Flowgates, Contingency-Constrained Dispatch, and Transmission Rights

There are many reasons to doubt the basic assumption of a flow-based market; including that there are only a few commercially significant flowgates and these have fixed capacities and power distribution factors. Perhaps the most important is that flow constraints arise under multiple contingencies, not just constraints on actual flows of physical elements.

Larry E. Ruff

Larry E. Ruff *holds a Ph.D. in economics from Stanford University, and is an internationally recognized expert on the restructuring of electric and gas utilities to create competition and on the operations of the resulting competitive markets. He has advised governments and utilities in the United States, Canada, Europe, the former Soviet Union, the Indian subcontinent, the Asia-Pacific region, and Latin America on competitive restructuring. He has served as a Senior Vice President at National Economic Research Associates and a Managing Director of Putnam, Hayes & Bartlett Inc. He has held senior positions in government, industry, and nonprofit organizations.*

Proposals for flow-based electricity markets are based on the assumption that there are only a few commercially significant flowgates (CSFs) with fixed capacities and fixed power distribution factors (PDFs) that describe how power flows over individual elements of the system. These assumptions are highly suspect for many reasons, including nonflow (e.g., voltage) constraints that can be more constraining than flow limits, nonlinear physical relationships that cause PDFs to change with operating conditions, and the widespread use of equipment (e.g., phase-shifters) specifically designed to change PDFs. But one reason to doubt the basic assumptions of a flow-based market is particularly noteworthy, if only because it is so consistently misunderstood and underappreciated: The actual operations of an electricity system are constrained not only by the actual flows on individual network elements, but also by the flows (and voltages, etc.) that would appear under any of many contingencies such as sudden loss of a network element or a critical generating unit.

  PII S1040-6190(00)00174-3

This article uses a simple example to illustrate the process of contingency-constrained dispatch (CCD),[1] discusses some of the implications for a flowgate/flowgate rights (FGR) market, and explains why these same problems do not arise in a market based on locational marginal pricing (LMP) and point-to-point financial (or firm) transmission rights (FTRs). It is shown that a flowgate/FGR market on any complex system must have either very many (hundreds of?) abstract, contingent CSFs or only many (scores of?) physical CSFs, but each with many (scores of?) contingent capacities and PDFs. The only way to make FGR trading easy and liquid in such a situation is for the RTO to define for trading purposes an artificially simplified and restricted set of CSFs with fixed capacities and PDFs.

If the RTO creates an artificial world to make FGR trading workable, the market solutions arising from the artificial forward trading will often be far from what is actually feasible or efficient on the system, requiring the RTO to incur significant costs in real time to close the gap. If the costs of closing this gap are paid by the specific traders whose infeasible or inefficient forward schedules create them, the flowgate/FGR market will not provide price certainty or good hedges. If these costs are socialized across system users as a whole through some sort of uplift, prices will be distorted and costs will be shifted, creating both short-run and long-run inefficiencies and inequities. Thus, a flowgate/FGR market faces an inescapable dilemma: impossible complexity on one horn of the dilemma, and unacceptable inefficiency and inequity on the other.

The theoretical literature advocating flow-based markets, while sometimes recognizing the fact of CCD,[2] has not acknowledged the severity of the dilemma this raises or proposed plausible resolutions. One prominent advocate of flow-based markets has simply asserted that defining a CSF for each network element/contingency pair—the natural interpretation of the mathematics—is inconsistent with the theory of flow-based markets, but has not dealt with the implication of non-contingent CSFs with constantly changing capacities and PDFs.[3]

Advocates of flowgate/FGR markets virtually never acknowledge and may not even recognize the dilemma they face.

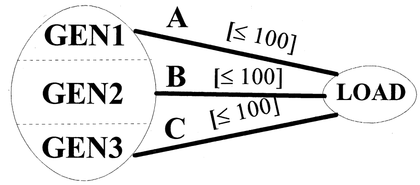
Whatever the flowgate theorists may say, advocates of flowgate/FGR markets in current policy discussions virtually never acknowledge and may not even recognize the dilemma they face. They continue to say that forward trading in a flowgate/FGR market would be easy, intuitive, and highly liquid because there would be only a few, physical CSFs with fixed capacities and PDFs. But they always hedge their bets by insisting that the system operator/dispatcher—assumed here to be a regional transmission organization (RTO)—should socialize the costs that will result when the basic assumptions of a flowgate/FGR market are not true.

CCD does not create the same sort of problems in an LMP/FTR market. In such a market, the LMPs are computed based on actual system operations using the full set of actual and contingency constraints, and the point-to-point FTRs are perfect hedges for point-to-point transactions for any LMPs, i.e., for any set of binding constraints (including nonflow constraints). The RTO deals with the technical complexity of flow and nonflow constraints to assure that the power gets from where it is produced to where it is consumed, and—as long as the set of FTRs outstanding is simultaneously feasible on the grid and the grid is in its standard condition[4]—can honor all the FTRs without socializing any costs. There are very many (thousands of?) logically possible point-to-point FTRs, but a single transaction can be perfectly hedged with precisely one FTR, and many similar transactions can be approximately hedged with one or a few FTRs; the problem of very many FTRs is fundamentally different than the problem of very many FGRs.

Electricity systems are inherently complex, much more so than the natural gas pipelines that are often used as analogies. It is only human to wish that somebody, somewhere, would invent some-

  PII S1040-6190(00)00174-3

thing that would somehow make electricity simple and more like natural gas. But a flow-based market is no more likely to perform this miracle of transubstantiation than is anything else. Once the many complexities of power flows—including the reality of many contingent and nonflow dispatch constraints—are taken into account, a flow-based electricity market would be anything but simple. The RTO could try to make it simple and profitable for traders at the expense of inefficient pricing and operations, large socialized costs, and an extensive role in the market for the monopoly RTO and its regulators, and the traders who would benefit would no doubt approve; indeed, they are actively advocating this in policy discussions now. But distorting prices and shifting costs would not benefit the consumers—probably small ones—who would pay the higher costs, and would not produce the efficient and effective competition that is the objective of electricity restructuring.



To LOAD from	PDF for Line		
	A	B	C
GEN1	0.5	0.3	0.2
GEN2	0.25	0.5	0.25
GEN3	0.2	0.3	0.5

Figure 1: The Three-Line System

I. Contingency-Constrained Dispatch in a Hybrid Market

This section describes the CCD process using a simple three-constrained-line example. It also discusses the concept of a "hybrid" market in which forward trading is based on flowgates and FGRs while real-time operations and pricing are based on LMP.

A. A Flow-Based Market on a Three-Constrained-Line System

The example system shown in **Figure 1** has three potentially congested transmission lines, A, B, and C, each with a 100 MW line rating. These lines connect a generation region GEN with three subregions, GEN1, GEN2, and GEN3 to a single load center LOAD. The PDFs indicate how energy generated in each of the three generation subregions distributes itself over the three lines to get to LOAD. For example, if an additional MW is generated at GEN1 and consumed at LOAD, an additional 0.5 MW will flow on Line A, 0.3 MW on Line B, and 0.2 MW on Line C. (Losses are being ignored.)

In simple explanations of flow-based markets, the system in Figure 1 would usually be described as having three potential CSFs, one for each potentially congested line or physical network element.[5] To establish a flow-based market, the RTO would simply sell 100 MW of FGRs on each CSF/line and publish the nine PDFs. Market participants would then trade the three types of FGRs (and energy) freely among themselves in forward markets to determine transactions or schedules that are consistent with the FGRs they hold. If the schedules submitted to the RTO are consistent with the outstanding FGRs and the PDFs, the scheduled/hedged flows will meet the constraints on each of the three lines from GEN to LOAD, and a market participant whose operations match its FGR portfolio will pay no real-time penalties or congestion charges.

For example, if a generator in GEN2 sells 10 MW to a buyer at LOAD, the generator (or the buyer) would need an FGR portfolio consisting of 2.5 MW (0.25×10 MW) of FGRs on Line A, 5 MW of FGRs on Line B, and 2.5 MW of FGRs on Line C in order to schedule or fully hedge the transaction.[6] If every schedule submitted is fully covered by FGRs in this way and the MW quantity of FGRs outstanding for any line does not exceed the physical flow limit on that line, the total scheduled flow on each line cannot exceed the line limit. Scheduled operations will be feasible and efficient.

B. A "Hybrid" Flow-Based/LMP Market

The early literature on flow-based congestion management implied or assumed that no markets other than the forward markets in energy and FGRs would be required, because the FGRs would accurately reflect the real constraints and trading would be efficient, so that the resulting schedules would be feasible and efficient. The RTO would have so little to do in real time that it need not—indeed, should not—operate a real-time market, but should just pay a few generators for a few ancillary services such as load-

  1040-6190/01/$–see front matter PII S1040-6190(00)00174-3

following and reactive power, and in the rare emergency use command-and-control methods such as transmission loading relief (TLR).

More recently, flowgate/FGR proponents have recognized that electricity is more difficult than natural gas, and hence forward trading will often produce schedules that are not fully feasible or efficient in real time, leaving more for the RTO to do. Indeed, most flowgate/FGR advocates now concede that the RTO should operate a real-time LMP market to manage and price real-time imbalances and congestion. This LMP market may be described as a residual market that will price and settle only small amounts of energy, but it will require most of the processes and systems required for a full LMP market[7]—although FTRs, which hedge against real-time LMPs, may not be needed if FGR trading leaves little real-time congestion.

