Regulation Initiative Discussion Paper Series
Number 46

Infrastructure Access Pricing and Lump Investments

Ralph Turvey, Associate Director, Regulation Initiative
London Business School

January 2002
Infrastructure access pricing and lumpy investments.

Ralph Turvey, Associate Director, Regulation Initiative.

Introduction

Consider the problem of financing and providing an incentive for an infrastructure provider to enhance the infrastructure. Much of the theoretical literature on access pricing is not applicable for two reasons,

Firstly, it ignores the difference between two cost concepts. One, which determines revenue requirements, is the accounting cost of the existing infrastructure, a sunk cost which reflects past conditions. The other, which is relevant to optimal resource allocation, is the forward-looking cost of enhancement, which depends upon current and expected future prices and technology.¹ A capital charge per unit of existing output consisting of:

\[
\text{The regulatory value of the infrastructure } \times \text{ the allowed rate of return + allowed depreciation}
\]

may be very different from the unit capital costs of future infrastructure enhancements².

Secondly, it assumes that the capacity of the infrastructure is continuously variable, so that there is such a thing as a strictly marginal cost of capacity. This paper deals with the contrasted case where capacity can be increased only in large chunks and where the forward-looking cost of increases may differ considerably from the sunk accounting cost of existing capacity.


² The belief that they are meaningfully related if Modern Equivalent Asset valuation is used is a chimera. It is briefly dismissed on p.41 of *What are Marginal Costs and how to estimate them*.
Capacity additions, that is to say enhancements of the existing infrastructure required to meet a growing demand, are often indivisible, entailing lumpy investment. When this is the case capacity cannot be gradually increased pari passu with demand but only in large amounts at considerable intervals. Before an increase, the pressure of demand on capacity grows, with the result that congestion costs are increasingly incurred. This justifies higher and higher access charges for use of the infrastructure, though, in practice, non-price rationing is frequently applied instead. Once a capacity increase occurs, congestion costs and the appropriate level of access charges fall drastically. Thus the problem arises that the investment may not be remunerative.

Section I describes the nature of congestion costs and the lumpiness of capacity increments in three industries, airports, railways and electricity transmission. Then the problem of paying for investment is posed with the aid of some simple static models in section II. Finally, section III examines some complications ignored in these simple models and some possible solutions to the problem.

I Examples.

Airports

Airport congestion arises from limitations of runway capacity, of aircraft stands, of terminal passenger capacity and, in some cases, of passenger access to the airport. A new terminal or runway is a classic example of a lumpy investment.

Congestion is manifested in aircraft delays which impose a cost upon airlines and passengers. The various delays which a flight can suffer are as follows:

<table>
<thead>
<tr>
<th>DELAY</th>
<th>CAUSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delayed pushback from gate</td>
<td>Late arrival of incoming aircraft</td>
</tr>
<tr>
<td>Delayed pushback from gate</td>
<td>Passengers late checking in or boarding³</td>
</tr>
<tr>
<td>Delayed pushback from gate</td>
<td>Ground handling delays cause turnaround time to be exceeded</td>
</tr>
<tr>
<td>Delayed pushback from gate</td>
<td>Aircraft maintenance problem</td>
</tr>
<tr>
<td>Delayed pushback from gate</td>
<td>Requested by Air Traffic Control to avoid airspace congestion</td>
</tr>
<tr>
<td>Delayed pushback from gate</td>
<td>Requested by Air Traffic Control because of congestion at destination</td>
</tr>
<tr>
<td>Takeoff delay</td>
<td>Takeoff queue in holding area or head of runway, or landing queue</td>
</tr>
<tr>
<td>Stack delay</td>
<td>Landing queue at destination</td>
</tr>
<tr>
<td>Approach delay</td>
<td>ATC requires lengthened path from stack to runway or reduced speed</td>
</tr>
</tbody>
</table>

³ An aircraft can be seriously delayed by the need to find and remove from the hold the baggage of a passenger who has failed to board the aircraft.
Airspace congestion arises when aircraft movements in an Air Traffic Control airspace sector reach that sector’s capacity\(^4\). This is set by safety considerations, being determined in terms of en route controller workload, so is a function of the number of routes traversing the sector, how much they intersect and the number of changes of flight level. Capacity can be increased by dividing one sector into two, but, since handover from one controller to another adds to workload, it will less than double. Revision of routes which reduce workload may sometimes be possible.

There can be interdependence between congestion affecting the traffic of one airport and congestion affecting the traffic of a neighbouring airport. Thus, although Stansted traffic has no impact on Heathrow traffic, it does have a small impact on Gatwick traffic, while traffic flows at Luton and Stansted are mutually constraining. Growth at City Airport could adversely affect Heathrow traffic, and growth at Northolt would certainly do so.

Congestion at the destination airport results in the Terminal Controller stacking aircraft at holding points in the vicinity of the airport, thus forming a landing queue. Optimisation of the the landing sequence may cause it to differ from the arrival sequence, and the same holds for takeoffs from congested airports. The reason for this is that the required lateral separation between any two aircraft and hence the interval between landings and the interval between takeoffs differs according to aircraft size because of differences in the wake turbulence they create.\(^5\) Separation distances may

\(^4\) Eurocontrol produces weekly reports on primary air traffic flight management delays in UK airspace

\(^5\) Aircraft in flight generate a pair of counter-rotating vortices trailing from the wing tips caused by the pressure differential above and below the wing. They can pose problems to following aircraft as they can impose rolling moments exceeding the roll control capability of some aircraft. The strength of the vortices is governed by the weight, speed, and shape of the wing of the generating aircraft, increasing proportionately with increase in aircraft operating weight. Required separation is thus a function of prevailing weather conditions. A crosswind will affect the lateral movement of the vortices. Thus, a light cross-runway wind could result in the upwind vortex remaining in the touchdown zone for a period of time and hasten the drift of the downwind vortex toward another runway. Similarly, a tailwind condition can move the vortices of the preceding aircraft forward. It takes time for a vortex to dissipate. (From an FAA Advisory Circular)
be more flexible for departures than for arrivals because departure paths may diverge
soon after takeoff. When they do not, a two minute wait between departures is
necessary in order to ensure radar separation.

