A Hierarchical Architecture for Optimal Unit Commitment and Control of an Ensemble of Steam Generators

S. Spinelli, M. Farina, and A. Ballarino

***Abstract*—A hierarchical architecture for the optimal management of an ensemble of steam generators is presented. The subsystems are coordinated by a multi-layer scheme for jointly sustaining a common load. The high level optimizes the load allocation and the generator schedule, considering activation dynamics by a hybrid model. At medium level, a robust tube-based Model Predictive Control (MPC) tracks a time-varying demand using a centralized - but aggregate - model, whose order does not scale with the number of subsystems. A nonlinear optimization, at medium level, addresses MPC infeasibility due to abrupt changes of ensemble configuration. Low-level decentralized controllers stabilize the generators. This control scheme enables the dynamical modification o f the ensemble configuration a nd p lug a nd p lay operations. Simulations demonstrate the approach potentialities.**

***Index Terms*—Hierarchical control of large-scale network systems, Model predictive control.**

Manuscript received on 03 July 2020. Revised: 22 March 2021. Accepted: 27 June 2021. *(Corresponding author: S. Spinelli.)*

S. Spinelli and A. Ballarino are with Istituto di Sistemi e Tecnologie Industriali Intelligenti per il Manifatturiero Avanzato, Consiglio Nazionale delle Ricerche, Milano, Italy (e-mail: name.surname@stiima.cnr.it).

S. Spinelli and M. Farina are with Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy (e-mail: name.surname@polimi.it).

I. Introduction and problem statement

Steam is widely used in industrial processes, playing a primary role in production. In industrial applications requiring a large and possibly time-varying steam demand, a flexible and efficient generation solution is mandatory. Since boiler operation close to the lower generation limit is largely inefficient, in high-fluctuating demand scenarios the production efficiency can be very unsatisfactory. In these cases, a virtual generation plant constituted of a set of smaller units working in parallel can be a viable alternative to operate a single boiler on a larger range [1]. A set of cooperative smaller units can be reconfigured to produce what demanded, enabling a quick and optimal connection/disconnection of subsystems, and considering the current and/or forecasted demand.

In line with this vision, the main objective of this work is the proposal of a hierarchical control scheme for the optimal unit commitment (UC) and management of a group of steam generators that work in a parallel configuration to sustain a cumulative steam demand.

A. State of the art

The coordination of independent (or interdependent) subsystems towards a main target characterizes different industrial applications, e.g., smart grids and electrical

generation systems [2], thermal energy grids [3], building heating and cooling systems [4], and distribution networks of steam, water, or compressed air [5]. These complex plants share similar features: several (homogeneous) systems work in parallel to commonly supply an overall demand; each subsystem and the whole plant operate in a constrained range and the subsystems must cooperate in a scenario of limited shared resources. The studies referred above focus on the optimal load sharing among the parallel systems. Actually, two main aspects must be addressed in this context: (i) the unit commitment and economic dispatch of the subsystems; (ii) the dynamic control of the overall plant and of the single subsystems.

The two problems are characterized by different timescales and are commonly addressed separately. The UC optimization problem has been extensively studied in the context of electrical generation systems, where the scheduling is optimized to minimize the plant operating cost, while satisfying process (and market) constraints. Several approaches have been studied, both in the deterministic and stochastic framework: an extensive discussion about solution techniques can be found in the review papers [6], [7]. While several (meta-) heuristic methods and mathematical programming approaches have been tested in the literature, in this paper we address the solution of UC optimization by Mixed-Integer Programming (MIP), as it guarantees an efficient, flexible and accurate modeling framework.

In the context of combined cycle power plants, a MIP formulation for the scheduling of thermal units has been presented in [8], while a tighter formulation reducing the number of binary variables is presented in [9]. An extended formulation, that provides a generalized-mode model for each unit, is discussed in [10]. Discrete-time state-space model formulations can be easily implemented in MPC strategy to manage the plant in a receding horizon way, as discussed in [11], whose formulation permits only to describe the unit dynamics by ON/OFF modes. The one presented in [12], based on a hybrid system approach - and specifically on a Mixed-Logical Dynamical (MLD) model - can generalize the unit dynamics. Based on a similar approach, in [13], the authors have formulated the high-level UC problem for a small Combined Heat and Power (CHP) unit, composed of a fire-tube boiler and an internal combustion engine for power generation.

In [14], the UC problem is presented for a CHP plant with eight steam boilers working in parallel, where maintenance issues of the flexible boiler array are integrated in the cost function. The authors of [15] focus on the boiler load allocation problem, uncoupled from electricity generation aspects, in a multi-boiler configuration: the optimization is addresses by gradient search methods considering boiler efficiency versus steam load. Crucially, these works focus only on the solution of the scheduling problem and do not consider the dynamic control of these units.

On the opposite front, other researchers are concentrating on dynamic control issues, with particular application on networked steam boilers operating in parallel. In [16] an optimal control scheme is presented for the energy loss minimization and the primary management of heat production for multi-boiler industrial systems, comparing the optimal approach to the traditional cascade control. The control of a multiple boiler configuration based on a MPC is discussed in [17], with application to a paper mill plant, or in [18] for a coal-fired boiler house, where maintaining stable header pressure and boiler availability is of critical importance for downstream consumers.

In the research work [19], a supervisory control, designed by LQR approach, is studied for a set of boilers in parallel configuration: a dynamic feedback strategy allows to continuously change each boiler set-point,

while minimizing a combined cost. Taking into account the dynamics of all the individual boilers, this optimal control can cope with general disturbances. However, the dimension of the model of the group of boilers grows with the number of units, thus encountering scalability issues. Moreover, the scheme is not flexible t o dynamically manage the variation of the boiler number, i.e., enabling plug-and-play capabilities.

In the recent years, some efforts have been devoted to provide unitary solutions to these problems. In this respect, decentralized, distributed, and hierarchical methods have many advantages over centralized ones, in view of their flexibility, r obustness (e.g., t o s ystem changes and demand variations), and scalability. In this work we focus on hierarchical methods, as the elective choice for optimal supervision and coordination of the system ensembles, e.g., as introduced in [20].

An extensive review of hierarchical and distributed approaches is reported in [21]. Recently, different solutions have been proposed based on the receding horizon approach. For example, [22] proposes a multi-rate solution for constrained linear systems based on reference governors, [23], [24]; on the other hand, in [25] a hierarchical scheme is introduced for coordinating independent systems with joint constraints and [26] extends the approach used in [25] in case of dynamically coupled units. Finally, [27] proposes a scalable solution based on finite i mpulse r esponse m odels e nabling plug-and-play operations, while [28] presents an application on power systems.

Notation: Calligraphic letters, $\mathcal{U}, \mathcal{Y}, \mathcal{W}, \mathcal{Z}$, indicate sets. The Minkowski sum of two sets is denoted by $\oplus$, while $\bigoplus_{i=1}^{N_\mathrm{g}} \mathcal{W}_i = \mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_{N_\mathrm{g}}$. Ensemble (resp. reference-model) variables are indicated with the notation $\bar{\cdot}$ (resp. $\hat{\cdot}$). Nonlinear models and linear counterparts are denoted by $\mathcal{S}$ and $\mathcal{L}$, respectively. Superscript $^{\mathrm{CL}}$ (resp. $^{\mathrm{OL}}$) connotes closed (resp. open) loop systems. Superscript $^{[\mathrm{M}]}$ (resp. $^{[\mathrm{H}]}$) denotes variables with sampling time T_M (resp. T_H), whose discrete time index is k (resp. h), referred to medium (resp. high) level. The floor operator is $\lfloor \cdot \rfloor$. Finally, for a generic variable $v(k)$, we denote $\Delta v(k) = v(k) - v(k-1)$.

B. Problem statement and paper contribution

In this work, we propose a hierarchical architecture for the management of an ensemble of steam generators. The aim is to manage a group of N_g steam generators, working in a parallel configuration to sustain a cumulative steam demand, $\bar{q}_\mathrm{s}^{\mathrm{Dem}}$. The objective is to guarantee the required steam flow rate, with the minimum operating cost. This implies both the minimization of fuel gas and the optimization of the network configuration (i.e., the partial contribution of each boiler to the overall demand), also considering the activation strategy.

The steam generator network is assumed to be composed of *similar* dynamical systems, i.e., having homogeneous quantities as inputs and outputs, but that might differ in physical dimensions, nominal production rate, consumption, and efficiency.

Each subsystem i is a water-tube boiler: a pressurized water, denoted feed-water $q_{\mathrm{f},i}$, circulates inside the tube coil, forced to flow by a displacement pump, and it is heated by a natural gas burner, whose flow rate is $q_{\mathrm{g},i}$. The heat, transmitted to the flowing fluid, induces a phase transition of the feed-water into steam. The steam flow rate generated is $q_{\mathrm{s},i}$. This design is characterized by extremely short start-up time and safe steam generation with respect to the fire-tube boiler configuration, due to the limited volume of water. The single subsystems and the network of generators are subject to input and output constraints. Both local and global variables are assumed to be defined in convex and compact sets, $\mathcal{U}_i$, $\mathcal{Y}_i$, $\bar{\mathcal{U}}$ and $\bar{\mathcal{Y}}$, i.e.

$$q_{\mathrm{s},i} \in \mathcal{U}_i \qquad\qquad q_{\mathrm{g},i} \in \mathcal{Y}_i \tag{1a}$$

$$\bar{q}_{\mathrm{s}} = \sum_{i=1}^{N_{\mathrm{g}}} q_{\mathrm{s},i} \in \bar{\mathcal{U}} \qquad \bar{q}_{\mathrm{g}} = \sum_{i=1}^{N_{\mathrm{g}}} q_{\mathrm{g},i} \in \bar{\mathcal{Y}} \tag{1b}$$

The proposed hierarchical control scheme consists of three layers.

The *high layer* (HL) extends the preliminary solution, proposed by the authors in [29] for a constant load demand, considering a time-varying demand and the discrete operating modes dynamics of the generators. To this aim, the model of the high-level behavior of the system is here defined i n d etail i n S ection II-D. This model is exploited by the top layer to optimize the strategy, i.e., the generator schedule and the working conditions, in order to minimize the operating costs. The activation/inactivation of units must consider the high-level state of each units and the transition costs. This layer computes the optimal number of units active and the best shares of production to be allocated to each boiler based on the time-varying profile o f the demand. With respect to [27] and [29], in this paper, the optimization program is reformulated on local steam flow r ates, i nstead o f d irectly o ptimizing t he sharing factors, which avoids to introduce mixed-integer bilinear constraints.

At the *medium layer* (ML), a robust MPC scheme is adopted, similarly to [20]. This layer, considering the ensemble model, allows to robustly track the overall demand. The ensemble model is an aggregate low-order model of the network of active systems, defined i n a scalable way. Differently from [20], in this work, we assume that the sharing factors can change during time. This condition must be opportunely handled by improving the formulation of the optimal control problem (OCP), in order to ensure at each time instant feasibility of the corresponding optimization program. We propose a procedure - based on an alternative nonlinear MPC program - to drive the ensemble to the new configuration when a sharp transition is not feasible.

At the *lowest layer* (LL), a set of decentralized controllers is used. Proportional-integral (PI) regulators, as currently used in industrial practice, stabilize the internal pressure to its set-point and track the individual requests. In this work, we opt for state-of-the-art regulators at low level, decoupled on pressure and flow-rate loops. This control layer exploits on purpose the embedded regulators, as provided by the generator producer, since the latter are actually neither open nor accessible for modifications, due to safety and regulatory issues. This choice permits to apply the proposed management architecture on brownfield, also on legacy systems.

II. The Boiler Models

In this section we present the dynamical model of the high-pressure steam generators used at the different layers. For notational simplicity, the index i will be dropped when clear from the context.