In current policy discussions in the Midwest Independent System Operator (MISO), the Southwest Power Pool (SPP), and elsewhere, a system in which forward markets trade FGRs while the RTO operates a real-time LMP market is called a "hybrid" system. There are many important issues involved in defining such a hybrid, but these issues are not the primary focus here. The objective here is to illustrate the CCD process and its implications for a flowgate/FGR market. For this purpose, it is easiest to assume a hybrid market process based on one proposed in *The Electricity Journal* by Chao, Peck, Oren, and Wilson.[8]

In the hybrid process proposed by Chao *et al.*, the RTO identifies the CSFs, issues the FGRs, defines the PDFs used for trading in forward markets, and then operates a real-time LMP market that prices and settles all residual congestion and imbalances. In settlements, both spot and scheduled transactions pay congestion charges equal to the difference in LMPs between the sink and source locations, and holders of FGRs are paid the value of their FGRs, with both LMPs and FGR prices determined using the same PDFs. The mathematics of this process guarantee that, *if all congestion is on CSFs and if the PDFs used to define hedging portfolios of FGRs are the same as the actual PDFs used for pricing*, a transaction exactly covered by FGRs will receive FGR payments exactly equal to its LMP-based congestion payments, and hence will be perfectly hedged against real-time congestion costs. A transaction approximately but not exactly covered by FGRs will receive FGR payments approximately equal to its LMP congestion charges, so it is still approximately hedged.[9]

More recently, flowgate/FGR proponents recognized that electricity is more difficult than natural gas.

Chao *et al.* accept the standard flowgate/FGR assumptions and hence do not discuss what happens when there is nonCSF congestion or when real-time CSF capacities and PDFs are not the same as those used to define FGRs and hedging portfolios. But if the flowgate/FGR assumptions are correct, there is no reason the RTO should not use all actually binding constraints and actual PDFs to determine the LMPs and FGR prices used in real-time settlements. This rule would require traders to pay congestion charges reflecting any nonCSF congestion and changes in CSF capacities and PDFs, but these congestion charges should be commercially insignificant—if the flowgate/FGR assumptions are correct.

Because the objective here is to illustrate the problems created when the flowgate/FGR assumptions are not correct, it will be assumed that the RTO computes real-time settlement prices that reflect all actual congestion, not just congestion on CSFs, and does so using the constraints and PDFs that apply to the actual dispatch. It is worth noting that these are not the same pricing and settlement rules being proposed in policy discussions now underway at MISO and perhaps elsewhere. MISO is trying to develop a more complex two-stage settlement process that would shift residual congestion costs to the RTO and then on to system users through an uplift. But the rules proposed here would prevent the RTO from being stuck with—i.e., socializing—large costs when the assumptions

PII S1040-6190(00)00174-3

of the flowgate/FGR market are not correct, and hence are useful for discussing the implications of such a situation.

C. Non-Contingent Dispatch and Pricing of the Three-Line System

In the simplest interpretation of the three-line system of Figure 1, forward trading of the three FGRs will determine a set of planned transactions or scheduled flows from each of the three GEN subregions to LOAD. These schedules, along with incremental/decremental (inc/dec) offers indicating each generator's willingness (or not) to produce more or less at various prices, will be submitted to the RTO an hour or so prior to real time. The RTO will then use these schedules and inc/dec offers to determine a dispatch—generation G_1 at GEN1, G_2 at GEN2, and G_3 at GEN3—that meets the net demand at LOAD at least cost subject to the three transmission constraints. If the line capacities and PDFs indicated in Figure 1 are the ones that actually apply to the real time dispatch, the RTO will use the following three mathematical constraints in the dispatch process:

$0.5 \times G_1 + 0.25 \times G_2 + 0.2 \times G_3 \leq 100$ (the Line A constraint)

$0.3 \times G_1 + 0.5 \times G_2 + 0.3 \times G_3 \leq 100$ (the Line B constraint)

$0.2 \times G_1 + 0.25 \times G_2 + 0.5 \times G_3 \leq 100$ (the Line C constraint)

The dispatch optimization will automatically determine the LMPs at each location and the value or price of each of the three types of FGRs, with the PDFs used for dispatch—the coefficients of the G_i in the above constraint equations—determining the relationship between the LMPs and the FGR prices. In settlements, each scheduled transaction from GEN to LOAD will pay a per-unit congestion charge equal to the LMP at LOAD minus the LMP at the generator's location, and will be paid the value of the FGR portfolio associated with that transaction.

An electricity system cannot be operated on the assumption that every network element is 100 percent reliable.

A transaction fully covered by its FGR portfolio will break even in these settlement calculations—if the PDFs used to determine the LMPs and FGR prices are the same as the PDFs used to define the FGR portfolios.

The assumption that PDFs are constant on a system with given topology may be "close enough" for some purposes, although it is grossly inaccurate if specialized equipment (e.g., phase shifters) is used to control flows, a problem not considered here at all. If PDFs do not change and only three FGRs are enough on a system with three potentially congested lines, perhaps the assumption that there are few CSFs with fixed PDFs is not so bad.

D. Contingency-Constrained Dispatch of the Three-Line System

Unfortunately for the concept of a flowgate/FGR market, an electricity system cannot be operated on the assumption that every network element or critical generator is 100 percent reliable and that the full capacity of each element can be used. System dispatchers realize that there is always some chance that network elements and critical generators will fail suddenly, so they dispatch the system in such a way that it will continue operating reliably even if one (or sometimes more) of these critical elements fail—the so-called "N–1" (or "N-more") contingency criterion.

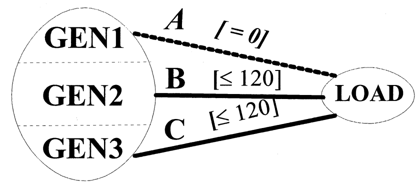
In terms of the three-line system of Figure 1, it is possible to dispatch the system so that 300 MW could flow from GEN to LOAD without violating any of the three line limits. But if anything close to 300 MW were flowing from GEN to LOAD and then one of the three lines failed, it would be impossible to shut down generation in GEN and increase generation at LOAD fast enough to prevent the flows over the remaining two lines from temporarily surging far above their individual ratings. If this were to happen, protective devices could disconnect or "trip" the remaining lines, starting a cascading failure that could shut down the entire system. To avoid this possibility, the RTO must assure that the level and location of gen-

  PII S1040-6190(00)00174-3

eration in GEN are such that a sudden failure of any of the three lines would not cause the flow on either of the remaining two lines to exceed their 100 MW limit by more than, say, 20 percent for as long as it takes for the RTO to ramp down generation in GEN and ramp up reserve generation in LOAD.[10]

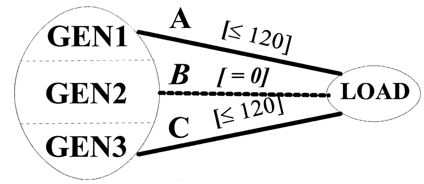
If the RTO dispatches the three-line system using an N–1 contingency criterion, the dispatch problem becomes much more complex even for the simple three-line system of Figure 1. The RTO must now consider not only the power flows on lines A, B, and C if the grid stays fully functional, but also how much power would flow on lines B and C if line A failed, on lines A and C if line B failed, and on lines A and B if line C failed. And in each of these four contingencies—counting "no-failure" as one contingency—the grid would have a different topology and hence a different set of line capacities and PDFs.

To apply the N–1 criterion, the RTO must consider each of the four contingencies and the associated capacities and PDFs as separate but simultaneous constraints on the actual dispatch. **Figure 2** shows the system the RTO would have to deal with if Contingency A, defined as failure of line A, were to occur. In Contingency A, all the power being generated in GEN would flow over lines B and C according to the PDFs indicated in Figure 2. With generation G_1 at GEN1, etc., the flow over line B prior to any contingency would be $0.3 \times G_1 + 0.5 \times G_2 + 0.3 \times G_3$ according to the three-line PDFs in Figure 1. But if Contingency A occurred, the flow over line B would immediately jump to $0.6 \times G_1 + 0.7 \times G_2 + 0.4 \times G_3$ according to the Contingency A PDFs in Figure 2. To assure that the *post*-contingency flows never violate instantaneous flow limits, the RTO must dispatch the *pre*-contingency system so that post-contingency flow limits are not exceeded given the post-contingency PDFs, i.e., so that $0.6 \times G_1 + 0.7 \times G_2 + 0.4 \times G_3 \leq 120$. A similar post-Contingency A flow constraint would apply to line C, and similar post-contingency constraints would apply in each of Contingency B and Contingency C, shown in **Figures 3** and **4**.



To LOAD from	PDF for Line		
	A	B	C
GEN1	0.0	0.6	0.4
GEN2	0.0	0.7	0.3
GEN3	0.0	0.4	0.6

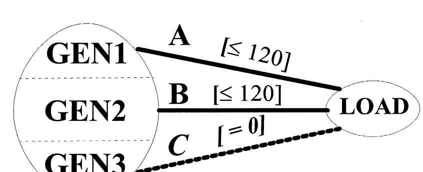
Figure 2: Contingency A



To LOAD from	PDF for Line		
	A	B	C
GEN1	0.7	0.0	0.3
GEN2	0.5	0.0	0.5
GEN3	0.3	0.0	0.7

Figure 3: Contingency B

It is critical to understand here that contingency constraints are not things that come into play only on the rare occasions when something actually fails, but that they must be taken into account and could affect the dispatch and prices any time the contingences *could* occur—i.e., all the time. The dispatcher does not remove the contingency constraints from its dispatch optimization if all the lines are still in place when the dispatch is being determined—although some contingencies may be considered important in some circumstances but not in others[11]—but leaves them there as actual constraints on the actual dispatch, even though in all probability none of the contingencies will actually occur. In the rare event that one of the contingencies does occur, the RTO adjusts the dispatch and the LMPs to reflect the actual situation, perhaps using some "post-contingency contingency constraints" to protect against a new set of contingencies, and adjusting transmission rights (or not) depending on the contractual terms in these rights.