Runway capacity is defined as the maximum number of aircraft movements
that can be scheduled to use a runway such that delays resulting solely from ensuring
the safe separation of aircraft using the runway (arrival delays in the holding stacks
and the departure delays at the runway holding points) are acceptable — as agreed
between the airport and the airlines in terms of the average delay, measured over a
period of given length and the maximum delay experienced. Gradual improvements
in operating procedures and pilot training have produced and can continue to produce
a small annual increase in runway capacity. Expressed in terms of aircraft movements
per hour (which averages 82 at Heathrow and 48 at Gatwick, with its single runway)
it is thus determined by the intervals which have to elapse between successive
movements in order to ensure safety — or by the further safety requirement that only
one aircraft at a time may be on a runway. In the latter case, additional runway exits
and entrances can sometimes increase capacity by reducing the time an aircraft has to
be on the runway. Furthermore, movements at night may be limited or prohibited
because of noise nuisance. This is greater for arrivals than for departures, as aircraft

---

6 Capacity declarations are made by BAA twice a year. The runway, terminal and apron are viewed as
a hierarchy of constraints. At present ACL take the first capacity constraint as being the runway.
Runway constraints are decided at a meeting of a sub-committee of the Airport Co-ordination
committee which comprises representatives from National Air Traffic Services, the Airport and the
airlines. The group review the evidence available, which would include the results of runway capacity
simulation by National Air Traffic Services’ Department of Research Analysis. Given the delay
criteria the group then arrive at a consensus view of the capacity. The responsibility for the capacity
declaration lies with the airport, who issue the formal statement of capacity. An average delay of ten
minutes is accepted. Since capacity partly depends upon the mix of aircraft types, it varies somewhat
by time of day. Improvements in Controller and airline efficiency can raise capacity, but the future
possible annual increase is expected to amount to no more than one or two aircraft movements per day
at the major London airports.
arrive at a 3° angle with changes in thrust. In summary, the constraints on runway use are: the arrival/departure mix, the aircraft type mix, wake vortex separations, departure routs, time on runway, Air Traffic Control procedures, runway configuration, weather conditions and noise restrictions. Mixed mode operation, where arriving aircraft and departing aircraft use the same runway, could raise capacity if introduced at Heathrow with its two runways, but is not allowed on noise grounds.

Runway congestion, i.e. congestion which would be relieved by the addition of another runway, therefore takes the form of queues of aircraft waiting for takeoff or landing. The takeoff queue consists of aircraft waiting their turn at the runway holding point or in other holding areas, plus aircraft whose pushback from their gate has been postponed on account of the queue. The landing queue consists of aircraft in holding stacks plus aircraft at origin airports whose departure has been delayed on account of the expected delays at the destination airport.

Queues can arise even at times when the hourly number of slots multiplied by the average required interval between movements is less than one hour, because of uneven intervals between arrival times and between pushback times. However, standard queueing theory is inappropriate for analysing arrival and departure delays, since it relates to steady states, whereas there is a daily cycle in aircraft movements. In the following diagram the horizontal axis relates to the part of the 24 hours when the airport is open. The continuous line shows the cumulated total of aircraft wanting to land or take off starting in the early morning and ceasing to grow late in the evening, while the dotted line shows the cumulative total of actual landings and takeoffs, its slope representing the average required interval between movements. The area between them represents total delay during the day. It is clear that an additional or decremental aircraft will not only incur or avoid its own queueing time but will also impose additional queueing time, or save queueing time, for all subsequent movements, thus adding to or reducing their costs, both for the airlines and as delays imposed upon their passengers. These externalities are the part of marginal

---

7 “Slot” is rather confusingly used in two senses: (1) The quarter-hour allocated well in advance by Airport Co-ordination Limited for a takeoff or landing for a season, reflecting an airport’s capacity constraints and negotiation between airlines; (2) A takeoff time, with a latitude of +15 minutes or –10 minutes, allocated on the day by Air Traffic Controllers to avoid airway congestion.
congestion cost not borne or saved by the incremental movement itself. This part is the *excess* of marginal over average congestion cost, since the latter is incurred by each user. A similar point applies to railway congestion discussed below.

Let the average interval required between touch-downs or takeoffs be $\tau$. Then a marginal plane arrival or push-out will hold up every plane that succeeds it by $\tau$ so long as the queue (i.e. the number of planes stacked or in a line-up for takeoff) endures. Let the number of planes joining the queue until it is dissipated $H$ hours later be $n$. Then the total delay that the marginal plane arrival or push-out imposes on these other planes is $n\tau$.

But $\tau = H/n$, so the total delay imposed upon others = $H$. Multiply by the value of time for the passengers and airline of an average plane, $V$, and we have the excess of marginal social cost over marginal private cost = $HV$. Note that this will be greatest just after the queue has first formed and smallest just before it ends, when cumulative touch-downs or takeoffs have caught up with cumulative arrivals or push-outs.

At Heathrow, at least in summer, queues usually endure from 6am till 10pm; at Gatwick from 6am till 1pm and again from 5pm till 7pm. Traffic growth at Stansted now means that queues endure for some two hours starting at 8am, midday and 5pm. At Manchester, the building of a second runway has ended congestion.

Next comes stand (apron) capacity. Stands may have a pier and jetty or be remote off-pier stands requiring the use of busses for passenger embarkation and
disembarkation. In aggregate, stand capacity is simply defined as the number of stands — parking places where aircraft can be loaded and unloaded, serviced and so on. In total this is required to meet the number of arrivals multiplied by average turnaround time\(^8\). However, stand capacity can limit not only the total of aircraft movements but also their composition by aircraft type and terminal used. Larger aircraft require more stand space than smaller aircraft, though two small aircraft can use a large stand, and each airline wants its allocated stands to be at, or readily accessible from, the terminal they use. Hence the stand constraint is more complex than a single simple upper limit; it may restrict particular movements. If a runway slot becomes available, a secondary check may be necessary to establish whether it can be given to a particular aircraft type at the terminal the airline uses. If not, then it may be given to another airline, possibly at another terminal.

