Model Predictive Control in Combined-Cycle Power Plants: Market Implications

Victor M. Zavala and Fernando J. D’Amato

Abstract—We analyze the implications of a widespread deployment of model predictive control (MPC) for combined cycle generation in electricity markets. We claim that MPC can lead to significant improvements in grid reliability and market efficiency because it enables a more consistent quantification of the plant dynamic limitations. We also claim that, if combined with a self-commitment structure, MPC can lead to dramatic reductions in the computational complexity of central commitment and dispatch procedures.

I. INTRODUCTION

The main objective of electricity markets is to reliably serve distributed and dynamic electric energy demands at efficient costs. Under current market designs, the delicate balance between generation and demand is handled by a series of resource allocation processes distributed among different time scales. Starting at relatively long time scales, demand and renewable generation forecasting is used to commit resources. As forecast conditions approach real-time, uncertainties and prediction inaccuracies are handled by shorter time markets.

Central to these resource allocation processes is the solution of large mixed-integer linear optimization problems known as unit commitment (UC) and economic dispatch (ED) problems [3], which are solved by the independent system operators (ISOs) to determine schedules of generation plants, loading of transmission lines, and electricity prices. In order to manage large computational requirements, current UC/ED uses coarse physical information about capabilities and dynamic behavior of generating plants, mostly in the form of capacity and ramp limits. These simplifications introduce structural uncertainties that can be avoided by using more detailed physical and dynamic constraint of generating assets. This paper presents a framework to minimize these uncertainties and address the computational complexity of UC/ED by using a widespread deployment of model predictive control (MPC) at the generation plant level. We focus on the special case of combined-cycle (CC) plants that provide high dynamic flexibility to absorb variations of renewable power and enable significant emission reductions. We claim that the proposed MPC approach has the potential to dramatically improve grid reliability and market efficiency.

II. COMBINED-CYCLE GENERATION

Combined-cycle plants are characterized by the integration of gas and steam turbines to generate electricity. The principles of operation can be summarized as follows. The gas turbine compresses ambient air, mixes it with a gas fuel, and burns the mixture in combustion chambers. The gases resulting from combustion are expanded in the turbine, generating mechanical power. The exhaust gases are passed through a heat recovery steam generator (HRSG) to produce steam, which is then expanded in a number of steam turbines to generate additional mechanical power. In large CC plants, the steam turbine system includes three steam turbines, denoted as high-, intermediate- and low-pressure steam turbines. The electric power is obtained from electrical generators mounted in the same shaft as the turbines.

CC units are available in many different configurations, depending on the rated power, number of gas turbines per steam turbines, number of shafts, type of HRSG, and condenser technology. One of the most common configurations in North America is a system with two gas turbines per steam turbine, denoted as 2 on 1 and shown in Figure 1. Each plant may have more than one operational mode to generate power, depending on the plant configuration and the presence of optional equipment. Plants with more than one gas turbine can operate with any combination of gas turbines turned on or off. Optional equipment can include duct burners to allow fuel injection in the HRSG path to generate extra power. The transfer between any two operating configurations is constrained by plant capabilities, transition processes, and plant operating guidelines (see Figure 2). Each configuration has different cost structure, efficiency, and physical constraints that must be considered in order to determine the most efficient way to serve a given demand schedule.

A. Normal Operation

1) Objectives: The normal operation of a CC plant focuses on satisfying a time-varying load (demand) profitably and securely. If the plant participates in energy markets, it needs to satisfy an hour-by-hour, day-ahead schedule determined by central market clearing through a bidding process. Typically, the day-ahead schedules are corrected by the real-time market to compensate for grid power imbalances due to changes in the demand and fluctuation of renewable sources. Coarse real-time corrections are usually addressed by following price and load signals that are sent by the ISO every few minutes (see Sections III-A and III-B). The operator can decide to correct the power levels in
Fig. 1. Simplified schematic of a CC power plant with 2 gas turbines and 1 steam turbine with a 3-pressure HRSG. Each gas turbine (GT) is mounted in the same shaft as its corresponding compressor (COMP) and electrical generator (GEN). The steam turbine (ST) system with high (H), intermediate(I) and low(L) pressure turbines has its own electrical generator. The components of the steam generator (economizer (EC), drum (D), superheater (SH), reheater (RH)) are identified as high, intermediate or low pressure component. The steam actuation includes control valves (CV), bypass valves (BV), admission valves (AV).