A. Nonlinear physical model

The continuous-time nonlinear dynamical model of the steam generator is derived from the drum-boiler model presented in [30]. Here the equations are adapted to the considered configuration: differently from drum-boilers, no accumulation exists in the water tubes and the drum is absent. In particular, the feed-water is forced

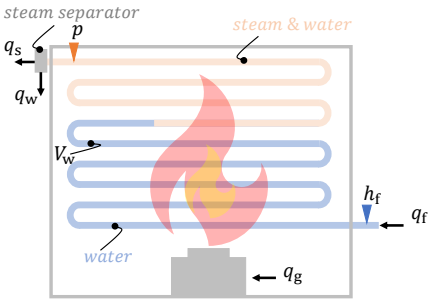

Figure 1: Steam generator functional scheme.

to flow at high-pressure through the heated tubes, with

flow-rate q_{f}. The heat transfer transforms the feed-water either totally or partially into steam. Therefore, the mass conservation equation on the water-tube control volume reads $q_{\mathrm{f}} = q_{\mathrm{s}} + q_{\mathrm{w}}$, where q_{s} is the steam flow-rate. The portion of flow that persists in liquid phase at the outflow, q_{w}, is assumed to be at saturated temperature, see Figure 1.

The i-th steam generator is characterized by a nonlinear dynamic model $\mathcal{S}_i^{\mathrm{OL}}$.

$$\dot{p} = \frac{1}{\phi}(\eta\lambda_H q_{\mathrm{g}} + q_{\mathrm{f}}(h_{\mathrm{f}} - h_{\mathrm{w}}) - q_{\mathrm{s}}(h_{\mathrm{s}} - h_{\mathrm{w}})) \quad (2)$$

$$\dot{V}_{\mathrm{w}} = \frac{1}{(\rho_{\mathrm{w}} - \rho_{\mathrm{s}})}(\frac{\partial \rho_{\mathrm{w}}}{\partial p}V_{\mathrm{w}} + \frac{\partial \rho_{\mathrm{s}}}{\partial p}V_{\mathrm{s}})\dot{p} \quad (3)$$

where

$$\begin{aligned}\phi = V_{\mathrm{s}}(h_{\mathrm{s}}\frac{\partial \rho_{\mathrm{s}}}{\partial p} + \rho_{\mathrm{s}}\frac{\partial h_{\mathrm{s}}}{\partial p}) + V_{\mathrm{w}}(h_{\mathrm{w}}\frac{\partial \rho_{\mathrm{w}}}{\partial p} + \rho_{\mathrm{w}}\frac{\partial h_{\mathrm{w}}}{\partial p}) + \\ V_{\mathrm{T}} + M_{\mathrm{T}}c_{\mathrm{p}}\frac{\partial T_{\mathrm{s}}}{\partial p} - (\frac{\partial \rho_{\mathrm{w}}}{\partial p}V_{\mathrm{w}} + \frac{\partial \rho_{\mathrm{s}}}{\partial p}V_{\mathrm{s}})\frac{(\rho_{\mathrm{w}}h_{\mathrm{w}} - \rho_{\mathrm{s}}h_{\mathrm{s}})}{(\rho_{\mathrm{w}} - \rho_{\mathrm{s}})}\end{aligned} \quad (4)$$

In equations (2)-(4), the subscripts $_{\mathrm{f,\,g,\,s,\,w}}$ refer to feed-water, fuel gas, steam, and internal water, respectively. Steam and internal water are assumed to be at saturated conditions. Therefore, the density ρ, the enthalpy h, and the temperature T are only function of internal pressure p.

The system is characterized by some specific parameters: the burner efficiency η, the gas low heat value λ_H, the total tubes internal volume V_{T}, the mass M_{T}, and the specific heat coefficient c_{p}.

The states of the nonlinear dynamical model (2)-(3) are the internal pressure p and the water volume V_{w}. The manipulable inputs are the feed-water flow rate q_{f} and the natural gas flow rate q_{g}, while the steam demand q_{s} is considered, at the low-level, as a disturbance term. Similarly, the enthalpy h_{f} of the feed-water is considered a known measured disturbance.

B. Low-level closed-loop model

An embedded controller is devoted to the regulation of the pressure at the set-point level, and to guarantee a constant water volume V_{w} for each subsystem $\mathcal{S}_i^{\mathrm{OL}}$. This controller acts on the local input variables q_{f} and q_{g}. Commercially-available boilers are already provided with low-level controllers for pressure regulation, designed on industrial standard configuration: a feedback PI regulator $\mathbf{R}$ on the fuel flow-rate, to steer the pressure p to a set-point p_{sp}, and a disturbance compensator $\mathbf{C}$ working, with an open-loop action, on the feed-water flow-rate to follow the steam demand, as depicted in Figure 2. The closed-loop system of the i-th boiler can

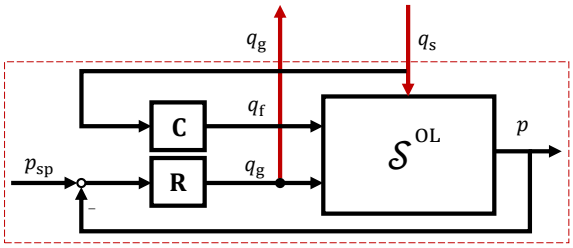

Figure 2: Closed-loop boiler function block diagram.

be described as a nonlinear dynamic model $\mathcal{S}_i^{\mathrm{CL}}$, in short denoted as $q_{\mathrm{g},i} = \mathcal{S}_i^{\mathrm{CL}}(q_{\mathrm{s},i})$.

One peculiarity of this closed-loop system is the possibility of considering the steam flow rate as input of the controlled system and the gas flow rate as an output, as shown in Figure 2. This closed-loop representation of the boiler enables the problem formalization in the framework of hierarchical control of ensemble systems, as in [20].

In Figure 3, the input/output static map at steady state is shown: historical static data are compared with data generated simulating the response of the system $\mathcal{S}_i^{\mathrm{CL}}$ with a multiple step input profile. An affine approximation is also shown. Note that, although this linear model is valid during production where the pressure is regulated at its set-point, non-linearity is still relevant during start-up.

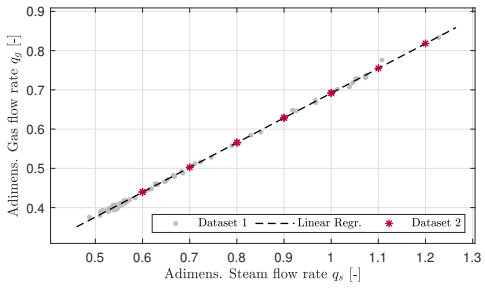

Figure 3: Input-output static map of the controlled steam generator at steady state.

C. Affine model for medium-level control

Consistently with the data reported in Figure 3, in production the boiler is maintained close to the nominal conditions, thus the dynamics of $\mathcal{S}_i^{\mathrm{CL}}$ can be well represented by an affine dynamic model, used to account for transient response. A discrete-time affine system $\mathcal{L}_i^{\mathrm{CL}}$ with output $y(k) = q_{\mathrm{g},i}(kT_{\mathrm{M}})$ and input $u(k) = q_{\mathrm{s},i}(t)$ constant $\forall t \in [kT_{\mathrm{M}}, (k+1)T_{\mathrm{M}})$ is identified with the simulation error minimization approach using the data drawn by exciting the controlled nonlinear model $\mathcal{S}_i^{\mathrm{CL}}$ with multiple-step inputs. Note that the sampling time is T_{M} and the time index k is the one used for control at medium hierarchical level. The identified discrete-time transfer function (plus constant) is denoted G_i^{CL} and is of the type

$$y(k) = \frac{\sum_{j=1}^{n_{\mathrm{b}}}(b_j z^{-j})}{1+\sum_{j=1}^{n_{\mathrm{f}}}(f_j z^{-j})}u(k) + \gamma \tag{5}$$

where γ is the identified bias when $u(k) = 0$. The corresponding state-space form is

$$\mathcal{L}_i^{\mathrm{CL}} : \begin{cases} x(k+1) = Ax(k) + Bu(k) \\ y(k) = Cx(k) + \gamma \end{cases} \tag{6}$$

with state vector, $x(k) = [\delta y(k), ..., \delta y(k-n_{\mathrm{f}}+1), u(k-1), ..., u(k-n_{\mathrm{b}}+1)]^{\mathrm{T}} \in \mathbb{R}^{n_{\mathrm{f}}+n_{\mathrm{b}}-1}$ and $\delta y(k) = y(k) - \gamma$. The matrices are $B =$ $\begin{bmatrix} b_1 & 0_{1\times(n_{\mathrm{f}}-1)} & 1 & 0_{1\times(n_{\mathrm{b}}-2)} \end{bmatrix}^{\mathrm{T}}$, $C = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}$, and

$$A = \left[\begin{array}{cc|cc} -f_1 & \cdots -f_{n_{\mathrm{f}}-2} & -f_{n_{\mathrm{f}}} & b_2 \dots b_{n_{\mathrm{b}}-1} \; b_{n_{\mathrm{b}}} \\ I_{n_{\mathrm{f}}-1} & & 0_{(n_{\mathrm{f}}-1)\times 1} & 0_{(n_{\mathrm{f}}-1)\times(n_{\mathrm{b}}-1)} \\ \hline & & 0_{1\times(n_{\mathrm{b}}-2)} & 0 \\ \multicolumn{2}{c|}{0_{(n_{\mathrm{b}}-1)\times n_{\mathrm{f}}}} & I_{n_{\mathrm{b}}-2} & 0_{(n_{\mathrm{b}}-2)\times 1} \end{array}\right]$$

Assumption 1: System (6), consistently with [26], enjoys the following properties:

1) A is Schur stable;
2) $m = p = 1$;
3) $g = C(I_n - A)^{-1}B \neq 0$

This model will be used to derive the ensemble model, as discussed in detail in Section III-C.

D. Hybrid automaton for high-level optimization

The boiler model used by the high-level optimizer operates on a coarser discrete-time grid, with a sampling time T_{H} and time index h. A hybrid automaton [31] is used, including both discrete and continuous states. The discrete variable m defines the operating modes: shut down (OFF), start-up (ST), and production (ON), i.e., $m \in \{\mathrm{OFF}, \mathrm{ST}, \mathrm{ON}\}$.

A simplified model is considered in each mode, where the fuel flow rate q_{g} depends on the steam demand q_{s}. In this paper, due to the small settling time t_{st} of the dynamic system (6) with respect to the high-level sampling time, $t_{\mathrm{st}} \ll T_{\mathrm{H}}$, a static input-output map for each operating mode is assumed. Thus, the dynamic state of the Hybrid Automaton (DHA) is the number of sampling times χ spent in the present operation mode. Namely, variable $\chi \in \mathbb{Z}_0^+$ is used for correctly model transitions. A continuous evolution, $\mathsf{f} : \chi(h+1) = \chi(h) + 1$ is valid when no transition occurs, i.e., $m(h+1) = m(h)$. Instead, any transition forces a reset to zero of the dynamic state, $\mathsf{r} : \chi(h+1) = 0$. A mode transition - see the Finite State Machine (FSM) in Figure 4 - depends

on the time spent in the current operating mode, $\chi(h)$, and possibly on switching binary input $\beta(h) \in \{0,1\}$ to 1. More specifically, a transition happens whenever a guard condition, g, is met:

$$\mathrm{g}: \begin{cases} \{\chi(h) \geq \chi_{\mathrm{OFF}\to\mathrm{ST}}\} \wedge \{\beta(h)=1\} & m(h)=\mathrm{OFF} \\ \{\chi(h) \geq \chi_{\mathrm{ST}\to\mathrm{ON}}\} & m(h)=\mathrm{ST} \\ \{\chi(h) \geq \chi_{\mathrm{ON}\to\mathrm{OFF}}\} \wedge \{\beta(h)=1\} & m(h)=\mathrm{ON} \end{cases}$$