New stands will be provided at a new terminal, but the number of existing stands can be adversely affected by construction work or, more permanently, reduced by the need to provide more clearance for new large aircraft or by taking space for new roadways. The stand composition is adapted to forecasts of airlines’ fleet composition. Construction work on stands can impede current airport activity. The result of all these factors is that stand development planning cannot be a formalised optimisation procedure.

Runway and stand constraints relate to aircraft but can be expressed in terms of passenger movements, known as PAX, given data on the number of passengers per aircraft movement. This average number has been steadily rising and the rise is expected to continue. Concentration in the morning of the long-haul arrivals of large aircraft and their subsequent departures can produce morning peaks in passenger flows not reflected in the number of aircraft movements which, at least at Heathrow, does not vary much between the early morning and late evening.

---

\(^8\) Turnaround time is higher for large aircraft, for which disembarkation, servicing, refuelling and loading take longer, and for aircraft flying long-distance routes, the latter sometimes requiring a longer stay for timetable reasons. Thus a move towards larger aircraft will raise average turnaround time. An airline granted an arrival slot earlier than it wants may keep its aircraft on the ground until an appropriate departure slot, hoping in time to get a later arrival slot. Turnaround time is shorter for aircraft based abroad than for home-based aircraft as the latter may spend more time on refuelling, loading etc. and may be scheduled to fly a different route on each successive departure. Aircraft staying more than three and a half hours are usually requested to move to a remote parking place.
Terminal passenger capacities, expressed in PAX per hour and distinguishing arrival flows from departure flows, constitute the third kind of constraint that can rule out a proposed landing or departure slot at a terminal at a particular time.

Complex simulations are needed to analyse actual terminal passenger capacity and the effect upon it of terminal enhancements or extensions\(^9\). Passengers move through a series of locations, any of which can constitute or contain a capacity limitation. Terminal capacity can be ascertained as the capacity of the tightest constraint. Each is defined in such terms as the rate at which passengers can be handled or the space required, that is to say in terms of target levels of service. Passenger flow in excess of capacity simply means that some of these service targets are not met, imposing a cost upon passengers in the shape of discomfort. Higher or lower standards would naturally entail higher or lower costs. Examples of these standards (they are many) include requirements that 95% of passengers have to wait no more than \(x\) minutes to check in, that lounge space per passenger is to be at least \(y\) square metres, and that baggage delivery is to be completed within \(z\) minutes of first passenger entering reclaim.

The three sets of constraints determine the number and composition of slots that can be allocated, a number which varies slightly during the day, with a small drop in the middle of the period to allow for catchup.

The cost to airlines of delays can be looked at from what can be termed a short-run and a long-run perspective. The former examines the addition to an airline's cost from delay to an individual flight, a cost which can be estimated case by case. It comprises some or all of the following items, most of which will be larger the larger the aircraft and the number of passengers:

- Overtime pay for cabin crew if arrival falls outside their normal hours
- Premium pay for flight deck crew if arrival falls outside their normal hours\(^10\)
- Extra fuel costs if the delay is caused by rerouting or involves time spent


\(^{10}\) But if the delay would bring either group outside their legal hours they will not be able to fly and may have to be provided with hotel accommodation,
in a stack at the arrival airport; in either case additional flight-hours related
maintenance expenditure will be incurred.

- Extra parking charges if the delay involves extra parking time at the
departure airport.
- The late arrival may necessitate use of off-pier parking on arrival, so that
bus costs are incurred.
- If some passengers’ connections are missed they may have to be
compensated or re-booked on another airline; if their baggage misses the
connection there will be a cost of forwarding it to them.
- If the delay exceeds the buffer time allowed in scheduling, the aircraft’s
next departure will be delayed (known as “rotational” delay) possibly
involving extra parking charges and staff costs.

Given that nominal flight times are fixed so as to allow for possible delays it
might be thought that a fraction of flights would arrive early, possibly providing a cost
saving. However this rarely happens, flight speed being adjusted to avoid early arrival
which could often find ground services that were not ready and might upset stand
allocation.

It can be seen, therefore, that valuing an airline’s average short-run cost for
each type of aircraft per minute of delay in excess of, say fifteen minutes, would be
possible only on the basis of very extensive (and confidential) data. There may
therefore be a temptation to estimate it by dividing some total of operational costs for
each type of aircraft by the total hours of operation. This would clearly provide a very
poor approximation to the right answer.

The long-run perspective is relevant when delays, and what is more important, their
variance, are increased or decreased by a long-lasting increase or decrease in the
number of aircraft movements. If they were to fall:

- Shorter buffer times (extended turnaround times used in the schedule to protect
departure punctuality) could be allowed, possibly reducing the required number of
standby aircraft and creating the opportunity for rescheduling.
- Shorter buffer times would allow rescheduling, resulting in improved aircraft
utilisation, fewer aircraft being required to operate the same schedule.
- Reduced departure and arrival delays would allow lower aircraft block times,
diminishing flight times and, in consequence, reducing operating costs.
Fewer aircraft movements would result in fewer off-pier departures, reducing the need to coach passengers and tow aircraft.

Estimation of the resulting savings would be a very complex undertaking; scheduling is often done route by route, since specialist knowledge is required to take account of the manifold complications of flight crew rostering, airport charges and time constraints.

**Railways**

Analysis of the current operating, maintenance and replacement costs of train operation and of the rail network is a complex matter. Here the discussion is limited to congestion costs, that is the costs of delays and cancellations. These rise with the volume of traffic, imposing costs both on the train operators and on passengers (as the value of their time losses).

