



Final Report

A Transitional Non-LMP Market for California: Issues and Recommendations

PREPARED FOR

California ISO – CAISO

151 Blue Ravine Road

Folsom, CA 95630

PREPARED BY

Charles River Associates

5335 Collete Avenue

Oakland, CA 94618-2804

September 16, 2004

Table of Contents

1. INTRODUCTION.....	1
1.1 Context and Purpose	1
1.2 Summary and Conclusions.....	2
• Any UMP Market Is No Better Than “Second Best”	3
• “Successful” UMP Markets Are Uncongested and Simple	3
• The Logic of a UMP Market Requires Compensation – and More.....	3
• The Second-Best UMP Depends on Compensation Policies	3
• A UMP Market Without CDPs Needs Scheduling Priorities	4
• Operational Complexity Implies UMP-Market Complexity.....	4
• A Two-Settlement UMP Market Requires Special Settlement Rules.....	4
2. OVERVIEW OF ALTERNATIVES TO LMP	5
2.1 Some Theoretical Considerations.....	5
2.1.1 Competitive Market Clearing and Constrained Optimization.....	5
2.1.2 “First-Best” and “Second-Best” Solutions	7
2.1.3 LMPs As the First-Best Prices	7
2.1.4 UMPs and Second-Best Solutions.....	8
2.2 Real-World Experience with UMP Markets	9
2.2.1 The Chilean Model.....	9
2.2.2 England and Wales.....	10
2.2.3 The PJM UMP Market.....	12
2.2.4 The NEPOOL/ISO-NE UMP Market.....	12
2.2.5 Australia	13
2.2.6 Alberta.....	15
2.2.7 Ontario	15
2.2.8 ERCOT	18
2.2.9 California.....	19
2.2.10 Summary of Experience with UMP Markets.....	20
• Zonal UMP Markets “Work” When Intrazonal Congestion Is Small	20
• UMP Markets Include CUPs “Always” and CDPs “Almost Always”	20

Table of Contents (Continued)

- UMP Markets Use Locational Grid Access Charges.....20
- UMP Market ISOs Do Not Seek Global Efficiency20
- UMP Markets Do Not Use Two-Settlement Systems20

3. SUMMARY OF THE TAPAS UMP APPROACH..... 22

- 3.1 Principal Elements of the TAPAS UMP Market..... 22
 - 3.1.1 MRTU Timetables and Procedures.....22
 - Pre-IFM Reliability and Market Power Mitigation Runs22
 - Day-Ahead (DA) Integrated Forward Markets (IFMs)22
 - DA Residual Unit Commitment (RUC)22
 - Simplified Hour-Ahead (HA) Market23
 - Pre-Dispatch and Real-Time (RT) Dispatch.....23
 - 3.1.2 MRTU Scheduling/Dispatch Processes.....23
 - 3.1.3 Settlement Prices for Energy (UMPs) and for A/S.....23
 - Options for Zonal Energy UMPs.....23
 - Options for A/S Settlement Prices.....24
 - 3.1.4 UMP-Market Schedules and Operating Profits24
 - 3.1.5 Compensation for Constrained-Up Resources24
 - 3.1.6 Uplift Paid by Scheduling Coordinators.....25
 - 3.1.7 Congestion Charges and Congestion Revenue Rights (CRRs)25
 - 3.1.8 Principal Design Issues Raised by the TAPAS Approach.....25
- 3.2 The TAPAS Proposal Compared to Other UMP Markets 25
 - 3.2.1 The ISO in Other UMP Markets25
 - 3.2.2 The CAISO in the TAPAS Approach.....26

4. A TAPAS-TYPE UMP MARKET IN THEORY 28

- 4.1 ISO Compensation Payments..... 28
 - 4.1.1 CUPs and CDPs as Second-Best Price Corrections28
 - 4.1.2 UMP-Market Schedules and CUPs/CDPs29
 - 4.1.3 The Effects of Not Making CDPs.....32
 - 4.1.4 CUPs, CDPs and UMP-Market Transmission Rights.....33

Table of Contents (Continued)

4.1.5	CDP Access Right/Charges.....	34
4.1.6	Long-Run Effects of a UMP Market	36
4.2	Determining a Second-Best UMP	37
4.2.1	Choosing UMP to Minimize Short-Run Inefficiencies.....	37
4.2.2	UMP Option 1: UMP = UMCP.....	40
4.2.3	UMP Option 2: UMP = Weighted Average of LMPs.....	42
4.2.4	Recommendations for the Second-Best UMP	43
4.3	Interzonal Congestion Charges and CRRs.....	44
4.4	The Effects of Intertemporal Constraints	45
4.4.1	A “Pure Market” Approach To Intertemporal Constraints	45
4.4.2	Intertemporal Optimization (IO) by the ISO	46
4.4.3	Multi-Step IO and Unit Commitment	48
4.5	Ancillary Service (A/S) Schedules and Payments	50
4.6	Issues in a Two-Settlement UMP Market	52
4.7	Other Issues in a TAPAS-Type UMP Market	54
4.7.1	Recovering Congestion Costs	54
4.7.2	Market Power Mitigation.....	55
5.	OPERATION OF THE TAPAS UMP MARKET.....	57
5.1	The Day-Ahead Processes	57
5.1.1	The Pre-IFM Step.....	57
5.1.2	IFM Unit Commitment.....	58
5.1.3	IFM Scheduling and Pricing	58
5.1.4	IFM Settlements and CUPs/CDPs	59
5.1.5	DA Residual Unit Commitment (RUC)	60
5.1.6	DA Settlement	60
5.2	Simplified Hour-Ahead “Market”	60
5.3	The Real-Time Processes.....	61
5.3.1	Real-Time Dispatch and Pricing	61
5.3.2	Real-Time Settlements	61

Appendix A: Second-Best UMPs when Compensation Is Not Paid

Table of Contents (Continued)

Appendix B: CUPs and CDPs with Intertemporal Optimization

A Transitional Non-LMP Market for California: Issues and Recommendations

Charles River Associates

September 16, 2004

1. INTRODUCTION

1.1 CONTEXT AND PURPOSE

The California Independent System Operator (CAISO), has been developing a new electricity market for California that would use locational marginal pricing (LMP) to manage and price transmission congestion, with congestion revenue rights (CRRs) to hedge congestion price risks. Implementation of this market may, however, be delayed by concern about the potential cost increase to buyers under “seller’s choice” contracts.¹ To prepare for this possibility, the CAISO is considering using a Transitional Alternative Pricing and Settlement (TAPAS) approach² during a transition period, and has commissioned Charles River Associates (CRA) to analyze the TAPAS approach and make recommendations concerning some design options within it.

Under the TAPAS approach, the CAISO would use all the concepts and systems developed for the LMP market in the Market Redesign – Technology Upgrade process (MRTU, previously MD02) to determine the quantities of energy and ancillary services (A/S³) in forward schedules and the real-time dispatch, but would settle the energy quantities, not at the associated LMPs, but at uniform market prices (UMPs), one for each of three pricing zones within the state. The CAISO would make compensation payments to generators (and other

¹ A seller’s choice contracts allow the seller – e.g., a generator – to designate, within limits, where on the CAISO-controlled grid the energy will be deemed to be delivered to the buyer. In an LMP market, a generator at location A can designate contract delivery at some lower-LMP location B and be paid for the (notional) counterflow from A to B, while the buyer must pay congestion charges for the (notional) flow from B to the buyer’s probably-higher-priced location C – even though the energy actually flows from A to C along many paths that have nothing to do with the notional delivery point B.

² CAISO “Draft Work in Progress” entitled “Transitional Alternative Pricing and Settlement (TAPAS) Approach to Locational Marginal Pricing (LMP)” dated September 3, 2004, referred to here as the “CAISO TAPAS draft.”

³ This paper deals only with the A/S that are scheduled and priced in the market processes described here, i.e., regulation and operating reserves.

resources⁴) that are “constrained up” to produce energy for which the bid price is greater than the relevant zonal UMP. The costs of the ISO’s compensation costs in each zone would be recovered with a \$/MWh charge or “uplift” on some measure of energy delivered to or taken from the CAISO-controlled grid in that zone by scheduling coordinators (SCs). When zonal UMPs differed, interzonal transactions would pay a congestion charge equal to the difference, which CRRs available to hedge the risks of the interzonal congestion charges.

The principal alternatives within the zonal-UMP TAPAS approach concern how the UMP is determined for each zone, whether and how market participants affected in various ways by CAISO congestion management actions are compensated and how A/S are paid. Whatever choices are made on these issues, more detailed issues such as how the uplift is allocated among SCs and how market power is mitigated will have to be resolved.

This paper analyzes the TAPAS approach and the principal alternatives within it. A major theme of the analysis is that any non-LMP market is by definition no better than “second best,” and that second-best situations inherently require difficult, often judgmental decisions among theoretically unsatisfactory alternatives. Analyses such as the one presented here can identify issues, pros and cons, and potential risks, but cannot make the final judgments required for choosing among the alternatives. The choice of a second-best alternatives to LMP, and between this alternative and the LMP market itself, must ultimately be made by the CAISO and its regulators informed by analysis, the preferences of stakeholders, and considerations of administrative and even political feasibility.

1.2 SUMMARY AND CONCLUSIONS

This paper begins with an overview of alternatives to LMP, drawing on both theory and real-world experience with UMP markets. It then describes the TAPAS approach and compares it to functioning UMP markets. The main part of the paper discusses the principal issues raised by a TAPAS-type market. A concluding section describes the principal steps in the daily cycle of the TAPAS approach, including some options at each of the steps. Two appendices contain some detailed discussions and examples better dealt with there.

The basic conclusion of the analysis here is that the TAPAS approach, while clearly “second-best” compared to the full LMP approach developed by the CAISO, would almost surely be better than the current California market – assuming that some difficult implementation issues can be resolved. The issue discussed here is not whether a TAPAS market would be better than an admittedly bad market; it is how to make such a market second-best to the first-best LMP market.

The implementation issues with the TAPAS approach are unusually difficult because the TAPAS market would do what no UMP market in the world has done, or even really tried to

⁴ For simplicity of exposition, this paper deals explicitly only with generators, even when the discussion also applies to (e.g.) importers and/or to dispatchable loads.

do: Determine forward and dispatch schedules using very complex and sophisticated processes, including simultaneous intertemporal optimization (IO) of energy and A/S with a full network model (FNM) and a two-settlement system, all of which were developed for the LMP market – and then use “simple” UMPs to settle the energy quantities in those schedules. Because transmission congestion in California is significant and is unlikely to become less so – another difference between California and places where UMP markets have been successful or have at least survived for some years – the sophisticated scheduling/dispatch processes are probably necessary. But it will be a challenge to attach a simple UMP façade onto such complex LMP machinery without something falling off or into the gears.

The principal conclusions of the analysis can be summarized as follows:

- **Any UMP Market Is No Better Than “Second Best”**

LMPs are the only settlement prices that are “incentive compatible” with efficient and secure operations of a transmission-constrained electricity system. Any other settlement prices create imperfect, even perverse, incentives that can be corrected only imperfectly with some combination of compensation, additional markets and administrative enforcement.

- **“Successful” UMP Markets Are Uncongested and Simple**

Virtually all long-lived UMP markets have little congestion (and aggressive investment programs to keep it that way), make constrained-up and (with a few exceptions) constrained-down payments, and impose locational grid access charges on generators and loads. None of these markets uses the kind of sophisticated unit commitment/scheduling/dispatch processes or two-settlement system contemplated in the TAPAS approach.

- **The Logic of a UMP Market Requires Compensation – and More**

A UMP market is second-best compared to a LMP market, but does have an economic logic of its own. This logic calls for both constrained-up payments (CUPs) and constrained-down payments (CDPs) to maintain short-run incentive compatibility, along with other payments, such as locational grid access charges and capacity payments, to correct the longer-term incentives that are distorted by the combination of a UMP, CUPs and CDPs. If the logic of a second-best solution is compromised by (e.g.) a decision not to make CDPs, it is necessary to look for “third-best” ways to offset the reduction in short-run incentive compatibility

- **The Second-Best UMP Depends on Compensation Policies**

Second-best considerations suggest that the value of the UMP should be determined to minimize the short-run operational inefficiencies inherent in any UMP. General economic reasoning suggests that the second-best UMP in this sense for dispatchable resources is the Unconstrained Market-Clearing Price (UMCP) if both CUPs and CDPs are paid, but is significantly lower if CUPs are paid but CDPs are not. The second-best UMP for price-

taking loads is a load-and-price-elasticity-weighted average of the LMPs associated with the efficient dispatch whatever is done for dispatchable resources.

- **A UMP Market Without CDPs Needs Scheduling Priorities**

If generators who want to run and collect the UMP are constrained off without compensation they will have incentives to bid large, negative prices and to self-schedule, which can badly distort the “economic” unit commitment and dispatch processes . The CAISO will need a system of non-price scheduling priorities to deal with this problem.

- **Operational Complexity Implies UMP-Market Complexity**

UMP markets that use simple procedures for determining operations can use similarly simple procedures for determining appropriate UMPs and compensation payments. But the TAPAS approach uses the entire MRTU multi-step IO apparatus to determine efficient and secure unit commitments, forward schedules and the real-time dispatch. The only logical and practical way to determine the appropriate UMP and compensation payments for settlement purposes is to develop transmission-unconstrained versions of these same IO processes. If the MRTU processes turn out to be simpler in practice than they are said to be in concept, the pricing and settlement complexity will also be less than suggested here.

- **A Two-Settlement UMP Market Requires Special Settlement Rules**

A two-settlement UMP market complicates incentive problems because resources that are constrained up/on or down/off in the forward market have incentives to act in the real-time market to undo their forward market commitments. The logical way to correct this problem is to make forward-market CUPs and (if made) CDPs contingent on actual performance in real time.

2. OVERVIEW OF ALTERNATIVES TO LMP

Before discussing the TAPAS approach itself, it is useful to provide some conceptual and practical context. This section begins by reviewing some basic economic and electricity market concepts to explain why and in what sense LMP can be said to be the “first-best” prices while any others are no better than “second-best,” and how possible second-best solutions can be identified and evaluated. It then reviews experience with non-LMP pricing methods in electricity spot markets worldwide.

2.1 SOME THEORETICAL CONSIDERATIONS

2.1.1 COMPETITIVE MARKET CLEARING AND CONSTRAINED OPTIMIZATION

Markets developed naturally in much of the economy long before Adam Smith explained how the self-serving and apparently uncoordinated actions of many competitors can serve the interests of the larger society. Over the next two centuries, economists and mathematicians developing Smith’s ideas demonstrated the duality between the outcome in a competitive market and the mathematical solution to a constrained optimization problem. In the specific and simplified form most useful for the purposes here, this duality result says the following:

***The Duality Result:** Under certain conditions (which are never met fully but are often met approximately), the quantities and prices that clear a competitive market will (approximately) minimize the cost of meeting market demand given the relevant constraints; and, conversely, mathematically minimizing the cost of meeting market demand subject to the same constraints will yield quantities and prices that (approximately) clear a competitive market.*

This duality between a competitive market and a constrained optimization problem is an interesting but not very useful result for most of the economy. In most economic sectors, competitive, decentralized markets can find a solution that is (approximately) efficient in the sense that it minimizes the cost of meeting demand subject to the constraints – the definition of “efficiency” used here – while it would be impossible to formulate, solve and enforce the related constrained optimization problem centrally with any degree of efficiency. But where electricity is concerned, there is no practical way to find even a feasible solution – i.e., one that satisfies the physical supply, demand and transmission constraints without regard to cost – without centrally collecting and processing a lot of information virtually continuously and in real time, which a decentralized market by definition cannot do.

For a century after Thomas Edison built the first central electricity systems, it seemed obvious that there was little positive role for competitive markets in electricity system operations. It was generally accepted that reliable operation of an electricity system required a central entity to gather information, find a feasible and reasonably efficient solution, and implement that solution. Because the central entity could get the required information and could control short-run operations effectively only if it directly controlled all the required resources, it had

to be vertically integrated; thus, the need for short-run centralization of real-time operations precluded any significant role for competition even in longer-term markets.

The breakthrough that allowed competitive market processes to play a large role in electricity was based on the duality result outlined above. There is no escaping the reality that operating an electricity system requires a central monopoly to collect and process a lot of information virtually in real time. But this monopoly need not own or directly control the production resources; it can be an independent system operator (ISO⁵) that gets the required cost and value information from decentralized competitors in the form of price bids,⁶ combines these bids with system information to formulate a constrained cost-minimization problem, solves that problem mathematically to determine (approximately) efficient quantities and the associated market-clearing prices, and then relies (primarily) on those prices to motivate the competitors to produce and consume the efficient quantities. An ISO doing this is not determining a *different solution* than a competitive market would, but is merely using a *different process* to determine the market-clearing quantities and prices that a decentralized competitive market would determine if such a market were possible in electricity.

It is critical to recognize that a competitive market-clearing solution, whether determined by decentralized trading or centralized calculation, consists of both the market-clearing quantities and the associated market-clearing prices that equate market demand to market supply in all interrelated parts of the market. Only when the market-clearing/efficient quantities are priced at the logically associated prices is the overall solution “incentive compatible,” in the sense that it is in the self-interest of each (competitive⁷) market participant to provide bids that reflect its actual costs/values and then to produce or consume the efficient quantities. It is not enough to use a mathematical process to find efficient quantities and then slap arbitrary prices on them for trading and settlement purposes, because such non-market-clearing prices will not be incentive compatible with the efficient quantities; such prices will make it difficult to enforce the efficient solution in the short run, will stimulate changes in behavior that will make it hard to find efficient solutions in the medium run, and will change investment and location decisions in the long run.

⁵ This paper deals only with electricity systems that are based on more-or-less open and efficient spot markets. The term “independent system operator” (ISO) here refers generically to the central entity (or entities, where system operations and the spot market are separated) that operates the power system and the spot markets in such a system. The term “California ISO” (CAISO) refers specifically to the organization with that name.

⁶ The term “bids” here includes both bids and offers. A bid generally includes price-quantity pairs that define a demand or supply curve, as well as operating constraints such as ramping limits, minimum and maximum quantities, etc.

⁷ Except where market power mitigation is discussed, it is assumed here that individual market participants are acting competitively, i.e. they are responding to market prices that they cannot much affect by their individual actions.

2.1.2 “FIRST-BEST” AND “SECOND-BEST” SOLUTIONS

In practice, it is never possible to implement fully a theoretically “first-best” market-clearing process (or anything else); compromises and approximations are always necessary in a complex and imperfect world. But market designers and operators charged with approximating a market-clearing solution at least have some well-understood and workable principles to guide them – e.g., prices should equate market supply to market demand – and can take these principles as far as practical. In contrast, if market designers and operators must work under some hard constraint on pricing that effectively rules out a market-clearing solution – e.g., electricity prices must be uniform over large areas even in the presence of congestion – they have little to go on.

Economists use the term “second best” to describe economic outcomes that are about as good as can be expected given some hard constraint on prices or other variables. In searching for a second-best solution, even basic and usually valid economic principles such as “equate supply to demand” cannot be trusted, because it may not be prudent in the short run or feasible in the long run to let markets respond freely to badly distorted prices.⁸ Finding a second-best solution is essentially a matter of formulating a range of plausible *ad hoc* alternatives, each of which will create its own set of problems and inefficiencies, and comparing these alternatives in terms of the overall inefficiencies they cause.

In complex situations, there is no practical way to identify, quantify and evaluate all the short-term and long-term effects of alternatives, so choosing among them is largely a matter of judgment based as far as practical on the facts as perceived at the time. Such pragmatic, fact-based judgments can and will be continually challenged and difficult to defend, particularly because such judgments *should* be reexamined as perceived conditions change and then modified if appropriate. As a result, second-best situations are inherently fraught with conflict, uncertainty and instability that will persist and even increase until the constraints creating the situation are eased.

2.1.3 LMPs AS THE FIRST-BEST PRICES

The mathematical optimization that simulates a competitive electricity market produces an efficient, market-clearing schedule⁹ and associated energy prices. When congestion and/or losses are important,¹⁰ the energy prices are different at every node on the grid, i.e., they are

⁸ The mathematical Theory of Second Best says that when there is a hard constraint on even one of the many standard efficiency conditions, maximizing economic welfare subject to that constraint will require violating many/most of the other standard efficiency conditions in unexpected ways.

⁹ The term “schedule” here refers to any set of generation and load quantities, whether in forward schedules or a real-time dispatch. A “feasible schedule” is a schedule that satisfies all of the physical operational and security constraints in the model used to determine the schedule.

¹⁰ For simplicity of exposition, the qualification “when congestion and/or losses are important” is implied here even when not explicitly stated.

LMPs. These LMPs are “first-best” in the sense that they are the only settlement prices that are incentive compatible with the associated schedule; only when all energy injections and withdrawals are settled at these LMPs will energy prices alone give each competitive market participant the right incentives to submit cost-reflective market bids and then to comply with the resulting schedules.

2.1.4 UMPs AND SECOND-BEST SOLUTIONS

A requirement that an electricity market use uniform market prices (UMPs) for settlements over large regions is the kind of hard constraint on pricing that creates a second-best situation, with all the difficulties and inefficiencies this implies. When congestion is significant, no UMP will clear the market for or be incentive compatible with an efficient schedule; any efficient schedule will require some generators to produce more or less than the quantities that would maximize their operating profits for any UMP. The ISO can and should try to reduce the incentive incompatibility between an efficient schedule and any UMP by adding supplementary markets and compensation payments and then choosing a value for the UMP that is second-best within this combination of policies. The combination that is second-best in terms of short-run inefficiencies may create additional problems in other markets and longer time frames, but the best response to any additional problems is to add additional second-best elements that attack these problems directly rather than to drop elements that are important for a second-best solution in the short run.

