customers are connected via a concentrator to local exchanges (via remote-local links), which route calls according to destination numbers. Local exchanges are in turn connected to tandem exchanges[24] (via local-tandem links), which interconnect different local exchanges in their area and interconnect with tandem switches in remote areas (via inter-tandem links) in order to allow for communications between different areas. Tandem switches are also connected to foreign networks for international calls.

Traffic is carried from the originating customer to the local node where it is routed according to destination. Traffic directed to a terminal connected to the same local node can be directly routed to the end destination. If the receiving party is not connected to the same local node as the originating party, then traffic is routed to a higher node (a tandem switch), where traffic is further routed towards the local node where the destination party is connected (maybe requiring going through a number of tandem switches if the destination local exchange is not directly connected to this particular tandem switch) and from there to its final destination, or to the destination network at a suitable interconnection point if the destination party is not connected to BT's PSTN.

[24] Local exchanges may be connected to other nearby local exchanges where it is cost-effective to bypass the tandem switch, however, for the sake of simplicity, we will not consider this special case.

Figure B-1: BT's current PSTN structure (simplified)



T=end user terminal

LE=local exchange

TS=tandem switch

There are two types of tandem switches: local tandem switches and main tandem switches. Local tandem switches have limited interconnection, linked to a number of local exchanges and other tandem switches. Main tandem switches are fully interconnected. As all local exchanges are connected to at least on main tandem switch, all calls can be routed going through a maximum of two tandem switches; however, in practice traffic may be directed via longer routes involving a larger number of tandem switches in order to avoid tandem switches with a heavy traffic load and congestion.

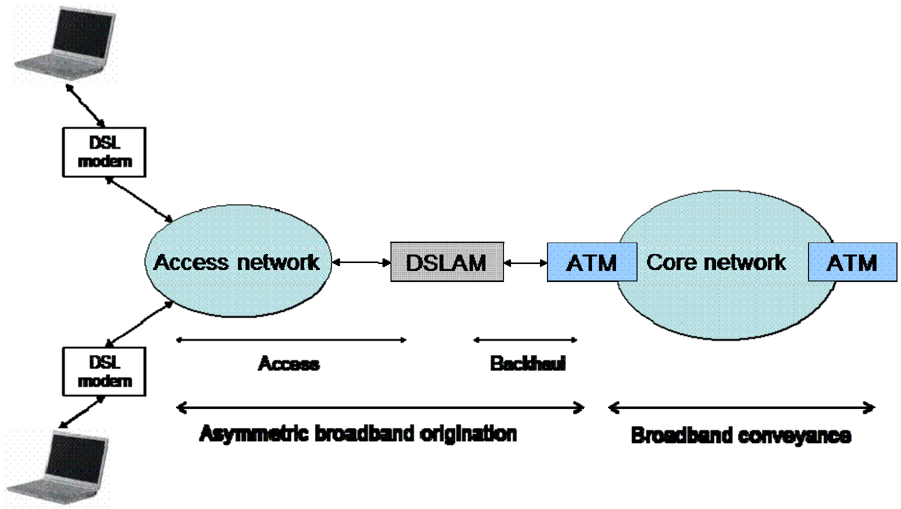
Thus, local exchanges may potentially route local calls to another customer connected to the same local exchange, or through one or more tandem switches interconnecting with the local exchange (directly or through another tandem switch) to which the end customer is connected.

Other networks interconnect with the PSTN at a local exchange level and/or at a tandem switch level.