Fig. 2. Operating configurations and possible upward transitions for a 2 on 1 CC plant with duct burners.

real-time if doing so would result in an economic surplus for the plant [14]. This decision is made according to the day-ahead and real-time prices. We describe market procedures in Section III.

2) Constraints: The most relevant operational constraints are the maximum temperatures that both gas and steam turbine materials can sustain and the maximum emissions allowed on site. The temperature limits impose a maximum power constraint, while local emission regulations impose a minimum power at which the plant can be dispatched, usually referred to as emissions compliance load. Also, across the operating range, the air flow into the combustion chamber has to be coordinated with the fuel flow to avoid lean mixtures that can lead to combustion instabilities or even suppression of combustion.

3) Uncertainties: Changes in ambient temperatures modify the normal plant operation. An increase in ambient temperature will tend to increase the gas turbine exhaust temperature. However, since the control system regulates the fuel and air flow to control turbine temperatures, the resulting effect is that whenever the ambient temperature surpasses a threshold value (typically 40 °C), the maximum generated power is lower than the rated power. Plant operations depend strongly on other exogenous factors such as ISO electricity prices, which fully determine the participation level of the plant in the day-ahead and real-time markets. In addition, on-site disturbances may impair the ability of the plant to comply with generation commitments. The most relevant on-site disturbances are frequency/speed variations, which can limit the maximum achievable power [12], [8].

B. Startup Operation

1) Objectives: The startup process involves drastic transients that carries the plant operating point from zero flows, zero speed, and initial temperatures state. Its main objective is to reach dispatch conditions. A representative example of a startup process of the plant in Figure 1 is described below. For simplicity, only the first part of a sequential startup is given, ending when the first gas turbine reaches emission compliance power.

Right after the gas turbine turbine rolls, the speed is held constant for a few minutes until the air in the HRSG is renewed, eliminating possible traces of unburnt hydrocarbons. When this purging process is completed, the gas turbine is ignited, and acceleration proceeds until the gas turbine reaches full speed condition, upon which the electrical generator is
connected to the grid. After this point, the gas turbine is kept at low power output while the HRSG continues to warm up and the drum pressures rise to recommended values for effective steam separation. When the steam production starts, it is diverted to the condenser by means of a steam bypass system. When steam production is enough to guarantee a sustained steam turbine acceleration, the steam control valves open slightly to accelerate it to the rated speed. When this condition is reached, the electrical generator mounted to the steam turbine shaft is connected to the grid. From this point on, the steam bypass valves are gradually closed, the steam control valves are gradually open to the fully open position (see Figure 1), and the gas turbine power is increased to reach the dispatch point, which is considered the end of the start process. These startup stages (see Table I) have to be followed in sequential order; each stage has its own objective and terminal constraints to enable transition to the next stage.

### TABLE I

**REPRESENTATIVE STARTUP MODES FOR CC PLANT**

<table>
<thead>
<tr>
<th>Start Mode</th>
<th>Main Goal</th>
<th>Duration (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRSG purge</td>
<td>renovate air in HRSG</td>
<td>20</td>
</tr>
<tr>
<td>GT acceleration</td>
<td>GT at 100% speed</td>
<td>5</td>
</tr>
<tr>
<td>HRSG, ST warmup</td>
<td>reach minimum steam production</td>
<td>20-60</td>
</tr>
<tr>
<td>ST acceleration</td>
<td>ST at 100% speed</td>
<td>7-30</td>
</tr>
<tr>
<td>Loading</td>
<td>GT at emission compliance</td>
<td>10-70</td>
</tr>
</tbody>
</table>

2) **Constraints:** Because of the delicate nature of the startup procedures, additional constraints must be considered. One constraint is associated with elements with thick metallic walls, such as elements exposed to thermal transients: steam turbine rotors, high pressure drums, and steam pressure headers. During startup, these elements can develop large thermal stresses due to fast temperature changes. The cumulative effect of thermal stresses through the cyclic operation of the plant contributes to a low cycle fatigue (LCF) damage (i.e., aging effect). Excessive stress and fatigue can ultimately take the plant out of operation for long periods of time. For an idea of the economic impact of this problem, we note that an off-line plant with a rated power of 500 MW can lose over 1,000,000 USD of revenue in a single day.