All electricity markets have rules requiring generators to follow ISO instructions within tolerance bands and/or under certain conditions, enforced by administrative and financial penalties. But such rules should be back-up devices, not the primary motivator for market participants to do what the system needs. The basic logic of markets depends on the fact that market participants will and should respond to market prices in ways that advance their own commercial interests. If an ISO routinely tells market participants to do things that reduce their market profits and penalizes them for responding to market incentives, the logic of and justification for the market is weakened and the ISO must take over more and more responsibilities from market participants.

The problems created by imposing uncompensated congestion costs on individual market participants have long been recognized, so in most non-LMP markets the ISO makes CUPs and (usually) CDPs to those market participants who incur significant commercial costs when they follow ISO schedules. Such compensation payments can reduce the problems involved in finding and implementing an efficient schedule despite the non-LMP prices, but – like any *ad hoc* procedures adopted in pursuit of a second-best solution – create new problems that require more second-best fixes. Most importantly, the combination of non-LMP prices and compensation payments can create perverse long-term incentives that require additional fixes.

An ISO constrained to use non-LMP prices can always find some way to deal with the resulting problems at least for awhile; the challenge is to find second-best solutions that minimize the inevitable inefficiencies and other problems. Before looking for specific second-best solutions, however, it is important to recognize that different solutions imply different market roles and institutional forms for the ISO; conversely, the role and form of the ISO limit

the options available when searching for a second-best solution. For example, an ISO that is allowed to issue orders that impose significant, uncompensated commercial costs on market participants has options – and creates problems – that a different kind of ISO does not. An ISO that is required to operate according to well-defined and transparent rules has constraints that a different kind of ISO does not. An ISO that is a private, profit-seeking corporation can be given financial incentives to exercise discretion constructively that are not practical for a non-profit ISO. The relationships between ISO characteristics and the alternatives available for dealing with non-LMP prices are particularly important when trying to draw lessons from experience elsewhere, where the market institutions may be very different.

2.2 REAL-WORLD EXPERIENCE WITH UMP MARKETS

This section provides a high-level, partially historical summary of the experience with markets that, in one way or another, are – or, in some cases, were – based on locationally uniform settlement prices.

2.2.1 THE CHILEAN MODEL

The first more-or-less competitive electricity market based on a central pool began operating in Chile in 1982, and was initially successful enough that very similar markets were adopted elsewhere in Latin America during the 1990s. The Chilean model has not been adopted outside Latin America and has evolved differently in different countries there. The best examples are now in Chile and Peru.

The Chilean model is based primarily on bilateral contracts, with a central pool operated by a Committee for Economic Operation of the System (Spanish acronym “COES”) that is essentially a generator’s club with opaque and informal procedures and processes. It is difficult even for smaller COES members – large members dominate the operating committees – to find out what is really going on and almost impossible for outsiders, even regulators, to do so.

Generators use the COES pool to determine a reasonably efficient dispatch and then settle differences between contract and dispatch quantities among themselves at a UMP. LSEs and large customers cannot buy and sell in the COES spot market, but must buy from generators under contracts, with regulated energy and capacity prices for LSEs serving all but large “free” customers.

COES decides which generators will provide A/S and be constrained up and down to manage congestion, and then compensates them somehow. There are interminable battles within COES about dispatch procedures, settlement prices and payments for A/S and congestion management, but the basic principle is that generators who provide such services should be paid for them, i.e., that both CUPs and CDPs should be made.

All payments for A/S and congestion management are among generators within COES. In effect, COES levies an uplift on all its generator-members to pay some of them to provide A/S and congestion management. LSEs and large consumers make no payments to COES

directly, and pay generators only for energy and capacity at contract prices. Not surprisingly, generators feel that regulated contract prices are too low to cover their COES costs as well as their own generation costs; they want higher contract prices for energy and capacity and/or separate payments for A/S and congestion management.

The Chilean model was very successful at first, largely because it reduced the role of corrupt and incompetent state-owned entities and politically dominated regulation. It has many problems and has continually been challenged and changed. But congestion management does not seem to be one of the biggest problems, as least as seen from the outside or reported from inside COES. The principle that all generators who incur costs to provide A/S and congestion management should be paid for these services has let a UMP market function, however inefficiently, for over 20 year in Chile and almost 10 years in Peru.

2.2.2 ENGLAND AND WALES

The original England and Wales (E&W) Pool began operating in 1991 and was replaced by the New Electricity Trading Arrangements (NETA) in 2001. The Pool was and NETA is operated by the National Grid Company (NGC), which also owns and operates the grid.

The Pool was a day-ahead spot market that used a system-wide UMP for each half hour of the following day determined by clearing a hypothetical unconstrained market; it was a “gross” pool, in the sense that all energy on the system had to be bought and sold through it, with all bilateral contracts taking the form of contracts for differences (CfDs). NGC managed real time imbalances and congestion by buying incs and decs from generators at their offer prices in the day-ahead market, which is equivalent to making both CUPs and CDPs. When generators with local market power began taking advantage of it to extract exorbitant CUPs and (less so) CDPs, NGC, with the help of regulators, negotiated contracts with the critical generators to control their spot market actions; these contracts were analogous to the reliability-must-run (RMR) contracts later used in California and elsewhere.

NGC recovered its balancing and congestion management costs each hour with an uplift on all energy purchased from the Pool, i.e., all energy taken from the system in a gross pool. The uplift increased rapidly in the early years of Pool operations, until NGC negotiated with the Pool an incentive arrangement under which NGC had large discretion to negotiate bilateral payments or contracts with market participants and was allowed to keep a share of any reductions in total uplift costs below a target value. The combination of NGC flexibility and financial incentives resulted in significant reductions in the uplift – and a powerful, opaque and highly profitable NGC.

Although by most objective measures the Pool worked reasonably well, for a variety of reasons (with pure politics playing a large role when Labour replaced the Tories) it was replaced by NETA in 2001. The main objective of NETA was to replace the central spot market with bilateral contracting and trading. Under NETA, scheduling entities submit balanced schedules for each half-hour at “gate closure” four hours ahead of the operating half-hour. In real time, NGC manages imbalances and congestion by buying incs and decs in a balancing mechanism (BM) that was deliberately designed not to be an efficient market, i.e.,

there is a large gap between buy and sell prices so that spot trading is penalized. NGC recovers its BM costs with an uplift on loads, with incentives to reduce the uplift.

NETA is not exactly a UMP market because there are no actual transactions at a UMP, but it does have the most fundamental characteristic of a UMP market: market participants can transact as though there were no operational congestion and the ISO – NGC in this case – resolves congestion by, in effect, making CUPs and CDPs and socializing the costs.

Under both the Pool and NETA, generators and loads each pay half of NGC’s fixed costs through grid access charges. To help deal with the long-term effects of a UMP, these grid charges vary by location to reflect the long-run marginal cost of the optimal grid to serve generation and load at each location. (See section 4.1.3 for more discussion of this NGC policy.) New generation and load facilities are required to pay their own direct connection costs, but NGC pays the costs of “deep” connection assets needed to accommodate them; generators and large loads that begin operating before NGC has completed the required deep investments can be curtailed without compensation when congestion arises.

The UMP-like markets in E&W, both in the Pool and NETA forms, have “worked,” in the sense that the lights have stayed on, prices have not been bad,¹¹ new generation (too much at times) has been built and transmission has been expanded more or less as needed. The principal problems have been the bias against undiversified generators inherent in NETA’s bilateral market with high imbalance penalties, and the inflexibility of the firm transmission rights implicit in a UMP market with CUPs and CDPs. (See section 4.1.3 below.) Subsidies for small-scale renewable generation have been used to offset the first of these problems, but the transmission rights problem has proven more difficult.

To at least some extent the relative success of the UMP market in E&W has been due to the fact that the E&W system is an island with few significant internal constraints – and a NGC policy of investing to eliminate these when or before they do arise – and no external loops.¹² The scheduled extension of NETA into Scotland to create the British Electricity Transmission and Trading Arrangements (BETTA) will change this because transmission capacity between the E&W and Scottish systems is limited. Because NETA is not a true UMP market, it is not possible “simply” to add another zone with its own UMP and create financial transmission

¹¹ Prices have gone up and down, but there are continuing debates about the extent to which these movements have been due to market arrangements as opposed to other, more fundamental factors such as technology and fuel cost changes and the supply-demand balance. In particular, NETA advocates say that NETA has decreased prices while POOL defenders say that the decline in prices – which began before NETA began operating – was due to other factors, such as the excess capacity left over from the POOL era.

¹² NGC deals with a lot of minor congestion in real time, but does so in a flexible, opaque way under its uplift-reduction incentive. There is a single DC link with France, which is treated as generation or load at the entry point into England; a separate commercial entity manages transmission rights across the link. The AC interconnection with Scotland has been managed much the same way.

rights on the interconnector. NGC and the regulator have been trying for years to define some sort of tradable “injection” and “withdrawal” rights, with little success. NGC is supposed to produce a proposal for such rights any time now.

2.2.3 THE PJM UMP MARKET

Before the successful PJM LMP market there was the abortive PJM UMP market. This market did not make CDPs, and the first time significant congestion arose on the east-west interface the ISO lost control of the system within hours and had to suspend the market. This experience is usually cited as an excellent argument for LMP, which it is; but it is also an instructive lesson on the dangers of operating a UMP market without CDPs.

The initial PJM market got into trouble so quickly because of the combination of three factors: (1) no CDPs; (2) the right to self-schedule; and (3) no congestion charges for scheduling across a constraint. When congestion on the interface developed and PJM constrained off some generators in the west, these generators quickly contracted with load in the east, self-scheduled to meet that load and kept running. PJM then had to constrain off some other generators that quickly did the same thing. Within hours all of the still-excess generation running in the west was self-scheduled and PJM had to suspend the market.

Eliminating any of the three factors listed above would have prevented such a rapid collapse, although slower-acting poisons would have had their effects eventually. If PJM had made CDPs, constrained-off generators would have had little incentive to contract bilaterally and self-schedule. If generators had not been allowed to self-schedule they would have had to try less-effective ways to shift congestion costs to others, such as bidding low or negative prices into the PJM spot market. And if self-schedules had had to pay congestion charges – which would have required some form of zonal market to produce the price differentials that define congestion charges – self-scheduling would not have been attractive. The main lesson for TAPAS in California is obvious: self-scheduling is potentially a serious problem that will have to be controlled somehow.

2.2.4 THE NEPOOL/ISO-NE UMP MARKET

The New England electricity market developed from the previous tight power pool among the integrated utilities, NEPOOL. The original NEPOOL market, operated by ISO-NE, was based on a system-wide UMP with no CDPs. Although the market had many problems and was ultimately abandoned in favor of LMP, it did operate from 1997 to 2003 without CDPs and without the kind of instantaneous meltdown experienced by the PJM UMP market. For this reason, the ISO-NE UMP market is sometimes cited as one that “worked” without CDPs; the questions are in what sense it worked, and why.

In the initial ISO-NE market, the real-time marginal price (RTMP) was set at the bid price of the most expensive dispatched resource that was really marginal – i.e., that was not up against one of its operating constraints and hence could produce more and/or less – and that was not “flagged” as running to resolve congestion. A constrained-up resource was paid the lower of its bid price and a maximum established under the market power mitigation procedures.

Constrained-down resources were paid the RTMP for what they were dispatched to (and did) produce, but nothing for their constrained-down amounts.

In practice, when congestion arose ISO-NE put congestion “flags” on essentially all resources running in constrained-up (high-cost) regions, leaving only resources that were actually running in constrained-down (low-cost) regions eligible to set the RTMP. Because the higher-cost resources in low-cost were the ones more likely to be constrained down, the RTMP was set by the lower-cost resources in low-cost regions. At such low RTMPs few resources were constrained down without CDPs – but many were constrained up and received CUPs based on their bid prices subject to cost-based market power mitigation caps.

Occasionally, there was so much excess generation in constrained regions that some had to be constrained down even at low prices; this was particularly true in Maine in 2001, when many new gas-fired combined cycle plants came on-line behind the transmission constraints into the southern New England load centers. When this occurred, ISO-NE used the threat of severe penalties to enforce compliance with its constrained-down dispatch instructions. The “bidding-down” game never became so serious that ISO-NE could not determine a feasible dispatch; and self-scheduling, although allowed, did not become the problem in ISO-NE that it did in PJM.

A major reason ISO-NE was able to enforce its (infrequent) constrained-down instructions despite the lack of CDPs had to be that RTMPs were so low when congestion arose that there was little incentive to find ways to avoid being dispatched down.¹³ But these low prices depressed generator profits everywhere, and the cost-based price caps used to determine CUPs kept generation in constrained-up regions from recovering their costs. New generation was being added where it was not needed despite the absence of CDPs, while old generation was being shut down where it was needed despite CUPs. The market design was seriously flawed, so NEPOOL decided to replace it with a LMP market.

The principal lessons from the ISO-NE experience are that a UMP market, even with significant congestion, can function for awhile if it sets the UMP low enough that few resources are constrained-down, but that such a low UMP will discourage investment generally without necessarily eliminating the incentive a UMP market creates for resources to locate in the wrong places.

2.2.5 AUSTRALIA

Australia, even more than the United States, is a federation of independent states that have primary responsibility for their own electricity systems (and other things). The first electricity market began operating in Victoria 1996. With encouragement and some “bribery” from the federal government in the form of subsidies for interconnections, the southeastern states

¹³ This is empirical support for the theoretical conclusion in section 4.2.1 that the UMP that minimizes dispatch inefficiencies when CUPs but not CDPs are made is something like the lowest LMP at a constrained-down node.

agreed to create the National Electricity Market (NEM). The NEM, which began operating in 1998, now includes Victoria, New South Wales, South Australia and Queensland, with Tasmania to be added if and when an interconnector is built under the Bass Strait.

The NEM is a zonal UMP market with interregional congestion charges based on zonal UMP differences and a (not very effective) form of CRRs for hedging these. The four original (and current) zones closely followed state lines, which was politically convenient but also electrically appropriate given that the Australian system consists of a few dense urban areas, one per state, interconnected by a few long transmission lines. The interconnections among the states were also primarily radial initially, so it was reasonable to model them as generic interconnections between zonal reference nodes and to assume that loop flow effects would not create significant LMP differentials within zones in the absence of intrazonal congestion.

A premise of the market design was that new zones would be created as necessary to keep intrazonal congestion at insignificant levels. The NEM Code says that NEMMCO – the ISO in Australia – will study the costs and benefits of creating new and internally unconstrained zones whenever a constraint within a zone is binding more than 50 hours in a year. There have been a few instances of intraregional constraints hitting this trigger and a few studies of possible subdivisions of regions and some form of intraregional congestion pricing, but the states, concerned about the political impact of intrastate price differentials, have blocked implementation of all such ideas. When intrastate congestion has arisen it has been transitory, as the network has been quickly upgraded to eliminate it.

The Australian NEM may be the only UMP market in the world operating with neither CUPs nor CDPs. NEMMCO can and does constrain generation up and pay CUPs in emergency situations, but the philosophy is to use longer-term solutions. Where network upgrades are too costly or are delayed, a regional (state) grid owner can use network support contracts that pay a generator to run just enough to relieve congestion, with the costs of such contracts recovered through region-wide (state-wide) tariffs; the few such contracts are used to support isolate loads at the end of long lines.

Explicit noncompliance with constrained-up/down instructions is rare despite the absence of CUPs/CDPs, because dispatch is mostly automated and there are severe penalties for noncompliance. There have been and continue to be instances of generators bidding low/negative prices or using other measures to try to avoid being constrained down. But the primary response to uncompensated congestion has been pressure to upgrade the grid and socialize the costs, which would be even easier under proposed changes to the regulatory regime.

The Australian electricity system and institutional context are so different from those in California that there may be few lessons from Australia relevant to the TAPAS proposal. Some possible lessons are: (1) a system with a simple network can spend enough on it to keep congestion minimal; (2) poor congestion management increases pressure to spend too much on the network; and (3) a zonal UMP market finds it very difficult to evolve by adding additional zones.

2.2.6 ALBERTA

The Alberta electricity market, which began operations in 1996, is unique in that there is a Transmission Administrator (TA) in addition to a Pool and the transmission owners (TOs). The Pool consists of a Power Pool Administrator (PPA) and the System Operator (SO). The PPA determines a UMP for the whole province based on a hypothetical unconstrained dispatch, and the SO manages real-time congestion by making CUPs but not CDPs. The TA plans transmission upgrades and awards contracts that effectively subsidize generators to locate where they reduce congestion if this is more cost-effective than new transmission. The TA is a regulated profit-making contractor with some financial incentives to reduce the total costs of losses, its congestion-reducing contracts and the cost of grid assets. The TA's costs are recovered from system users through a combination of energy uplift and transmission tariffs.

The basic philosophy of transmission and congestion management in Alberta is that the TA will plan and implement a cost-effective combination of new transmission and subsidized generation in critical locations so that the system can expand in response to the market – meaning the combination of energy prices, grid access charges, TA payments and contracts, etc. – without creating significant real-time congestion under normal conditions. When real-time congestion does arise, usually because of an unexpected grid outage or other problem, the SO decides which resources to constrain up and down using the bid-based merit order from the PPA as far as practical. But in the absence of constrained-down payments, generators may try to avoid being constrained down by lowering their bids to zero (the lowest bid allowed in Alberta), making merit-order dispatch largely meaningless if not impossible.

When merit-order dispatch cannot resolve congestion the SO curtails generation and load. The first to be curtailed are any “trigger participants,” defined as new generation or load in transmission-constrained zones where the TA has not yet expanded the grid enough to handle it. After all trigger participants have been curtailed the SO curtails on a *pro rate* basis. Such curtailments have been rare.

Generators and loads each pay half of fixed grid costs through grid access charges. Because a UMP, even without CDPs, gives generators no incentive to avoid locating in congested areas, grid access charges for generators vary across twelve zones, with higher charges in zones with excess generation, i.e., where constrained-down instructions and curtailments are more likely.

The Alberta market has operated under this system for eight years now and, despite recurring problems and continual conflict, has rejected proposals to change the congestion management approach by implementing LMP or some sort of transmission right. The main reasons this system has “worked” are that the Alberta grid is highly radial, with three large urban areas separated by long transmission lines, and the TA has used socialized transmission investments and network support contracts to keep real-time congestion small.

2.2.7 ONTARIO

The Ontario electricity market began operating in 2002. In this market, the Independent Market Operator (IMO) determines a real-time dispatch and associated LMPs using a myopic

– i.e., period-by-period – security-constrained economic dispatch program that simultaneously optimizes energy and A/S, but then settles all spot energy transactions at a UMP that clears a hypothetical unconstrained market. The IMO manages congestion by making both CUPs and CDPs. There is no IMO-operated day-ahead or hour-ahead market, no (formal) intertemporal optimization by the IMO and no capacity payments. The IMO market is as close to a real-time, energy only spot market as any in the world.

Most of the generation in Ontario is owned and operated by Ontario Power Generation (OPG), a state-owned enterprise created from the generation assets of Ontario Hydro. OPG operates under a market-power mitigation agreement (MPPA) that is in substance an aggregate one-way contract-for-differences (CfD) between OPG and consumers, with a state agency acting as agent for consumers. Under the MPPA, if average prices over a year exceed a specified target level OPG must pay the state agency the price difference multiplied by an amount of energy equal to about 85 percent of OPG’s expected output given its capacity mix, and this money is rebated to consumers through the IMO’s settlement system the following year. The amount of energy to which the rebate applies is reduced as OPG sells or otherwise “decontrols” capacity.

The Ontario electricity market ran into serious political problems soon after opening, largely because spot prices increased rapidly, the MPPA rebates that would offset these high prices were not scheduled to arrive until the following year, and the political commitment behind the market was always weak. The surge in prices was due primarily to rapid demand growth, the failure of OPG to restart some large nuclear units as planned, and the lack of other generation investment, probably attributable to the depressing effect of the long-planned but oft-delayed nuclear restarts. The government reacted to the price surge by capping prices and promising big changes in the market structure and design. Political turmoil and uncertainty continue.

The main criticism of the IMO market itself has been related to the Congestion Management Settlement Credits (CMSCs) – CUPs plus CDPs plus some other stuff – which have been much larger than expected. Most of the unexpected increase in CMSCs was due to the fact that the IMO decided or was compelled to make compensation payments for things other than congestion. For example, the IMO often constrains off energy-limited hydro generators early in the day to assure that their energy will be available later in the day when it will be needed for “reliability” reasons; this usually has little to do with congestion, but allows a clever hydro generator to collect CDPs for the same energy in every dispatch period of the day until the IMO finally releases that energy.¹⁴ Imports into Ontario are treated in a way that results in

¹⁴ An energy-limited hydro generator that bids “too low” in a morning hour will be constrained off by the IMO, but will then still have its energy so that it can do the same thing in the next hour, and the next, ... until the IMO finally releases the energy. Of course, a rational generator will not offer its limited energy at a low price in the morning if it expects a high price in the afternoon – unless it knows that the IMO will constrain it off if it does so, in which case it will bid low prices early in the day and create exactly the problem the IMO is concerned about. The IMO says that it never constrains off hydro energy for economic reasons, only for “reliability,” but the effects are the same.

large and inappropriate CDPs.¹⁵ And there are some technical details in the way uninstructed deviations are handled that can result in inappropriate CUPs and (more often) CDPs.¹⁶ There has also been concern about gaming to take advantage of CUP and CDP payments, but this has not been a major problem, perhaps because OPG has little to gain from it under the MPPA.

The Ontario Market Surveillance Panel (MSP) conducted an extensive stakeholder consultation process in 2002-03 on the question of whether to continue CDPs. The MSP began with the premise that CDPs should be eliminated immediately because they pay generators not to do something, distort incentives, are subject to gaming, etc. The final MSP report on the subject in July 2003 concluded that all those reasons were valid (although it did not identify gaming as a major problem) and in principle CDPs should be eliminated, but that because Ontario was considering moving to LMP soon it would not be worth the effort to end CDPs now. The MSP recommended that the specific problems mentioned above (and some others) be fixed, and that if Ontario decided to stay with a UMP market CDPs should be eliminated.¹⁷

The UMP-CMSC market in Ontario was intended to be a transition to a LMP market, probably with a day-ahead forward market. But there has been little support for moving to LMP, partly but not entirely because of the political uncertainty about the future of the whole market in Ontario. It has not been decided what to do about CDPs if, as seems likely, the UMP market continues, but the IMO is considering adding a day-ahead market.

