Optimal Charging Arrangements for Energy Transmission: Final Report

Ross Baldick,[1] James Bushnell,[2] Benjamin F. Hobbs,[3] and Frank A. Wolak[4]

Report Prepared for and Commissioned by
Project TransmiT, Great Britain Office of Gas & Electricity Markets

1 May 2011

[1] Professor in the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX, USA; IEEE Fellow; Former Chair of the System Economics Subcommittee of Power Systems Analysis, Computing, and Economics Committee of the IEEE Power Engineering Society.

[2] Cargill Endowed Chair in Energy Economics and Director, Biobased Industry Center, Iowa State University, Ames, IA, USA; Member, Market Surveillance Committee of the California Independent System Operator.

[3] Theodore K. and Kay W. Schad Professor of Environmental Management and Director, Environment, Energy, Health & Sustainability Institute, The Johns Hopkins University, Baltimore, MD, USA; IEEE and INFORMS Fellow; Member, Market Surveillance Committee of the California Independent System Operator.

[4] Holbrook Working Professor of Commodity Price Studies in the Department of Economics at Stanford University and Director, Program on Energy and Sustainable Development (PESD) in the Freeman-Spogli Institute (FSI) for International Studies, Stanford University, Stanford, CA, USA; Former Chair of the Market Surveillance Committee of the California Independent System Operator.

Summary of Recommendations

Four principles underlie our recommendations for Great Britain's (GB) transmission charging arrangements:

1. Charges for the usage of the network should reflect the incremental costs imposed by that usage.

2. Charges to recover historic (sunk) capital costs and other fixed costs should distort usage as little as possible.

3. Environmental objectives are most efficiently pursued through mechanisms that directly address those objectives.

4. Objectives for equitable distribution of costs and risks can be addressed while still preserving incentives for efficient use of the network.

Our recommendations are as follows. First, incentives for efficient congestion relief should be provided by energy prices that are differentiated locationally. At a minimum, there should be transparent zonal or nodal pricing of energy in the balancing market in which:

- all imbalances at a particular network location are cleared at the same price,
- a balanced schedule submitted to a locational market would pay (or receive) the difference in prices between the locations of injections and withdrawals, and
- prices are determined by the interaction of offers by incremental supply and bids by incremental demand, submitted schedules, and the physical and security constraints of the network. A price at a location will then reflect the as-offered cost of meeting an incremental MW of demand (or the value of an incremental MW of supply) at that place and time.[5]

One possible implementation of a locational pricing scheme that would be minimally disruptive to present trading arrangements would be to extend the current balancing mechanism to report locational prices and use them for settlement. Locational pricing could also be extended to creating a formal day-ahead market in which producers either submit energy-only offers (internalizing start-up and other lumpy costs as well as operating constraints), or, alternatively, producers could submit information about these costs and their operating constraints, in which case the market is cleared using a security-constrained unit commitment.

The reason for clearing all imbalances at a particular location and time using the same price stems from the fungibility of electricity. Paying different up- and down-balancing prices to different generation units at the same time and location (or different locations if there is no congestion) ignores the fact that injections from those generation units are perfect substitutes for meeting load and should therefore be paid the same price. Otherwise, the load would purchase more energy from the lower-priced generator and less from the higher-priced generator, and thereby reduce its total wholesale energy purchase costs.

[5] See Appendix A.1 for an introduction to the philosophy and mechanics of locational pricing.

Zonal pricing may suffice for the present system, in which most generation is from thermal power plants and the network can be reasonably approximated as radial in form. However, as the amount of intermittent and dispersed renewable generation grows and the grid is expanded, congestion patterns will become more variable and unpredictable and security constraints will become more complex. As a result, intrazonal congestion will become increasingly unmanageable. Therefore, we anticipate that a transition to locational pricing at the nodal level will eventually become necessary. The experience with zonal pricing in organized markets in the US indicates that zonal pricing is a transition phase that ends with implementation of fully disaggregated locational pricing by node.[6]

Second, we recommend that the philosophy of transmission construction be changed. We would characterize the present approach as largely one of a "transmission follows (is driven by) generation" approach or "price, connect, and manage". Simply stated, fixed Transmission Network Use of System (TNUoS) charges are applied annually; generators choose where they connect; the expense of any resulting transmission congestion is socialized; and the grid is then assumed to be expanded to follow generation if the upgrade can be economically justified. We instead recommend adoption of a philosophy of "generation follows transmission" or "plan and then price" that recognizes that transmission additions generally have a much longer lead time than generation additions.[7]

Under this approach, a rational planning process would be undertaken in which transmission reinforcements are made that are anticipated to yield the lowest overall expected system-wide cost of generation and transmission, subject to environmental and security constraints and accounting for generation investment and operating decisions that result from the upgrades. Generators would not be subject to locationally differentiated fixed annual charges for sunk transmission costs, with the possible exception of shallow connection charges and payments for transmission reinforcements requested by generators. Instead, projections of locational energy prices would incent siting of generation in the most economically efficient locations.

The "generation follows transmission" philosophy is partially implemented in GB, in the form of the Transmission Investment Incentives project that provides funding for "anticipatory" transmission investment. We recommend that such anticipatory planning, conducted using appropriate benefit-cost analysis methods and in a public process, be the primary basis for transmission planning. Such a forward-looking transmission planning process implies that it is an extremely rare event that expansion of the backbone transmission network occurs as a subsequent response to the entry decision of a generation unit owner. Such an event would be considered a failure of an anticipatory transmission expansion planning process.

---

[6] See, for instance, the experiences in California and Texas. In those states, rapid growth of intrazonal constraint costs and issues related to zone definition motivated transitions to nodal pricing (Section 3.3, Appendix A.3).

[7] Not only are transmission lead times greater than for transmission, but locational variations in total delivered cost of energy are to a great extent determined by transmission. Although the total cost of generation is much more than the cost of transmission, variations in generation expenses due to locational factors is generally a small fraction of total generation costs. It is increasingly recognized that the cost of transmission is the major component of renewable integration costs (E. Kahn, "Wind Integration Studies: Optimization vs. Simulation," *The Electricity Journal*, Nov. 2010).

Third, we recommend that all of costs of the existing network be allocated to load, rather than the present roughly one-quarter/three-quarters split between generation and load. In the end, costs paid by generators are passed on to consumers in the prices charged by generation unit owners, which can also lead to distortions from the least cost supply of wholesale energy. Together with elimination of locational differentiation, full allocation of these costs to load will considerably simplify the TNUoS system and limit the risk that the transmission charging mechanism reduces the efficiency of the wholesale energy market. Furthermore, this change would bring GB's charging system into closer alignment with those in neighboring countries, which will help level the playing field in the international competition between GB and non-GB generation.

Fourth, we recommend that a system of financial transmission rights (FTRs) be created to enable generators to hedge uncertainty in congestion costs. These rights would be defined as point-to-point or point-to-region rights that would pay the difference between one location's price and the price at the second location or a weighted average of prices for several locations, respectively. We recommend that these rights be sold or auctioned by the system operator, and the revenues refunded to loads. If socializing congestion costs is an important objective, then allocation of some FTRs to generation unit owners can be used to achieve this, as long as this is done in a manner that does not impact generation unit siting and retirement decisions.

Fifth, we recommend that the possibility of transmission projects sponsored by generator unit owners or independent transmission providers be allowed. If a group of generator unit owners believes that the congestion relief benefits will more than compensate for the cost of a line that is not in the central transmission plan, then they should be allowed to sponsor construction of the line, receiving in return the additional financial transmission rights made possible by the upgrade. This will provide a check against "false negatives", in which a centralized planning process overlooks an economically beneficial reinforcement. Of course, this does not protect against "false positives" in which the plan endorses construction of a transmission facility that is in fact not economically or otherwise justified.

Because (1) National Grid is implementing optimal power flow software from the New York ISO that can automatically generate locational marginal prices, and (2) software used elsewhere for managing and settling LMP-based markets could be adapted for use in GB, the cost of transitioning to a balancing market with locational prices need not be high. We encourage Ofgem to investigate the practicality and benefits of implementing such a market. Such an investigation would inform UK involvement in ongoing EU-level discussions of a 2014 target electricity model, which indicate a strong interest in developing short-term wholesale electricity markets with spatially differentiated prices within nations.

1. Introduction

Project TransmiT is a public review of the principles and mechanisms of the Great Britain (GB) electricity and gas transmission charging and connection system, and is being conducted by the GB Office of Gas and Electricity Markets (Ofgem). Its goal is to examine whether the present charging arrangements are responsive to the growing demands on the power and gas system for economically efficient, reliable, and transparent incorporation of increased amounts of low carbon and, especially, renewable energy sources. New renewable resources will often be sited in remote locations not presently served by the high voltage grid, and their intermittency implies that the power grid will experience different patterns of usage and congestion than in the past. At the same time, the GB power market is increasingly integrated with markets in continental Europe and Ireland.

These developments raise questions as to whether the present charging arrangements will encourage efficient siting and operation of generation, cost-effective decarbonization of the power sector, and a level playing field for GB generation in the international markets. Thus, Project TransmiT is a timely review of the principles and procedures of the charging and connection systems administered by National Grid Electricity Transmission, P.L.C. and National Grid Gas, P.L.C.

Project TransmiT has involved a public consultation process, including a call for evidence, as well as commissioning of reports by three teams of academic consultants. In response to the call for evidence, approximately sixty stakeholder and other submissions were received in November, 2010. The public consultation process is to continue through the spring of 2011 with an evaluation of options for retaining or changing transmission charging and connection arrangements. This process included a roundtable discussion of the draft consultants' reports on 4 March 2011, to be followed by publication of the final versions of those reports. The Project TransmiT recommendations will then be published by Ofgem in the summer of 2011.

As one of the teams of academic consultants, we have been asked by Ofgem to provide three deliverables: a draft statement of principles (submitted in December 2010), a draft final report (delivered in February 2011), and a revised version of the final report (this submission). This final report is to provide our views on the following issues and questions:

- Appropriate guiding principles for transmission charging that are consistent with meeting the objectives of economic efficiency in operations and investment, facilitation of carbon reduction, integration with international power markets, security of supply, cost-recovery, and ease of administration.
- The broad building blocks of a suitable target charging model that would best achieve the objectives as a whole, accounting for possible trade-offs among those objectives.
- The interdependencies between our recommended target charging model and other aspects of the regulatory regime for electricity and, if relevant, gas networks, including cross-European regulatory and policy developments.

We would like to express our appreciation to the stakeholders and observers who made submissions in November 2011 in response to the call for evidence, as well as for the extensive comments made in response to our draft report. We have found these contributions to be extremely helpful contributions to our understanding of the issues and concerns, and we are grateful for the considerable effort that the submitters put into preparing them. We are

also grateful for the tremendous help provided by Ofgem staff in response to our requests for information.

In this final report, we first summarize several broad principles that we believe should be satisfied by an efficient charging system for energy networks (Section 2). Then in Section 3, we discuss short-run congestion management. We discuss incentives for generation and transmission investment in Section 4, proposed modifications to the charging methodology in Section 5, and consistency with policy objectives in Section 6. The Appendices include supplementary details. These include tutorial descriptions of locational marginal pricing (Appendix A.1) and the interaction of LMP with market power and related issues (Appendix A.2). Appendix A.3 describes experiences with zonal and locational pricing in the Texas (ERCOT) market, noting that large intrazonal constraint costs forced a transition to full nodal pricing.

2. Principles for Charging Systems

Enormous investments will be needed to achieve the UK's low carbon and renewable energy objectives. For this reason, it is imperative that each of these investments achieve the greatest possible value. It is important, for instance, that network charges motivate investors in renewable energy to choose the mix of technologies and locations that achieve the UK's renewable targets at the lowest possible cost. Ideally, transmission charging arrangements should encourage the market to make the efficient tradeoff between more productive but also more remote renewable facilities with local investments that may be less productive, but impose lower transmission costs.

Two decades ago the UK began its transition to reliance upon market mechanisms for stimulating investment and creating incentives for the efficient production and consumption of electricity and natural gas. Market mechanisms rely, at least in part, on communicating prices that reflect private benefits and costs through the interaction of production and demand. For contestable activities such as the construction and operation of power plants or the development of natural gas wells, markets have been a powerful force in increasing efficiency in the energy sector.

These competitive activities need to operate within the context of transmission and distribution networks where formal regulatory mechanisms are used to plan, expand, and price access to these networks. This means that, in order to reap the full benefits of markets and competition in the energy sector, the competitive activities must be coordinated with regulatory oversight and planning of the transmission and distribution infrastructure.

The growing prominence and urgency of environmental goals, particularly mitigating greenhouse gas emissions, does not change this picture fundamentally, although it does change many of the standard assumptions in transmission planning, such as an emphasis on peak load conditions, as opposed to peak renewable production conditions. Market mechanisms are still powerful tools, and can applied to both maintaining adequate supplies of energy and the mitigation of undesirable emissions. The environmental goals add a new "value" to certain types of energy, while reflecting additional costs inherent in others. Again, these activities are happening within a complicated network that, now even more, requires coordination between infrastructure planning and the investment activities of individual market actors.

In later sections of this report, we will discuss aspects of the network infrastructure planning problem. Before that discussion, we will first cover a few high level principles that guide our thinking on how to approach these questions. The first two principles reflect the conventional wisdom of economics on the efficiency of marginal cost pricing. The last two principles concern the least cost achievement of environmental and equity goals.

Underlying all of our principles is a broad notion of economic efficiency: the maximization of the net benefits from the electricity market subject to constraints posed by policies addressing environmental, renewable, and equity goals. We do *not* frame our discussion as a debate over efficiency *versus* the other social goals. Rather, good market design follows from a broad definition of efficiency that focuses on achieving policy goals at minimum cost to the markets. This makes the economic pie as large as possible; compared to less efficient policies, this allows society to either meet the goals at lower costs, or to set and achieve more aggressive goals at the same cost.

Principle 1: Charges for the usage of the network should reflect the incremental costs imposed by that usage.

This principle is a network-specific version of the general economic result that prices should reflect the incremental (or marginal) costs of providing a good or service, including environmental and other external costs. These are the prices that promote efficient consumption of the goods and services. Prices that are too low (*e.g.*, set below incremental costs) incent over-consumption. This in turns leads to either an inefficient over-use of resources, as costs of meeting demand are greater than the value to consumers, or a shortage of resources as producers are unable to meet demand at those prices. Prices that exceed incremental cost, on the other hand, discourage consumption whose value more than compensates for the cost of production.

In the network context, as we will argue, this means that incremental costs of using the network should be reflected in balancing charges. The short run marginal cost structure of energy transmission can be highly variable and usually driven by the presence or absence of congestion together with the cost of losses. When there is congestion, additional demand for the network imposes a high cost on other users. When there is no congestion, additional demand imposes very little incremental cost, aside from resistance losses.[8] Following this logic, fees for using a network during unconstrained periods should be minimal, while they may be quite high during periods of congestion.

It is misleading to think of the capital cost of constructing existing or new infrastructure as the cost created by congestion. In most cases, the "costs" imposed by congestion is itself derived from the lost opportunities of those who cannot use the network. For example, if transmission congestion forces an expensive plant to operate in place of an inexpensive one, the differences in these generation costs are the true costs of congestion.

If these congestion costs are not properly reflected in the costs of using the network, the common result is that in some locations demand exceeds the supply that the network can deliver from willing sellers, while in other locations there is a surplus of supply. Given the fixed capability of a network, this usually leads to a rationing of access to the network in one form or another, and redispatch of production via constrained-on and constrained-off generation units. If not performed systematically based on relative costs, this rationing can often result in inefficient use of the network, which implies higher than necessary costs of meeting system demand.

Principle 2: Charges to recover historic (sunk) capital costs and other fixed costs should distort usage as little as possible.

The complementary principle to that of marginal cost pricing is that costs that are *not* incremental should *not* be present in the prices of goods and services. If fixed costs are included into the price of a service but there is unused capacity, consumption is too "low," in

[8] The relative importance of congestion and resistance losses depends on the system configuration and conditions; for instance, one study of California locational marginal costs found that they were of roughly equal importance (J. E. Price, Market-Based Price Differentials in Zonal and. LMP Market Designs", *IEEE Transactions on Power Systems*, 22(4), November 2007). Meanwhile, marginal costs of natural gas transmission would include not only congestion but also relatively minor costs for operating compressors to maintain desired pressure levels.

the sense that there are customers whose value for the service is much higher than the marginal cost of providing it, yet are not consuming it because of these high prices. For capital intensive projects, this can be quite counter-productive, as building the facilities can lead to high capital costs, and including those costs in prices can leave the expensive facility unused.

This is the obvious danger in energy networks where, absent congestion, marginal costs are typically low and capital costs quite high. Yet, despite the efficiency concerns, capital costs do need to be recovered *somehow*. One way is through scarcity pricing, when prices rise to ration demand when facility capacity is fully used. But revenues received during scarcity periods may be insufficient. Absent subsidies from public funds (which can create their own distortions), economists typically argue for the recovery of costs to be achieved in a way that minimizes the inevitable distortions in economic activity that such charges must create.

This is a well known problem in the context of pricing to recover fixed costs. Following the insights of what is known as *Ramsey Pricing*, the general strategy is to direct the higher charges toward the less price-responsive, or price elastic, activities. To summarize, unlike marginal costs, which are ideally fully reflected in prices, fixed and capital costs should be recovered in a way that maximizes total economic surplus subject to full cost recovery.

The first two principles above are concerned primarily with economic efficiency, defined as the ability to provide the appropriate level of network services at the lowest possible cost to final consumers. The next two principles concern goals beyond simply providing network services, and address approaches for meeting those goals at the minimum cost.

Principle 3: Environmental objectives are most efficiently pursued through mechanisms that directly address those objectives.

In the UK and many other parts of the world, ambitious environmental goals, particularly the reduction of greenhouse gas emissions, are coming to dominate policy objectives in the energy sector. As a means of obtaining these objectives, many economists have argued for policies that focus on rewarding the specific attributes that we as a society wish to expand, as well as discouraging those we wish to reduce. In many cases, environmental policy has promoted broad classes of activities, for example promoting renewable energy production, or discouraged other broad classes of activities, such as reducing car ownership or vehicle miles-travelled. In many cases these policies address activities that are closely related to the desired attributes, but not necessarily focused on the ultimate goals. For example, renewable energy production yields little or no greenhouse gases, but this is true also of some non-renewable technologies.

