Computers and Chemical Engineering 76 (2015) 117–136

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Optimization of steel production scheduling with complex time-sensitive electricity cost

Hubert Hadera [a,b], Iiro Harjunkoski [a,*], Guido Sand [a], Ignacio E. Grossmann [c], Sebastian Engell [b]

[a] ABB Corporate Research, Wallstadter Str. 59, 68526 Ladenburg, Germany
[b] Technische Universität Dortmund, Emil-Figge Str. 70, 44221 Dortmund, Germany
[c] Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form 5 February 2015
Accepted 8 February 2015
Available online 18 February 2015

Keywords:
Scheduling
Steel plant
Energy optimization
Demand-side management
Continuous-time models

ABSTRACT

Energy-intensive industries can take advantage of process flexibility to reduce operating costs by optimal scheduling of production tasks. In this study, we develop an MILP formulation to extend a continuous-time model with energy-awareness to optimize the daily production schedules and the electricity purchase including the load commitment problem. The sources of electricity that are considered are purchase on volatile markets, time-of-use and base load contracts, as well as onsite generation. The possibility to sell electricity back to the grid is also included. The model is applied to the melt shop section of a stainless steel plant. Due to the large-scale nature of the combinatorial problem, we propose a bi-level heuristic algorithm to tackle instances of industrial size. Case studies show that the potential impact of high prices in the day-ahead markets of electricity can be mitigated by jointly optimizing the production schedule and the associated net electricity consumption cost.

1. Introduction

In many countries, renewable energy sources contribute a significant share of the overall electric power consumption and due to the volatility of their availability and their privileged role on the market, this may cause high fluctuations of the energy cost for the final user. At the grid level, the demand should always match the supply; otherwise the grid infrastructure is stressed, possibly causing expensive failures. Therefore it is of interest to the supply side of the grid to achieve flexibility of the demand, which traditionally has been assumed to be inelastic in the short-term. This is largely because the consumers of electricity were not getting incentive signals that could trigger changes of the consumption pattern when shortages or oversupply occur.

In recent times, however, smart grid technologies and the liberalization of the energy markets have provided new ways of communicating such signals, both for dispatchable loads (the user is given direct signals to change the consumption) and for non-dispatchable loads (the user decides whether to change the consumption) (NERC, 2007). The latter signals are considered in this work in the form of financial incentives and different pricing contracts. These fall within the Demand-Side Management technology, which aims at supporting an active shaping of the patterns of energy use. In particular, industrial Demand-Side Response (iDSR) involves activities defined as a temporary change in electricity consumption in response to market or supply conditions. In non-dispatchable iDSR a consumer, e.g. a steel plant is allowed to decide whether it wants to react to a changing situation within the grid, potentially gaining financial benefits, or to stick to the production plan. This implies the need of proper day-to-day scheduling and planning of plant operations, and for making use of incentive and price based schemes, such as for example intra-day or day-ahead spot market pricing since changes in the prices of energy might significantly affect the profitability as shown for a stainless-steel production plant in Hadera et al. (2014). If it is assumed that the goal of the plant managers is to deliver the same amount of final products over a certain time horizon, the production schedule can be modified in favor of a lower cost of energy procurement only when the process-specific constraints are always satisfied, and when at the same time the plant faces a certain under-utilization of its production capacity. As shown in Fig. 1, the capacity utilization of the US-based energy-intensive primary metal sector went down by nearly 20% in recent years compared to 1990s (BGFRS, 2013). This excess capacity is an enabler for flexibility, provided that the products can either be delivered earlier than their due dates or stored without quality degradation. In many cases the due dates in steel

* Corresponding author. Tel.: +49 6203 716014.
E-mail address: Iiro.Harjunkoski@de.abb.com (I. Harjunkoski).

http://dx.doi.org/10.1016/j.compchemeng.2015.02.004
0098-1354/

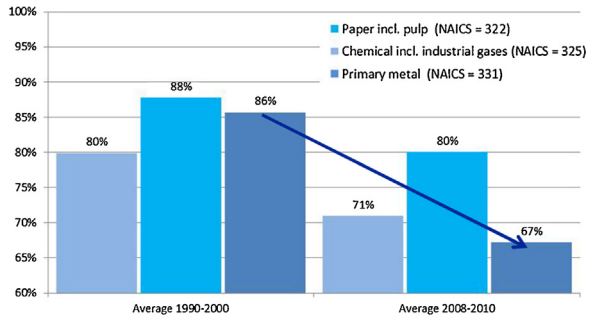

Fig. 1. Capacity utilization of US energy-intensive industries.
Source: ABB study based on BGFRS (2013).

making are provided on a weekly basis and are thus not critical due to a make-to-stock policy.

This creates a potential to optimally shift the production to times when the consumption of electricity is cheaper. This is especially valid for energy-intensive process industries where the raw material and energy cost can account for up to 90% of the total production cost. Demand-response technology on the production scheduling and planning level has the advantage of a potentially low investment cost for the final user, since very often it does not require the purchase of new equipment. Other selected positive outcomes of a more flexible Demand-side Response that are reported in the literature are the following (NERC, 2007; CRA, 2005; Todd et al., 2009):

- Plant level: direct cost savings on the electricity bill;
- Grid level: increase of reliability, e.g. reduction of outages;
- Grid level: reduction of expensive peak load hours in the short-term;
- Environment: potential emission savings by reducing the grid's peak generation (only for regions with fossil fired peak generation plants);
- Environment: potential reduction of emissions by enabling the installation of larger renewable generation capacities;
- Market: market-wide wholesale electricity price reduction in the long-term;
- Market: market performance benefits, e.g. mitigating the suppliers' ability to raise prices significantly above production costs.

Except for the direct energy bill cost savings at the plant site, quantification of the above benefits is difficult and strongly depends on assumptions. However, industrial and academic studies conclude that the potential exists (DOE, 2006; NERC, 2007; DENA, 2011). When investigating DSR of industrial production, it is important to consider the technical potential of Demand-side Response shaping capabilities and not only the total consumption of the process, as pointed out by Paulus and Borggrefe (2011). An ideal industrial plant should have large consumers of electric power that operate in a preplanned fashion and a degree of process flexibility. Both hold true for the steel plant considered in this paper.

While the iDSR technology is recognized as beneficial for both the power supplier side and for the energy-intensive industry, it should be noted that it cannot compensate long-term deficits or surplus of electricity generation in regional grids.

2. Scope and methodology

The goal of this work is to find an optimal production schedule of a part of a steel making process that is operated in batch mode which minimizes a weighted combination of the electricity bill and the lead times of product delivery, while satisfying complex production constraints. In the problem considered here, a continuous-time based general precedence scheduling approach has already been developed for the plant which is extended here to include awareness of the cost of electric energy. The goal is to enable an energy-intensive process plant to realize its demand-side response potential at the production scheduling level, finding a compromise between production delays and the cost of electricity. The main contributions of this work concerns the development of the following items:

- A general strategy for energy-aware scheduling, accounting for time-depending cost of energy in general precedence continuous-time scheduling models;
- Extension of multiple purchase contracts optimization to energy-aware scheduling;
- A bi-level heuristic for obtaining good solutions in reasonable times for industrial scale combined production scheduling and energy cost minimization problems.

The benefit of using a continuous-time formulation is that the exact timing of the production tasks is accounted for within the scheduling horizon. This is in contrast to discrete-time approaches, which discretize the time horizon into discrete time intervals. From industrial practice, we consider 5-min discretization steps as the desired level of time granularity. Such small time window leads to very large discrete-time models that have computational limitations. Other studies showed that a 15-min discretization (Castro et al., 2013) can still be efficient for solving 24 h scheduling horizon with a Resource-Task Network (RTN) based monolithic model approach. However, compared to the discrete-time formulation, continuous-time models also have some drawbacks. Due to the structure of the pricing contracts, it is much easier to account for the cost of the consumption of electricity in discrete-time scheduling models. Extending it to continuous-time formulations is not straightforward since the use of electricity has to be accounted for in fixed time intervals in which the resource prices are constant.

In this paper, we consider purchase optimization of multiple sources of electricity, including the possibility to sell the electricity back to the grid. Also, the challenge of responding to a committed load curve with penalties incurred for both under- and over-consumption is addressed. The combination of these two features has not received much attention in the process scheduling literature. For the given multi-stage steel plant with parallel machines at each stage, the resulting monolithic formulation of the problem is computationally intractable when the scheduling decisions (assignment and sequence binaries) are degrees of freedom for the optimization. To overcome the computational limitations we propose a simple bi-level heuristic approach. The problem is modeled using mathematical programing with Mixed-Integer Linear Programing (MILP) and implemented in the GAMS modeling environment using the CPLEX solver.

In the remaining sections, the background and previous work in the area and the industrial problem are reviewed first. Then the section of the process that is considered and the corresponding energy purchase situation are explained (Section 3). Based on the continuous-time approach for batch processes, the scheduling problem is formulated so that all production constraints of the use case are satisfied (Section 4.2). In Section 4.3, we introduce a strategy for embedding energy-awareness into the continuous-time general precedence formulation. We test the resulting monolithic models (Section 5) and show that they are not able to cope with sizes of real-world problems. Therefore, in Section 6, we describe a bi-level heuristic for the solution of the application problem. Numerical experiments and results are discussed in Section 6.5, followed by conclusions and recommendations for further work (Section 7).

3. Literature review

The field of scheduling and planning has grown rapidly in the last decades. A large number of studies have emerged using both time representation approaches: discrete and continuous. For a general overview concerning the scheduling problems we refer the reader to review papers, such as for example Floudas and Lin (2004), Méndez et al. (2006), Maravelias (2012) and Harjunkoski et al. (2014). The latter focuses especially on the industrial aspects of the scheduling methods. Pochet and Wolsey (2006) present an overview of MILP methods used for production planning.

Scheduling of steel plants has been studied quite extensively as well, as it is recognized one of the most difficult industrial scheduling problems. For handling complex process constraints and optimizing traditional objective functions such as makespan or earliest task completion time an efficient multi-step decomposition approach for the industrial-size scheduling of the melt shop area of a stainless steel plant has been reported by Harjunkoski and Grossmann (2001) based on MILP and LP models. Tang et al. (2001) gives an overview of planning and scheduling systems for integrated steel plants, including Artificial Intelligence, Expert Systems, intelligent search and Constraint Programing methods. In Li et al. (2012) the focus was on the last continuous-casting stage where particular operational features have to be addressed and a rolling horizon was used.

In recent years scheduling under energy constraints has gained increasing attention. It has been also recognized as one of the challenges for industrial implementation of advanced scheduling solutions (Harjunkoski et al., 2014). Optimizing operations with regard to the response to a deterministic single time-varying price of electricity can be found in the literature, including also the stochastic nature of the prices, such as in Li et al. (2003) or in Ierapetritou et al. (2002).

The Demand-side Response has to deal with different time scales. A fast response is required in some iDSR schemes, for example in network ancillary services. Here control techniques rather than scheduling might be better suited. As reported by Vujanic et al. (2012), robust optimization might help creating flexible schedules to support the ancillary services of cement plants. Since energy availability and prices can be treated like any other resource in the scheduling models, many of the formulations in the literature use a discrete-time approach. Zhang and Tang (2010) introduce a discrete-time scheduling formulation using a Lagrangian relaxation algorithm based on the subgradient method. The model includes constraints concerning power availability and minimization of the energy cost. Similarly, in a study by Ashok (2006) a discrete-time formulation is used to schedule a mini steel plant where the operating cost is optimized. The operating cost includes the price of power consumption under different tariffs, charges for registered maximum demand and additional operating cost due to the shifting of loads.

In recent years, models based on the RTN representation have gained attention as an efficient way to deal with resource consumption. Castro et al. (2009) proposed a new strategy for handling variable electricity cost in continuous plants using a continuous-time formulation. Comparison of both continuous- and discrete-time RTN representations showed that the latter's computational performance is better for handling industrial-size instances. The work has been extended by an efficient rolling horizon algorithm in Castro et al. (2011) using an aggregate model, where time intervals of the same resource cost are aggregated into one interval. A steel plant scheduling problem similar to the one studied in this paper, but with response to a single price curve, has been successfully reported by Castro et al. (2013) for a time granularity of 15 min intervals.

Nolde and Morari (2010) proposed a strategy for the modeling of electricity consumption with time-dependent prices in continuous-time models based on precedence variables. It was applied to a stainless-steel process with parallel Electric-Arc Furnaces. The formulation uses six different binary variables to capture the relation of a production task to its placement within a grid of uniform time intervals. For these intervals, electricity consumption is individually accounted for, which makes it possible to track the process load and to optimize the deviation from a pre-agreed consumption curve. Hait and Artigues (2011a) proposed an improvement to Nolde and Morari's approach replacing the set of six binary variables by binaries indicating whether or not an event takes place before or during a time interval. For the same steel case problem, the resulting continuous-time MILP model introduced fewer number of constraints and binary variables. As a follow up study on scheduling of a foundry, Hait and Artigues (2011b) proposed a hybrid heuristic combining Constraint Programing (CP) for solving the assignment and sequencing problem with an MILP model for solving the remaining energy-cost scheduling problem. In addition, the detailed scheduling of the Electric Arc Furnace stage and human operator availability were taken into consideration. Castro et al. (2014) applied the concept of the six cases of task-time interval relations as in Nolde and Morari (2010) to optimize the maintenance of a gas-fired power plant. Using Generalized Disjunctive Programming, Castro and co-workers found a tighter formulation for accounting the electricity consumption. The continuous-time strategy was applied to find a schedule under constraints of operator availability and cost, maximizing profits from electricity sales under time-sensitive demand and pricing. A steel plant has also been considered in a study by Boukas et al. (1990), using a hierarchical approach with separation of operation and secondary resource scheduling in two steps. Constraints were subject to a global limitation of the power delivered to the furnaces.

Apart from the steel industry, demand-side response strategies have been investigated for other energy-intensive processes. Mitra et al. (2012) proposed a discrete-time formulation for process plants with an emphasis on switching the operating modes of the plant units. Responding to a single time-sensitive price curve of electricity the model was successfully applied to air separation and cement production processes. The same solution strategy was also applied in the context of optimal scheduling of an industrial Combined Heat and Power (CHP) plant (Mitra et al. (2013). Underutilization of the CHP plant and its response to time-sensitive electricity prices were investigated.