There are three principal lessons from the Ontario experience with CUPs and CDPs that are relevant to the TAPAS approach for California: (1) compensating market participants for ISO actions can make it easier for an ISO to act arbitrarily, because it allows the ISO to buy off those directly affected and socialize the costs; (2) the details of the CUP/CDP calculation are

¹⁵ In simple terms, importers are allowed to schedule imports into the IMO market that are known to be undeliverable in real time because of constraints in New York, and the difference between schedules and actual imports results in CDPs; these CDPs are clearly inappropriate and should be eliminated. The IMO also guarantees importers that they will be paid at least the UMP as forecast two hours in advance, and the IMO's losses on this one-sided bet are collected via the CMSC; this is not a CDP problem, but critics do not make such fine distinctions.

¹⁶ This problem arises because of the intertemporal and ramping limit issues discussed in section 4.4 below. The transmission-unconstrained market schedule begins each period assuming that generators are producing where last period's schedule instruction told them to be, while the constrained schedule begins with where they really are. The resulting difference between market and constrained schedule can result in CUPs or (more frequently) CDPs.

¹⁷ The author of this CRA report prepared for the IMO an analysis of the MSP's initial analysis of CDPs, making the arguments made here, i.e., that CDPs are a logical and necessary part of a UMP market. What role, if any, that analysis played in the MSP's final recommendation to live with CDPs at least for awhile is unknown.

important but difficult to get right; and (3) supposedly transitional arrangements have a tendency to last a long time.

2.2.8 ERCOT

The ERCOT market opened in 2001. It is primarily a bilateral market with a real-time zonal UMP market used for balancing. Qualified Scheduling Entities (QSEs) submit balanced schedules and A/S bids a day ahead, along with balancing energy bids for each of the four (originally three) zones. The balancing bids are not resource- or location-specific within the zone, but are taken in merit order to balance load within each zone and to resolve interzonal congestion. The 15 minute zonal real-time price for each zone is the marginal balancing bid taken in the zone. Initially there were no congestion charges on schedules between zones.

After each zone is in balance and interzonal flows are within transmission limits, resource- and location-specific inc/dec bids are used to resolve intrazonal congestion but do not affect the zonal market price. Generators are paid their bid prices – cost-based if there is insufficient competition, as is usually the case – for the incs/decs, the equivalent of making both CUPs and CDPs.

Initially, both intrazonal and interzonal congestion costs were recovered through an uplift paid by all market participants without regard to causation. A trigger mechanism was established whereby congestion costs would be directly assigned to those who cause them if congestion costs for either intrazonal or interzonal congestion exceeded \$20 million in any one rolling year after the first six months. Interzonal congestion costs were \$137 million in the first month after the market opened (August 2001) and totaled \$188 million over the first 17 months. As a result, starting in February 2002 interzonal congestion costs were assigned to the QSEs scheduling over constrained interfaces. This resulted in a dramatic reduction in interzonal congestion costs, largely because it eliminated the incentive for QSEs to self-schedule flows that caused interzonal congestion so that they could be paid to resolve their own congestion – a version of the “dec game”.

Intrazonal congestion costs were \$212 million over the first 17 months. A major cause of these high costs was the “dec game”, in which a QSE submits a schedule that causes congestion along with dec bids to relieve that same congestion. Another, longer term problem is the lack of incentives to avoid locating new resources where they contribute to congestion (although a generator or resource that does so risks seeing those congestion costs come back if they are large enough to trigger causation-allocation rules). To reduce these problems the Market Oversight Division of the Texas PUC has proposed a LMP market for generators, with loads having the choice of paying LMPs or a load-weighted average of LMPs within a zone. This market redesign, which also includes a day-ahead market and CRRs, is not expected to be complete before October 2006.

The principal lessons of the ERCOT experience relevant to the TAPAS proposal for California are that a zonal UMP market, particularly when self-scheduling across constraints without congestion charges is allowed, can cause problems, and that LMP is much better. But the ERCOT UMP-market design had many flaws that are obvious at least in retrospect and

that the TAPAS proposal avoids; for example, the three-zone TAPAS UMP market would include congestion charges on interzonal flows from the beginning. It is unclear to what extent CDPs were a major cause of ERCOT's problems or that eliminating them would not have caused larger problems.

2.2.9 CALIFORNIA

The California experience is well-known in California so it will not be described here. It is useful, however, to identify in broad terms the features of the current California market that have caused its principal problems for comparison later to the TAPAS proposal.

The present California market is a form of three-zone UMP market with congestion charges on transactions between zones and the equivalent of CUPs for managing congestion. At this high level it is similar to the TAPAS proposal.

The real differences between the current market and TAPAS concern how the forward schedule and real-time dispatch are determined. In the current market the ISO does not optimize unit commitment in advance or find and implement an efficient DA schedule or RT dispatch. Instead, it starts with DA balanced schedules from SCs that are often not feasible on the real grid, and then makes a series of adjustments to these for reliability reasons using network models that (by design) do not accurately reflect reality, a series of A/S and energy markets that do not take the interactions properly into account, and a non-economic dispatch process that is subject to market separation rules explicitly designed to prevent the CAISO from identifying and implementing efficiency-increasing trades among SCs. The result is the market schedules are highly inefficient, even unreliable, forcing the CAISO to fix them as best it can as real time approaches and arrives, using ancillary services, RMR contracts and (in effect) CUPs/CDPs.

One critical characteristic of this process is that it requires market participants to base their bids to the various, separated CAISO markets on their own forecasts of outcomes in later markets rather than on objective assessments of their own costs and capabilities. The resulting bids, even for a highly competitive market participant playing no games, may have little relationship to observable costs and hence are difficult to distinguish from the strategic bids of a market participant exercising market power or trying to game idiosyncrasies in the market rules. Effective market power mitigation, either automatically in real time or after the fact, is very difficult or impossible.

The TAPAS proposal eliminates virtually all the constraints on the CAISO processes that now preclude determination of efficient, reliable forward schedules and real-time dispatch, which should make it easier to identify and mitigate market power and to find efficient and secure forward schedules and the real-time dispatch. However, the main points of the analysis here are that the complexity and sophistication of the MRTU processes raises some thorny implementation issues and that without CDPs market participants will still have incentives to use strategic bidding and self-scheduling to influence CAISO schedules. It is surely possible to develop a market based on the general TAPAS framework that would be better than the

current market in California. The real issue is how to make the TAPAS approach not just better than an admittedly bad market, but second-best to the first-best LMP market.

2.2.10 SUMMARY OF EXPERIENCE WITH UMP MARKETS

These diverse experiences with different kinds of UMP markets in different parts of the world suggest some generalizations about UMP markets, including the following:

- **Zonal UMP Markets “Work” When Intrazonal Congestion Is Small**

The only UMP markets that have lasted for any significant time have been on systems that had little intrazonal congestion to start with and that are able and willing to invest in grid assets and/or to subsidize generation in critical locations to eliminate intrazonal congestion as soon as – or before – it becomes significant.

- **UMP Markets Include CUPs “Always” and CDPs “Almost Always”**

Of the markets discussed above, only Australia does not routinely make CUPs, and even here emergency CUPs and network support contracts are used when necessary. Only Australia and Alberta have lasted for long without CDPs, and this is largely because they are sparse, radial systems that invest to keep congestion small. The ISO-NE UMP market “worked” for some years without CDPs primarily by setting the UMP so low that few resources were constrained down and many were constrained up.

- **UMP Markets Use Locational Grid Access Charges**

All of the UMP markets that have survived for some years require generators and loads to pay grid access charges that are higher where their operations increase congestion. Some such mechanism is needed to provide the long-term locational decisions that a UMP energy market, with or without CDPs, does not provide.

- **UMP Market ISOs Do Not Seek Global Efficiency**

UMP market ISOs do not use sophisticated unit commitment or scheduling tools based on intertemporal optimization (IO) to find globally optimal solutions subject to complex constraints. In UMP markets it is largely up to market participants acting in bilateral and exchange-based contract markets to decide how they want to operate and then to submit self-schedules and inc/dec bids reflecting what they want to do, and the ISO uses these in relatively simple processes to manage imbalances and congestion. The ISO is not as concerned with global efficiency as with finding a secure solution that is as consistent as possible with the self-schedules and bids.

- **UMP Markets Do Not Use Two-Settlement Systems**

None of the UMP markets discussed above uses the combination of day-ahead (DA) and real-time (RT) markets that has become standard in LMP markets (although Ontario is trying to develop a DA market to supplement its RT market). This is largely due to the fact that UMP markets rely primarily on decentralized trading to determine outcomes and tend to regard an ISO-operated DA market as unnecessary or even as unfair competition

for private forward markets. But it is also because UMPs and the associated CUPs/CDPs make it hard to manage the interactions between a DA and RT market, as discussed in section 4.6.

3. SUMMARY OF THE TAPAS UMP APPROACH

This section begins by describing the Transitional Alternative Pricing and Settlement (TAPAS) UMP approach and some design options within it. It then discusses some important differences between the TAPAS approach and other UMP markets.

3.1 PRINCIPAL ELEMENTS OF THE TAPAS UMP MARKET

3.1.1 MRTU TIMETABLES AND PROCEDURES

The market timetable and procedures developed in the MRTU project are to be maintained unless these must be changed or supplemented to implement the TAPAS approach. The principle features of these timetables and procedures are outlined in this section.

- **Pre-IFM Reliability and Market Power Mitigation Runs**

After receiving bids for the DA integrated forward markets (IFMs) markets but before clearing those market, the CAISO determines a market-based unit commitment and schedules of resources that meets the CAISO's demand forecasts (as opposed to demand bid into the market); only resources committed at this stage are eligible for later stages of the DA process. If the pre-IFM schedule indicates that congestion is likely to create local market power or reliability problems, the CAISO implements market power mitigation procedures and/or schedules RMR resources as appropriate.

- **Day-Ahead (DA) Integrated Forward Markets (IFMs)**

The CAISO operates integrated energy and A/S forward markets at the DA stage¹⁸ that determine physical unit commitments and forward schedules for both energy and A/S, along with the prices at which these forward-traded quantities are settled. The forward schedules and prices are based on the self-scheduled and bid-in demand from market participants (as opposed to the CAISO's own demand forecasts). Resources committed in the IFM are guaranteed that if they do not recover their start-up costs from market prices over the "day"¹⁹ the CAISO will make up the difference – a generalized form of CUP.

- **DA Residual Unit Commitment (RUC)**

If, after the IFM closes, the CAISO determines that insufficient resources have been scheduled to meet the CAISO's demand forecast within security constraints (taking into

¹⁸ In the CAISO's MD02 submission to FERC in July 2003 the term IFMs referred to both the DA market and an HA market with full pricing and settlement, creating a three-settlement system. In the current CAISO MRTU proposal the HA process is used only for planning purposes, not to produce settlement quantities or prices, so the only IFMs now are the integrated DA energy and A/S markets.

¹⁹ The period over which a resource is guaranteed cost-recovery is referred to here as a "day," although in practice it depends on the resource and the situation and is not necessarily 24 hours.

account how long it takes to start various resources), the CAISO may commit additional resources. Resources committed by the CAISO at this stage are guaranteed cost recovery over the day (or so).

- **Simplified Hour-Ahead (HA) Market**

The CAISO updates schedules and price forecasts in an hour-ahead (HA) process, based on its own demand forecasts. The results of this process are used in HA versions of the DA reliability, market power and residual unit commitment (RUC) processes described above, and to schedule market resources that cannot respond to real-time dispatch instructions. There are no HA prices used for settlements.

- **Pre-Dispatch and Real-Time (RT) Dispatch**

The CAISO performs a pre-dispatch approximately forty-five minutes in advance of real time to update schedules for resources that cannot respond to intra-hour instructions, and then performs a unit commitment for fast-start resources every fifteen minutes and a dispatch every five minutes that produces the real-time (RT) dispatch instructions that are transmitted to dispatchable market participants.

3.1.2 MRTU SCHEDULING/DISPATCH PROCESSES

In all the scheduling and dispatch processes outlined above – e.g., DA, HA, RUC and RT pre-dispatch/dispatch – the CAISO determines the physical energy and A/S quantities using the concepts, processes and systems developed for the LMP market in the MRTU/MD02 processes. These scheduling/dispatch processes are based on simultaneous intertemporal optimization (IO) of energy and A/S using security-constrained unit commitment/dispatch (SCUC/SCED) models, a full network model (FNM), and bids from market participants (mitigated to control market power when so indicated) that include supply or demand curves, start-up and no-load costs, minimum and maximum operating levels, and technical data such as ramping limits, start-up times, and minimum run and shut-down times. The MRTU processes automatically determine LMPs for energy and regional²⁰ A/S marginal prices (ASMPs) that are logically associated with the efficient schedules.

3.1.3 SETTLEMENT PRICES FOR ENERGY (UMPs) AND FOR A/S

After each of the DA IFM and the RT dispatch/market, zonal UMPs are determined for energy and settlement prices are determined for A/S. There are options for determining prices in each case.

- **Options for Zonal Energy UMPs**

An UMP for energy is determined for each of the three pricing zones within the state. There are two options for determining the zonal UMPs: (1) clear a hypothetical market

²⁰ In the full LMP market A/S are not priced by node but they may be priced by region. The A/S regions would probably not be the same as the three energy pricing zones used.

that has no constraints within zones but radial flow constraints between the three zones; or (2) compute the zonal weighted averages of the LMPs associated with the constrained schedules.

- **Options for A/S Settlement Prices**

There are two basic options for determining A/S prices: (1) use the ASMPs directly from the constrained solution (the option proposed in the CAISO TAPAS draft); or (2) determine regional “uniform A/S prices” (UAPs) using a hypothetical unconstrained schedule analogous to – or the same as – the schedule used to compute UMCPs.

3.1.4 UMP-MARKET SCHEDULES AND OPERATING PROFITS

However the UMPs and A/S prices are determined, compensation for congestion costs – even if only CUPs are paid – requires knowing how much the daily operating profit of each resource is reduced when it follows CAISO schedules instead of optimizing its commercial position given those prices. The need for such a calculation is often not recognized, because in simple cases the calculation for each resource involves little more than (e.g.) multiplying the amount of constrained-up energy in each hour by the difference between the bid price and the UMP in that hour and summing over all the hours in the day. But given the complexity of the sophisticated MRTU LMP processes, with their intertemporal optimization and joint optimization of energy and A/S, it will probably be necessary to use similarly sophisticated and almost equally complex methods to determine how much energy and A/S each resource would have provided and how much operating profit it would have made (based on its bids) in both energy and A/S markets over the day given the UMPs and A/S prices in the absence of congestion. The hypothetical schedules of energy and A/S in such a calculation are called here the “UMP-market schedules,” and the corresponding operating profits are called “UMP-market operating profits.”

3.1.5 COMPENSATION FOR CONSTRAINED-UP RESOURCES

The CAISO’s settlement system uses the UMP-market schedules and UMP-market operating profits to calculate and make CUPs to resources that are scheduled to supply some energy that costs more (as measured by bid prices) than the UMP they are paid for it, but does not make CDPs to resources that are scheduled *not* to supply some energy that costs less than the UMP they would be paid for it.

The CAISO also makes CUPs to dispatchable demands that are scheduled to take some energy that is worth less (as measured by bid prices) than the UMP they must pay for it, but does not make CDPs to dispatchable demands that are scheduled *not* to take energy for which they bid more than the UMP, i.e., dispatchable demand can be curtailed without compensation..²¹

²¹ CAISO TAPAS draft, p. 10. It is not clear why this compensation policy is adopted for dispatchable demand instead of just the opposite, but such a policy is likely to discourage loads from becoming dispatchable precisely where they are most valuable, i.e., in high-LMP areas. Why

(continued)

3.1.6 UPLIFT PAID BY SCHEDULING COORDINATORS

The total costs of CUPs in each pricing zone in each hour are recovered from SCs using a uniform \$/MWh “uplift” charge levied on some MWh measure of their price-taking load, price-taking generation, imports and/or exports in that zone and hour.

3.1.7 CONGESTION CHARGES AND CONGESTION REVENUE RIGHTS (CRRs)

Any SC that sells spot energy in one zone and buys the same amount of spot energy in another zone will automatically pay a congestion charge equal to the difference between zonal UMPs (which will be negative for counterflows from a high-UMP zone to a lower-price zone), and the same congestion charges will be imposed on self-scheduled interzonal flows. Congestion revenue rights (CRRs) will be available to hedge these interzonal transactions.

3.1.8 PRINCIPAL DESIGN ISSUES RAISED BY THE TAPAS APPROACH

There are three two principal issues raised by the TAPAS approach that are discussed in this paper: (1) the logic and implications of the decision not to pay CDPs; (2) how to compute the energy UMPs and the A/S prices; and (3) how to compute and make CUPs when the ISO uses intertemporal optimization of energy and A/S simultaneously and uses a two-settlement system. However these principal issues are resolved, there will be more detailed issues such as the allocation of uplift costs and market power mitigation that are touched on but not analyzed in detail in this paper.

3.2 THE TAPAS PROPOSAL COMPARED TO OTHER UMP MARKETS

Most UMP markets were designed from the beginning to be just that, and developed procedures and processes that, while conceptually inferior to and operationally less efficient than those used in a LMP market, were suited for the purposes of a UMP market. But the TAPAS approach keeps all the procedures and processes developed in the MRTU process for a sophisticated LMP market, including security-constrained unit commitment/dispatch (SCUC/SCED) both based on intertemporal optimization (IO), simultaneous optimization of energy, A/S and congestion management, and a two-settlement system, and then “simply” uses UMPs for settlement. Unfortunately, it may not be so simple to bolt a UMP façade onto the complex MRTU LMP machinery without something falling into the gears.

3.2.1 THE ISO IN OTHER UMP MARKETS

As discussed in section 2.2.10 above, in existing UMP markets the ISO uses some combination of (1) a myopic and/or heuristic central bid-based dispatch with simple hour-to-hour intertemporal constraints such as ramping limits, and (2) contract schedules with inc/dec bids to manage imbalances and congestion. It is largely up to each market participant to decide for itself how it wants to operate based on its own forecasts of the UMPs and its own

would a load offer to reduce its demand if the price exceeds (say) \$500/MWh if by doing so it increases the risk that it will be curtailed without compensation when the LMP exceeds \$500/MWh even if the UMP is only \$100/MWh?

bilateral or exchange-based contracts, and then to submit bids and/or self-schedules reflecting what it wants to do. The ISO is not primarily concerned with finding and implementing a truly least-cost unit commitment or forward schedule, but with using the inc/dec bids to find a secure real-time dispatch that is reasonably consistent with the self-schedules and bids given the UMPs.

Most importantly for the issues here, existing UMP-market ISOs do not decide when specific resources should start up, shut down or use limited energy, do not optimize energy, A/S and congestion simultaneously, and do not use a two-settlement system. The limited scope of ISO decisions both limits the commercial impact of those decisions and makes it relatively easy to determine what those impacts are so that the ISO can pay compensation – which most UMP-market ISOs do, whether those decisions require a resource to produce or consume more or less than it would do on its own given the UMPs. Such a UMP market is arguably inefficient because the ISO does so little, and cannot last long unless real-time congestion is kept small as evidenced by the fact that all UMP markets have aggressive investment and contracting programs designed to accomplish just that. But such a market can “work” as long as congestion is kept small and the inefficiencies are tolerated, even though – or perhaps because – the ISO does not strive for sophistication or efficiency.

3.2.2 THE CAISO IN THE TAPAS APPROACH

Things are very different in the TAPAS approach, where the CAISO uses sophisticated SCUC/SCED processes based on IO to tell market participants when to start up, when to shut down and when to use limited energy in both a DA forward market and a RT market. The resulting schedules and dispatch instructions will often require individual market participants to do things that are very different, and will have very different commercial results, from what those market participants would do to optimize their commercial results given the UMPs and A/S prices used for settlements.

It is true that the CAISO would use the same SCUC/SCED processes to find the same schedules in the full LMP market defined by the MRTU process, but with a huge difference: when the LMPs associated with those schedules are used for settlements, the CAISO schedules do not impose commercial costs on market participants but essentially tell each market participant what it should do to optimize its commercial results given the LMPs. Market participants may not like the LMPs, but they cannot blame the CAISO for those LMPs or for their own inability to predict or failure to hedge them. Furthermore, to the extent that the CAISO schedules do not optimize commercial results given the LMPs – i.e., when these schedules commit a resource that does not recover its costs from LMPs over a day (or so) – the CAISO makes what are in effect CUPs equal to the difference.

Thus, the MRTU procedures accept the principle that the CAISO should compensate those who incur costs when they follow CAISO schedules. If such compensation is not a big issue when LMPs are used for settlement, it is because in that case the CAISO seldom imposes constrained-up costs (and when it does so finds it easy to estimate and pay the required compensation) and virtually never imposes constrained-down costs (because any resource that is not committed or scheduled in the CAISO procedures would not make money at the LMPs

anyway). But in the TAPAS UMP market, compensation for the commercial effects of its decisions will be a big issue for several reasons: (1) CAISO schedules will often impose significant constrained-up and constrained-down costs on individual resources; (2) estimating these costs will require processes similar in complexity to the process used to determine the schedules; and (2) if ISO-imposed costs are not compensated market participants will face large risks and will have strong incentives to take defensive actions.

The fact that the TAPAS approach is very different from other UMP markets does not imply that it is impossible or unwise to use it during a transition to a full LMP market. But it does suggest that implementing the TAPAS approach is likely to require policies and procedures significantly different from those used in other UMP markets or contemplated when the TAPAS approach was first proposed.