The risk of policies that are too broad in their application or that focus on rewarding specific activities rather than the ultimately desired attributes is that they can produce consequences that can work at cross-purposes to the objectives behind the policy in the first place. As an example, subsidizing renewable energy and fuels through tax credits can lower the costs of desirable fuels, but also contribute to lowering the cost of consuming all energy and thus promote more energy use.

In particular, adjusting transmission charges at particular locations is likely to be an inefficient way of promoting greenhouse gas reduction and renewable energy objectives. This is because carbon emissions and renewable energy output depend on *how* not *where* electricity is produced. At a given location or region there is generally more than one way to produce power.[9] It is therefore more efficient, and certainly more transparent, to promote directly the desired attribute of production (whether by pricing carbon or paying for high renewable content) rather than to attempt to subsidize it by adjusting transmission charges away from the cost of providing transmission services.

The goals of renewable and low-carbon production are more likely to be efficiently achieved if the policies to promote them focus specifically on these goals. Mandates for renewable production and markets such as the EU emissions trading system do this by placing value on desired attributes (renewables) and penalizing undesirable ones (CO_2 emissions). If policymakers commit to enforcing these kinds of mechanisms, they can be effective at achieving their goals without additional support through indirect, less targeted, and, as will be explained below, ultimately less cost-effective policies.

The other benefit of market-based mechanisms for renewable and climate policies is that they provide information about the costs of those policies and allow them to be compared with other options that address the same goals. Layering indirect supports on top of such policies distorts this information. To the extent that indirect support, through infrastructure subsidies for example, creates an artificial impression about the cost of producing low-carbon energy, this can distort the consumption of energy and discourage conservation and energy-efficiency efforts.

In summary, there are a suite of attractive policy tools already in place for directly addressing renewable and climate goals. We believe these goals would be most efficiently met if these policies are allowed to work as intended, without distortions introduced by indirect support such as discriminatory network pricing. In this way, economic efficiency in the broad sense (maximization of market benefits while meeting policy goals) is more likely to be achieved.

Principle 4: Objectives for equitable distribution of costs and risks can be addressed while still preserving incentives for efficient use of the network.

In addition to environmental and efficiency goals, there are also stakeholder preferences for less volatile and more predictable network costs, and promoting an equitable distribution of those costs. The goals of designing a charging structure that is both efficient and equitable need not substantially conflict with each other. Whatever the societal definition of "equitable" may be, it can be at least partly accommodated with mechanisms that still preserve the correct marginal incentives for users of the system. Thus, a goal of "socializing" the costs of network investments can be largely honored without deviating from setting the efficient marginal price signal.

[9] For instance, a proposal to promote renewable energy by subsidizing transmission charges to regions rich in such resources will, in general, increase the social cost of meeting the renewable target. This is because renewable developments in areas that could also host other power generation sources (say, combined cycle units or distributed generation) would not receive that subsidy. Subsidizing some renewable energy sources but not others will, in general, shift the mix of such sources to one that is less efficient because the effective marginal payment for the "renewableness" of a resource will then depend on location.

For example, the risks of very high consumer bills associated with time-varying electricity pricing can be largely eliminated when coupled with instruments that bestow the right (but not requirement) to consume a certain amount of energy at a fixed price. This can closely resemble an individualized contract for difference (CFD). An individual can have the "right" to 10 kWh of energy at a fixed price. If they choose to consume more, they must pay the marginal price for the additional power. If they consume less, they can implicitly "sell back" that energy to the network by consuming less than their allotted amount. Once a party has "ownership" of the policy in this way, they can choose individually whether consumption or resale is the best choice.[10]

Similar examples, such as the allocation of emissions allowances, abound in the policy arena. As long as the allocation quantity is fixed, the recipient must confront the opportunity cost of consuming the allocated good, rather than reselling it. In this way, incentives for efficient utilization of resources are preserved, but a wide variety of risk-management and equity goals can be addressed. By contrast, bestowing the "right" to consume unlimited quantities at a fixed price can address equity concerns, but creates problematic incentives for over-consumption.

As we discuss below, in the context of energy networks, the instrument of choice for heding transmission costs is a financial transmission right. In electricity networks, such rights capture many of the desirable quantities described in this section, and we will argue that their deployment in the GB electricity network can go a long way toward promoting both equity and efficiency goals.

[10] See S. Borenstein, "Customer Risk from Real-Time Retail Electricity Pricing: Bill Volatility and Hedgability," *Energy Journal*, 28(2), 2007, pp. 111-130.

3. Short-Run Congestion Management

If locational marginal pricing is not employed, the process used by a wholesale market to schedule generation units and set prices can often yield schedules that are infeasible given the configuration of the transmission network, geographic distribution of demand, and available generation capacity. The resulting prices paid to generation units can also create financial incentives for a unit to deviate from its final schedule. The wholesale market mechanism that is used to reconcile the financial market results with the physical constraints of the network is typically referred to as a *congestion management process*.

Wholesale markets differ substantially in terms of the process used to price and manage transmission congestion and other operating constraints. Some wholesale markets assume that all generation units are equally effective at meeting demand at any location in the transmission network and that all transmission paths have infinite transmission capacity. These "copper plate" assumptions imply that a single price can be set for energy for the entire market. However, because neither of these assumptions usually hold for the actual transmission network, these markets must rely on an extensive congestion management process to produce a feasible real-time generation dispatch. This can increase generation costs, depending on the quantity and predictability of schedule adjustments. This can also pose a threat to system reliability, by decreasing the ability of operators to predict the final configuration of schedules and dispatch and requiring more last-minute adjustments which might not all be feasible.

An intermediate market design might apply these two assumptions to smaller geographic areas, typically called *congestion zones*, within the larger geographic market, but recognize that the transmission capacity connecting these zones is finite and that losses occur when energy is transferred from one zone to another. This market design can require a less extensive adjustment process, because some of the infeasibilities are managed in the zonal scheduling and pricing process.[11] Finally, a *nodal-pricing*, or locational marginal pricing (LMP), market attempts to account for all relevant transmission network constraints in determining generation schedules and the prices paid to generation units. This market design requires the least amount of adjustments to obtain feasible final generation schedules.[12]

The combination of generation unit scheduling, pricing and congestion management methods that are chosen can influence both overall efficiency and system reliability. For example, a producer that knows one of its units is required to operate because of the configuration of the transmission network and geographic location of demand will take that fact into account in constructing the offer curves it submits to the wholesale market for its generation units. Depending on the details of the congestion management process, these actions can result in a significantly more costly dispatch of generation units, as well as a less reliable transmission network.

[11]Although the adjustments in generator schedules may be less extensive than under copper plate pricing, the zonal congestion management system itself can be as complex or more complex than a locational pricing system that provides individual bus-level prices. For instance, a very complex procedure has been proposed to produce consistent zonal prices for the four country-based zones in the Netherlands-Belgian-France-Germany market coupling now under development while maintaining feasibility of flows (M. Van Vyve, The impact of network modelling on prices in power markets," Presentation at *Recent Advances in Energy Economics, A Conference in Honour of Yves Smeers*, CORE, Université Catholique de Louvain, June 17-18, 2010). The computational procedures are much more complex and less transparent than nodal pricing.

[12] See Appendix A.1 for a tutorial on the calculation and interpretation of locational marginal prices (LMPs).

The fact that generation schedules must ultimately be physically feasible implies a trade-off in the design of a wholesale electricity market. As noted above, the more simplified the financial market, the more likely significant adjustments will be necessary, and the more opportunities producers will have to take actions that, although privately profitable, will reduce overall market efficiency. One way to avoid the need to make these adjustments is to construct sufficient transmission capacity and procure sufficient energy from reliable sources of production so that it is unnecessary to alter the schedules that emerge from a market that does not consider all network constraints. However, this approach can be extremely expensive, because of the cost of the additional transmission capacity necessary to obviate the need to make these adjustments to final generation schedules. Moreover, as more renewable resources are added to the transmission network, there will be a greater need to make adjustments to the generation schedules that emerge from wholesale markets that assume a simplified network structure.

This section discusses the trade-offs that exist in the design of a transmission charging and congestion management system for a wholesale electricity market. We proceed by first analyzing the current system, which combines the connect-and-manage interconnection policy with socialized balancing charges. We then discuss the benefits of integrating the congestion management process into the energy market through locational pricing of energy. We believe the current system can unnecessarily raise the cost of wholesale electricity, in part because it creates a bias in favor of investments in transmission capacity. It is also economically unstable in the sense that unhedgeable transmission charges can vary significantly across years because of changes in generation entry decisions and load growth. The problems with the current system are likely to be compounded by a large-scale expansion of the amount of intermittent electricity generation resources.

3.1. The current charging system.

The current system of charges is structured in a way that assumes, and emphasizes, investments in transmission capacity to ensure limited use of a congestion management mechanism. Generation unit owners are allowed to enter and connect their generation units to the transmission network with the expectation that sufficient transmission capacity will eventually be built to accommodate the full capacity of their generation unit. Moreover, the pound per kW-year TNUoS transmission charge paid by the generation unit owner entitles it to receive compensation, through constrained-off payments, for network congestion that prevents that unit from producing at its desired level. This payment scheme and the Balancing Services Incentive Scheme (BSIS) faced by NGET together provide a signal for the construction of new transmission capacity.[13]

[13] This signal for reducing congestion costs is a strong one. The following was stated by an eminent US economist who was also a member of the NGET board: “The electricity sector in England and Wales now has nearly a decade of experience with incentive-based transmission regulatory mechanisms governing the revenues that the National Grid Company (NGC) receives for providing services. There is much to learn from this experience. Of particular interest is the (BSIS) that provides NGC with financial incentives to reduce “transmission uplift” costs (these are costs associated with out-of-merit dispatch to manage congestion, thermal losses, and ancillary services costs). The important analytical insight embodied in this regulatory mechanism is that it gives NGC a financial stake in reducing the direct and indirect costs associated with the operation of and investments in transmission. The evidence indicates that the regulatory mechanisms applicable to NGC have encouraged substantial new investment in the network, facilitated generator interconnections, reduced transmission uplift costs, while increasing the reliability of the network.” ("Comments of Professor Paul L.

These capital investments imply that ultimately there will be a limited need to make adjustments to final generation schedules that emerge from the wholesale market to make them physically feasible. Under such a structure, additional capital investments will be necessary to build out the network to accommodate new load and new generation sources to limit the need to adjust schedules to obtain physical feasibility. As long as the network capacity is large enough, the market inefficiencies introduced by a simplified market structure that in effect assumes infinite transmission capacity and no transmission losses will remain small. Moreover, a system with strong incentives to limit congestion through investment would give the misleading appearance of little economic benefit from locational price signals. In summary, if investment is sufficient to ensure that infeasibilities are rare, there are few benefits to accounting for and explicitly pricing these infeasibilities in the wholesale market. Paradoxically, locational price signals—by providing generators with incentives to relieve congestion, and planners with better information about its consequences—can reveal that some of those same network investments may in fact be unnecessary.

Our understanding is that the current transmission paradigm emphasizes building sufficient transmission capacity so that congestion is rare; such a paradigm may have fit the many parts of the GB system of the past two decades reasonably well. As long as reasonably low-cost options for incremental transmission capacity expansion were available, there are minimal distortions from using a generation scheduling and pricing system that in effect assumes infinite transfer capacity across all paths in the network. However, there are several key transmission paths in the current system that already cause significant infeasibilities that must be dealt with. There are indications that more significant capital expenditures would be necessary to meet further growth in load. In addition, the future policy emphasis on intermittent and dispersed energy sources will mean that a paradigm of building the network to accommodate all possible patterns of generation with limited congestion would have excessively high incremental costs relative to the incremental benefits of the energy delivered.[14] All of these factors make the current day-ahead scheduling assumption of infinite transfer capacity between all locations in the transmission network increasingly unrealistic. Therefore, this growing divergence between how the wholesale market is assumed to operate and how the transmission network is actually operated is likely to get even larger.

The current approach provides a very strong incentive to limit the frequency and magnitude of congestion. It also has created an expectation by generation unit owners that sufficient transmission will be built to accommodate their output, with the transmission network operator bearing at least part of the cost of congestion if that transmission is not built. As a result, that approach creates incentives to over-emphasize capital investment, which would thereby raise the total cost of the transmission network while simultaneously lowering costs of managing transmission network constraints. A natural analogy is to decisions

---

Joskow," Docket No. RM99-2-000, Notice of Proposed Rulemaking on Regional Transmission Groups, Before the US Federal Energy Regulatory Commission, August 16, 1999.)

[14] See G. Strbac, C. Ramsay, & D. Pudjianto, "Framework for development of enduring UK transmission access arrangements," Centre for Distributed Generation and Sustainable Electrical Energy, Imperial College London, July 2007. We also note that, for example, for the planning of the "Competitive Renewable Energy Zone" transmission buildout in the Electric Reliability Council of Texas (ERCOT) system, a target was set on overall wind curtailment, rather than on delivering all possible patterns of wind generation.

concerning investment in generation capacity and the benefits of time-varying pricing to load. If enough baseload capacity is built to ensure that marginal generation costs seldom fluctuate with demand, then there is little point to dynamic pricing—*i.e.*, confronting customers wholesale prices that vary with real-time system conditions. However, the resulting total generation (fixed and variable) costs are much larger than in a system with dynamic pricing. The least-cost generation mix with dynamic pricing will not only have both low and high marginal cost units (with conversely high to low fixed costs), but demand will be managed so that loads can be reduced during times of tight supply. For the very few hours of high demand, it will almost always be cheaper to reduce demand rather than to build generation units that would be idle, say, 99.9% of the time.

The same logic applies to the case of transmission investments. Building transmission capacity that has excess capacity 99.9% of the time in order to limit the need to manage what would be only occasional congestion is likely to be far more expensive for electricity consumers than using locational pricing in the wholesale market to achieve feasible generation schedules. The increased transmission capacity must be paid for during all hours of the year, whereas the congestion costs must be paid by users of the congested transmission path only during the hours of the year when these infeasibilities arise.

In an environment of significant capital expenditures to meet both load and large growth in intermittent and dispersed renewable generation, minimal congestion is a luxury that GB consumers may no longer be able to afford. We believe that the transmission pricing paradigm needs to be adjusted in recognition that, at times, operating with significant congestion on parts of the grid may in fact be the system-wide least-cost solution. In contrast, continuing with the past approaches (of providing NGET with a strong incentive to build to reduce congestion together with a "connect and manage" policy that creates the strong expectation that transmission will be built to follow generation) will, in our opinion, likely yield total costs of generation and transmission that are considerably higher than the least-cost solution that allows congestion to occur when it is economic to do so.

The current approach to transmission pricing sets annual locational prices for access to the transmission network that are fixed for the year based on system conditions that rarely if ever occur during the period the prices are applied. Moreover, because the current methodology used to determine these locational prices depends on the location of demand,, the location of generation units, configuration of the transmission network, and other modeling assumptions, these prices can, and have, changed substantially across years. The increased penetration of renewable resources and transmission expansions necessary to accommodate them are expected to change these prices significantly in the future.

It is unclear what market efficiency goal is served by setting locational TNUoS charges that change yearly for existing generation units, because their entry decision was typically in the distant past. Doing so is also inconsistent with our Principle 2 that sunk costs should be recovered in a manner that does not distort usage of the transmission network. This is because a high enough TNUoS charge could cause an otherwise economic-to-operate generation unit owner to cease operating to avoid having to pay this charge to inject power in the transmission network.[15] The risk of future changes to the TNUoS at a location is also

[15] A decision to close a generating unit should consider not sunk capital costs, but instead the opportunity cost of the transmission network, in terms of accommodating generation from other units. If there is spare export transmission capacity from the unit's region, or the unit produces nonzero output at times when other units are not using that capacity, then high TNUoS charges would provide too much incentive to shut down. On the

likely to dull the incentive generation unit owners have to enter. In summary, although a locational TNUoS is likely to provide incentives for a generation unit owner to avoid entering at locations where it expects to pay high TNUoS charges, it is unclear how to set the locational TNUoS charges to provide the appropriate long-run locational pricing signals to achieve the least-cost mix wholesale electricity.

In contrast, locational pricing of energy with no locational differences in the price paid to access the transmission network does provide the appropriate signal for where new generation units should be built or how they should operate. Each hour of the year the transmission constraints that are valid for that hour are incorporated into the price paid to each generation unit and paid by each electricity consumer. This logic implies that, all else equal, profit-maximizing new entrants will build generation units at locations where they expect future energy prices to be highest and avoid locations where they expect them to be lowest. The hourly locational energy prices set will depend on actual configuration of the transmission network, the geographic location and actual levels of load, and the geographic location and actual availability of generation units during that hour. This differs from the case of the locational TNUoS charge which is fixed for entire year based on assumptions about the transmission network and the level and locations of demand and generation unit availabilities (accounting, e.g., for wind capacity factors) that are likely to be valid for few if any of the hours of the coming year.

3.2. Incentive issues in congestion management

If one assumes that more congestion is inevitable in future years and will occur in less predictable places and times,[16] then it becomes very important to develop a charging system to provide stronger incentives to market participants to manage this congestion cost effectively in the short-term energy market. This contrasts with the present "connect and manage" system, where a producer that gains access to any point on the network is effectively bestowed a right to sell its energy at an uncongested system-wide price, with the grid operator being responsible for paying generation unit owners to manage the resulting congestion.

When congestion actually occurs in a system that provides all generation units with the right to sell at a system-wide price, the inevitable result is that out-of-market side payments are necessary to pay for the re-dispatch of generation units needed to avoid overloading certain portions of transmission network. Under such as a scheme, market participants are able to benefit from being paid to remove congestion that is caused by their own actions. The experience in all US jurisdictions is that the incentives provided to market participants by such a system results in significant side payments that tend to grow unless

---

other hand, if the unit generates mainly at times when the export capacity is congested, then TNUoS charges may actually understate the opportunity cost of transmission capacity, and there could be too little incentive to retire the unit.

[16] A study of California congestion shows that peak locational prices for particular buses often did not coincide with peak average prices across large region (F.A. Wolak, "Quantifying the Benefits of Spatial vs. Temporal Granularity in Retail Electricity Pricing," Presentation, Meeting of the CAISO Market Surveillance Committee, Oct. 10, 2010, available at www.stanford.edu/~wolak). That is, important within-region congestion often occurs off-peak. With increased penetration of intermittent renewables, this phenomenon will increase.

significant transmission network expansions are undertaken.[17] In the next section, we summarize this experience, with additional detail provided in Appendix A.2.