The values of $\chi_{\mathrm{OFF}\to\mathrm{ST}}$, $\chi_{\mathrm{ST}\to\mathrm{ON}}$, and $\chi_{\mathrm{ON}\to\mathrm{OFF}}$, are suitably-defined thresholds. The model output is given by:

$$q_{\mathrm{g}}(h) = g \cdot q_{\mathrm{s}}(h) + \gamma_{\mathrm{ON}} \qquad \text{if } m=\mathrm{ON} \tag{7a}$$

$$q_{\mathrm{g}}(h) = \gamma_{\mathrm{ST}} \qquad \text{if } m=\mathrm{ST} \tag{7b}$$

$$q_{\mathrm{g}}(h) = 0 \qquad \text{if } m=\mathrm{OFF} \tag{7c}$$

where $g = C(I_n - A)^{-1}B$, γ_{ON}, and γ_{ST} are the static gain of the closed-loop system $\mathcal{L}^{\mathrm{CL}}$, the constant fuel gas consumption in production and in start-up modes, respectively. Note that, consistently with the model derived in the previous sections, the affine map (7a) is the one depicted in Figure 3. To make the model

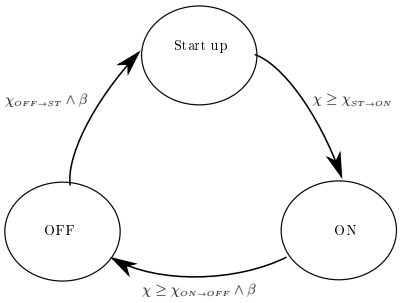

Figure 4: Boiler operation mode transitions.

easily manageable in a suitable optimization program, the DHA model is converted into the MLD one [32]. The MLD model is an extended state-space dynamical system where the state vector, $x^{[\mathrm{H}]} = \{\chi, x^{[\mathrm{H}]}_{\mathrm{OFF}}, x^{[\mathrm{H}]}_{\mathrm{ST}}, x^{[\mathrm{H}]}_{\mathrm{ON}}\} \in \mathbb{Z}\times\{0,1\}^3$, includes integer and Boolean variables. The inputs are the Boolean command and the steam flow-rate, $u^{[\mathrm{H}]} = \{\beta^{[\mathrm{H}]}, q^{[\mathrm{H}]}_{\mathrm{s}}\} \in \{0,1\}\times\mathbb{R}$, while the output is the consumed gas $y^{[\mathrm{H}]} = \{q^{[\mathrm{H}]}_{\mathrm{g}}\} \in \mathbb{R}$, which depends on the active mode, as in (7).

A set of Boolean and continuous auxiliary variables $\{\delta^{[\mathrm{H}]}, z^{[\mathrm{H}]}\} \in \{0,1\}^{n_\delta}\times\mathbb{R}^{n_z}$ is added to model the FSM evolution, the transition guards, and the reset maps. The MLD model takes the general form:

$$x^{[\mathrm{H}]}(h+1) = A^{[\mathrm{H}]}x^{[\mathrm{H}]}(h) + B^{[\mathrm{H}]}_u u^{[\mathrm{H}]}(h) + B^{[\mathrm{H}]}_z z^{[\mathrm{H}]}(h) + B^{[\mathrm{H}]}_\delta \delta^{[\mathrm{H}]}(h)$$

$$y^{[\mathrm{H}]}(h) = C^{[\mathrm{H}]}x^{[\mathrm{H}]}(h) + D^{[\mathrm{H}]}_u u^{[\mathrm{H}]}(h) + D^{[\mathrm{H}]}_z z^{[\mathrm{H}]}(h) + D^{[\mathrm{H}]}_\delta \delta^{[\mathrm{H}]}(h)$$

$$E^{[\mathrm{H}]}_x x^{[\mathrm{H}]}(h) + E^{[\mathrm{H}]}_u u^{[\mathrm{H}]}(h) + E^{[\mathrm{H}]}_z z^{[\mathrm{H}]}(h) + E^{[\mathrm{H}]}_\delta \delta^{[\mathrm{H}]}(h) \leq E^{[\mathrm{H}]}_{\mathrm{aff}}$$

III. The hierarchical control scheme

In the previous section we derived the single subsystem models to be used at the different levels. Now, we explain how to manage and control them in a unitary and coordinated way.

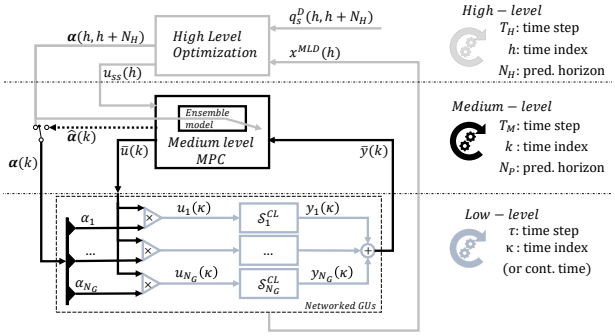

Figure 5: Steam generator ensemble and hierarchical scheme. Typically, $T_{\mathrm{H}} \in [10, 30]$ min, $T_{\mathrm{M}} \in [30, 60]$ s, $\tau \in [1, 10]$ s.

A. Sketch of the proposed control architecture

As shown in Figure 5, the medium and high levels of the hierarchical scheme are designed to concurrently define the input u_i of each subsystem (i.e., local steam flow-rate $q_{\mathrm{s},i}$) as

$$u_i = \alpha_i \bar{u} \qquad i = 1, \ldots, N_{\mathrm{g}} \tag{8}$$

where α_i is the sharing factor used to partition the overall ensemble input $\bar{u}$.

Sharing factors α_i are computed by the optimization layer. Here, thanks to the DHA models defined in Section II-D, they are optimized in a receding horizon way, minimizing the operating cost of the ensemble to supply the steam demand forecast. The sharing factors are time-varying and defined according to the slow time-scale (i.e., T_{H}).

The ensemble input $\bar{u}$ is instead computed by a dynamic optimal reference tracking problem at medium level. To do so, an aggregate model of the whole ensemble is derived by considering the subset of active generation units. The ensemble dynamical model is built by combining opportunely the closed-loop models of the controlled generators. By considering a unique ensemble model, medium level exhibits interesting scalability properties, as its dimensions do not grow with the number of subsystems. A robust reference-tracking MPC scheme is implemented to define the overall gas consumption of the ensemble, operating with a faster time-scale with respect to the high level, with sampling time $T_{\mathrm{M}} = T_{\mathrm{H}}$.

B. High-level Optimization

The high hierarchical level aims to optimize the sharing factor profiles $\alpha_i^{[\mathrm{H}]}(h)$ and the modes of all the subsystems by minimizing the operating expenses, including the subsystem activation costs, the actual start-up time, and other constraints, as the ones related to mode transitions and operational range of each subsystem in the ensemble.

The algorithm presented here extends the one presented in [20], [27], and [29] by solving the unit commitment in receding horizon along a prediction window with time-varying demand. We assume its profile to be known for the entire prediction horizon and approximated by a piece-wise constant function, $\bar{q}_{\mathrm{s}}^{\mathrm{Dem}}(h)$.

In [29], where both the sharing factors $\alpha_i^{[\mathrm{H}]}$ and the ensemble steady-state input $\bar{u}_{\mathrm{ss}}(h)$ were considered as decision variables, we obtained a MIP with bilinear inequality constraints. In this work the problem is reformulated as a simpler MIP with linear constraints by considering as optimization variables the partitioned steam flow rates $q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h)$. In this formulation, the optimal sharing factors are computed as $\alpha_i^{[\mathrm{H}]}(h) = q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h)/\bar{u}_{\mathrm{ss}}(h)$.

The optimization problem at high-level reads:

$$\min_{\boldsymbol{\beta}_i^{[\mathrm{H}]}, \mathbf{q}_{\mathrm{s}\,i}^{[\mathrm{H}]}} \quad \sum_{h=0}^{N_{\mathrm{H}}} \sum_{i=1}^{N_{\mathrm{g}}} l_i(h, \beta_i(h), q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h)) \tag{9a}$$

$$\text{s.t.} \quad \sum_{i=1}^{N_{\mathrm{g}}} q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h) \geq \bar{q}_{\mathrm{s}}^{\mathrm{Dem}}(h) \tag{9b}$$

$$\begin{cases} \sum_{i=1}^{N_{\mathrm{g}}} q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h) = 0 \\ \qquad \text{iff} \quad \sum_{i=1}^{N_{\mathrm{g}}} x_{\mathrm{ON}\,i}^{[\mathrm{H}]}(h) = 0 \\ \bar{u}_{\mathrm{m}} \leq \sum_{i=1}^{N_{\mathrm{g}}} q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h) = 0 \leq \bar{u}_{\mathrm{M}} \\ \qquad \text{otherwise} \end{cases} \tag{9c}$$

$$\begin{cases} 0 \leq \sum_{i=1}^{N_{\mathrm{g}}} q_{\mathrm{g}\,i}^{[\mathrm{H}]}(h) \leq \bar{y}_{\mathrm{ST}} \\ \qquad \text{iff} \quad \sum_{i=1}^{N_{\mathrm{g}}} x_{\mathrm{ON}\,i}^{[\mathrm{H}]}(h) = 0 \\ \bar{y}_{\mathrm{m}} \leq \sum_{i=1}^{N_{\mathrm{g}}} q_{\mathrm{g}\,i}^{[\mathrm{H}]}(h) \leq \bar{y}_{\mathrm{M}} \\ \qquad \text{otherwise} \end{cases} \tag{9d}$$

and, $\forall i = 1, \ldots, N_{\mathrm{g}}$

MLD model of unit i

$$u_{\mathrm{m},i} x_{\mathrm{ON}\,i}^{[\mathrm{H}]}(h) \leq q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h) \leq u_{\mathrm{M},i} x_{\mathrm{ON}\,i}^{[\mathrm{H}]}(h) \tag{9e}$$

$$\begin{aligned} & y_{\mathrm{m},i} x_{\mathrm{ON}\,i}^{[\mathrm{H}]}(h) + \gamma_{\mathrm{ST}\,i} x_{\mathrm{ST}\,i}^{[\mathrm{H}]}(h) \leq y_i^{[\mathrm{H}]}(h) \\ & \leq y_{\mathrm{M},i} x_{\mathrm{ON}\,i}^{[\mathrm{H}]}(h) + \gamma_{\mathrm{ST}\,i} x_{\mathrm{ST}\,i}^{[\mathrm{H}]}(h) \end{aligned} \tag{9f}$$

$\forall h = 0, \ldots, N_{\mathrm{H}}$

The decision variables are defined as a sequence of vectors along the optimization horizon, i.e., $\forall h = 0, \ldots, N_{\mathrm{H}}$: steam flow-rate, $\mathbf{q}_{\mathrm{s}\,i}^{[\mathrm{H}]}(h) = [q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h), \ldots, q_{\mathrm{s}\,i}^{[\mathrm{H}]}(h + N_{\mathrm{H}})]$ and the Boolean command for FSM transitions $\boldsymbol{\beta}_i^{[\mathrm{H}]}(h) = [\beta_i^{[\mathrm{H}]}(h), \ldots, \beta_i^{[\mathrm{H}]}(h+N_{\mathrm{H}})]$ of each boiler, i.e. $\forall i = 1, \ldots, N_{\mathrm{g}}$.

The cost function $J : \mathbb{R}^n \to \mathbb{R}$ is defined by summing the subsystems' stage costs $l_i(h)$ i.e., the operating cost related to the fuel consumption - based on natural

gas price λ_{g} - fixed operating cost connected to the production mode $\lambda_{\mathrm{ON}\,i}$ and the fixed startup costs $\lambda_{\mathrm{ST}\,i}$. The fixed costs are in general specific for each generator: they can include personnel, maintenance and degradation costs, that might increase for frequent start and stops.

$$\begin{aligned} l_i(h) = \;& \lambda_{\mathrm{ON}\,i}(x^{[\mathrm{H}]}_{\mathrm{ON}\,i}(h)) + \lambda_{\mathrm{ST}\,i}(x^{[\mathrm{H}]}_{\mathrm{ST}i}(h)) + \\ & \lambda_{\mathrm{g}}\tfrac{T_{\mathrm{H}}}{\rho_{\mathrm{g}}} \;\left[\left(g_i q^{[\mathrm{H}]}_{\mathrm{s}\,i}(h) + \gamma_{\mathrm{ON}\,i}\right) x^{[\mathrm{H}]}_{\mathrm{ON}\,i}(h) + \;\gamma_{\mathrm{ST}\,i} x^{[\mathrm{H}]}_{\mathrm{ST}\,i}(h)\right] \end{aligned} \tag{10}$$

Note that constraints (9c)-(9d) - enforced to guarantee (1b) - are defined by logical conditions. A so-called "Big-M" reformulation can be adopted to transform these conditional constraints in a set of mixed-integer inequalities [32].