4. A TAPAS-TYPE UMP MARKET IN THEORY

This section discusses the principal conceptual issues raised by the TAPAS approach. The initial subsections use examples and illustrations that oversimplify the situation but that are useful for discussing issues and illustrating concepts. The later sections outline how such things as intertemporal constraints, multi-step IO and a two-settlement system complicate implementation of those concepts.

It might seem natural to start a discussion of UMP market concepts with the question of how the UMP is determined. But the biggest issues raised by UMP markets in general and the TAPAS approach in particular concern the whether and how of CUPs and especially CDPs, which has implications for many other things, including how the UMP should be determined. Thus, the analysis here begins with the issues surrounding CUPs and CDPs.

4.1 ISO COMPENSATION PAYMENTS

The need for and value of CUPs in a UMP market are widely recognized, but CDPs are often strongly criticized and sometimes rejected as unnecessary, illogical and distorting. This section summarizes some general principles relevant to CUPs and CDPs, and concludes that the second-best solution is to include both, along with other, complementary measures that can reduce some of the longer-term effects of the combination of a UMP, CUPs and CDPs. A decision to exclude CDPs amounts to a hard constraint on the second-best solution, creating what might be called a “third-best” situation.

4.1.1 CUPS AND CDPs AS SECOND-BEST PRICE CORRECTIONS

If market prices – or, for that matter, the entire market – are to mean much, market participants must be expected, allowed and even encouraged to respond to prices in ways that best advance their individual commercial interests. Markets should be designed and operated to ensure that their prices and other payments are incentive compatible with reliable and efficient short-run system operations and with long-term system development, so that market participants can be left free to respond to these prices and other payments because when they do so their actions will be good for the system as well as for themselves. To the extent that a hard constraint on settlement energy prices make these prices incentive incompatible with system needs, the second-best solution is usually to add supplementary markets and/or payments to improve incentive compatibility so that administrative rules become back-up mechanisms that are usually easy to enforce rather than primary motivators that are strongly resisted.

In a UMP energy market the ISO must manage congestion by scheduling some market participants to produce or consume more or less energy (or A/S) than the amounts that would optimize their commercial positions given the UMP (and A/S prices). An ISO that can order individual market participants to follow its instructions and eat their own out-of-pocket or opportunity costs is not a neutral market operator but is instead a major and disruptive force in the market. Such an ISO must issue and somehow enforce instructions that, no matter how

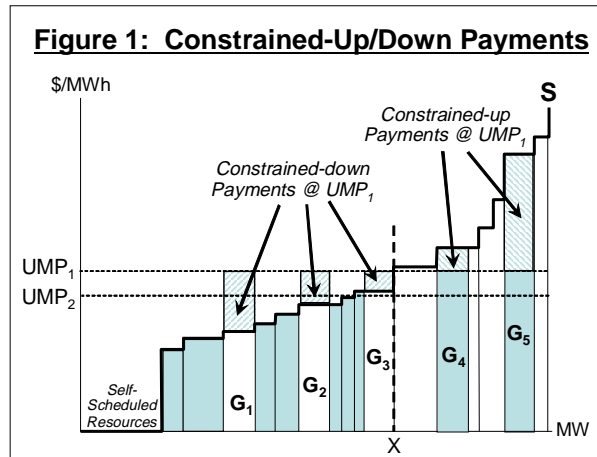
objective and logical they may seem to the ISO, will be unpredictable and risky for, and will look arbitrary and discriminatory to, those affected by them.²² Market participants will have no effective way to predict, manage or hedge the risks of ISO actions. The rational response, even for very competitive participants, will be to bid strategically to try to shift congestion costs to others, to find “good” reasons for not following instructions, to challenge the ISO at every opportunity, and to require higher risk premiums to stay in or enter the business.

Instead of relying purely on administrative penalties to enforce compliance with its schedules, the ISO can and should supplement the UMP with compensation for any out-of-pocket or opportunity costs caused by complying. Such compensation should be paid, not primarily because it is fair or because market participants have some inherent right to it, but because it makes total ISO payments in a UMP market incentive compatible with short-run system needs; it rewards, or at least eliminates any penalty for, bidding true costs and capabilities and then responding to dispatch instructions whatever these may be, and hence is in the interest of the system and ultimately of consumers. The combination of a UMP and such compensation is not a perfect substitute for LMPs, and in particular creates some longer-run problems, but second-best considerations suggest that such problems should be dealt with directly in other ways, not by sacrificing short-run incentive compatibility.

4.1.2 UMP-MARKET SCHEDULES AND CUPS/CDPS

Leave aside for now the question of how the UMP is determined or what its value is. Given any UMP and the market bids from resources, it is possible to determine an unconstrained “UMP-market schedule,” defined here as the schedule of energy quantities that would optimize the commercial position of each resource given the UMP in the absence of transmission constraints. This is illustrated in Figure 1, for the simple case in which each hour can be dealt with in isolation and energy is not optimized simultaneously with A/S.

In Figure 1, the (unconstrained) supply curve S consists of all market supply bids without regard to location on the grid, including dispatchable demand as negative supply but excluding resources scheduled to provide A/S



²² For example, suppose generator A at node A bids its \$30/MWh marginal costs (MC) while generator B at nearby node B bids its \$31/MWh MC, and it turns out that LMP_A = \$29/MWh, LMP_B = \$32/MWh and UMP = \$40/MWh. How will the ISO explain its efficient but apparently discriminatory decision to schedule the higher-bidding generator B to produce and make \$9/MWh while telling the lower-bidding generator A that it cannot produce and make \$10/MWh? Once A understands the logic, how will it bid the next time this situation is likely to arise? Then how will B respond? Then what happens? There is no way to predict behavior in such a game, but the result is unlike to be efficient, stable or tolerable for long.

instead of energy. When the UMP has the value UMP_1 , the UMP-market schedule consists of all resources with bid prices less than or equal to UMP_1 – those to the left of X – and excludes all resources with higher bid prices – those to the right of X. Notice that X and the UMP-market schedule depend on the value of UMP_1 and the shape of the supply curve S, and that X is not necessarily total demand; a UMP-market schedule in the sense used here can be defined for any arbitrary value of the UMP.

When the ISO uses its physical scheduling process to determine a schedule that minimizes total as-bid costs subject to transmission constraints, it will usually find that this “ISO schedule” requires some market participants to produce more and some to produce less than they would produce in the UMP-market schedule. Figure 1 illustrates a situation in which, at UMP_1 , all of G_1 and parts of G_2 and G_3 are constrained down to produce less than they would in the UMP-market schedule, while part of G_4 and all of G_5 are constrained up to produce more than they would in the UMP-market schedule. As a result, the ISO schedule uses the shaded resources in Figure 1.

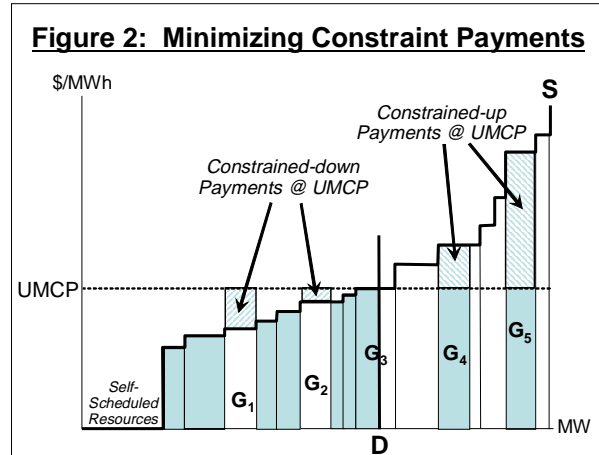
The argument in the preceding section says that an ISO should compensate those affected by its schedules, whether they are constrained up or down, so that they are indifferent between the UMP-market schedule and the ISO schedule. For both practical and conceptual reasons, such compensation amounts (or even if CUPs but not CDPs are made) must be determined automatically within the ISO’s settlement system. In general, as discussed in sections 4.4 and 4.5 below, this must be done by computing the operating profit each resource would make over the entire “day” from both energy and A/S under its UMP-market schedule – called here the “UMP-market operating profit” – and subtract the operating profit for that resource under its constrained energy and A/S schedule over the day – the “ISO-schedule operating profit.” In the simple single-hour, single-product case illustrated in Figure 1, the differences between UMP-market operating profit and ISO-schedule operating profit, and hence the required CUPs and CDPs, are just the areas of the cross-hatched rectangles in Figure 1.

It is important to recognize that, where short-run incentives are concerned, there is no commercial or economic difference between constrained-up and constrained-down costs: both reduce the amount of money a resource earns over the year to (try to) cover its fixed operating costs, pay interests and taxes, and earn some profit for shareholders; and both must be compensated by the ISO if incentive compatibility between total ISO payments and the efficient constrained schedule is to be maintained. A constrained-up generator may have to find hard cash then and there to pay fuel costs that are not covered by the UMP; but a constrained-down generator may have to find hard cash then and there to pay for replacement energy needed to cover its contract obligations. There is no logical or practical basis for the common view that out-of-pocket costs are somehow more real or immediate than “mere” opportunity costs.

In Figure 1, more resources are constrained down than are constrained up and CUPs appear to be roughly equal to CDPs. But this is purely because of the assumed shape of the supply curve and the arbitrary value chosen for UMP_1 . At lower values of the UMP, more resources would be constrained up while fewer would be constrained down, and CDPs would be lower

while CUPs would be higher; for example, if the UMP is UMP_2 , none of resource G_3 is constrained down but some of G_3 is constrained up. Conversely, at higher values of the UMP there are more constrained-down and fewer constrained-up resources, lower CUPs and higher CDPs (if these are paid).

An intuitively appealing and commonly used way to determine the UMP is to clear a hypothetical unconstrained market to determine the “unconstrained market-clearing price” (UMCP). This is illustrated in Figure 2, where D is total system price-taking demand (plus losses) on the system. Total supply in the hypothetical unconstrained market equals demand D when the UMP equals UMCP; at this value of the UMP, resource G_3 (hypothetically) provides marginal amounts of energy with no need for either constrained-up or constrained-down payments.



An important feature of the UMCP is that, if the ISO makes both CUPs and CDPs, the total of such payments is minimized when the UMP is set at UMCP. This is because when the $UMP = UMCP$ the total amount of constrained-down energy equals the total amount of constrained-up energy (because the actual constrained schedule and the hypothetical UMP-market schedule must meet the same system demand D), so small changes in UMP either way do not change total constraint payments to the resources that receive such payments when the $UMP = UMCP$, but do require either CUPs (for lower UMP) or CDPs (for higher UMP) for resource G_3 ; thus, a UMP either higher or lower than UMCP would imply higher total constraint payments.

The total costs – or the “redispatch” costs²³ – caused by congestion is most naturally defined as the total cost of the actual, constrained schedule less the total cost of a hypothetical unconstrained schedule. The ISO’s total constraint payments are a good estimate of these total congestion costs if, but only if, the ISO makes both CUPs and CDPs as above and the UMP is set at UMCP (and market offers reflect real costs). Unless both of these conditions are met, the ISO’s constraint payments will contain some economic rents and/or will not include some costs and hence may either over- or underestimate total congestion redispatch costs.

The facts that setting $UMP=UMCP$ minimizes the sum of CUPs plus CDPs and makes that sum a good estimate of total congestion costs do not necessarily imply that UMCP is a good

²³ The term “redispatch” reflects the fact that traditionally resources were first dispatched as though there were no congestion and then *re*dispatched to deal with congestion.

second-best UMP or that CDPs should be paid. The second-best UMP in the short-run sense is one that minimizes short-run inefficiencies, and the principal reason to make CDPs is to create incentive compatibility, not to make it easier to estimate congestion costs. It is shown in section 4.1.1, however, that if the ISO adopts the second-best policy of making both CUPs and CDPs, the second-best UMP (in the short-run sense) probably is UMCP.

4.1.3 THE EFFECTS OF NOT MAKING CDPs

If it is decided not to make CDPs to resources scheduled to produce (or consume) less than they want to given the UMP, ways must be found to offset the reduction in short-run incentive compatibility. Obviously, it will be necessary to penalize energy produced (consumed) in excess of scheduled amounts outside some tolerance band – except under emergency conditions when such production (consumption) is valuable. But the most important problems created by not making CDPs in a UMP market are more subtle than blatant deviations from schedules. Without CDPs, all rational resources, even very competitive ones, have strong incentives to use strategic bidding and self-scheduling to try to get into the ISO's constrained schedules, and these incentives must be countered somehow.

The effect of CDPs on the bidding incentives of competitive resources is frequently misunderstood. It is true that making CDPs creates incentives for some constrained-down resources to bid down in order to increase their CDP amounts; but this incentive is limited to resources that have local market power. Obviously, a generator in a location where it must be constrained down and will collect CDPs equal to the UMP minus its own bid price has incentives to submit the lowest bid prices it can get away with. But this incentive to bid down to increase CDPs is perfectly analogous to the incentive a critical generator in a load pocket has to bid up; in both cases, the incentive declines with the extent of local competition. If there are several generators in the generation pocket, only some of whom need to be constrained down, a generator that bids down risks bidding below its competitors and being scheduled to run, in which case it makes only the same difference between the UMP and its own incremental costs it could make by not bidding down at all. If there are a small number of generators competing for CDPs a complex oligopoly game will result, but if there are many generators competing for CDPs the best any of them can do is bid its own marginal costs and wait to see what happens.

In the absence of CDPs, the incentive to bid down is not limited to resources with local market power but exists for all rational players. Even a very small, highly competitive generator knows it will make nothing unless it gets into the ISO's constrained schedule, so it has nothing to lose and much to gain by bidding the lowest price it can get away with. In fact, a smaller generator has more incentive to bid down, because it is less likely to depress the UMP. Better yet, if allowed to do so all such generators (particularly small ones) will self-schedule so that the ISO must take their energy and pay them the UMP – the strategy that caused the initial PJM UMP market to collapse within hours when congestion first appeared. Thus, while a UMP market may require market power mitigation procedures to begin monitoring and mitigating bids that are too low as well as bids that are too high, only the lack of CDPs might require such procedures to be extended to all resources, not just those with local market power.

If CDPs are not made in a UMP market, other mechanisms will be needed to offset the reduction in short-run incentive compatibility. As just mentioned, market power mitigation procedures may have to be extended to all resources, not just those with local market power. Self-scheduling by generators (and dispatchable demands) who expect to be constrained down will have to be limited, and/or special dispatch and pricing rules will be needed to establish scheduling priorities for when too many potentially constrained-off resources bid the minimum allowed. More and/or different RMR contracts may be needed to support resources that are often constrained down but are occasionally needed. Resource adequacy requirements (RARs) may have to be used to impose some restrictions on bidding down or self-scheduling. Lower values of the UMP may be desirable to reduce the quantity of constrained-down resources and their incentives to avoid or evade constrained-down instructions. A competent ISO will find ways to keep the lights on without CDPs; the question is whether the mix of *ad-hoc* fixes needed to deal with the lack of CDPs is really better than living with CDPs.

A market without CDPs does reduce the incentives that resources – both generators and dispatchable demands – have to continue operating or to locate where they are frequently constrained down, but does not eliminate these incentives. Even the few UMP markets that do not make CDPs (e.g., Alberta and Australia) use locational grid access charges to offset the effects of too-high UMPs in such areas.

4.1.4 CUPS, CDPs AND UMP-MARKET TRANSMISSION RIGHTS

It is generally agreed that participants in a competitive electricity market should be able to get rights to use the system that protect them from the risks of congestion, including the risks of unpredictable and apparently arbitrary and discriminatory actions by the ISO. Market participants should be able to get such rights, not because they have some inherent right to them, but because logic and experience demonstrate that if such rights are not available the risks of doing business can be large, reducing investment and increasing costs to final consumers in the long run. But transmission rights are valuable, so those who get them should pay for them one way or another, unless they have a preexisting explicit or implied contractual right to use the grid without paying congestion costs.

If an ISO in a UMP market issues and (tries to) enforce its constrained-up/down instructions without compensation, individual market participants are fully exposed to the risks of real-time congestion and the unpredictable actions the ISO will have to take to manage it. Because the combination of CUPS and CDPs largely eliminates these risks in a UMP market, this combination can be thought of as a form of commercially firm transmission right. The transmission rights inherent in a UMP-CUP-CDP market are a distant second-best alternative to CRRs in a LMP market, but at least protect market participants from the risks of unpredictable and seemingly arbitrary ISO congestion management actions.

The commercially firm transmission rights implicit in the combination of a UMP and both CUPS and CDPs are commercially valuable, particularly for generators that are often constrained down. In a LMP market, such generators would often experience low LMPs and/or pay high prices for CRRs from their location to higher-LMP locations. It is both fair

and efficient, therefore, that such generators, or at least new ones, be required to pay for the valuable transmission rights implicit in CDPs.

In principle, the amount a resource in a UMP-CUP-CDP market should pay each (say) year for the transmission rights implicit in such a market is the expected value of the increase in total system congestion costs caused by the existence of that resource at that location over the year, as measured by the increase in the ISO's total CUP and CDP payments.²⁴ For a resource located where it or other resources are frequently constrained down, these incremental congestion costs are at least as large as the total CDPs paid to that resource and may be much larger; for example, a resource that using bidding-down or self-scheduling to avoid being constrained off itself simply shifts the ISO's CDPs to other resources. It is fair and efficient, therefore, that a resource located where its existence results in additional CDP payments by the ISO pay an annual transmission rights charge equal to the expected increase in those CDPs over the year.

For constrained-up resources the situation is different. The transmission rights implicit in CUPs are valuable to a resource in a UMP market if otherwise it would have to respond to constrained-up instructions without compensation. But for a resource that is frequently constrained up, a UMP market even with CUPs is no better and probably worse than a LMP market. Thus, if a UMP market is to be implemented instead of a LMP market, resources should not have to pay much or anything for the right to CUPs.

A resource receiving CUPs in a UMP market is not adding to congestion costs by existing where it is and in fact may be reducing them, in the sense that if the resource were not there the ISO would have to do something at least as costly as scheduling this resource and making a CUP to it. CUPs make the total compensation to a constrained-up resource closer to but no greater than the real value of that resource to the system. Because resources receiving CUPs are being paid no more and perhaps less than the benefits they provide to the system, they should not be required to pay anything for the right to receive such CUPs.

4.1.5 CDP ACCESS RIGHT/CHARGES

Combining the above analyses of CUPs and CDPs leads to the conclusion that, in concept, each resource (or at least new ones) in a UMP-CUP-CDP market should, in exchange for the firm transmission rights implicit in such a market, make (e.g.) annual payments equal to the incremental CDPs the ISO makes to that resource and all others over each year because that resource is where it is. The most natural form of such payments is a grid access charge (or adjustment to an existing grid access charge used to recover fixed grid costs) that is different for each type of resource at each location on the grid. Because such access charges are for the right to access the grid, to sell at UMP and to be paid CDPs when constrained off, call them "CDP access charges" and call the related rights "CDP access rights."

²⁴ It is shown above that the total of CUPs and CDPs is a good measure of total congestion costs only when the UMP is set at the UMCP. But even when the UMP is set some other way, differences in total constraint payments will approximate differences in total congestion costs.

Valuing CDP access rights and assessing the resulting CDP access charges would not be easy and in practice could be done only approximately – just like all other non-LMP charges related to transmission in any electricity market. But the ISO could determine annual \$/MW CDP access charges that vary by zones and resource type; the CDP access charges would be lower for peakers and in load pockets but higher for baseload plants and in generation pockets. The CDP access charges could be determined prospectively for each year based on expected values or retrospectively based on actual outcomes, although the former option is probably better.

Preexisting resources that have explicit or implicit transmission rights in the grid at the time the new market arrangements are implemented could be exempted from CDP access charges for some years, just as such resources can be allocated free CRRs in a LMP market. But any such grandfathered exemptions should be limited in term, so that resources do not stay in business when it would be more economic for them to retire and make room for others. This tendency for grandfathered transmission rights to keep uneconomic resources in business need not be a serious problem with CRRs in a LMP market because a retiring resource can sell its CRRs,²⁵ but can be a major problem with CDP access rights because these cannot really be transferred except in special circumstances, such as when a new resource replaces a similar old one at the same location.

CDP access rights should not be regarded as the right for an individual generator to receive CDPs when constrained down, but as the right to operate and sell at the UMP where doing so increases CDPs no matter who receives them. And the CDP access rights cannot be optional, in the sense that an individual generator willing to forego CDPs when constrained down could decline to pay for them. If CDP access rights were optional, resources that could avoid constrained-down instructions and thereby shift CDPs to others would not buy the rights. Resources (at least new ones) located where they add to CDPs should be required to pay for access to the UMP even if they do not expect to collect much in CDPs themselves.

Many practical problems would have to be solved to convert the concept of CDP access rights and charges into a practical proposal, but these problems are not necessarily more difficult than the problems that will arise for a UMP market in the long run if something very like CDP access rights/charges is not implemented. A decision to replace LMPs with a UMP has many logical and practical implications that will have to be dealt with one way or another. It is not possible to change only one thing in a complex market.

It is worth noting that the England and Wales market described in section 2.2.2, one of the few long-lasting UMP markets in the world, makes what are in effect CUPs and CDPs (although no longer called that) and imposes grid access charges that vary by location in ways that reflect the expected value of CDPs at different locations (although not estimated that way). NGC's grid access charges for generation (load) at various locations are based on

²⁵ However, the same problem will arise with CRRs in a LMP market if grandfathered resources get free CRRs (say) annually only as long as they stay in business.

estimates of the cost of expanding a hypothetical optimal grid to handle an increment of generation (load) there; because the incremental cost of grid capacity should equal incremental constraint costs at each location on an optimal grid and NGC is responsible for grid planning and investment, locational differences in grid access charges can be interpreted as reflecting, at least roughly, locational differences in expected constraint costs.