3.1. Problem description

The industrial problem that is addressed in this work concerns the optimal scheduling of a part of the stainless-steel production process. The production starts with the scrap melting phase in an Electric Arc Furnace (EAF) to form a so-called heat which is the object of scheduling. The process of melting is carried out by passing large amounts of electricity through electrodes in order to form high-temperature electric arc (up to 3500 °C) that is capable of melting scrap metal. After a full heat is formed in the EAF, it is transported to the next stage, the Argon Oxygen Decarburization (AOD), where the carbon content of the molten steel is reduced by injecting an argon-oxygen gas mixture. In order to ensure specific parameters of the molten steel for the final stage of casting, a heat goes through the Ladle Furnace (LF) stage to adjust the chemistry and temperature to their specified values. Finally, the heat is casted in the Continuous-casting (CC) stage, where specific rules about the sequences of heats apply. The process is shown in Fig. 2.

There are several production constraints that have to be satisfied by the scheduling model. We consider two parallel, non-identical machines at each stage. For all stages, except of the CC, processing

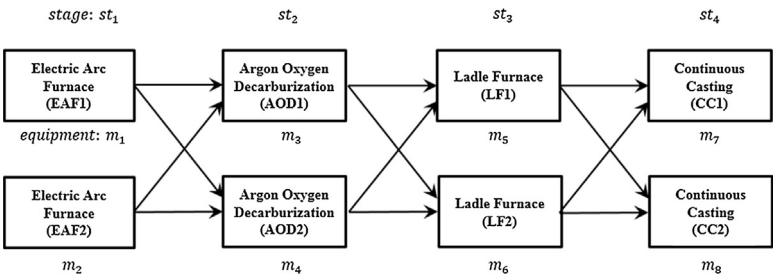

Fig. 2. Stainless-steel production process (melt shop section).

of a subsequent heat can be carried out only after an equipment specific setup has been performed. Between subsequent stages, a heat is transported with some minimum time requirement which differs depending on the two units considered. The time spent by a heat waiting between two subsequent stages is restricted by a maximum allowed hold-up time in order to avoid a too-large drop of the temperature of the molten steel. Heats of the same heat group are casted subsequently on the CC without waiting times.

3.2. Electricity demand considerations

The above mentioned production process consumes large amounts of electricity, in the considered case up to 192 MW. The energy demand for this process must always be met, i.e. the plant is assumed to purchase at least the amount of electricity needed to satisfy the load curve that results from the production schedule. We consider demand-side response strategies which preserve the total production output over some given time horizon, in the computational studies over one day. Therefore, the total energy consumption stays roughly the same regardless of different energy prices, since the same number of products is considered to be processed. The only change in the total energy used might be due to non-identical specific electricity consumption of parallel machines at the same production stage, which is not the case in the considered example. Potential energy losses due to different waiting times in-between the production stages are neglected. It is assumed that the electricity consumption of any production task is constant over the time span of the task, following previous studies by for example Nolde and Morari (2010), Hadera and Harjunkoski (2013), Castro et al. (2013). This is a widely used simplification since exact modeling of e.g. the scrap melting phase would make the model more complex without bringing significant benefits. The challenge addressed in this work is to determine simultaneously an optimal purchase and sales policy for the electricity, with complex time- and load-sensitive purchasing options as shown in Fig. 3 and a production schedule that defines the demand of electricity.

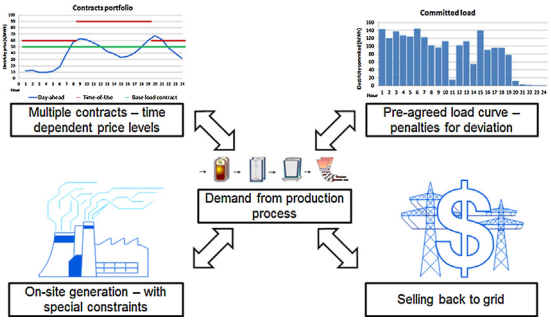

Fig. 3. Electricity bill structure.

For the industrial case study the purchasing contracts include:

- long-term contract (base contract or base load) – constant price, constant amount of electricity delivered over time;
- short-term contract (Time-of-Use or TOU) – two price levels (on- and off-peak);
- spot market (day-ahead) – hourly-varying prices, assumed to be known 24 h ahead;
- onsite generation– constant price with additional start-up costs.

The long-term contract is agreed with a provider usually for a period of 3–12 months. Over this time a certain fixed amount of electricity is available for the production plant at all times. The agreed amount must be purchased by the plant at an agreed advance price. Therefore, in a situation where there is no load consumption planned at some time interval, the surplus of electricity must be sold back to the grid. Establishing a long-term contract is usually considered profitable for the plant since the provider is able to offer a lower price for such a constant delivery over a long period.

The short-term contract (TOU) usually covers up to 3 months. Therefore, the price offered by the supplier is normally higher compared to the long-term contract and also agreed in advance. Here, we assume that the contract has two different price levels corresponding to on- and off-peak times. The off-peak price is lower than the on-peak price which applies during the daylight period. Another contract considered in the case study is the day-ahead spot market. Here, the price follows regional fluctuations of electricity availability; therefore, it varies on an hourly basis. Whether the exact prices of the day-ahead contract are known in advance depends on the market the plant enters and contractual commitments with the local operator. For some cases, they are settled and published exactly one day ahead. Often, the price of the day-ahead market is not known in advance for the scheduling activities at the plant, because the price is settled through a bidding policy. However, based on historical data and the weather forecast, it can be predicted through appropriate forecasting tools. In practice, the uncertainty in the prices might, however, have a significant impact on the scheduling problem, but only in the case of very rare unexpected events on the power supply side, e.g. power plant shut down due to an equipment failure. Otherwise, the day-ahead electricity price forecasts are considered precise enough. In this work the uncertainty of the prices is not considered to have a noticeable impact on the problem due to the short-term nature of the scheduling.

Apart from purchase contracts, the plant may have the possibility of producing electricity internally, which is subject to additional constraints. A start-up cost needs to be accounted for in the total cost of onsite generation for each time the onsite power generation is started up. Also, minimum runtime and downtime restrictions apply to avoid frequent start-ups and shut-downs of the power plant which lead to an accelerated deterioration of the plant.

The total electricity bill can be reduced by selling electricity back to the grid. The price of selling electricity also differs on an hourly

basis, depending on the regional situation in the grid. In the case of low availability of electric power, the plant can decrease its internal demand, to use the negotiated contracts and to use onsite generation in order to sell the electricity with a profit. This might happen especially in regions with heavy industry and at low temperatures during winter time. Often, as in the case of the day-ahead market, it is not known in advance at what rate the electricity can be sold. However, in practice this can also be predicted based on the day-ahead market price forecast.

The electricity bill, apart from the electricity purchase costs, includes so called deviation penalties, included also earlier in for example Nolde and Morari (2010), Hadera and Harjunkoski (2013), Castro et al. (2013). The plant is assumed to predict its load consumption for a period of 24 h minimum one day before the actual load occurs. In practice, the load curve (so called pre-agreed or committed load curve) is computed based on the optimized schedule to provide a forecast for the next 24-h energy consumption. This forecast is sent to the energy supplier, committing the plant to a certain load profile. If the actual consumption differs from this profile, the plant is obliged to pay penalties. Here, we assume that both under- and over-consumption are penalized, but with a penalty-free tolerance margin of a few percent. Following other studies, it is also assumed that the level of penalties (i.e. the price per deviating kWh) is known. In reality, very often their level is forecasted and determined after the actual load occurs, since only then the electricity market operator is able to quantify their true cost. However, it is important to note that usually there is a correlation between the penalty levels and the day-ahead market prices, i.e. for example sometimes underconsumption during high peak price hours may lead to negative penalties.

4. Monolithic model

The proposed MILP formulation describes a power intensive steel making process that produces a set of products (heats) $p \in P$ on a set of units $m \in M$, while satisfying various operational constraints. The plant is assumed to deliver a fixed number of products that are known in advance. The power consumption is both unit and product specific. The goal is to determine a one day production schedule that minimizes the total (net) cost of electricity and the weighted starting times of the tasks (i.e. a throughput related criterion). Electricity purchase includes different options and is subject to hourly price-variations. The optimization should determine the optimal amounts to be transferred from or to the electricity sources or sinks $i \in I$ at any given time interval $s \in S$. The end of the last time slot is equal to the scheduling horizon. Penalties due to the deviation from a pre-agreed load curve are incurred when a certain penalty-free buffer is exceeded and may differ for under- and over-consumption. The electricity bill can be reduced by selling the surplus of electricity. The monolithic models are described using the notation shown in Table 1. Additional notation introduced for the bi-level solution heuristic is given in Section 6.

4.1. Structure of the monolithic model

For the problem described in Section 3 we describe a monolithic model in this section. It consists of several components that are shown in Fig. 4. First, to ensure that all process specific constraints are satisfied, a scheduling model is formulated using the continuous-time general precedence approach (Section 4.2). The use of this approach is motivated by the required level of precision stemming from the specification by the industrial end-user. In order to optimize the purchase of electricity, and to augment the schedule in order to express potential changes of load pattern, a strategy for expanding the scheduling model with

Table 1
Model notation.

Sets:	
P	heats (products) to be produced
HG	heat groups with defined sequence of casting
$HGP(HG,P)$	subset of heats p mapped to corresponding heat group hg
$L(HG,P),F(HG,P)$	subset of heats p cast respectively last or first in a heat group casting sequence hg
M	equipment (machines)
EAF, AOD, LF, CC	subsets of equipment
S	time intervals
ST	production stage
$SM(ST,M)$	production stage st mapped to corresponding equipment m
$Node$, I, J	nodes in flow network denoting sources and sinks of electricity
$Pur(Node)$	purchase contracts node
$Dem(Node)$	production process electricity demand node
$Gen(Node)$	onsite generation node
$Bal(Node)$	balancing node
$Sale(Node)$	electricity sale sink node
$ARC_{i,j,s}$	defined arc between nodes i and j in time slot s
Parameters:	
$\theta_{p,m}$	processing duration of heat p on equipment m
t_m^{setup}	setup time for machine m
$t_{m,m'}^{min}$	minimum transport time from equipment m to m'
$t_{p,st}^{max}$	maximum hold-up (waiting) time after stage st
a_s	pre-agreed (committed) load curve
τ_s	electricity consumption time slot boundary
$h_{p,m}$	specific power consumption of processing heat p on equipment m
$c_{s,i,j}$	electricity cost of flow from i to j in time slot s
$f_{s,i,j}^{min}$, $f_{s,i,j}^{max}$	minimum and maximum flow between nodes i and j
r^{min}, d^{min}	minimum run- and down-time of onsite generation
c^{start}	startup cost of onsite generation
k	coefficient of delivered power reduction due to startup of onsite generation
c	coefficient of task start time weight in the objective function
Variables:	
$t_{m,p}^s$, $t_{m,p}^f$	positive continuous variables of starting and finishing time of heat p on equipment m
$t_{p,st}^s$, $t_{p,st}^f$	positive continuous variables of starting and finishing time of heat p at stage st
$w_{p,st}$	positive continuous variables of waiting time of heat p after stage st
q_s	positive continuous variables of electricity consumed in time slot s
$X_{m,p}$	binary variable, true when heat p is assigned for processing on equipment m
$V_{st,p,p'}$	binary variable, true when heat p' is processed after heat p on stage st
$Y_{p,st,s}^s$, $Y_{p,st,s}^f$	binary variable, true when heat p starts or finishes on stage st in the slot s
$G_{s,i,j}$	binary variable, true when generation is running in time slot s
$g_{s,i,j}^s$	pseudo-continuous positive variable denoting if onsite generation start-up occurred in time slot s
$y_{p,m,st,s}^{saux}$, $y_{p,m,st,s}^{faux}$	auxiliary continuous positive variable true when heat p is assigned for processing and started or finished processing on stage st in time slot s
$a_{p,m,st,s}$, $b_{p,m,st,s}$, $c_{p,m,st,s}$, $d_{p,m,st,s}$	positive continuous variables accounting for processing time of heat p on equipment m on stage st spent within a slot s
b_s	positive continuous variables of buffer level for allowed deviation from committed load in time slot s
b_s^o, b_s^u	positive continuous variables of upper and lower bounds for buffer in time slot s
c_s^o, c_s^u	positive continuous variables of actual over- and under-consumption in time slot s
$f_{s,i,j}$	positive continuous variables of flow from node i to j in time slot s
c_s^{gen}	positive continuous variables of cost of onsite generation in slot s
μ	continuous variable of net electricity consumption cost
δ	positive continuous variables of deviation penalties cost

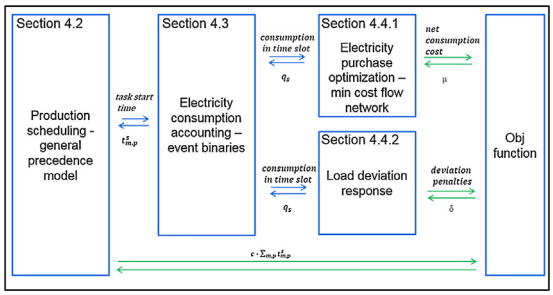

Fig. 4. Monolithic model structure.

energy-awareness was formulated (Section 4.3). This part of the monolithic model uses the continuous variable (used in the scheduling part) of task start time $t^s_{m,p}$ in order to find the contribution of a task to the electricity consumption within a given time interval s.

When applying this strategy for all tasks, the total electricity consumption q_s of the process in a given time interval can be computed, which is needed for the optimization of the cost of electricity (Section 4.4). This part computes optimal values in a flow network representing possible flows of electricity $f_{s,i,j}$ from sources to sinks. The optimization results in an optimal cost structure of the available purchase contracts with the exact amount of the electricity to be bought or sold under each contract. The knowledge of the process consumption during the time slots also enables to account for potential penalties δ paid due to deviations from pre-agreed load curve, and to determine when it is profitable to under- or over-consume electricity.