To LOAD from	PDF for Line		
	A	B	C
GEN1	0.6	0.4	0.0
GEN2	0.3	0.7	0.0
GEN3	0.4	0.6	0.0

Figure 4: Contingency C

The RTO will determine an optimal dispatch and the associated LMPs and FGR prices by minimizing bid-based costs subject to the instantaneous flow limits and PDFs in all of the contingencies that have been judged to be worth worrying about—or "monitoring"—for operational purposes. Even with the simple three-line system being considered here, the four contingencies and three potential line constraints in each contingency mean that up to 12 (4×3) separate linear flow constraints must be monitored in the contingency-constrained dispatch (CCD) process, with up to 36 (3×12) PDFs as coefficients (although not all 36 PDFs are mathematically independent, e.g., each row in the PDF matrices must sum to 1.0 in this no-loss case). The 12 mathematical equations corresponding to the 12 constraints potentially needed for this three-constrained-line system are given in **Table 1**.

Given the specific assumptions made here about the contingencies that the RTO monitors in dispatching the three-line system, the three constraints corresponding to post-contingency flows on the failed lines ([4], [8], and [12]) reduce to "$0 \leq 0$," and the three constraints on actual operations ([1], [2], and [3]) turn out to be always less stringent than the post-contingency constraints (see **Figure 5**), making only six of the 12 constraints "interesting." But there is no way to know which of the 12 potential constraints are interesting without careful analysis of the specific contingencies even on the same system.[12] Changes in the monitored contingencies have precisely the same effect on dispatch and pricing as do changes in the physical system, and can occur even if the physical system does not change. Thus, any or all 12 constraints (or others) could come into play under different circumstances, even on a system with only three physical lines that can be congested.

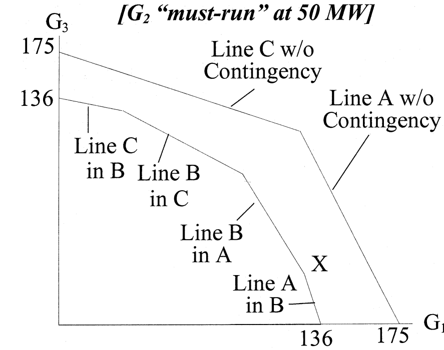


Figure 5: Example Nomograms ("Line C in B" means the flow constraint on Line C would be binding in Contingency B)

Although there can be many interesting constraints even on a simple system, only a few of them will be binding in any particular dispatch. For example, with only three dispatch variables (G_1, G_2, and G_3) in this case, no more than three of the six interesting constraints will actually be binding at any time (except by coincidence). The problem is that the set of binding constraints, and hence the PDFs used for pricing, can vary widely from one dispatch and one hour to the next, not because the PDFs associated with a given grid topology change—which might or might not be a relatively minor problem—but because a different contingency with a different topology becomes a binding constraint.

It is important to note that there is no difference anywhere in this process between the constraints on actually expected flows and the constraints on contingent flows that are not really expected to occur. The mathematical equations above make no distinction between actual and contingent constraints, suggesting that there is no logical basis for treating these differently in a flow-based market that is supposed to produce schedules or hedging portfolios that

Table 1: The Twelve Dispatch Constraints on the Three-Constrained-Line System

[1]	$0.5\times G_1 + 0.25\times G_2 + 0.2\times G_3 \leq 100$	(the Line A constraint in Actual Operations)
[2]	$0.3\times G_1 + 0.5\times G_2 + 0.3\times G_3 \leq 100$	(the Line B constraint in Actual Operations)
[3]	$0.2\times G_1 + 0.25\times G_2 + 0.5\times G_3 \leq 100$	(the Line C constraint in Actual Operations)
[4]	$0.0\times G_1 + 0.0\times G_2 + 0.0\times G_3 \leq 0$	(the Line A constraint in Contingency A)
[5]	$0.6\times G_1 + 0.7\times G_2 + 0.4\times G_3 \leq 120$	(the Line B constraint in Contingency A)
[6]	$0.4\times G_1 + 0.3\times G_2 + 0.6\times G_3 \leq 120$	(the Line C constraint in Contingency A)
[7]	$0.7\times G_1 + 0.5\times G_2 + 0.3\times G_3 \leq 120$	(the Line A constraint in Contingency B)
[8]	$0.0\times G_1 + 0.0\times G_2 + 0.0\times G_3 \leq 0$	(the Line B constraint in Contingency B)
[9]	$0.3\times G_1 + 0.5\times G_2 + 0.7\times G_3 \leq 120$	(the Line C constraint in Contingency B)
[10]	$0.6\times G_1 + 0.3\times G_2 + 0.4\times G_3 \leq 120$	(the Line A constraint in Contingency C)
[11]	$0.4\times G_1 + 0.7\times G_2 + 0.6\times G_3 \leq 120$	(the Line B constraint in Contingency C)
[12]	$0.0\times G_1 + 0.0\times G_2 + 0.0\times G_3 \leq 0$	(the Line C constraint in Contingency C)

  PII S1040-6190(00)00174-3

approximate actual operations. The logical implication is that an FGR is required for each combination of a physical network element and a contingency that is potentially binding in any dispatch.

The effects of contingency constraints on the scope and complexity of the feasible dispatches are illustrated in the two-dimensional nomogram in Figure 5. In order to get a two-dimensional nomogram, it is necessary to fix one of the three generation levels, so it is assumed in Figure 5 that G_2 in GEN2 is fixed at 50 MW. If the rest of the system is dispatched ignoring contingency constraints, the three-line system and PDFs in Figure 1 apply and any of the dispatches within the outer boundary are feasible, with only two (three, when G_2 can vary) relevant flowgate constraints. If, however, the rest of the system is dispatched subject to contingency constraints—as any real system always is—the feasible dispatches are limited by the inner boundary in Figure 5, and any of four (six, when G_2 is allowed to vary; nine, if flows on "failed" lines can be non-zero; 12, if actual flows can also be constraining[13]) different network element/contingency pairs can be binding.

Because only a few of the many possible constraints will be binding at any time, most of the FGR prices—mathematically, the constraint multipliers—will be zero most of the time and will have a high value only occasionally. The problem is that there is no practical way to predict which combination of contingent FGRs is likely to have value at any time, and it can even be difficult to explain after the fact why some particular contingencies were binding. It is not even true that only a single contingency will be binding at any time. In Figure 5, for example, the optimal dispatch will usually be at a "kink" in the nomogram, such as point X, where there are two binding constraints: Line A in Contingency B and Line B in Contingency A. It will be the rare trader—or even system operator—who will be able to predict such an outcome in advance or even provide an intuitive explanation for it after the fact.

The problem is that there is no practical way to predict which combination of contingent FGRs is likely to have value at any time.

II. Some Implications for a Flow-Based Market

The only basis for the claim that a flowgate/FGR market will result in easy and liquid trading is the assertion that there are only a few, predictable, intuitively obvious CSFs, and these will have stable, predictable capacities and PDFs. But when every possible combination of a potentially constrained network element and a dispatch contingency is treated in system operations just as though it were a separate constraint on the actual flow on a network element, the assumption of a few CSFs with fixed PDFs becomes untenable. And this is without considering nonflow constraints or the other factors that can make CSF capacities and PDFs vary.

Not every potentially constrained network element/contingency pair will be commercially significant at any time, but for operational purposes the RTO must include every element/contingency pair that has any non-negligible probability of being binding. On a system of any complexity there will be at least scores of potentially constrained network elements and at least scores of potentially important dispatch contingencies,[14] implying that hundreds of network element/contingency pairs may be at least operationally and potentially commercially significant. Furthermore, these potential CSFs are not intuitively obvious things or places, but abstract combinations of a network element and a contingency, e.g., "Transformer T when Line L is out of service." This reality has serious implications for the assertion that trading in a flowgate/FGR market will be easy and highly liquid.

A. Treating Each Element-Contingency Pair as a CSF

The mathematics of the CCD problem suggest that each combination of a potentially congested network element and a contingency should be treated as a separate flowgate with its own FGRs, capac-

  PII S1040-6190(00)00174-3

ity, and PDF matrix. In principle, a flowgate/FGR market could operate this way, at the expense of having very many potential CSFs/FGRs. In the three-line system being discussed here, there are 12 logically possible line/contingency combinations, six of which turn out to be interesting given the specific assumptions in the example.