4.1.6 LONG-RUN EFFECTS OF A UMP MARKET

The combination of CUPs and CDPs is the second-best way to minimize the short-run inefficiencies inherent in a UMP market, but will produce long-run investment and locational incentives different from those in a first-best LMP market. The second-best solution to this problem is to find and implement additional mechanisms that can improve long-run incentives without distorting the short-run incentives resulting from the UMP-CUP-CDP combination.

The most serious long-run problem caused by the UMP-CUP-CDP combination is the incentive it creates for resources to remain in or enter regions where resources are frequently constrained down. This problem is usually ascribed to, and is clearly made worse by, CDPs. But CDPs give constrained-down resources only what they would get selling at the UMP; the real problem in constrained-down regions is that the UMP is higher than the LMPs on average. Even in the absence of CDPs, low-running-cost resources that will not be constrained down much have too much incentive to enter a constrained-down region, and something must be done to deal with this problem. ISOs that do not make CDPs – i.e., in Alberta and Australia – have realized this and have adopted other mechanisms to improve long-run incentives.

The most logical solution to the long-run incentive problems caused by the UMP-CUP-CDP combination is to require all resources, at least new ones, to pay something like the CDP access charges discussed above. This would effectively reduce a resource's net long-run revenue to what it would get in a LMP market, thereby eliminating any uneconomic incentive to remain or begin operating in a constrained-down regions, but without reducing incentive compatibility with short-run dispatch instructions.

If the second-best option of imposing something akin to CDP access charges is not acceptable, it may be necessary to adopt a third-best option that accepts higher short-run inefficiencies in order to reduce longer-term problems. The most obvious and common such option is to eliminate CDPs altogether; some of the effects of this option are discussed in section 4.1.3.

If a UMP market is being considered only as a transition to a LMP market, as is said to be the case for the TAPAS approach in California, the long-run effects of the UMPs and any CUPs or CDPs should be less important than the short-run effects. This suggests that a transitional UMP market should include CDPs to improve the efficiency of dispatch even if something like CDP access rights are not implemented. New generators would presumably not locate in regions where they expect LMPs to be low just because they could collect CDPs for a few years.

4.2 DETERMINING A SECOND-BEST UMP

The decision to use a UMP for settlements is precisely the kind of hard constraint on pricing that creates a second-best situation, where there are few principles except high-level generalities such as “try to minimize the inevitable inefficiencies.” Choosing a second-best solution in the presence of such a hard constraint is ultimately a matter of listing all the plausible alternatives, trying to predict and estimate the inefficiencies each will cause, and then using judgment to pick one that arguably creates less, or at least no more, inefficiency than the others.

The CAISO has proposed two options for setting the UMP: UMP Option 1: $UMP = UMCP$; and UMP Option 2: $UMP =$ a weighted average of the LMPs. Before analyzing these specific options it is worth trying to develop the general second-best objective – “minimize inefficiencies” – into something more specific.

4.2.1 CHOOSING UMP TO MINIMIZE SHORT-RUN INEFFICIENCIES

Because the UMP is used for settling short-run transactions and other mechanisms will be needed to deal with the longer-run problems no matter what the precise level of the UMP is, second-best principles suggest that the UMP should be chosen to minimize short-run inefficiencies. General economic reasoning can be used to develop this high-level proposition into some more specific recommendations regarding the UMP.

The most important short-run inefficiencies caused by a UMP arise from the fact that any value of the UMP will differ from the LMP at many/most nodes, which gives those who pay or are paid the UMP incentives to consume or produce more or less than they would if they paid or were paid the LMP at their node. Unless these short-run incentives are offset by other short-run payments or charges, dispatchable resources will adopt strategies to influence the schedules they get back from the ISO and/or to let them deviate from their schedules without being caught or penalized. Price-taking loads (and any price-taking generation) who pay the UMP instead of LMPs and simply respond to the UMP as they choose will consume (or produce) more or less than the efficient amounts corresponding to the LMP.

The ISO can and should try to offset the short-run inefficiencies created by a UMP with a combination of second-best mechanisms. The most important such mechanisms are the CUPs and CDPs that the ISO makes – or not – to dispatchable resources. The precise level of the UMP itself is not as critical as the ISO’s CUP and CDP policies, but can be important, particularly if the ISO’s CUP/CDP policies do not effectively correct the inefficient incentives inherent in a UMP. Thus, the second-best value of the UMP for dispatchable resources depends on the ISO’s CUP/CDP policies.

Consider first the case in which the ISO makes neither CUPs nor CDPs to dispatchable resources. (See Appendix A for more details of the argument summarized here.) In this case, the incentive for a dispatchable resource at any node to try to affect its ISO schedules or to deviate from them once it gets them is proportional to the UMP-LMP difference at that node. General economic and optimization principles suggest that the expected dispatch inefficiencies at a node are roughly proportional to the *square* of the UMP-LMP difference at that node (for

differences that are “not too large”) and to the amount of dispatchable generation or dispatchable demand at that node. Thus, the total dispatch inefficiencies associated with any UMP can be approximated by a weighted sum of squared UMP-LMP differentials over all nodes, i.e.:

$$\text{The Total Dispatch Inefficiencies Caused by any UMP} = \sum_j G_j \times (\text{UMP} - \text{LMP}_j)^2$$

where G_j is the amount of dispatchable generation and dispatchable demand at node j , LMP_j is the LMP at node j and the sum is over all nodes. The second-best UMP is one that minimizes the sum in the expression above, which leads to the conclusion that the second-best UMP is the weighted arithmetic average of the LMPs with the G_j as weights, i.e.

$$\text{Second-Best UMP for Dispatchable Resources} = \text{LMP}_{\text{GAV}} \equiv (\sum_j G_j \times \text{LMP}_j) / (\sum_j G_j).$$

It is hardly revolutionary to suggest that the UMP should be a weighted average of LMPs; this is one of the options suggested by the CAISO, and a (load-weighted) average of LMPs is used to settle load in some markets that calculate LMPs. But it is worthwhile to realize that this definition of the UMP can be derived from general principles and – more importantly – when and why is the second-best UMP: *when* the ISO makes neither CUPs nor CDPs and hence resources have incentives to manipulate or deviate from schedules that increase with UMP-LMP differences; and *because* a UMP equal to a weighted average of LMPs minimizes the average of such UMP-LMP differences. But this result raises an obvious question that is addressed next: Is a weighted-average of LMPs the second-best UMP under other compensation policies, and if not, what is?

Suppose the ISO makes both CUPs and CDPs designed to compensate resources fully for any commercial costs incurred in following instructions, as described in section 4.1.2 above. In this case, UMP-LMP differences create no significant incentives for dispatchable resources to try to influence or deviate from their schedules, because any resource constrained either up or down is compensated for any commercial costs it incurs in responding. Compensation will never be perfect, but as long as the ISO tries to provide full compensation any errors will be small, random and unlikely to have any systematic effects on operational incentives or efficiencies, so the precise level of the UMP will (within broad limits) have no effect on dispatch inefficiencies. But if the ISO makes both CUPs and CDPs, total ISO payments will be large, and any such payments create generalized administrative costs. Such administrative costs are roughly proportional to the total amount of ISO payments, without regard to the precise level of the UMP or the distribution of LMPs, suggesting that the UMP should be chosen to minimize total ISO constraint payments. It is shown in section 4.1.2 that the UMP that minimizes total payments when both CUPs and CDPs are paid is the UMCP that clears the hypothetical unconstrained market. Thus, the second-best UMP when the ISO makes both CUPs and CDPs is the UMCP.

Again, it is hardly revolutionary to suggest that the UMP should equal the UMCP; that is the case in most UMP markets. But, again, it is worthwhile to realize when and why the UMCP can be interpreted as the second-best UMP: *when* the ISO makes both CUPs and CDPs; and *because* this policy creates incentive compatibility for any value of the UMP (within broad limits) so the UMP might as well be chosen to minimize total compensation payments and the

generalized inefficiencies such payments always create, which is what a UMP equal to UMCP does.

Consider next the case where the ISO makes CUPs but not CDPs. In this case, a resource that expects to be constrained up has little incentive to do anything except submit cost-reflective bids and follow instructions, because it will be compensated for any commercial effects of responding to constrained-up instructions. But a resource that is likely to be constrained down has strong incentives to try to avoid or evade constrained-down instructions because it must eat all costs caused by complying with them. Given these very different incentives on constrained-up and constrained-down resources, lower UMPs (up to a point) decrease dispatch inefficiencies for constrained-down resources without having much effect on other resources. Taking this argument to its logical conclusion implies that, when CUPs but not CDPs are paid, the UMP that minimizes dispatch inefficiencies is the lowest bid price of any constrained-down resource, call this LMP_{CD} . If the UMP is equal to LMP_{CD} , no resources will be constrained down and many will be constrained up.

Although a UMP equal to LMP_{CD} may minimize dispatch inefficiencies when CUPs but not CDPs are made, it will result in large CUPs that will produce the generalized administrative inefficiencies that increase with the size of total ISO payments. This suggests that the UMP that minimizes the sum of short-run dispatch and administrative inefficiencies may be something greater than LMP_{CD} . But the marginal dispatch inefficiencies caused by increasing UMP-LMP differences are probably greater than the marginal administrative inefficiencies caused by increasing the total size of constraint payments, so the short-run-second-best UMP in this case is probably not much above LMP_{CD} , and is almost surely well below UMCP.

It may seem bizarre to suggest that, when CUPs but not CDPs are made, the UMP should be so low that no resource is constrained down while many resources are constrained up. But the ISO-NE's initial UMP market used a UMP that had almost this effect (See section 2.2.4.) and, at least partly as a result, was able to enforce constrained-down instructions even though it did not make CDPs; this was in sharp contrast to the experience in PJM's abortive UMP market with no CDPs and a UMP equal to UMCP, which collapsed within hours when congestion appeared. If CDPs are not made under the TAPAS approach, the option of setting a low UMP should be seriously considered.

The efficiency arguments above regarding dispatch inefficiencies apply in a slightly different way to consumption inefficiencies caused by differences between LMPs and the UMP paid by price-taking load (and nondispatchable generation) that responds to the UMP or not as it chooses: A UMP creates incentives for price-taking load to consume too much or too little, depending on whether the UMP is lower or higher, respectively, than the LMP at that load's node. But a price-taking load does not respond to those incentives by trying to affect or deviate from an ISO schedule; instead, it just responds to the UMP to an extent measured by the price-elasticity of its demand. This suggests that the expected inefficiency at each node is proportional to the square of the UMP-LMP difference, and – following the same logic used for dispatchable resources – the UMP that minimizes the sum of such inefficiencies over all nodes is a weighted average of the LMPs, with the weights equal to the amount of price-

taking load at each node multiplied by the price-elasticity of demand for that load. On the assumption that it is not practical to measure nodal price-elasticities of demand or they do not differ much, the second-best UMP for load is:

$$\text{Second-Best UMP for Loads} \equiv \text{LMP}_{\text{LAV}} = (\sum_j L_j \times \text{LMP}_j) / (\sum_j L_j)$$

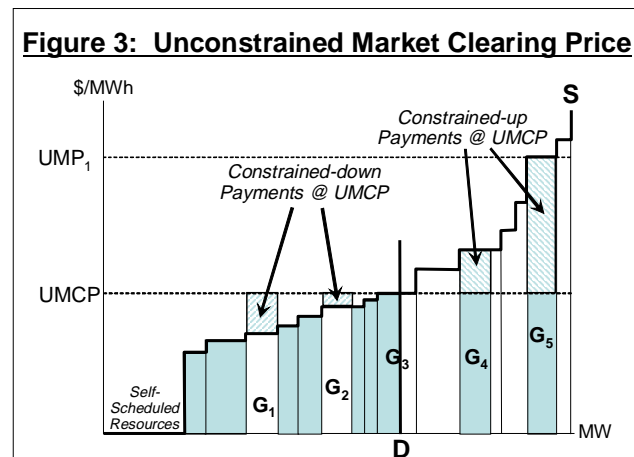
where L_j is the amount of price-taking load (and nondispatchable generation) at node j . If it is possible to estimate nodal price elasticities and they differ significantly, in principle they should be taken into account; for example, a node that contains primarily price-responsive but nondispatched industrial load should have a larger weight in the above calculation.

Notice that the arguments regarding the second-best UMP for price-taking load do not depend on whether or what compensation payments the ISO makes to generators and dispatchable demand, because such payments have no direct effect on price-taking load. This suggests that the second-best UMP for load is LMP_{LAV} no matter how the UMP is determined for generation.

4.2.2 UMP OPTION 1: UMP = UMCP

UMP Option 1 is the most common in UMP markets: Set UMP at the UMCP that clears a hypothetical unconstrained energy market. It is shown in the preceding section that UMCP is the second-best UMP if the ISO makes both CUPs and CDPs, but the reasons for this result have little to do with the fact that UMCP clears a hypothetical unconstrained market. The only “magic” in the UMCP is that this is the value of the UMP that minimizes total constraint payments when both CUPs and CDPs are made, and when both CUPs and CDPs are made the precise value of the UMP (within broad limits) has no effect on dispatch efficiency, so the UMP should be chosen to minimize total constraint payments and the general administrative inefficiencies associated with these, i.e., the UMP should be the UMCP.

Figure 3 illustrates how the UMCP is determined when there is only one pricing zone, when each hour can be considered in isolation and when energy and A/S are not jointly optimized. Supply curve S includes all energy resources (i.e., excluding resources scheduled at an earlier stage to provide A/S) without regard to location; D is total system demand plus losses; and UMCP is the price that equates total supply to total demand. UMCP is also the UMP that minimizes total constraint payment when both CUPs and CDPs are made.



Although UMCP can be determined by inspection in this simple case, in more complex cases this is not true. The ISO must perform a separate run of the same scheduling/dispatch models it uses to determine actual operations, with the same bids from market participants and the

same demand (plus losses), but with all transmission constraints removed. If these models are based on LMPs, as the CAISO's models will be in the TAPAS approach, the result will be a set of LMPs that all have the same value (ignoring marginal losses²⁶), and this value can be used as the UMCP.

In the TAPAS approach, the CAISO uses intertemporal optimization (IO) to manage intertemporal constraints and to optimize energy and A/S simultaneously using a full network model (FNM) over the whole "day" simultaneously, and then settles energy using three UMPs, one for each zone. If the zonal UMPs are to be zonal UMCPs, the only fully logical and the most practical way to find the UMCPs (and the A/S prices and the CUPs/CDPs) given these complications is to create versions of these same models, but with the FNM replaced by a transmission model that has no intrazonal constraints and purely radial flow constraints between the zones;²⁷ processes and results based on such a model are referred to here as "three-zone-unconstrained" or simply "unconstrained" processes and results. The (three-zone) unconstrained versions of the CAISO pricing models will produce many LMPs, but all the LMPs within any zone will have the same value that can be used as the UMCP in that zone (as long as all interzonal connections are radial). Deciding how to do this, including what approximations to make and short-cuts to take, will be a serious practical challenge for the TAPAS approach, but in concept it can and must be done.

One apparent but illusory *dis*advantage of using the UMCPs for the UMPs is that determining the UMCPs requires the additional unconstrained model runs just outlined, while UMPs based on (e.g.) a weighted average of the (constrained) LMPs can be computed directly from the constrained solution. This would be a serious disadvantage of the UMCP if neither UMPs nor CDPs were to be made; but even if only CUPs are made the ISO must compute the UMP-market schedules and UMP-market operating profits for each resource no matter how the UMP is determined. In very simple cases, such as those illustrated in Figures 1-3, this might be done using a few rules – e.g., "CUP = (Bid Price – UMP) × (Constrained-Up Energy)" – but in the TAPAS commitment and scheduling processes are so complex that it is not possible to determine (e.g.) the amount of constrained-up energy in any hour by looking at that hour in isolation. As discussed in Appendix B, in this case the only logical and practical way to determine the appropriate compensation is as outlined in the preceding paragraph, even if the UMP itself is determined some other way.

²⁶ If the LMPs reflect marginal losses the marginal-loss term in the LMP formula can be ignored. This will introduce some arbitrariness, because the disaggregation of LMPs into energy, congestion and loss components depends on choice of the reference node, which is essentially arbitrary. But for a "reasonable" choice of the reference node this should not be a serious problem – particularly because the precise value of the UMP is not critical anyway. The qualifier "ignoring marginal losses" is implied where appropriate in this paper even when not explicitly stated.

²⁷ Interconnections among zones are "radial" when there is no more than one way to get from one zone to another, either within the state or through other neighboring states. When the interzonal connections are not radial, the LMPs within a zone can differ even when there are no intrazonal constraints.

Using the UMCP as the UMP is sometimes criticized on the grounds that resources that are constrained-down and hence “cannot get to market” – e.g., G_1 and part of G_2 in Figure 2 – are included in the supply curve, where they “artificially” or “inefficiently” depress the UMCP. If this criticism means that a UMCP computed this way is too low to be a second-best UMP, it has no basis without at least some argument that a higher UMP would create less inefficiency. Given the argument above for the UMCP as the second-best UMP when both CUPs and CDPs are made, and for an even lower UMP if only CUPs are made, it would be hard to argue that the second-best UMP is higher than the UMCP.

Trying to determine an *unconstrained* market clearing price that reflects the effects of *constraints* not only misses the whole point of a UMCP, but leads to the strange result that the UMCP should be the bid price of the highest-cost resource running anywhere on the system (for energy, not A/S, reasons) in the constrained dispatch. Whatever this price or whatever it is called, it would be neither an unconstrained market-clearing price or a reasonable second-best UMP. It is shown in section 4.4.3, however, that something similar to this could occur in the TAPAS approach if the UMCP is determined using only those resources that are selected in the constrained SCUC process.

UMCP is analogous to and probably close in value to the median of the distribution of the LMPs. Like the median, UMCP is somewhere “in the middle” of the LMPs, because when the UMP = UMCP the amount of constrained-up energy (where LMPs are greater than UMCP) must equal the amount of constrained-down energy (where LMPs are less than UMCP). And, like the median, UMCP is unaffected by the “skewness” of the LMP distribution, i.e., by the number or magnitude of very high or very low LMPs.

4.2.3 UMP OPTION 2: UMP = WEIGHTED AVERAGE OF LMPs

In UMP Option 2, the UMP in each zone is equal to a weighted average of the constrained LMPs within the zone. It is shown in section 4.2.1 above that the generation-weighted average of LMPs, called here LMP_{GAV} , is the second-best UMP for dispatchable generation and loads if the ISO makes neither CUPs nor CDPs, while the load(-and-elasticity)-weighted average of LMPs, called here LMP_{LAV} , is the second-best UMP for price-taking loads (and non-dispatchable generation) whatever the ISO does on the generation side.

LMP_{GAV} is higher than the UMCP if the distribution of LMPs at generator nodes is skewed toward high LMPs and is lower than the UMCP if this distribution is skewed toward low LMPs. This is what makes LMP_{GAV} a better UMP for dispatch than UMCP when neither CUPs nor CDPs are paid; intuitively, LMP_{GAV} , unlike UMCP, “follows” the generator LMPs when they are skewed, thereby reducing the average of the uncompensated UMP-LMP differences and the associated dispatch inefficiencies.

LMP_{GAV} would have a major advantage over UMCP as the UMP if neither CUPs nor CDPs were to be made, because it can be computed directly from the constrained dispatch and LMPs with no need for the additional unconstrained model runs needed to determine the UMCPs. But, as discussed in the preceding section, this advantage evaporates if CUPs are to

be made, because determining the appropriate CUPs (or CDPs, if made) requires essentially the same the unconstrained model runs no matter how the UMP itself is determined.

LMP_{LAV} is probably higher than either LMP_{GAV} or UMCP, so if load is settled at LMP_{LAV} and generation is settled at either LMP_{GAV} or UMCP the ISO settlement is likely to produce a surplus. This surplus is analogous to – in fact, is equal to, if generation is settled at LMP_{GAV} – the settlement surplus resulting from LMP differentials in a LMP market. It is appropriate and efficient that loads pay more for energy than generators receive in the presence of congestion, and the difference can be used to pay ISO constraint costs with no need to impose an uplift on SC's, with the surplus used to reduce (e.g.) transmission access charges for loads.

Basing the zonal UMPs on the average of LMPs within each zone can cause some problems if LMPs include marginal losses, because then the zonal average of LMPs can be different even if there is no congestion on the interconnections between zones. This problem can be solved by running a version of the scheduling/pricing models without losses (and adding losses to demand), by ignoring the marginal-loss component of the LMPs, or more simply by averaging over the LMPs in both (all) zones when the interconnection(s) between them is (are) not congested. The fact that these different ways of computing an average of LMPs may produce slightly different results is not important, given that the precise level of the UMP is not critical anyway.

4.2.4 RECOMMENDATIONS FOR THE SECOND-BEST UMP

The analysis above concludes that for settling price-taking loads (and any price-taking generation), the second-best zonal UMP is the load-weighted average of the constrained LMPs in the zone, LMP_{LAV} , whatever is done for dispatchable resources. This UMP is easy to compute and understand and is used to settle loads in PJM and elsewhere. Thus, it is recommended here as the second-best UMP for settling price-taking loads and generation, even though it will be higher than the UMPs recommended for dispatchable resources. The settlement surplus resulting from settling loads at LMP_{LAV} and generation at a lower UMP should be used to pay constraint costs (instead of recovering these costs from SC's through an uplift) with the balance used to reduce other charges paid by loads, such as grid access charges. If it is not acceptable to use a different UMP for loads than for generation, the generation UMP defined below can be used as a third-best UMP for loads.

For dispatchable generation and loads, the second-best UMP depends on the ISO's constraint compensation policies. If neither CUPs nor CDPs are paid, the second-best zonal UMP is the generation-weighted average of the constrained LMPs within each zone, LMP_{GAV} . But the option of paying no compensation has not been suggested under the TAPAS approach for California, so this method of determining the generation UMP is not considered further here.