In plants using drum-type boilers, the drums levels need to be limited from above and from below. The upper limit prevents water injection in the steam pipes, heat exchangers, and turbines, which can cause erosion and corrosion damage. The lower limit prevents dryout of evaporator tubes, which can cause severe material damage from overheating. Maximum steam temperatures along the steam pipes and heat exchangers are typically handled by water spray systems to decrease local steam temperatures and prevent material damage. Maximum limits in condenser pressure are imposed to prevent damage in the last stages of the low pressure steam turbine due to overheating.

3) **Uncertainties:** Significant uncertainty exists in the startup process because of the limited set of sensors available to determine the true state of critical plant components (e.g., wall temperatures). In addition, some of these sensors are not effective when the plant is operating at low loads, flows, and pressures. The total startup time also has limited predictability because a multiplicity of small events can delay the process. For example, the speed hold events during acceleration of the steam turbine or the warm-up time required before allowing steam in the steam turbine is not known before the start. Many other startup events such as the time necessary to draw vacuum in the condenser or the time to reach acceptable water chemistry, also contribute to the uncertainty in the startup time. From the plant operator’s point of view, the consequence of startup time uncertainties is that the startup process is typically longer and costlier than necessary. From the ISO’s point of view, the startup uncertainty represents variations in the expected power injected to the grid, leading to reliability issues.

### C. Modeling

Normal and startup operations of a CC plant are characterized by having different modes in which certain equipment or functions are turned on or off. Because of this, there exist hybrid or integer decisions that need to be made. Depending on the optimization objectives and operational domain, one can formulate the model as either a hybrid or multi-stage dynamic system. A general hybrid system takes the following form:

\[
\dot{z}(t) = \sum_{j=1}^{M} \omega_j(t) \cdot f_j(z(t), u(t), \omega(t)), \quad \Gamma(\omega(t)) = 0, \quad (1)
\]

with \( t \in [0, T] \). Here, \( \omega_j(t) \in [0, 1] \) is a binary control decision describing operating mode \( j \) at time \( t \), \( z(t) \) are the states, \( u(t) \) are the continuous controls, and \( \omega(t) \) are the disturbances. Each model component \( f_j(\cdot) \) describes the operation of the plant for a given mode \( j \). In this formulation, \( \omega_j(t) \) and \( u(t) \) are the degrees of freedom. In plant operations there will also exist operational logic dictating which mode can be active before or after another; this is considered from the constraint \( \Gamma(\omega(t)) = 0 \). For instance, this constraint captures the transition dependencies depicted in Figure 2, and other logic such as the fact that only a single mode can be active at each point in time so that \( \sum_{j=1}^{M} \omega_j(t) = 1 \), \( t \in [0, T] \).

The models can also include mode-dependent constraints of the form

\[
g_j(z(t), u(t), \omega(t)) \geq (1 - \omega_j(t)) \cdot K, \quad t \in [0, T], \quad (2)
\]

where \( K \) is a large, positive number. Typically, the control functions \( \omega_j(\cdot) \) and \( u(\cdot) \) are discretized by using piecewise linear functions over \( N \) intervals \( t \in [t_j, t_j + T_j] \), \( j = 0, \ldots, N - 1 \), where \( N \) is the number of intervals, or stages, and \( T_j \) are the stage durations. If the number and sequence of modes \( \omega_j(t) \), \( j = 1, \ldots, M \), \( t \in [0, T] \), are given, one can formulate the above problem as a multi-stage
dynamic system of the form
\begin{align}
\dot{z}(t) &= w_j(t) \cdot f_j(z(t), u(t), \omega(t)) \tag{3a} \\
w_j(t) &= \text{given}, \quad t \in [t_j, t_j + T_j] \tag{3b} \\
z(t_{j+1}) &= z(t_j + T_j), \tag{3c}
\end{align}
with \( j = 0, \ldots, N - 1 \). We also have \( t_0 = 0 \) and \( t_{N-1} = T \). In this formulation, the stages' duration \( T_j \) and the controls \( u(t) \) are the degrees of freedom. Note that the multi-stage formulation is a special case of the more general hybrid formulation (1). In normal operation, the binary decision \( w(\cdot) \) represents, for example, the set of operating configurations and the transition processes described in Section II. In this case, there is no pre-established mode sequence, but there are transition rules, as depicted in Figure 2. Therefore, the normal operation can be described as a hybrid system (1, 2). In the case of startup operation, the binary decision variables define at which stage of the startup sequence the plant is evolving at any time \( t \in [0, T] \) (see Section II-B). For startup operation, there is a well-defined sequence for operating modes that has to be respected for every start. The startup operation therefore can be represented as a multi-stage problem (3, 2).