The objective function of the monolithic model takes into account the weighted task start times, the net electricity consumption cost μ and penalties δ paid for deviations. By choosing the weights in the summation, potential losses in the process (e.g. heat losses due to waiting time between the stages) or delays of the production can be traded off against the cost of electricity purchase and sales.

4.2. Production scheduling model

The general precedence scheduling model for the stainless-steel plant used in this study is largely based on the model introduced by Harjunkoski and Grossmann (2001). This model was further extended to a more flexible formulation introducing stages and multiple machines in Harjunkoski and Sand (2008). The scheduling part of the model uses assignment and precedence binaries following Eqs. (1)–(17) in Hadera and Harjunkoski (2013).

The scheduling model is based on the precedence variables and assignment variables that determine which of the parallel machines on each stage shall process a given heat. The general precedence $V_{st,p,p'}$ is true if a product p is processed before a product p' on a stage st. The assignment $X_{m,p}$ is true only when a given product p is processed on machine m. The sum (Eq. (1)) states that exactly one machine should process a heat per stage.

$$\sum_{m \in SM_{st,m}} X_{m,p} = 1 \quad \forall p \in P, st \in ST \tag{1}$$

Eq. (2) defines the finishing time $t^f_{m,p}$ as the starting time $t^s_{m,p}$ plus the selected processing length $\theta_{p,m}$.

$$t^f_{m,p} = t^s_{m,p} + X_{m,p} \cdot \theta_{p,m} \quad \forall m \in M, p \in P \tag{2}$$

Since a product can be processed only once on a given machine, the unassigned machines get a zero starting time (Eq. (3)).

$$t^s_{m,p} \le M \cdot X_{m,p} \quad \forall m \in M, p \in P \tag{3}$$

The stage starting and finishing times $t^s_{p,st}, t^f_{p,st}$ are synchronized with the corresponding machine times in Eqs. (4) and (5).

$$t^s_{p,st} = \sum_{m \in SM_{st,m}} t^s_{m,p} \quad p \in P, st \in ST \tag{4}$$

$$t^f_{p,st} = \sum_{m \in SM_{st,m}} t^f_{m,p} \quad p \in P, st \in ST \tag{5}$$

The scheduling model handles maximum hold-up times after processing has been completed on a given stage, equipment specific setup t^{setup}_m times and minimum transportation times. The processing on the next stage can be performed only after the processing of the previous stage has finished plus some waiting time $w_{p,st}$, which serves here as a slack variable which is determined by the optimization. The production flow between subsequent stages is established in Eq. (6).

$$t^s_{p,st+1} = t^f_{p,st} + w_{p,st} \quad \forall p \in P, st \in ST, st < |ST| \tag{6}$$

Due to process restrictions, it is necessary to enforce lower and upper bounds for the waiting times. The minimum corresponds to the physical possibility of transferring the product to the next stage, and it is equal to the minimum transportation time between machines $t^{min}_{m,m'}$ as stated in Eq. (7). The upper bound $t^{max}_{p,st}$ of the waiting time reflects the process constraint that a heat should not cool off below a certain level.

$$t^{min}_{m,m'}(X_{m,p} + X_{m',p} - 1) \le w_{p,st} \le t^{max}_{p,st} \quad \forall p \in P, m, m' \in M, st \in ST, \{st, m\} \in SM, \{st+1, m'\} \in SM, st < |ST| \tag{7}$$

The precedence of the products is characterized by the fact that either p is processed after p' or p' is processed after p. Therefore, only one of the two binaries can be true. Eq. (8) enforces a correct sequencing.

$$V_{st,p,p'} + V_{st,p',p} = 1 \quad \forall p, p' \in P, st \in ST, p < p' \tag{8}$$

In order to impose the common practice that the sequence of the products that are casted on a CC must propagate back to the other production stages, Eq. (9) is introduced.

$$V_{st,p,p'} = V_{st+1,p,p'} \quad \forall p, p' \in P, st \in ST, p < p', st < |ST| \tag{9}$$

The precedence constraint in Eq. (10) for other stages than CC restricts that a next heat should be processed only after the previous one has finished plus a setup time.

$$t^s_{m,p'} \ge t^f_{m,p} + t^{setup}_m - (M + t^{setup}_m)(3 - V_{st,p,p'} - X_{m,p} - X_{m,p'}) \quad \forall p, p' \in P, m \in M, st \in ST, \{st, m\} \in SM, p \ne p', st < |ST| \tag{10}$$

At the CC-stage no setup time t^{setup}_m should occur to ensure continuous casting (Eq. (11)). However, a setup must be carried out between the last $L(HG,P)$ and first $F(HG,P)$, heats of different heat groups (Eq. (12)).

$$t^s_{m,p'} \ge t^f_{m,p} - M(3 - V_{st,p,p'} - X_{m,p} - X_{m,p'}) \quad \forall p, p' \in P, m \in M, st \in ST, \{st, m\} \in SM, p \ne p', st = |ST| \tag{11}$$

$$t^s_{m,p'} \ge t^f_{m,p} + t^{setup}_m - (M + t^{setup}_m)(3 - V_{st,p,p'} - X_{m,p} - X_{m,p'}) \quad \forall p \in L(HG, P), p' \in F(HG, P), m \in M, st \in ST, \{st, m\} \in SM, p \ne p', st = |ST| \tag{12}$$

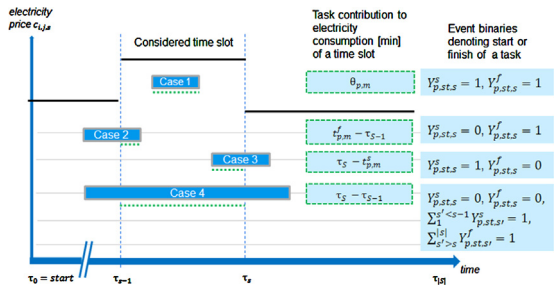

Fig. 5. Event binaries model to describe the consumption of electric energy in time slots of the price.

Constraint (13) ensures that heats of the same heat group are assigned to the same caster.

$$t^{\max UL1}_{p',st1} = \max_{p\in P}\{t^{\max}_{p,st1}\} + \max_{p\in P}\{\tau_{p,AOD1}, \tau_{p,AOD2}\} + \max_{p\in P}\{t^{\max}_{p,st2}\} + \max_{p\in P}\{\theta_{p,LF1}, \theta_{p,LF2}\} + \max_{p\in P}\{w^{\max}_{p,st3}\} \quad \forall p' \in P \tag{13}$$

As the heats are pre-ordered within a casting sequence, Eq. (14) ensures that next heat in a sequence starts immediately after the previous one has finished.

$$t^s_{p+1,st} = t^f_{p,st} \quad \forall p \in P\backslash L(HG,P), st \in ST, st = \left|ST\right| \tag{14}$$

From technical process requirements, the heat sequence within one heat group is known. The precedence of heats within one heat group is enforced and redundant values are eliminated in Eq. (15). Redundant sequencing variables are eliminated when comparing two identical products in Eq. (16).

$$V_{st,p,p'} = 1 \quad \forall p,p' \in P, p < p', st \in ST, hg \in HG, \{hg,p\}, \{hg,p'\} \in HGP(P) \tag{15}$$

$$V_{st,p,p'} = 0 \quad \forall p,p' \in P, st \in ST, p = p' \tag{16}$$

Since the goal of the production plant is to meet the production targets as soon as possible, minimizing the makespan (or tasks completion time) can be specified as an objective function in the MILP model.

4.3. Energy-awareness in precedence based scheduling

In continuous-time models, it is challenging to account for resource consumption. In this work we extend the scheduling model described above to account for the electricity consumption by each task within given time intervals of interest. The time grid with intervals in our use case corresponds to volatile electricity prices and committed load values. Therefore the length of the intervals is 1 h. The scheduling model uses continuous task start time variables, which are linked to the energy-aware part of the model, leading to the computation of the overall electricity consumption within a given time interval. Once the model is complemented by energy-awareness, both the electricity purchase and the load commitment can be optimized.

4.3.1. Model with event binaries

They key idea in the approach developed here is the use of the two event binaries representing whether a given task started ($Y^s_{p,st,s}$) or finished ($Y^f_{p,st,s}$) in or before or after particular time slot s related to the energy pricing (Fig. 5).

Since the boundaries of the time slot s are known, Big-M constraints in Eqs. (17)–(20) force the event binaries to be true in case the start or finish variable takes a value between the time slot's upper bound τ_s and lower bound τ_{s-1}.

$$t^s_{p,st} \geq \tau_{s-1}\cdot Y^s_{p,st,s} \quad \forall p \in P, st \in ST, s \in S \tag{17}$$

$$t^s_{p,st} \leq \tau_s + (M-\tau_s)(1-Y^s_{p,st,s}) \quad \forall p \in P, st \in ST, s \in S \tag{18}$$

$$t^f_{p,st} \geq \tau_{s-1}\cdot Y^f_{p,st,s} \quad \forall p \in P, st \in ST, s \in S \tag{19}$$

$$t^f_{p,st} \leq \tau_s + (M-\tau_s)(1-Y^f_{p,st,s}) \quad \forall p \in P, st \in ST, s \in S \tag{20}$$

However, the use of the stage set st in the definition of the event binaries does not indicate which of the available equipment of this stage is actually processing. Therefore, together with the assignment variable $X_{p,m}$, we can introduce two additional auxiliary pseudo-binary variables $y^{saux}_{p,m,st,s}$ and $y^{faux}_{p,m,st,s}$ that will be true only in case the respective event binary is true and the assignment is true as well (Eqs. (21)–(26)).

$$y^{saux}_{p,m,st,s} \geq X_{m,p} + Y^s_{p,st,s} - 1 \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{21}$$

$$y^{saux}_{p,m,st,s} \leq X_{m,p} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{22}$$

$$y^{saux}_{p,m,st,s} \leq Y^s_{p,st,s} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{23}$$

$$y^{faux}_{p,m,st,s} \geq X_{m,p} + Y^f_{p,st,s} - 1 \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{24}$$

$$y^{faux}_{p,m,st,s} \leq X_{m,p} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{25}$$

$$y^{faux}_{p,m,st,s} \leq Y^f_{p,st,s} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{26}$$

The two auxiliary binaries have indices representing product p, machine m, stage st and time slot s. That enables us to introduce continuous variables that are used to capture different cases of how a particular task (here a heat processed on a unit) relates to a time slot. As shown in Fig. 5 there are four different scenarios:

4.3.1.1. A task is processed entirely within the time slot. Processing within a time slot means that the start and finish time of the task must be placed within the time slot upper and lower boundary, both event binaries need to hold true (case 1 in Fig. 5). To capture this case we introduce an auxiliary variable $a_{p,m,st,s}$ described in Eqs. (27)–(29).

$$a_{p,m,st,s} \geq y^{saux}_{p,m,st,s} + y^{faux}_{p,m,st,s} - 1 \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{27}$$

$$a_{p,m,st,s} \leq y^{saux}_{p,m,st,s} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{28}$$

$$a_{p,m,st,s} \leq y^{faux}_{p,m,st,s} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{29}$$

4.3.1.2. A task starts before and finishes within the time slot. For the second case (Fig. 5) the start binary shall be zero for the considered slot. However the finish binary must hold true. That combination of the two binaries is enough to capture the time contribution $b_{p,m,st,s}$ of the task, as shown in Eqs. (30)–(33).

$$b_{p,m,st,s} \geq t^f_{p,m} - \tau_{s-1} - (M-\tau_{s-1})(1 - y^{faux}_{p,m,st,s} + y^{saux}_{p,m,st,s}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st,m\} \in SM \tag{30}$$

$$b_{p,m,st,s} \le t^{f}_{p,m} - \tau_{s-1} + \tau_{s-1}(1 - y^{faux}_{p,m,st,s}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{31}$$

$$b_{p,m,st,s} \le (\tau_s - \tau_{s-1}) y^{faux}_{p,m,st,s} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{32}$$

$$b_{p,m,,st,s} \le (\tau_s - \tau_{s-1})(1 - y^{saux}_{p,m,st,s}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{33}$$

4.3.1.3. A task starts within and finishes after the time slot. In this situation (case 3 in Fig. 5), the task should start between the lower and the upper bounds of the time slot and finish sometime after the upper bound. That means the start event binary is true for the slot and the finish event binary is false. The variable $c_{p,m,st,s}$ is defined by Eqs. (34)–(37).

$$c_{p,m,st,s} \ge \tau_s - t^{s}_{p,m} - \tau_s(1 - y^{saux}_{p,m,st,s} + y^{faux}_{p,m,st,s}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{34}$$

$$c_{p,m,st,s} \le \tau_s - t^{s}_{p,m} + (M - \tau_s)(1 - y^{saux}_{p,m,st,s}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{35}$$

$$c_{p,m,st,s} \le (\tau_s - \tau_{s-1}) y^{saux}_{p,m,st,s} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{36}$$

$$c_{p,m,st,s} \le (\tau_s - \tau_{s-1})(1 - y^{faux}_{p,m,st,s}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{37}$$

4.3.1.4. A task over-spans the full time slot. This occurs only when the start time of the task is placed before the lower bound of the time slot and at the same time the finish time of the task occurs after the upper bound of the time slot (case 4 in Fig. 5). This translates into zero values for both of the event binaries. In addition, the start event binary is true in one of the earlier slots before the considered one, and similarly, the finish event binary is true in one of the later time slots. The variable $d_{p,m,st,s}$ is defined by the constraints in Eqs. (38)–(42). If the task either started or finished entirely before the considered slot or after the slot it does not contribute to the electricity consumption within the slot.