A related issue is that the out-of-market payments are typically not coupled with a systematic overall optimization of transmission-constrained dispatch, so that the resulting dispatch is more costly than the least cost dispatch subject to the actual transmission network and other operating constraints. Specifically, the adjustments made to obtain feasibility typically attempt to either minimize the total change in dispatch compared to schedule, or the total cost of the change in dispatch to obtain a feasible solution, relative to the final schedules that emerge from the wholesale market. The market operator is typically prohibited from re-dispatching feasible generation unit schedules that only reduce the total cost of serving load. For this reason, even though the congestion management process for the existing wholesale market is used infrequently, the actual redispatch of generation units is likely to be higher than the one that would result from a wholesale market that used locational pricing.

The incentives to cause congestion and then extract payments to remove it (the so-called INC and DEC game) are inherent in any system that does not differentiate prices locationally based on congestion. Zonal systems have been designed and implemented to capture some of the locational variation as a partial solution to this issue. However, a nodal (locational marginal) pricing system has been the end-state of all such systems in the US. The fundamental reason for the evolution to nodal pricing is that using uniform prices over a region also results in opportunities for market participants to cause congestion within that region in order to be paid to remove the congestion. Inevitably, without significant transmission upgrades, the magnitude of the transfer payments associated with managing this congestion grows to the point where they are no longer deemed acceptable.

It is important to emphasize the costs of such activities. First there are sizable balancing market transfers to resources located in constrained regions, which can be made worse if those resources possess local market power. Second, generation investment can be distorted because there is no short-run incentive to site generation where it is most needed and, indeed, the prospect of receiving constrained-off payments provides an incentive to site in precisely the wrong locations. The third consequence is that by inflating the costs of managing congestion in real-time, there are incentives and signals to invest more in transmission infrastructure. These investments are not necessarily efficient—they are capital expenditures made to correct incentive problems rather than to address actual transmission capacity needs.

To summarize, we expect that congestion in GB will increase in future years due to load and renewable energy growth. Unless the energy market system represents the locational value of energy, out-of-market side payments will be necessary to avoid overloading the transmission network. This will create incentives for market participants to cause congestion in order to be paid to remove it, resulting in ever-increasing transfer payments and, likely, inefficient dispatch and, quite possibly, inefficient siting of new generation facilities.

[17] For example, in the Electric Reliability Council of Texas (ERCOT) market opened in 2001, there was no representation of congestion in market prices. Congestion re-dispatch costs reached $20 million in the first 15 days after opening of the market, necessitating a lengthy market re-design process. See R. Baldick and H. Niu, "Lessons Learned: The Texas Experience," in J. Griffin and S. Puller, Editors, *Electricity Deregulation: Where To From Here?* University of Chicago Press, 2005. This case and others are discussed at length in the Appendices.

3.3. Congestion management incentive problems under zonal pricing in US Markets

The United States has seen several power markets make the transition from single- or multiple-zonal spot pricing systems to nodal (bus) locational pricing. These transitions resulted from the recognition that disregarding transmission constraints in short-run pricing of energy results in increasing gaming opportunities, operating inefficiencies, and congestion management systems that wind up being more complex and less transparent than full locational pricing.

A useful and relevant discussion of incentive and inefficiency problems associated single-zone and zonal management systems appears in Hogan (1999), who describes the early experiences with the PJM and New England markets.[18] The PJM market originally had a commercial network model with just a single zone, and without any representation of the effect of transmission constraints into the market prices. The emphasis on bilateral contracting, similar to the GB system, resulted in a strong incentive for load serving entities to contract bilaterally with generation resources having low costs and located remotely from the demand. The resulting flows from scheduling such contracts exceeded actual transmission capacity, resulting in the necessity of administrative procedures to cope with the need for re-dispatch compared to the bilateral schedules. The significant drawbacks of such a system led to implementation of a nodal locational pricing system in PJM within approximately one year of the opening of the single zone system.

Similar difficulties plagued the initial intentions of the New England Power Pool (now ISO New England) to implement a single zone pricing system. In particular, in response to the originally announced market design, 30,000 MW of new generating plant was proposed. Recognizing that the resulting flows from this generating plant would be infeasible, the New England Power Pool then proposed rules that would require transmission studies and upgrades for new generation. These rules would have both delayed the development of new generation and also thereby protected incumbent generation from competition. The Federal Energy Regulation Commission found these conditions unacceptable, and ISO New England subsequently adopted a nodal locational pricing system.

The first dash to gas on the East coast of England can also be viewed in a similar light to the experience of New England. In particular, investors made generation siting decisions that responded to the incentives provided by a single zone system. The generators located to avoid costs of gas transportation, but this resulted in the need for transmission system reinforcements that were socialized to electricity consumers.[19] The regulator in the Netherlands has also recognized the damaging siting incentives created by "copper plate" transmission pricing. As a result, it has been considering a range of congestion management

[18] W.W. Hogan, "Restructuring the Electricity Market: Institutions for Network Systems," John F. Kennedy School of Government, Harvard University, April 1999, www.hks.harvard.edu/ fs/whogan/hjp0499.pdf.

[19] See Hogan, *ibid.*

options that might correct the lack of siting incentive,[20] all of which are either less effective or more complex, or both, compared to locational pricing.[21]

Not all US organized power markets switched quickly from zonal to nodal locational pricing. Texas (ERCOT) and California, in particular, retained their zonal systems until just recently. Their experience illustrates the following standard problems with wholesale market designs that do not account for all relevant operating constraints in the scheduling and pricing process:

- predetermined zones become increasingly poor approximations of the actual patterns of congestion,
- periodic adjustment of zones, to improve the approximation of the actual patterns of congestion, leads to significant risks for generation assets that are located near to zone borders, and
- without transmission upgrades, an increasing amount of wholesale market costs go to making the final generation unit schedules from the wholesale market physically feasible.

The experiences of the ERCOT and California zonal market designs are instructive on this point. These are summarized below, with the ERCOT situation discussed in more detail in Appendix A.2.

For these markets, the typical result is that intra-zonal congestion costs—the additional fuel costs due to operating constraints not accounted for in the wholesale market scheduling and pricing process—tend to be higher than inter-zonal congestion costs—the additional fuel costs due to operating constraints that are explicitly priced in the scheduling and pricing process.

In the ERCOT market, a zonal pricing system was used until November 2010. The annual costs for "out of market" re-dispatch actions necessary to make generation schedules within each congestion zone physically feasible (that is, the congestion costs of intra-zonal congestion management) are given below for the years immediately prior to the implementation of nodal locational pricing:

Year	**2006**	**2007**	**2008**	**2009**
Intra-zonal Congestion Costs	$190m	$169m	$191m	$179m

Estimates of inter-zonal congestion rent—the difference between the total market costs paid by electricity consumers and total revenues receives by generation unit owners—are listed below.

[20] R. Hakvoort, D. Harris, J. Meeuwsen, & S. Hesmondhalgh, "A system for congestion management in the Netherlands, Assessment of the options," Brattle Group, June 2009, www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2009/07/08/a-system-for-congestion-management-in-the-netherlands-assessment-of-the-options.html.

[21] For analyses of the congestion management proposals, see J.S. Hers, O. Ozdemir, C. Kolokathis, F.D.J. Nieuwenhout, "Net Benefits of a New Dutch Congestion Management System," ECN-E-09--075, ECN, Petten, The Netherlands, Submitted to the Netherlands Ministry of Economic Affairs, November 2009 (www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2009/11/16/ecn-rapport-net-benefits-of-a-new-dutch-congestion-management-system.html); and J. Dijk & B. Willems, "The effect of counter-trading on competition in the Dutch electricity market," *Energy Policy*, 39(3), March 2011, 1764-1773.

Year	**2006**	**2007**	**2008**	**2009**
Inter-zonal Congestion Rent	$60m	$60m	$400m	$130m

As a general observation, at any given time, the congestion rent is typically larger, and usually much larger, than the corresponding congestion cost (*i.e.*, the additional fuel cost associated with redispatching units to clear congestion). With the exception of 2008, annual intra-zonal congestion costs are greater than annual inter-zonal congestion rents. Consequently, we observe that annual intra-zonal congestion costs are greater than annual inter-zonal congestion cost (which is generally less than the rents). This high level of intra-zonal congestion costs persisted in ERCOT in spite of the fact that a number of transmission expansions came on line during this time period.

Meanwhile, the original zonal-pricing wholesale electricity market in California had a similar experience with intra-zonal versus inter-zonal congestion costs. During two of the last three calendar years that the zonal market was in operation, intra-zonal congestion costs vastly exceeded inter-zonal congestion rents, so that the annual intra-zonal congestion costs were likely to have greatly exceeded annual inter-zonal congestion costs.[22]

Year	**2006**	**2007**	**2008**
Intra-zonal Congestion Costs	$207m	$96m	$174m
Inter-zonal Congestion Rent	$56m	$85m	$176m

The GB market does not operate a zonal congestion management process so there is no analogous comparison of intra-zonal to inter-zonal congestion costs. However, the total reported costs of constraints on the system have increased from 70 million pounds in 2007/08 to 263 million pounds in 2008/09, while subsiding to a forecasted 206 million pounds in 2009/10, but increasing again to a forecasted 477 million pounds in 2010/11, according to National Grid Ofgem figures.[23] The high levels of these constraint costs are due in part to generation unit owners taking advantage of the fact that final generation schedules are typically infeasible so that a number of generation unit owners must then be paid their offer by the system operator to increase their output or pay the system operator their offer to reduce their output.[24]

---

[22] An example of the intrazonal congestion and resulting costs arising from the "dec" game was the Mexican generation situation within the southern zone of the California ISO system. Heavy congestion arose at the Miguel constraint, which was the bottleneck for three new combined cycle generation units sited just on the Mexican side of the boarder, totaling 1070 MW. This resulted in significant consumer losses due to the "dec" game. In particular, those generators would overschedule into the day-ahead southern California zonal market, and then would have to be backed down in real-time (by having those generators buy back the power at a reduced price). Units in San Diego would be constrained up to provide the energy the Mexican units were unable to inject. The Mexican generators submitted very low (very negative) bids to provide decremental energy,and in reaction, California ISO Tariff Amendment 50 was passed in March 2003, which mitigated "dec" bids by prohibiting the most extreme negative bids. Nevertheless, until the Miguel substation was upgraded in 2005, Miguel congestion management costs due to the "dec" game averaged $3-$4 million/month even with this mitigation mechanism. The value of this game was about $5/MW/hour for the Mexican generators, based on the the aggregate cost figure and the units' capacity.

[23] See entry under "Constraints" in Table on page 2 of National Grid, "Historic and Forecast Balancing Services Incentive Scheme Costs," available from: www.nationalgrid.com/NR/rdonlyres/1B6B81A0-7583-4EC0-B16D-A814E2100546/38603/ElectricitySOIncentivesHistoricForecastCosts.pdf, accessed April 22, 2011.

[24]Reportedly, the reduction in 2009/10 costs may in part be due to self-restraint by generator owners in the face of prospective license conditions concerning market abuse that were in the process of being adopted by the

The current market design provides opportunities for generation unit owners to schedule in a way that would result in congestion on transmission lines in order to then receive a lower price to reduce their output in the balancing mechanism. These actions have led the Government to propose a license condition in 2010 for generation unit owners that prohibits them from taking advantage of being behind a constraint and being the only generation unit owner able to reduce its energy schedule.[25] This license condition would also prohibit the generation unit owner from earning excess profits from extremely high offers to supply additional energy when it is the only unit able to supply the additional energy on the import side of a constraint. However, such license conditions put firms in the uncomfortable position of being asked not to take actions that are in their own self-interest.

This license condition would penalize generation unit owners that are unwilling to test the boundary of acceptable behavior. The firms that are willing to test this boundary will be rewarded by earning higher profits than the less aggressive firms. Evidence from other markets suggests that a superior market design is one that aligns the self-interest of market participants with the goals of maximizing market efficiency and system reliability.

3.4. Locational pricing

Under a locational pricing scheme, prices differ across locations in the transmission network for two reasons. The first reason is due to the existence of lines losses when energy is injected one location and withdrawn at another location in the transmission network. For example, even if there is available additional transfer capacity between two locations in the transmission network, the prices at these locations can differ because one is farther from the major load center and therefore subject to more line losses.

The second reason for spatial price differences is insufficient transfer capacity between two locations to allow all of the lower-priced generation units at one location to be used to serve demand at another location. Under these circumstances, in order to prevent the transmission line connecting these two locations from overloading, the price at the demand location must be increased to cause a local generation unit owner to supply more energy while the price at the generation-rich location must be reduced.

Because of the underlying objective for cost minimization, locational pricing markets typically compute generation schedules to minimize the as-offered cost of serving demand while honoring all relevant transmission network and other operating constraints. The locational prices are typically calculated as the change in the minimized objective function value associated with serving one more unit at each location in the transmission network. If

---

government. Part 3 of Energy Act 2010 (www.legislation.gov.uk/ukpga/ 2010/27/pdfs/ukpga_20100027_en. pdf), which came into force in April 2010, gave the government the power to insert a Market Power License Condition in the licenses of electricity generation companies to prevent companies exploiting market power that may arise as a result of constrained capacity in the electricity transmission system. These conditions are still pending parliamentary approval as for this writing. We note, however, that the "dec" game, that arises when a generator deliberately sells more than it knows the transmission system can actually deliver, and then buys back part of its obligation at a lower price in the balancing market, is a game that is, strictly speaking, not an example of market power. Rather, it is an artificial arbitrage opportunity between two markets that arises from a failure to price congestion in the forward market; even small generators who cannot affect price can also play that game.

[25] See previous footnote.

line losses are explicitly accounted for in this optimization process, then prices differ across all locations in the transmission network. If lines losses are not explicitly accounted for, then unless there is transmission congestion, the prices will be the same at all locations.

This locational pricing approach can be applied at varying degrees of spatial granularity. For example, a small number of uniform pricing zones can be created and this locational pricing algorithm can be applied to set prices for each of these zones. As noted above, this approach is effective only if the transmission network operator makes sufficient investment to ensure that all generation units within each of these pricing zones is equally effective at meeting demand. Otherwise, there will still be a need to make adjustments to ensure a feasible dispatch after the close of the wholesale market.

It is also important to note that greater spatial granularity changes the incentives for, and the proper regulatory response to, exercising local market power. Locational pricing changes the nature of local market power. Certain strategies such as the DEC game are no longer possible with locational pricing (see Appendix A.2). At the same time, under locational marginal pricing, congestion can impact the price paid to all generation unit owners in a region, rather than the price paid to the single constrained-on plant. Therefore, if the owner of the constrained-on plant owns several nearby plants, locational marginal pricing can increase the incentive that the owner has to raise the unit's offer price. Consequently, regardless of the spatial granularity in pricing allowed, all generation units should be subject to an appropriate local market power mitigation mechanism to limit the opportunities of producers to take advantage of their favorable location in the transmission network.

4. Incentives for Generation and Transmission Investment

4.1. Introduction

Having discussed the need for proper incentives for use of existing generation and transmission infrastructure, we now turn to the question of how short-term congestion management practices provide incentives for investment in both generation and transmission. We also provide a critique of the use of capacity-based fixed charges to provide locational incentives. Before discussing these issues in detail we would like to highlight a few points of emphasis.

First, there are important differences between how market prices can provide incentives for generation investment and the ability of those prices to properly incent transmission investment. In general, the latter is a much harder challenge for a market process. While we take the focus of the TransmiT process to emphasize the role that transmission pricing places on *generation* investment decisions, as we describe below it is impossible to completely separate the two issues. Therefore, we discuss the interplay between the two issues.

Second, we would like to make the potentially contentious observation that the correct amount of average congestion *should not be minimal or zero*. Transmission investments are costly in many dimensions and therefore need to be weighed against their benefits. In terms of both economics and system reliability, these benefits are reflected in the costs of managing congestion during real-time operations. If there are negligible congestion costs, then there is, in effect, no benefit to additional transmission capacity. It may be least cost to design a system with no congestion *some of the time*, or even for years at a time. For instance, scale economies in constructing transmission capacity may dictate that the efficient increment of transmission capacity will be well in excess of flows until generation investments or demand growth catch up. But the fact remains that a process that is designed to have a long-run steady state goal of *no congestion* is in fact designed to over-invest in transmission, and is therefore not matching costs to benefits.[26]

Third, the goals of designing a charging structure that is both efficient and equitable need not conflict with each other. Whatever the societal definition of "equitable" may be, it can largely be accommodated with mechanisms that still preserve the correct marginal incentives for users of the transmission system. In particular, a financial transmission rights

[26] We are *not* saying that the present policy is that the GB grid is to be expanded to achieve zero congestion; transmission reinforcements will be subjected to a benefit cost analysis under the process, and investments will not automatically be made to eliminate all congestion when it arises. However, examination of submissions in response to the TransmiT "Call for Evidence" as well as our discussions with selected stakeholders indicates that there is a widespread impression that "connect and manage" means that congestion that arises from interconnecting new generation is intended to be temporary until transmission expansion "catches up" (in the words of one group of stakeholders), bringing the congestion down to zero or very low levels. Our point is that any such policy is likely to be suboptimal.

The Province of Alberta (Canada) situation is a case in point; the Provincial Parliament's approval of several lines designed to eliminate congestion in the province has been shown to result in costs that are approximately $2 billion (Canadian) in excess of the least cost mix of generation and transmission investments together with congestion management; this works out to over $500 (Canadian) for each resident of the province (J. Church, W. Rosehart, and J. MacCormack, Transmission Policy in Alberta and Bill 50, SPP Research Paper, School of Public Policy, University of Alberta, Calgary, Alberta, Canada, 2 Nov. 2009.)

system can be used both to help market participants hedge risks and to compensate parties harmed by pricing changes. At the same time, the holding of such rights does not affect the incentive for market participants to respond appropriately on the margin to locational price signals.[27] Thus, a goal of "socializing" the costs of network investments can be largely honored while at the same time avoiding incentives that increase transmission costs.