To implement a flowgate/FGR market in this example, the RTO could define six CSFs, issue six different types of FGRs in different MW quantities, and publish the three PDFs corresponding to each type of FGR, or 18 PDFs in all. Market participants could then trade the six types of FGRs among themselves to assemble FGR portfolios that would hedge their expected transactions. For example, using the PDFs from Figures 2 through 4 above, a trader wanting to hedge a 1 MW transaction from GEN1 to LOAD would simply assemble the following portfolio of the six FGRs:

Line-B-in-Contingency-A FGRs 0.6 MW

Line-C-in-Contingency-A FGRs 0.4 MW

Line-A-in-Contingency-B FGRs 0.7 MW

Line-C-in-Contingency-B FGRs 0.3 MW

Line-A-in-Contingency-C FGRs 0.6 MW

Line-B-in-Contingency-C FGRs 0.4 MW

If six (or perhaps 12, with different assumptions) contingent FGRs are required to hedge a transaction in a simple three-constrained-line example, a system with "only" a few dozen potentially congested physical network elements could easily require scores or hundreds of contingent FGRs to hedge each transaction fully. A trader considering alternative transactions would have to compare the prices of different portfolios, each consisting of many FGRs, and then buy the entire portfolio needed to

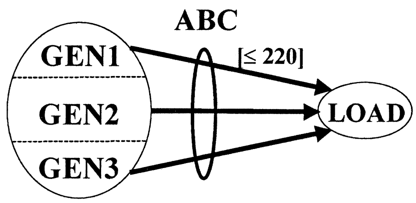
hedge the transaction it chooses. And each FGR would apply to an abstract network element/contingency pair, not some physical thing or place. This is a far cry from trading a few, simple, intuitively meaningful FGRs.

The problems involved in trading FGRs, compared to the problems involved in trading FTRs, are discussed below. But whatever the wonders of modern technology, it is likely that a market in which each individual transaction may need a specific combination of scores of the potentially hundreds of financial instruments could be complex, costly, inefficient, and not very liquid. Even proponents of flowgate/FGR markets do not usually argue otherwise. What they usually say is that there are ways to simplify trading so that this nightmare of very large numbers of FGRs would not arise. But what are the possible simplifications and how well are they likely to work?

B. Simplifying the System Assumed for FGR Trading

One way to simplify FGR trading would be to create a simplified model of the real system, use that simplified model for trading purposes, and let the RTO deal with the fact that the trading model ignores some—perhaps a lot—of reality. The simplified system model might be created by combining several network elements into compound or proxy flowgates and building into the capacity assigned to each compound flowgate some of the contingencies that would otherwise have to be considered explicitly.

For example, the three-line system in Figure 1 could be simplified for trading purposes into the one-proxy-flowgate system in **Figure 6**. The RTO could define a single flowgate, ABC, with a PDF of 1.0 from anywhere in GEN to LOAD and sell (say) 220 MW of tradable FGRs on this flowgate to account for the contingency that one of the lines may fail at any time leaving a maximum of 240 MW of instantaneous transmission capacity. Market participants could then trade these FGRs among themselves and schedule up to 220 MW of transactions from anywhere in



To LOAD from		PDF for Line ABC	
GEN		1.0	
~~GEN2~~			
~~GEN3~~			

Figure 6: A 1-Flowgate Proxy

GEN to LOAD. The RTO would then deal with the actual system and its constraints in real-time, and settle payments using some agreed set of settlement rules.

The primary problem with simplifying FGR trading in this way is that the artificially simplified market could easily result in schedules/hedges that poorly reflect reality and leave a large gap for the RTO to close somehow.[15] The actual capacity of this system to move power from GEN to LOAD can be anywhere from 171.4 MW (if any of G_1, G_2, or G_3 = 171.4 MW and the others are zero) to 240 MW (if $G_1 = G_3 = 120$ MW and $G_2 = 0$). The assumption that the proxy ABC flowgate has a capacity of 220 MW might seem a reasonable compromise between 171.4 MW and 240 MW. But if a large fraction of the 220 MW of FGRs were traded to generators in any subregion and used to schedule 220 MW of transactions to LOAD—which looks OK, given the proxy system of Figure 6—the scheduled flows would violate several of the actual constraints. The RTO would then have to either curtail some transactions even though they are covered by FGRs—i.e., call a TLR—or redispatch the system so that all 220 MW of FGRs could be honored, and recover the redispatch costs from somewhere.[16]

Under other market conditions, the 220 MW of FGRs could be used to schedule 220 MW of generation distributed across the three GEN subregions in such a way that none of the lines would be congested in any contingency. To assure efficient use of the whole system, the RTO would then have to sell the unused capacity in the spot market, implying that some transactions would not be hedged. If the RTO reduced the underscheduling/hedging problem by selling more FGRs, it would find itself dealing with infeasible schedules more often.

In short, if the RTO tries to simplify FGR trading by creating compound or proxy flowgates that do not reflect reality, it will often have to take action in real time to reduce the gap between the solution determined in the simplified market and the dispatch that is feasible and efficient given reality. If the cost of these RTO gap-closing actions is allocated to the traders whose unrealistic forward schedules caused the problem—as they would be if the RTO settled all transactions using LMPs that reflect all congestion—the FGRs are poor hedges. If the RTO socializes these costs through an uplift, prices are distorted and costs are shifted, creating short-term and long-term inefficiencies and inequities. If very many network element/contingency pairs are truly commercially significant, the RTO must choose between defining so many CSFs that trading is unworkable or so few that large costs are socialized—or perhaps some intermediate number that results in both difficult trading and inefficient prices. The dilemma created by a complex reality does not go away just because the RTO creates an artificially simplified market for trading purposes.

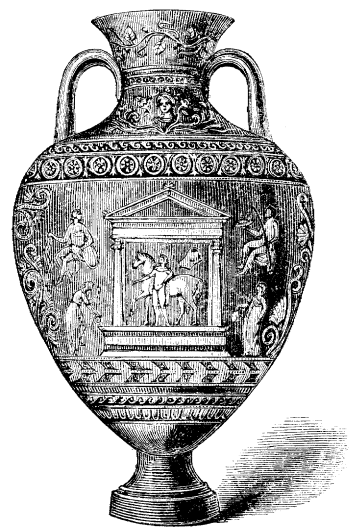
C. Forecasting/Guessing the Binding Contingencies and PDFs

Another way to try to make FGR trading easy even though reality is complex is to define noncontingent CSFs and FGRs for the physical network elements that are often constrained, and let market participants decide how to assemble a portfolio of these noncontingent FGRs that will hedge against prices determined in the CCD process. In the three-line system, for example, the RTO could issue 100 MW (or 90 MW or 80 MW) of noncontingent FGRs on each of Lines A, B, and C, perhaps inform traders about the RTO's best guess about which constraints will be binding, but provide no guarantees. A generator at (say) GEN1 wanting to hedge a transaction to LOAD could do so by buying a mix of the three available FGRs that it expects to have a value in the CCD process that approximates the LMP congestion payments that will be determined in that same process.

In principle, a trader that knows which constraints will be binding and hence which PDFs will be used for pricing can use the mathematical relationship between the LMPs and FGR prices to compute the combination of the three FGRs needed for a perfect hedge. But there would be nothing easy or intuitive about this process, which would involve selecting PDFs from different contingencies depending on which combination of constraints was expected to be binding. More importantly, the only reason to hedge against congestion prices is that it is not possible to predict them, i.e., there is no way to know which constraints will turn out to be binding. A trader smart enough to be able to predict which contingencies will be binding could make a fortune without bothering to hedge actual transactions.

Clearly, nothing is solved by defining noncontingent FGRs and telling market participants to guess which PDFs will be used for pricing. Nonetheless, this "solution" is sometimes proposed, at least implicitly, in the theoretical literature. For example, at one point in the Chao and Peck text,[17] FGRs are not indexed by contingencies and hence must be treated the same in all contingencies, even though the PDFs used for pricing and settlements will be those corresponding to the actually binding contingencies. The implication is that a trader can select an FGR portfolio based on any PDFs it chooses, but its transactions will be hedged against actual congestion only if (to the extent that) those PDFs turn out to be (similar to) the PDFs that apply in the actual dispatch, which are not known until the actual dispatch is known.

Chao and Peck do not acknowledge that this arrangement creates any commercial problems for traders; after all, the mathematics still work. But how are market participants supposed to know which PDFs to use for scheduling/hedging when the relevant PDFs will not be known until the dispatch and prices are known? Even



if market participants can guess at the hedging portfolio that works for the next hour, a different portfolio may be required the hour after that, because different contingencies may be binding in each hour; an FGR portfolio cannot provide even a medium-term term hedge if the binding dispatch contingencies can change often. This leaves market participants with only two options: (1) They can guess at a hedging portfolio and hence are unlikely to be well hedged, particularly over any meaningful time period; or (2) they can pressure the RTO to give them guaranteed PDFs and socialize the costs when different contingencies turn out to be binding.

If the RTO tries to go down the road of issuing noncontingent FGRs in an effort to make a flowgate market workable, the problem of unstable and unpredictable PDFs will quickly arise. The RTO will then come under strong pressure to tell traders how many of these FGRs they will need to hedge a given transaction—i.e., to specify the PDFs that apply to the FGRs—and then to socialize the costs of guaranteeing that anybody fully hedged based on RTO-provided information will pay no congestion charges. In fact, the pressure to guarantee PDFs and socialize costs is already arising in MISO and perhaps elsewhere, where flowgate proponents are insisting that the RTO guarantee that anybody hedged using FGRs and PDFs issued a month or so in advance will be protected against real-time congestion no matter what happens to binding contingencies, flowgate capacities, or PDFs within the month. It is not clear how or whether this can be done in any administratively feasible manner, or how it can provide meaningful long-term transmission rights, but this is clearly the direction in which MISO is being pushed now and other RTOs will be pushed soon.