$$d_{p,m,st,s} \ge (\tau_s - \tau_{s-1}) \cdot \left(\sum_{1}^{s'<s} y^{saux}_{p,m,st,s'} + \sum_{s'>s}^{|S|} y^{faux}_{p,m,st,s'} - 1 \right) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{38}$$

$$d_{p,m,st,s} \le (\tau_s - \tau_{s-1}) \cdot \sum_{1}^{s'<s} y^{saux}_{p,m,st,s'} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{39}$$

$$d_{p,m,st,s} \le (\tau_s - \tau_{s-1}) \cdot \sum_{s'>s}^{|S|} y^{faux}_{p,m,st,s'} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{40}$$

$$d_{p,m,st,s} \le (\tau_s - \tau_{s-1})(1 - y^{saux}_{p,m,st,s}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{41}$$

$$d_{p,m,st,s} \le (\tau_s - \tau_{s-1})(1 - y^{faux}_{p,m,st,s'}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{42}$$

The constraints based on the above cases yield the continuous variables $b_{p,m,st,s}$, $c_{p,m,st,s}$, $d_{p,m,st,s}$ accounting for how much time a given processing task spends within the considered time slot. Since the specific electricity consumption of the processing task is known, a proper summation of a product of the continuous variables and machine-specific electricity consumption parameter accounts for the total consumption in a given time slot (Eq. (43)). Note again that the electricity consumption of a given task is assumed to be constant over time, otherwise multiple steps for each task or a non-linear expression would need to be introduced. The approximation using a constant consumption rate still creates an acceptable error. A more accurate model, with varying consumption rates, would also be much more vulnerable to production disturbances and delays. The above described approach yields fewer binaries than the one used by Nolde and Morari (2010), where six binary variables are used to describe the relation between the task and the time slot. Here only two event binaries are needed.

$$q_s = \frac{\sum_{p \in P, m \in M} h_{p,m}(a_{p,s,st,m}\tau_{p,m} + b_{p,s,st,m} + c_{p,s,st,m} + d_{p,s,st,m})}{60} \quad \forall s \in S \tag{43}$$

A set of tightening constraints can help to speed up the computational performance of the model. In Eqs. (44) and (45), we restrict that for only one slot within the entire time horizon, the event binary is active. Additionally, it is true only when a task exist, i.e. when a product is assigned to be processed on a machine. Eq. (46) accounts for total consumption of the schedule to be equal to sum of total consumption of those tasks that has been assigned. Note that the potential energy losses due to higher waiting time (thus higher energy requirement due to the cooling effect) are not modeled here.

$$\sum_{s \in S} Y^{s}_{p,st,s} = \sum_{m \in M, \{st,m\} \in SM} X_{m,p} \quad \forall p \in P, st \in ST \tag{44}$$

$$\sum_{s \in S} Y^{f}_{p,st,s} = \sum_{m \in M, \{st,m\} \in SM} X_{m,p} \quad \forall p \in P, st \in ST \tag{45}$$

$$\frac{\sum_{p \in P, m \in M} X_{m,p} \cdot h_{p,m} . \pi_{p,m}}{60} = \sum_{s \in S} q_s \tag{46}$$

4.3.2. Literature model

Other formulations of resource consumption accounting in continuous-time based scheduling models have already been reported in the literature. Nolde and Morari (2010) presented a formulation that introduces six binaries to capture six different cases of the position a task might have relative to a time interval on the time axis. This approach was later reformulated by Hadera and Harjunkoski (2013) to account for parallel machines at each

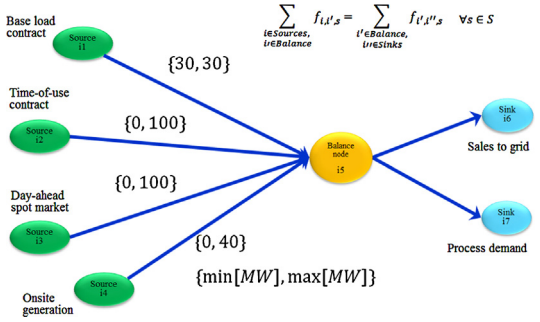

Fig. 6. Formulation of the electricity purchase and sale optimization problem with example bounds on the flows.

production stage with the goal of optimizing the cost of electricity for a single price curve and load deviation penalties. To reduce the model size, starting and finishing times of tasks were replaced by the corresponding stage starting and finishing times. The resulting model formulation is presented in Appendix A and is later used in the computations (Section 5.3) to compare its performance to the event binaries formulation described in Section 4.3.1.

4.4. Optimization of the cost of electricity

Tracking of the consumption of electricity over the time intervals can be used for optimizing the purchase and sales strategy. Once the scheduling model has been extended by the corresponding values of electricity consumption in the time slots, the purchase optimization can influence the schedule in such way that a mixed criterion (with e.g. task start times as used later) that includes also the cost of electricity is minimized.

4.4.1. Multiple purchase sources optimization

The idea for purchase optimization is based on a minimum cost flow network formulation (Fig. 6) with a balancing node for which all inflows are equal to all outflows (Eq. (47)). The inflow nodes represent the possible sources of electricity. The outflow nodes are the process demand and the selling of electricity.

$$\sum_{i\in Node} f_{s,i,j'} = \sum_{j\in Node} f_{s,j',j} \quad \forall (i,j'),(j',j) \in Arc, j' \in Bal, s \in S \tag{47}$$

The balancing node is connected with the sink and the source nodes by arcs that are characterized by parameters and variables. An arc exists only if there is a cost defined for it. The parameters are the minimum and maximum levels of the flows between two given nodes (Eq. (48)) and the cost function.

$$f^{\min}_{s,i,j} \le f_{s,i,j} \le f^{\max}_{s,i,j} \quad \forall (i,j) \in Arc, i,j \in Node, s \in S \tag{48}$$

The network is used to identify the most economical flows while satisfying the load from the process demand node (Eq. (49)).

$$q_s = \sum_{i\in Node, j\in Dem} f_{s,i,j} q_s = \sum_{i\in Node, j\in Dem} f_{s,i,j} \quad \forall (i,j) \in Arc, s \in S \tag{49}$$

The onsite generation is modeled using a binary variable $G_{s,i,j}$ that denotes whether the plant is in production mode (Eq. (50)) and an auxiliary pseudo-continuous variable $g^s_{s,i,j}$ indicating generation start-up (Eqs. (51) and (52)). Here, the Big-M value M_2 should not be less than the maximum flow on the arc between the onsite generation and the balancing node.

$$G_{s,i,j} \le f_{s,i,j} \le M_2 \cdot G_{s,i,j} \quad \forall (i,j) \in Arc, i \in Gen, j \in Bal, s \in S \tag{50}$$

$$G_{s,i,j} - G_{s-1,i,j} \le g^s_{s,i,j} \le G_{s,i,j} \quad \forall (i,j) \in Arc, i \in Gen, j \in Bal, s \in S \tag{51}$$

$$0 \le g^s_{s,i,j} \le 1 - G_{s-1,i,j} \quad \forall (i,j) \in Arc, i \in Gen, j \in Bal, s \in S \tag{52}$$

The onsite generation constraints are kept simple by considering a constant generation cost with additional start-up cost (Eq. (53)) and a reduced production rate by a factor k for those time intervals where a start-up occurs (Eq. (54)).

$$c^{gen}_s = \sum_{i\in Node, j\in Gen} f_{s,i,j} \cdot c_{s,i,j} + c^{start} \cdot g^s_{s,i,j} c^{gen}_s$$

$$= \sum_{i\in Node, j\in Gen} f_{s,i,j} \cdot c_{s,i,j} + c^{start} \cdot g^s_{s,i,j} \quad \forall (i,j) \in Arc, s \in S \tag{53}$$

$$f_{s,i,j} = f^{\max}_{s,i,j} \cdot G_{s,i,j} - k \cdot f^{\max}_{s,i,j} \cdot g^s_{s,i,j} \quad \forall (i,j) \in Arc, i \in Gen, j \in Bal, s \in S \tag{54}$$

Moreover, a minimum runtime r^{min} and a minimum downtime d^{min} are enforced (Eqs. (55) and (56)). The implementation of more detailed constraints that are available in literature would also be possible here, including accounting for steam flows and more detailed electricity production rates as for example in Mitra et al. (2013).

$$\sum_{s'=s}^{s+r^{\min}-1} G_{s',i,j} \ge r^{\min}(G_{s,i,j} - G_{s-1,i,j})$$

$$\forall (i,j) \in Arc, i \in Gen, j \in Bal, s \in S, s < |S| - r^{\min} \tag{55}$$

$$\sum_{s'=s}^{s+d^{\min}-1} G_{s',i,j} \le d^{\min}(1 + G_{s,i,j} - G_{s-1,i,j})$$

$$\forall (i,j) \in Arc, i \in Gen, j \in Bal, s \in S, s < |S| - d^{\min} \tag{56}$$

The final net electricity purchase cost (Eq. (57)) is composed of the cost associated with purchase from contracts, the cost of the generation and the revenues from the electricity sold.

$$\mu = \sum_{s\in S} \Big(\sum_{i'\in Node, j'\in Pur} f_{s,i',j'} \cdot c_{s,i,j} + c^{gen}_s - \sum_{i\in Node, j\in Sale} f_{s,i,j} \cdot c_{s,i,j} \Big)$$

$$\forall (i,j),(i',j') \in Arc \tag{57}$$

4.4.2. Load deviation problem

As a steel plant is a large consumer, the suppliers of electricity impose that it commits to certain hourly varying levels of load. We only consider here the situation where there is only one supplier and the committed load is the total load of the steel plant. Instead it could also be the load that is covered by one of the contracts, or there could be several such curves. In case the actual consumption deviates from the pre-agreed values, financial penalties are incurred. The part of the model that accounts for the penalties is similar to the formulation in Hadera and Harjunkoski (2013). For the load tracking error penalties, it is assumed that there is a penalty-free deviation (buffer) b_s that is relative to the committed consumption a_s and limited by relative upper and lower bounds b^o_s and b^u_s as stated in Eq. (58).

$$-a_s \cdot b^u_s \le b_s \le a_s \cdot b^o_s \quad \forall s \in S \tag{58}$$

The actual levels of over- and under consumption (c^o_s and c^u_s) are determined by Eq. (59).

$$q_s = a_s + c^o_s - c^u_s + b_s \quad \forall s \in S \tag{59}$$

The penalty term δ calculating the fines p^o, p^u for over- and under consumption is given by Eq. (60).

$$\delta = p^o \cdot \sum_{s \in S} c_s^o + p^u \cdot \sum_{s \in S} c_s^u \tag{60}$$

The final objective function of the monolithic model in Eq. (61) minimizes the net electricity cost μ, the deviation penalties and the weighted sum of the task starting times $t^s_{p,m}$ with c being a weighting factor. Two out of three objective function components are related directly to a quantitative energy cost. In contrast, the lead times (e.g. makespan) are not easily transferable into money. It is difficult to assess how much each production hour is worth, this also differs for different plants. That is why the flexibility is left to the final user to determine the weight, depending on how much production throughput should be emphasized. With different coefficients, different schedules can be generated and the plant can select the most preferred one. For later numerical studies it is assumed $c = 1$ in order to emphasize the energy over the throughput, since the plant pursues to augment its schedule according to the energy market conditions.

$$\min \left(\mu + \delta + c \cdot \sum_{p \in P, m \in M} t^s_{p,m} \right) \tag{61}$$

The part of the model that concerns the deviation problem can easily be used in load commitment of one particular contract. For example, when changing the variable representing the total consumption in a time slot q_s to the amount drawn from Time-of-Use source $f_{s,i,j}$, where $i \in TOU$ and $j \in Bal$ we get the committed load problem of the TOU contract.

5. Industrial case study

The assumptions on the constraints of the steel making process include the knowledge of the sequences and assignments of products to the last stage, the Continuous-Casting (CC) stage. We assume that it is known which products must be processed on one of the casters. However, the assignment of the heats to other units in other stages must be determined by the optimization. We also assume that the sequence of the heats that must be processed on a particular caster is known. However, it is up to the optimization to determine the sequence of those products that can be processed on two different CC machines. This assumption reduces the size of the search space. It is a reasonable assumption because very often the sequence of the products to be processed is dictated by higher level planning solutions (e.g. mill-wide planning) that are directly linked with customer orders and knowledge concerning in-house inventory levels.

For integrated steel plants, the further processing of the steel slabs is carried out in the Hot Rolling Mill (HRM) after the Melt Shops section. At the HRM section it is important to define a sequence of steel slabs to be rolled such that the cost of reheating using natural gas is minimized. This challenging optimization problem of coordination between Melt Shop and Hot Rolling Mill (Xu et al., 2012) can also determine the sequence of the products on the CC stage. Usually, the assignments and the sequences on the casters reflects the quality requirements for steel, i.e. one of the casters processes certain high quality types of steel, while the other one might not be able to deliver the same qualities. The timing of the tasks to avoid production delays and to minimize the cost of energy-related problem is subject of the optimization.

Table 2
Heat group definition.

Group	Heat (product)
HG1	P1–P3
HG2	P4–P7
HG3	P8–P12
HG4	P13–P16
HG5	P17–P20

Table 3
Processing times and electricity consumption.

	EAF1, EAF2	AOD1, AOD2	LF1, LF2	CC1, CC2
P1–P20	85 [min] 85 [MW]	8 [min] 2 [MW]	45 [min] 2 [MW]	60 [min] 7 [MW]

5.1. Calculation of lower and upper bounds of task start times

To tighten the MILP model, we calculate lower and upper bounds for the task start variables. For each heat group we solve two optimization problems, minimizing and maximizing the task start time of the first product in the heat group at the CC stage. In this way we check what is the minimum value of the variable when a given heat group finishes as soon as possible on the CC. Similarly, we check what is the maximum value of the variable when a given heat group finishes as late as possible on the CC. Based on this knowledge, it is possible, using process parameters, to calculate the earliest start times and the latest start times of each task at the other stages as shown in Appendix B. The bounds obtained from the above optimization are then propagated to the monolithic model and to the heuristic optimization in order to impose upper and lower bounds for task start and finish time variables. Finding tighter bounds helps to speed up the solution of the MILP model since then many of the energy-related binaries can be set to zero.

5.2. Case study data

Following the Demand Response strategy the plant has the goal of delivering a fixed number of heats (products) within a scheduling horizon of 24 h. Due to the continuous casting requirement, products are divided within heat groups as defined in Table 2.