4.2. Locational pricing and investment in generation

If one accepts the premise that a planning process should not be designed to simply eliminate (or drastically reduce) congestion, but instead to balance costs with benefits, then it naturally follows that it would be helpful to manage the remaining congestion as efficiently as possible. As we have argued above, there is overwhelming evidence around the world that a system of real-time locational pricing, such as the LMP systems used in the US, can go a long way towards achieving that operational efficiency. In addition to efficiently using existing assets, and thereby reducing congestion costs overall, the locational prices also provide important information about the value of investments in both generation and transmission.

A very useful attribute of locational pricing is that it can combine in a single price information about the value of energy to the system as a whole as well as information about the congestion costs and marginal losses imposed by withdrawing additional energy at that specific location. Thus the congestion costs and energy costs are bundled into a single locational price.[28]

For a small generation plant that lacks the ability to significantly change prices (that is, for a plant that has no ability to exercise market power), the local price provides the fundamental building block for investment decisions. Just as expectations about future energy prices in an uncongested system provide the basis for decisions about whether and what type of generation capacity to build, so do future locational prices provide a basis for decisions in a constrained network by combining information about the market value of both the capacity and location of generation. Under locational pricing, prices are higher in congested-in regions and lower in congested-out regions.

Simply put, a firm should build a plant and enter a market if its expected future net-revenue stream exceeds the investment costs, properly adjusted for risk. The locational pricing version of this heuristic is that investment should follow if the net-revenues based upon the discounted future prices at a given location exceed the costs of investing in that location. For example, consider the locational decisions for a generation technology that has relatively high capital and/or operating costs but can be built in a small footprint. Because LMPs will typically be higher on average closer to a demand center, this will tend to encourage such generators to locate closer to the demand center. In contrast, a technology

[27] A holder of a fixed MW of financial transmission rights (for instance, 100 MW of point-to-point rights that pays the price difference between two buses) cannot affect the payoff of that right unless it can exercise local market power. As a result, the revenue it receives for generating at a location is affected only by the price there, and not by whether or not it holds a FTR to or from that location; see Note 33, *infra.*, for more explanation. The effect is similar to that of any other forward contract for a fixed MW quantity. Efficient short-run operations are incented, but risks are hedged.

[28] See Appendix A.1 to this document, or F.C. Schweppe, M.C. Caramanis, R.D. Tabors, and R.E. Bohn, *Spot Pricing of Electricity*, Kluwer, Boston, 1988. A few locational pricing systems have deliberately excluded losses.

that requires a large footprint, but has low operating costs will tend to locate further from the demand center.

Currently, the GB system relies upon the principle that bilateral energy markets can provide sufficient incentives for investment in generation capacity. Of course few if any plants are built in the expectation of selling on the spot market every day. Nevertheless, the market relies upon expectations of future prices by both buyers and sellers to generate the long-term supply contracts that in turn finance the construction of new power plants.

If we could consider the future stream of locational prices at a given location as *independent* of an individual generator's investment decision, then the principle of comparing anticipated revenue streams to investment costs would apply in exactly the same fashion under a locational pricing system as it does today. The market will place a different value on energy at different locations, and the incentive to build generation at a particular location will therefore be automatically internalized in the expectation of future energy prices in that location. No further locational incentives would be necessary.[29]

Just as investors today might anticipate that the revenues from new plant in, the GB market would exceed revenues from, say, the French or Norwegian markets, investors could forecast locational prices at different possible locations to build plants and combine those with information on output distributions and costs in order to decide where and when to build and when to retire.

When the investment in generation is seen as *causing* costly investments in transmission, this can weaken the power of locational prices to provide the proper incentive to locate generation plants, or load. For instance, if locational prices were implemented under the present "connect and manage" regime, generators might anticipate that there would be a reasonable chance that transmission expansion would follow, diminishing or erasing congestion costs between the plant site and the rest of the system. We therefore discuss in Section 4.4 the role of transmission investment in setting a charging methodology.

4.3. Role of transmission rights

Financial transmission rights (FTRs) are an important feature of locational pricing-based markets for two reasons. First, they provide a hedge against uncertain differences in locational prices, by providing a payment equal to the price difference between two points (or between a point and an aggregation of points). Appropriately defined rights can thereby offset any potential short-term transmission charges relating to specific point-to-point transactions. Second, they provide a means of refunding to market participants a portion of the congestion rent earned by the system operator, defined as the difference between what load pays for energy minus what generators are paid under a locational pricing system. This might be done, for instance, on grounds of equity in order to cushion the transition to a locational pricing system or because the benefiting parties have previously helped pay for the

[29] It is possible for various imperfections in the energy market to prevent energy revenues from fully reflecting the value of generation capacity in a particular location. For this reason, some US markets (PJM, NYISO, ISO-New England, and CAISO) have established capacity markets or resource adequacy contracting requirements that include a locational component. These markets provide generators revenue in addition to what they receive from energy and ancillary services markets. A capacity mechanism has been recommended as part of the UK Government's proposed reforms (DECC, Electricity Market Reform Consultation Document, Dec. 2010), and discussions are ongoing concerning whether or not a locational component should be included.

grid. FTRs can also affect generator investment decisions, as we will describe after defining FTRs and how they work.

Due to their role in hedging locational energy prices, FTRs that are designed for a locational marginal pricing system must be *point-to-point* instruments, or bundles of those instruments. In other words, the FTR payments are based on the price difference between two points in the network.[30] This is in contrast to *path-based* rights that are defined by the cost of sending energy through a single transmission path or pipeline. Because physical electricity transactions always impact multiple parallel paths, it is much simpler from a user perspective to hedge local energy prices (with bundled transmission costs) than to acquire a large number of path-based rights. From the network-operator perspective, however, defining the number of rights than can simultaneously exist becomes more challenging than under a path-based system. Despite the complexities, much progress has been made and systems for defining and distributing point-to-point FTRs are commonplace in LMP markets.[31]

Financial transmission rights work as the locational analog to a contract for differences (CFD). The combination of a long-term FTR with a contract for difference that clear against the price at the point of withdrawal of the energy can allow for both parties to a bilateral transaction to secure a fixed price for a fixed quantity of energy at both the point of injection and point of withdrawal. Importantly, as with a CFD, the payments are linked to the contract, not to any specific physical action undertaken by the parties to a contract. Therefore, while financial instruments can allow parties to "lock-in" a price it also allows the freedom of adjusting dispatch if such an adjustment results in lower costs. In this fashion, a system of FTRs can help provide the price-certainty desired by market participants while still maintaining incentives to manage generation and transmission resources on the margin as efficiently as possible and also allowing the market operator the flexibility to seek optimal dispatch.[32,33]

---

[30] Point-to-point rights may also include transaction *hubs*, either defined as a specific single location in the network or as an aggregation of nodes in the network. Hubs can easily be accommodated in an FTR distribution process. For example a set of FTRs that link generation nodes to hubs that coincide with a complementary set of FTRs linking that same hub to consumer nodes is simply a disaggregation of point-to-point rights from the generation nodes to the consumer nodes.

[31] The various FTR allocation schemes for US organized markets ensure that the great majority of feasible FTRs (in terms of ensuring revenue adequacy, *i.e.*, FTR payments that do not exceed ISO congestion revenues) are distributed to market participants. Various mechanisms are implemented to increase liquidity; for instance, in California, the ISO automatically reallocates FTRs to retail suppliers when consumers change suppliers. The market monitoring reports for these markets indicate that large volumes of FTRs are exchanged. Improvements in market design and increased confidence of market participants in the benefits and availability of FTRs has resulted in steadily increasing volumes of trade in the MISO, PJM, and ISO-New England markets, although not in the NYISO market. Compared to 2004 volumes, volumes in the former markets in 2008 were about five times as large (DC Energy, Strategic Energy Investments, "Organized Electricity Markets Providing Benefits to Customers," Testimony before the Pennsylvania Public Utility Commission *en banc* hearings on "Current and Future Wholesale Electricity Markets", Dec. 2008.)

[32] For example, consider a plant owner at location A that wishes to contract with a consumer at B to provide 100 MW of power. If the plant simply wishes to obtain p_B, the locational price at B, rather than p_A, the locational price where it generates (which might usually be lower), then this can be accomplished with a FTR for 100 MW from A to B. Such a FTR would pay $100*(p_B - p_A)$ which, when added to the locational payment of $100*p_A$ by the system operator to the generator results in revenue of $100*p_B$. Now if furthermore the generator would like to instead earn a fixed price of (say) £50/MWh, it could in addition sign a contract for differences with the consumer that would pay $100*(50-p_B)$. Consequently, the generator would earn its desired £100*50, which is also what the consumer would pay (as its CFD payment of $100*(50-p_B)$ to the generator would be added to the

FTRs can influence investment decisions by mitigating the risk of congestion costs by allowing the generator to earn (a possibly higher, on average) price at a load center rather than its local price. FTRs can either be auctioned by the system operator (backed by congestion revenues),[34,35] or purchased from other parties. The availability of FTRs can lessen the risk associated with siting in a remote location of the network, if the generator is concerned about the predictability of its local prices.[36] Otherwise, however, the possibility of buying FTRs will not change the relative financial attractiveness of different possible sites because sellers will demand a higher price for FTRs that are sourced in locations in which lower locational prices are expected. The incentives for choosing efficient locations for new generation (that represent the best balance of generator availability with construction, fuel, and transmission costs) will remain.[37]

Because the FTRs sold by the system operator are backed by congestion revenues, FTRs provide a ready hedge for locational price risk. In contrast, there is no ready source of revenues available to the system operator to provide hedges to market participants for differences in TNUSoS charge. Moreover, because TNUoS charges are the result of a methodology that depends on the future evolution of load growth, generation entry, configuration of the transmission network and other modeling assumption, this implies that it is much less likely that independent market parties would be willing to sell contracts to hedge such risks. Although TNUoS charges obviously do not vary as frequently as

---

payment to the operator of $100*p_B$). That the incentives on the margin are unaffected by the FTR and CFD contracts can be illustrated by considering what would happen, for instance, if the generators marginal cost was £40/MWh, but the locational price p_A turned out to be £30/MWh—then the generator would simply turn off, and meet its obligation to the consumer by buying the cheaper power on the spot market rather than generating it more expensively.

[33] For a detailed tutorial on how FTRs function to hedge risk, how they affect incentives for investment, and their relationship to other types of electricity contracts, see J. Bushnell and S. Stoft, "Transmission and Generation Investment In a Competitive Electric Power Industry," POWER Working Paper PWP-030, University of California, Berkeley, Jan. 1996, www.ucei.berkeley.edu/PDF/pwp030.pdf .

[34] Under certain conditions, net FTR payouts by the system operator will not exceed its congestion revenues as long as the allocated set of FTRs pass a so-called "simultaneous feasibility" test with respect to the network (W. Hogan, "Contract networks for electric power transmission," *Journal of Regulatory Economics*, 4(3), 1992, 211-242). In US FTR systems, the system operator's allocation of FTRs is designed to satisfy that test.

[35] FTR auctions can also be used to signal where investment in transmission is desirable, if the value of the bids for FTRs over and above what would be feasible in the existing grid would exceed the cost of expanding the grid to provide those rights (J. Bushnell and S. Stoft, "Improving Private Incentives for Electric Grid Investment," *Resource and Energy Economics*, 19, 1997, 85–108).

[36] A challenge for intermittent generation sources is that point-to-point FTRs are usually, for simplicity, specified as constant over time or sold in simple blocks (*e.g.*, 100 MW for 9 a.m. – 5 p.m.). This is unlikely to match the pattern of output for intermittent generators, overhedging the congestion risk when output is low and underhedging when it is high. However, when both renewable and thermal generation are in a "gen pocket" (from which power is generally exported), then there is a complementarity of interests that can allow a FTR sharing agreement that would better hedge both of their risks. For instance, an agreement could be made that the FTR payout is given to the intermittent source when it generates. But when the intermittent source is unavailable, the payment could instead be made to the thermal unit; these are likely to be the times when the thermal unit's generation is greater because export capacity is freed up. Furthermore, hubs can help to match needs for FTRs with underlying physical constraints.

[37] However, if FTRs are freely given by the system operator to new generators, then the locational incentive is diminished or eliminated. Therefore, any free distribution of FTRs should not be contingent on decisions to build (or, for that matter, to not retire) generators.

locational spot prices, although there is year-to-year variation that can be difficult to predict, as several market parties noted in their filings in response to the TransmiT call for evidence. In addition, as we noted earlier, the use of a locational price signal to recover a a fixed cost is likely to induce significant distortions from least cost behavior in the short-term market. The year-to-year changes in the locational TNUoS prices are likely to increase the incentives for distortions from least cost behavior.

4.4. Transmission investment and its impact on locational prices

Because a potentially important aspect of the decision to invest in generation is the revenue from a given location, a key consideration is whether that investment will cause a change in the expectations of local prices and revenues. With regards to the locational incentives, this concept translates to the question of whether transmission investment is required to follow the generation or whether the planning process plays in out the opposite order.

There are many complications to transmission investment that make it difficult to leave it completely up to a market process or to rely solely on market prices for decision-making. These factors include *significant economies of scale.*[38] This means that it often makes sense to build out infrastructure in *lumpy* increments, expanding beyond immediate needs in anticipation of future growth in demand. One consequence of this is that new infrastructure can be highly disruptive to locational pricing patterns. Paths that were previously frequently congested can transition to no congestion at all, for a while, once an expansion is complete. Note again the contrast to generation. The entry of a single generation unit, outside of small local markets, is not nearly as disruptive to the time pattern of energy prices as is the construction of transmission projects.

While these attributes greatly complicate the process of transmission investment, they do not on their own imply that locational prices would provide insufficient locational incentives for generation. Even if a lumpy transmission investment causes congestion costs to plummet at least temporarily, it is appropriate for generation unit owners to internalize the likelihood of this new circumstance in making their investment decisions. Once the transmission investment is made, it is efficient for generators to not include such sunk costs in their location decisions. As noted above, the difficulties instead stem from the perception, correct or not, that siting new generation in a location can force transmission investment into that region.

If there is a policy (such as "connect and manage") that generation is to in fact lead transmission, it is easy to demonstrate how economies of scale in the construction of transmission can create problems. Construction of generation in Scotland creates congestion that in turn leads to large-scale investments to relieve that congestion. Due to economies of scale, those transmission investments can leave minimal congestion in their wake. Under this example, generators pay little locational penalty in the balancing market for locating in the north, despite the fact that their decisions forced considerable capital investment that is paid by all market participants under the current TNUoS charging mechanism.

[38] In most contexts, transmission network congestion is also accompanied by large degrees of uncertainty and generator local market power, further complicating the role of market prices in dictating investment decisions. See P. Joskow and J. Tirole, "Merchant Transmission Investment," *Journal of Industrial Economics*, LIII(2), 2005, and Bushnell and Stoft, 1997, *op. cit.*.

This cautionary tale helps to illustrate the role that the transmission planning process can play in determining how transmission charging drives incentives. Two planning approaches representing opposite extremes are:

1. ***Transmission follows generation.*** This might be called "*Connect and Invest*." If transmission follows generation, the transmission planning process has to make whatever investments necessary to accommodate the full output of any generator locating in a given location. Then locational prices in the spot market will provide inaccurate incentives for locating generation in the most efficient places, as the planning regime will guarantee those congestion costs will be minimal. Locational incentives under this system might more be provided more effectively if generators reveal their willingness to pay for transmission services either (a) by paying a preset fixed capacity-based fee that is an approximation of incremental transmission expansion costs,[39] or (b) by allowing generators or other market parties to sponsor transmission investments themselves, receiving FTRs in return.

2. ***Generation follows transmission:*** This might be called "Coordinated Infrastructure Master Plan." If the transmission planning process instead operates under a regional process focused on a set of regional policy objectives, then it is less likely that a single generation plant location can trigger substantial network investments. Combinations of such planning regimes with locational pricing can be found in the United States. Locational prices then reflect the relative value of power at different locations, in terms of increases or decreases in redispatch costs necessary to manage congestion; if a generator chooses to site in a location in which congestion would be significant, then it would receive a lower stream of revenue and, all else being equal, construction there would be less profitable.

A question for the TransmiT process is where a philosophy of "connect and manage" fits along the spectrum delineated by these two extremes. We believe it resembles but is not identical to the above "transmission follows generation" caricature. It differs because transmission investment is not automatically triggered by generator construction; instead, such investment is subject to Ofgem regulatory approval, in which economic justification (based on avoided congestion costs) plays an important role. However, the general philosophy of "connect and manage" is closer to one of generation leading transmission than vice versa; and there is a clear expectation articulated to us directly by stakeholders and through their TransmiT filings that congestion that arises from generation siting will only be temporary until transmission investment "catches up." If this is generally true, then if locational pricing signals are provided through the spot market, they would likely affect siting decisions less than in a "generation follows transmission" scheme.

Timing is also important here. The time horizon for planning and implementing transmission investments can be far longer than the time it takes to site and build new generation plants. Therefore, transmission planning is often leading generation in timing, even if transmission planning is undertaken with an eye toward future needs – as it usually is. If generators are coming second in the process, then transmission costs are sunk from their perspective, and it could be inefficient to charge them disproportionately for those costs. Consider the prospect of a large capital investment with considerable capacity, due to

[39] As TNUoS fees attempt to do, but see Section 4.5 below for our critique of those fees.

economies of scale. Large fixed TNUoS costs could discourage new generators from locating in a region where there is now *excess* transmission capacity.

4.5. The alternative of capacity-based annual locational price signals

Energy prices vary considerably over the year, and adding a congestion component to create locational prices increases that variability. The increase is not large, but is enough to influence siting decisions. Variations from hour to hour (which is already reflected in GB prices) tend to be much greater than variation from place to place. In PJM, for instance, interzonal variations in prices are approximately an order of magnitude smaller than intertemporal variations,[40] and the story has been similar in California.[41] However, there is a distinct spatial pattern to average prices that indicates, for example, that capacity is more valuable in, for instance, the San Diego region and Humboldt County than elsewhere in California due to well understood transmission constraints, and that urban load pockets experience higher prices as well. These price patterns signal a relative need for new generation capacity in those areas.

In this section we consider how well TNUoS-type charges can reflect actual transmission costs. In theory, there can exist simple circumstances in which such charges can provide a reasonable approximation to actual costs. But as the sources of power in a system diversify, and as transmission additions become larger in size, the relationship between a TNUoS-type charge and actual costs that a generator imposes on a transmission system becomes difficult or impossible to establish.