III. Electricity Market Operations

Capacity markets (day-ahead and real-time) are managed centrally at a given geographical region by the ISO. In U.S. markets, the ISO receives piece-wise linear bidding curves from suppliers (GENCOs) reflecting the price to be paid for a given generation level. These curves are expressed as piece-wise linear cost functions that we represent as \( p(q, \alpha^*) \) [13]. Here, \( q \geq 0 \) is the generation quantity, and \( \alpha^* \geq 0 \) are the supply curve parameters. The ISO also receives demand bid curves from utility companies and other consumers of the form \( p(d, \beta^*) \), where \( d \geq 0 \) is the demand quantity and \( \beta^* \geq 0 \) are the demand curve parameters. Another important cost is the startup cost of the unit which can be represented simply as \( c(u) \), where \( u \) is a binary variable indicating the on/off status of the unit, and \( c(u) > 0 \) if \( u = 1 \) and \( c(u) = 0 \) otherwise. Certain consumers and suppliers do not have control over their demands and generation capacities (e.g., renewable generation) so they provide expected values (forecasts) of their quantities, which we denote as \( w \).

The ISO will also receive parameters reflecting the operational limits of the units. Typical parameters include minimum and maximum outputs \( q \leq q \leq \bar{q} \) and the so-called ramping limits \( \underline{r} \leq \Delta q \leq \bar{r} \), which bound generation decrements/increments \( \Delta q \) between consecutive time periods [20]. We denote the operational limits of a unit using a single parameter \( \bar{q}^* \). In the case of CC units, the ISO will receive parameters representing the operational limitations of the system for different modes of operation [11], [10], [4].

A. Day-Ahead Markets

In the so-called day-ahead market, the ISO will clear the market by balancing supply and demand revenue over a horizon of 24-72 hours with time increments of one hour. This is done by solving the day-ahead market clearing problem (unit-commitment problem):
\begin{align}
\min_{u, q, d} \quad & c(u) + u \cdot \int_{0}^{d} p(Q, \alpha^*) dQ - \int_{0}^{d} p(D, \beta^*) dD \tag{4a} \\
\text{s.t.} \quad & \Pi_1(q, d, w) = 0 \quad (\pi) \tag{4b} \\
& \Pi_2(u, q, \bar{q}^*) \geq 0. \tag{4c}
\end{align}
Here, (4a) is the negative welfare and (4b) is the clearing condition balancing supply and demand over the transmission network given by the connectivity function \( \Pi_1(q) \). Constraints (4c) are logic constraints capturing minimum up/down time constraints and hybrid operational logic. These constraints implicitly assign operational limits depending on the units mode of operation. For a detailed structure of this problem and more detailed formulations, we refer the reader to [3], [10]. The solution of the unit commitment problem yields a commitment \( u^* \), generation \( q^* \), and demand \( d^* \) schedules over the time horizon. The multiplier of the clearing constraint (4b) are the locational marginal prices \( \pi^* \) over the network, which are the prices the suppliers will be paid for their committed generation and the consumers will be charged for their demands. We note that one of the key reasons for having a day-ahead market is that starting generation plants can take several hours. In addition, since grid storage is limited, ramping generation up and down is insufficient to balance the grid at short time scales.

For our discussion, we distinguish between two types of commitment structures: central commitment and self-commitment. In central commitment, the ISO receives physical information about the generators to compute their schedules. This is the structure that prevails in current operations. In self-commitment, the ISO allows suppliers to bid into the market by providing their supply curve, but no physical information is required. These types of players include firms with multiple generation facilities and virtual bidders. The supplier is responsible for providing the committed quantities either by physically generating that quantity or by buying it from other market participants.

B. Real-Time Markets

Several uncertainties are present during the day-ahead clearing procedure. Of special interest in our discussion is the fact that operational limits \( \bar{q}^* \) provided in the bidding procedure do not often capture the actual physical limitations of the units. For instance, PJM has reported that generation schedules cannot often be tracked efficiently because of misspecification of capacity and ramp limits [21]. Misspecification of operational limits is mostly due to the fact that actual physical limits depend on the actual generation level (i.e., they are state-dependent) and to the fact that the plant behavior is nonlinear. In addition, on-site operational limits introduce additional nonlinear behavior. Consequently, operational limits must not be assumed to be constant or state-independent. We observe that these simplifications are usually made because of limited computational capabilities at the ISO level. These limitations can in principle be ameliorated by delegating complexity to market players through...
self-commitment practices that would lead to a more scalable solution. We discuss this issue in more detail in Section III-D.