For the test cases with fewer products the last heats were excluded. Processing times and specific electricity consumption of the tasks are given in Table 3, while setup times are reported in Table 4. Minimum transportation times and maximum waiting (hold-up) times after processing on a given stage are shown in Tables 5 and 6 respectively.

The input data concerning the electricity purchase limits are shown in Fig. 6. Note that the base load contract has a fixed amount of delivery for each hour of the day, regardless whether the electricity is needed for the production process or not. The prices of electricity and the committed load curve are shown in Figs. 7 and 8. The electricity prices of both day-ahead contract cases, low price (EPEX Spot, 2013, Germany/Austria 23/09/2013) and high price (EPEX Spot, 2013, France 10/02/2012) are taken from a real spot market. The pre-agreed load curve comes from a valid production

Table 4
Setup times [min].

Machine	Setup time
EAF1, EAF2	9
AOD1, AOD2	5
LF1	15
LF2	5
CC1	50
CC2	70

Table 5
Transportation times [min].

	AOD1	AOD2	LF1	LF2	CC1	CC2
EAF1	10	25				
EAF2	25	10				
AOD1			4	20		
AOD2			20	4		
LF1					20	45
LF2					45	20

Table 6
Maximum waiting times after stages.

	ST1	ST2	ST3
P1-P20	60	90	60

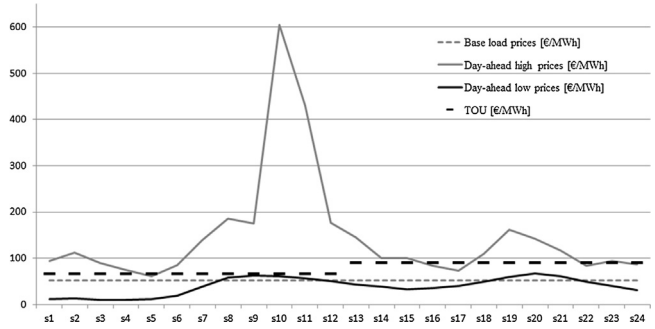

Fig. 7. Electricity prices for case studies.

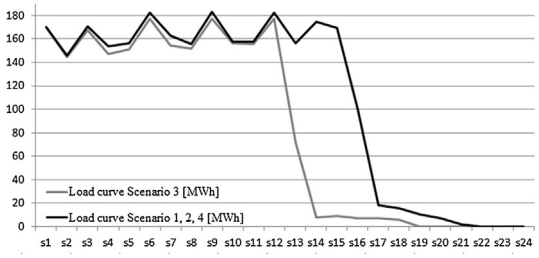

Fig. 8. Pre-agreed load curves.

Table 7
Onsite generation parameters.

Cost of onsite generation [€/MWh]	61
Minimum run time [h]	3
Minimum down time [h]	3
Start-up cost [€]	1000
Reduced production rate due to start-up	20%

schedule which was computed not considering the energy cost in the optimization, but in our case only the lead times optimization ($c \cdot \sum_{m\in M, p\in P} t^s_{m,p}$). This follows previous studies (Castro et al., 2013) where the schedule with optimized production-specific cost (makespan) served as a basis for the comparison with an energy-driven schedule to assess the iDSM benefits.

We assume that the income from selling back electricity to the grid is also time sensitive and is equal to 75% of the cost of day-ahead market in the same time slot. The cost of onsite generation and other related parameters are shown in Table 7. The allowed range for over- and under-consumption and the corresponding penalties can be found in Table 8, together with other load-deviation problem related parameters.

Table 8
Load deviation problem parameters.

Over-conumption penalty [€/MWh]	100
Under-consumption penalty [€/MWh]	80
Buffer for over-conumption	3%
Buffer for under-conumption	4%

5.3. Numerical results obtained with the monolithic models

Numerical tests have been performed on a 4-core Intel Xeon 2.53 GHz with 16GB of RAM using GAMS/CPLEX 23.7.3. The monolithic formulations shown in Section 4 and Appendix A were tested on the same instances with identical process assumptions and input data from Section 5.2. Table 9 gives the overview of the test cases involving different numbers of products to be scheduled, different electricity prices and pre-agreed load curves considered. The first two scenarios have a typical production target with the day-ahead spot market prices are high and low. In the third scenario the production target is lowered to represent underutilized capacity. The fourth scenario is similar to the latter. However the pre-agreed load curve is used as for the first two scenarios. This simulates a situation where due to unexpected equipment break-downs (known before the schedule is computed) the plant cannot deliver the planned number of products and it has therefore overcommitted the load curve.

The two monolithic model strategies are compared for different problem instances (scenarios) with computational limitations of 600 s and 3600 s. Allowing a higher computation time than 1 h does not yield noticeable improvements of the solution quality. In addition, higher solution times are not acceptable from the practical point of view, since the desired solution time is 5–10 min. The resulting computational statistics are shown in Table 10. The MIP solution quality is described by the value of the total weighted objective function value.

The results show that for all of the problem instances the new event binaries models (HM1-4) perform better than the literature model (NM1-4), especially for larger problems (see Tables 11 and 12). This statement holds true for the short computational time limit (600 s) where the differences of the relative gaps between the solutions is usually around 8 – 13% (scenario 1–3). For longer solution times the improvement obtained with the event binaries model is similar, around 7–15%. In Table 11 the economic assessment of the resulting solutions is shown, for the computation time limit of 600 s. The purchase strategy obtained from the solutions of the two monolithic models differs. However, in some cases for some contracts both models seem to be choosing the same levels of the flows in the network. For example, both models recognize that it is preferable to generate electricity from the power plant and to buy from TOU market when the day-ahead prices are high as in the case of scenarios 1 and 4.

Table 12 gives the economic results for the test runs under the computational limitation of 3600 s. When discussing the economic results it is important to note that the final objective function consists of three components: net electricity cost, penalties for load deviation and lead times (summation of task start times with the weight factor $c = 1$). Therefore, the solver can find two similar solutions in terms of the objective function values. However, the distribution of the costs among the elements of the objective function might differ. In Scenario 3 for 3600 s computation time the objective value for both model types is quite similar, but the six binaries model solution chooses to decrease the net purchase cost at the expense of higher deviation penalties cost, while the event binaries model follows better the pre-agreed load curve causing the net purchase cost to increase.

Table 9
Test case description.

Scenario	Horizon	Products	Electricity sources and sinks
1	24 h	20	all possible, day-ahead with high prices
2	24 h	20	all possible, day-ahead with low prices
3	24 h	16	all possible, day-ahead with high prices
4	24 h	16	all possible, day-ahead with high prices, overcommitted load curve (as for 20 products)
Name	Model type		
NM	monolithic six binaries model (Nolde and Morari, 2010; Hadera and Harjunkoski, 2013)		
HM	monolithic event binary model (Hadera et al., 2014)		

Table 10
Numerical results of monolithic models.

Scenario	Model statistics						
	Binary vars	Total vars	Equations	MIP solution 600 s	Relative gap 600 s	MIP solution 3600 s	Relative gap 3600 s
NM1	13,017	29,508	98,095	313,128	43.78%	290,708	38.97%
HM1	4065	29,508	102,335	247,838	29.30%	241,136	26.80%
NM2	13,017	29,508	98,095	223,887	32.30%	222,167	31.20%
HM2	4065	29,508	102,335	200,038	24.90%	180,023	16.10%
NM3	10,181	23,428	77,136	174,227	32.63%	156,986	24.63%
HM3	3229	23,428	80,528	155,226	22.81%	146,339	17.93%
NM4	10,181	23,428	77,136	234,643	31.53%	221,454	27.20%
HM4	3229	23,428	80,528	204,173	22.50%	180,965	12.10%

The new model requires fewer binaries, while the total number of variables remains the same. The number of equations increases in the event binaries model. Unfortunately, the proposed model is not able to cope with the size of the industrial problem. Therefore, in the next sections we propose a heuristic strategy for the solution of the integrated scheduling and energy cost optimization problem.

6. Bi-level heuristic

When trying to solve an instance of the problem with significant flexibility in the process, i.e. when the optimization is free to assign and to sequence all products, the computational performance of the monolithic models from Section 4 is not sufficient. This is mainly due to large number of binary variables in the scheduling formulation that makes it difficult to solve and to the loose Big-M constraints, which is specific for precedence-based continuous-time models. To overcome the computational limitations we describe a heuristic decomposition strategy.

For large scale scheduling problems decomposition techniques have long been recognized as possible solution approaches. Starting from fundamental studies by Benders (1962) and Dantzig (1963) with row and column generation approaches, strategies have been developed for solving problems in an iterative fashion that could not be solved using a monolithic formulation. Decomposition approaches can be categorized into approaches that can be shown to converge to the true optimum (even if convergence may be slow) and approaches that discard a part of the solution space so that optimality cannot be guaranteed (decomposition heuristics).

Table 11
Economic assessment results of the monolithic models – 600 s computation time limit.

Scenario	Economic assessment – 600 s						
	Lead times [min]	Net electricity cost [€]	Electricity purchase [€]	Deviation penalties [€]	Day-ahead market [MWh]	TOU [MWh]	Onsite generation [MWh]
NM1	60,784	149,832	161,098	102,512	172.55	1471.825	912
HM1	51,990	133,972	151,905	61,876	173.49	1421.44	952
NM2	61,210	119,440	98,505	43,236	1608.917	4	352
HM2	53,946	120,989	98,592	25,103	1514.51	95.78	432
NM3	42,283	86,071	88,984	45,872	1417.983	54.267	0
HM3	33,038	96,508	128,208	25,679	82.604	1241.156	952
NM4	43,598	96,266	134,759	94,780	77.207	1266.043	952
HM4	36,152	72,846	140,130	95,175	142.48	1318.8	952

Table 12
Economic assessment results of the monolithic models – 3600 s computation time limit.

Scenario	Economic assessment – 3600 s						
	Lead times [min]	Net electricity cost [€]	Electricity purchase [€]	Deviation penalties [€]	Day-ahead market [MWh]	TOU [MWh]	Onsite generation [MWh]
NM1	58,763	146,373	162,528	85,572	203.72	1466.61	952
HM1	50,796	142,452	14,396	47,888	234.22	1349.66	952
NM2	60,282	116,872	98,942	45,014	1649.35	16.83	312
HM2	51,598	115,937	98,229	12,489	1635.81	56.11	352
NM3	39,444	83,531	89,620	34,012	1547.368	0.597	0
HM3	32,753	95,840	130,217	17,746	83.779	1270.351	952
NM4	42,269	100,104	127,986	79,081	110.87	1164.93	952
HM4	35,396	94,723	128,534	50,846	46.66	1256.92	952

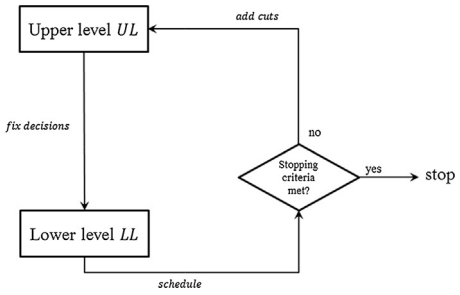

Fig. 9. General idea of bi-level heuristic approach.

Wu and Ierapetritou (2003) described a number of different heuristic decomposition approaches for scheduling problems. For example, one may use time decomposition where the long time horizon is divided into several smaller sub-periods with resulting sub-problems. Another important class of approaches make use of Lagrangean decomposition (Guignard and Kim, 1987) to relax the original problem into a problem that is easier to solve, systematically providing a lower bound for the solution. For problems with a clear separation of planning level decisions and scheduling level decisions these can be represented in a bi-level setup where first in upper level the planning variables are determined and then fixed to solve the more detailed lower level scheduling problem. This scheme was used for example by Bassett et al. (1996). Similarly in another example, Erdirik-Dogan and Grossmann (2008) use the bi-level concept for continuous multiproduct plants first solving an aggregate model to obtain an upper bound for the profit and then solving a scheduling problem to obtain a lower bound. Xu et al. (2012) developed a bi-level decomposition scheme for the coordination of a Melt Shop process with Hot Rolling section of a stainless steel plant. In this paper, we also employ a bi-level scheme, where the solution procedure consists of two problems that are solved in an iterative manner, as shown in Fig. 9.

First, an aggregate model (upper level *UL*) that approximates the original monolithic model is solved in order to obtain feasible values of some binary decisions. These binary decisions are passed to the full model (lower level *LL*) with a restriction to keep some of the variables fixed, optimizing some other continuous and binary variables, in our case the starting times and the event binaries. The full model should provide a feasible schedule and an objective function value which represents an upper bound of the optimal value. A new iteration of the algorithm starts by solving the upper problem again, with some new restrictions in the form of integer cuts that exclude previous solutions of the full model. The search space can be reduced based on the knowledge about the optimal solution provided from *LL*. In our particular case, since for the full problem with some of the decision variables fixed a feasible solution is obtained, the combination of the binaries of that solution can be removed from *UL* so that new values of these binaries are generated by the upper level model and the new solution is again refined by the lower level model. The algorithm iterates until a stopping criterion is met, e.g. until a time limit is exceeded. For the following sections, additional notation specific for the decomposition approach is given in Table 13.

6.1. Upper level problem

The upper level problem *UL* consist of solving two models *UL*1 and *UL*2 as shown in Fig. 10. The *UL*1 is a simplified model of the original problem and it is computed to obtain a valuable guess of some binary decisions, while the *UL*2 is a pre-computation step for the *UL*1 starting from the second iteration as explained later. The algorithm starts with solving *UL*1, which is constructed in such way

Table 13
Model notation for the heuristic.