As in the case of airports, congestion which takes the form of delays is essentially stochastic. If all trains were always ten minutes late, real timetables would differ from nominal timetables and passengers’ time costs and some railway costs would be greater than if trains were never late. But the main problem is the variance of delays. Passengers and operators have to make allowances for possible delays in scheduling their journeys and in planning a timetable. Thus the passenger travelling to an important appointment will allow more time than he or she expects will most likely be required, depending upon his or her experience of delays and upon the degree of inconvenience expected from arriving late.

Delays to trains are caused by unpredictable incidents such as points or signal failures, rolling stock defects, late arrivals of train crews and other station delays. Such delays are independent of the number of trains running. However, unless the timetable is full of white spaces, they have reactionary or knock-on effects in addition, delaying other trains by causing them to lose their train path or wait for another late...

---

11 This section partly repeats some of the section on railway marginal costs in Ralph Turvey, *What are marginal costs and how to estimate them*, (Centre for the Study of Regulated Industries,2000)

12 This becomes clear from the discussion in the Technical Consultation document *The periodic review of Railtrack’s access charges: Usage charges*. (Office of the Rail Regulator, November 1999)
train or wait because of station congestion. Delays can be recorded and those due to these immediate causes can distinguished from the reactionary delays. The social marginal cost of a train path includes the value of these external effects. They generally increase exponentially with the degree of track capacity utilisation a good approximation in the relevant range being furnished by

\[ Total\ Delay = \text{Constant} \times e^{\text{Constant} \times \text{capacity utilisation}} \]

i.e. congestion related reactionary delays.

Such losses can be internalised and expressed in monetary terms reflecting time valuations, if:

— train operators pay monetary compensation to passengers for delays,
— passengers reduce their demand for train journeys, and hence train operators’ revenues, if not compensated,
— and/or the Regulator imposes fines on train operators for delays;
— when they are separate undertakings, the network owner has to pay compensation to train operators for those delays caused by infrastructure faults;
— compensation has to be paid by one train operator to another on which it imposes delays (when there are several operators).

Clearly the time valuations used to determine fines and compensation can only be arbitrary. It is difficult enough to estimate the value of timetabled travel time; it is even more difficult to put a value on unforeseen delays! Yet the argument is the same as in the case of the value of saving lives —valuations are implicit in many decisions and introducing an unavoidably arbitrary valuation does at least secure consistent decisions. It is better to internalise externalities approximately than not at all. The determination of compensation rates and fines can be viewed as one of those tasks best left to politicians because it is too difficult for economists.

Some railway network enhancements are designed to increase safety or to allow higher train speeds. Others are planned to relieve congested parts of the network or to increase capacity so that planned new services do not create congestion. The

\[ 13 \text{ Measured by taking all the trains along a section in a time period, moving them as closely together as minimum permitted headway allows without changing their order and dividing the total time between first and last arrivals at the section end by the length of the period.} \]
The most important kinds of enhancements consist of track upgrades, realignments and recants, signalling upgrades, additional tracks and crossovers, bridge improvements and power supply upgrades. Obviously, some of these enhancements are lumpy investments.

The effects on congestion of a route enhancement can be estimated by using Operational Research tools which employ advanced stochastic optimisation heuristics to generate the new timetable which the enhancement makes possible. Given:
— The layout of a route’s enhanced infrastructure,
— Specified frequency of trains and their stopping patterns,
— Requirements for departure time, turnaround times and connection needs,
a new timetable can be generated which minimises a weighted sum of divergences from these requirements. The effects on delays can then be predicted by using Monte Carlo simulation techniques.

The results of a congestion-relieving enhancement may be: (i) more trains per day, (ii) higher speeds (reduced timetabled journey times) and (iii) improved performance (lower delays and fewer cancellations arising from incidents.) In each case, the result has to be valued in monetary terms in order to allow comparison of the gain with the cost. With a small fraction of enhancements there are tradeoffs between these three dimensions of output, so that the gain can only be expressed as a single figure if their relative values have been determined. However, in a majority of projects the gains appear to be predominantly of one kind.

The gains from different enhancement projects lying along a major route are not always simply additive. For example, the ability to run more trains per day along just one part of a route may not be very valuable, but the benefit from several projects along the whole of that route can exceed the sum of their enhancement benefits considered separately. Thus even when the individual enhancements are small, the realistic choice may lie between undertaking all or none of them. Their aggregate cost constitutes a large indivisible lump of capital expenditure and the benefits are provided by a sizeable increase in capacity. This may result in a drop in the value of access rights or, if there are congestion-related access charges, in the revenue they yield. Hence the problems of indivisibility and lumpiness appear when selecting what railway infrastructure enhancements are to be undertaken and how they shall be financed.
Electricity Transmission

The flow of electricity through a transmission network, given the amount and precise location of generation injections into it and deliveries from it, follows physical laws. A change in generation at one point matched by a change in demand at another will affect flows all over a meshed network, not just in those transmission lines directly connecting the two points. Similarly, the addition to the network of a new line connecting two points, with unchanged generation injections and demands everywhere, can result in changes in flows even in quite remote parts of the network.

This phenomenon makes the notion of transmission capacity a complex one. The thermal capacity of any given line, a function of ambient temperature, is readily ascertainable. But where several lines in parallel link two parts of the system, their aggregate capacity can fall short of the sum of their individual thermal capacities, since, depending on the precise constellation of generation and demands, one of these lines may be fully loaded even though the others are not, so that no increase in the total flow between the two parts of the system is possible.

Generation to meet demand has to be despatched in such a way that sudden unexpected credible outages of transmission links or of generating sets do not result in thermal overload of any line. What is more, such outages may infringe other limits, those set for system voltage level. Also, account has to be taken of the possible effect of disturbances on system stability. Thus “capabilities” have been defined as the maximum power transfer across boundaries which can be accepted in order to ensure that voltage and thermal limits would continue to be observed and stability preserved in the event of any one or two of a set of plausible contingencies.