Because of uncertainties and simplifications of physical behavior, a recursive market called the real-time market is needed to balance supply and demand at shorter time scales. The real-time market normally clears every five minutes with predictions horizons of a few hours or less (look-ahead dispatch) [23]. Typically, the real-time market corrects generation and demand levels in the system while operational modes are fixed. This is done by receiving real-time bidding curves and operational limits information, which now reflect the fact that generators are positioned at their scheduled levels $q^*$. Variations of the clearing procedure sometimes allow for intraday commitment decisions to turn on peaking units (such as CC plants) that can be started within a few hours and ramped up and down in minutes. For simplicity, we can represent the real-time market clearing problem as

$$
\min_{\delta_q, \delta_d} c(u^*) + u^* \cdot \int_0^{q^* + \delta q} p(Q, \alpha_r) dQ - \int_0^{d^* + \delta d} p(D, \beta_r) dD
$$

s.t.

$$
\Pi_1(q^* + \delta q, d^* - \delta d, w) = 0 \quad (\pi_r) \quad (5a)
$$

$$
\Pi_2(u^*, q^* + \delta q, \eta_r) \geq 0. \quad (5b)
$$

Here, $\alpha^*_r, \beta^*_r, \eta^*_r$ are the bidding parameters and operational limits in the real-time market, while $\delta q, \delta d$ are correction quantities for generation and demands. After clearing, the real-time locational marginal prices are $\pi^*_r$, and the scheduled generation and demand quantities are $q^* + \delta q^*$ and $d^* - \delta d^*$, respectively. Depending on prevailing market conditions, the generator has the flexibility to decide not follow the scheduled generation quantity $q^*$ and to instead pay for the difference at the prevailing real-time market price.

In the absence of uncertainty, a real-time market is unnecessary. This ideal scenario represents maximum market efficiency reflected by a convergence of day-ahead and real-time prices [14]. The limited physical knowledge of the units by the ISO, however, will ultimately result in divergence of day-ahead and real-time prices [22]. This divergence can in turn diminish incentives to participate in the market due to an increased perceived risk by the participants.

### C. Bidding Strategies

Suppliers bid into the market to maximize their economic surplus. To do so, they need to compute their bidding parameters $\alpha$ so as to maximize profit. A complication is that, at the moment of bidding, both suppliers and consumers do not know the clearing price so they cannot assess their profit with full certainty [22]. Typically, price forecasts are used to determine the supply function parameters $\alpha$ while operating limits $\eta$ are chosen based on operational experience and plant knowledge. The bidding problem for the day-ahead market can be posed in abstract form as

$$
\max_{\alpha} \quad p \cdot q(p, \alpha) - \varphi(q(p, \alpha)) \quad (6a)
$$

s.t.

$$
\Xi_1(q(p, \alpha), u, \omega) = 0 \quad (6b)
$$

$$
\Xi_2(q(p, \alpha), \eta, u, \omega) \geq 0. \quad (6c)
$$

Here, $q(p, \alpha)$ is the inverse supply function specified in terms of the expected value of the price $p$. The operating costs are given by $\varphi(\cdot)$, and the mapping $\Xi(\cdot)$ represents the actual power plant system. This maps the power output $q(p, \alpha)$ to the the operational degrees of freedom $u$ and the uncertain disturbances $\omega$. The degrees of freedom can include set-points of operating conditions and manipulated variables such as scheduling logic and fuel flow rates, while the disturbances include ambient conditions. The solution of the bidding problem results in a bidding strategy given by $\alpha^*$ and $\eta^*$.