Sets:	
$DY_0^r, DY_1^r, DX_0^r, DX_1^r, DV_0^r, DV_1^r$	dynamic sets used in bi-level heuristic for false and true decision of the respective binaries
Variables:	
$X_{m,p}^{UL1}, X_{m,p}^{LL}, X_{m,p}^{UL2}$	binary variable in respective models UL1, LL and UL2, true when heat p is assigned for processing on equipment m
$V_{st,p,p'}^{UL1}, V_{st,p,p'}^{LL}, V_{st,p,p'}^{UL2}$	binary variable in respective models UL1, LL and UL2, true when heat p' is processed after heat p on stage st
$Y_{p,st,s}^{sUL1}, Y_{p,st,s}^{sLL}, Y_{p,st,s}^{sUL2}$	binary variable in respective models UL1, LL and UL2, true when heat p starts on stage st in the slot s
Parameters:	
$t_{m,m'}^{\min UL1}, t_{m,m'}^{\min LL}$	minimum transport time from equipment m to m' in respective models UL1 and LL
$t_{p,st}^{\max UL1}, t_{p,st}^{\max LL}$	maximum hold-up time after stage st in respective models UL1 and LL
RHS	sum of binary variables
α	number of neighboring slots to be evaluated
β	desired optimality gap
r	iteration number

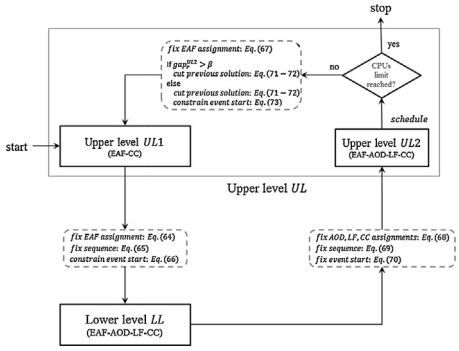

Fig. 10. Bi-level heuristic algorithm.

that it represents the full monolithic problem as closely as possible, while at the same time reducing the size of the MILP. The main component of the objective function value is the electricity-related cost. It depends directly on the load pattern that results from the processing of the tasks. In the stainless-steel production process investigated in the case study, the EAF stage consumes about 88% of the total electricity needed to deliver one product. Therefore, potential changes in the assignments, sequences or especially the timing of different products on that stage will have a significant impact on the final consumption pattern. Therefore, the energy-intense melting task is included in the upper level problem. A rough approximation of the lower level problem can be generated by simply scheduling the EAF stage alone, maintaining all energy-related constraints. However, the tasks on the first stage cannot be timed arbitrarily and must be sequenced according to the special continuous-casting constraints on the CC stage. For example, two subsequently casted products should be processed within a reasonable time interval in the EAF stage in order to ensure the proper delivery of the heats to continuous-casting, while at the same time satisfying all transfer and waiting time constraints between the stages.

Since the EAF and CC stages together account for around 95% of the total Melt Shop electricity consumption, scheduling of these two stages alone should produce a good guess of the values of the

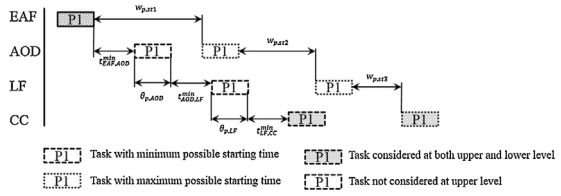

Fig. 11. Transportation and waiting time between EAF and CC stage in the upper level *UL1* problem.

variables related to the EAF and the CC of the full problem. If the last production stage is considered together with the EAF, the casting constraints are not violated and the remaining stages of AOD and LF can be scheduled on the lower level. In order to ensure feasibility of the lower model concerning decisions for these two stages, the upper level problem needs to account for the range of possible delays between processing on the EAF and on the CC stage. The equations of the *UL*1 problem are the same as in the corresponding monolithic model (and in lower level problem), apart from removing elements and cuts as follows:

$\min^{UL1}$ (Eq. (61))
s.t.:

Eqs. (1)–(16)	Scheduling model equations with new sets *M* and *ST* (Section 4.2)
Eqs. (17)–(46)	Energy-awareness extension with new sets *M* and *ST* (Section 4.3.1)
Eqs. (47)–(57)	Electricity sources optimization (Section 4.4.1)
Eqs. (58)–(60)	Load deviation problem (Section 4.4.2)
Cuts	New constraints for other iterations than the initial one (Section 6.4)

The equipment AOD and LF are eliminated from the equipment set *M*. The stages *st*2 and *st*3 are eliminated from the set of production stages *ST*. Therefore, new values of the minimum transport times and maximum hold-up times between EAF and CC stage in the upper level problem need to be calculated based on the parameters of the full model as shown in Fig. 11.

The new $t^{\max UL1}_{p,st}$ and $t^{\min UL1}_{m,m',CC}$ replace the original $t^{\max}_{p,st}$ and $t^{\min}_{m,m'}$ from the monolithic model. The maximum hold-up time corresponds in the full model to the maximum time after which a heat can be processed on CC after finished on EAF as in Eq. (62) below.

$$\begin{aligned} t^{\max UL1}_{p',st1} &= \max_{p\in P}\{t^{\max}_{p,st1}\} + \max_{p\in P}\{\tau_{p,AOD1}, \tau_{p,AOD2}\} \\ &+ \max_{p\in P}\{t^{\max}_{p,st2}\} + \max_{p\in P}\{\theta_{p,LF1}, \theta_{p,LF2}\} \\ &+ \max_{p\in P}\{w^{\max}_{p,st3}\} \quad \forall p' \in P \end{aligned} \tag{62}$$

Similarly, the minimum transportation time between EAF and CC corresponds to the minimum possible time between these two in the full problem as in Eq. (63).

$$\begin{aligned} t^{\min UL1}_{EAF,CC} &= \min_{\substack{m\in SM(EAF,m)\\ m'\in SM(AOD,m')}} \{t^{\min}_{m,m'}\} \\ &+ \min_{p\in P}\{\theta_{p,AOD1}, \theta_{p,AOD2}\} + \min_{\substack{m\in SM(AOD,m)\\ m'\in SM(LF,m')}} \{t^{\min}_{m,m'}\} \\ &+ \min_{p\in P}\{\theta_{p,LF1}, \theta_{p,LF2}\} + \min_{\substack{m\in SM(LF,m)\\ m'\in SM(CC,m')}} \{t^{\min}_{m,m'}\} \end{aligned} \tag{63}$$

In the upper level model *UL*1, the EAF stage is the first stage, followed by the CC which is the second and last production stage. Another modification of the input data of the upper level problem concerns the pre-agreed load curve. For the original full problem, the agreed curve is calculated based on a pre-defined schedule. For the same schedule, it is possible to eliminate the AOD and LF stages to obtain a load curve for the other two stages.

The second model *UL*2 of the upper level is solved after the lower level *LL* problem as shown in Fig. 10. From the latter, most binary decisions are fixed and transferred to *UL*2, which essentially is the same problem as *LL* discussed in the next section. However, within *UL*2 the only binary decision to be determined by optimization is to find better assignments of heats to EAFs in order to pre-compute new assignment decisions on EAFs for the next iteration of *UL*1. In this way the search space of the approximate model *UL*1 is reduced and it is no longer a relaxation of the original problem in the later iterations, which might prevent it of finding the optimal solution. However, speeds up the computational time significantly.

6.2. Lower level problem

The constraints and sets of the lower level *LL* problem are not changed compared to the monolithic problem. However, the lower level problem is solved with some fixed decisions which improves its computational performance, as discussed in details in the next Section. The model *LL* serves as an evaluation model for the decisions that were determined by the upper level *UL*1. After fixing some decisions, as described in the next section, *LL* is solved with a limitation on the solution time to avoid spending too much time in closing a small optimality gap.

6.3. Information exchange between the levels

Since the EAF stage is the most power intensive one, the decisions taken with regard to the assignment $X_{p,m}$ to machines of the melting stage are fixed for the *LL* problem as in Eq. (64), which helps to speed up the solving time. Further, for the same reason another variable is fixed, the sequence $V_{st,p,p'}$ on the casting stage as in Eq. (65). In contrast to the process assumptions where the sequence in the particular caster is known a priori only if only one caster can be used as described in Section 5, we fix the sequence relation of the products between the two casters here which is a degree of freedom of the monolithic problem.

$$X^{LL}_{m,p} = X^{UL1}_{m,p} \quad \forall p \in P, m \in EAF \tag{64}$$

$$V^{LL}_{st,p,p'} = V^{UL1}_{st,p,p'} \quad \forall p, p' \in P, p \neq p', st = |ST| \tag{65}$$

Since the upper problem should provide a good approximation of the full problem, it would be beneficial to use also the energy-related information obtained from it for fixing some decisions in the lower level problem. A natural choice is the event binaries. However, since it is expected that these have a large impact on the value of the objective function, the kind of fixing needs to be carefully chosen. The fixing should still allow for providing flexibility to the model, and at the same time reduce the computational time of the full problem. After experimenting with different options, we developed a fixing decision that if the upper level problem is solved close to optimality (i.e. the gap is lower than $\beta = 2\%$) the variables of event binary start of the lower level problem $Y^{sLL}_{p,st,s}$ should be true within a neighborhood of the slots for which the binary holds true in the *UL*1 solution as shown in Eq. (66) where s^* denotes the time slot when the event binary starts to hold true in the *UL*1 solution.

$$\sum_{s=s^*-\alpha}^{s^*+\alpha} Y^{sLL}_{p,st,s} = 1, \quad \text{where} \quad s^* : Y^{sUL1}_{p,st,s^*} = 1 \quad \forall p \in P, st \in ST \tag{66}$$

If the upper level problem determined that the start of a product should occur in the *n*th time slot, then the start of that product in the *LL* solution should occur in one of the time slots within $(n-\alpha\,;n+\alpha)$. For the particular case, we choose α to define a neighborhood of 3 slots, which is a wide range of 7 h in total. Since the decision of the

event start binary has a direct impact on the event finish binary, there is no need for further fixing of the latter.

With the above exchange of information between models *UP*1 and *LL*, the most important degrees of freedom in the lower level problem are the timing of EAFs (but also all the other units), while keeping the sequence determined by the upper level.

In order to update the *UL*1 problem with new assignment decisions on EAFs (Eq. (67)), the *UL*2 problem is solved with fixed decisions of the other binaries, as shown in Eqs. (68)–(70).

$$X^{UL1}_{m,p} = X^{UL2}_{m,p} \quad \forall p \in P, m \in EAF \tag{67}$$

$$X^{UL2}_{m,p} = X^{LL}_{m,p} \quad \forall p \in P, m \in M \backslash EAF \tag{68}$$

$$V^{UL2}_{st,p,p'} = V^{LL}_{st,p,p'} \quad \forall p, p' \in P, p \neq p', st \in ST \tag{69}$$

$$Y^{sUL2}_{p,st,s} = Y^{sLL}_{p,st,s} \quad \forall p \in P, st \in ST, s \in S \tag{70}$$

The *UL*2 model has very few degrees of freedom since it can only change the binaries related to the EAF assignment. In contrast, in *UL*1 is the one where many important decisions are made since this model finds the timing and sequence of products on the most important units, especially on the EAF. The latter are then fixed at the lower level.

6.4. Cuts and stopping criteria

In the proposed approach, cuts imposed in each iteration are related to the scheduling decisions ($X_{m,p}$, $V_{p,p',st}$) and the energy-awareness ($Y^s_{p,st,s}$). Of course the latter are strongly related to the former since it is the timing of a task start which links both. In the case when *LL* is not proven to have a desired level of optimality, we can suspect that the decisions obtained from it might not be good enough to later cut off the neighborhood of the obtained solution of event binary start variables from the solution space of *UL*1. Therefore, if for a particular iteration the desired optimality level is not obtained in the *LL* problem, the cut for the *UL*1 involves only removing a particular solution of *LL*, which means a particular combination of the binaries $X_{m,p}$, $V_{p,p',st}$, $Y^s_{p,st,s}$ obtained in *LL* as there is no need of evaluating that solution again in new iteration in the upper level problem. The cut is achieved by the constraints shown in Eqs. (71) and (72), similar to those reported (Balas and Jeroslow, 1972) and successfully used in the literature (Iyer and Grossmann, 1998) for the elimination of existing binary solutions. In case where *LL* is solved to optimality we can enforce a stronger cut, removing also the neighborhood of the event binary start as shown in Eq. (73), following the fixing in Eq. (66) coming from the *UL*1.

$$RHS = \sum_{p\in P, s\in S} Y^{sUL1}_{p,st1,s} + \sum_{m\in M, p\in P} X^{UL1}_{m,p} + \sum_{p,p'\in P, p\neq p'} V^{UL1}_{st4,p,p'} \tag{71}$$

$$\sum_{(p,s)\in DY^r_1} Y^{sUL1}_{p,st1,s} - \sum_{(p,s)\in DY^r_0} Y^{sUL1}_{p,st1,s} + \sum_{(m,p)\in DX^r_1} X^{UL1}_{m,p} - \sum_{(m,p)\in DX^r_0} X^{UL1}_{m,p} - + \sum_{(p,p')\in DV^r_1} V^{UL1}_{st4,p,p'} - \sum_{(p,p')\in DV^r_0} V^{UL1}_{st4,p,p'} \leq RHS - 1$$
$$DY^r_0 = \{(p,s) | Y^{sUL1}_{r,p,st1,s} = 0\}\, DY^r_1 = \{(p,s) | Y^{sUL1}_{r,p,st1,s} = 1\},\, DX^r_0 = \{(m,p) | X^{UL1}_{r,m,p} = 0\}\, DX^r_1 = \{(m,p) | X^{UL1}_{r,m,p} = 1\}, \tag{72}$$
$$DV^r_0 = \{(p,p') | V^{UL1}_{st4,p,p'} = 0\}\, DV^r_1 = \{(p,p') | V^{UL1}_{st4,p,p'} = 1\},\, p \neq p'$$

$$\sum_{(p,s')\in DY^r_1} Y^{sUL1}_{p,st1,s'+\gamma} - \sum_{(p,s')\in DY^r_0} Y^{sUL1}_{p,st1,s'+\gamma} + \sum_{(m,p)\in DX^r_1} X^{UL1}_{m,p} - \sum_{(m,p)\in DX^r_0} X^{UL1}_{m,p} + \sum_{(p,p')\in DV^r_1} V^{UL1}_{st4,p,p'} - \sum_{(p,p')\in DV^r_0} V^{UL1}_{st4,p,p'} \leq RHS - 1$$
$$\forall \gamma \in (-\alpha; +\alpha), \alpha = 3\, DY^r_0 = \{(p,s') | Y^{sUL1}_{r,p,st1,s'} = 0\}\, DY^r_1 = \{(p,s') | Y^{sUL1}_{r,p,st1,s'} = 1\},\, DX^r_0 = \{(m,p) | X^{UL1}_{r,m,p} = 0\} \tag{73}$$
$$DX^r_1 = \{(m,p) | X^{UL1}_{r,m,p} = 1\},\, DV^r_0 = \{(p,p') | V^{UL1}_{st4,p,p'} = 0\}\, DV^r_1 = \{(p,p') | V^{UL1}_{st4,p,p'} = 1\},\, p \neq p'$$

The algorithm performs the iterative steps as shown in Fig. 10. The upper problem *UL*1 is not a strict mathematical relaxation except of the first iteration. Therefore, we cannot use the objective function to systematically close the gap between the lower and the upper bounds as it was the case for example in Iyer and Grossmann (1998). In the later iterations, *UL*1 is not a relaxed problem of the monolithic model because it considers the assignment of the EAFs as fixed and as long as this fixing is not optimal the solution from the upper level problem is not a valid lower bound. The assignments coming from *UL*2 to *UL*1 are used to speed up the computation time of solving *UL*1, which most of the times is not solvable to near-optimal solutions in short times, thus giving weak solution without the fixing. It is reasonable to use the fixing also because of its much lower importance on the objective function compared to the degrees of freedom that the *UL*1 is handling, namely timing and sequencing.