In the following circumstances, we believe that a locationally differentiated annual per-MW charge applied to generation capacity is not an unreasonable representation of transmission costs:

- Generation leads transmission, so that transmission is expanded to accommodate new power plant capacity.
- Transmission is sized to accommodate maximum flows on the system, which occur at times of peak demand when all generator outputs are at or close to their full capacity.
- The magnitude and direction of flows and any congestion are reasonably predictable.
- Increments of transmission capacity are not large compared to the amounts of generation capacity added.

---

[40] Data from the old PJM market (prior to its expansion to the Midwest) for April 1, 2002 to December 31, 2003 was analyzed in J. Popova, Spatial Patterns in Modeling Electricity Prices: Evidence from the PJM Market, Department of Economics, West Virginia University, *Proc. of 24th USAEE and IAEE North American Conference*, Washington, DC, 2004, www.iaee.org/en/students/best_papers/popova_washington_dc_2004.pdf. She showed that the 24 hourly average prices for that period vary from 16.4 to 48.6 $/MWh, and the standard deviation of those 24 average values was 10.4 $/MWh. In contrast, average prices for the 12 eastern PJM zones varied from 32.7 to 36.8 $/MWh; the standard deviation of those 12 averages was 1.2 $/MWh, an order of magnitude smaller than the standard deviation over the hourly averages.

[41] Data from the new California locational pricing market was analyzed by Wolak (2010), *op. cit.*. Prices for 3000 electrical buses was analyzed, and it was found that 90% of the buses had average day-ahead prices between 33.25 and 36.87 $/MWh, while the same interval for real-time prices was 39.7-42.9 $/MWh (more extreme buses tended to be locations with very little generation or load). Meanwhile, the median standard deviation of locational prices across those buses was approximately 13.3 $/MWh for day-ahead prices and 70 $/MWh for real-time prices.

- There are not large changes in generation mixes that result in permanent shifts in flow patterns that leave formerly intensely used segments of the network uncongested.

As a result, as new generation capacity is added in remote areas to serve loads elsewhere, transmission capacity can added in a relatively smooth manner to accommodate it, all transmission facilities are used at close to their capacity, and all generation capacity contributes to the need for that transmission. TNUoS-type charges, based on a MW-km calculation, can then be an acceptable approximation of the marginal investment cost needed to accommodate more generation.

Unfortunately, these assumptions are decreasingly applicable to the GB situation.

- Transmission will not always be built to erase congestion, as any transmission additions will be subject to Ofgem regulatory oversight and economic analyses. Thus, significant congestion can persist, in part because of the lag between when generation investment takes place and transmission follows, and in part because the regulatory process will find some congestion uneconomic to relieve. As a result, a significant proportion of the cost that generators impose on the transmission system will be in the form of re-dispatch costs to manage congestion, rather than capital expenditures.
- As more generation is obtained from intermittent and remote sources as well as from storage, congestion will become increasingly unpredictable in location and timing, and is less likely to coincide with peak loads.[42] Security constraints in the main interconnected system will become more complex, and system conditions during peak demand may become less important as drivers of transmission investment. Some adjustments (such as for intermittent sources or storage) could be made to TNUoS-type charges to account for the coincidence or lack of coincidence of generation from particular facilities with peak system flows. However, stakeholder comments indicate that this will be a contentious process (with claims and counter-claims about biases towards particular plant types), and shifting patterns of congestion over space and time will make estimates of the relevant scaling factors increasingly difficult. Locational energy prices would be no easier to predict, but have the distinct advantage of representing the actual marginal as-offered cost of power at a location, rather than a (potentially highly inaccurate) approximation to this magnitude. Furthermore, in an FTR auction process, market participants can hedge their exposures to uncertainty based on their own assessment of their risks, contrasting significantly with a contentious administrative adjustment of TNUoS-type charges.
- As outmoded thermal facilities are retired (*e.g.*, under the EU Large Combustion Plant Directive), some transmission facilities will likely become permanently underused. At the same time, other facilities such as those associated with North to South flows will need to be expanded due to new generation investment, but may not be fully utilized at peak demand conditions.
- Transmission investments can be large compared to individual generation facilities. As a result, new facilities may be oversized for many years until usage increases to catch up. Rather than increasing capacity in a fairly regular manner over the entire

[42] Even in a thermal-only system, peak flows may not coincide with peak loads. For instance, during times of peak demands, it can be the situation that more of the local cheap generation in exporting regions is needed to meet local demands and the most expensive generators in importing regions are turned on. As a result, the flow from the exporting to importing region decreases relative to shoulder or offpeak periods when more of the exporting region's capacity can be devoted to exports.

grid as usage grows, particular "pinch points" will arise. The costs associated with relieving those pinch points and their relationship to generator siting decisions are decreasingly likely to be related to the MW-km metrics of grid use that underlie TNUoS-type charges.

We have just described some general concerns with the use of fixed charge to incent location decisions. There are also practical implementation difficulties. In particular, the MW-km methodology that is an input to the setting of TNUoS charges is subject to a number of limitations that give us more reason to question whether these charges are cost-reflective. As a general observation, MW-km approaches involve a number of very arbitrary choices, most particularly about the flows due to incremental generation and therefore about the need for incremental transmission investment. Consequently, results are difficult to relate to actual transmission expenditures necessitated by incremental generation and are therefore extremely unlikely to provide efficient incentives.[43] Despite the specific intention and claims of "cost reflection," MW-km approaches as used around the world, including in the UK, are simply complex mechanisms for cost recovery.

The particular rationalizations for what are essentially arbitrary decisions and parameter values in "The Statement of the Use of System Charging Methodology"[44] are typical of this approach, which has been tried and discarded in various other jurisdictions, including the Electric Reliability Council of Texas (ERCOT). The many choices and parameters in the model determine the sharing of the charges among market participants. Negotiation about the choice of values of parameters allows plenty of scope for sophisticated market participants to shift costs to other market participants. Examples of such choices in the MW-km calculations underlying TNUoS include (with page references to "The Statement of the Use of System Charging Methodology"):

- Scaling of generation, as described in Section 2.12, page 14. Generation capacity is apparently scaled down so that total generation equals total demand, with demand specified as peak winter conditions. However, a baseload generator is likely to be running at full output when demand is at peak, whereas peaking generators are likely to be below their maximum generation, because of the need for reserves. Moreover,

[43] Arguably, the MW-km approach *might* give a reasonable approximation to the needed transmission for incremental thermal generation necessary to meet peak demand under particular idealized and unrealistic assumptions such as the existing transmission system being "fully" utilized. However, transmission requirements for remote renewables are unlikely to be well-represented by a scaled fraction of the capacity necessary to allow flow from the wind farm to the demand center. The reason is due to the inherent sharing of some, but not all, of the transmission capacity amongst renewables (and other technologies) over time. For example, for a wind farm in Scotland, the transmission capacity needed from the wind farm to a Scottish "hub" in the main existing interconnected system in Scotland is likely to be equal to the capacity of the wind farm. That is, this transmission capacity cannot be easily shared with other wind farms. However, the incremental capacity from Scotland to England to support increased renewables is likely to be on the order of the total capacity of the incremental renewables multiplied by somewhat more than their capacity factor (perhaps around 50%), allowing for considerable sharing. Moreover, if the wind production is not coincident with the system peak, then essentially no additional capacity in England into demand centers would be required to support incremental renewable production. Since the transmission capacity into the demand centers is likely to be the most expensive on a MW-km basis, and since the Scottish to English capacity is likely to be the longest amount of capacity (and the most expensive on a MW-km basis, as well, if it is undersea cable), then failing to account for this sharing will not result in a correct assessment of the needed transmission for the wind farm and *a fortiori* will not result in efficient prices for incremental transmission.

[44] www.nationalgrid.com/uk/Electricity/Charges/chargingstatementsapproval/

renewables may be operating below capacity at the peak of demand and may operate at their capacity during off-peak demand conditions. Scaling of capacities results in flows on the network that would not match actual flows. Furthermore, flows based on scaled capacity do not represent the way in which resources can and cannot share transmission capacity over time because they fail to represent the temporal distribution of generation. That is, scaling does not represent the implications for the needs for incremental transmission capacity. Allowing for different scaling factors for different technologies without reference to actual dispatch over time simply inserts another potentially arbitrary parameter value choice into the methodology without making a direct connection to the required incremental transmission to support the incremental generation.

- Split of revenue between generation and demand as mentioned in Section 2.63, page 23. The arbitrariness of this is recognized in the document, as evidenced by the word "correct" appearing in quotes. This has less of an effect on spatial differentials in TNUoS charges than on their overall level, but is another potentially arbitrary parameter value choice in the methodology.
- The MW-km method as used to determine the TNUoS does not check for security,[45] nor does it check that pre-contingency flows are within ratings. It does not distinguish flows on the existing network from the need to construct additional capacity. So, the MW-km method does not in fact reveal at all the incremental investment needed to support additional flows. As a number of filings submitted in response to the TransmiT call for evidence indicate, there are concerns about particular assumptions made about treatment of onshore substation facilities, HVDC converter stations, various voltages, and the distinction between local connection costs and wider connection costs. These assumptions have important implications for the charges but are a challenge to justify in terms of cost-reflectivity. Particular choices will have significant effects on market participants, providing yet another set of arbitrarily chosen set of parameters that will be a cause for contention.

In summary, as is quite common in the justification of MW-km approaches around the globe, what is actually just a cost recovery mechanism is erroneously being promoted as having relevance to evaluating incremental transmission investment costs. A TNUoS-type policy of MW-km charges related to some approximate calculation of average costs of transmission is unlikely to lead to the correct incentives for siting baseload, intermittent, storage, and peaking generation, as their use of the transmission system is very different over time. For various reasons, the MW-km methodology and subsequent adjustments used to obtain the TNUoS charges are unlikely to bear more than the roughest relationship to incremental transmission and congestion costs resulting from a siting decision. The various parameters and modeling assumptions affect the outcomes but are only indirectly connected to transmission planning, Because these parameters do not arise from a fundamental design process, but are determined in an administrative manner, they primarily have the role of adjusting the payoffs to determine winners and losers.

[45] A fixed "security factor" is applied to all lines to account for security. The security factor is at best an average for the whole network, underestimating the case for radial connections and likely overestimating it for the main interconnected system.

5. Proposed Modifications to the Charging Methodology

This section outlines a charging methodology that we believe would reduce short-term market operating costs and provide more efficient locational price signals to new generation investment decisions. Briefly, we believe that the GB system would benefit from modifying the current planning and charging system in three ways. First, we believe that the emphasis on locationally varying network charges should shift from long-term transmission network charges to short-term energy prices. This would require integration of congestion management with short-term system operation either through a zonal pricing or full nodal-pricing and eliminating locational differences in TNUoS charges. Second, we recommend that a system of financial transportation or transmission rights be introduced to allow generation unit owners and load-serving entities to hedge locational price risk.

Third, we believe that the process for network planning should shift its emphasis entirely to a philosophy that transmission network investments anticipate least cost generation unit location decisions, as opposed to the present mix of "connect and manage" (resulting in at least some transmission following generation) and anticipatory planning (as embodied in the Transmission Investment Incentive). The transmission planner knows the location of the major renewable energy sources, points of access to natural gas pipelines and other sources of input fossil fuels, the locations of load growth, and, most important, the characteristics of the existing transmission network. For this reason, the transmission planner is in the best position to determine the network upgrades that will facilitate the entry of generation units that results in the least-cost supply of electricity to final consumers. A forward-looking transmission expansion policy that anticipates the expected profit-maximizing generation unit owner entry decisions will achieve a lower cost of supply of wholesale electricity to load throughout GB, relative to a policy that expands the transmission network in response to generation location decisions. Although it is unlikely that a generation unit owner would find it profitable to construct new units at locations not anticipated by a forward looking transmission planner, if generation unit owners would like to enter at locations that do not meet the planners' estimation of the least cost configuration for the transmission network, they can construct the necessary facilities and pay for them and in exchange receive financial transmission rights for the additional transmission capacity their upgrade creates.

5.1. Introduce locational pricing of energy

We believe that the introduction of locational pricing of energy will prove the most critical to improving the performance of both short- and long-term wholesale market. There are locational aspects to the current balancing market in the sense the producers are paid as-bid for incremental and decremental energy relative to their final energy schedules. However, as discussed in Section 3, this approach to ensuring feasible final schedules can create opportunities for producers to raise their bids if they expect to be asked the system operator to provide additional energy, or lower them if they expect to be asked to provide less energy. Both these actions increase the costs to load of achieving feasible generation schedules and increase the costs of balancing the system in real time. As also noted in Section 3, similar methods to those used in GB balancing mechanism, such as the intra-zonal congestion management process in US zonal pricing markets, have proven unable to manage efficiently more than minimal levels of transmission congestion, and are vulnerable as well to inflated congestion costs due to the DEC game. We anticipate that growth in demand and renewable energy will increase the frequency and magnitude of congestion in the GB system. Without

the introduction of locational pricing, the cost to consumers of constrained on- and off-payments is very likely to grow as the amount of renewable generation capacity increases. Because NGET cannot simply pass on these costs to final consumers under Ofgem rules, as those payments become more volatile, they will pose increasing financial risks to NGET.

Under locational pricing of energy, producers should no longer have a right to inject pre-specified quantities into the network regardless of underlying system conditions. Instead, producers should have to compete with each other to inject energy based on their willingness to supply energy at their location in the network. Locational prices and the amount of energy that individual producers inject will rise and fall over time, depending on the characteristics of the network, the spatial distribution and level of demand, and the availability of other sources of generation. These characteristics of the transmission network continually change. It is impossible to guarantee any specific generation unit owner the right to inject a pre-specified quantity of energy at a zero or fixed variable price, regardless of system conditions. Consequently, a more realistic and lower cost solution is an energy market that sets spatial prices that reflect the current as-bid cost of energy withdrawn or injected at each location in the network.

As we pointed out in Section 3.3, although a number of US markets started as zonal-pricing markets, all of them have found dealing with intra-zonal congestion to be increasingly costly without substantial transmission upgrades within and across these congestion zones. International experience in Australia and the Nordic countries does demonstrate it is possible to operate a zonal-pricing market, if there is adequate transmission capacity within each zone to make all generation units equally effective at meeting load at all locations in that zone,[46] However, the cost of the transmission expansions necessary to limit the incidence and magnitude of intra-zonal congestion is likely to become increasingly high as the penetration of intermittent renewable generation increases. For these reasons, it is very likely to make more sense for GB to directly implement a nodal pricing short-term market, rather than first implement a zonal-pricing market design as was the case in all US wholesale markets.

A locational marginal pricing real-time market could be implemented in GB with no changes to the existing scheduling and real-time dispatch process. Generating companies would continue to submit balanced schedules to the system operator. The system operator would continue use the same mechanism to achieve feasible real-time operating points for all generation units. However, the system operator would then use the willingness-to-adjust curves submitted by all generation units to compute *ex post* locational marginal prices that support the actual real-time operating points of all generation units by minimizing the as-bid cost of meeting actual demand at all locations in the transmission network for the configuration of the transmission network for that half-hour. These locational marginal prices would then be charged to all load consumed in that delivery period and paid to all

[46]Yet the Nordic experience also shows that zonal systems can be very inflexible, in that it can be extremely difficult from an institutional point of view to modify zones to match physical reality when there is a large amount intrazonal congestion. Only last year did Sweden finally split into more than one zone, despite the extensive and persistent congestion between its hydropower-rich north and southern population centers. California experienced similar problems; the Mexican generation pocket within the southern California zone resulted in several million dollars per month of ratepayer costs due to the "dec" game, but it proved to be impractical to carve out a separate zone that would have eliminated it. The short-term solution was a transmission reinforcement, and the long run solution was implementation of full nodal locational pricing in 2009.

energy generated during that period. This approach would have the advantage of involving no changes in scheduling process and real-time system operation. Only the prices and quantities at which generation units and loads settle would change. Generation units at low-priced areas would receive the LMP at their location only for the amount of output they actually produced, not for amount they scheduled. Generation units at high-priced areas would be paid this price for their actual output, not for what was scheduled. Similar logic would apply to all loads. They would pay the real-time price for their actual consumption. This approach would create strong financial incentives for all market participants to submit physically feasible generation schedules.

This single settlement LMP market based on the energy offer curves of generation unit owners is very similar to the current New Zealand nodal market design. Under the New Zealand market design, none of the generation unit and load schedules submitted to the system operator before real-time are financially binding. Generators are only paid for their actual real-time output and loads pay for their actual consumption at the real-time price.

The disadvantage of this single settlement approach is that it may be difficult to schedule a number of long-start generation units in a least-cost manner. Given the small number of long-start units in the hydroelectric-dominated New Zealand electricity supply industry, their cost of a single settlement market is unlikely to be very large. However, the GB market has significantly more long-start units, so substantial operating cost savings should result from transitioning to a two-settlement LMP market with a day-ahead scheduling process that sets locational marginal prices and financially binding day-ahead schedules. This would then be followed by a real-time imbalance market that sets LMPs by minimizing the as-bid cost of meeting the deviations from day-ahead demand and generation schedules. This market design would be similar to the current multi-settlement US market design, except for the fact that only energy offer curves are used to dispatch units in the day-ahead and real-time markets. The US markets allow generators to submit start-up cost offers and energy offer curves to both the day-ahead and real-time markets and guarantees start-up cost offer recovery for all units dispatched by the system operator. This approach has the advantage of allowing the system operator to consider both costs in determining whether to start-up a generation unit. If the GB system decides to transition to a two-settlement LMP market, then it should also consider allowing generators to submit start-up cost offers that are used in the day-ahead and real-time markets.

5.2. Develop a system of financial transmission rights

A second change we recommend is a system of financial transportation and transmission rights. These rights would allow generators and customers to reduce spatial price risks for their injections into and withdrawals from the network, while also providing all participants with proper incentives to address congestion in a least cost manner. The ability to introduce financial transmission right (FTRs) to hedge spatial energy price risk is another advantage of introducing locational pricing of energy relative to spatial pricing via TNUoS charges. As noted earlier, because there is no revenue stream accruing to the system operator (similar to the congestion rents in an LMP market) that would allow it to hedge changes in the TNUoS, entities interested in hedging changes in the TNUoS charges would have to rely on private parties to provide these hedging instruments. Because the primary source of risk in TNUoS charges is regulatory risk, it is unlikely the private parties will provide a hedge against this risk at a price that a market participant would be willing to pay.