Dispatching generation to stay within these contingent limits may require that some more expensive generating plant has to be run on the import side of a boundary, while some less expensive generating plant on the export side has to be turned down. Thus taking account of the contingent capability constraints imposes an extra cost, a congestion cost.

A transmission system built without any such constraints would be unnecessarily large, for there must be a point beyond which the cost of adding to the size of the transmission network would outweigh the saving in constraint costs, possibly limited to only a few hours per year, which would result from the addition. Because transmission investment is lumpy, there can be no question of equating its marginal cost with the marginal value of reducing the present worth of future
constraint costs.

Under some schemes for transmission pricing, a revenue for the transmission owners would be produced equal, at times when the constraint was binding, to the difference in marginal generation costs on either side of it multiplied by the flow across it. A lumpy investment to relieve that transmission constraint would reduce that revenue. The same would apply if transmission rights were sold by auction. These might be defined as entry and exit rights to the whole system, similar zonal rights, rights to transmit to or from a central Balancing Point or rights to inject on one side of a transmission constraint and withdraw on the other. Thus the design of incentives for transmission investment and its finance pose very difficult problems when transmission and generation are in different ownership.

Similar problems arise with gas transmission.

II The basic problem

Consider a lumpy investment which raises capacity provided to a single user. For any given capacity, output increases sooner or later entail a rising marginal congestion cost and/or a rising marginal operating cost. This rising cost may include a cost to the infrastructure provider, a cost to the user or an externality which may or may not be internalised through a system of compensation payments or fines. Any actual case will lie between two extremes, one where the infrastructure provider bears all the operating and congestion cost, and one where the user bears it all.

Take the first extreme, where the whole of the area under the marginal cost curve is borne by the infrastructure provider, and consider the effect of a capacity enhancement. The situation where the price charged to the user equals marginal cost is shown in the first diagram. The enhancement shifts the marginal cost curve to the right.
As drawn, the enhancement is desirable if M-P, the net annual reduction in the infrastructure provider’s cost, plus N, the gain in consumer surplus\(^\text{14}\), exceeds the annuitised cost of the enhancement. But it will be profitable for the infrastructure provider only if, in addition to meeting that annuitised cost plus a bit of profit, the user agrees to pay L as a fixed sum to compensate the provider for the excess of annual revenue loss over its annual cost saving.. It is evident that if, instead, the infrastructure provider sought to keep the price sufficiently high to avoid any loss, the investment would prove nugatory.

A situation where the price charged to the user does not equal marginal cost and is not changed is shown in the second diagram.

The enhancement is both desirable and profitable for the infrastructure provider if the annual cost saving Q+R+S exceeds the annuitised cost of the

\(^{14}\) Plus a rise in producer surplus if the marginal cost curve is not flat.
enhancement. Whereas output exceeded optimal output previously, it will fall short of the new optimal level afterwards — unless the price is lowered. If it is to be lowered to marginal cost, the user will have to pay the infrastructure provider a fixed annual sum equal to the annuitised cost of the enhancement plus the infrastructure provider’s loss of revenue, L+S, net of the cost saving, Q+R+S, plus a bit of profit.

Now take the other extreme, shown in the third diagram, where the user bears the whole of the cost under the marginal cost curve. Assume, again, that the user can freely choose output and assume that he pays a fixed unit price to the infrastructure provider, which has zero marginal operating costs.

The enhancement is desirable if the gain in the user’s consumer surplus Q+S+U plus his net cost saving R+T exceeds the annuitised cost of the enhancement. But it will be profitable for the infrastructure provider only if the user meets any excess of that annuitised cost over the increase, U, in the annual revenue of the infrastructure provider, and provides the infrastructure provider with a bit of extra profit. Alternatively, if for the infrastructure provider sought to recover such an excess by raising the price, the social gain from the enhancement would be reduced.

These three simple examples suffice to demonstrate that, with lumpy investment, the infrastructure provider may only find it profitable to undertake an economically desirable investment if, in addition to paying a user charge proportional to output, the user also pays a share of the capital cost. This share must be paid as a fixed amount, not varying with output. If user charges are limited to charges which vary with use, there will be some desirable infrastructure investment which will be underutilised or which cannot be made profitable.
III Complications and solutions

A single user.

To simplify the exposition, the argument above ran in terms of a single user, though the conclusion applies when there are several users. In order to examine the complications so far ignored, take first the case of a single user whose right to use the infrastructure extends for a long time into the future.

If both this user and the infrastructure provider expect the future value of the gains to exceed the projected cost, then they should be able to come to an agreement as to how the cost should be shared which will allow the investment to go ahead. This is so even if their forecasts are not identical. The agreement may share out all risks, giving the infrastructure provider an equity interest in the future user output facilitated by the enhancement. Alternatively, the infrastructure provider may bear only the construction risk leaving the demand risk to be borne by the user.

But if the user’s right is a franchise with a shorter remaining duration than the life of the new infrastructure, the user will be prepared to meet only a smaller part of the capital cost unless it can be assured (i) that its franchise will be renewed, or (ii) that a successor user will be made to reimburse it for part of its capital outlay, or (iii) if its contribution to the capital cost consists of an annual sum payable to the provider. The infrastructure provider will only be willing to meet its share of the capital cost if it can be assured it will be added to its regulatory asset value. Thus the Regulator will have to be involved in the decision as to whether or not to undertake the enhancement. Furthermore, a commitment to renewal of the franchise or imposition of a liability upon any future franchised user may be needed, though this raises the awkward problem that a Regulator whose appointment is for a term of years will not be in a position to commit his or her successor.

Multiple users.