After generation is committed to the levels $q^*$ in the day-ahead market at price $p^*$, the supplier determines its bidding in the real-time market by solving the following problem:

$$
\max_{\alpha_r} \quad p^* \cdot q^* + p_r \cdot (\delta q(p_r, \alpha_r)) - \varphi(q^* + \delta q(p, \alpha)) \quad (7a)
$$

s.t.

$$
\Xi_1(q^* + \delta q(p_r, \alpha_r), u, \omega) = 0 \quad (7b)
$$

$$
\Xi_2(q^* + \delta q(p_r, \alpha_r), \eta_r, u, \omega) \geq 0. \quad (7c)
$$

Here, $p_r$ is the expected value of the real-time price. The solution of this problem gives the bidding strategy $\alpha^*_r$ and $\eta^*_r$. Note that if the generator decides to produce less than committed in the day-ahead market, then $q^* + \delta q(p_r, \alpha_r) < q^*$, and this induces the real-time penalty $p_r \cdot (\delta q(p_r, \alpha_r)) < 0$. Depending on the generation costs $\varphi(q^* + \delta q(p, \alpha), u, \omega)$ and the prices $p_r$, it might be worth paying this penalty.

![Fig. 3. Inconsistency of ramp approximation of dynamic response.](image)

### D. Inconsistencies in ISO-GENCOs Interface

One of the most critical inconsistencies arising in the coordination of ISO and GENCOs arises from the limited physical information captured in clearing operations. An important parameter used by the ISO is the ramping limit $r > 0$, which bounds the generation move between consecutive time points:

$$
-r \leq q(t + \Delta) - q(t) \leq r. \quad (8)
$$

The ramp concept was introduced to capture the dynamic limitations (e.g., response time) of the plant [20]. A problem
with this concept is that it assumes that the plant dynamics are state-independent. To see this, we write the above system of inequalities as

\[ q(t + \Delta) - q(t) = \Delta \cdot q(t), \quad -\Delta \cdot r \leq \Delta q \leq \Delta \cdot r. \]  

This is a discretization of the system \( \dot{q}(t) = \Delta q(t) \) for which the right-hand side is state-independent. Note that this formulation is even weaker than the traditional first-order dynamic model \( \dot{q}(t) = \kappa \cdot (q(t) - \bar{q}(t)) \), where \( \kappa \) is the controller gain and \( \bar{q}(t) \) is the set-point signal. Consequently, the ramp limit is a highly coarse representation of the plant dynamics. In Figure 3 we illustrate the approximation error of a first-order response. In the actual physical system, the dynamic response and limitations are affected by many factors, including metal capacitances and combustion efficiencies (implicitly captured in the dynamic model) and thermal stress constraints (captured by operational constraints). These factors lead to nonlinear behavior that cannot be captured through ramp limits. In CC plants, the situation is even more complex because each mode of operation presents a different set of dynamics and constraints. A similar situation occurs with the plant capacity limit, which is normally assumed to be state-independent. Currently, the plant capacity is typically correlated to ambient temperature by the ISOs, but this might not be sufficient to capture the entire physical complexity of the plant [18].

The limited ability of the ISO clearing formulation to capture the plant physical constraints introduces unreachable generation signals [21]. These in turn lead to distorted prices, market efficiencies, and grid reliability issues. As an example of the magnitude of this problem, we consider the Illinois system reported in [23]. For an average day the total generation is around 2,600 GW, with a share of 8% from natural gas peaking units. A typical average cost for units not using natural gas (baseload) is 30 USD/MWh while that for natural gas units is 80 USD/MWh. This gives an average cost (price) of 34 USD/MWh and a total cost of 88,400,000 USD. If an error due to misspecification of ramp limits or capacity requires an increase in the real-time market of 10% over the projected day-ahead capacity, the total generation increases to 2,860 GW with a share of 17% for natural gas since the peaking units would have to adjust the mismatch. This translates into a price of 38.5 USD/MWh and a total cost of 110,110,000 USD, an increase of 24%. As can be seen, since the cost difference between peaking and baseload units is high, misspecifications in day-ahead clearing can introduce important economic penalties.

IV. MODEL PREDICTIVE CONTROL

A. Supervisory Functions

The capabilities of MPC as a supervisory control system for tracking output set-points are well understood [16]. MPC enables the handling of constraints and multivariable interactions in a systematic manner. In CC plants, MPC has been used for load-following and startup functions. Typically, the power plant output is specified, and optimal control policies are determined to reach it in minimum time. Recently, additional functions have been proposed to optimize a given economic objective. For instance, fuel minimization has been used as a typical objective in startup operations [1], [9], [6]. We can specify a canonical cost-oriented MPC formulation as follows:

\[
\min_{w_j(\cdot), u(\cdot)} \int_0^T \varphi(q(t))dt \tag{10a}
\]

\[
\text{s.t. } \dot{z}(t) = \sum_{j=1}^M w_j(t) \cdot f_j(z(t), u(t), \omega(t)) \tag{10b}
\]

\[
\Gamma(w(\cdot)) = 0 \tag{10c}
\]

\[
g_j(z(t), u(t), \omega(t)) \geq (1 - w_j(t)) \cdot K \tag{10d}
\]

\[
q(t) = \Xi(z(t), w(t), u(t), \omega(t)) = \bar{q}(t). \tag{10e}
\]