As mentioned in Section 4.3, FTRs are readily available to hedge locational price risk in US markets. This is because locational energy price risk arises naturally from economic fundamentals of dispatch and FTRs are readily backed by the system operator's congestion rent. In the absence of loss charges, this rent equals the difference between the total amount paid by loads and the total amount paid to generation unit owners.[47] If FTRs are allocated to satisfy a simultaneous feasibility test, then the congestion rents are guaranteed to be adequate to cover the FTR obligations, so that the system operator can provide FTRs without taking on significant financial risks.[48]

In contrast, the TNUoS charges are the result of a regulatory process whose future evolution is uncertain, and it is much less likely that independent market participants would be willing to sell contracts to hedge such regulatory risks. Moreover, there is no obvious counterparty able to sell these products and cover the payments and revenues implied by a product that hedges locational transmission charge risk. Although TNUoS charges obviously do not vary as frequently as locational spot prices, there can be year-to-year variation in their magnitude. To summarize, there are no financial instruments to hedge TNUoS charges and such instruments are unlikely to emerge in GB. This is because the risks relating to variation of these charges would have to be taken on by the issuer of such instruments. This contrasts with the situation for FTRs, where the system operator has a natural hedge in the form of its congestion rents. Further, financial institutions who might issue such instruments are unlikely to be eager to hedge regulatory uncertainties, as opposed to market risks that arise from changes in market fundamentals. Finally, as noted earlier, because the TNUoS charges are primarily used to recover sunk costs and are unlikely to provide an efficient signal for new generation unit location decisions, changing them on a year-to-year basis is likely to introduce additional incentives for deviations from the least-cost supply of wholesale electricity and transmission services to final electricity consumers.

We emphasize that, particularly for electricity, any transmission rights issued should be purely financial in nature. FTRs, by providing a payment equal to the locational price difference between two points, provide a hedge against any potential short-term price differences relating to specific point-to-point energy transactions. Transmission rights should only pay and/or obligate market participants to locational energy prices differences for a fixed quantity of energy. Allowing participants to hold only financial transmission rights rather than physical rights ensures that transmission capacity cannot be withheld to benefit the holder of this right at the expense of overall market efficiency.

Financial transmission rights work as the locational analog to a contract for differences (CFD). The combination of a long-term FTR with a contract for difference can allow for both parties to a bilateral transaction to secure a fixed price for a fixed quantity of energy at their respective locations. As with a CFD, the payments are linked to the terms of the financial contract and not to any specific action undertaken by the parties to the contract. Therefore, while financial instruments can allow parties to "lock-in" a price it also allows the generation unit owner the freedom to adjust it output level if such an adjustment results in

[47] Congestion rent is also referred to as "merchandising surplus."

[48] This is true as long as the transmission network can accommodate the flows implied by the allocated transmission rights. To be conservative, some systems allocate slightly fewer rights than simultaneous feasibility would permit. Despite this, due to transmission outages, congestion rents are sometimes inadequate to cover FTR payments in US systems; the common remedy in that situation is to discount payments so that the system operator is not in deficit.

lower costs. In this fashion, a system of FTRs can help provide the price certainty desired while still maintaining incentives to manage generation and transmission resources as efficiently as possible.

We acknowledge that the hedging capability that can be provided by FTRs is of a different nature and in some ways more limited than under the current system. The amount of FTRs that can be issued by a grid operator is limited by the physical capabilities of the network if the operator is to maintain a financially balanced position, known as "revenue adequacy." This means that the transmission network operator recovers sufficient revenues from all market participants to pay all of its obligations to FTR holders. Transmission rights are also issued for a fixed capacity, and because the revenue streams from FTRs are independent of the generation unit's output, they do not always provide a "perfect hedge" for locational price differences for the generation unit's actual output level.

By contrast the current arrangements could be seen as a form of "full requirements contract," one where the purchaser (of transmission access) is in theory entitled to consume as much as they require at a fixed price whenever they wish to consume it. Historically, most retail energy arrangements in the United States also operated in this way.[49] However, it is critical to remember that this "guarantee" of transmission access is itself a financial fiction. Rather than bestowing true unlimited access to the network, the current arrangements really create an obligation to compensate generators when the inevitable congestion does occur (with limited exceptions when congestion is due to line outages).

Because of their role in hedging locational energy prices, FTRs that are designed for a LMP system must be *point-to-point* instruments. In other words, the FTR payments are based upon the locational price difference between two points in the network. This is in contrast to *path-based* rights that are defined by the cost of sending energy through a single transmission path or pipeline. Because electricity transactions frequently impact multiple paths, it is much simpler from a user perspective to hedge local energy prices (with bundled transmission costs) than to acquire a large number of path-based rights. From the network-operator perspective, however, defining the number of rights than can simultaneously exist becomes more challenging than under a path-based system. Despite the complexities, much progress has been made and systems for defining and distributing point-to-point FTRs are commonplace in LMP markets.

Point-to-point rights may also include transaction *hubs*, either defined as a specific single location in the network or as an aggregation of nodes in the network. Hubs may easily be accommodated in an FTR distribution process. For example a set of FTRs that link generation nodes to hubs that coincide with a complementary set of FTRs linking that same hub to consumer nodes is simply a disaggregation of point-to-point rights from the generation nodes to the consumer nodes.

5.3. De-emphasize locational charges for fixed and capital costs of the network

The third change is to limit the locational aspects of annual transmission charges. This means that the TNUoS would have much less variation across locations in the GB system. This should be accompanied by a change in philosophy for building out the

[49] These traditional retail energy arrangements also create analogous incentive problems in that users have no incentive to reduce consumption when system conditions are tight.

transmission network in GB. Rather than the current "connect and manage" approach, the new paradigm would be one where transmission planning leads new generation investment. As we discuss below, a transmission planning and expansion policy that anticipates the profit-maximizing entry decisions of generation unit owners is likely to yield a lower average delivered price of electricity (including transmission and distribution costs) than a policy that where generation investment lead transmission expansions.

Transmission planning has become more complex in restructured markets, in part, because of the separation of generation from transmission planning. "Connect and manage" is one approach to this separation, where the advent of new generation would then prompt new transmission development in response. However, with significant new potential renewable generation sites in the GB, it is likely that the total amount of possible renewable resources greatly exceeds the immediate or even medium-term levels of renewable development. In this context, connect and manage provides no reasonable expectation that the most economical overall resources are developed, by which we mean the expansion of renewable generation and the associated transmission that minimizes the overall cost of both generation and transmission.

The transmission network in the wholesale market regime serves another role in addition to the ones it served in the vertically-integrated monopoly regime. It can increase the number of distant competitors that local generators face in selling electricity to final electricity consumers. If transmission capacity into a load center is limited, this can bestow market power on generators located in this load center. Consequently, if generation unit investments lead transmission investments, the transmission planner is likely to have to focus on investing in upgrades to limit the ability of new generation entrants to exercise local market power. This will lead to both higher average wholesale energy costs (due to the exercise of local market power) and higher transmission costs (because of the transmission upgrades made in response to this market power), relative to a regime where transmission investments lead generation unit investments and generators are paid locational marginal prices. By committing only to upgrading the transmission network at locations that are likely to yield the lowest wholesale energy costs consistent with the long-term financial viability of the industry, the transmission expansion process can avoid having to engage in costly transmission expansions in response to the exercise of local market power by new generation unit entrants. Moreover, new generation units at locations anticipated by the forward-looking transmission planning process will have less ability to exercise unilateral market power, which will lead to lower average wholesale electricity prices.

The forward-looking transmission expansion plan in a "transmission lead generation" regime would need to consider multiple scenarios of future policy, technology, and economic developments. Benefit-cost analysis of potential transmission additions would account for how generation siting and operating decisions would likely be affected by these developments as well as the incentives provided by locational pricing.[50] Tallied benefits could include, for example, reduction in generation costs and prices to consumers, amelioration of market power in isolated markets, environmental improvement (as measured

[50] A prototype of this type of analysis was demonstrated for GB in in A.H, van der Weijde and B,F. Hobbs, "Planning electricity transmission to accommodate renewables: Using two-stage programming to evaluate flexibility and the cost of disregarding uncertainty," Working Paper EPRG1102, Electricity Policy Research Group, University of Cambridge, Jan. 2011, www.eprg.group.cam.ac.uk/category/publications/working-paper-series/.

by the market value of renewable energy and carbon reductions that the investment makes possible), enhancements in reliability, and hedging against uncertainties concerning future technological, economic, and policy developments.[51]

An example of this type of approach was carried out in the ERCOT power region, where wind regions were identified and potential resources in each region grouped and prioritized for transmission planning from the perspective of wind potential.[52] Any such process is likely to be imperfect. For example, the ERCOT process did not fully consider off-shore wind and mostly focused on on-shore wind and there were limitations in the optimization of the transmission planning activity itself and its coordination with the wind expansion. However, given economies of scale of transmission expansion, given the various uncertainties with prospective renewable resources, we argue that a principled effort to plan access to the specific regions on the basis of estimated costs and benefits will likely result in a superior overall plan compared to *ad hoc* planning such as is inherent in a connect and manage regime. Although it is impossible for the transmission planner to know precisely the least-cost mix of future generation capacity, an imperfect forward-looking planning process should lead to lower delivered energy costs relative to a regime where transmission investments occur primarily in response to the profit-maximizing entry decisions of generation unit owners. Under a planning approach where transmission expansions are not made in response to past generation unit investment decisions, there is little danger that socializing the fixed costs of the transmission network through a TNUoS with little spatial variation will cause generation unit owners to site where their presence reduces overall wholesale market efficiency. All capital costs for existing infrastructure (*e.g.*, sunk costs), ongoing fixed costs of network operation, and future investments that are not forced by a single or small group of generation sources should be allocated to load.

Under this scheme, shallow connection costs should still be borne by the owner of the facility making the connection, and this will preserve meaningful spatial differences in total transport charges. There is ambiguity for where a shallow connection ends and a transmission

[51] California quantifies most of these benefits in its Transmission Economic Assessment Methodology while anticipating the siting, operational, and bidding responses of generation (M. Awad, K.E. Casey, A.S. Geevarghese, J.C. Miller, A.F. Rahimi, A.Y. Sheffrin, M. Zhang, E. Toolson, G. Drayton, B.F. Hobbs, and F.A. Wolak, "Economic Assessment of Transmission Upgrades: Application of the California ISO Approach", Ch. 7, in X.-P. Zhang, *Restructured Electric Power Systems: Analysis of Electricity Markets with Equilibrium Models*, Power Engineering Series, J. Wiley & Sons/IEEE Press, July 2010). That approach uses market simulation models to determine the reactions of generator decisions and the resulting changes in costs (as does van der Weijde and Hobbs, *op. cit.*). This can be contrasted to the much simpler analysis done of GB reinforcements as part of the study by the Electricity Strategies Network Group (ENSG, "Our Electricity Transmission Network: A Vision for 2020, Full Supporting Data", Full Report, Report ENSGR 2009-026, July 2009, http://webarchive.nationalarchives.gov.uk/20100919181607/http:/www.ensg.gov.uk/assets/ensg_transmission_pwg_full_report_final_issue_1.pdf). In that analysis, simplifying assumptions included: no consideration of CO_2 costs; generation construction is assumed fixed (so no investment cost savings; and the constraint cost at the Scottish border was assumed to uniformly equal 90£/MWh during hours of congestion, without justification using production costing or market simulation results. We would recommend that such analyses be done in the future using widely available market simulation software in order to obtain more defensible estimates of congestion cost and other savings.

[52] The ERCOT CREZ study did not focus on individual wind farms, but rather indentified a number of particular regions within West Texas that had high wind potential. Several overall target wind capacity levels were considered, and plans were developed to enable bulk transport at these target capacity levels from these regions to demand centers. In the UK context, relevant regions for wind and tidal could be identified and plans developed based on national objectives for renewables. Radial connections from individual wind and tidal projects would be based on the subsequent pattern of actual development.

line begins, but the basic idea is any dedicated line to interconnect to the bulk transmission network is a shallow connection and any bi-directional transmission line used by a number of parties is a transmission line.

As we have discussed above, the goal of providing investment incentives for generation location using fixed access charges is laden with difficulties. Under standard economic principles, sunk cost should explicitly *not* influence location decisions. Instead recovery of those costs should be accomplished in a way that influences the actions of market participants the least. This is the principle of *Ramsey pricing*, which directs such charges to the users of the network who would be least elastic, or responsive, to those charges. It follows from Ramsey's principles that such costs should not influence location decisions and therefore there is not an efficiency argument to be made for a strong locational component to the TNUoS for recovery of sunk or fixed costs.

The question of how to deal with going-forward investments to relieve congestion that is caused by or anticipated by the location decision of specific generation plants is more complex. Because of the lumpiness of transmission investment, there is an argument that the entry of such plants can "cause" fixed costs to be incurred that will not be recovered (or deterred) by LMPs. This is because the LMPs will reflect little congestion costs *after* the transmission investment is made. Weighed against this argument is the stubborn fact that it is extremely difficult to pinpoint exactly which investments are caused by which plants, and which investments benefit which users. Further, just as the lack of a locational charge may encourage inefficient entry to a region that would not otherwise require network reinforcement, so can the presence of a locational charge discourage the entry of plants that could take advantage of that reinforcement, once the decision to pursue it has been made.

We do not see an overwhelming efficiency argument for attempting to impose locational aspects to the TNUoS for anything but shallow connection charges as long as the transmission planning and investment process can be viewed as holistic and not responding to the reinforcement demands of individual generators and only undertaking upgrades that satisfy a social cost versus benefits tests. We emphasize that this is not a central planning process for both transmission and generation; rather transmission is anticipating how decentralized expected profit-maximizing generation investments will respond to the incentives created by the transmission planning and expansion process. Specifically, in a wholesale market where generation units are paid LMPs, if generation unit owners know that transmission expansions are very unlikely, if ever, to occur in response to generation entry decisions, then generators will be more likely to locate where their presence enhances rather than detracts from wholesale market efficiency.

5.4. Apply these fixed network charges 100% to load

Consistent with principles of setting prices to recover fixed costs with minimal distortion from efficient pricing, we also support applying all non-shallow transmission charges to loads, because loads ultimately pay for the cost of the transmission network[53] and direct assignment of these costs to load is unlikely to distort the behavior of all but the largest electricity consumers. In contrast, direct assignment of these costs to generation unit owners

[53] This is true for a closed system; to the extent that the GB trades power with other countries, and these exchanges are sensitive to prices, this is less true.

can distort generation entry and operating decisions. For these reasons, we favor direct assignment of these costs to loads.[54]

With load covering the cost of the transmission network, generators can focus their entry decisions on the most profitable location in terms of expected future energy prices, without having to worry about the risk of future changes in the TNUoS at that location relative to others. Therefore, this approach lowers the future price risk faced by potential new entrants relative to a scheme that also allows for spatial prices of the TNUoS.

An additional advantage of allocating fixed network charges entirely to load is that this will make the GB charging system more consistent with those of its immediate neighbors on the European continent. This harmonization should result in more efficient cross-border competition between generators in the GB and elsewhere.

5.5. Targeted Investments for grid enhancement to be funded by generators

We envision that the vast majority of transmission expansions would take place through the planning and evaluation process. However, there may be instances where market participants have other reasons besides the satisfaction of a social cost benefit test as their reason for undertaking an expansion, or where the planning process might overlook an efficient expansion. Therefore, we recommend that interested stakeholders be allowed to undertake any transmission expansion they would like in exchange for FTRs for the increments transmission capacity created by this expansion.

The above approach allows transmission investments to follow generation investments, but only if a specific market participant is willing to fund the transmission project and only receive FTRs for its investment expenditures. Although it is difficult to believe that this path to transmission expansion will be taken very often, the option should remain open just in case an instance arises.

5.6. A hybrid approach: LMP with deep connection charges

We note that the notion of a system with forward looking planning for transmission investments whose costs would be socialized need not preclude other investments whose costs may be allocated to specific beneficiaries. The two approaches could be combined in a hybrid of socialized and locational (deeper) capital charges. A hybrid model would still feature a long-term forward-looking planning process for expansion of a "backbone" transmission network. Other investments could be triggered by specific generation investments. Costs for these generation-led investments would be recovered through location, or even facility-specific capital charges. We would favor a system that determines a capital charge up-front, to be recovered through an amortized schedule. These annual costs would therefore be set for the duration of the recovery period and not a source of cost uncertainty. Short-run LMPs would apply to all facilities as an integral part of the congestion-

[54] An issue in this context is the treatment of distributed generation located in the distribution system, since the charging methodology will have different incentives if charged to load versus being charged to net load. Issues related to such embedded generation our outside the scope of this report; however, we recognize that these should be carefully considered so as to minimally distort the investment decisions of developers of embedded generation.

management process, but the payment of deep connection charges would entitle the payer to FTRs as a hedge against LMP differentials.

6. Consistency of Recommendations with Policy Objectives

This section examines the consistency our recommendations with the government's policy objectives for the electricity sector. The license conditions for National Grid state that the goal of the review of the charging system is to achieve the following objectives:[55]

1. *Facilitation of competition in the sale, distribution, and purchase of electricity.*
2. *Cost reflectivity in transmission charges, except for transmission congestion costs which are to be socialized.*

In addition, the TransmiT "Call for Evidence" mentions the following objectives:[56]

3. *Provision of value for money for consumers* (which we interpret as economic efficiency in construction and operation of the transmission system).
4. *Facilitation of a timely move to a low-carbon energy sector.*
5. *Delivery of safe, secure, and high quality network services.*
6. *Integration of GB electricity markets with the wider European market.*

Finally, our review of the extensive submissions made in response to the TransmiT "Call for Evidence" indicates that the following additional objectives are also widely, but not unanimously viewed as important by stakeholders:

7. *Predictability and stability of transmission charges and minimization of regulatory risks.*
8. *Technology neutrality.*
9. *Consistency with predominantly bilateral market structure.*
10. *Transparency and ease of administration.*

We discuss the consistency of our recommendations with each of these objectives.

6.1. Facilitation of competition in the sale, distribution and purchase of electricity

Locational pricing, by sending transparent signals about the cost and value of power at different locations at different times, allows investors to assess the market value of their generation and storage assets in a way that is consistent with the actual value to the system, in terms of energy, losses, and avoided congestion costs. Although these prices are variable, they are closely related to technology and market fundamentals and can be hedged by financial instruments.