D. Selecting a Few of the Very Many Contingent FGRs

Yet another way to try to deal with very many contingent CSFs and FGRs is for the RTO to issue contingent FGRs with associated PDFs for every significant network

  PII S1040-6190(00)00174-3

element/contingency pair, and then let market participants decide which of the contingent FGRs are worth buying. The problem with this “solution” is that no small number of contingent FGRs will provide a reasonable hedge for a specific transaction. Most FGRs for specific element/contingency pairs will have no value most of the time, but will occasionally be very valuable, with no practical way for anybody to predict just which contingent FGRs will have value at any time.

For example, when the optimal CCD solution is at point X in Figure 5, above, only two, apparently unrelated FGRs have any value—the Line-A-in-Contingency-B FGR and the Line-B-in-Contingency-A FGR. Relatively small changes in the market could make one or both of these FGRs worthless and make others valuable. The only way to get a reasonable hedge is to hold a portfolio that hedges against many element/contingency pairs, i.e., that contains many contingent FGRs. Holding only a few of the very many contingent FGRs cannot give a good hedge.

It is worth noting here a critical point that is discussed in more detail below: Although point-to-point FTRs are often criticized because there are potentially so many of them, any feasible transaction can be perfectly hedged with a single FTR, and many similar transactions can be approximately hedged with one or a few FTRs. The value of an FTR does not switch from zero to some large number depending on which contingency turns out to be binding in the dispatch. If there are ten similar-price nodes in region A and ten similar-price nodes in region B, there are 100 possible (two-way) point-to-point FTRs between region A and region B, but all 100 of them will have approximately the same value in every hour and hence any one of them will provide a reasonable hedge for any A-to-B transaction. It is perfectly logical and reasonable to define hundreds or

thousands of potential FTRs and let market participants select the few they need to hedge important transactions, e.g., from hub to hub. The same strategy does not solve the problem of very many contingent FGRs, because even a single hub-to-hub transaction could need scores of contingent FGRs.

Again, it should be obvious that it solves nothing to tell traders to pick the few FGRs that will have value if there is no way to predict which few will have value. Nor does this solve the problem of long-term hedging, because the few contingent FGRs that will be valuable in an hour can be different in every hour. Even so, this nonsolution is sometimes proposed in the theoretical literature advocating flow-based markets.

For example, Chao and Peck suggest in a footnote[18] that the RTO could issue many contingent FGRs, and—because only a few of them will have a positive value at any time—these FGRs can be bundled together somehow to simplify trading. It is not clear whether the suggestion is to create bundles of FGRs for each physical CSF or for each contingency, but neither of these approaches would help much. Every transaction would need a different bundle of contingent FGRs on each CSF, depending on the contingent PDFs that apply to that transaction.[19] A small number of instruments each of which is a bundle of many contingent FGRs is still a small number of instruments, and no small number of instruments can capture the full complexities of many flowgates, many contingencies, and many possible transactions.

E. Are there Few, Predictable CSFs with Stable, Predictable PDFs?

Although proponents of flow-based markets assert that there are few, predictable CSFs with stable, predictable capacities and PDFs, they have not defined how many CSFs are “few,” how CSFs would be defined, how CSF capacities and PDFs would be predicted, or what happens when the predictions turn out to be wrong—other than saying the RTO will deal with it somehow. Nor have they presented any evidence suggesting that there are few, predictable CSFs with stable PDFs. In fact, their record at predicting the amount or location of congestion is very bad. For example, one study by flowgate proponents predicted that some 28 constraints could capture most of the congestion in the Pennsylvania–New Jersey–Maryland (PJM) Interconnection, but in the first six months of LMP operations in PJM there were 43 important constraints—*not one of which* was on the list of 28 constraints predicted beforehand.[20]

The best available evidence on the number and predictability of CSFs comes from operating experience to date in PJM. As reported by Andy Ott of PJM,[21] this experience suggests that it would take more than 50 and perhaps as many as 100 contingent CSFs to capture 80–90 percent of the annual congestion costs in PJM. Perhaps more importantly, the set of element/contingency pairs that captured most of the congestion costs in one year has so far not been a very good predictor of the commercially significant pairs for the following year. The limited but rapidly expanding experience to date suggests that the number of CSFs in PJM is more than 100 and still increasing, with new element/contingency pairs becoming commercially significant all the time.

It should not be surprising that it is so hard to predict how many or which network element/contingency pairs will turn out to be important. If CSFs really could be defined in terms of physical network elements alone, without regard to contingencies, one might expect the number and identity of CSFs to settle down at some point. After all, there is only a finite—albeit, very large—number of physical elements that can be congested, and one would expect a "Top 100" list of congested physical elements to stabilize over time. But when constraints must be defined as network element/contingency pairs, there is virtually no limit to the number of pairs that can be com-

mercially significant over a period of time.

Even if the same, say, 50 physical network elements are congested every year and the same, say, 50 contingencies are monitored in the CCD process every year, changes in the pattern of load and generation will bring new element/contingency pairs into play all the time. With $50 \times 50 = 2{,}500$ possibilities, there could be a totally different Top 100 list every year for 25 years. More realistically, an unpredictable 50 CSFs could drop off the Top 100 list every year and a different unpredictable 50 CSFs could appear. There is certainly little reason to think that this year's Top-100-of-2,500 element-contingency pairs will be a very good predictor of next year's Top 100.

Perhaps recognizing that dispatch contingencies undermine the whole notion of a flow-based market, some flowgate advocates have said that CSFs should not be defined as network element/contingency pairs. For example, according to Shmuel Oren, "defining flowgates as pairs of a monitored element and a contingency element would not allow definition of unique flowgate rights and is not consistent with the theory of flowgate based congestion management."[22] This apparently means that CSFs and FGRs should be defined for specific physical elements (or compound elements) without regard to contingencies. As discussed above, CSFs and FGRs can be noncontingent—but only if their associated capacities and PDFs are contingent, which invalidates the fundamental assumption of a flow-based market: that flowgate capacities and PDFs are stable and predictable. A flowgate/FGR market with a relatively few (but probably still several score) CSFs, each with constantly changing capacities and PDFs, will either provide very poor hedges against actual congestion or require the RTO to play an active role in the market and socialize large costs—or both. There is no escape from the dilemma.

According to Oren, the theory of flow-based congestion management does not even allow CSFs to be defined in terms of element/

contingency pairs. But Chao and Peck do allow for contingent CSFs—albeit, only in a footnote, and without providing a plausible solution to the resulting problem of too many CSFs—raising questions about just what the theory of flow-based congestion management really is. As a mathematical matter, however, constraints on contingent flows enter into the CCD process in precisely the same way that constraints on scheduled flows do, so it is unclear why Oren would not treat them the same when defining FGRs—unless it is that doing so would make obvious the impracticality of the whole idea of a flow-based market.

III. FTRs and Dispatch Contingencies

The current proposals for flow-based markets appear to be motivated by a desire to find something—maybe anything—as an alternative to a market based on LMPs and FTRs, preferably something in which transmission rights are "physical" as they are said to be in natural gas. It is noteworthy that, as flowgate advocates have thought more about the problem, they have come to accept that FGRs should be financial instruments with little or no operational effect and that the RTO should use a real-time LMP process to manage and price physical operations and to settle FGRs. As they think about the problem some more, they are likely to (re)discover the advantages of FTRs—which were, after all, developed specifically to avoid the still apparently insurmountable difficulties of a market that depends on defining stable paths through the network.

A. Development of LMP/FTR Theory and Practice

The principal developer of LMP/FTR theory was William Hogan of Harvard.[23] Building on the work of Fred Schweppe and his colleagues at MIT,[24] the LMP/FTR approach exploits the concept

that a (approximately) consistent set of LMPs could be derived from an (approximately) optimal operational dispatch.[25] This was an important variation on the Schweppe idea that LMPs can and should determine actual operations, because it allowed (approximately) efficient prices to be determined from a dispatch that was influenced by the judgment of human operators.

The complexity of electricity systems goes beyond the contingency issues discussed in this article, to include the changing patterns of power flows and transmission limits that arise from voltage restrictions, the inherently nonlinear character of the network, and the increasing use of equipment—i.e., phase shifters—specifically designed to change PDFs. As a practical matter, it is essentially impossible to define long-term transmission rights that depend on the pattern of flows through the grid—as evidenced by the fact that the industry has been searching for such flow-based methods for years without success.

To avoid the many problems of a flow-based approach to defining transmission rights, Hogan developed the concept of point-to-point FTRs, which are rights to the LMP differentials between two defined points. He noticed that if system users are given FTRs corresponding to a specific dispatch, both individual users and the settlement system as a whole are largely unaffected financially even by large changes in the pattern of injections and withdrawals, even when these changes resulted in large changes in flows, PDFs, and LMPs. These observations motivated the development of the theory of LMPs and FTRs.

LMP/FTR theory was explicitly designed to avoid the difficulties associated with a flow-based approach by basing pricing and transmission rights on the outcome of the actual dispatch rather than on specific network flows and PDFs that are essentially unpredictable over any commercially interesting period of time. The specific LMPs and LMP differentials that define congestion charges depend on which of the many constraints are actually binding in the dispatch, and hence are volatile and difficult to predict—although not nearly as much so as the individual constraint multipliers/FGR prices. But if a market participant holds point-to-point FTRs that exactly cover its point-to-point transactions, it is perfectly hedged against congestion charges no matter what happens to LMPs. This allows a market participant to hedge a transaction with a single, easy-to-understand financial instrument without worrying about flows, dispatch constraints, contingencies, PDFs, constraint multipliers/FGR prices, or any of the other arcane network details about which market participants have little knowledge and over which they have virtually no control.