If both CUPs and CDPs are made, the second-best UMP is clearly the UMCP, not only because this minimizes total CUP+CDP payments and the inefficiencies associated with these, but because the UMCP is relatively easy to understand and is widely used in UMP markets – virtually all of which make CUPs and most of which make CDPs. The difficult question is

what to do if the CAISO makes CUPs but not CDPs, which is the compensation policy currently preferred by the CAISO under the TAPAS approach.

The second-best UMP in terms of dispatch inefficiencies when CUPs but not CDPs are made is well below the UMCP; in fact, it is the lowest bid price of any constrained-down resource on the system, called here LMP_{CD} . A UMP this low may be unacceptable because it would result in large CUPs and would put a heavy burden on resource adequacy requirements (RARs) and other payments to provide adequate revenue to cover generators' costs and encourage needed investment, so a third-best option is probably needed. The third-best UMP in this case is probably the UMCP, because it is easily understood and is recommended when both CUPs and CDPs are paid. The other obvious option, LMP_{GAV} , would probably be even higher on average because LMP distributions tend to be skewed toward high LMPs. Something between UMCP and LMP_{CD} would probably be better, but there is no nonarbitrary way to define such a value.

In summary, it is recommended here that the zonal UMPs used for settling dispatchable resources should be the UMCPs determined using the (three-zone) unconstrained approach outlined above, whether the CAISO makes both CUPs and CDPs or only CUPs, and in the balance of this paper the terms "UMP" and "UMCP" are used essentially interchangeably when referring to the value of the UMP in the TAPAS approach. If only CUPs are made, however, this UMCP/UMP is too high to minimize dispatch inefficiencies, creating strong incentives even for competitive resources to use strategic bidding and self-scheduling to try to avoid being constrained down and requiring some of the third-best fixes discussed in section 4.1.3 above; if these problems are serious enough, a lower value of the UMP should be considered.

4.3 INTERZONAL CONGESTION CHARGES AND CRRS

Dividing a UMP market into three pricing zones, each with its own UMP, does not change any of the conclusions above regarding CUPs and CDPs, or the relative merits of the various UMP rules given compensation policies. The discussion of the UMP options above indicates how to compute zonal UMPs for the two viable options, i.e., a UMCP and weighted-average LMPs. The main effect of a three-zone UMP market is to create the need for interzonal congestion charges equal to the interzonal UMP differences and congestion revenue rights (CRRs) to hedge these congestion charges.

Defining CRRs and simultaneously feasible sets of CRRs is much simpler in the (three-zone) unconstrained market model used in the TAPAS approach than it would be in the full LMP market model. It will be necessary to decide how to allocate the risk that an interzonal connector might be down or that unmodeled loop flows may reduce its effective capacity, but these problems are essentially the same or harder in a full LMP market.

If zonal UMPs are based on averages of LMPs that include marginal losses, they can differ even if there is no interzonal congestion, complicating settlement of CRRs and some other matters. These problems can be dealt with by computing loss-less LMPs or by average across

both/all zones when interzonal constraints are not binding. In any case, the recommendation here is that zonal UMPs be based on a hypothetical three-zone-unconstrained market-clearing process, so these issues should not arise (unless the modeled interconnections between zones include loops).

4.4 THE EFFECTS OF INTERTEMPORAL CONSTRAINTS

The simple methods for computing CUPs/CDPs and the UMCP described above implicitly assume that the price bids and hence the supply curves for each dispatch/pricing “hour” reasonably reflect the actual options available for that hour and their economic implications. But if what happens in one hour affects the alternatives available later – for example, if the change in output from one hour to the next cannot exceed specified ramping limits, or if using limited energy or pollution allowances this hour affects costs or supply in later hours – the options available for any hour are constrained by what happened in earlier hours, and choices among the options available in any hour should consider their effects on later hours. It is generally understood that such intertemporal effects must be considered in the actual dispatch and pricing. But they must also be considered in determining the UMCP (if used as the UMP, as recommended here) and the CUPs/CDPs.

4.4.1 A “PURE MARKET” APPROACH TO INTERTEMPORAL CONSTRAINTS

The “pure market” way to manage intertemporal constraints is to allow/require each resource to do so for itself. In this approach, each resource submits to the ISO hourly bids and/or self-schedules reflecting what the resource really can do and wants to do in each hour given its current situation, its own forecasts of the future and its own intertemporal constraints. In the two-settlement TAPAS approach, the CAISO would use these bids first to determine the DA forward schedules, transactions and prices for each hour, clearing one hour at a time, and would then use updated bids to determine the actual dispatch for each hour as it came up in real time. The CAISO could deal with each hour in isolation as assumed in the simple conceptual analysis above, with no need (in principle) to adjust bids to reflect the current situation or to look ahead to consider what might happen in future hours, because market participants would have included all those factors in their bids.

Most UMP markets use not-quite-so-pure single-settlement versions of this approach. Market participants are expected to manage most of their own intertemporal constraints and to submit bids and schedules reflecting how they want to operate in each hour, and the ISO deals with each hour more or less in isolation as it determines the real-time dispatch. This approach can be logical and in principle can achieve efficiency, but requires each resource to make its own forecasts for much of the system and then to bid strategically so that it is dispatched the way it wants to be. This puts a heavy burden on market participants to get things right, makes it hard to detect and control gaming and the exercise of market power, and requires the ISO to be relaxed about the security implications of such individual market-driven decisions.

4.4.2 INTERTEMPORAL OPTIMIZATION (IO) BY THE ISO

The other logical way to manage intertemporal constraints is for the ISO to do it. In this approach, each resource submits bids that apply to the whole “day,” consisting of supply curves and/or energy limits for the day along with maximum and minimum output levels and intertemporal constraints such as start-up times and costs, ramping limits, etc. It is then up to the ISO to optimize the entire system for the whole day by considering all hours simultaneously and taking into account the fact that what happens in one hour affects what happens in other hours. The result is a daily schedule of hourly outputs/consumption for each resource and for the system, and (in a MRTU-like process) LMPs – and even A/S schedules and prices in more sophisticated versions that optimize energy and A/S simultaneously, as in the TAPAS approach.

When the ISO manages intertemporal constraints as just outlined, it would be very difficult or impossible to define hourly supply curves analogous to those illustrated in Figures 1-3 above and use them as described there to determine the UMCP and CUPs/CDPs. (See Appendix B for more discussion.) But neither is it necessary to do so, because the same multi-hour IO processes and models used by the ISO to determine transmission-constrained solutions and LMPs can be used to determine the UMP-market schedules, UMCPs and CUPs/CDPs for the entire day as a whole.

In concept, determining (three-zone) unconstrained UMCPs and compensation payments when the ISO uses IO to determine forward schedules and dispatch involves the following principal steps:

- Create (three-zone) unconstrained versions of the IO processes and models, defined as versions that have no transmission constraints within pricing zones but radial constraints between zones, and that retain resource-specific intertemporal constraints such as ramping limits and daily energy limits.
- Each time the constrained IO models are used to determine a constrained solution and the associated LMPs, run the unconstrained versions of these models to determine the UMP-market schedules and the associated LMPs. Because there are no transmission constraints within zones and all interzonal interconnections are radial, all the LMPs for each hour within each zone should be the same (ignoring marginal losses), and this common value can be used as the UMCP/UMP for that hour and zone.
- Compute the daily UMP-market operating profit for each resource, defined as the operating profit that resource would make over the day operating as called for in the UMP-market schedule given the UMCPs, and compare this to the daily operating profit that resource makes in the ISO’s constrained schedule; any difference is the amount of daily compensation necessary to make that resource indifferent between the ISO’s constrained schedule and the market schedule; this amount will include both CUPs and CDPs.

The concept of paying compensation based on outcomes over a day (or other multi-hour period) is not new: the MRTU LMP market would make what is essentially a daily CUP

when a resource is committed by the CAISO but LMPs over the day do not cover its full costs including start-up costs. In the MRTU LMP market there is no analog to constrained-down resources, because a resource that does not make the cut in the SCUC process would not make money over the day at the LMPs anyway. But in a UMP market, a resource might be rejected for commitment even though it could make money over the day given the UMPs over the day, which is a generalized version of being constrained off/down.

The use of a daily UMP-market schedule to determine daily compensation payments is logical, but raises several difficult issues, including the following:

- **Defining “Constrained Up” and “Constrained Down”:** When compensation is determined by comparing total constrained and unconstrained results over a full day, there may be no non-arbitrary way to decide how much of the total compensation represents CUPs and how much represents CDPs. The difference may be obvious for resources that are constrained only up or down during the day, but if a resource is (apparently) constrained up in some hours and down in other hours there may be no clear way to allocate the total daily amount of compensation between CUPs and CDPs. (See Appendix B for examples.) This is no problem if both CUPs and CDPs are made, but if CDPs are not made it will be necessary to develop rules and automated systems for distinguishing allowed CUPs from disallowed CDPs – which will not be easy and may require essentially arbitrary judgments with the problems these imply.
- **The Starting Positions:** If the starting position – i.e., the ending position of the previous day – of a resource is influenced by congestion in the previous day, this should be taken into account. The start-time of the dispatch day should be chosen so that this is a rare occurrence.
- **Differences Between UMP-Market and Constrained Schedules:** The unconstrained UMP-market schedules can deviate from constrained schedules in a significant and cumulative way. For example, the UMP-market schedule could imply that a resource ramps up and down several times over the day even though the constrained schedule constrains it off all day; or the UMP-market schedule could imply that a resource is off all day even though the constrained schedule ramps it up and down over the day.
- **Uninstructed Deviations:** When the process outlined above is applied to determine real-time dispatch (as opposed to forward market schedules) the possibility of noncompliance with dispatch instructions must be considered. Situations can arise in which a resource that fails to follow dispatch instructions appears to be constrained up or down by the ISO. Special rules and processes must be developed to deal with such situations.²⁸

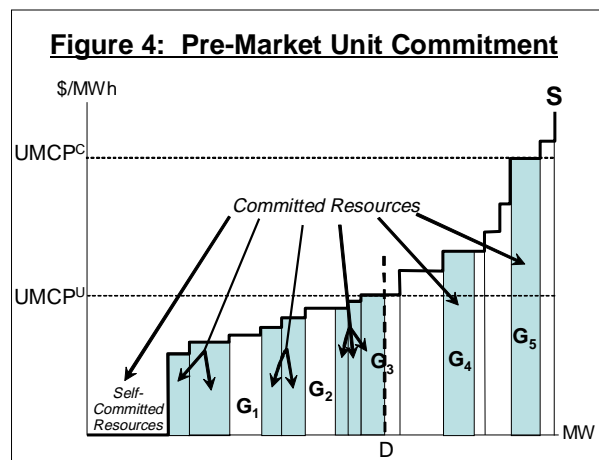
²⁸ For example, during a period when system demand and UMPs are falling both the ISO’s constrained dispatch schedule and the UMP-market schedule may call for generator G to ramp down as rapidly as it can given its down-ramp limit. If G does not ramp down as instructed in the first hour, the ISO’s actual dispatch instructions for the second hour will tell G to start ramping down, while the UMP-market schedule may assume G has already done so. In this case, the ISO

(continued)

4.4.3 MULTI-STEP IO AND UNIT COMMITMENT

The issues discussed in the previous section arise in a single-step IO process, in which the ISO uses market bids to determine schedules and prices for the whole day in a single, integrated process. In practice, however, IO processes are so complex that it is often necessary to break the problem into steps. For example, the TAPAS approach uses a security-constrained unit commitment (SCUC) model to commit enough resources to meet expected demand securely, and then only these committed resources are used in an IO-based scheduling/pricing model to determine the market-clearing schedules and the associated LMPs. This all works well when the LMPs from this process are used to settle the quantities in the schedules (and the ISO makes daily CUPs to resources that do not recover all their start-up costs from LMPs), but can create problems when settlements are based on a UMP.

The basic problem created for a UMP market by the type of two-step IO process just described is that some resources may be constrained off at the SCUC step before they ever get to the market-clearing step. This can be illustrated with the help of Figure 4, which is familiar from earlier discussions. Supply curve *S* represents all the resources that bid into the market initially; $UMCP^U$ is the price that equates unconstrained supply to demand (plus losses) *D*; resources G_1 and G_2 are constrained down; and G_4 and G_5 are constrained up. If the ISO uses a SCUC



process to decide which resources to consider in the constrained scheduling and pricing step, it will exclude many/most of the resources it expects to be constrained off as well as many/most of those resources it does not expect to clear at the $UMCP$ or to be constrained on. In fact, if the SCUC process had perfect foresight it would eliminate all except the shaded and self-committed resources in Figure 4, resulting in a “transmission-constrained unconstrained market-clearing price” $UMCP^C$ equal to the bid price of the highest-cost resource selected (for energy, not A/S reasons) in the constrained solution, e.g., a combustion turbine at the end of a long line somewhere. The result would be a $UMCP^C$ that is (perhaps much) higher than the “truly-transmission-unconstrained” market-clearing price $UMCP^U$, no constrained-down/off resources (because they would all be rejected in the SCUC step), and no constrained-up/on resources (because at this UMP all resources would be happy, even eager, to run, including many/most of those rejected by the SCUC screen).

In a LMP market, the main purpose of a SCUC process should be to identify resources that need to be constrained on for system security purposes, not to eliminate resources that might

will tell *G* to produce more than its UMP-market schedule amount, which looks like a constrained-up instruction deserving of a CUP.

want to self-commit. In Figure 4, for example, the ISO might commit the shaded resources and promise to cover their full-day costs, but allow all the others to self-commit if they want to take their chances in the LMP market. Allowing such self-commitment would create few problems in a LMP system, because resources rejected in the SCUC process would not expect to make money at the LMPs anyway and hence would probably not self-commit.

In a UMP market, however, resources that are not part of the least-cost unit commitment might be able to make money at the expected UMPs, and hence might bid strategically or self-commit and then self-schedule so that they could fit into the constrained ISO schedules. This might not affect system security directly, because the ISO's processes can always commit and then schedule enough resources to meet security standards. But strategic bidding and extensive self-commitment will reduce the efficiency and stability of operations and the UMPs, and create incentives to deviate from ISO dispatch instructions that, if serious enough, could affect system security.

It is not possible to know in advance how serious the strategic bidding and self-commitment/scheduling problem might be in practice. The strategy of lowering bid prices might be controllable with market power mitigation methods, although these methods would have to be extended to monitor and mitigate generator bids that are too low, not just too high. The self-commitment/self-scheduling problem could be more serious, requiring tight restrictions on the ability of resources to self-schedule – self-commitment is probably no problem if self-scheduling is not allowed – to avoid the problems that required the initial PJM UMP market to shut down within hours when significant congestion first appeared.

In principle, the correct way to determine UMCPs and compensation payments in a multi-step IO process is to operate constrained and unconstrained processes in parallel at each step. In the SCUC/scheduling process under discussion here, the same market bids used in the constrained SCUC process should be used in the (three-zone) unconstrained version of the SCUC model to identify the resources that would be committed in this hypothetical situation.²⁹ Then the resources committed in the unconstrained SCUC process should be used in an unconstrained scheduling/pricing model to determine the zonal UMCPs and the UMP-market schedules for each resource. The compensation payments for the day should be determined as discussed in the preceding section, by comparing operating profits for the entire day in the ISO's constrained schedules and in the UMP-market schedules given the UMPs.

It may be that in the absence of intrazonal congestion there would seldom be any need for the ISO to commit resources over and above those committed by the market, because there are almost always enough fast-start units somewhere on the system to meet even the ISO's demand forecasts even with interzonal transmission constraints. If so, then the unconstrained versions of the SCUC process can be skipped and the unconstrained scheduling/pricing

²⁹ A SCUC process without transmission constraints should, by definition, find no need for RMR resources or local market power mitigation, but might commit some inflexible resources that might otherwise not start early enough or at all.

processes run using all available resources. This is the assumption used in section 5 when discussing how the TAPAS approach might work in practice.

If UMCPs, UMP-market schedules and CUPs/CDPs are determined using either the conceptually correct or the simplified approach just described, it should be relatively easy to find and enforce the efficient outcome, because the combination of a UMP and CUPs/CDPs is incentive compatible with the ISO-determined unit commitment, forward schedules and dispatch. But daily CDPs would sometimes/often be made to resources that do not start up for a whole day or longer because they would only be constrained off if they did. This would no doubt be regarded as unnecessary and unfair, and could invite gaming by resources with local market power who know they are likely to be constrained off.³⁰

Given the widespread aversion to CDPs even for resources that start up and are available for real-time dispatch, it is probably acceptable to let uncommitted resources affect the UMP or receive CDPs. It is likely, therefore, that if the ISO uses the kind of multi-step IO process discussed here for unit commitment and scheduling/pricing, it will constrain off some resources without compensation in the SCUC step, and will then determine UMCPs and compensation payments using only those resources that made the cut there. If the lack of CDPs does not affect bidding and self-commitment/scheduling decisions, the result will be UMCPs that are “too high,” many resources constrained off without compensation in the SCUC step even if CDPs are made in the scheduling/pricing process, and strong incentives for resources to use strategic bidding and self-commitment/scheduling to get access to the too-high UMCPs; if these defensive responses are strong enough, they will depress the UMCPs to some unpredictable extent and make the schedules inefficient and unstable.

4.5 ANCILLARY SERVICE (A/S) SCHEDULES AND PAYMENTS

In the MRTU procedures developed for the LMP market, A/S are scheduled and priced simultaneously with energy, subject to the constraints in the full network model (FNM) and resource-specific constraints such as ramping limits. These same A/S and energy schedules would be used in the TAPAS approach to define the quantities in forward market schedules and real-time dispatch. The question is: what should A/S providers be paid for the A/S they provide?

³⁰ Special provisions regarding compensation (or CRR payments) are needed to deal with extended transmission outages, whether in a UMP-CUP-CDR market or a LMP-CRR market. If this situation can arise under normal grid conditions the resources receiving the compensation should, as discussed in section 4.1.5 above, be required to pay (say) annual CDP access charges that would be at least as large as the compensation they expected to receive over the year. As discussed in section 4.1.3, if CDPs are made bidding down is a profitable strategy only for resources with local market power, but if CDPs are not make bidding down is a necessary strategy for every potentially constrained-down resource, whether or not it has market power. This suggests that controlling bidding down is a larger problem if CDPs are not made than if they are.

One possible answer to this question, and the one proposed in the CAISO TAPAS draft, is that A/S providers should be paid the A/S marginal prices (ASMP) associated with the A/S schedules from the constrained MRTU processes. But these ASMPs reflect opportunity costs based on local LMPs, which have no direct relationship to opportunity costs in a UMP market. Paying A/S suppliers in a UMP market the ASMPs from the MRTU models would be inconsistent with what is done on the energy side and would not be incentive compatible with efficient schedules. Some ASMPs would be so high relative to energy UMPs that resources not scheduled to supply A/S would want to do so, while other ASMPs would be so low that the resources scheduled to provide A/S would prefer to produce nothing or to get the UMP for energy. Such incentive incompatibility would stimulate strategic bidding and self-scheduling intended to influence MRTU scheduling decisions.

In the TAPAS approach, energy and A/S schedules are determined *simultaneously* by the constrained MRTU models that minimize total bid-costs over the whole day. Then (assuming the analysis in this paper is correct and its recommendations are adopted) unconstrained versions of these same MRTU models are run to determine the zonal UMCPs, UMP-market schedules and UMP-market operating profits. But the unconstrained MRTU models will still simultaneously optimize energy and A/S, and hence will produce A/S schedules and prices that are consistent with the UMP-market energy schedules and UMCPs. The most logical and practical way to deal with A/S in the UMP market is to use these UMP-market A/S schedules and prices the same way the UMP-market energy schedules and prices are used.

The principal steps in a TAPAS process that treats energy and A/S symmetrically are the following:

- **Constrained Schedules and Prices:** The MRTU's I/O processes are used to determine the constrained schedules for energy and A/S for the entire day, along with the energy LMPs and the A/S marginal prices (ASMPs) for each of the A/S regions.
- **UMP-Market Schedules and Prices:** The unconstrained versions of the MRTU processes are used to determine UMP-market schedules for energy and A/S for the day, along with the associated LMPs and ASMPs. For each hour, all the (unconstrained) LMPs within an energy pricing zone will have the same value, which is the zonal UMCP/UMP used for settling the energy quantities in the constrained schedules. The (unconstrained) ASMPs for each A/S-region are the "uniform A/S prices" (UAPs) used for settling the A/S in the constrained schedules.
- **UMP-Market Operating Profits and Compensation Payments:** The UMP-market operating profit for each resource, defined as the operating profit that resource would make for the day under its UMP-market energy and A/S schedules given the UMCPs and UAPs, is determined and compared to the operating profit resulting from operating according to the constrained energy and A/S schedules. If the UMP-market operating profit is larger the difference (perhaps less any CDP amounts determined in the next step) is paid as compensation.
- **Separating CUPs from CDPs:** If CDPs are not made for either energy or A/S or both, the total compensation amounts determined in the previous step must be divided into

CUPs and CDPs, and perhaps for energy and A/S separately. But the total compensation amounts for the day will include both CUPs and CDPs, for both energy and A/S, mixed together in an omelet that may be impossible to unscramble. In simple cases, such as where a resource provides either energy or A/S during the day but not both, and is constrained either up or down all day, it may be clear that the total amount is either an allowed CUP or a disallowed CDP. But in complex cases, a resource may provide both energy and A/S in different ratios over the day in either the UMP-market or the constrained schedule or both, may (appear to) be constrained up in some hours and constrained down in others, or may even (appear to) be constrained up for energy and constrained down for A/S at the same time. In general, separating CUPs from CDPs, for either energy or A/S or both, will require difficult and largely arbitrary judgments that must be implemented in complex systems.