We denote by $\bar{\cdot}_{\mathrm{m}}$ ($\bar{\cdot}_{\mathrm{M}}$) the minimum (maximum) values of inputs and outputs, while $\bar{y}_{\mathrm{ST}} = \sum_{i=1}^{N_{\mathrm{g}}} x^{[\mathrm{H}]}_{\mathrm{ST}\,i}(h)\gamma_{\mathrm{ST}\,i}$. At each step h, the optimizer computes the optimal trajectory of the sharing factors $\alpha^{[\mathrm{H}]}(j)$ for all $j = h, \ldots, h + N_{\mathrm{H}}$. Based on the receding horizon principle, the configuration $\alpha^{[\mathrm{H}]}(h)$, related to the first step, is broadcast to the network, while the rest of the trajectory is discarded (or, better said, kept as backup solution). At the subsequent step, $h+1$, the status of the GUs is retrieved, as well as an updated forecast of the future demand, moving forward the prediction horizon by one step. This strategy permits to correct the demand forecast of remote steps as soon as they come closer, thus adjusting inaccurate estimations. A new profile $\alpha^{[\mathrm{H}]}(j)$, with $j = h+1, \ldots, h+N_{\mathrm{H}}+1$, is computed by (9) and the solution $\alpha^{[\mathrm{H}]}(h+1)$ sent to the GUs.

Remark 1: The hard constraint (9b) can be tightened to equality, accelerating the solution convergence - if any feasible solution exists. Otherwise, if the program is infeasible, as the constraint (9b) cannot be satisfied for certain demand profiles, it can be relaxed thanks to a slack variable $\varepsilon \geq 0$ with the modified objective function (9a), $\hat{l} = l + \lambda_\varepsilon \varepsilon^2$ with the constraint $\sum_{i=1}^{N_{\mathrm{g}}} q^{[\mathrm{H}]}_{\mathrm{s}\,i}(h) \geq \bar{q}^{\mathrm{Dem}}_{\mathrm{s}}(h) - \varepsilon$.

Remark 2: We solve (9) in a centralized way, since the solution must be available with a frequency $f_{\mathrm{H}} = 1/T_{\mathrm{H}}$. However, its computational complexity scales with the number of generation units, which can be very large in some applications. To overcome this, one may implement (9) in a distributed fashion, as in [33], partitioning the set of generators in clusters.

Remark 3: The accuracy of demand forecast strongly impacts on the solution quality: since the reference is an additional decision variable, the feasibility is guaranteed. However, whenever the mismatch between demand forecast and its actual value is greater than a given threshold, the execution of the HL optimization can be triggered at an event-based "asynchronous" fashion to foster optimal tracking performances.

A good demand forecast is indeed one of the main challenges for practical implementation of any scheme aiming to schedule the generation units. Small scale generators, for medium-pressure steam, are usually operated in the industrial context where steam is considered a commodity resource. Therefore sometimes no demand forecast is available and neither considered for the generator management. Actually, accurate forecasts can be easily obtained by historical data and future production scheduling. Nowadays, companies that aim to implement energy efficiency strategies are increasing their awareness on energy utilization, through the analysis of historical data, and are pushed to implement procedures to correlate energy demand with production, giving the tools for deriving approximated evaluations of future steam demand to be used as input of the proposed management architecture.

C. Medium-level control

The ML controller regulates the ensemble based on the operating modes and the sharing factors defined by the higher layer, driving the ensemble input $\bar{u}^{[\mathrm{M}]}(k)$ to the

steady-state value, $\bar{u}_{\rm ss}^{[\rm H]}(h) = \sum_{i=1}^{N_{\rm g}} q_{{\rm s}\,i}^{[\rm H]}(h)$, computed by the HL optimizer. The medium-level MPC deals with an aggregate - small scale - model of the whole ensemble.

1) Reference models and consistency requirements: Medium level controller design requires, first of all, to devise an aggregate model of the ensemble. According to [20], a *reference* model must be derived for each subsystem, defined as

$$\hat{\mathcal{L}}_i : \begin{cases} \hat{x}_i^{[\rm M]}(k+1) = & \hat{A}\hat{x}_i^{[\rm M]}(k) + \hat{B}_i u_i^{[\rm M]}(k) + \hat{w}_i^{[\rm M]}(k) \\ \hat{y}_i^{[\rm M]}(k) = & \hat{C}\hat{x}_i^{[\rm M]}(k) + \hat{\gamma}_i \end{cases} \tag{11}$$

where this alternative model can be built on a possibly reduced state, defined as $\hat{x}_i^{[\rm M]} = \beta_i x_i^{[\rm M]}$, where $\beta_i \in \mathbb{R}^{\hat{n}\times n_i}$ is a suitable map, with $\hat{n} \leq n_i$. In addition, a term $\hat{w}_i^{[\rm M]}(k)$ is introduced to embed the error due to the mismatch between the reference model (11) and the identified system (6).

By design, the state matrix $\hat{A}$ and the output matrix $\hat{C}$ can be generically defined: they just must be the same for all subsystems' reference models. Conversely, the input matrix $\hat{B}_i$ must be accurately defined. It is advantageous to select $\hat{A}$, $\hat{B}_i$ and $\hat{C}$ with the same canonical structure of A_i, B_i and C_i, as defined in Section II-C. Using this convenient choice, the state-reduction map $\beta_i \in \mathbb{R}^{\hat{n}\times n_i}$ is merely a selection matrix, whose rows are basis vectors of the new canonical space. In this way the state of the reference models, is $\hat{x}^{[\rm M]}(k) = [\delta y^{[\rm M]}(k), \delta y^{[\rm M]}(k-1), ..., \delta y^{[\rm M]}(k-\hat{n}_{\rm f}+1), u^{[\rm M]}(k-1), ..., u^{[\rm M]}(k-\hat{n}_{\rm b}+1)]^{\rm T}$.

The input matrix of the reference system must be defined in order to satisfy the so-called *gain consistency* conditions (see [20]): the reference model (11) and the model (6) must guarantee to have the same static gain and a consistent output map. This is verified by imposing:

$$\hat{\gamma}_i = \gamma_{{\rm ON}\,i} \tag{12a}$$

$$\hat{b}_{i,1} = \frac{\sum_{j=1}^{n_{\rm b}} b_{i,j}}{1+\sum_{j=1}^{n_{\rm f}} f_{i,j}}(1+\sum_{j=1}^{\hat{n}_{\rm f}} \hat{f}_j) - \sum_{j=2}^{\hat{n}_{\rm b}} \hat{b}_j \tag{12b}$$

where $(b_{i,j}, f_{i,j})$ and $(\hat{b}_{i,j}, \hat{f}_{i,j})$ are the parameters of the i-th models (6) and (11), respectively.

2) Disturbance $\hat{w}_i^{[\rm M]}(k)$: As discussed, the term $\hat{w}_i^{[\rm M]}(k)$ embeds the error due to the mismatch between the reference model (11) and the original one (6) induced by the selection of the same state matrices for the reference models. To apply robust MPC for ensemble control, we need to ensure that $\hat{w}_i^{[\rm M]}(k)$ is bounded. In [20], it is shown that the set where $\hat{w}_i^{[\rm M]}(k)$ lies (i.e., $\mathcal{W}_i$) can be made small by properly restricting the set of $\Delta u_i^{[\rm M]}(k) = u_i^{[\rm M]}(k) - u_i^{[\rm M]}(k-1)$, i.e., $\Delta\bar{\mathcal{U}}_i$. However, the definition of $\mathcal{W}_i$ used in [20] requires the definition of a suitable invariant set, used to define the low-level MPC controller, which is here absent.

In any case the fact that $\mathcal{W}_i$ depends upon $\Delta\bar{\mathcal{U}}_i$ remains valid also in this framework, i.e., when the low-level controller is unconstrained. This is supported by the fact that $\hat{w}_i(k) = \beta_i x_i(k+1) - \hat{x}_i(k+1) = \beta_i[\delta y_i(k+1), \delta y_i(k), ..., \delta y_i(k-n_{\rm f}+2), u_i(k), ..., u_i(k-n_{\rm b}+2)]^T - [\delta y_i^o(k+1), \delta y_i^o(k), ..., \delta y_i^o(k-\hat{n}_{\rm f}+2), u_i(k), ..., u_i(k-\hat{n}_{\rm b}+2)]^T = [\delta y_i(k+1) - \delta y_i^o(k+1), \delta y_i(k) - \delta y_i^o(k), ..., \delta y_i(k-\hat{n}_{\rm f}+2) - \delta y_i^o(k-\hat{n}_{\rm f}+2), 0, ..., 0]^T$, where $\delta y_i^o(k)$ is defined as the output of the "unperturbed" reference system

$$\hat{\mathcal{S}}_i^o : \begin{cases} \hat{x}_i^o(k+1) = & \hat{A}\hat{x}_i^o(k) + \hat{B}_i u_i(k) \\ \delta\hat{y}_i^o(k) = & \hat{C}\hat{x}_i^o(k) \end{cases} \tag{13}$$

In view of this, each non-zero component of vector $\hat{w}_i(k)$ is a lagged version of $e_y(k+1) = \delta y(k+1) - \delta y^o(k+1)$. It is possible to show that, thanks to the gain consistency condition, there exists a transfer function $\Delta\mathcal{G}_i(z^{-1})$ such that[1]

$$e_i(k+1) = -\Delta\mathcal{G}_i(z^{-1})\Delta u_i(k) \tag{14}$$

[1]To retrieve (14) we can write, from (6) and (13), that

$$e_i(k+1) = G_i(z^{-1})u_i(k) - \hat{G}_i(z^{-1})u_i(k)$$

Following [34], the set $\mathcal{W}_i$ can be explicitly computed based on (14). However, in this work, to quantify set $\mathcal{W}_i$, due to its convexity, we have taken the convex hull of the points - given simulating with a signal $\Delta u_i(k)$ sampled from $\Delta\bar{\mathcal{U}}_i$ - to approximate the set. This solution has permitted to apply the robust MPC approach defined in this section with no constraint violation on the real variables.

3) Ensemble model: To define the ensemble dynamics, the reference models must be opportunely combined. The state of the ensemble dynamical model $\bar{\mathcal{L}}$ is composed of the states of the active generators, i.e., with $x^{[\mathrm{H}]}_{\mathrm{ON}\,i} = 1$. When a boiler is switched off, its contribution to ensemble steam production is immediately removed: in practice, during the transient, its steam is diverted from the ensemble output. Similarly, during start-up, the produced steam is not conveyed to the ensemble output, due to low steam quality - with high percentage of transported condensate. Accordingly, we define the ensemble state as $\bar{x}^{[\mathrm{M}]} = \sum_i^{N_\mathrm{g}} x^{[\mathrm{H}]}_{\mathrm{ON}\,i}\hat{x}^{[\mathrm{M}]}_i$, its input as $\bar{u}^{[\mathrm{M}]}$, and its output as $\bar{y}^{[\mathrm{M}]} = \sum_i^{N_\mathrm{g}} x^{[\mathrm{H}]}_{\mathrm{ON}\,i}\hat{y}^{[\mathrm{M}]}_i$.

Considering the reference models (11), we can write

$$\bar{\mathcal{L}}: \begin{cases} \bar{x}^{[\mathrm{M}]}(k+1) = & \hat{A}\bar{x}^{[\mathrm{M}]}(k) + \bar{B}\bar{u}^{[\mathrm{M}]}(k) + \bar{w}^{[\mathrm{M}]}(k) \\ \bar{y}^{[\mathrm{M}]}(k) = & \hat{C}\bar{x}^{[\mathrm{M}]}(k) + \bar{\gamma} \end{cases} \tag{15}$$

where $\bar{B} = \sum_i^{N_\mathrm{g}} \alpha_i^{[\mathrm{H}]}\hat{B}_i$, $\bar{\gamma} = \sum_i^{N_\mathrm{g}} x^{[\mathrm{H}]}_{\mathrm{ON}\,i}\hat{\gamma}_i$, and $\bar{w}^{[\mathrm{M}]} = \sum_i^{N_\mathrm{g}} x^{[\mathrm{H}]}_{\mathrm{ON}\,i}\hat{w}^{[\mathrm{M}]}_i$. We also define the static gain of the ensemble as $\bar{g} = \sum_i^{N_\mathrm{g}} \alpha_i^{[\mathrm{H}]} g_i$.