Furthermore—and this is the real key to FTRs—as long as the pattern of injections and withdrawals implied by the set of FTRs issued by the RTO would be simultaneously feasible on the grid,[26] the LMP-based congestion charges associated with *any* optimal dispatch on that grid will be at least enough to cover all the FTR payments the RTO must make given that dispatch and LMPs, no matter which of the many possible actual or contingent flow constraints (or nonflow constraints, or phase-shifter settings, or . . .) apply in the new dispatch. Thus, the RTO can issue a set of FTRs that fully hedges any simultaneously feasible set of transactions without fear that the RTO's settlement system will run a deficit if the market prefers a different set of transactions that creates a different set of binding constraints, contingencies, or operating parameters. Market participants can get good hedges against any feasible set of transactions—with only one FTR required to hedge one transaction—with no risk that the RTO will have to socialize large costs when market conditions and binding dispatch contingencies change.[27]

This is a remarkable result that greatly simplifies both the opera-

tional and the commercial problems of market-based congestion management. Neither LMP/FTR theory nor its application is easy in any absolute sense or even relative to what is required in most other markets. But that is because electricity itself is not easy. LMP/FTR theory and practice are easy only relative to the logical alternatives for market-based congestion management on a complex electricity system. In particular, an LMP/FTR market is easy relative to any flow-based market that makes a serious effort to price all congestion in the market so that market participants will have good price signals and incentives to seek cost-effective trade-offs between congestion costs and other costs. The only way to make a flowgate/FGR market anywhere near as easy as an LMP/FTR market is to get the RTO to define an artificially simplified set of constraints for trading purposes and then to manage and socialize the costs of the real congestion outside the market.

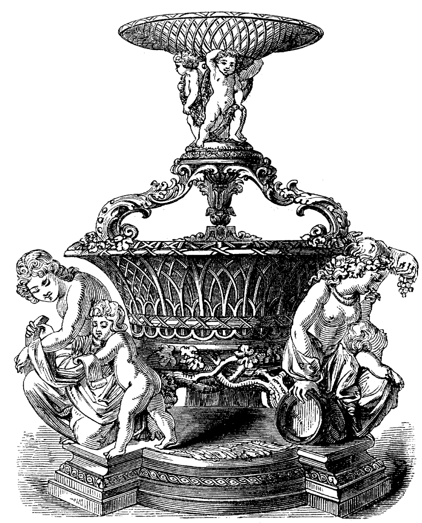
B. The Problem of Very Many FTRs

The efforts to develop a flow-based alternative to an LMP/FTR market are motivated largely by the belief that there are too many potential point-to-point FTRs to make FTR trading easy and liquid. Flowgate/FGR advocates like to say, for example, that with 100 nodes there are 100^2 or 10,000 possible point-to-point FTRs.[28] They do not say that a flowgate/FGR market with only 20 potentially congested physical network elements and only 20 contingencies could need 20^2 or 400 contingent CSFs/FGRs, and even if only 50 of these contingent CSFs/FGRs are important and any point-to-point transaction can be reasonably hedged with FGRs on only 10 of these, there are more than 10 *billion* possibly interesting point-to-point FGR portfolios—even without considering variations in the quantity of each type of FGR.[29] Such numbers games prove nothing, but if FGR proponents want to play them they should be prepared to lose.

The essential characteristic of real transactions on the grid is that they are point-to-point: Power is injected somewhere for delivery somewhere else. The pattern of flows that results from a specific injection and withdrawal depends on the laws of physics—and, when such things as phase-shifters are considered, the actions of system controllers—not the preferences or actions of the traders involved in the transaction. To hedge a specific point-to-point transaction against all possible combinations of flows it could "cause" under all contingencies that can affect dispatch could require a portfolio consisting of a very large number of contingent FGRs.

The number of FGR portfolios actually needed to hedge all possible transactions is, of course, limited by the number of possible point-to-point transactions. On a system with 100 nodes there can be no more than $100^2 = 10{,}000$ interesting portfolios of (one-way) FGRs—which is precisely the same as the number of possible (one-way) point-to-point FTRs. Not all of these possible FGR portfolios would be commercially interesting, because there will not be any transactions between most point-to-point pairs, many of the FGR portfolios could serve as reasonable-if-not-perfect hedges for all transactions from one region to another region, etc. But all of this is just as true with point-to-point FTRs as it is with point-to-point FGR portfolios. There is no difference.

Some advocates of flow-based markets respond to such observations by saying that a flowgate/FGR market will trade individual FGRs, not point-to-point FGR portfolios that are analogous to FTRs. But this brings into play the other horn of the dilemma facing a flowgate/FGR market: How efficient or liquid will a market be if every point-to-point deal needs a specific combination of many of the very many contingent FGRs? Some flowgate/FGR proponents say it will be no problem trading dozens or maybe even hundreds of individual FGRs using modern



technology, and even demonstrate prototype software that allows a trader to specify a point-to-point transaction, get an instant quote on the price of the entire portfolio of FGRs necessary to hedge this transaction, and then buy that portfolio with the push of a button or click of a mouse. But such electronic gadgetry begs the commercial question of who will be holding and pricing hundreds of different FGRs so that there is somebody on the other end of the mouse-click.

It is highly unlikely that a flowgate/FGR market would function as some sort of electronic FGR bazaar in which many, competitive FGR mongers would be making markets in individual FGRs by holding inventories and posting prices, particularly if a use-it-or-lose-it rule were in place making it very risky to hold FGR inventories.[30] So perhaps there would be a centralized, sophisticated, multi-round auction process similar to electromagnetic spectrum auctions, as suggested by Chao *et al.* [31]—although this seems unlikely, given that FGRs would have to be traded continuously whereas spectrum auctions are one-off, multi-day affairs.

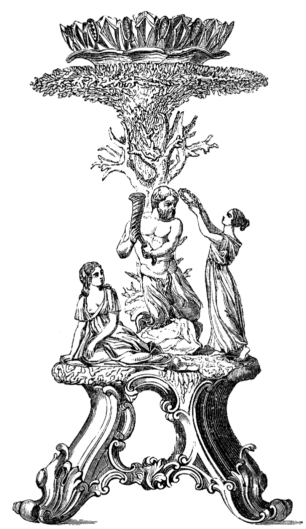
As these examples show, flowgate/FGR proponents cannot decide whether FGR trading, which is supposed to be easy, cheap, and highly liquid, will be a highly decentralized, simple, over-the-counter entrepreneurial affair or a highly sophisticated, centralized, probably monopolized (and almost surely Securities and Exchange Commission-and/or Federal Energy Regulatory Commission-regulated) activity. The most likely outcome is that FGR trading would evolve to include one or a few centralized exchanges trading a few standard FGR portfolios that hedge point-to-point transactions between important hubs, with specialized brokers assembling FGR portfolios for transactions between the major hubs and nearby locations. In other words, FGRs would be traded in portfolios that are awkward and incomplete versions of FTRs.

If FGR trading is likely to evolve into a difficult form of FTR trading, it would seem preferable to skip the FGR step entirely and go straight to FTRs. Conceptually and mathematically, FTRs are portfolios of all the FGRs needed for a perfect point-to-point hedge against all congestion on all network elements in all contingencies considered for dispatch and pricing—plus hedges against all congestion arising from nonflow constraints and against PDF changes due to nonlinear effects and operator actions such as changing phase-shifter settings. There are very many possible point-to-point FTRs, just as there are very many possible point-to-point FGR portfolios. But any point-to-point transaction can be perfectly hedged with a single FTR, and many similar transactions can be approximately hedged with one or a few FTRs.

It is far easier and more natural to work with FTRs than with FGRs. The operators of LMP/FTR systems and the traders using such systems still have a lot to learn about how to use FTRs to accomplish various legitimate (and perhaps illegitimate) commercial objectives. But there is no reason to think that this learning process would be any easier or the ultimate results would be any better if point-to-point FTRs were replaced with hundreds of contingent FGRs that must be assembled into thousands of point-to-point portfolios to work as well.

C. Short-Term Hedging and Trading

Much of the current criticism of LMP/FTR markets comes from market participants who want to use such markets for short-term trading. Such traders say that LMP/FGR systems such as PJM make it difficult to get the FTRs they need to hedge transactions that change from day-to-day or even hour-to-hour. A market based on flowgates and FGRs, although not even defined in concept much less demonstrated in practice, is held out as a solution to these



problems. But it is unreasonable to blame LMP/FTR markets generically for the implementation problems on specific systems, and unrealistic to think the same problems would not arise for the same reasons in a flowgate/FGR market.

The fact that traders naturally want a system that allows near-perfect short-term hedging does not mean that such a system should or could exist. In any real commodity market there are many logically possible hedging instruments that could exist but do not, presumably because they are not worth what they would cost to develop and use. The fact that some financial instrument or market process might lower costs and risks for somebody if only somebody else would pay for it is no proof that such an instrument or market should exist. Certainly there is a lot of room for improvement in short-term trading and hedging arrangements in electricity markets. But one should be realistic about how rapidly these can develop and how far they can go.