Therefore, the solution of *UL*1 does not provide an increasing lower bound. At the same time the lower level problem *LL* and *UL*2 provide upper bounds as feasible solutions of the monolithic problem - the latter is always at least as good as the one from *LL*. Since the proposed algorithm does not guarantee to convergence to the optimal solution, reasonable stopping criterion for the iterative execution is the total time spent on computations or the desired number of iterations, which is acceptable for industrial practice as long as the algorithms yields good quality solutions in reasonable computation times.

Table 14
Upper level *UL*1 problem maximum waiting times.

	ST1	ST2
P1-P20	161	90

Table 15
Upper level *UL*1 problem minimum transportation times.

	CC1	CC2
EAF1	155	161
EAF2	161	155

6.5. Application of the heuristic to the industrial case study

We tested the bi-level heuristic on the same problem instances as the monolithic model. In the decomposition scheme, some modifications of the input data are needed, due to the elimination of the AOD and LF stages. The new maximum waiting times and minimum transportation times of the upper level *UL*1 that were calculated using Eqs. (62) and (63) are shown in Tables 14 and 15.

The committed load curve for the upper level problem *UL*1 is modified by considering the lower consumption due to omitting the AOD and LF stages as shown in Table 16.

Table 16
Pre-agreed load curve for the upper level *UL*1 problem.

Time interval	Load curve UL1 Scenario 1, 2, 4 [MWh]	Load curve UL1 Scenario 3 [MWh]
s1	170	170
s2	144.5	144.5
s3	167.17	167.167
s4	147.33	147.33
s5	151.03	151.03
s6	177	177
s7	157.1	154.183
s8	151.97	151.96
s9	177	177
s10	153.37	156.167
s11	152.9	155.817
s12	177	177
s13	151.5	72.167
s14	168.5	7.817
s15	163.15	9.333
s16	94.83	7
s17	12.02	7
s18	14	6.183
s19	8.98	0
s20	7.00	0
s21	1.98	0
s22	0	0
s23	0	0
s24	0	0

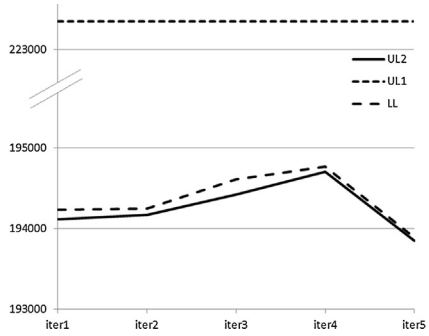

Fig. 12. Objective function value change in each iteration for all models of Scenario 1.

6.5.1. Numerical results of the heuristic approach

Since the heuristic approach does not guarantee to systematically provide a better upper bound with each iteration the best solution of *UL*2 among all iterations is considered to be the bi-level algorithm's solution. Therefore, *UL*2's solution statistics are reported in Table 17 which shows that the approach is always able to find better quality solutions, within the given time limit, compared to the monolithic formulation. To assess the solution obtained from the heuristic (relative gap) we compared it with a best bound (LP relaxation reported by the solver) obtained from optimization runs of the corresponding monolithic model with the heuristic solution being provided as the initial solution for the solver.

The heuristic decomposition always obtains better solutions than the monolithic model, usually by around 14%. The quality of these solutions is expected to be better. However, it is difficult to find an optimal solution or a best bound to asses them. Very good results of the heuristic decomposition are achieved already in the first iteration. Often after up to 3–4 iterations the best solution is found. In Fig. 12 the evolution of objective function values for all event binaries models in Scenario 1 is shown. It can be observed that the *UL*1 values in each iteration of the algorithm are constant, even though due to the cuts, each iteration finds different solution

Table 17
Numerical results for monolithic (HM) and bi-level heuristic (H) approaches – 600 s computation limit.

Scenario	Model statistics							Economic assessment						
	Binary vars (UL2)	Total vars (UL2)	Equations (UL2)	MIP solution (UL2)	MIP solution LL	MIP solution UL1	Relative gap	Lead times [min]	Net electricity cost [€]	Electricity purchase [€]	Deviation penalties [€]	Day-ahead market [MWh]	TOU [MWh]	Onsite generation [MWh]	No. of iterations (best)
HM1	4065	29,508	102,335	247,838	–	–	29.30%	51,990	133,972	151,905	61,876	173.49	1421.44	952	–
H1	1458	29,508	102,335	193,845	193,888	227,293	9.89%	46,176	147,087	156,853	582	176.349	1456.514	952	5(5)
HM2	4065	29,508	102,335	200,038	–	–	24.90%	53,946	120,989	98,592	25,103	1514.51	95.78	432	–
H2	1458	29,508	102,335	165,196	165,285	160,707	9.09%	45,472	119,443	101,531	281	1512.868	196.161	392	5(3)
HM3	3229	23,428	80,528	155,226	–	–	22.81%	33,038	96,508	128,208	25,679	82.604	1241.156	952	–
H3	1276	23,428	80,528	134,588	134,749	169,197	9.87%	30,626	103,057	133,980	904	133.033	1243.566	952	3(1)
HM4	3229	23,428	80,528	204,173	–	–	22.50%	36,152	72,846	140,130	95,175	142.48	1318.8	952	–
H4	1276	23,428	80,528	176,006	176,244	153,727	8.71%	35,289	98,990	123,190	41,727	13.13	1228.59	952	4(3)

from all the previous ones. This is due to the fact that a slight change in the timing and assignment or sequence of products (while satisfying the cuts) is very likely to give the same objective value since there are many similar solutions in *UL*1. However, when solving the more detailed *LL* model the values are changing in each iteration in response to the different decisions taken in *UL*1. For the same reason the objective function value of *LL* can improve in further iterations since there are AOD and LF stages added as well as the new load deviation curve. It can be noted that the solution quality is not expected to improve significantly in further iterations as the bi-level solution method is based on the idea that the upper level should provide a very good rough schedule already in the first iteration. The objective function value of *UL*2 always improves the solution from *LL* slightly by finding a better assignment on EAF units. It should be also noted that for Scenarios 2 and 4 the objective function value of *UL*1 is lower than *LL* and *UL*2; however, this is not true for Scenario 1 and 3. The reason that higher values might appear in approximated *UL*1 is larger deviation penalties paid than in detailed *LL* and *UL*2 due to the changed pre-agreed load curve. A general behavior of the algorithm very similar to the one shown in Fig. 12 was observed for all of the investigated scenarios.

7. Conclusions and remarks

In this paper, we have proposed a new strategy for embedding energy-awareness into a continuous-time scheduling approach which optimizes the production schedules of energy-intensive plants (Section 3) under consideration of to time-sensitive prices of electricity and load commitment penalties (Section 4). The proposed approach was compared to the model by Nolde and Morari (2010). The numerical experiments (Section 5.3) show that the use of the new event binaries is more efficient. However, both monolithic models cannot be solved within the available computation times for large-scale industrial problem instances. Therefore, we developed a bi-level decomposition-based heuristic (Section 6) to obtain good quality results in reasonable computation times.

The proposed solution scheme benefits from the exact timing of the tasks by the continuous-time scheduling representation. The model is able to capture complex price structures and to optimally determine the exact amount of electricity to be purchased and sold. The flexible part of the purchase optimization can be further extended to more complex dependencies between the contracts. The model might help assessing different price levels of negotiated contracts, as well as reducing the risk associated with volatile electricity markets. An important restriction is that the plant needs to make commitments on the amounts to be bought and sold on the day-ahead markets. Even more important factors are the disturbances and the technical capability to implement the optimized schedule.

To address further the limitations of the model concerning computational performance for large instances, a scheduling horizon of several days could be investigated with a rolling horizon approach. Decisions for longer time windows should be done with higher level short- and long-term planning solutions taking into account different factors than those considered by the scheduling level. Further work could also deal with improvements of the developed algorithm toward a more rigorous scheme.

Acknowledgments

We would like to acknowledge the Marie Curie FP7-ITN project "Energy savings from smart operation of electrical, process and mechanical equipment – ENERGY-SMARTOPS", Contract No: PITN-GA-2010-264940 for financial support.

Appendix A.

The literature based extension of energy-awareness for continuous-time scheduling models uses six different cases of how a task can contribute to electricity consumption within a considered time slot:

A.1. A task is processed entirely within the time slot

Processing within a time slot means that stage's finishing time $t^f_{p,st}$ occurs before the time slot's finishing time τ_s and stage's starting time $t^s_{p,st}$ occurs later than the time slot's starting time τ_{s-1}. For this case the binary variable $A_{p,s,st}$ will be true, thus equations using Big-M formulation are written as in Eqs. (A.1) and (A.2). The duration of processing within the slot will in this case be equal to the processing time of the task itself.

$$t^f_{p,st} \le \tau_s + (M - \tau_s)(1 - A_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.1}$$

$$t^s_{p,st} \ge \tau_{s-1} - \tau_{s-1}(1 - A_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.2}$$

A.2. A task starts before and finishes within the time slot

Second case occurs if stage's start time $t^s_{p,st}$ occurs before the lower boundary of the considered slot (Eq. (A.5)), however the stage's finish time $t^f_{p,st}$ is placed within the slot (Eqs. (A.3) and (A.4)). For this case the binary variable $B_{p,s,st}$ will be true. Processing time contribution of the task within the slot is equal to the tasks' finishing time $t^f_{p,m}$ minus the lower boundary τ_{s-1} of the considered time slot.

$$t^f_{p,st} \ge \tau_{s-1} - \tau_{s-1}(1 - B_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.3}$$

$$t^f_{p,st} \le \tau_s + (M - \tau_s)(1 - B_{p,s,st} - A_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.4}$$

$$t^s_{p,st} \le \tau_{s-1} + (M - \tau_{s-1})(1 - B_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.5}$$

A.3. A task starts within and finishes after the time slot

Similarly to the second case, the task's start time $t^s_{p,st}$ occurs within the considered time interval (Eqs. (A.7) and (A.8)) and at the same time finishing time $t^f_{p,st}$ is placed after the upper boundary of the slot (Eq. (A.6)). For this case the binary variable $C_{p,s,st}$ will be true. The time a task spent within the slot will equal to the upper boundary τ_s of the slot minus the start time $t^s_{p,m}$ of the task.

$$t^f_{p,st} \ge \tau_s - \tau_s(1 - C_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.6}$$

$$t^s_{p,st} \ge \tau_{s-1} - \tau_{s-1}(1 - C_{p,s,st} - A_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.7}$$

$$t^s_{p,st} \le \tau_s + (M - \tau_s)(1 - C_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.8}$$

A.4. A task over-spans the time slot

When duration of the task is longer than the time interval itself there might be a case when it over-spans the interval. This occurs only when the start time of the task $t^s_{p,st}$ is placed before the lower boundary of the time slot (Eq. (A.10)) and at the same time the finish time $t^f_{p,st}$ of task occurs after the upper bound of the slot (Eq. (A.9)). For this case the binary variable $D_{p,s,st}$ will be true. Then, the

amount of time the task contributed to the time slot will be equal to the length of the time slot itself ($\tau_s - \tau_{s-1}$).

$$t^f_{p,st} \geq \tau_s - \tau_s(1 - D_{p,s,st} - C_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.9}$$

$$t^s_{p,st} \leq \tau_{s-1} + (M - \tau_{s-1})(1 - D_{p,s,st} - B_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.10}$$

A.5. A task starts and finishes before the considered time slot

Here both the starting time $t^s_{p,st}$ and finishing time $t^f_{p,st}$ takes place before the starting of the considered time interval τ_{s-1}. For this case the binary variable $E_{p,s,st}$ will be true when finishing time $t^f_{p,st}$ occurs before the considered time slot, as in Eq. (A.11).

$$t^f_{p,st} \leq \tau_{s-1} + (M - \tau_{s-1})(1 - E_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.11}$$

A.6. A task starts and finishes after the considered time slot

Here both the starting time $t^s_{p,st}$ and finishing time $t^f_{p,st}$ takes place after the finishing of the considered time interval τ_s. For this case the binary variable $F_{p,s,st}$ will be true when starting time $t^s_{p,st}$ occurs later than upper bound of the considered time slot, as in Eq. (A.12).

$$t^s_{p,st} \geq \tau_s - \tau_s(1 - F_{p,s,st}) \quad \forall p \in P, s \in S, st \in ST \tag{A.12}$$

The big-M value is set to be the end of the scheduling horizon. The formulation is improved compared to Nolde and Morari (2010) by introducing second binary in the Big-M equations of similar boundary conditions as in Eqs. (A.4), (A.7), (A.9) and (A.10). To complete the formulation, an important constraint ensuring that there is only one of the six binaries true for a task has to be enforced, as in Eq. (A.13).