Here, \( q(\cdot) \) is the plant power output (a function of the states, controls, and disturbances), and \( \bar{q}(\cdot) \) is the desired output sent by the ISO. Typically, the objective is to minimize operational cost while tracking the desired output. On a daily operation, costs include fuel costs, variable operations and maintenance cost, and startup cost. For a representative CC plant the relative size of these costs are fuel (85%), variable O&M (8%), and startup costs (7%).

One of the limitations of the above formulation is that it tracks power output strictly. Market operations, however, enable the operator to decide whether to track this signal and instead buy/sell power at real-time price. This is a decision that the operator typically performs independently of the control function. For this, the operator uses basic knowledge of the capacity and dynamics of the plant to determine how much to produce and how much to buy/sell in the market to maximize profit and then use the control system to track the resulting generation policy. One can extend the cost-oriented MPC formulation (10) to optimize the GENCO profit directly, thus avoiding the need to decompose the decision process in two hierarchical levels. The profit-
oriented formulation takes the form

\[
\max_{\delta q(t)} \int_0^T \left[p^*(t) \cdot q^*(t) + p_\alpha^*(t) \cdot (\delta q(t)) - \varphi(q^*(t) + \delta q(t))\right] dt
\]

s.t. (10b) - (10e), \quad (11a)

Where \( q^*(\cdot) \) is given by day-ahead commitment, the day-ahead market prices are given by \( p^*(\cdot) \), and the real-time prices are \( p_\alpha^*(\cdot) \). After the market has cleared, the day-ahead prices are known a day or two in advance while the real-time prices are known five minutes in advance. Consequently, the real-time price signals need to be forecast if the horizon \( T \) goes beyond five minutes. One can envision using a stochastic formulation to account for uncertainty in the real-time prices. This would also be useful in accounting for uncertainty in the disturbances \( w(\cdot) \). A stochastic formulation would take the following form:

\[
\max_{\delta q(t)} \mathbb{E}_{p(\cdot), w(\cdot)} \int_0^T \left[p^*(t) \cdot q^*(t) + p_\alpha^*(t) \cdot (\delta q(t)) - \varphi(q^*(t) + \delta q(t))\right] dt
\]

s.t. (10b) - (10e), \quad \forall \omega(\cdot) \in \Omega. \quad (12a)

Here, \( \Omega \) is the support of the joint distribution \( \mathbb{P} \) for \( p(\cdot) \) and \( w(\cdot) \), and the constraints hold almost-surely (for all realizations in \( \Omega \)). In the above formulation, the controls \( u(\cdot) \) and \( w(\cdot) \) are here-and-now decisions (scenario-independent) since these need to be implemented.

B. Market Functions

An important advantage of MPC is that, through the model, it provides important information about physical and dynamic constraints of the system. This can be used to enable more informed and profitable bidding decisions. Consider the day-ahead bidding formulation:

\[
\max_{w_j(\cdot), u(\cdot), \alpha(\cdot)} \int_0^T \left[p(t) \cdot q(p(t), \alpha(t)) - \varphi(q(p(t), \alpha(t)))\right] dt \quad (13a)
\]

s.t. (10b) - (10e). \quad (13b)

The solution of this problem yields the bidding policy \( \alpha(\cdot) \) to be sent to the operator. This bidding policy is an implicit function of the optimal schedule \( w_j(\cdot), u(\cdot) \). The solution also gives the production policy \( q(p(\cdot), \alpha(\cdot)) \). A limitation of this formulation, however, is that the clearing prices and thus the production quantities are uncertain. If the clearing prices turn out to be different from those expected, then the resulting production policy will be at risk of not being able to be reached because of the dynamic limitations of the plant. Also, unexpected disturbances might render the cleared policy unreachable or suboptimal. To address these limitations one can embed uncertainty in the price to ensure that the range of potential clearing policies is reachable.