The claims that there is little liquidity in functioning LMP/FTR markets such as PJM are at least overstated, if not flatly wrong, given that PJM appears by most measures to have the most liquid electricity and transmission markets in the US.[32] There may currently be little short-term trading of the valuable FTRs from western PJM to eastern PJM, but that is because there is limited capacity for west-to-east transfers and the FTRs on that capacity are held by entities—the utilities with load-serving obligations—who need those FTRs to hedge their own transactions and/or have regulatory disincentives to trade them.[33] These specific transitional and implementation problems have nothing to do with the nature of FTRs themselves and would not be solved by replacing FTRs with FGRs.

Whatever the reality or cause of short-term trading problems with FTRs in functioning LMP/FTR markets, there is no logic or evidence suggesting that short-term trading would be any better in a flowgate/LMP market. The only basis for thinking that FGR trading would be easy and liquid is the belief that there would be few FGRs with stable capacities and PDFs, a belief that has no foundation for many reasons, including the CCD process considered here. As the RTOs (independent system operators) in PJM, New York, and New England continue developing short-term FTR auctions—including the FTR auction implicit in the day-ahead energy markets—as traders learn more about FTRs, and as the transitional disincentives to trading fade away, short-term FTR trading can be expected to become easier and more liquid in those markets. It may never be easy and liquid enough to satisfy day-traders; but it should serve the needs of the generators and consumers who are the intended beneficiaries of competitive restructuring.

D. Long-Term Hedging and Transmission Investments

One of the principal objectives of any transmission rights regime is to reduce long-term congestion cost risks and to encourage long-term investment in generation and transmission. This objective is underemphasized in current policy discussions, which are dominated by the concerns of short-term traders. But providing long-term transmission rights is a critical objective of any electricity market—and one for which FTRs are ideally suited and FGRs are virtually worthless.

To see this, consider first a generator G at GEN who holds 100 MW of a single long-term FTR from GEN to LOAD. This FTR, along with real-time settlements based on LMP, creates the following situation:

- G has 100 MW of transmission service from GEN to LOAD at no cost (beyond the cost of the FTR itself) for the term of the FTR and—as long as all outstanding RTO-issued FTRs are simultaneously feasible—without the RTO socializing any costs except perhaps those due to unexpected changes in the physical network. If

market conditions change so that the pattern of LMPs changes—implying that congestion is now appearing on different network elements and/or in different contingencies and/or due to nonflow constraints—G continues to get its 100 MW of transmission service at no cost automatically.

- If the grid is expanded, new FTRs can be sold or allocated to other users without affecting G's rights and without requiring the RTO to socialize any costs, as long as both existing and new FTRs are simultaneously feasible on the new grid. If part of the grid is removed from service or a new contingency is added to those considered in the CCD process,[34] the RTO or grid owner may have to buy back some FTRs to make the remaining FTRs simultaneously feasible on the reduced grid; but G does not have to sell any of its FTRs, and if it does not sell any will automatically continue getting its 100 MW from GEN to LOAD on the reduced grid.

- If G wants long-term rights for another (say) 50 MW from GEN to LOAD, G can pay for one or several grid expansion projects (whether or not they are on CSFs) that relieve congestion (in all contingencies) enough to support an additional 50 MW of GEN-to-LOAD transactions simultaneously with the transactions implied by the existing FTRs. GEN can then be given 50 MW of additional GEN-to-LOAD FTRs and these will fully hedge an additional 50 MW of transactions in all contingencies.

Now consider this same generator G in a flowgate/FGR market in which the RTO defines CSFs and FGRs for only a subset of the possibly congested network element/contingency pairs in order to try to make FGR trading workable, but does not socialize the costs of guaranteeing the capacities and PDFs for the entire term of the FGRs. If G holds the portfolio of FGRs that perfectly hedges a 100 MW transaction from GEN to LOAD given the CSF capacities and PDFs that exist today, the situation is as follows:

- G has 100 MW of transmission service from GEN to LOAD at no cost (beyond the cost of the FGR portfolio itself)—except when non-CSF congestion arises or PDFs change, which could happen every hour. If long-term changes in the market create persistent congestion on different network element/contingency pairs, new contingent CSFs will have to be defined with their associated capacities and PDFs, or new capacities and PDFs will have to be defined for the existing noncontingent CSFs. Either way, G must get a new portfolio of FGRs to hedge the same GEN-to-LOAD transaction, implying that G's original FGR portfolio was not worth much as a long-term transmission right.

• If the grid is either expanded or reduced physically, some network element/contingency pairs may be dropped as CSFs while others may be designated as new CSFs. And the new grid configuration will change most of the PDFs. So G will need a different set of FGRs to hedge the same 100 MW GEN-to-LOAD transaction. It may be desirable to hold G harmless against changes in the grid that G cannot predict or control, but this would require redefining and reallocating FGRs whenever a change in the grid changes CSFs, capacities, contingencies, and PDFs.

• If G wants long-term rights for another (say) 50 MW from GEN to LOAD, G can pay for one or several grid expansion projects and, as long as they are on CSFs, can be given FGRs on those CSFs corresponding to the increase in capacity in each contingency (if FGRs are contingency-specific). But in order to hedge an additional 50 MW of GEN-to-LOAD transactions G will still need a portfolio of FGRs on many CSFs, with the required portfolio changing as CSFs and PDFs change over time. The RTO will not be able to assure G that any set of grid expansions will provide incremental FGRs that will hedge an additional 50 MW of GEN-to-LOAD transactions as different contingencies become binding.

The critical point is that a

flowgate/FGR system must, as a practical matter, deal with only a subset of the potentially congested network element/contingency pairs and hence cannot provide reliable long-term transmission rights when binding flow (or non-flow) constraints—or other things, such as use of phase-shifters—change. Even if the RTO is prepared to guarantee long-term FGRs and socialize large costs over their terms, it is not clear how this could be done given the difficulty of continually defining new and reallocating old FGRs. Unless contingent CSFs and FGRs are defined for every one of the potentially thousands of network element/contingency pairs that have any chance of becoming important over the term of the FGRs, and these FGRs are made contingent on other factors that can affect CSF capacities and PDFs, the portfolio of FGRs needed to hedge the same transactions must be expected to change over time, perhaps often and significantly.

These difficulties would make it troublesome to provide reliable grandfathering for existing long-term rights, to sell new long-term rights, or to grant long-term rights to those who pay for grid expansions. Some flowgate/FGR proponents have recognized this problem and have begun trying to think of a solution. In one preliminary proposal, generator G above would assemble a portfolio of long-term FGRs that hedge 100 MW from GEN to LOAD given the CSFs and PDFs that exist today, and register that point-to-point transaction with the RTO. Then, whenever new market conditions forced the RTO to add CSFs or change PDFs, the RTO would exchange that FGR portfolio for the new one required to hedge the registered transaction given the new CSFs and PDFs.

Such proposals raise more questions than they answer. What are the "long-term FGRs" that would be assembled into point-to-point portfolios and then registered to get long-term point-to-point rights? How does the RTO guarantee that all the registered point-to-point transactions will be feasible in combination with the still-outstanding long-term FGRs

  PII S1040-6190(00)00174-3

when new CSFs and/or PDFs must be defined? What about non-CSF congestion or changes in capacities and PDFs in between the times when CSFs, capacities and PDFs are redefined? What will be the process for defining new CSFs and PDFs, when every market participant will be affected by such changes? Or does the RTO guarantee the long-term FGRs and registered point-to-point transactions by buying back long-term rights that become infeasible as new CSFs and/or associated PDFs emerge over time and socializing the costs? Perhaps the most difficult question raised by such proposals is: Why is there so much effort to find unnatural and complex ways to do with FGRs what is natural and relatively easy to do with FTRs?

The virtual impossibility of providing even short-term price certainty with FGRs is driving flowgate/FGR advocates in, e.g., the MISO discussions, to propose that the RTO guarantee FGRs and PDFs for periods as long—or as short, depending on how one looks at it—as a month, and socialize all the costs of making good on that guarantee. It is not clear how or whether this could be administratively feasible, but it is reasonably clear that a guarantee of FTRs and PDFs for even a month is likely to create more cost socialization than RTOs (or public utility commissions or FERC) are likely to find acceptable. If a flowgate/FGR system cannot guarantee FGRs and PDFs for even a month without a lot of administrative complexity and cost socialization, how will it ever provide the kind of long-term transmission rights needed in any reasonably efficient and effective market?

IV. Conclusions

The basic assumption of a flow-based market is that there are only a few CSFs and these have fixed capacities and PDFs. There are many reasons to doubt this

assumption, but one of the most important is that modern electricity systems are operated subject to constraints on the flows that would arise under multiple contingencies, not just constraints on actual flows on physical elements. The constraints on contingent flows have precisely the same effects on operations, congestion, and pricing as the constraints on actual flows. If forward FGR trading is to produce schedules or hedges that are close to operational reality, there must be a different FGR for every network element/contingency pair that has any significant chance of having significant value over the term of the FGR. This implies that scores of contingent FGRs could be required for even an approximate hedge over a "long-term" such as a month, with perhaps hundreds required for multi-year transmission rights.

The need for very many contingent FGRs (or fewer noncontingent FGRs but with contingent capacities and PDFs) is not some trick invented by the evil advocates of FTRs, but is a straightforward logical implication of contingency-constrained dispatch on a complex system. In fact, FTRs were developed largely to deal with the implications of this and other complex realities of electricity systems. The point-to-point nature of FTRs, which mirrors the point-to-point reality of electricity transactions, allows a single FTR to hedge a transaction perfectly against congestion due to flow limits on every network element in every contingency, nonflow constraints, and changes in PDFs due to nonlinear physical relationships and operator actions. Requiring market participants to use FGRs to try to assemble their own hedges against so many things that they cannot predict, understand, or control would be unnatural, ineffective, and pointless—although it might benefit those who stand to gain from high transactions costs and market inefficiency.