When there is more than one user, an additional problems arise. Sharing out the capital cost of infrastructure enhancements between the various users and the infrastructure provider is a complex matter. Unless it can be sorted out, some desirable enhancements will not be undertaken and those which are undertaken may

---

15 These issues are discussed in chapter 18 “Enhancement framework” in Periodic review of Railtrack’s access charges: Final conclusions, Volume I. (Office of the Rail Regulator, October 2000)
be used suboptimally. Thus suppose that multiple users each pay the provider a share of the capital sum, or make an annual equivalent payment. The apparently simple principle of making these payments proportional to the output of each user following the enhancement raises two problems:

1. Some users may not change their output, yet benefit from a reduction of congestion caused by the enhancement. For example, airlines which will continue to use existing terminals will benefit from construction of a new terminal to which other airlines move.

2. Payments which vary according to each user’s output constitute marginal costs to the users. This will prevent the optimal output level being attained. It could be attained only if users each paid fixed shares unrelated to their current outputs, for example amounts proportional to their shares in the previous output. However some of them may refuse to agree to this if they expect to gain differentially, so that reaching agreement may prove impossible. Again, the Regulator may have to be involved to settle the matter. The best may be the enemy of the good, and it may be that undertaking the investment and charging a price above marginal cost will nevertheless be preferable to not undertaking the investment.

Where capacity can be clearly defined, it may be possible to allocate the rights to use capacity by auctioning them periodically. But this will not resolve the basic problem. A lumpy increase in capacity will lower the auction prices and may raise the infrastructure provider’s revenue obtained from the auction insufficiently, or even lower it, making a desirable investment unprofitable. Furthermore, auctions may raise another problem for regulation. The revenue obtained from them may be different from that allowed under the price control, requiring a change in some other component of the infrastructure provider’s charges by an amount sufficient to offset the over or under recovery. It may be difficult to find a way of doing this which is both acceptable to users and has no distorting effects. Determining fixed contributions in addition to the auction prices in a way accepted as fair by users may be impossible, while charging a price per unit of use above marginal operating cost would lead to underutilisation of the infrastructure.

An alternative way of allocating rights to users yields no revenue to the infrastructure provider, instead providing quasi-rents to the users. This is some form of rationing, where rights are allocated by the infrastructure provider, by the Regulator or (as in the case of airport slots) by a committee of users. Restricting rights in this way will keep down congestion cost, but will yield the same allocation of rights as would auctions only if those users who are excluded are those who would have made the lowest auction bids. This is unlikely to be the case if the primary allocation principle is “grandfathering” whereby rights are awarded to users who had them before.

**New users appear subsequently.**

The difficulties are compounded if there is a possibility that new users will appear and start to use the infrastructure in later years. Any such subsequent entrants will gain from the enhancement, but there will be uncertainty about when and how much they will contribute towards reimbursing its cost. Meanwhile the whole of that cost will have to be covered, involving the risk that new users may not appear and make a contribution. This risk could be borne by:

- Existing users, each being reimbursed by the infrastructure provider from the charges it levies on subsequent entrants, or, alternatively, each obtaining capacity rights some of which new users would have to buy from them in order to be allowed to operate.
- The infrastructure provider, which would be entitled to obtain a contribution to the cost from new users when they appear.

In either case, economic inefficiency could result from requiring a contribution from a newcomer to the (by now sunk) cost of the enhancement if it dissuaded the newcomer from entry, even though the benefits exceeded the costs of this entry. Periodic auctions of capacity rights, if feasible, would avoid this, the risk arising from the uncertainty about their future yield being borne by the infrastructure provider.

**The time dimension.**

The discussion so far has run in static terms. It has ignored the fact that demand can vary by hour, day, week and year and may grow through time. To take account of within-year variations would make the exposition more complicated, but distinguishing, for example, peak, shoulder and off-peak periods would not alter it essentially. Long term growth, on the other hand, does raise a significantly new and
additional problem. What has to be considered now is not only how enhancements are
to be paid for, but also their timing. (This point serves to remind us that future costs
and outputs have to be forecast and that such forecasts are uncertain.)

If demand is growing, there will come a time when a capacity increase
becomes desirable and a time when an increase becomes profitable for the
infrastructure provider (often much later). This suggests that the present worth of
time-streams of costs and revenues should be looked at, not just a series of single
annual values.\footnote{This is what UK Regulators do, though only to a limited extent, in that they fix \( P_0 \) and \( X \) so that the
present worth of forecast revenues just covers the initial Regulatory Asset Value plus Capex and Opex
for the next few years less the terminal Regulatory Asset Value. However the price limits that they
impose have to be consistent with acceptable levels of critical annual financial indicators. None of them
has dealt explicitly with a case of the sort discussed here.}

This means that the best way of treating the problems of getting users to pay
some part of the capital cost of enhancements is to express optimisation in present
worth terms as subject to a revenue constraint, requiring the present worth of
infrastructure provider’s revenue to cover the present worths of both operating cost
and capital charges, plus a bit of profit. Both the target variable, the sum of producer
and consumer surpluses and the constraint thus \textit{need to be formulated in present worth
terms rather than separately for each year.}

Consider a case like that illustrated in the first diagram, where the price
charged to the user is always kept equal to marginal cost. Growth in demand will now
make capacity expansion first become desirable in the year when:

\begin{enumerate}
\item The present worth of the future time stream of \( M \)-\( P \) (the annual
net saving in the infrastructure provider’s cost) plus \( N \) (the gain in
consumer surplus) exceeds the capital cost of the enhancement,
\end{enumerate}

but will first satisfy the present worth budget constraint in the year when:

\begin{enumerate}
\item The accumulated worth of the past excess of \( L \) over capital charges
for the existing assets including a bit of profit exceeds or equals the
capital cost of the enhancement.
\end{enumerate}

If condition (2) is met in an earlier year than (1), but (1) determines when the
expansion is undertaken, then the infrastructure provider will earn more than the
budget constraint requires. If it is later, the infrastructure provider’s earnings will fall