A stochastic bidding formulation has the form in (14). In this formulation, the bidding parameter policy \( \alpha(\cdot) \) is a here-and-now decision (sent to the ISO) while the rest of the variables are scenario-dependent. This formulation yields a bidding policy that remains feasible for all realizations of uncertainty in prices and disturbances. A similar extension is possible for real-time market bidding.

\[
\max_{w_j(\cdot), u(\cdot), \alpha(\cdot)} \mathbb{E}_{p(\cdot), \omega(\cdot)} \int_0^T \left[p(t) \cdot q(p(t), \alpha(t)) - \varphi(q(p(t), \alpha(t)))\right] dt
\]

s.t. (10b) - (10e), \quad \forall \omega(\cdot) \in \Omega. \quad (14a)

V. TECHNICAL ISSUES

A. Self-Commitment vs. Central Commitment

From the GENCO’s perspective, self-commitment provides more flexibility in exploiting multiple resources to provide capacity. From the ISO perspective, an advantage of enabling self-commitment is that all the complexity of the power plant does not need to be handled in the clearing formulation. This is particularly critical in the case of CC plants, which exhibit nontraditional operations with multiple modes. This makes, for instance, capacity and ramp limits state- and mode-dependent. Currently, the ISO develops specialized mixed-integer linear formulations to capture this additional complexity. This approach comes at the expense of a drastically larger mixed-integer program with at least an order of magnitude more integer variables [4]. This computational complexity will dramatically increase as more CC units are installed because of the expected expansion in renewable generation. We note that even if such an optimization problem can be solved, capturing physical complexities using mixed-integer models is complicated and can lead to unreachable clearing signals. Consequently, we believe that deploying MPC at a large scale can dramatically reduce the computational complexity and mitigate unnecessary uncertainty in ISO clearing problems.

B. Computational Issues

The proposed MPC formulations pose many computational challenges. The resulting formulations are optimal control problems with the following characteristics:

- **Nonlinearity**: The use of MPC in high-level economic functions requires detailed first-principles models. First-principles models of a CC plant can contain hundreds of dynamic states and thousands of algebraic states with significant nonlinearity and algebraic coupling. These models share many of the complexities of other industrial models such as stiffness and strong nonlinearity introduced by thermodynamic relationships. Also, CC plants are tightly interconnected because of more stringent requirements in heat recovery. Consequently, models are not as sparse and are difficult to decompose by linear algebra routines.
• **Hybrid Decisions**: The hybrid nature of combined cycle units poses many challenges in the number of integer decisions and operational logic that needs to be handled in the formulation. Recent advances in mixed-integer control [17], complementarity formulations [2], and differential variational inequalities [15] can be leveraged to handle this type of problem. Another possible approach, commonly used in the polymer industry, is to decompose the hierarchy of scheduling and control [7]. The key challenge is to determine the right decomposition and aggregation strategy. Currently, no such formulations exist for CC plants. An alternative is to use multi-stage formulations [19] although this formulations introduce significant ill-conditioning.

• **Time Resolutions**: The horizons and time resolutions coupled to the relatively slow dynamics of certain power plants components pose important challenges. The day-ahead market runs at a resolution of one hour with horizons of up to 72 hours. The real-time market is currently desired to run with similar horizons but at resolutions of five minutes or less [23]. This is necessary to achieve consistency between markets and to handle dynamic limits more efficiently. Extending the horizon of the real-time market problem leads to extremely high-dimensional problems (in the number of controls or the integration horizons or both). It is not clear which solution approach (direct discretization, shooting) would be appropriate.

• **Uncertainty**: MPC formulations must be extended to account for uncertainties resulting from market interactions. In addition, these formulations need to be coupled with appropriate forecasting and uncertainty quantification capabilities. Weather is a major uncertainty that affects the plant capacity, economics of a power plant, and grid loads [5]. Recent advances in weather forecasting can be leveraged to tackle this uncertainty.

### VI. Conclusions

This paper discusses the implications of deploying MPC at a large scale in electricity markets. We claim that deploying MPC in conjunction with self-commitment practices can dramatically reduce computational complexity of commitment and dispatch decisions, as well as structural uncertainties introduced by misspecification of plant dynamic limitations.

### Acknowledgments

This work was supported by the U.S. Department of Energy, under Contract No. DE-AC02-06CH11357.

### References