It is possible that the complications suggested in the preceding paragraph are more theoretical than real. It may be that the CAISO's unit commitment, IO-based scheduling and pricing, and simultaneous optimization of energy and A/S are much simpler in practice than they appear to be in theory. If so, something simpler than suggested above may be adequate. But if the MRTU processes are anywhere near as sophisticated and complex in practice as they appear to be based on high-level descriptions of them, some version of the process outlined above will be needed, and trying to do anything much simpler will create more problems than it will solve. In this case, it may be necessary to reconsider the decision not to make CDPs and/or to use the full MRTU apparatus in the TAPAS approach.

4.6 ISSUES IN A TWO-SETTLEMENT UMP MARKET

The analysis above implicitly assumes that the energy and A/S schedules produced by the constrained and unconstrained processes are used for real-time dispatch, not for forward trading. When the ISO schedules are used for settling forward markets things become more complex, because now a resource that is constrained on or off in the forward market has incentives and opportunities to act in the real-time market to change forward market outcomes. To counter these incentives, the ISO must either rely on administrative coercion to force compliance with forward schedules or implement special settlement processes to try to maintain incentive compatibility between market payments and efficient operations.

The problem can be illustrated by considering what happens if generator "G" is constrained up in an ISO energy schedule/market, i.e., is scheduled to produce energy even though the UMP is less than G's bid price/marginal cost. The need to pay G both the UMP and a CUP equal to the difference between its bid price/marginal cost and the UMP is generally accepted. But in a single-settlement UMP market, G is not paid UMP+CUP based on its scheduled/dispatched MWh independent of what it actually does; G is paid UMP+CUP only for the MWh it actually produces in real time, up to its scheduled quantity and perhaps subject to penalties for dispatch deviations. Something similar is needed in the two-settlement TAPAS approach.

Suppose G is constrained up by the ISO in the forward (DA) market. In this case, the ISO's DA settlement process will credit G with the DA UMP (UMP_{DA}) plus the DA CUP (CUP_{DA})

on its full DA scheduled quantity (Q_{DA}). Suppose that the DA market is a perfect predictor of RT market outcomes, so that the real time UMP (UMP_{RT}) and G's real-time dispatch quantity (Q_{RT}) are the same as UMP_{DA} and Q_{DA} , respectively. In this case, G is constrained up again in real time, in the sense that given the UMP_{RT} and its bid price/marginal cost (MC), G would prefer to produce less than Q_{RT} . If the two-settlement UMP market worked the way two-settlement LMP markets do, G would be able to buy back its forward commitment at UMP_{RT} , which is the same price G was paid for it in the DA market, leaving G with the CUP_{DA} and no obligation to deliver anything in the RT market. Obviously, this cannot be allowed; a two-settlement UMP market needs something that a two-settlement LMP market does not.

In a two-settlement system, a resource that has made contract commitments in the DA market is encouraged to submit to the RT market inc bids prices at which it will produce more than its contracted amount and dec prices at which it will buy back some of its contract commitment, and the ISO uses these inc/dec bids to manage RT imbalances and congestion. In the example here, if G bids its actual marginal costs MC to buy back its DA contract commitment Q_{DA} , its bid will not clear in the RT market and G will be dispatched to produce Q_{DA} . But if G raises its dec bid to slightly above the LMP at its node, that bid will be accepted; G will then be dispatched to produce nothing in real time, will buy back its Q_{DA} for the same price it was paid for it, and will pocket the CUP_{DA} for doing nothing.

The ISO's market-power mitigation procedures could try to prevent G from bidding in a non-cost-reflective manner in the RT market, but this would be difficult and not very effective – particularly because G's incentive to bid up as in the previous paragraph has nothing to do with market power, but is due purely to incentive incompatibilities in the RT market. And even if G does bid its MC in the RT market, the RT LMP at G's node could fall below G's MC, producing the same result with no help from G, i.e., the ISO would dispatch G to produce the efficient amount – 0 MW – in real time, and G would buy back its forward commitment and pocket the CUP_{DA} for doing nothing. Or G could have an unexpected outage before real time; should G be allowed to buy back its Q_{DA} contract amount at UMP_{RT} and keep the CUP_{DA} as long as it has a “good enough” excuse for not being available in real time?

Obviously, something other than – or in addition to – administrative enforcement of real-time dispatch instructions is needed in a two-settlement UMP market. The logic of the situation suggests that, because G was paid $UMP_{DA}+CUP_{DA}$ for its forward contract amount Q_{DA} , if G does not deliver Q_{DA} in real time *for any reason whatsoever*, it should have to buy back any shortfall at an analogous “uplifted” price. The only buy-back price that is fully incentive compatible with efficient RT operations is LMP_{RT} , but this is ruled out by definition in a UMP market. A second-best approach is to require that G buy back any contracted but undelivered energy $UMP_{RT}+CUP_{DA}$.

If G is constrained down in the DA market to Q_{DA} , the situation is just reversed. When G shows up in real time it will again be constrained down unless UMPs or LMPs have changed significantly, and in the absence of any CDP_{RT} s will have incentives to use strategic bidding and/or self-scheduling to try to produce more than Q_{DA} , whether or not it was paid a CDP_{DA} .

But if G did receive a CDP_{DA} in exchange for its schedule/promise not to produce more than Q_{DA} in RT, that amount should be paid only if G actually does limit its output in real time; if G produces more than Q_{DA} in RT *for any reason whatsoever*, G should be paid only $UMP_{RT} - CDP_{DA}$ for anything RT production in excess of Q_{DA} up to its DA UMP-market schedule amount.

In summary, the DA market should make only contingent CUPs (or CDPs, if these are made at all). A resource that is constrained up (down) in the DA market should be told that it will recover any (opportunity) costs caused by differences between its own costs/bids and the UMP_{DA} on DA-scheduled quantities if, but only if, it actually incurs those (opportunity) costs in real time.³¹ This is essentially what happens in a single-settlement system, where CUPs (and CDPs, if made) are based on actual metered quantities rather than on dispatch instructions. It is hard to see how this is unfair, creates incentives for strategic bidding or exacerbates whatever market power some market participants may have anyway.

4.7 OTHER ISSUES IN A TAPAS-TYPE UMP MARKET

4.7.1 RECOVERING CONGESTION COSTS

Because the CAISO is a non-profit entity, the total costs incurred by the ISO in making constraint payments must be recovered from market participants somehow.³² This section describes some general principles of cost allocation and what they imply for the allocation of uplift.

It is in the nature of a UMP market that many or most of the costs of managing congestion must be socialized across market participants somehow, because without something like LMPs it is difficult to determine what actions cause how much of total congestion costs. But some congestion costs can and should be allocated on a cost-causation basis, with only the balance recovered through what is essentially a tax – the uplift. In particular, it is argued in section 4.1.5 above that, in a UMP market with CDPs, generators located where they add to

³¹ This does not fully solve the incentive problem if UMP_{DA} and UMP_{RT} are different. In this case, a resource constrained up in the DA market can make $(UMP_{RT} - UMP_{DA}) \times (Q_{DA} - Q_{RT})$ if it can find ways to produce less in RT than its DA commitment, and is exposed to the DA-RT price risk if it is scheduled to produce more or less than Q_{DA} in RT. The incentives and risks resulting from DA-RT price differences should be small given that the DA market should be a good predictor of RT outcomes.

³² A profit-making ISO can share some of these costs in order to give it an incentive to minimize these costs. This makes sense only if the ISO has a lot of flexibility in deciding how to manage and reduce congestion and probably only if it is a profit-making entity that also owns or at least makes operational and investment decisions regarding the grid. Such arrangements have worked well in England and Wales, where NGC is a profit-making private company that owns the grid and has a lot of flexibility in negotiating with individual market participants – although not to impose significant uncompensated costs on them.

the ISO's expected CDP payments should pay locational CDP access charges that approximate their incremental effects on the ISO's total CDP payments. This would require those who cause CDP costs to pay them, albeit through periodic charges rather than in each pricing interval. Any such CDP access charges should be used to reduce something like periodic grid access charges for those market participants who pay the uplift in each pricing interval. Even if CDPs are not made there are good reasons to impose locational grid access charges reflecting differences in incremental congestion costs, but in this case there is no compelling reason to use the revenue from such charges for the specific benefit of those who pay the uplift.

It is argued in section 4.2.1 above that the second-best UMP for settling loads is the load-weighted average of LMPs, LMP_{LAV} . Settlement of loads at LMP_{LAV} and of generation at a UMCP or LMP_{GAV} should yield a settlement surplus in excess of the ISO's congestion costs, eliminating the need for an uplift charge to SCs and even providing a surplus that can be used to reduce grid access or other charges paid by loads.

If loads are not settled at higher prices than generation, there will be no settlement surplus to pay the ISO's congestion costs, making it necessary to recover these costs through charges on SCs. There are two general principles for allocating such costs: it is *fair* to allocate such costs to the broad groups that benefit from them; and it is *efficient* to allocate such costs where they do not stimulate inefficient reactions. ISO congestion management reduces costs and risks both for generators as a group and for consumers as a group, and it is hard to say that one group benefits more than the other; for this reason, some markets split congestion costs (and the fixed costs of grid assets) more-or-less evenly between generators and loads. But allocating costs to generators without clear efficiency benefits increases the costs generators must recover from consumers through energy or other prices in the long run, and is likely to stimulate inefficient actions. For these reasons, it is probably better to allocate the uplift primarily to price-taking loads (and exports, and perhaps price-taking generators, who share many of the same characteristics), as has been the practice in California and elsewhere.

The CAISO has proposed some specific methods for allocating uplift costs among SCs in the TAPAS approach. These methods appear to be broadly consistent with the general principles above and, probably more importantly, are based on past and/or agreed practices. If uplift costs in the TAPAS approach turn out to be much larger than anticipated – which is possible – it may be necessary to reconsider some of the uplift allocation decisions, but for now there is no compelling reason to do so.

4.7.2 MARKET POWER MITIGATION

Under the MRTU procedures the CAISO conducts pre-IFM and pre-dispatch processes to determine whether market power mitigation (MPM) procedures should be activated. These processes identify any resources whose bid prices both (1) exceed administratively-determined reference prices and (2) would increase local LMPs by more than specified threshold amounts. If such bids are identified, they are automatically mitigated – i.e., reduced – based on the reference prices. These mitigated bids are then treated as the market bids of those resources in the subsequent scheduling/dispatch and pricing procedures.

The details of the MRTU MPM procedures are still being developed. Whatever the merits of these (or any such) procedures in general, there is no compelling reason to change them just because UMPs instead of LMPs are used for settlements – except for the need to extend them to monitor and mitigate bids that are *too low* as well as bids that are *too high*.

Whether or not CDPs are made in a UMP market, some resources will have incentives to bid below their costs, but with an important difference that is explained in more detail in section 4.1.3: paying CDPs encourages bidding down only by resources that have local market power, while not-paying CDPs encourages bidding down by all resources, even small and highly competitive ones. The MPM procedures will have to be extended to monitor and mitigate unreasonably low bids as well as unreasonably high bids whether or not CDPs are paid, but in the absence of CDPs these procedures may have to cover even small and competitive resources, not just those with local market power.

Mitigating bids based on the impact on local LMPs is reasonable even though settlement is at UMPs, because the effect on local LMPs is the best measure of the success of an effort to exercise local market power. The fact that the effects of using local market power are socialized across many locations instead of being focused on a few actually strengthens the argument for mitigation, because when the effects are socialized there is no effective demand-side response to help limit the exercise of local market power.

One issue with the MPM procedures is how bidding-down strategies might affect the reference prices and *vice versa*, given that reference prices are in some cases based on past bidding behavior and not just costs. A generator that bids low or even negative prices to take advantage of CDPs or to avoid being constrained down in the absence of CDPs may lower its future reference prices, making it more subject to mitigation of high bids in the future. This would seem to be a deterrent to bidding-down strategies, at least for resources that are not consistently constrained down. But how the details of the MPM procedure might affect gaming opportunities and incentives cannot be determined without more detailed analysis than is possible here.

5. OPERATION OF THE TAPAS UMP MARKET

In the TAPAS approach, the UMP market will include the same steps and use the same procedures and processes proposed in the MRTU for use in the LMP market, as summarized in section 3.1. This section discusses how the UMP market would operate at each step in the daily CAISO market process as proposed in the CAISO TAPAS draft, focusing on the differences between that proposed process and the process suggested by the preceding analysis in this paper.

5.1 THE DAY-AHEAD PROCESSES

5.1.1 THE PRE-IFM STEP

After receiving all bids and self-commitments/schedules for the IFM but before clearing that market, the CAISO determines an optimal commitment and dispatch of energy and A/S resources to meet the CAISO's forecast of demand (as opposed to self-scheduled and bid-in demand, which will usually be lower than expected actual demand), using the same constrained SCUC process to be used in the IFM. The results of this pre-IFM process are used for market power mitigation purposes; if certain conditions are met, some bid prices will automatically be reduced (or, presumably, increased if necessary to control bidding-down strategies in a UMP market) to administratively predetermined levels and those mitigated bids will be used in the subsequent IFM steps. The pre-IFM process is also used to decide which RMR resources to schedule, if any. Only the resources identified by the CAISO as part of the pre-IFM least-cost unit commitment will be eligible for unit commitment in the DA IFM and DA RUC processes.³³

In principle, the CAISO should use a (three-zone) unconstrained version of the pre-IFM SCUC process to determine a "UMP-market unit commitment" of the resources that will be eligible for the unconstrained versions of the IFM processes. In practice, the unconstrained SCUC step will probably be skipped, because CDPs will be judged unacceptable for resources constrained off at this stage even if – as is also unlikely – they are accepted for resources constrained off later. An unconstrained SCUC might have little effect even if it were computed, because in the absence of transmission constraints there should be no need for local market power mitigation or for RMR resources, and there may be little/no need for the

³³ The CAISO's July 2003 MD02 submission to FERC says (para. 43) that the resources "committed" in the pre-IFM SCUC process will "represent the optimal commitment decisions" and "therefore" only these resources will be eligible for the IFM. It is not clear if a resource that is not part of the least-cost SCUC solution can continue into the IFM if it is willing to take its chances without CAISO cost-recovery guarantees, or if not, why not. Allowing this would not cause any obvious problems, at least not in a LMP market or a UMP market with both CUPs and CDPs, because these markets maintains incentive compatibility between total payments and efficient schedules. But in a UMP market that does not make CDPs it may be necessary to preclude such self-commitment.

CAISO to commit resources that would not clear in a (three-zone) unconstrained market anyway.

5.1.2 IFM UNIT COMMITMENT

The CAISO uses the constrained SCUC process again to find a unit commitment that meets the load that self-schedules or clears in the IFM energy market plus A/S requirements (as opposed to the CAISO demand forecasts used in the pre-IFM process). The only resources eligible for the constrained IFM SCUC process are those that were committed in the constrained pre-IFM process. The result of this process is a constrained IFM unit commitment that defines the resources available for the IFM scheduling and pricing process. The CAISO guarantees each resource committed in this step that if does not recover all of its costs, including start-up costs, from the UMPs and A/S prices over the day (or other appropriate period) the CAISO will make up the difference. Resources that were committed in the pre-IFM process but not in the IFM process are available for commitment in the RUC process that follows the IFM.

In principle, the CAISO should use an unconstrained version of the IFM SCUC process to select, from those resources selected in the unconstrained version of the pre-IFM SCUC, the resources that will be eligible for the unconstrained IFM scheduling/dispatch processes. In practice, the unconstrained SCUC process will probably not be used – and might have little effect even if it were.

5.1.3 IFM SCHEDULING AND PRICING

The final constrained IFM run uses self-scheduled and bid-in generation (from resources committed in the constrained IFM SCUC) to meet self-scheduled and bid-in load plus A/S requirements. Energy and A/S are optimized simultaneously for all hours of the day simultaneously, to determine constrained energy and A/S schedules and the associated (constrained) LMPs and regional A/S marginal prices (ASMPs).

In principle, the market bids from all resources committed in the unconstrained IFM SCUC process – or all resources that bid or self-commit in the IFM if the unconstrained SCUC processes are skipped – should be used in unconstrained versions of the IFM scheduling/pricing models, to produce UMP-market energy and A/S schedules, and associated UMCPs and UAPs; these schedules and prices can be called “truly-unconstrained” because transmission constraints have no effect on them at any step in the process.

In practice, it is likely that the unconstrained IFM scheduling/pricing models will be run using only those resources selected in the constrained IFM SCUC step, i.e., not constrained off in that step. As discussed in section 4.4.3 above, these UMCPs will usually be higher than the “truly-unconstrained” UMCPs in the previous paragraph, because they are based on a subset of resources that excludes some low-cost resources that were constrained off in the SCUC step. This relationship between the UMCPs will probably hold even if CDPs are not made and there is some bidding down and self-scheduling, because these strategies will depress UMCPs in both cases – as well as make the constrained schedules (which are based on the same bids and self-schedules) inefficient and unstable.

According to the CAISO TAPAS draft (p. 7), the zonal UMPs might be an average of the constrained LMPs in each zone rather than a UMCP as recommended here, and the UAPs would be the ASMPs from the constrained scheduling model. Zonal UMPs determined this way would probably be higher than the “truly-unconstrained” UMCP in the preceding paragraph, even if CDPs are not made and there is a lot of bidding down and self-scheduling. And UAPs equal to the constrained ASMPs would, as discussed in section 4.5, not be incentive compatible with the efficient schedules and the UMPs.

5.1.4 IFM SETTLEMENTS AND CUPS/CDPS

The energy and A/S quantities in the constrained schedules are settled at the zonal UMPs and regional UAPs, whether these are determined using the unconstrained market-clearing approach recommended here or some other way. The difficult problem is how to determine compensation payments, even if only CUPs are made.

As discussed in sections 4.4 and 4.5 above, the most logical and perhaps the only practical way to determine compensation payments given the complex processes used to determine energy and A/S schedules in the TAPAS approach is to compute the total commercial results for each resource given the UMPs and the UAPs over the day in energy and A/S markets under two different schedules: (1) the unconstrained UMP-market schedules; and (2) the constrained ISO schedules. If a resource would make more money under the UMP-market schedule, the difference should be paid as compensation. If CDPs are not to be made, or if either CUPs or CDPs are to be made for energy but not A/S or *vice versa*, some way must be found to divide this total compensation between CUPs and CDPs for energy and A/S, which may require complex and largely arbitrary rules of thumb except in simple cases.

The CAISO TAPAS draft does not propose such a process for determining CUPs (CDPs are not contemplated). Instead, it suggests a very simple process for energy CUPs (p. 10): “The supply bids (mitigated were relevant) accepted in the scheduling run that are higher than the relevant zonal MCP are paid uplift to make them whole.” (p. 10) The implication is that energy CUPs can be determined essentially by inspecting each hour in turn, finding all suppliers that are scheduled to provide energy at bid prices that exceed the UMP in that hour, multiplying the bid price-UMP differences by the constrained-up energy amounts for each hour, and adding up over the hours. But this ignores all the intertemporal complications discussed above and illustrated in Appendix B. Unless the MRTU processes are much simpler in practice than they are said to be in theory, the suggested hour-by-hour inspection process will not be adequate. UMP-market schedules must be determined somehow.

Furthermore, the CAISO TAPAS draft proposes no CUPs or CDPs for A/S suppliers even though the UAPs based on the constrained ASMPs may be far from incentive compatible with the schedules. The MRTU processes are likely to schedule some suppliers to produce A/S when, given the UAPs and UMPs, it would be much more profitable for them to produce energy (or nothing), or *vice versa*. If no CUPs or CDPs are paid for A/S, resources may have strong incentives to develop and implement complex bidding and self-scheduling strategies to affect the outcome of the CAISO’s unit commitment and scheduling processes.

5.1.5 DA RESIDUAL UNIT COMMITMENT (RUC)

If the demand that clears in the IFM is less than the CAISO's forecast of actual demand, the CAISO applies the SCUC models in a post-IFM process to determine if additional resources should be committed to meet the CAISO's demand forecast. The resources eligible for this incremental or residual unit commitment (RUC) process are those that were committed in the pre-IFM step but not in the IFM step. Resources are not committed in the DA RUC process if they can start up rapidly enough to be committed in the hour-ahead (HA) process if they are needed then. No settlement prices are determined in the RUC process, but resources committed there are guaranteed that if they do not recover their daily costs in the real-time energy and A/S markets the CAISO will make up the difference.

In principle, the CAISO should run an unconstrained version of the SCUC model in parallel with the constrained version, using as eligible resources those that were committed in the unconstrained pre-IFM SCUC but not in the unconstrained IFM SCUC, to determine which resources would be committed at this stage if there were no intrazonal congestion. It seems unlikely, however, that an unconstrained SCUC would find any need to commit additional resources at this stage; there should be enough fast-start flexible resources somewhere on the system, even with interzonal constraints, to meet incremental demand without having to schedule and make cost-recovery guarantees to additional resources at the DA stage. If this is correct, no unconstrained RUC run is needed.

5.1.6 DA SETTLEMENT

The results of the DA IFM are the DA energy and A/S schedules – or DA contract commitments – for each resource, the hourly UMPs and UAPs at which those commitments will be settled and the CUPs (and CDPs, if made) for each resource. This information must be carried forward to the RT market so that any CUPs and CDPs determined in the DA IFM are made only if the corresponding forward commitments are actually fulfilled. As discussed in section 4.5, if a resource deviates from its DA contract commitments in real time its DA CUPs/CDPs should be adjusted accordingly.

5.2 SIMPLIFIED HOUR-AHEAD “MARKET”

No settlement prices are determined in the simplified HA “market,” so it is not really a market. But an hour ahead of the real-time dispatch hour, the CAISO uses updated self-schedules and inc/dec bids (relatively to IFM schedules for resources or loads that have such schedules) in an IO process with a five-hour look-ahead to determine a real-time dispatch that meets the ISO's demand forecasts. Resources scheduled in the IFM to provide A/S are not used in the HA energy scheduling process; if A/S requirements have increased since the IFM because load has surged or if an A/S resource has failed, the CAISO may procure additional A/S at this stage. The results of this HA process become dispatch targets for those resources that cannot respond to dispatch instructions within the hour; such resources are treated as must-run in the real-time dispatch and pricing process, and are guaranteed that if they do not recover their total costs, including start-up costs, in the RT energy market over some number of hours the CAISO will make a CUP equal to the difference.