Remark 4: The gain consistency conditions (12) are necessary to guarantee that the ensemble gain correctly reflects the overall gains of the subsystems, given the specified load partition.

The set containing the reference deviation $\bar{w}^{[\mathrm{M}]}(k)$ is defined as $\bar{\mathcal{W}}$. It can be computed as discussed in [20]. More specifically, we can enforce - as discussed in Section III-C4 - $\Delta u_i^{[\mathrm{M}]}(k) \in \Delta\bar{\mathcal{U}}$, for all $i = 1, \ldots, N_\mathrm{g}$ and for all values of $\alpha_i^{[\mathrm{H}]}$, where $\Delta\bar{\mathcal{U}} = [-\Delta\bar{u}, \Delta\bar{u}]$ for a given threshold $\Delta\bar{u}$. As discussed, this is done to guarantee that $\bar{w}_i^{[\mathrm{M}]}(k) \in \mathcal{W}_i$, and also that $\bar{w}^{[\mathrm{M}]}(k) \in \bar{\mathcal{W}} = \bigoplus_{i=1}^{N_\mathrm{g}} \mathcal{W}_i$ in all possible system configurations.

4) Medium-level controller design: The ML MPC objective is to track the global fuel flow-rate target $r = \bar{q}_\mathrm{g}^{\mathrm{Dem}}$, that depends on the HL solution. At any time instant k, the HL share and mode signals, $(\alpha_i^{[\mathrm{H}]}, x^{[\mathrm{H}]}_{m\,i})$, are re-sampled with sampling time T_M, as $\alpha_i^{[\mathrm{M}]}(k) = \alpha_i^{[\mathrm{H}]}(\lfloor k/\mu \rfloor)$, and are assumed to remain constant, e.g., $\alpha_i^{[\mathrm{M}]}(k+l) = \alpha_i^{[\mathrm{M}]}(k)$ for the whole control horizon, i.e., $\forall l = 1, \ldots, N_\mathrm{M}$. This implies that the ensemble model $\bar{\mathcal{L}}$, (15), is invariant during the optimization horizon, N_M. To cope with disturbance $\bar{w}^{[\mathrm{M}]}(k)$ in the ensemble model $\bar{\mathcal{L}}$, the ML must be designed according to a robust tube-based implementation. The system is augmented and written in velocity form, as in [35]

$$\xi^{[\mathrm{M}]}(k+1) = \mathcal{A}\xi^{[\mathrm{M}]}(k) + \mathcal{B}\Delta\bar{u}^{[\mathrm{M}]}(k) + \mathcal{H}\Delta\bar{w}^{[\mathrm{M}]}(k) \tag{16}$$

with state vector $\xi^{[\mathrm{M}]}(k) = [\Delta\bar{x}^{[\mathrm{M}]}(k), \varepsilon^{[\mathrm{M}]}(k)]$, input $\Delta\bar{u}^{[\mathrm{M}]}(k)$, and disturbance $\Delta\bar{w}^{[\mathrm{M}]}(k)$. Matrices $\mathcal{A}, \mathcal{B}, \mathcal{H}$

where G_i and $\hat{G}_i$ are the transfer functions of systems (6), (13), respectively, while z^{-1} is the discrete time backward shifting operator. According to the gain consistency property, $\hat{G}_i(1) = G_i(1)$. Therefore

$$e_i(k+1) = \Big((G_i(z^{-1}) - G_i(1)) - (\hat{G}_i(z^{-1}) - \hat{G}_i(1))\Big)\, u_i(k)$$

Considering the canonical structure of the model (6), the term $G_i(z^{-1}) - G_i(1)$ has the following form:

$$G_i(z^{-1}) - G_i(1) = \frac{\sum_{j=1}^{n_\mathrm{b}}(b_{i,j}(z^{-j} - 1))}{\sum_{j=1}^{n_\mathrm{f}}(f_{i,j}(z^{-j} - 1))}$$

By the rational root theorem, every binomial $(z^{-j} - 1)$ can be factorized by $(z^{-1} - 1)$, therefore we can write

$$G_i(z^{-1}) - G_i(1) = (z^{-1} - 1)\mathcal{G}_i(z^{-1})$$

and, similarly,

$$\hat{G}_i(z^{-1}) - \hat{G}_i(1) = (z^{-1} - 1)\hat{\mathcal{G}}_i(z^{-1})$$

Therefore,

$$\begin{aligned} e_i(k+1) &= \Big[\mathcal{G}_i(z^{-1}) - \hat{\mathcal{G}}_i(z^{-1})\Big]\,(z^{-1} - 1)u_i(k) \\ &= -\Delta\mathcal{G}_i(z^{-1})(u_i(k) - u_i(k-1)) \end{aligned}$$

can be trivially derived from (15).

The added state is $\varepsilon^{[\mathrm{M}]}(k) = \bar{y}^{[\mathrm{M}]}(k) - \hat{r}$, where the reference output $\hat{r}$ is set as a decision variable of the OCP, as in [36], to ensure recursive feasibility and offset-free tracking capabilities in presence of continuous variations of the target values (which can be possibly infeasible): in a few words, $\hat{r}$ is the closest feasible set-point to r, at least in stationary conditions.

A nominal (undisturbed) model, used to formulated the OCP, can be associated to (16):

$$\tilde{\xi}^{[\mathrm{M}]}(k+1) = \mathcal{A}\tilde{\xi}^{[\mathrm{M}]}(k) + \mathcal{B}\Delta\tilde{u}^{[\mathrm{M}]}(k) \tag{17}$$

whose variables are denoted by $\tilde{\cdot}$.

To guarantee the feasibility in the disturbed case, the constraints for the OCP with nominal model must be opportunely tightened. The tube-based approach requires the computation of a Robust Positively Invariant (RPI) set $\mathcal{Z}$ - computed based on [37] - where $\xi^{[\mathrm{M}]}(k) - \tilde{\xi}^{[\mathrm{M}]}(k)$ is guaranteed to lie if the following control law is applied to the real system,

$$\delta\bar{u}^{[\mathrm{M}]}(k) = \delta\tilde{u}^{[\mathrm{M}]}(k) + \mathcal{K}(\xi^{[\mathrm{M}]}(k) - \tilde{\xi}^{[\mathrm{M}]}(k)) \tag{18}$$

where $\mathcal{K}$ a gain matrix that makes the matrix $\mathcal{A} + \mathcal{B}\mathcal{K}$ Schur stable. Namely, the real system is kept close to the nominal state, i.e.,

$$\xi^{[\mathrm{M}]}(k+j) \in \tilde{\xi}^{[\mathrm{M}]}(k) \oplus \mathcal{Z} \qquad \forall j \geq 1$$

So, the robust MPC problem is formulated on the nominal system (17), leading to a quadratic program (QP), where the optimization variables are the future nominal input trajectory, $\delta\tilde{\mathbf{u}}(k) = [\delta\tilde{u}^{[\mathrm{M}]}(k) : \delta\tilde{u}^{[\mathrm{M}]}(k+N_{\mathrm{M}}-1)]$, the initial condition of the nominal system, $\tilde{\xi}^{[\mathrm{M}]}(k) = (\delta\tilde{x}^{[\mathrm{M}]}(k), \tilde{y}^{[\mathrm{M}]}(k) - \hat{r})$, and the output reference point, $\hat{r}$.

$$\min_{\substack{\tilde{\xi}^{[\mathrm{M}]}(k),\hat{r},\\ \delta\tilde{\mathbf{u}}(k)}} \|\hat{r} - r\|_T^2 + \sum_{j\in\mathcal{J}} \left\{ \|\tilde{\xi}^{[\mathrm{M}]}(j)\|_Q^2 + \|\delta\tilde{u}^{[\mathrm{M}]}(j)\|_R^2 \right\} \tag{19a}$$

$$\text{s.t.} \quad \xi^{[\mathrm{M}]}(k) - \tilde{\xi}^{[\mathrm{M}]}(k) \in \mathcal{Z} \tag{19b}$$

$$\tilde{\xi}^{[\mathrm{M}]}(j+1) = \mathcal{A}\tilde{\xi}^{[\mathrm{M}]}(j) + \mathcal{B}\delta\tilde{u}^{[\mathrm{M}]}(j) \tag{19c}$$

$$\tilde{u}^{[\mathrm{M}]}(j) \in \tilde{\mathcal{U}} \tag{19d}$$

$$\alpha_i^{[\mathrm{H}]}(h)\tilde{u}^{[\mathrm{M}]}(j) \in \tilde{\mathcal{U}}_i \tag{19e}$$

$$x_{\mathrm{ON}\,i}^{[\mathrm{H}]}(h)\,[g_i\alpha_i(h)\tilde{u}^{[\mathrm{M}]}(j) + \hat{\gamma}_{\mathrm{ON}\,i}] \in \tilde{\mathcal{Y}}_i \tag{19f}$$

$$\alpha_i^{[\mathrm{M}]}(j)\tilde{u}^{[\mathrm{M}]}(j) - \alpha_i^{[\mathrm{M}]}(j-1)\tilde{u}^{[\mathrm{M}]}(j-1) \in \Delta\tilde{\mathcal{U}} \tag{19g}$$

$$\forall j \in \mathcal{J}$$
$$\forall i = 1, \ldots, N_{\mathrm{g}}$$

$$\tilde{\xi}^{[\mathrm{M}]}(k + N_{\mathrm{M}}) = 0 \tag{19h}$$

$$\tilde{x}^{[\mathrm{M}]}(k + N_{\mathrm{M}}) = \tilde{x}_{\mathrm{ss}} \tag{19i}$$

$$\tilde{u}^{[\mathrm{M}]}(k + N_{\mathrm{M}} - 1) = \tilde{u}_{\mathrm{ss}} \tag{19j}$$

where $\mathcal{J} = \{k, \ldots, k + N_{\mathrm{M}} - 1\}$. Moreover, $\tilde{x}_{\mathrm{ss}}$, $\tilde{u}_{\mathrm{ss}}$ are given by

$$\begin{bmatrix} \tilde{x}_{\mathrm{ss}} \\ \tilde{u}_{\mathrm{ss}} \end{bmatrix} = \begin{bmatrix} I_n - \hat{A} & -\bar{B} \\ \hat{C} & 0_m \end{bmatrix}^{-1} \begin{bmatrix} 0_{n\times p} \\ I_p \end{bmatrix} (\hat{r} - \bar{\gamma})$$

The constraints (19i)-(19j) requires the calculation of $\tilde{x}^{[\mathrm{M}]}(k-1)$, $\tilde{u}^{[\mathrm{M}]}(k-1)$ that can be evaluated based on

$$\begin{bmatrix} \tilde{x}^{[\mathrm{M}]}(k-1) \\ \tilde{u}^{[\mathrm{M}]}(k-1) \end{bmatrix} = \begin{bmatrix} \hat{A} - I_n & \bar{B} \\ \hat{C}\hat{A} & \hat{C}\bar{B} \end{bmatrix}^{-1} \begin{bmatrix} \Delta\tilde{x}^{[\mathrm{M}]}(k) \\ \tilde{y}^{[\mathrm{M}]}(k) \end{bmatrix}$$

Differently from [35], the terminal constraint is a *steady-state* condition for (17) in the last step of the prediction horizon. The computation of a terminal steady-state condition guarantees that the MPC problem is practically recursively feasible, with auxiliary control law $\Delta\tilde{u}^{[\mathrm{M}]}(k) = 0$. This formulation avoids the computation of the Maximal Output Admissible Set (MOAS) required in [35]. It is worth noting that - similarly to the computation of the RPI [37] - the calculation of the MOAS [38] is an iterative time-consuming procedure. Any variation of the configuration requires the re-computation online of both the RPI and the MOAS. At least the latter is avoided by forcing the system to reach a steady-state condition at the end of the prediction window; on the other hand, it might affect the promptness of the controller, reducing

the optimal control action, since $\Delta\tilde{u}^{*[T]}(k) \rightarrow 0$ as $k \rightarrow N_{\mathrm{M}}$. This can be mitigated by selecting a longer prediction window.