\footnote{This is what UK Regulators do, though only to a limited extent, in that they fix \( P_0 \) and \( X \) so that the
present worth of forecast revenues just covers the initial Regulatory Asset Value plus Capex and Opex
for the next few years less the terminal Regulatory Asset Value. However the price limits that they
impose have to be consistent with acceptable levels of critical annual financial indicators. None of them
has dealt explicitly with a case of the sort discussed here.}
short of the budget constraint. Thus, with price always equal to marginal cost, before it is commissioned, the infrastructure provider may save up more than enough or insufficient to pay for the enhancement.\textsuperscript{18}

The same holds in a case like that examined in the third diagram but with multiple users. Each of them will treat average congestion cost as an addition to the price he pays. Unconstrained optimality would instead require each of them to take account of marginal congestion cost, paying a price which covered the excess of marginal over average congestion cost, in addition to marginal operating cost. The result could again be that the present worth of the infrastructure provider’s revenue, exceeds or falls short of the present worth of operating cost, capital charges on existing assets and the present worth of the cost of increasing capacity.\textsuperscript{19}

Now suppose that there is a budget constraint. If it were formulated separately for each year, then before capacity is increased, price would be kept too low to reflect marginal operating cost plus marginal minus average congestion cost. Conversely, afterwards, price would be much too high. But under a financial constraint

---

\textsuperscript{18} If the requirement that price always equal marginal cost is dropped, or if marginal cost continues to exceed average cost (though by less) after the additional capacity is commissioned, the optimal budget-constrained solution might be to commission the additional capacity before (2) is satisfied, with some of its capital cost recovered by an excess of price over average cost in succeeding years. Thus the first order conditions for maximisation of the present worth of the sum of producer and consumer surpluses subject to the present worth of revenue covering the present worth of capital charges on the existing capacity plus the present worth of the capital cost of the addition to capacity would be more complex. But the basic point remains that the budget constraint should be formulated in present worth terms, not separately for each year, so that part of the cost of a lumpy investment should be met from the congestion rents of the years before it is commissioned.

\textsuperscript{19} T.H Oum and Y Zhang have investigated these possibilities in their paper “Airport Pricing: Congestion Tolls, Lumpy Investment and Cost Recovery” Journal of Public Economics, 43 (1990) pp 353-74. They assume that airport charges equal marginal congestion cost, growing annually as demand expands. The optimal year for introducing an additional lump of capacity, when demand is growing monotonically, is when its annuitised plus operating cost first falls below the saving in congestion cost that it will enable. This, they point out, holds whatever was the history of congestion revenue in preceding years. This could have been such as to yield far more than the cost of the investment or much less. They then provide some ingenious numerical simulations, applying a simple queueing model and using data appropriate for Toronto airport in the mid ’eighties and estimated delay costs for passengers and aircraft. Most of their computed illustrative results showed that congestion charges preceding investment in an additional runway would yield much more than it cost.
reformulated in present worth terms instead of annually, the Regulator could allow some of the investment cost to be met in advance, by allowing a substantial increase in the price towards marginal operating cost plus the excess of marginal over average congestion cost. This would leave less of the investment cost to be met afterwards, allowing a reduction in the price towards marginal operating cost plus the new, lower, level of marginal minus average congestion cost.

A formal treatment of this problem is presented in the Appendix. It makes various simplifications and rests on the assumption (imposed by recourse to algebra) of complete knowledge of future demand and costs.

Part of the revenue generated prior to the investment could be set aside in a fund not belonging to the infrastructure shareholders. Then, when the expenditure on the lumpy increase in capacity commences, part of it would met from this fund. This part of the capital expenditure would not be added to regulatory capital, which would be increased only on account of the remainder of the expenditure, financed by the infrastructure provider. The two effects of this would be:

1) To raise charges before the new capacity is commissioned, reducing demand and so diminishing economic inefficiency.
2) To lower charges after the new capacity is commissioned, reducing their excess over the new reduced level of marginal congestion costs, so diminishing economic inefficiency.

Under this scheme, the present worth of charges paid by the users should be approximately the same as if charges were increased only after the new capacity is commissioned, then being made sufficiently high to cover the whole of the annuitised cost of the capacity increase in addition to the capital charges on the existing capacity.

IV Conclusions

A lumpy infrastructure capacity enhancement will eliminate much of the high revenue that results from marginal cost pricing before the enhancement. Thus, while the initial amount of that revenue can signal the need for the investment, the subsequent revenue will either fall far short of what is needed to meet the annualised investment costs or, being obtained by imposing a price much exceeding marginal congestion costs, will result in inefficient use of the enhanced capacity. To avoid this and for users to meet the cost of enhancements, a contract between the infrastructure provider and the users may be necessary. Such a contract may be difficult to agree
when there is more than one user and will raise even more difficult problems if new
users appear who were not parties to the original contract. Use of the surplus raised
previously by the intial charges to finance at least part of the cost of enhancements
should be considered. This follows naturally from approaching the problem as
optimisation subject to a budget constraint expressed in present worth terms. A budget
constraint set out as a series of annual targets is thus inappropriate in the context of
lumpy infrastructure investment.

Appendix.
Consider the simple case where a single type of output is produced at a
constant variable cost per unit of \( v \) and where congestion cost, borne entirely by the
users, is an increasing function of a period’s output, \( X_t \). The infrastructure provider’s
revenue requirement in each period is initially \( vX_t \) plus remuneration of his sunk
capital costs; after capacity has been increased its capital costs will also have to be
remunerated. The present worth of the financial charges on the sunk costs of the
capacity existing at the beginning is \( S \), and the present worth of the cost of the
capacity increase is \( I \). This capacity increase is planned to occur at the end of period \( r \),
so that the congestion cost function changes from \( C^b(X_t) \) before then to \( C^a(X_t) \) after.
Demand for the output in each period is a decreasing function of both price and of
congestion cost, represented for simplicity by using a Willingness to pay function
\( W_t(X_t) \) (the integral of a demand function) and assuming that the demand price is
\( dW_t/dX_t \) minus average congestion cost \( C(X_t) \).