In principle, the CAISO should perform an unconstrained version of this same HA process to determine (among other things) the unconstrained UMP-market schedules for resources that cannot respond to within-hour dispatch instructions. However, if CDPs are not to be made this may not be necessary, because the dispatch process may be simple enough – e.g., it takes A/S schedules as given rather than co-optimizing energy and A/S – that CUPs can be determined as the CAISO proposes, i.e., as any amount necessary to make up any difference between a scheduled resource’s full costs and the revenue from RT energy and A/S prices over some number of hours.

5.3 THE REAL-TIME PROCESSES

5.3.1 REAL-TIME DISPATCH AND PRICING

In real time, the CAISO commits fast-start resources every fifteen minutes and uses security-constrained economic dispatch (SCED) with a FNM to determine energy dispatch instructions every five minutes. The CAISO TAPAS draft proposes that zonal RT UMPs for energy be determined as an injection-weighted average of LMPs within each zone, using the LMPs either from the constrained SCED run (if market power mitigation is “adequate”) or from a (three-zone) unconstrained run of the SCED model using all available resources not scheduled for A/S (if market power mitigation is “inadequate”). The CAISO TAPAS draft proposes that the RT UAPs be the ASMPs from the constrained SCED.

The MRTU’s RT unit commitment and dispatch processes are complex, with many details either still undecided or unclear from available CAISO documents. In concept, however, the constrained RT SCED process should use inc/dec bids from resources (at least for energy, if not for A/S) to determine optimal RT energy and A/S schedules simultaneously, and an unconstrained version of this same process should be run to determine the UMP-market energy and A/S schedules and the zonal UMPs and regional UAPs. If the RT unit commitment and scheduling process does not involve IO and other complexities it may be easy to determine the RT UMP-market energy and A/S schedules, but this must be done in some form.

5.3.2 REAL-TIME SETTLEMENTS

Resources that are scheduled to produce more or less energy or A/S in real time than called for in their DA schedules should be paid or should pay the real-time UMPs and UAPs for it, with adjustments for any quantities for which CUPs or CDPs were made in the DA market. As explained in section 4.6, a generator (for example) that was constrained up in the DA market to produce some quantity Q_{DA} in exchange for a total credit of $UMP_{DA} + CUP_{DA}$ (in \$/MWh) should be debited $UMP_{RT} + CUP_{DA}$ for any RT shortfall no matter what the reason for that shortfall. Recovering the CUP_{DA} in this way is not a penalty; it is just recognition of the fact that the generator was paid for a forward commitment that it did not keep and that it should have to buy back at an appropriate price. With the suggested adjustment, resources constrained up in the DA market (or down, if CDPs are made in the DA market) will not be unfairly rewarded or penalized for any difference between DA and RT schedules; they will

make or lose money on any changes in the UMP between the DA and RT market, but this is a desirable feature of a two-settlement market.³⁴

An important issue regarding RT settlements is whether CDPs will be paid to resources that are dispatched in real-time to produce less than the UMP-market schedule quantity amount that would maximize their RT operating profit given their bids and the RT UMP. The CAISO TAPAS draft suggests (p. 15) that both CUPs and CDPs would be made at least to resources with DA schedules. In particular, a resource that is “dec-ed” in real time to produce less than its DA scheduled amount when its dec bid is less than the RT UMP would be compensated for the difference between the RT UMP and its RT dec bid; this is a RT CDP that seems inconsistent with the general rejection of CDPs in the CAISO TAPAS draft. If there are good reasons for such CDPs here, perhaps they apply just as well elsewhere.

³⁴ The fact that a DA commitment can be sold (or added to) in the RT market at a profit if the RT price is less (more) than the DA price is desirable in a market with efficient RT prices, because it encourages market participants to respond to RT prices whatever their DA positions. This advantage is less important in a UMP market, because the RT UMPs may be far from the efficient RT LMPs that would encourage efficient responses.

Appendix A: Second-Best UMPs when Compensation Is Not Paid

Consider the following situation:

- The ISO schedules generators to minimize the total as-bid costs of meeting demand given a full network model (FNM), which produces LMPs consistent with the efficient dispatch.
- The ISO makes neither constrained-up payments (CUPs) nor constrained-down payments (CDPs).
- Generator G_j at node j , where the LMP is LMP_j , is scheduled by the ISO to produce the amount Q_j .
- Generator G_j can produce either more or less than Q_j , i.e., G_j is a marginal generator that can and presumably will respond to incentives.

If G_j can change or deviate from its efficient schedule amount Q_j by a small amount ΔQ_j and sell the additional output at the UMP, its operating profit will increase by $\Delta Q_j \times (UMP - MC_j)$, where MC_j is G_j 's marginal cost. But, because G_j is a marginal generator, its marginal cost MC_j at its scheduled amount Q_j must equal the LMP at its node, LMP_j . Thus, G_j 's incremental profit from deviating from its efficient schedule by an amount ΔQ_j is $\Delta Q_j \times (UMP - LMP_j)$.

It is reasonable to assume that the more money G_j can make by deviating from its efficient output amount, the harder it will try to find ways to do so and the more successful it will be, either in bidding or self-scheduling to affect the schedule it gets from the ISO or in deviating from the ISO's schedule without being caught or penalized. This assumption, along with the result above, suggests that the expected deviation from efficient dispatch at node j , $E(\Delta Q_j)$, is approximately proportional to the difference $(UMP - LMP_j)$.

It is a general principle of optimization that the objective function being optimized is, near the optimal values of the independent variables, flat and rounded (in mathematical directions defined by the binding constraints). This implies that very small (feasible) deviations in the independent variables from their optimal values have virtually no effect on the value of the objective function because the deviations are (for a maximization problem) along the flat top of the "hill", but larger deviations move further down the hill where it becomes increasingly steep. For deviations that are "not too large," the cost of a deviation in terms of the value of the objective function increases approximately with the square of the deviation.

Generalizing this concept to the problem of choosing a second-best UMP suggests that the expected cost of dispatch inefficiencies – i.e., the increase in total dispatch costs, which is what is being minimized here – for generator G_j at node j is proportional to the square of the expected deviation $E(\Delta Q_j)$ from the efficient dispatch. Because $E(\Delta Q_j)$ is proportional to $(UMP - LMP_j)$, it follows that expected dispatch inefficiencies are proportional to the square of this difference. These assumptions imply that, as long as $(UMP - LMP_j)$ is not "too large,"

$$\text{The Expected Cost of Dispatch Inefficiencies at Node } j = k_j \times (UMP - LMP_j)^2,$$

where k_j is a positive number that depends on characteristics of the resources at node j . The factor k_j must increase with the quantity of dispatchable resources at node j , most of which will be generation (with some dispatchable demand at some nodes). Assuming that the other characteristics that might influence k_j are unknown, unobservable and/or randomly distributed, it is reasonable to say that:

$$\text{The Expected Cost of Dispatch Inefficiencies at Node } j = k \times G_j \times (\text{UMP} - \text{LMP}_j)^2,$$

where G_j is the amount of dispatchable generation (and any dispatchable load) at node j and k is a positive number that is the same for all nodes. The total expected cost of dispatch inefficiencies as a function of UMP, $\text{Cost}(\text{UMP})$, is then:

Total Expected Cost of Dispatch Inefficiencies = $\text{Cost}(\text{UMP}) = k \times \sum_j G_j \times (\text{UMP} - \text{LMP}_j)^2$,
where the sum is over all nodes j (at which there is dispatchable generation or dispatchable demand).

The second-best UMP, call it UMP^{SB} , is the value of UMP that minimizes $\text{Cost}(\text{UMP})$. Applying basic calculus, UMP^{SB} is the value of UMP at which the derivative of $\text{Cost}(\text{UMP})$ with respect to UMP is zero. Thus

$$d\text{Cost}/d\text{UMP} = 2k \times \sum_j G_j \times (\text{UMP}^{\text{SB}} - \text{LMP}_j) = 2k \times (\text{UMP}^{\text{SB}} \times \sum_j G_j - \sum_j G_j \times \text{LMP}_j) = 0,$$

or, solving for UMP^{SB} :

$$\text{UMP}^{\text{SB}} = (\sum_j G_j \times \text{LMP}_j) / (\sum_j G_j)$$

where G_j is the amount of dispatchable generation and demand at node j and the sum is over all nodes.

Thus, when the ISO makes neither CUPs nor CDPs, the UMP that is second best in the sense that it minimizes dispatch inefficiencies is the weighted average of the LMPs, with the weights equal to the amount of dispatchable resources at each node.

Appendix B: CUPs and CDPs with Intertemporal Optimization

In the simplest examples of constrained-up payments (CUPs) and constrained-down payments (CDPs), such as those illustrated in the figures in the text, an ISO in a UMP market can determine CUPs and (if made) CDPs simply by comparing energy bid prices to energy UMPs for each pricing/dispatch period – call it an “hour” – in isolation. For example, a generator that is scheduled in an hour to produce some amount of energy for which the bid price exceeds the UMP is entitled to a CUP for that hour equal to that amount of energy multiplied by the difference between the UMP and the bid price. This type of calculation can be made for each hour separately and the CUPs summed over the day to determine a daily CUP. Nothing could be easier.

Unfortunately, things are never that easy in practice and are immensely more difficult in the complex TAPAS approach. The basic problem is that in the TAPAS approach the CAISO does not determine energy schedules one hour at a time based on energy-only bids that reflect each resource’s actual situation in that hour and its own forecasts for future hours. Instead, the CAISO (says that it³⁵) determines hourly schedules for energy and ancillary services (A/S) simultaneously, and for all the hours in a multi-hour period – call it a “day” – simultaneously, taking into account the sometimes-complex interactions between energy and A/S, and complex intertemporal constraints ranging from simple hour-to-hour ramping limits through daily energy limits for (e.g.) hydro generators, to limits on how long it takes and how much it costs for a unit to start up or shut down and how long the unit must remain up once up or down once down.

When an ISO optimizes energy and A/S simultaneously for multiple hours simultaneously, it is impossible, as both a logical and a practical matter, to determine the appropriate energy CUPs or CDPs by looking only at energy bids and schedules for individual hours in isolation. In fact, it can be difficult even to define whether a resource is constrained up as opposed to down, for energy as opposed to A/S, in some specific hour, and impossible to do so by looking only at bids and schedules for that hour. This implies (among other things) that it may be difficult to implement a decision not to make CDPs.

The easiest way to illustrate the problems caused by intertemporal constraints is to consider a simple example in which the only intertemporal constraint is a ramping limit on a generator and the ISO schedules only one hour at a time, i.e., does not look ahead to consider how its decisions in this hour affect later hours (even though such a myopic scheduling process will not achieve a full optimum, in general, because decisions in one hour should consider their effects in later hours).

³⁵ The discussion here is based on what the CAISO says it will do under the MRTU and TAPAS proposals. Things may be much simpler in practice, if only because it may be impossible for the CAISO (or any real-world entity) to do everything it hopes to do.

Suppose generator G's bid for every hour of the day is to produce anything between 0 and 100 MW at \$30/MWh but with a ramping constraint of 30 MW/hour in either direction, and that G produces 50 MWh in hour h.³⁶ In hour h+1 there is no congestion, so all the LMPs and hence the UMP (for any reasonable UMP rule) are the same, say \$25/MWh. Because the UMP is below G's bid price of \$30/MWh, G would maximize its operating profit in hour h+1 by producing 0 MWh. But, because the ISO's schedule must take into account all constraints, including resource-specific constraints, the ISO will schedule G to ramp down only as far as it can, to produce 20 MWh in hour h+1.

The ISO's automated settlement system must now use well-defined criteria and formulas to determine the amount of any CUPs or (if made) CDPs that should be made with respect to hour h+1. The settlement system will observe that in hour h+1 the ISO has scheduled G to produce 20 MWh for which G's bid price is \$30/MWh even though the UMP is only \$25/MWh. If each hour is considered purely in isolation, G appears to deserve a CUP equal to $20 \text{ MWh} \times (\$30/\text{MWh} - \$25/\text{MWh}) = \$100$. The amount of this CUP can be determined virtually by inspection in this simple case, but to determine CUPs (or CDPs) in more complex cases the settlement system must compute G's operating profit under two different schedules and take the difference. Specifically, the settlement system must calculate G's UMP-market operating profit (which is \$0 in this case), defined as the profit G would make operating under its profit-maximizing UMP-market schedule given its bid and the UMP, and G's ISO-schedule operating profit (a loss of \$100 in this case), defined as what G makes under the ISO's schedule given its bid and the UMP, and subtract the latter from the former to calculate the CUP ($\$0 - (-\$100) = \$100$ in this case).

So it is possible to compute some sort of CUP by looking at hour h+1 in isolation. The question is whether this is the *right* way to calculate a CUP. In particular, is the \$100 CUP appropriate in this case? After all, the ISO scheduled G to produce 20 MWh at a loss instead of its profit-maximizing amount of 0 MWh only because G could not ramp down rapidly enough, and the purpose of CUPs/CDPs is not to protect resources against the adverse consequences of their own inflexibilities. Maybe the ISO should consider more than just what happened in hour h+1 in determining G's CUP for hour h+1.

A slightly better way to determine G's CUP for hour h+1 is to take into account where G was in hour h and its ramping limit. To do this, the ISO's settlement system subtracts G's ISO-market operating profit (still a loss of \$100) from its UMP-market operating profit, as before. But now G's UMP-market operating profit for hour h+1 is defined as the maximum profit G could make in hour h+1 given the UMP and its bid price *and given where G was in hour h*

³⁶ In practice, schedules usually take the form of target MW levels to which the resource is supposed to ramp linearly by the end of the pricing/dispatch "hour" and payments are based on metered MWh over the hour. In this case, even if the pricing/dispatch period is an hour, if a generator is ramping up or down over the hour the MWh for which it is paid is not the same as its scheduled MW, i.e., its end-of-hour target. These complications are ignored here, so that MW and MWh can be used interchangeably.

and its own ramping limit; given that G was producing 50 MWh in hour h and can ramp down no faster than 30 MW/hour, G could not produce less than 20 MWh in hour h+1 no matter what the ISO did. Thus, G's UMP-market profit for hour h+1 is a loss of \$100, just as in the ISO schedule, so maybe the appropriate CUP for hour h+1 is \$0.

But it is not enough to calculate G's CUP for hour h+1 by considering only G's ramping limit and *where* G was in hour h; it is also necessary to consider *why* G was where it was in hour h. After all, there are two reasons why the ISO cannot schedule G to produce less than 20 MWh in hour h+1. One reason is G's own down-ramping limit, which is its own problem; but the other reason is that G was producing 50 MWh in hour h, presumably because that is what the ISO scheduled G to do in hour h. Suppose G's UMP-market schedule amount in hour h was 25 MWh, but the ISO constrained G up to produce 50 MWh in hour h because of congestion. G may have received CUP payments in hour h to compensate it for its extra costs then, but if the ISO had not constrained G up in hour h G would have been able to ramp down to its UMP-market schedule amount of 0 MW in hour h+1. Because the reason G is scheduled to produce 20 MWh in hour h+1 at a \$100 loss is because G was constrained *up* in hour h – it is important to note that the problem here has nothing to do with G being constrained *down* – perhaps G does deserve the \$100 CUP for hour h+1 after all.

It might be argued in response to this example that the \$100 CUP for hour h+1 should not be made, because G's own inflexibility is at least partly to blame for the fact that G could not produce 0 MWh in hour h+1. But if resources who respond to constrained-up (or constrained-down) instructions in hour h are not compensated for any resulting (out-of-pocket or opportunity) costs in hour h+1 or h+2 or ... , they will begin looking ahead themselves, and when they see a potential future problem will use strategic bidding and self-scheduling to avoid being constrained up (or down) in hour h. And this is not just a theoretical issue: given the inflexibilities inherent in large generating units, most of the out-of-pocket or opportunity costs caused by an action in one hour may actually be incurred in later hours.

The delayed effects of constrained-up/down instructions may be positive as well as negative. For example, suppose everything in the above example is the same except that the UMP in hour h+1 is \$35/MWh compared to G's bid price of \$30/MWh, so that G wants to produce as much as it can in hour h+1. Because G produced 50 MWh in hour h, it can ramp up to 80 MWh in hour h+1; but if G had not been constrained up to 50 MWh from 25 MWh in hour h, it could have produced no more than 55 MWh in hour h+1. Thus, G's operating profit in hour h+1 is higher by $(80 \text{ MWh} - 55 \text{ MWh}) \times (\$35/\text{MWh} - \$30/\text{MWh}) = \125 because G was constrained up in hour h, which suggests that G's CUP for hour h+1 is a negative \$125, i.e., G should pay \$125 to the ISO.

Now suppose everything is the same as in the preceding paragraph, except that there is congestion in hour h+1 and the ISO constrains G down as far as possible, to 20 MWh. If CDPs are made in this market, what should G be paid – or perhaps pay – in hour h+1? G could be producing 80 MWh in hour h+1, so perhaps it should receive a CDP of $(80 \text{ MWh} - 20 \text{ MWh}) \times (\$35/\text{MWh} - \$30/\text{MWh}) = \300 . But if G had not been constrained up from

25 MWh to 50 MWh in hour h it could produce no more than 55 MWh in hour h+1, so maybe the “correct” CDP for hour h+1 is $(55 \text{ MWh} - 20 \text{ MWh}) \times (\$35/\text{MWh} - \$30/\text{MWh}) = \175 .

Now suppose everything in hours h and h+1 is as indicated in the preceding paragraph and that in hour h+2 there is no congestion and the UMP decreases to \$25/MWh. At this UMP, G wants to reduce its output as far as possible, and because it was producing 20 MWh in hour h+1 it can ramp down enough to produce 0 MWh in hour h+2. But if G had not been constrained down in hour h+1, it would have been producing 80 MW in hour h+1 and hence would have been unable to reduce its output in hour h+2 below 50 MWh; so perhaps G owes the ISO $(50 \text{ MWh} - 0 \text{ MWh}) \times (\$30/\text{MWh} - \$25/\text{MWh}) = \250 . But then, if G had been neither constrained up in hour h nor constrained down in hour h+1, it would have been producing 0 MWh in hour h+1 and hence could produce 0 MWh in h+2 as well; so maybe neither party owes the other anything in hour h+2. Or does this calculation double-count something that was already considered in hour h+1?

The lesson from these examples is that even with the simplest type of intertemporal constraints and simple myopic (and intertemporally non-optimal) scheduling by the ISO there is no logical way to determine CUPs or CDPs by looking at individual hours. It is necessary to consider all hours of the day as a whole to determine a CUP amount – or a CDP amount, if CDPs are made – for the day as a whole, and then to deal with the “end effects” created by the transition from one day to another.

Energy-Limited Resources.

Even more complex issues are raised by energy limited resources, such as hydro plant with a limited amount of water that it can use any time during the day but perhaps with complex constraints related to (e.g.) the level and continuity of stream flows. If the operator of such a resource wants the ISO to allocate its limited energy over the day based on the results of the ISO’s complex intertemporal optimization and full network models, the operator will specify the energy limits and other constraints but will bid a low or zero price in all hours, i.e., the energy-limited resource will become a price taker. The ISO will then schedule the energy over the day to optimize its value to the system, given the LMPs, without regard to the UMPs, but the resource will be paid for the resulting energy at the hourly UMP.

Except in very simple cases, the ISO schedule resulting from this process will imply a daily operating profit at the UMPs that is less, perhaps much less, than the maximum daily operating profit the energy-limited resource could earn given the UMPs and the resource-specific constraints. But if the ISO pays only CUPs determined on an hour-by-hour basis, it will pay little or nothing in compensation, because the UMP will be above the low/zero bid price in most hours; it will look like the energy-limited resource is often/always constrained off – for which it is not compensated – but is seldom/never constrained on to receive CUPs. In the absence of CDPs, energy-limited resources will quickly learn that it is does not pay to let the ISO schedule their energy, and will start making their own forecasts of UMPs and submitting complex bids that get the limited energy scheduled when the resource thinks UMPs will be high, not when the ISO, with its much better information, knows that LMPs are high. Not

only will system efficiency suffer, but resources will have to engage in the kind of strategic bidding and self-scheduling that makes it very hard to distinguish actions necessary for competitive survival from the exercise of market power or the exploitation of loopholes in the rules.

If the ISO decides to make both CDPs at least for energy limited resources so that they will let the ISO schedule them, it cannot use an hour-by-hour process to determine the CDPs, because an energy-limited resource with a very low or zero bid price will appear to be constrained off/down most of the time. For example, suppose an energy-limited resource offers to produce 100 MW in any hour of the day with a bid price of \$0/MWh, but can do so only in one of the day, and the ISO schedules that 100 MWh for late in the day. In any positive-UMP hour early in the day, it will appear that the resource is constrained off and deserves a CDP equal to $100 \text{ MWh} \times \text{UMP}$; in effect, the resource will “sell” the same energy in every hour of the day until the ISO finally takes it (and maybe beyond, depending how the hour-by-hour calculation is done). This is clearly nonsense.

The only logical way to determine CUPs and CDPs in this case is to consider the day as a whole. The daily operating profit for an energy-limited resource under the ISO-market schedule must be subtracted from the maximum daily operating profit that resource could make given the UMPs, the bids and all resource-specific constraints, and the difference paid as the compensation for the day. Whether that difference is CUP or a CDP, for the day as a whole or for specific hours within the day, is an interesting question. But there should be little question that unless compensation is determined in some version of this full-day process, energy limited resources will have strong incentives not to let the ISO allocate their energy, but to bid or self-schedule in an attempt to get their energy used when they think UMPs will be high, not when the energy is most valuable for the system.