The initial condition of the nominal state is enforced by (19b). For all the time steps $j \in \mathcal{J}$, the ML is committed to impose the constraints (1) through (19d)-(19f). Moreover, as discussed in Section III-C1, in order to keep the disturbance term $\bar{w}^{[\mathrm{M}]}(k)$ bounded, we need to ensure that for each generator the input variation is limited thanks to (the tightened) constraint (19g).

Also constraints (19d)-(19g) are imposed on the nominal system variables: this requires a proper tightening [35] of the original sets $\bar{\mathcal{U}}$, $\bar{\mathcal{U}}_i$,$\Delta\bar{\mathcal{U}}_i$ allowing us to define $\tilde{\mathcal{U}}$, $\tilde{\mathcal{U}}_i$, and $\Delta\tilde{\mathcal{U}}_i$.

Note also that, while in [20] constraints on local outputs are not considered, in our application scenario they play a key role. In fact they represent limitations in the gas available to each burner. To enforce $y_i^{[\mathrm{M}]} \in \mathcal{Y}_i$, we use its simplified "quasi steady-state" version (19f). Set $\tilde{\mathcal{Y}}_i$ is computed by suitably tightening set $\mathcal{Y}_i$.

5) Transitions among configurations: When configuration transitions occur, i.e., when the high hierarchical level returns a new optimal value of sharing factor $\alpha_i^{*[\mathrm{M}]}(k) \neq \alpha_i^{*[\mathrm{M}]}(k-1)$ at least for some subsystems, infeasibility issues may occur due to two reasons: (i) the ensemble model is varying with respect to the one used at the previous time step, since $\bar{B} = \bar{B}\left(\alpha^{*[\mathrm{M}]}(k)\right)$; (ii) it is not guaranteed that constraints (19d) and (19g) can be enforced in a recursive manner. The procedure adopted when configuration changes occur is the following one:

- Apply $\alpha_i^{[\mathrm{M}]}(k) = \alpha_i^{*[\mathrm{M}]}(k)$, $\forall i = 1, \ldots, N_{\mathrm{g}}$, and solve the corresponding MPC optimization problem. If it is feasible, then the configuration change is accepted.
- If the optimization problem formulated at the previous time step does not result feasible, then reformulate the MPC optimization problem using the actual sharing factors $\alpha^{[\mathrm{M}]}(k)$ (under the assumption to keep them constant during the whole control horizon) as further - temporary - optimization variables and adding the term $\sum_{i=1}^{N_{\mathrm{g}}} \|\alpha_i^{[\mathrm{M}]}(k) - \alpha_i^{*[\mathrm{M}]}(k)\|^2$ to the cost function, in order to steer $\alpha_i^{[\mathrm{M}]}(k)$ to the values $\alpha_i^{*[\mathrm{M}]}(k)$, selected as the optimal ones by the high-level optimizer.

Remark 5: The introduction of the sharing factors as additional decision variables transforms the program (19) from QP to a nonlinear one. In fact, the dependence of the model on $\alpha_i^{[\mathrm{M}]}$ implies that a number of elements of problem (19) are dependent upon $\alpha_i^{[\mathrm{M}]}$ in a non-trivial way, e.g., the gain $\mathcal{K}$, the RPI set $\mathcal{Z}$ (to be used in the constraint (19b)), and the set tightening.

We can here address this issue by reformulating (19) in a slightly different, but consistent, way, to be applied exclusively during the transitions. First of all, to avoid the use of $\mathcal{Z}$, we replace (19b) with the equality $\tilde{\xi}^{[\mathrm{M}]}(k) = \xi^{[\mathrm{M}]}(k)$. Also, due to Assumption 1, $\hat{A}$ is Schur stable. Thus, we can adopt, during the transition, an auxiliary law with $\mathcal{K} = 0$. So, the input applied to the model ensemble is not corrected by (18). A final remark is in order: to support transitions, the tightening operations to be performed on sets $\bar{\mathcal{U}}$, $\Delta\bar{\mathcal{U}}$, $\bar{\mathcal{U}}_i$, and $\bar{\mathcal{Y}}_i$ should be sufficiently general to be compatible with all ensemble models of interest to avoid possible feasibility losses.

IV. Simulations

The hierarchical control scheme is tested in simulation, considering a use-case with $N_{\mathrm{g}} = 5$ steam generators that operate at a pressure of 57 bar and cooperate to serve a common load. The boilers that form the ensemble are slightly different among each other, since they are characterized by dissimilar dimensions and efficiencies. Also, they are limited to work in different operating ranges, i.e. minimum/maximum generated steam. Their

parameters are reported in Table I.

The natural gas price λ_{g} is assumed fixed and constant

Table I: Boiler parameters.

Boiler n	1	2	3	4	5
V_{T} $[m^3]$	1.21	1.15	1.28	1.14	1.32
M_{T} $[t]$	5.49	5.22	5.83	5.06	5.99
η $[-]$	0.90	0.92	0.89	0.95	0.99
$q_{\mathrm{s}}^{\mathrm{Min}}$ $[kg/s]$	0.1	0.09	0.09	0.09	0.1
$q_{\mathrm{s}}^{\mathrm{Max}}$ $[kg/s]$	1.26	1.16	1.13	1.20	1.25
$q_{\mathrm{g}}^{\mathrm{Min}}$ $[kg/s]$	0.125	0.127	0.129	0.126	0.123
$q_{\mathrm{g}}^{\mathrm{Max}}$ $[kg/s]$	0.859	0.844	0.846	0.841	0.839
λ_{g} $[€/m^3]$	0.22	0.22	0.22	0.22	0.22
λ_{ON} $[€/T_{\mathrm{H}}]$	40	30	22	55	45
λ_{ST} $[€/T_{\mathrm{H}}]$	100	130	120	70	80

for all the generators, while the fixed operating cost in ON and startup modes $\lambda_{\mathrm{ON}\,i}$ and $\lambda_{\mathrm{ST}\,i}$, respectively, are different for each generator. Gas density is $\rho_{\mathrm{g}} = 0.71$ $[kg/m^3]$ and tube specific heat $c_{\mathrm{p}} = 0.5$ $[kJ/(kgK)]$.

The system is characterized by the following global constraints $\bar{\mathcal{Y}} = [0.1227, 4.220]$ $[kg/s]$ and $\bar{\mathcal{U}} = [0.089, 6.0]$ $[kg/s]$, determined by constraints of the distribution network.

The sampling times of the multi-layer architecture are reported in Table II.

The low-level controllers have been implemented in

Table II: Multi-layer sampling times.

Sampling time	τ	T_{M}	T_{H}
	10 s	30 s	10 min

discrete-time with a fast sampling time $\tau = 10$ s; their parameters are tuned to stabilize the system with a settling time of 120 s.

All systems are assumed to have the same compensator $\mathbf{C}$, with $K_{\mathrm{P}} = 0.30$ and $K_{\mathrm{I}} = 0.10$ and regulator $\mathbf{R}$, with $K_{\mathrm{P}} = 0.87$ and $K_{\mathrm{I}} = 3.5 \cdot 10^{-4}$.

The discrete-time linear model (5) is identified on a dataset generated by simulating the closed-loop nonlinear model $\mathcal{S}_i^{\mathrm{CL}}$, with sampling time $T_{\mathrm{M}} = 30$ s. For each boiler, the identified models, $\mathcal{L}_i^{\mathrm{CL}}$, are characterized by $n_{\mathrm{f}} = 3$, $n_{\mathrm{b}} = 2$, and $n_{\mathrm{k}} = 1$. So that systems $\mathcal{L}_i^{\mathrm{CL}}$ have the same order n. The high-level optimization is executed in receding horizon with a slow sampling time $T_{\mathrm{H}} = 10$ min.

The optimization (9) considers a prediction horizon of $N_{\mathrm{H}} = 10$, which is long enough to consider the high-level dynamics of the sub-systems - by considering their start-up dynamics - and forthcoming fluctuation of the users global demand, $\bar{q}_{\mathrm{s}}^{\mathrm{Dem}}(h)$ for $h = 0, \ldots, N_{\mathrm{H}}$.

The latter is given as a piece-wise constant forecast of users' demand, which can be opportunely updated at any iteration of the rolling window of the high-level optimization.

Regarding the high-level dynamic models of the steam generators, each unit is characterized by an hybrid automaton, as presented in Section II-D, with the dwell-times reported in Table III. In this case-study all gener-

Table III: Hybrid automaton dwell times (in HL steps T_{H}).

Boiler	$\chi_{\mathrm{OFF}\to\mathrm{ST}}$	$\chi_{\mathrm{ST}\to\mathrm{ON}}$	$\chi_{\mathrm{ON}\to\mathrm{OFF}}$
i	2	2	3

ators have the same transition times.

It is worth emphasizing that, as reported in Figure 5, the reference trajectory is naturally given in terms of steam demand $\bar{q}_{\mathrm{s}}^{\mathrm{Dem}}$ and converted into equivalent gas target using the static gain of the ensemble $\bar{q}_{\mathrm{g}}^{\mathrm{Dem}} = \bar{g}\bar{q}_{\mathrm{s}}^{\mathrm{Dem}} + \bar{\gamma}$. In particular, the reference target for the fuel flow-rate of the ensemble incorporates only the units in mode ON. While the high-level optimizer considers the consumption and the relative costs of the steam generators also in startup modes, we recall that the ensemble model considers just the producing boiler, i.e., in mode ON. The scope of the MPC layer is indeed a robust reference tracking for the ensemble and not an

economic optimization, which is the target of the high-level optimization.

As discussed in Section III-C1, a requirement is that all the reference models share the same dynamic and output matrices $\hat{A}_i$ and $\hat{C}_i$, respectively. Conceptually, they can be arbitrarily chosen by the designer, e.g., by imposing a desired dynamic matrix or an "averaged" one for all the subsystems of the ensemble. In this work, we select a specific unit as the reference dynamic model: therefore, we impose the matrices of the first steam generator for the reference model, i.e. $\hat{A}_i = A_1$ and $\hat{C}_i = C_1$. As a consequence, the state reduction map is simply $\beta_i = I_n$ for all the subsystems and gain consistency conditions reduce to (12).

In Figure 6, the comparison of the step response of each system $\mathcal{L}_i^{\text{CL}}$ with its reference model $\hat{\mathcal{L}}_i$ is shown: the gain consistency conditions (12) guarantee that at steady state the actual and reference models reach the same value.

The maximum amplitude of the disturbance $\bar{w}$ in the ensemble model is evaluated by imposing the maximum variation of the input equal to $\Delta\bar{u} = 0.4[kg/s]$, resulting in $\|\bar{w}\|_\infty \leq 1 \times 10^{-3}[kg/s]$.

We compare, in simulation, the performance of the proposed control scheme (HL OPT) with two alternative ones, obtainable with different strategies, see Table IV: NO HL, where the sharing factors are not optimized during operation, but predefined and fixed (e.g., by equally splitting the load on available units, $\alpha_i = 1/\sum x_{\text{ON}\,i}$), and HL η-OPT, where the units are activated in a round-robin fashion according to their efficiency ranking. For all these schemes, the robust MPC, see (19), is synthesized on the ensemble configuration defined at HL, using the steady-state terminal condition, with a prediction horizon $N_{\text{M}} = 10$. The constraints, imposed according to the tube-based paradigm, are enforced in a tightened way to the unperturbed system variables.

Table IV: Control architectures features for performance comparison.