BT's ATM network structure

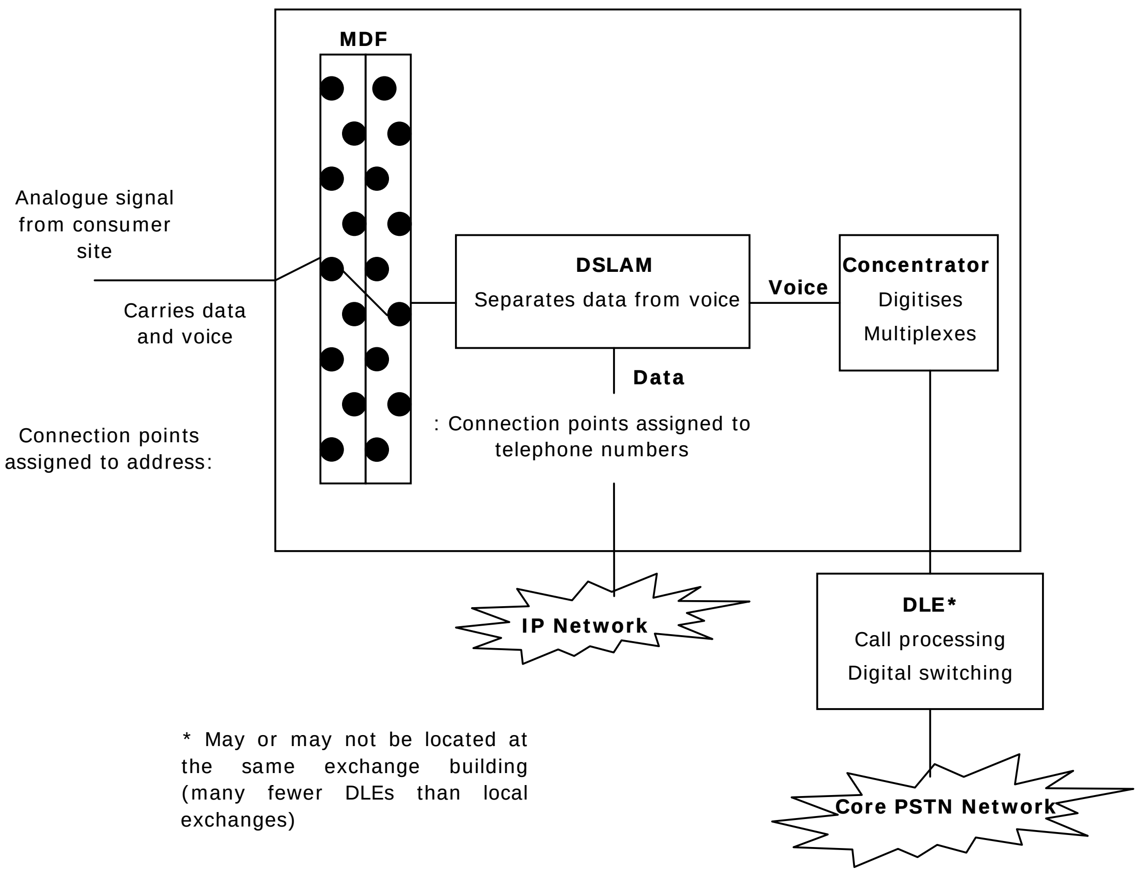
The network structure used for broadband products is simpler than for narrowband. Figure B-2 shows the typical network structure used for asymmetric broadband. End users are connected to a multiplexer (DSLAM) in their area, which is in turn connected to an ATM and from there to the Internet core network.