Moreover, if generators submit their marginal variable costs as their offers into the balancing mechanism, then the dispatch that results from the locational marginal pricing process is the least cost solution to meeting load at locations in the network subject to all relevant operating constraints included in the nodal-pricing process. Recent evidence from the California market suggests that this potential operating efficiency gain has been realized.

[55] http://epr.ofgem.gov.uk/document_fetch.php?documentid=15246 , Condition C5, Clause 5, pp. 158-159. We note that the cost reflectivity objective in the license condition specifically exempts congestion costs. Licence condition C5 also requires National Grid to account for developments in transmission licensees' transmission business.

[56] www.ofgem.gov.uk/Networks/Trans/PT/Pages/ProjectTransmiT.aspx

The switch to nodal pricing is estimated to have reduced the total fossil fuel energy used to meet load and the total variable operating costs associated with meeting load by approximately 2 percent.[57]

6.2. Cost reflectivity in transmission charges, except for transmission congestion costs, which are to be socialized

In contrast, we are recommending that energy transmission charges reflect transmission congestion costs. By definition, locational marginal prices fold the costs of energy, losses, and congestion into one transparent price, and represent the actual marginal value to the system of withdrawing energy at a particular location at a particular time. We propose that shallow interconnection costs, those caused by a generation unit connecting to the bulk transmission grid, be paid by the generation unit owner. However, we recommend that deep interconnection costs be socialized. This is because determining which market participant's actions caused what fraction of a transmission upgrade is typically difficult to determine with any precision in an increasingly meshed system in which security constraints are becoming more complicated and intermittent generation makes the timing and location of congestion more difficult to predict.

6.3. Provision of value for money for consumers (*i.e.*, economic efficiency in construction and operation of the transmission system)

Economic efficiency in operations will be improved because short run prices will equal the marginal value to the system of generation, if adequate protections against the exercise of local market power are in place. Because of the possibility that different prices can be set at all locations in the transmission network, the prices paid to individual generation unit owners can be adjusted to ensure that they are always compatible with system operator's desired operating point for the generation unit owner. For example, it is never the case that a generation unit owner will be paid a price below or above the minimum necessary to cause the market participant to operate at output level required by the system operation. This is in contrast to a "copper plate" pricing system in which generators sited in generation pockets have an incentive to over-schedule to gain the system-wide price, while generators in load pockets have an incentive to under-schedule so that they can obtain a higher constrained-on price.[58]

Under locational pricing, consumers do not pay inflated prices for generation that is overscheduled because of the failure to efficiently price congestion and then can buy back its obligation at a lower price than it receives. This DEC game, and other games, are eliminated or minimized (see Appendix A.2).

---

[57] F.A. Wolak, "Measuring the Benefits of Greater Spatial Granularity in Short-Term Pricing in Wholesale Electricity Markets," POWER Meeting, University of California, Berkeley, March 17, 2011, available at www.stanford.edu/~wolak.

[58] As explained in an earlier footnote, such behavior is not, strictly speaking, the exercise of market power (the unilateral taking of actions to move prices), as small generators have as much incentive to behave this way as large ones. An advantage of locational pricing is that there would be no need for rules to prohibit such behavior (examples of such rules are some of the provisions of the Market License Conditions proposed by the government in Part 3 of the Energy Act 2010), because it would no longer be profit-maximizing for generators. In general, market designs that pursue efficiency by enlisting rather than prohibiting profit-maximizing behavior are preferred.

Efficiency of generator investment is encouraged by a transparent and systematic transmission planning process that makes investment commitments in a timely manner to the most efficient areas for generator construction. Generators will then be incented to site there because they would earn higher locational prices in those areas than elsewhere.

Efficiency of transmission investment is most likely to be ensured in what we characterize as a "plan and price" approach through careful analysis and regulatory review of benefits and costs of alternative transmission plans. Tallied benefits could include, for example, reduction in generation costs and wholesale prices to consumers, amelioration of the ability of generators to exercise unilateral market power in isolated markets, environmental improvement (as measured by the market value of renewable energy and carbon reductions that the investment makes possible), enhancements in reliability, and hedging against uncertainties concerning future technological, economic, and policy developments. As a check against "false negatives" from this process, generators and other market parties should be allowed to propose their own transmission reinforcements, gaining the additional financial transmission rights that then become possible.

6.4. Facilitation of a timely move to a low-carbon energy sector

Effective market mechanisms for incenting reductions in carbon production, including tradable renewable obligation credits, European Union (EU) Emissions Trading System (ETS) CO_2 allowances, and other market instruments, mean that the price of power at different locations and different times will internalize the cost of the carbon produced at that location. All else being equal in costs of production, if power produced at one location at one time yields more carbon reductions (or less displacement of renewable energy) than power produced somewhere or sometime else, then the price at the former location will be higher. The patterns of locational prices will also provide a guide to where renewable and low carbon generation can be sited in order to provide the most value to the power system, where "value" will, again, internalize carbon costs.

This is different from the case of single price system, where market for the entire system might reflect the highest cost unit on the system, including the price of carbon. Under a locational pricing system, if the market separates because of increased carbon costs at one location, the price energy throughout the entire transmission will not increase. Only the prices at the high cost carbon location will increase, thereby providing strong incentives for lower carbon energy sources to enter at that location.

6.5. Delivery of safe, secure, and high quality network services

A power system that is priced in a way that is consistent with the physics and security constraints of the system-–that is, short-run locational pricing reflecting actual system conditions-–is more likely to incent generator behavior that contributes to rather than works against system reliability. In contrast, short run "copper plate" pricing (*i.e.*, giving all generators the right to fully schedule in the market with no short run incentive for congestion management) results in more infeasible schedules, and necessitates a greater number of operator actions close to real-time to adjust generation and power flows in order to satisfy system constraints. Generator investment will also be incented to take place where power is most valued by electricity consumers, recognizing its effect on those constraints.

6.6. Integration of GB electricity markets with the wider European market

According to stakeholder submissions in response to the TransmiT call for evidence, the greatest obstacle to integration (other than lack of interconnector capacity) is the 73%:27% split of TNUoS costs between load and generators. Most neighboring countries, in contrast, have load bear 100% of the costs;[59] as a result, their generators arguably have a competitive advantage against GB generators. Aligning transmission fixed cost allocation in a manner consistent with neighboring countries would level the playing field.

6.7. Predictability and stability of transmission charges and minimization of regulatory risks

Although locational pricing could increase the variability of energy prices, these risks are hedgeable by financial transmission rights backed by congestion revenues. Generators and purchasers of power can sign bilateral contracts that reduce both of their risks from congestion. It is well-known that fixed-price long-term contracts sold by generation unit owners to electricity retailers limit the incentives of generators to exercise unilateral market power in the short-term energy market. As noted earlier, FTRs can be used to transfer a long-term contract clears against the price at the node where the retailer withdraws energy from the network to the point where the generation unit owner injects power into the network.

Congestion risks are the result of technical and economic drivers which, although not perfectly predictable, at least reflect market fundamentals. Furthermore, the system operator is the natural counterparty for hedges of congestion costs. This is because congestion rents that the operator earns generally equal or exceed the operator's payment obligations, as long as the operator does not overallocate financial transmission rights. In contrast, TNUoS charges, which have problems of predictability largely due to regulatory uncertainty, cannot be hedged, in part because there is no natural counterparty, unlike the case of locational energy prices. In contrast, the system operator is the natural counterparty for hedges of congestion costs. This is because congestion rents that the operator earns generally equal or exceed the operator's payment obligations, as long as the operator does not overallocate financial transmission rights.

6.8. Technology neutrality

The congestion costs faced by different technologies will depend appropriately on the timing and nature of their use of the network, and not directly on how the power is produced. No artificial and arbitrary scaling factors will be needed to adjust charges to account for the nature of a technology's network use; locational prices will automatically reward those technologies that produce when congestion is less of a problem.

6.9. Consistency with predominantly bilateral market structure

The GB market is dominated by bilateral transactions between generators and suppliers (load-serving entities, in US parlance), with only a small fraction of power

[59] In Belgium, France, the Netherlands, and Germany, all or nearly all of transmission fixed costs are paid by load, while in the Republic of Ireland, 80% is borne by load (ENTSO-E, "Overview of transmission tariffs in Europe: Synthesis 2010," Sept. 2010, www.entsoe.eu/fileadmin/user_upload/library/Market/Transmission_Tariffs/20100914_Transmission_Tariffs_Synthesis_2010.pdf)

exchanges consisting of direct sales or purchases between a market party and the grid operator. This is also the case in the US locational pricing systems. Most output of US generators is contracted for sale to other parties. Market parties who have made a forward transaction can choose to submit a fixed MW schedule for the amount of the transaction. However, it is in fact profitable for all parties to accompany a schedule with adjustment bids. As a result, such bids are the rule rather than the exception in US markets, and they benefit the system as a whole by providing flexibility to the operator to adapt to changing conditions by seeking the lowest cost dispatch at each time.

6.10. Transparency and ease of administration

Locational prices, by including the full cost of energy, losses, and re-dispatch necessary to meet system constraints, are the most transparent way to operate a power transmission system economically and feasibly. In contrast, the disregarding of operating constraints when pricing energy will imply that complex and non-transparent mechanisms will be necessary to manage the resulting infeasibilities. Currently, market license conditions are needed in GB to prevent the "dec" game and the deleterious effects of other incentives that arise because of the mismatch of physics and financial incentives (*i.e.*, the disregarding of congestion in pricing). Under locational pricing, no such market license conditions are necessary, although a local market power mitigation mechanism would be needed.[60]

This is not to say that a transition by GB to a locational pricing system would be inexpensive. Experience in the US shows that the information systems and energy management systems required can be quite costly, although their expense can be quickly recovered by the greater operating efficiencies that result.[61] Extensive negotiations among stakeholders, regulators, and the system operator are necessary to settle details of design of energy, ancillary services, and financial transmission rights markets, including definitions of property rights, allocation of those rights, creation of rules for exchange and financial settlements, and mechanisms to mitigate local market power. To the extent that systems already exist for optimal system dispatch and power flow, and for short term monitoring and control of generation and loads, the transition is simplified.[62] Use of software and market mechanisms developed elsewhere (as ISO-New England did, by borrowing ideas and software from PJM) also lowers costs and transition times.

---

[60]As Appendix A.2 points out, local market power is not created by locational energy pricing, and is a problem whether prices are differentiated spatially or not. As a result, if such mitigation mechanisms are needed under locational pricing, they would be needed under zonal or copper plate pricing systems as well.

[61]The "gains to trade" from the expansion of the PJM locational pricing system to the US Midwest paid for the costs of the required software in approximately one year (F. Mansur, & M. White, "Market Organization and Market Efficiency in Electricity Markets," Yale School of Management Working Paper, June 2009. bpp.wharton.upenn.edu/mawhite/papers/MarketOrg.pdf). Elsewhere, it was estimated that the annual operating costs savings from the introduction of LMP in California was approximately $100 million (Wolak, 2011, *op. cit.*).

[62]For this reason, the transition by PJM (which already had a system-wide dispatch system) to locational pricing in 1998-99 was rapid and relatively inexpensive compared to the California and ERCOT (Texas) systems in 2009 and 2010, respectively. The latter two systems required integration of energy management systems for several utility companies who did not already calculate locational marginal costs. It is our understanding, for example, that National Grid will be refreshing their balancing market system in the near future by implementing optimal AC load flow software (from the New York ISO) that will produce marginal cost information at every bus.

Appendices

Appendix A.1 Summary of locational pricing[63]

This tutorial on nodal or locational marginal pricing first describes offer-based economic dispatch in the absence of transmission constraints and then includes the effect of transmission constraints in a stylized example system.

In offer-based economic dispatch, an offer by a generator is a specification of the minimum price it is willing to receive to produce, versus the quantity of production. An offer applies for a particular hour or range of hours. To simplify the presentation, we consider a particular hour, ignoring intra-hour variations, and a particular type of offer, namely a "block" offer. Specification of a block offer requires a quantity and a price and can be interpreted as an offer:

- to generate at any level up to maximum power in the block in MW,
- for remuneration at least equal to the nominated price in $/MWh.

Figure A.1.1. Block offer.

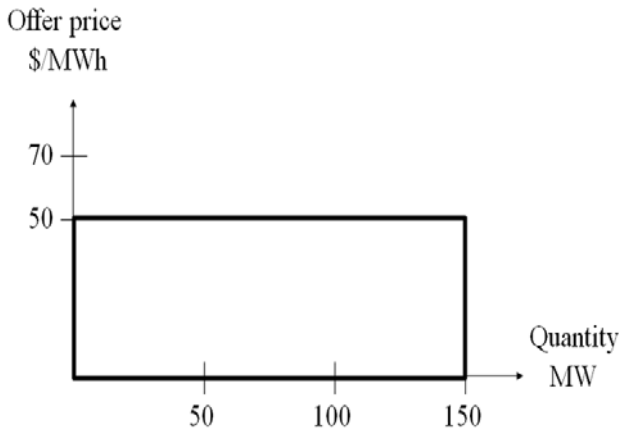


Figure A.1,1 shows an example block offer for 150 MW at a price of $50/MWh. The ISO receives offers from the various generators in the market and then selects amongst the offers to meet its short-term forecast of demand based on the offer prices. As a general principle, the ISO uses an offer with a lower offer price in preference to a higher offer price.

A significant issue for market participants is the model of price formation; that is, the rules under which prices are set. Roughly speaking, the highest *accepted* offer price or, equivalently, the offer price that would serve an additional MW of demand, sets the price for all energy sold. This description is somewhat loose and a more careful definition is needed if, for example, there are insufficient offers to meet demand, the demand is at a jump in prices

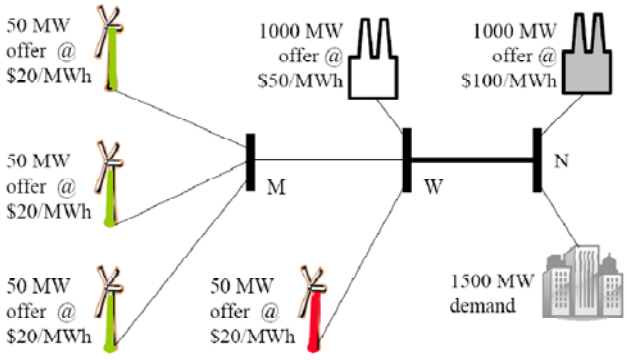
[63]This Appendix based on material in R. Baldick, "Wind and Energy Markets: A Case Study of Texas," to appear in *IEEE Systems Journal*, 2011. For a summary of the fundamentals of financial transmission rights, see Bushnell and Stoft (1996), *op. cit.*.

between blocks, and in the case of limiting transmission constraints ("congestion"), which will be considered in more detail below.

To illustrate the operation of a real-time market that includes wind resources, we will consider a very simple example system. The transmission in this system consists of just two lines joining three "buses," M, W, and N, which are the points of interconnection between generators, transmission lines, and substations that serve demand. This topology simplifies the situation compared to reality, but is useful as a start.

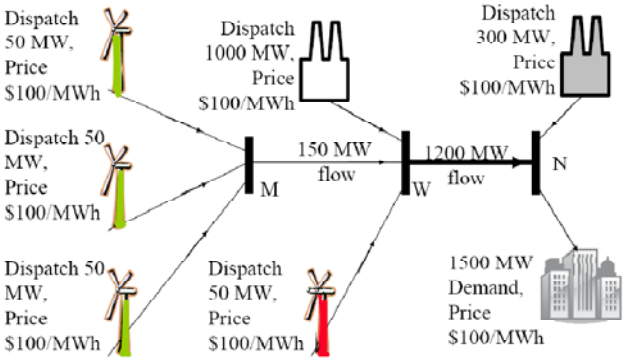
Although some market designs do not allow or do not require wind to make explicit offers, we will assume that wind generators offer into the market. Wind (at buses M and W) and thermal (at buses W and N) submit block offers into the real-time market to meet 1,500 MW of demand (at N). To distinguish between the various generators, the wind farms at M will be called the "green" wind farms, the wind farm at W will be called the "red" wind farm, the thermal generator at W will be called the "white" generator, and the thermal generator at N will be called the "gray" generator. The offer prices are illustrated in Figure A.2: wind offers at the lowest price of $20/MWh, then the "white" thermal at $50/MWh, and then the "gray" thermal at $100/MWh. We will first consider unlimited transmission and then consider limited transmission.

Figure A.1.2: Example with unlimited transmission, 1500 MW demand at N, and block offers.



To meet demand on the basis of using lower offer priced generation in preference to higher offer priced generation, all 200 MW of available wind generation, 1,000 MW of "white" thermal, and 300 MW of "gray" thermal generation are required. Amongst the dispatched generation, the highest accepted offer price was $100/MWh from the "gray" thermal generator at bus N. Note that to serve an additional MW of demand at any bus it would be necessary to use an additional MW of "gray" generation. The "green" and "red" wind and the "white" thermal generator are all fully dispatched. The situation is illustrated in Figure A.1.3, and to summarize, the price paid to all generators and paid by demand is $100/MWh.

Figure A.1.3. Dispatch and prices for 1500 MW demand, unlimited transmission capacity.



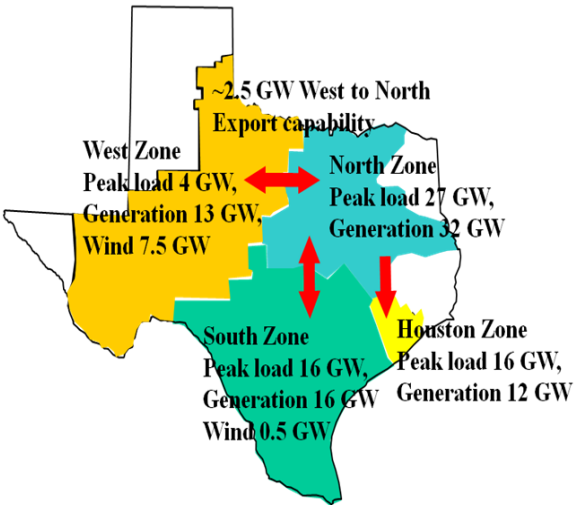
In practice, limitations on transmission capacity can limit the flexibility of the ISO to dispatch from the lowest priced offers. If the limited capacity of transmission prevents the use of an offer with a lower price then the highest accepted offer can be thought of as *varying* with the location of the bus. So-called "nodal" or "locational marginal prices" reflect this variation. Roughly speaking, the price at each bus is based on the offer price to meet an additional MW of demand *at that bus*. These prices provide incentives for market participants to act in accordance with optimal transmission-constrained dispatch, avoiding incentives to over-schedule as in the "inc and dec game."