Strategy	HL	ML-MPC	LL	Test
HL OPT	$\alpha_i \leftarrow$(9) (l_i eq.(10))	Ensemble	PI	1&2
NO HL	$\alpha_i \leftarrow 1/\sum x_{\text{ON}\,i}$	Ensemble	PI	1&2
HL η-OPT	$\alpha_i \leftarrow$(9) ($l_i = -\eta_i x_{\text{ON}\,i}$)	Ensemble	PI	1&2
C-MPC	-	Centralized	PI	1&2

Test 1: Unit 3 with scheduled maintenance, piece-wise demand

Test 2: All units available, noisy demand

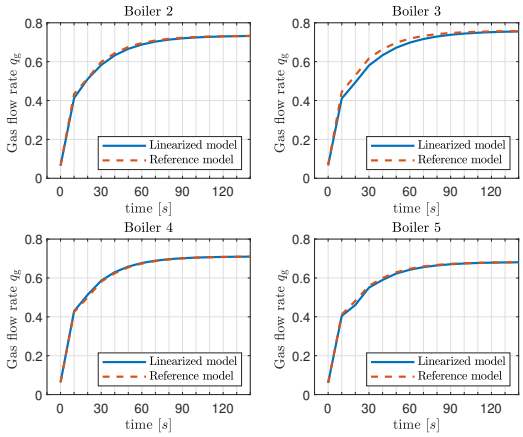

Figure 6: Step-response of the actual (solid line) and reference models (dashed line).

The tracking performances are also compared with the ones of a centralized MPC scheme (C-MPC), which controls directly all the subsystem inputs, u_i. Two scenarios are proposed: *Test 1* considers a maintenance schedule for Boiler 3, with a piece-wise constant demand; *Test 2* shows the behavior of the schemes with a noisy demand, to assess the operational cost and the tracking performance.

1) Test 1: We assume Boiler 3 to be unavailable in the time range $t = [50, 80)$ min and we compare the behavior of the proposed scheme with HL η-OPT and with NO HL. The idea here is to focus on the role of the HL control layer on the overall performances.

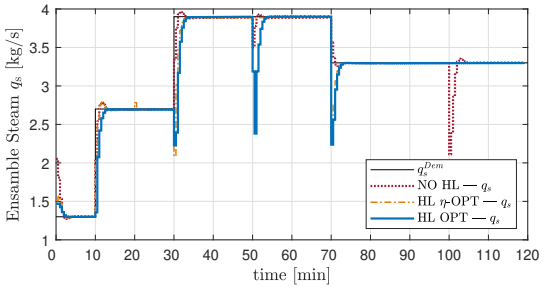

Figure 7: Steam demand for the ensemble $\bar{q}_{\mathrm{s}}^{\mathrm{Dem}}$ (black) tracked by Ensemble-MPC at ML, with HL OPT strategy (solid blue), HL η-OPT (dot-dashed orange), and NO HL (dotted red).

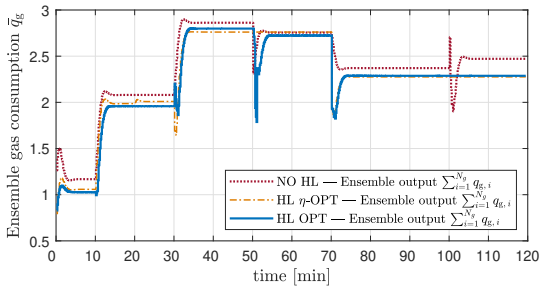

Figure 8: Ensemble gas consumption. The reference target depends on the ensemble configuration, since different sharing factors change the overall gain.

Figure 7 shows the tracking of steam demand with the three considered strategies. Note that the overshoots are due to the plug/unplug operations. Figure 8 shows the reference tracking performances of the natural gas signal. Due to the unequal overall gain of the ensemble in the three alternative configurations, n atural gas trajectories are different. This is more evident in the period $t = [30, 70)$ min even if the steam demand is the same. This is related to a difference in the subsystem's efficiency. R ecall t hat t he g as c onsumption i s g iven by (7a) for each boiler and the ensemble efficiency is given by the α-weighted combination of such equations.

In Figure 9, the operating modes of each unit are

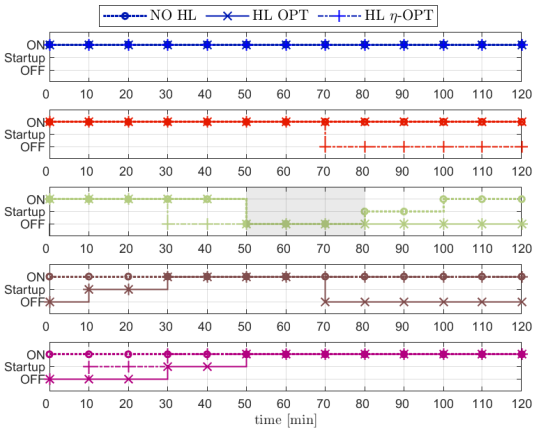

Figure 9: Operating mode of each subsystem. Boiler 3 temporal unavailability shown by gray region.

shown: at $t = 50$ min, Boiler 3 is forced to OFF mode, for prescribed unavailability (e.g., for maintenance reasons) shown by a gray area. With HL optimization, Boiler 4 is activated in place of Boiler 3. Note that, even if Boiler 5 has a higher efficiency with respect to Boiler 4, the latter is chosen in the first place due to its lower start-up costs $\lambda_{\mathrm{ST}\,i}$; with HL η-OPT, instead, the different efficiency-based criterion for boiler activation leading to slightly larger overall costs.

As shown in Figure 10, when the global steam demand rises, new generators are added to the ensemble based, not only on the subsystem efficiency rank, but also on the associated operating costs $\lambda_{\mathrm{ON}\,i}$ and $\lambda_{\mathrm{ST}\,i}$, which are different for each generator. In the NO HL scenario, the weights α_i are adjusted only to consider that just four generators are available. When the transition is sharp, the abrupt change of α_i could lead one of the subsystems out of its local ranges. If so, the MPC optimization problem may become infeasible. In response to that, the control architecture will compute a transient solution by considering the sharing factors as an additional set of optimization variables, as discussed in Section III-C5: the nonlinear

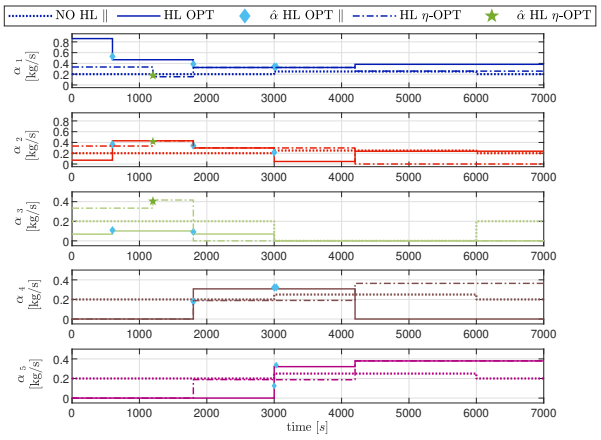

Figure 10: High-level sharing factors with HL OPT strategy (solid, transient $\hat{\alpha}$ diamonds), HL η-OPT (dot-dashed, transient $\hat{\alpha}$ stars), and NO HL (dotted).

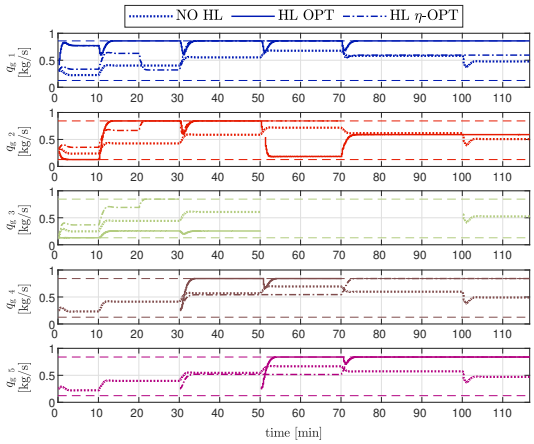

Figure 11: In each subplot, the gas flow-rate $q_{\mathrm{g},\,i}$ of the units.

program provides the closest feasible configuration to the target computed at the top-level. In Figure 10, the sharing factors computed at high-level, with the three strategies are shown by different line styles. When the high-level solution is reachable in one medium-level step, the optimal and actual points coincide and just target is shown, otherwise the ensemble is guided to the high-level optimal configuration by a smooth shift through temporary configurations (a diamond for the HL OPT and a star for HL η-OPT), computed solving the NLP. Figure 11 shows that also local constraints are respected.

The simulation is executed in Matlab on a Intel®Core™ i7-8550U CPU 1.80GHz, RAM 16 GB, with `SCIP` solver [39] for HL optimization and transitional configurations, and `quadprog` for medium-level QP. The HL optimization takes an average time of 3.28s ($\pm$0.53s), while the ML QP takes 0.13s ($\pm$0.02s), where the RPI computation requires 0.08s. Instead, the NLP for transitional configuration requires up to 40s.

2) Test 2: A second test is done to compare the performance of the proposed scheme and to demonstrate also the robustness of the control architecture in presence of possibly significant errors on the demand forecast. Here the focus is both on the HL, considering the operational cost, and on the ML, by measuring the tracking performance. The latter is assessed by considering a further alternative scheme consisting of a centralized MPC, which governs directly the input u_i of each subsystem. It is worth noting, that this controller cannot manage the HL dynamics related to mode transitions, i.e., start-ups, and plug-and-play operations, thus for this case it does not really make sense to quantify the related HL operational cost. However, given a fixed number of active generators, this represents the best tracking controller. The performance metric is given by $J_{\mathrm{M}}^{\mathrm{tr}} = \sum_k \|y(k) - r\|^2$. Instead, the operational cost J^{op}, is computed as (10).

The disturbed demand is given by $r\,(k) = \bar{q}_s^{\mathrm{Dem}}\,(\lfloor h/\mu \rfloor) + v\,(k)$ with the noise term $v\,(k) \sim n\,(0,\ \sigma)$, with $\sigma = 1.25\%\bar{q}_s^{\mathrm{Dem}}$. In addition, at $t = 90$ (resp. 100) min a downward (resp. upward) step disturbance is given, with $d = \pm 4\%\bar{q}_s^{\mathrm{Dem}}$ thus with the

noise term[2] $v(k) \sim n(d,\ \sigma/10)$.

In Figure 12, the tracking of the natural gas for the ensemble is reported for the different control strategies. Note that the different overall efficiency of the ensemble leads to distinct gas flow-rates, even if the steam demand is the same, see Figure 13. At medium level this demand is disturbed by an additive noise term, the mismatch between the piece-wise reference is managed by the MPC: as the reference $\hat{r}$ is a decision variable at ML, this MPC formulation can deal also with infeasible references. Typically, an increased demand might become unreachable, while lower actual demand can be easily managed: in $t = [90,\ 100)$ min, with the downward step disturbance, ML can track the actual demand by keeping the same sharing factors. Instead in $t = [100,\ 110)$ min, the global generation cannot reach the actual target. However, the controller robustly gives a feasible solution, which minimizes the distance from the target. The event-based optimization of the HL sharing factors is applied at $t = 102.5$ min, when a bias on the demand is detected: a new HL optimization is triggered on an updated demand forecast, which includes the bias. The sharing factors are adapted to achieve the increased demand. The best tracking performance is obtained by C-MPC, that does not operate on an overall ensemble model, but controls directly each subsystem. However, it does not provide flexibility to system changes and scalability properties. A good tracking performance is apparently given by the NO HL scheme, where demand at ML is tracked by an ensemble-MPC, with all the units sharing equally the load. The controller at ML is the same used for HL η-OPT and HL OPT, where the plug-&-play operations negatively impacts the tracking. Note that, instead, the overall operational cost J^{op} is better in case the HL optimization is performed. The operational and tracking costs of the four strategies are compared in Table V. Regarding the computational complexity, the proposed method maintains constant the dimension of the QP problem to be solved at medium level, by relying on the ensemble model. On the contrary, the

Table V: Operational cost, J^{op}, (scaled on NO HL cost) - and tracking cost, J^{tr}, (scaled on C-MPC cost).