$$A_{p,s,st} + B_{p,s,st} + C_{p,s,st} + D_{p,s,st} + E_{p,s,st} + F_{p,s,st} = 1 \quad \forall p \in P, s \in S, st \in ST \tag{A.13}$$

With the help of the binaries being true for respective cases of task-time slot relation, it is possible to capture the amount of time a given task was processed in a particular time slot. The task's consumption within the slot can be accounted for by multiplying time spent with a parameter of specific electricity consumption of the task. Therefore, with summation of all tasks the total electricity consumption in the time slot is captured with the Eq. (A.14). The equation is divided by 60 to convert the unit from *MWmin* into *MWh*. In the equation two problems arise. First, there are two nonlinearities from the product of binary and continuous variable. Second, the equation do not account for the fact that one of the machines in the stage does not process a task.

$$q_s = \sum_{p,st,m \in SM_{st,m}} h_{p,m}(A_{p,s,st} \cdot \tau_{p,m} + B_{p,s,st}(t^f_{p,st} - \tau_{s-1}) + C_{p,s,st}(\tau_s - t^s_{p,st}) + D_{p,s,st}(\tau_s - \tau_{s-1}))\frac{1}{60} \quad \forall s \in S \tag{A.14}$$

In order to deal with the latter problem, a set of auxiliary variables can be designed for which those tasks not processing a product will have the time contribution to the slot put to zero. That means, whenever a product is not assigned to a machine the binaries of respective six cases shall be put to zero. For the first case with $A_{p,s,st}$ binary, it can only be true when assignment binary $S_{m,p}$ is true, as in Eqs. (A.15) and (A.16). Similarly for the $D_{p,s,st}$ binary as in Eqs. (A.17) and (A.18).

$$a_{p,m,st,s} \geq A_{p,s,st} - (1 - X_{m,p}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.15}$$

$$a_{p,m,st,s} \leq A_{p,s,st} + 1 - X_{m,p}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.16}$$

$$d_{p,m,st,s} \geq A_{p,s,st} - (1 - X_{m,p}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.17}$$

$$d_{p,m,st,s} \leq A_{p,s,st} + 1 - X_{m,p} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.18}$$

For the other cases of $B_{p,s,st}$ and $C_{p,s,st}$ by designing the auxiliary variable we also can deal with the nonlinearities, by applying an exact linearization method. The auxiliary variables will have the value of the time contribution of the respective binary case only both the case binary is true and the assignment is true as well. The constraints for the two cases are shown in Eqs. (A.19)–(A.26).

$$b_{p,m,st,s} \geq t^f_{p,st} - \tau_{s-1} - (M - \tau_{s-1})(2 - B_{p,s,st} - X_{m,p}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.19}$$

$$b_{p,m,st,s} \leq t^f_{p,st} - \tau_{s-1} + \tau_{s-1}(2 - B_{p,s,st} - X_{m,p}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.20}$$

$$b_{p,m,st,s} \leq (\tau_s - \tau_{s-1})(1 - B_{p,s,st} + X_{p,m}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.21}$$

$$b_{p,m,st,s} \leq (\tau_s - \tau_{s-1}) \cdot B_{p,s,st} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.22}$$

$$c_{p,m,st,s} \geq \tau_s - t^s_{p,st} - \tau_s(2 - C_{p,s,st} - X_{m,p}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.23}$$

$$c_{p,m,st,s} \leq \tau_s - t^s_{p,st} + (M - \tau_s)(2 - C_{p,s,st} - X_{m,p}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.24}$$

$$c_{p,m,st,s} \leq (\tau_s - \tau_{s-1})(1 - C_{p,s,st} + X_{m,p}) \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.25}$$

$$c_{p,m,st,s} \leq (\tau_s - \tau_{s-1}) \cdot C_{p,s,st} \quad \forall p \in P, m \in M, st \in ST, s \in S, \{st, m\} \in SM \tag{A.26}$$

Table B1
Calculation of bounds for task start time.

The lower and upper bound of $t^{s}_{p,st}$
initialize $t^{s,\min}_{p,st} = t^{s,\max}_{p,st} = 0$
for $st = st_4$, $hg \in HG$, $p \in F(HG, P)$ do

$$\text{set } t^{s,min}_{p,st} = \begin{cases} min(t^{s}_{p,st}) \\ s.t. \\ Scheduler\ constraints \end{cases}$$

$$\text{set } t^{s,\max}_{p,st} = \begin{cases} \max(t^{s}_{p,st}) \\ s.t. \\ Scheduler\ constraints \end{cases}$$

end for
Find the bounds of the other products at the last stage
for $st = st_4$, $hg \in HG$, $p \in P/F(HG, P)$ do
for $p \in F(HG, P) + 1, \ldots, p \in L(HG, P)$ do

$$t^{s,\min}_{p,st} = t^{s,\min}_{p-1,st} + \min_{m \in SM(st,m)} \theta_{p-1,m}$$

$$t^{s,\max}_{p,st} = t^{s,\max}_{p-1,st} + \max_{m \in SM(st,m)} \theta_{p-1,m}$$

end for
end for
Set the upper bound in the other stages as equal to the upper bound at the last stage

$$t^{s,\max}_{p,st} = t^{s,\min}_{p,st4} \quad \forall p \in P,\ st \in ST$$

Find the upper bound in stages other than last stage
for $st \in st_4, \ldots, st_2$, $P \in P$ do
for $st' = st - 1$ do

$$t^{s,\max}_{p,st'} = t^{s,\max}_{p,st} - \min_{m \in SM(st',m)} \tau_{p,m} - \min_{m \in SM(st,m), m' \in SM(st',m')} t^{\min}_{m,m'},$$

end for
end for
Find the lower bound in stages 1–3
for $st \in st_1, \ldots, st_3$, $P \in P$ do
for $st' = st + 1$ do

$$t^{s,\min}_{p,st'} = t^{s,\min}_{p,st} + \min_{m \in SM(st',m)} \tau_{p,m} + \min_{m \in SM(st,m), m' \in SM(st',m')} t^{\min}_{m,m'},$$

end for
end for

With the help of the auxiliary variables the final constraint for electricity consumption accounting can be changed from Eq. (A.14) to the one shown in Eq. (A.27).

$$q_s = \sum_{p \in P, st \in ST, m \in SM_{st,m}} h_{p,m}(a_{p,m,st,s}\tau_{p,m} + b_{p,m,st,s} + c_{p,m,st,s} + d_{p,m,st,s}(\tau_s - \tau_{s-1}))\frac{1}{60} \quad \forall s \in S \tag{A.27}$$

Appendix B.

See Table B1.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compchemeng.2015.02.004.

References

Ashok S. Peak-load management in steel plants. Appl Energ 2006;83(5):413–24.
Balas E, Jeroslow R. Canonical cuts on the unit hypercube. SIAM J Appl Math 1972;23(1):61–9.
Bassett M, Pekny JF, Reklaitis GV. Decomposition techniques for the solution of large-scale scheduling problems. AIChE J 1996;42(12):3373–87.
Benders JF. Partitioning procedures for solving mixed-variables programming problems. Numer Math 1962;4(3):238–52.
BGFRS – Board of Governors of Federal Reserve System. Industrial Production and Capacity Utilization; 2013 http://www.federalreserve.gov/datadownload/Download.aspx?rel=G17&series=0fc8d1a1edcda88d7d3db1971fa6d4b8&filetype=spreadsheetml&label=include&layout=seriescolumn&from=01/01/1950&to=12/31/2013 [accessed 02.02.14].
Boukas EK, Haurie A, Soumis F. Hierarchical approach to steel production scheduling under a global energy constraints. Ann Oper Res 1990;26:171–84.
Castro P, Grossmann IE, Veldhuizen P, Esplin D. Optimal maintenance scheduling of a gas engine power plant using generalized disjunctive programming. AIChE J 2014;60:2083–97, http://dx.doi.org/10.1002/aic.14412.
Castro P, Harjunkoski I, Grossmann IE. New continuous-time scheduling formulation for continuous plants under variable electricity cost. Ind Eng Chem Res 2009;48(14):6701–14.
Castro P, Harjunkoski I, Grossmann IE. Optimal scheduling of continuous plants with energy constraints. Comput Chem Eng 2011;35(2):372–87.
Castro P, Sun L, Harjunkoski I. Resource–task network formulations for industrial demand side management of a steel plant. Ind Eng Chem Res 2013;52(36):13046–58.
CRA (Charles River Associates). Primer on demand-side management. The World Bank; 2005.
Dantzig GB. Linear programming and extensions. Princeton, NJ, USA: Princeton University Press; 1963.
DENA. Dena grid study II: integration of renewable energy sources in the German power supply system from 2015–2020 with an outlook to 2025. Deutsche Energie-Agentur GmbH (dena); 2011.
DOE. A report to the United States Congress Pursuant to Section 1252 of the Energy Policy Act of 2005 Benefits of demand response in electricity markets and recommendations for achieving them A report to the United States Congress Pursuant to Section 1252 of the Energy Policy Act of 2005; 2006 http://energy.gov/oe/downloads/benefits-demand-response-electricity-markets-and-recommendations-achieving-them-report [accessed 02.02.14].
EPEX Spot. European power exchange; 2013 http://www.epexspot.com/en/ [accessed 11.11.13].
Erdirik-Dogan M, Grossmann IE. Simultaneous planning and scheduling of single-stage multi-product continuous plants with parallel lines. Comput Chem Eng 2008;32:2664–83.
Floudas C, Lin X. Continuous-time versus discrete-time approaches for scheduling of chemical processes: a review. Comput Chem Eng 2004;28(11):2109–29.
Guignard M, Kim S. Lagrangean decomposition: a model yielding stronger Lagrangean bounds. Math Program 1987;39:215–28.
Hadera H, Harjunkoski I. Continuous-time batch scheduling approach for optimizing electricity consumption cost. Comput Aided Chem Eng 2013;32:403–8.
Hadera H, Harjunkoski I, Grossmann IE, Sand G, Engell S. Steel production scheduling under time-sensitive electricity cost. Comput Aided Chem Eng 2014;33:373–8.
Hait A, Artigues C. On electrical load tracking scheduling for a steel plant. Comput Chem Eng 2011a;12(14):3044–7.
Hait A, Artigues C. A hybrid CP/MILP method for scheduling with energy costs. Eur J Ind Eng 2011b;5(4):471–89.
Harjunkoski I, Grossmann IE. A decomposition approach for the scheduling of a steel plant production. Comput Chem Eng 2001;25(11-12):1647–60.
Harjunkoski I, Sand G. Flexible and configurable MILP models for meltshop scheduling optimization. Comput Aided Chem Eng 2008;25:677–82.
Harjunkoski I, Maravelias C, Bongers P, Castro P, Engell S, Grossmann IE, Hooker J, Méndez C, Sand G, Wassick J. Scope for industrial applications of production scheduling models and solution methods. Comput Chem Eng 2014;62:161–93.
Ierapetritou MG, Wu D, Vin J, Sweeny P, Chigirinskiy M. Cost minimization in an energy-intensive plant using mathematical programming approaches. Ind Eng Chem Res 2002;41:5262.
Iyer RR, Grossmann IE. A bilevel decomposition algorithm for long-range planning of process networks. Ind Eng Chem Res 1998;37(2):474–81.
Li J, Xiao X, Tang O, Floudas CA. Production scheduling of a large-scale steelmaking continuous casting process via unit-specific event-based continuous-time models: short-term and medium-term scheduling. Ind Eng Chem Res 2012;51(21):7300–19.
Li P, Wendt M, Wozny G. Optimal production planning under uncertain market conditions. In: Bingzhen Chen, Arthur W, Westerberg, editors. Computer aided chemical engineering, 15. Elsevier; 2003. p. 511–6.
Maravelias C. General framework and modeling approach classification for chemical production scheduling. AIChE J 2012;58(6):1812–28.
Méndez CA, Cerdá J, Grossmann IE, Harjunkoski I, Fahl M. State-of-the-art review of optimization methods for short-term scheduling of batch processes. Comput Chem Eng 2006;30(6–7):913–46.
Mitra S, Grossmann IE, Pinto JM, Arora N. Optimal production planning under time-sensitive electricity prices for continuous power-intensive processes. Comput Chem Eng 2012;38:171–84.
Mitra S, Sun L, Grossmann IE. Optimal scheduling of industrial combined heat and power plant under time-sensitive electricity prices. Energy 2013;54:194–211.
NERC. Data collection for demand-side management for quantifying its influence on reliability; 2007 http://www.nerc.com/docs/pc/drdtf/NERC_DSMTF_Report_040308.pdf [accessed 02.02.14].
Nolde K, Morari M. Electrical load tracking scheduling of a steel plant. Comput Chem Eng 2010;34(11):1899–903.
Paulus M, Borggrefe F. The potential of demand-side management in energy-intensive industries for electricity markets in Germany. Appl Energ 2011;88(2):432–41.
Pochet Y, Wolsey LA. Production planning by mixed integer programming. New York: Springer; 2006.
Tang L, Liu J, Rong A, Yang Z. A review of planning and scheduling systems and methods for integrated steel production. Eur J Oper Res 2001;133(1):1–20.
Todd D, Caufield M, Helms B, Starke M, Kirby B, Kueck J. Project report to US Department of Energy contract DE-AC05-00OR22725 Providing reliability services through demand response: a preliminary evaluation of the demand response

capabilities of Alcoa Inc Project report to US Department of Energy contract DE-AC05-00OR22725; 2009.
Vujanic R, Mariéthoz S, Goulart PJ, Morari M. Robust integer optimization and scheduling problems for large electricity consumers. In: American control conference; 2012. p. 3108–13.
Wu D, Ierapetritou MG. Decomposition approaches for the efficient solution of short-term scheduling problems. Comput Chem Eng 2003;27(8):1261–76.
Xu C, Sand G, Harjunkoski I, Engell S. A new heuristic for plant-wide schedule coordination problems: the intersection coordination heuristic. Comput Chem Eng 2012;42:152–67.
Zhang Y, Tang L. Production scheduling with power price coordination in steel industry. In: Power and energy engineering conference (APPEEC); 2010. p. 1–4.