From Figure A.1.4, illustrating the ERCOT system, it is clear that the wind is predominantly in the West zone and that the export capability from the West zone, which varies but is on the order of 2.5 GW, is significantly less than the wind generation capacity. These stylized facts will be represented into the example to qualitatively understand the effects of transmission constraints.

In particular, we will now assume that there is only 100 MW of transmission capability from bus M to bus W and only 1,000 MW of transmission capability from bus W to bus N. Optimal dispatch now involves only 100 MW of production in total from the three "green" wind farms, 50 MW from the "red" wind farm, 850 MW from the "white" thermal generation, and 500 MW from the "gray" thermal generation. The presence of transmission constraints has necessitated less use of low offer price resources and more use of higher offer price resources.

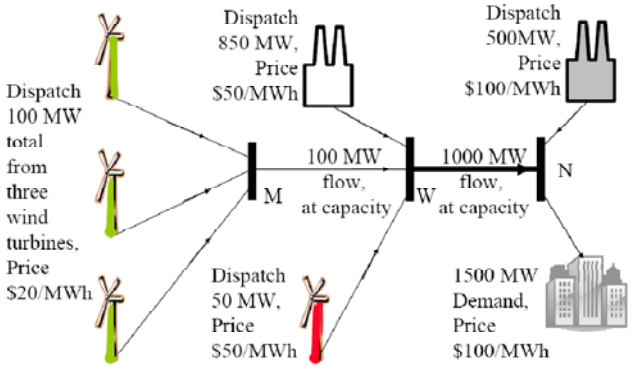
Again, the highest accepted offer price was $100/MWh from the "gray" thermal generator at bus N, but this no longer determines the price at the other buses because of the transmission limitations. While the "red" wind fully was dispatched at bus W, the "white" thermal generator at bus W was not fully dispatched. Moreover, the "green" wind farms at bus M were not fully dispatched.

Figure A.1.4. ERCOT zones. Source: ERCOT.



Since the wind farms are not generating at their full potential, they have been curtailed. Since the process is the result of a market-based process, we might refer to this as "economic curtailment" to distinguish it from, for example, a quantity rationing basis where the ISO decides on the amount of generation without (direct) reference to bids and offers. For example, until the advent of the ERCOT nodal market in December 2010, wind curtailment in ERCOT due to intra-zonal transmission limitations has typically been on the basis of quantity rationing rather than economic rationing; however, we will focus on economic rationing, consistent with locational marginal pricing.

Figure A.1.5. Dispatch and prices with limited transmission.



To translate the observations about dispatch into the LMPs, denote the LMPs at buses N, W, and M, by LMP_N, LMP_W, and LMP_M, respectively. Now note that:

- To meet an additional MW of demand at N, we would dispatch an additional MW of $100/MWh "gray" thermal generation, so LMP_N = $100/MWh at N,
- To meet an additional MW of demand at W, we would dispatch an additional MW of $50/MWh "white" thermal generation, so LMP_W = $50/MWh at W,

- To meet an additional MW of demand at M, we would dispatch an additional MW of \$20/MWh "green" wind generation, so LMP_M = \$20/MWh at M.

Because of these prices, the "green" wind is paid \$20/MWh and the "red" wind is paid \$50/MWh. These outcomes are shown in Figure A.1.5.

Although the example system is "radial," locational marginal pricing also applies to more typical meshed systems. The calculations of prices become more complicated due to the interaction of Kirchhoff's laws with the capacity limitations. However, the principles are similar and result in prices that provide incentives to market participants to behave consistent with optimal transmission-constrained dispatch.

Appendix A.2. Market power and related issues with locational marginal pricing

A criticism occasionally leveled at locational pricing is that market power issues will be worse due to the separation of markets into less liquid sub-markets when there are transmission constraints. It is important to understand that the transmission constraints exist under any system -- whether the "commercial market model" includes their effect directly in the energy prices or if instead an out-of-market procedure is used to ensure satisfaction of the transmission constraints.

For example, consider a system that has an abundance of low marginal cost resources in the North, but does not represent a particular North-to-South constraint in the commercial market model. That is, there is an "unconstrained" market price that is not differentiated by whether the generation is located in the North or the South. Whether the market is primarily arranged through bilateral contracting or through offer-based economic dispatch, it can be expected that the resulting generation schedules that ignore the constraint would result in flows that would exceed the transmission limitation. Inevitably, the system operator must then arrange for some out-of-market process to re-dispatch compared to the schedules into order to reduce production in the North and increase it in the South. To the extent that there are limited choices for re-dispatch, and in the absence of regulation on the actions of the market participants able to provide the re-dispatch services, those market participants can presumably extract a premium from the system operator compared to the competitive outcome.

In a market with a commercial model that does represent the constraint, there is still a concern about the market power of these same market participants. However, the concern is essentially no worse than in the case where the constraint is hidden from the market, although more market participants may be directly exposed to the effects of market power when the commercial network model includes representation of constraints. To summarize, market power mitigation will be necessary whether or not the commercial network model explicitly represents the constraint.

Three further issues tend to worsen the performance of systems that do not represent transmission constraints into the commercial network model. First, there is typically a requirement for the re-dispatch to minimally affect the initial "unconstrained" schedule, and in practice only some of the market participants may participate in the re-dispatch process. This worsens the efficiency of the resulting re-dispatch and exacerbates market power issues, by reducing the competition between market participants to provide the re-dispatch service

compared to the case where the market participants were all competing in a single-step transmission-constrained offer-based dispatch.

Second, when constraints are not represented in the prices, market participants have incentives to over-schedule compared to the capacity of the constraint since they anticipate that they will then be compensated for the re-dispatch that is necessitated by their own scheduling actions. (See Appendix A.3 for evidence of this activity in the context of the ERCOT market; it was also an essential part of the "inc and dec game" in the initial implementation of the California market.)

Finally, the market participants that do generate in what would be the lower priced market (the North in the above example) are being exposed to a price that is above the market clearing price; that is, they are over-compensated for their production, while the participants in the other market (the South in the above example) are being exposed to below market clearing prices. These prices will provide the wrong incentives for investment in the respective markets. (For example, wind generation in the McCamey area of West Texas was exposed to a zonal price, despite local transmission constraints on export of capacity, resulting in too much capacity siting in this region and eventually necessitating wind curtailment.)

Financial transmission right systems can also help decrease the exercise of local market power in locational energy price systems, compared to other systems. In general, if a party can exercise market power by changing the local price and thereby the payoff of an FTR, the ownership of FTRs can either increase or decrease the motivation to do so. Generators who own point-to-point FTRs sourced at the point of generation (*i.e.*, the payoff is the price at some other bus minus the generator's local price) will have less incentive to increase the local price. This is because the increase in LMP is matched by a decrease in the payoff of the FTR). On the other hand, generators owning FTRs that are instead sunk at their local bus have a magnified incentive to raise prices; U.S. system operators have position limits designed to prevent such situations from arising. Fortunately, the former situation is likely to more common than the latter.

Appendix A.3 Experience with locational pricing in Texas[64]

This appendix describes some of the context and experiences in congestion management in the Electric Reliability Council of Texas (ERCOT) balancing market since its beginning in 2001.

On May 21, 1999, the Texas Legislature passed Senate Bill 7 (SB7). Under SB7, the ERCOT Independent System Operator (ISO) was given the responsibility to develop the wholesale market structure, infrastructure, and business processes to facilitate retail competition in Texas. The ERCOT market began to operate as a single "control area" under the ERCOT ISO on July 31, 2001. Market participants, the "qualified scheduling entities" (QSEs), submitted "balanced schedules" of generation to meet specified demand, and the ISO also operated a "balancing" market to compensate for deviation between actual and

[64] The material is primarily based on R. Baldick and H. Niu, "Lessons Learned: The Texas Experience," in J. Griffin and S. Puller, Editors, *Electricity Deregulation: Where to from here*?, University of Chicago Press, 2005.

scheduled demand and between actual and specified generation. The ISO was also responsible for managing transmission congestion and the mechanisms for managing congestion evolved over several years, culminating in December 2010 with the implementation of a locational marginal pricing market that explicitly incorporates congestion into market prices.

However, when the ERCOT single control area originally began operation in 2001, congestion re-dispatch costs were "uplifted" (that is, charged) to market participants on a "load ratio share" basis and not explicitly incorporated into market prices. This presented an opportunity for profiting by over-scheduling and then being paid to relieve congestion. This is similar to the "Inc and Dec" game in the California market. Serious over-scheduling was observed in August 2001. The re-dispatch costs and the costs related to load imbalance, resource imbalance, and uninstructed deviation were aggregated in what are called Balancing Energy Neutrality Adjustment (BENA) charges. BENA charges for August 2001 alone were approximately $75.9 million. Six QSEs received more than $2 million each in load imbalance revenues for that month. A settlement was eventually reached with them agreeing to refunds gains from the ERCOT market.

The potential for this problem was anticipated[65] and the PUCT required ERCOT to switch to charging of inter-zonal congestion rents on nominated "commercially significant constraints" (CSCs) by January 1, 2003 or six months after inter-zonal re-dispatch costs rose above $20 million on a rolling twelve-month period, whichever came first. It also required ERCOT to implement a system of transmission congestion rights (TCRs), which would allow market participants to hedge their inter-zonal congestion charges on the CSCs.

The $20 million threshold for inter-zonal re-dispatch costs was reached on August 15, 2001, just 15 days after the beginning of operation as a single control area. A zonal balancing market, with each zone effectively a node in a locational marginal pricing system joined by CSCs, was implemented on February 15, 2002. Under that system, each QSE submitted bids and offers in each zone to adjust its portfolio of generation from scheduled levels and then ERCOT operated the balancing market as effectively a locational marginal pricing market with a handful of nodes, each of which represents a single zone. Under this design, the charge or payment to a QSE was based on the product of its scheduled flow and shadow prices on the congested CSCs. That is, a QSE was exposed to the variation of the shadow price for the CSC.

Transmission Congestion Rights (TCRs) and Pre-assigned Congestion Rights (PCRs) were implemented as financial hedges against the zonal congestion rent. The TCR and PCR holder received an amount equal to the congestion rent for an equivalent quantity of scheduled flow. TCRs were awarded in yearly and monthly simultaneous combinatorial auctions based on the auction clearing prices.

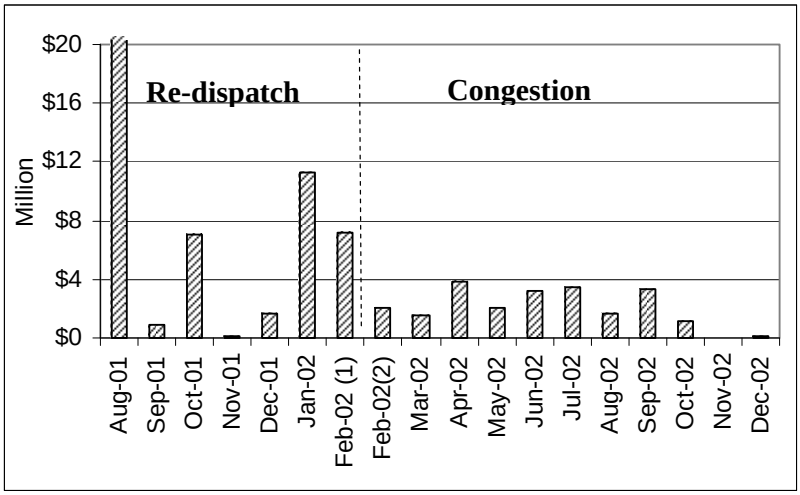
Figure A.2.1 shows the monthly zonal re-dispatch costs (until February 14, 2002) and congestion rent (after February 15, 2002) in ERCOT. Until February 14, 2002, zonal re-dispatch cost was uplifted to all QSEs in the system based on their load ratio share. From February 15, 2002, "direct assignment" of zonal congestion rent was implemented in ERCOT, under which the congestion charge or payment for each QSE is based on the shadow prices

[65]S. Oren, "Report to the Public Utility Commission of Texas on the ERCOT Proposals." February, 2001.

and power flow it scheduled on the congested CSCs. Zonal congestion rent after February 15, 2002 was significantly less than the re-dispatch cost prior to February 15, 2002.

Since it is normally the case that, at a given time, congestion rent is much greater than re-dispatch costs, the observations strongly suggest that significant over-scheduling was taking place prior to February 15, 2002. Over-scheduling across the CSCs has stopped and should not re-occur because the change to direct assignment of zonal congestion rent removed the incentives for QSEs to over-schedule load across the CSCs.

Figure A.2.1: Zonal Re-dispatch Cost (Aug. 1, 2001 – Feb. 14, 2002) and Congestion Rent (Feb. 15, 2002 – Dec. 30, 2002) (Source: ERCOT, "Market Information," www.ercot.com/ Participants/PublicMarketInfo/PublicMarketInformation.htm, Accessed March 28, 2003)



The problem that existed for zonal congestion prior to February 15, 2002 still remained for "local" or intra-zonal congestion after implementation of the zonal congestion management mechanism. In particular, generation was re-dispatched by ERCOT to compensate for schedules that would have resulted in violation of transmission constraints on lines internal to zones. However, ERCOT relies on a more detailed operational model to determine how each particular resource or load affects the transmission system and this model does not use portfolio offers and bids. Each resource was required to submit resource specific premiums (positive or negative) and the resource-specific dispatch ranges. The resource specific premiums and unit specific shift factor were used to relieve local congestion through a set of balanced adjustments to local resources in each zone. Resources in other zones may be chosen when there is no solution within local resources.

The ERCOT protocols defined a "market solution" for local congestion as when at least three unaffiliated resources, with capacity available, submit bids to the ERCOT ISO that can solve the local congestion and no one bidder is essential to solving the congestion. If there is no market solution then bid prices are mitigated based on verifiable operating costs.

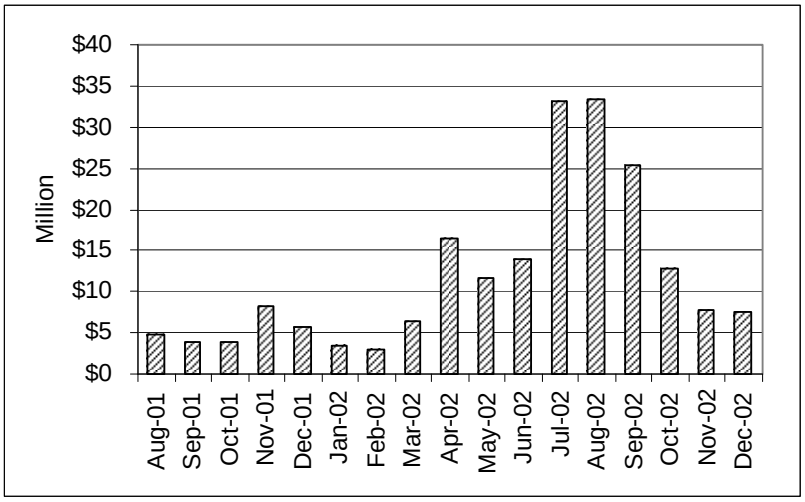
There has been no "market solution" for local congestion in ERCOT in most cases. That is, local market power is deemed to exist most of the time when local transmission constraints are binding. Instead of relying on a market process to determine prices, ERCOT obtains commitments to provide capacity and energy at a pre-specified cost level. These are

called Out of Merit Order Energy (OOME) and Out of Merit Order Capacity (OOMC). OOME services are provided by resources selected by ERCOT ISO outside the bidding process in order to resolve local congestion when no market solution exists. OOMC provides generation capacity needed such that balancing energy is available to solve local congestion or other reliability needs when a market solution does not exist. OOMC can be provided from any resource or load acting as a resource that is listed as available in the resource plan.

Sometimes a Reliability Must Run (RMR) unit was needed to provide generation capacity or energy resources when there was no market solution. A RMR unit is a generation resource unit operated under the terms of an annual agreement with ERCOT that would not otherwise be operated except that they are necessary to provide voltage support, stability, or management of localized transmission constraints under first contingency criteria where Market Solutions do not exist.

The local congestion cost was uplifted to each QSE based on the load ratio share of the QSE. Figure A.2.2 shows the local re-dispatch costs in ERCOT from August 2001 to December 2002. Evidently, the submission of schedules that would result in congestion shifted from a largely zonal phenomenon prior to February 15, 2002 to a largely local phenomenon, suggesting strongly that market participants were creating congestion in their schedules in order to be paid to alleviate the congestion.

Figure A.2.2: Local Re-dispatch Cost of ERCOT. Source: ERCOT (2003), *ibid.*



In Docket 23220 of 2001, the Public Utility Commission of Texas (PUCT) ordered the ERCOT ISO to implement "direct assignment" of local congestion costs if the re-dispatch costs for resolving local congestion rose above $20 million in a rolling twelve-month period. The $20 million threshold for local re-dispatch costs was met on March 5, 2002, after seven months of operation as a single control area. Several proposals were suggested for solving the local congestion problem, including implementing locational marginal pricing. Ultimately, in 2005 the PUCT mandated locational marginal pricing to manage congestion and the ERCOT locational marginal pricing market opened in December 2010.

The ten year evolution of the ERCOT market from not representing congestion in market prices to locational marginal pricing illustrates several issues:

1. Intermediate, zonal representations are able to solve part of the congestion management issue, but typically then focus market participants on exploiting intra-zonal congestion, as illustrated in Figure A.2.2,
2. Actions of market participants will respond to incentives, and
3. Failing to represent transmission congestion into market prices provides incentives to market participants to create congestion in their schedules that they are then paid to relieve.

There has only been a brief experience with the ERCOT locational marginal pricing market. However, already there is evidence that the interaction of renewables and transmission constraints is better managed under locational marginal pricing. In particular, under zonal pricing, lack of fine controllability of dispatch of various resources meant that some of the transmission from the West zone could not be utilized. Under the locational marginal pricing system, these limiting constraints have been more fully utilized.