Cost	NO HL	HL OPT	HL η OPT	C-MPC
$J^{\text{op}}/J^{\text{op}}_{\text{NO HL}}[-]$	1.00	0.78	0.80	-
$J^{\text{tr}}/J^{\text{tr}}_{\text{C-MPC}}[-]$	2.30	3.09	3.48	1.00

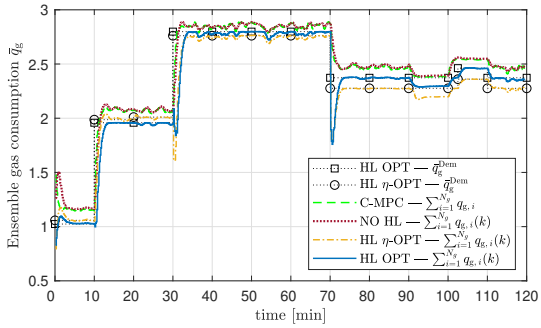

Figure 12: Natural gas consumption of the ensemble. Comparison of the four strategies.

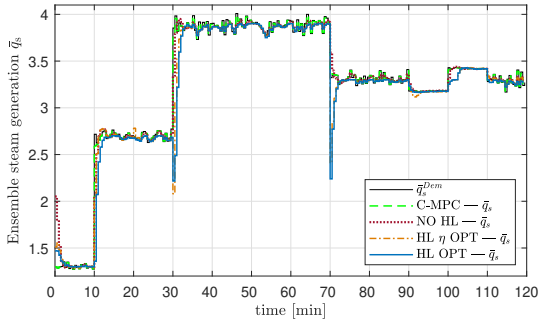

Figure 13: Ensemble steam flow-rate. Comparison of the four strategies.

[2]Note that, merely for clarity of the resulting plots, we have reduced the high-frequency component of the noise by setting a lower standard deviation.

dimension of the QP with the C-MPC grows linearly with respect to the number of considered units, see Table VI. A proportional dimension increase affects also the high-level MIP: in this case the computational impact is greater. Note that it still within the HL sampling time, $T_{\text{H}} = 10$ minutes. Note that, however, the HL should run offline, and so its computational complexity does not impact on the real-time feasibility of the hierarchical scheme.

Table VI: Computational complexity.

	No. of units	5	10	15
E-MPC	[vars.]	14	14	14
	CPU [s] mean	0.158	0.161	0.131
	CPU [s] min	0.106	0.102	0.023
	CPU [s] max	0.271	0.463	0.333
C-MPC	[vars.]	51	101	151
	CPU [s] mean	0.141	0.170	0.205
	CPU [s] min	0.102	0.127	0.162
	CPU [s] max	0.808	0.819	0.876
HL OPT	[vars.]	0.5k	1.1k	1.5k
	CPU [s] mean	3.28	6.54	21.44

V. Conclusions

In this paper a hierarchical control scheme has been proposed for the coordination of an ensemble of steam generators, which must cooperate to fulfill a common load. The definition of an ensemble reference model, as proposed here, permits to solve the medium level tracking MPC in a scalable and flexible way, as its dimension does not grow with the number of steam generators in the ensemble. Thanks to the model reformulation, the ensemble model can be simply obtained from the high level and updated online. The model configuration is determined by the high-level mixed-integer optimization that computes the optimal number of generators to be included in the ensemble and their shares of steam production by minimizing the operating cost and considering global and subsystem constraints. The accuracy of demand forecast impacts the solution quality: generally, forecast mismatch is managed at medium level, with a small degradation of the overall cost. However, if units are committed with a greedy policy with active ones working already at maximum, any higher actual demand cannot be fully sustained, as an additional boiler would be needed, but the start-up dynamics might impede it. This can be managed by tightening the subsystem input/output constraints, at HL level, to prevent such condition, even if the feasibility at medium level is guaranteed by the presence of the reference point among the decision variables. How to properly set this tightening will be studied. Future work will consider the improvement of the multi-layer scheme by comparing the overall performance with the implementation of an additional low-level shrinking MPC control to further address the local model mismatch. We also envision to solve the high level optimization in a distributed framework.

References

[1] Advanced Manufacturing Office, "Minimize boiler short cycling losses - energy tips: Steam," U.S. Department of Energy, Tech. Rep., 2012.

[2] P. Yang, P. Chavali, E. Gilboa, and A. Nehorai, "Parallel load schedule optimization with renewable distributed generators in smart grids," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1431–1441, 2013.

[3] W. L. Theo, J. S. Lim, S. R. W. Alwi, N. E. M. Rozali, W. S. Ho, and Z. Abdul-Manan, "An MILP model for cost-optimal planning of an on-grid hybrid power system for an eco-industrial park," *Energy*, vol. 116, pp. 1423–1441, 2016.

[4] B. Conte, J. C. Bruno, and A. Coronas, "Optimal cooling load sharing strategies for different types of absorption chillers in trigeneration plants," *Energies*, vol. 9, no. 8, p. 573, 2016.

[5] F. Paparella, L. Domínguez, A. Cortinovis, M. Mercangöz, D. Pareschi, and S. Bittanti, "Load sharing optimization of parallel compressors," in *2013 European Control Conference (ECC)*. IEEE, 2013, pp. 4059–4064.

[6] B. Saravanan, S. Das, S. Sikri, and D. Kothari, "A solution to the unit commitment problem—a review," *Frontiers in Energy*, vol. 7, no. 2, pp. 223–236, 2013.

[7] Q. P. Zheng, J. Wang, and A. L. Liu, "Stochastic optimization for unit commitment—a review," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1913–1924, 2014.

[8] J. M. Arroyo and A. J. Conejo, "Optimal response of a thermal unit to an electricity spot market," *IEEE Transactions on Power Systems*, vol. 15, no. 3, pp. 1098–1104, 2000.

[9] M. Carrión and J. M. Arroyo, "A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem," *IEEE Transactions on Power Systems*, vol. 21, no. 3, pp. 1371–1378, 2006.

[10] S. Mitra, L. Sun, and I. E. Grossmann, "Optimal scheduling of industrial combined heat and power plants under time-sensitive electricity prices," *Energy*, vol. 54, pp. 194–211, 2013.

[11] M. Tuffaha and J. T. Gravdahl, "Discrete state-space model to solve the unit commitment and economic dispatch problems," *Energy Systems*, vol. 8, no. 3, pp. 525–547, 2017.

[12] G. Ferrari-Trecate, E. Gallestey, P. Letizia, M. Spedicato, M. Morari, and M. Antoine, "Modeling and control of co-generation power plants: A hybrid system approach," *IEEE Transactions on Control Systems Technology*, vol. 12, no. 5, pp. 694–705, 2004.

[13] S. Spinelli, M. Farina, and A. Ballarino, "A hierarchical optimization-based scheme for combined fire-tube boiler/CHP generation units," in *2018 European Control Conference (ECC)*. IEEE, 2018, pp. 416–421.

[14] G. M. Kopanos, O. C. Murele, J. Silvente, N. Zhakiyev, Y. Akhmetbekov, and D. Tutkushev, "Efficient planning of energy production and maintenance of large-scale combined heat and power plants," *Energy Conversion and Management*, vol. 169, pp. 390–403, 2018.

[15] A. C. Dunn and Y. Y. Du, "Optimal load allocation of multiple fuel boilers," *ISA Transactions*, vol. 48, no. 2, pp. 190–195, 2009.

[16] J. Bujak, "Optimal control of energy losses in multi-boiler steam systems," *Energy*, vol. 34, no. 9, pp. 1260–1270, 2009.

[17] P. Austin, M. McEwan, J. Mack, J. Godsal, J. Tyler, and J. Maciejowski, "Optimisation of fuel usage and steam availability in the power and steam plant of a paper mill." *IFAC Proceedings Volumes*, vol. 43, no. 1, pp. 226–231, 2010.

[18] P. Rambalee, G. Gous, P. De Villiers, N. McCulloch, and G. Humphries, "Control and stabilization of a multiple boiler plant: An APC approach," *IFAC Proceedings Volumes*, vol. 43, no. 9, pp. 109–114, 2010.

[19] V. Costanza and P. S. Rivadeneira, "Optimal supervisory control of steam generators operating in parallel," *Energy*, vol. 93, pp. 1819–1831, 2015.

[20] F. Petzke, M. Farina, and S. Streif, "A multirate hierarchical MPC scheme for ensemble systems," in *Proc. Conference on Decision and Control (CDC)*, 2018, pp. 5874–5879.

[21] R. Scattolini, "Architectures for distributed and hierarchical model predictive control: A review," *Journal of Process Control*, vol. 19, no. 5, pp. 723 – 731, 2009.

[22] D. Barcelli, A. Bemporad, and G. Ripaccioli, "Hierarchical multi-rate control design for constrained linear systems," in *49th IEEE Conference on Decision and Control*, 2010, pp. 5216–5221.

[23] E. Garone, S. Di Cairano, and I. Kolmanovsky, "Reference and command governors for systems with constraints: A survey on theory and applications," *Automatica*, vol. 75, pp. 306–328, 2017.

[24] U. Kalabić, I. Kolmanovsky, J. Buckland, and E. Gilbert, "Reduced order reference governor," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE, 2012, pp. 3245–3251.

[25] B. Picasso, X. Zhang, and R. Scattolini, "Hierarchical model predictive control of independent systems with joint constraints," *Automatica*, vol. 74, pp. 99 – 106, 2016.

[26] M. Farina, X. Zhang, and R. Scattolini, "A hierarchical multi-rate MPC scheme for interconnected systems," *Automatica*, vol. 90, pp. 38 – 46, 2018.

[27] ——, "A hierarchical MPC scheme for coordination of independent systems with shared resources and plug-and-play capabilities," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 2, pp. 521–532, 2018.

[28] L. E. Sokoler, P. J. Dinesen, and J. B. Jørgensen, "A hierarchical algorithm for integrated scheduling and control with applications to power systems," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 2, pp. 590–599, 2016.

[29] S. Spinelli, E. Longoni, M. Farina, F. Petzke, S. Streif, and A. Ballarino, "A hierarchical architecture for the coordination of an ensemble of steam generators," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 11 557–11 562, 2020.

[30] K. J. Åström and R. D. Bell, "Drum-boiler dynamics," *Automatica*, vol. 36, no. 3, pp. 363–378, 2000.

[31] J. Lygeros, K. H. Johansson, S. N. Simic, J. Zhang, and S. S. Sastry, "Dynamical properties of hybrid automata," *IEEE Transactions on Automatic Control*, vol. 48, no. 1, pp. 2–17, 2003.

[32] A. Bemporad and M. Morari, "Control of systems integrating logic, dynamics, and constraints," *Automatica*, vol. 35, no. 3, pp. 407–427, 1999.

[33] A. Falsone, K. Margellos, and M. Prandini, "A distributed iterative algorithm for multi-agent milps: finite-time feasibility and performance characterization," *IEEE Control Systems Letters*, vol. 2, no. 4, pp. 563–568, 2018.

[34] I. Kolmanovsky and E. G. Gilbert, "Theory and computation of disturbance invariant sets for discrete-time linear systems,"

Mathematical problems in engineering, vol. 4, no. 4, pp. 317–367, 1998.

[35] G. Betti, M. Farina, and R. Scattolini, "A robust MPC algorithm for offset-free tracking of constant reference signals," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2394–2400, Sept 2013.

[36] D. Limon, I. Alvarado, T. Alamo, and E. Camacho, "MPC for tracking piecewise constant references for constrained linear systems," *Automatica*, vol. 44, no. 9, pp. 2382 – 2387, 2008.

[37] S. V. Rakovic, E. C. Kerrigan, K. I. Kouramas, and D. Q. Mayne, "Invariant approximations of the minimal robust positively invariant set," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 406–410, 2005.

[38] E. G. Gilbert and K. T. Tan, "Linear systems with state and control constraints: The theory and application of maximal output admissible sets," *IEEE Transactions on Automatic Control*, vol. 36, no. 9, pp. 1008–1020, 1991.

[39] G. Gamrath and et al., "The SCIP Optimization Suite 7.0," Optimization Online, Technical Report, March 2020.