The maximand is the present worth of Willingness to pay over periods
\( 1 \) to a final period \( f \), less the present worth of variable and congestion costs and the
investment cost of the capacity addition.

\[
\text{Maximise } \sum_{i=1}^{f} \frac{W_i(X_t) - vX_t}{(1 + i)^t} - \sum_{i=1}^{r} \frac{C^b(X_t)}{(1 + i)^t} - \sum_{r+1}^{f} \frac{C^a(X_t)}{(1 + i)^t} - I
\]

The present worth revenue constraint requires the present worth of revenue in
excess of variable costs over the whole set of periods \( 1 \) to \( f \) to cover not only the
present worth of the investment cost, \( I \), of the capacity increase (net of its residual
value at the end of the final period) but also the present worth, \( S \), of the financial
charges on the sunk costs of the capacity existing at the beginning of period \( 1 \).
Furthermore, in each period, output must be such as to equate the demand price with 
the price, $P_t$. Thus maximisation is subject to the constraints:

$$\sum_{t} X_t \frac{P_t - v}{(1 + i)^t} \geq (S + I) \quad \text{dual } R $$

$$\frac{dW_t}{dX_t} - \frac{C^b(X_t)}{X_t} = P_t \quad \text{for } t = 1 \text{ to } r \quad \text{duals } \alpha_t$$

$$\frac{dW_t}{dX_t} - \frac{C^a(X_t)}{X_t} = P_t \quad \text{for } t = r+1 \text{ to } f \quad \text{duals } \alpha_t$$

The Lagrangian, $L$, is:

$$\sum_{t} \left( \frac{W_t(X_t) - vX_t - \sum_{t}^{r} C^b(X_t) - \sum_{r}^{f} C^a(X_t)}{(1 + i)^t} \right)$$

$$+ \alpha \left( \frac{dW_t}{dX_t} - \frac{C^b(X_t)}{X_t} - P_t \right) \ldots \alpha \left( \frac{dW_f}{dX_f} - \frac{C^a(X_f)}{X_f} - P_f \right)$$

Setting its derivatives with respect to $X_t$ equal to zero yields the first-order 
conditions for $t = 1$ to $r$, with a similar result for $t = r+1$ to $f$:

$$\frac{dW_t}{dX_t} - v - \frac{dC^b}{dX_t} + R(P_t - v) \quad \left( \frac{d^2W_t}{dX_t^2} + \frac{1}{X_t} \frac{dC^b}{dX_t} - \frac{C^b(X_t)}{X_t^2} \right) = 0$$

The first-order conditions with respect to $P_t$ for each $t$ are:

$$\frac{RX_t}{(1 + i)^t} = 0$$

Using these to substitute for $\alpha$ in the preceding equation, gives, for each $t = 1$ to $r$:

$$\frac{dW_t}{dX_t} - v - \frac{dC^b}{dX_t} - R \left[ P_t - v - \left( X_t \frac{d^2W_t}{dX_t^2} + \frac{dC^b}{dX_t} - \frac{C^b(X_t)}{X_t} \right) \right]$$

$$\left( \frac{1 + i}{1 + i} \right)^t = 0$$

and correspondingly for each $t = r+1$ to $f$, $C^a$ again replacing $C^b$.

So, for each $t = 1$ to $r$:

25
\[
\frac{dW_t}{dX_t} - \nu - \frac{dC^b}{dX_t} = -R \left( P_t - \nu - X_t \frac{d^2W_t}{dX_t^2} - \frac{dC^b}{dX_t} + \frac{C^b(X_t)}{X_t} \right)
\]
giving:
\[
P_t = \nu + X_t \frac{d^2W_t}{dX_t^2} + \frac{dC^b}{dX_t} - \frac{C^b(X_t)}{X_t} - \left( \frac{dW_t}{dX_t} - \nu - \frac{dC^b}{dX_t} \right) \frac{1}{R}
\]
or, rearranging:
\[
P_t = \left[ \nu + \frac{dC^b}{dX_t} - \frac{C^b(X_t)}{X_t} \right] + \left[ X_t \frac{d^2W_t}{dX_t^2} - \left( \frac{dW_t}{dX_t} - \nu - \frac{dC^b}{dX_t} \right) \frac{1}{R} \right]
\]
and correspondingly for each \( t = r+1 \) to \( f \), \( C^a \) again replacing \( C^b \).

The first term on the right is what the optimal price would be in the absence of the revenue constraint, namely variable cost plus the marginal externality, i.e. the excess of marginal congestion cost \( \frac{dC^b}{dX_t} \) over average congestion cost, \( \frac{C^b(X_t)}{X_t} \). The second term on the right is therefore the required difference of price from that optimal price. It is the excess of marginal revenue over \( 1/R \) times marginal social gain. \( R \), being the dual of the revenue constraint, reflects the marginal effect upon the maximand of a tightening of that constraint (higher \( S \) or \( I \)). The advent of additional capacity at the beginning of period \( r+1 \) will lower the price because of the reduction in congestion cost from \( C^b(X_t) \) to \( C^a(X_t) \).

Further analysis would show the condition for the optimal timing of the additional capacity.\(^{20}\)

\(^{20}\) An analysis of this problem under different assumptions, namely that marginal costs are borne by the provider, that a single price has to be fixed for the whole interval but that there is no financial constraint, was provided in the Appendix to Ralph Turvey and Dennis Anderson, *Electricity Economics*, (Johns Hopkins University Press, 1977) and, more formally, by Ray Rees in his paper “Indivisibilities, Pricing and Investment: The Case of the Second Best” *Journal of Economics, Zeitschrift für Nationalökonomie*, 1986. Rees continues by introducing a present value financial constraint and considers the case (still with all costs borne by the provider) where one price is to be charged until the lumpy capacity expansion is installed and another price thereafter. He then analyses the optimal timing of the capacity increase (here taken as given) which involves treating time as continuous and demands the application of higher mathematics than I can muster.