The proponents of flow-based markets have not explicitly acknowledged their dilemma of either very many CSFs or ever-changing CSF capacities and PDFs, but they must realize the dilemma is there. If they really believed that a flowgate/FGR market could capture all commercially significant congestion with a few CSFs with fixed capacities and PDFs, they would not object to requiring traders to pay LMP-based congestion charges on all actual congestion not hedged by a few FGRs with fixed capacities and PDFs, because the unhedged congestion costs would not be commercially significant. The fact that proponents of flowgate/FGR markets are so adamant that the RTO must guarantee a few CSFs with fixed capacities and PDFs and socialize the costs of delivering on that guarantee strongly suggests that they know these costs would not be commercially insignificant and would probably be very large. In this, at least, they are almost surely correct. ■

Endnotes:

1. The term "security-constrained dispatch" (SCD) is more commonly applied to this process, but the term CCD is used here to emphasize the role of dispatch contingencies.

2. Hung-Po Chao and Stephen Peck, *A Market Mechanism for Electric Power Transmission*, 10 J. REG. ECON., 1996, at 25–59. As discussed below, Chao and Peck propose two, mutually inconsistent ways to deal with CCD in a flowgate/FGR market, neither of which resolve the dilemma.

3. Shmuel Oren in an e-mail to Andrew Ott of the Pennsylvania–New Jersey–Maryland Interconnection (and a large cc list), Sept. 20, 2000. See discussion below.

4. The dispatch is constrained by contingencies even when the grid is in its standard condition, i.e., when none of the contingencies has actually occurred. The issue of who should bear the costs when the grid is not in its standard condition is an important issue but is not considered here.

5. In principle, each of the three line constraints implied by Figure 1 could be interpreted as either an actual or a contingent constraint. But virtually nobody draws such a simple diagram and then says that each constrained line in that diagram actually represents an abstract "element/contingency pair." A "line" almost always represents a physical network element.

6. Whether FGRs are needed for scheduling or only to hedge transactions depends on whether an FGR is regarded as a "physical" right or a "financial"

right. Although FGRs are still sometimes called physical rights, it is now generally accepted that FGRs should not be required for scheduling, but are primarily financial hedges. It is assumed here that FGRs are financial instruments but that market participants seek FGR portfolios that reflect their expected physical operations.

7. Even if market participants must submit balanced schedules covered by FGRs, the RTO will have to use a full LMP process to determine the incremental dispatch and the associated prices. Energy traded under balanced schedules covered by FGRs will not be priced or settled in this process, but will affect actual transmission congestion and prices and hence must be an input to the LMP process.

8. Hung-Po Chao, Stephen Peck, Shmuel Oren, and Robert Wilson, *Flow-Based Transmission Rights and Congestion Management*, ELEC. J., Oct. 2000, at 38.

9. Mathematically, if P_f is the price of flowgate f and PDF_{gf} is the PDF from location g to the "hub" H over flowgate f, the LMP at location g is $LMP_g = LMP_H + \Sigma P_f \times PDF_{gf}$, where the sum is over all flowgates f. The congestion charge from any location g to any other location k is then $LMP_k - LMP_g = \Sigma P_f \times (PDF_{kf} - PDF_{gf})$. A 1 MW transaction from g to h that is covered by $(PDF_{kf} - PDF_{gf})$ FGRs on all flowgates f will be paid for its FGRs exactly what it pays in congestion charges and hence will be perfectly hedged. Approximate coverage on flowgates with "significant" prices will result in an approximate hedge.

10. This is not just a hypothetical situation, but arises all the time on real systems. For example, operation of the New York system is often constrained by the risk that one of several cables into Manhattan and Long Island might fail and the flows over the remaining cables would surge above line limits.

11. For example, the RTO may impose more stringent—e.g., N–2—contingency constraints when thunderstorms are in the area, on the grounds that lightning strikes may disable several transformers simultaneously. Who bears the costs when such changes in dispatch contingencies or actual grid conditions reduce grid capacity is an important issue that is essentially the same with either FGRs or FTRs, and hence is not discussed here.

12. For example, in this case if a line could fail partially it would be necessary to consider post-contingency flows on the failed line in each contingency; and larger post-contingency instantaneous flow limits on failed lines—e.g., 150 percent of steady-state limits—could make the constraints on actual flows binding in some dispatches.

13. If the limits on instantaneous, post-contingency flows were higher, the inner boundary in Figure 5 would expand outward, and at some point the two boundaries would overlap.

14. PJM dispatchers reportedly monitor approximately 50 contingencies.

15. A secondary problem is that it is very difficult even to define a set of settlement rules that make sense here. The difficulty of defining a sensible settlement process is creating difficulties for the efforts to develop a workable hybrid at MISO and perhaps elsewhere, but is not discussed here.

16. In this simple case, redispatch would require paying some scheduled/hedged generation in GEN not to run and paying some unscheduled/unhedged generation at LOAD to run out-of-merit.

17. *Supra* note 2, at 44–46. As discussed below, Chao and Peck suggest a different nonsolution to the problem of too many FGRs in a footnote within their article.

18. *Supra* note 2, at note 18. As discussed above, Chao and Peck suggest a different nonsolution in the text of their article.

19. There is no FGR portfolio that perfectly hedges congestion on flowgate A in every contingency for all transactions, because different PDFs will apply to each different point-to-point transaction. For the same reason, there is no FGR portfolio that perfectly hedges congestion on all flowgates in contingency X for all transactions. In principle it is possible to create an FGR portfolio that perfectly hedges against congestion on all flowgates in all contingencies for a specified point-to-point transaction. These are called FTRs.

20. William Hogan, *Flowgate Rights and Wrongs*, working paper, John F. Kennedy School of Government, Harvard University, Aug. 20, 2000, at 20.

21. Andrew L. Ott, *Can Flowgates Really Work? An Analysis of Transmission Congestion in the PJM Market from April 1, 1998–April 30, 2000*, PJM Interconnection, Sept. 15, 2000.

22. *Supra* note 3.

23. William W. Hogan, *Contract Networks for Electric Power Transmission*, 4 J. REG. ECON., 1992, at 211–42. Grant Read of Canterbury University in New Zealand was an early collaborator with Hogan. Read spent some time in the late 1980s and early 1990s developing a link-based approach to transmission pricing and congestion management, but says that he gave it up as hopeless (personal communication to author).

24. F.C. SCHWEPPE, M.C. CARAMANIS, R.D. TABORS, AND R.E. BOHN, SPOT PRICING OF ELECTRICITY (Norwell, MA: Kluwer, 1988).

25. William W. Hogan, E. Grant Read, and Brendan J. Ring, *Using Mathematical Programming for Electricity Spot Pricing*, in ENERGY MODELS FOR POLICY AND PLANNING, International Transactions of Operational Research, Vol. 3, No. 3/4, 1996.

26. A dispatch or set of FTRs is said to be simultaneously feasible if it satisfies all of the actual and contingent constraints used for dispatch. To be explicit about what should be obvious but is sometimes confusing, a simultaneously feasible dispatch does not have to be feasible even if all contingencies were to occur simultaneously. No dispatch could meet these conditions.

27. Given the nonconvex nature of electricity systems—economies of scale, integer choices, etc.—such sweeping assertions require some technical qualifications, and it is possible to create counterexamples using unrealistic numerical examples. But the important assumptions of LMP/FTR theory appear to be valid over wide ranges of actual conditions on real systems, in sharp contrast to the assumptions of flowgate/FGR theory, which are virtually never correct (e.g., PDFs are simply not constant).

28. Actually, this is true only if an FTR is a one-way option, so that an A-to-B FTR and a B-to-A FTR are different. With N nodes, there are $N\times(N-1)/2$ different point-to-point pairs if A-to-B is the same as B-to-A. Thus, with 100 nodes there could be "only" 4,950 bi-directional FTRs.

29. If a portfolio of n FGRs can include any of $N > n$ different FGRs, there can be $N!/[n!\times(N-n)!]$ qualitatively different portfolios. If $N = 50$ and $n = 10$, the resulting number is 10,272,278,170.

30. Under a use-it-or-lose-it rule, anybody holding an inventory of FGRs must find a buyer for all of its FGRs or be left with nothing of value in real time. But buyers could always wait to buy transmission in the spot market by paying real-time congestion charges. It is hard to imagine how anybody could afford to make a market in such FGRs. An FGR bazaar might be more plausible without a use-it-or-lose-it rule, but even so a small market maker would be at a tremendous disadvantage. It seems inevitable that FGR trading would become a centralized, monopoly activity.

31. *Supra* note 8.

32. See *PJM Market News*, Vol. 1, No. 1, Oct. 2000, for information on the volume of energy and FTRs traded in PJM markets.

33. If something is valuable to its current owners one would not expect it to be offered for sale unless the prospective buyers are willing to offer high prices for it. And if the current owners fear that they will not be allowed to keep any trading profits but will be stuck with any trading losses—a dilemma commonly created by regulation—they will be reluctant to sell even when they "should" according to normal commercial and economic criteria.

34. For example, the regional reliability process may decide that the RTO should replace an N–1 standard with an N–2 standard on some elements. This would reduce the usable capacity of the grid just as a grid outage would.