Figure B-2: Broadband network structure



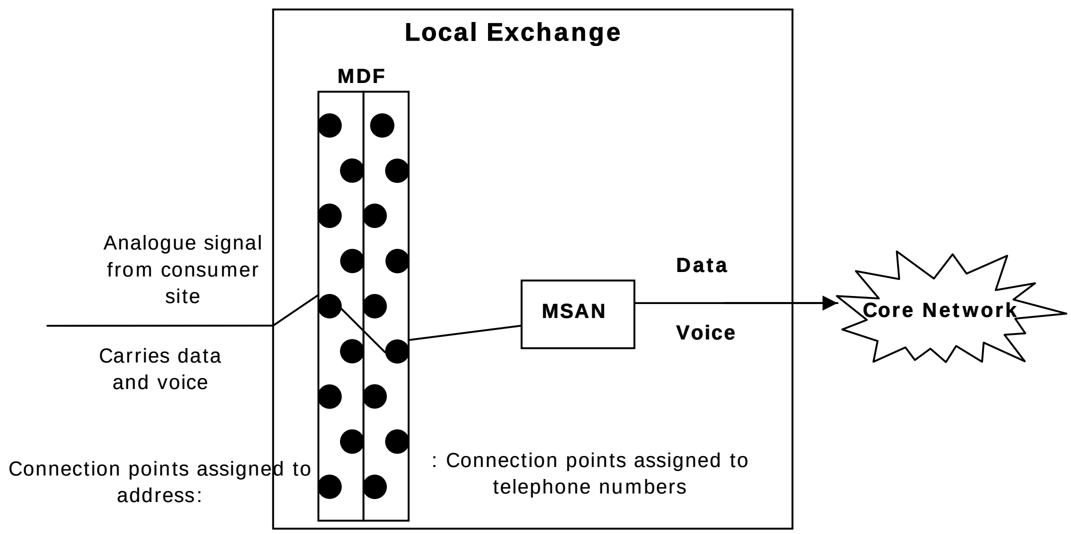
BT's current local node

Figure B-3: Local node in current BT's networks



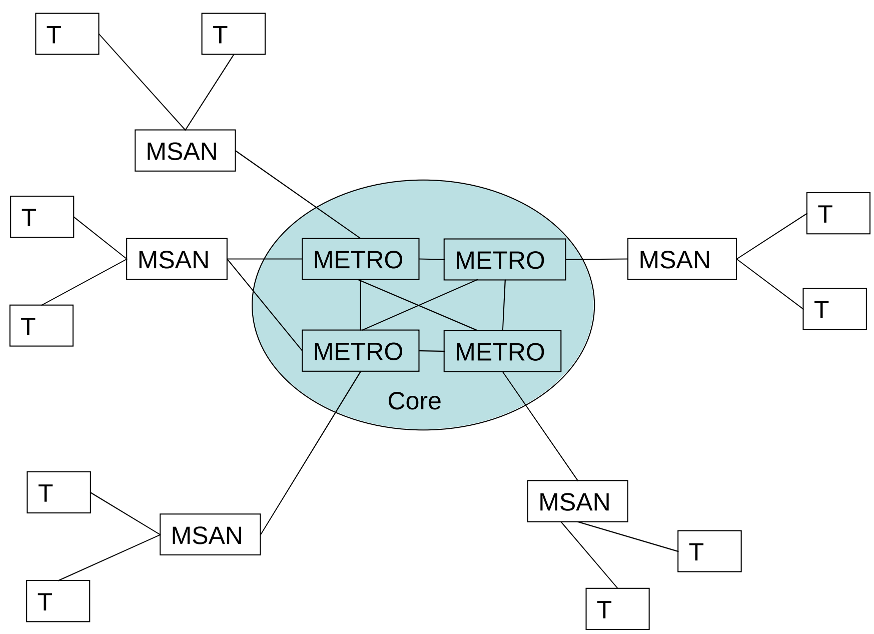
BT's proposed NGN local node

Figure B-4: Local segment in BT's 21CN



BT's proposed 21CN

Figure B-5: BT's 21CN structure



Annex C

Cost of conveyance per traffic unit

Assuming that each network element i has a fixed cost of infrastructure (F_i) plus a constant cost per traffic unit utilising that element (v_i), we can write the cost per traffic unit for network element i as:

$$c_i = \frac{F_i}{t_i} + v_i$$, where t_i is the traffic using the element.

Traffic going through different network elements

In order to simplify our analysis, we consider that all calls are ended on the same network, and therefore that $\rho_L + \rho_M + \rho_C = 1$.

We can calculate the theoretical average traffic for each type of element, assuming that:

- each call could be provided using the more direct route without requiring any detour in order to avoid congested nodes (cost implications deriving from irregular traffic distribution and congestion of particular nodes are discussed below); and
- traffic is uniformly distributed and symmetrical (i.e. all users generate exactly the same amount and type of traffic).

In order to do so, we also need to define the aggregation ratio (α_i), providing the average number of links aggregated at node i, thus $\alpha_i > 1$.

We denote the traffic generated and received by each terminal (user) as t, which is the average traffic at each local link element (L).

At the local node (LN) traffic from and to terminals connected to that node is aggregated at the ratio α_L , and thus the traffic at local node is equal to $\alpha_L\, t$.

The proportion ρ_L (or equivalently $\rho_M + \rho_C$) of the traffic going through local nodes may be directly routed to a terminal connected to that same node. Thus, assuming local nodes are connected to a single core node, traffic at the local node to core node link element (M) is $t\alpha_L(\rho_M + \rho_C)$. Similarly, traffic at the core node (MN), aggregating traffic from a number of M links, is $t\alpha_L\alpha_M(\rho_M + \rho_C)$.

Traffic traversing the core network (C), will be equal to the traffic at the core node not directly routed to a terminal connected to a local node directly linked to that core node; thus traffic at MN minus a proportion ρ_M , hence $t\alpha_L\alpha_M\rho_C$. However, as core nodes are interconnected with a number of other core nodes, only a proportion of the traffic can be expected to travel on each link. Therefore, on average, the traffic on each core node is $\frac{t\alpha_L\alpha_M\rho_C}{n_c}$, where n_c is the number of links going out from each core node.

Impact of collapsing networks into one

In absence of any cost savings from network consolidation, one could expect unit costs to be equal to the weighted average of the standalone costs of each network. Cost savings can therefore be modelled as a change in the fixed cost/traffic ratio in each segment, and define

$$s_i = \frac{\dfrac{F_i^{\text{combined}}}{t_i^{\text{combined}}}}{\dfrac{\sum F_i^{\text{standalone}}}{\sum t_i^{\text{standalone}}}}.$$

Thus the new cost per traffic unit of each segment would be given by:

$$\frac{c_i}{t_i} = s_i \frac{\sum F_i}{\sum t_i} + v_i \quad .$$

One expects the increase in costs to be smaller than the increase in traffic, as otherwise it would not be profitable to combine the separate networks, and thus we assume $s_i \leq 1$. In addition, we expect the benefits from scale economies to be more substantial at the edges of the network, and therefore assume that s_i becomes smaller as i tends towards the edges.