N-SIDE

Co-optimization of energy & ancillary services in GB

A report for NESO

May 2024

Authors	Contact
Anthony Papavasiliou	apa@n-side.com +30 694 784 3105
Marcelo Torres	mto@n-side.com +30 695 192 1831
Mehdi Madani	mma@n-side.com +32 495 31 6505
Yves Langer	yla@n-side.com +32 476 60 8181

Client

NESO

Date

2 May 2024

Legal notices

Table of Contents

Executive Summary

This report explores a set of market design questions related to the topic of co-optimization of energy and ancillary services. It complements FTI's analysis for the ESO, providing background information on technical and economic aspects of different design options and qualitatively assessing their potential benefits and drawbacks. The report also discusses key implementation considerations, which would need to be further assessed should co-optimization of energy and ancillary services be taken forward.

Scope of co-optimization

Co-optimization is a broad term, and generally refers to the simultaneous allocation of multiple products and services, while jointly accounting for their interdependencies in an integrated fashion. The specific products and services that we cover in this report are energy, transmission network capacity, and ancillary services. The term ancillary services is also fairly broad. This document is centered on capacity procurement for reserve and response services, which we refer to as "balancing capacity". Our assessment focusses on day-ahead and real-time markets.

Choices related to fundamental elements of wholesale and balancing market design influence materially the feasibility and the effectiveness of co-optimization of energy, ancillary services and transmission (and there are stark differences in the European design relative to other international examples, such as Australia or the US). The most relevant market design choices relate to:

- *How the different products are priced;*
- *What the bidding language is;*

- *How the transmission network is considered in price formation;*
- *Whether there is a real-time market for balancing capacity and how it interacts with the energy market.*

For this reason, chapters 2 and 3 provide an extensive introduction to non-convex problems[1], multi-product auctions, pricing rules, and how US and European market designs have evolved to meet these challenges. Chapter 4 then dives deeper into aspects that relate to considering the transmission network capacity in the co-optimization process, while chapter 5 focuses on real-time co-optimization of energy and balancing capacity.

Benefits of co-optimization - theoretical basis

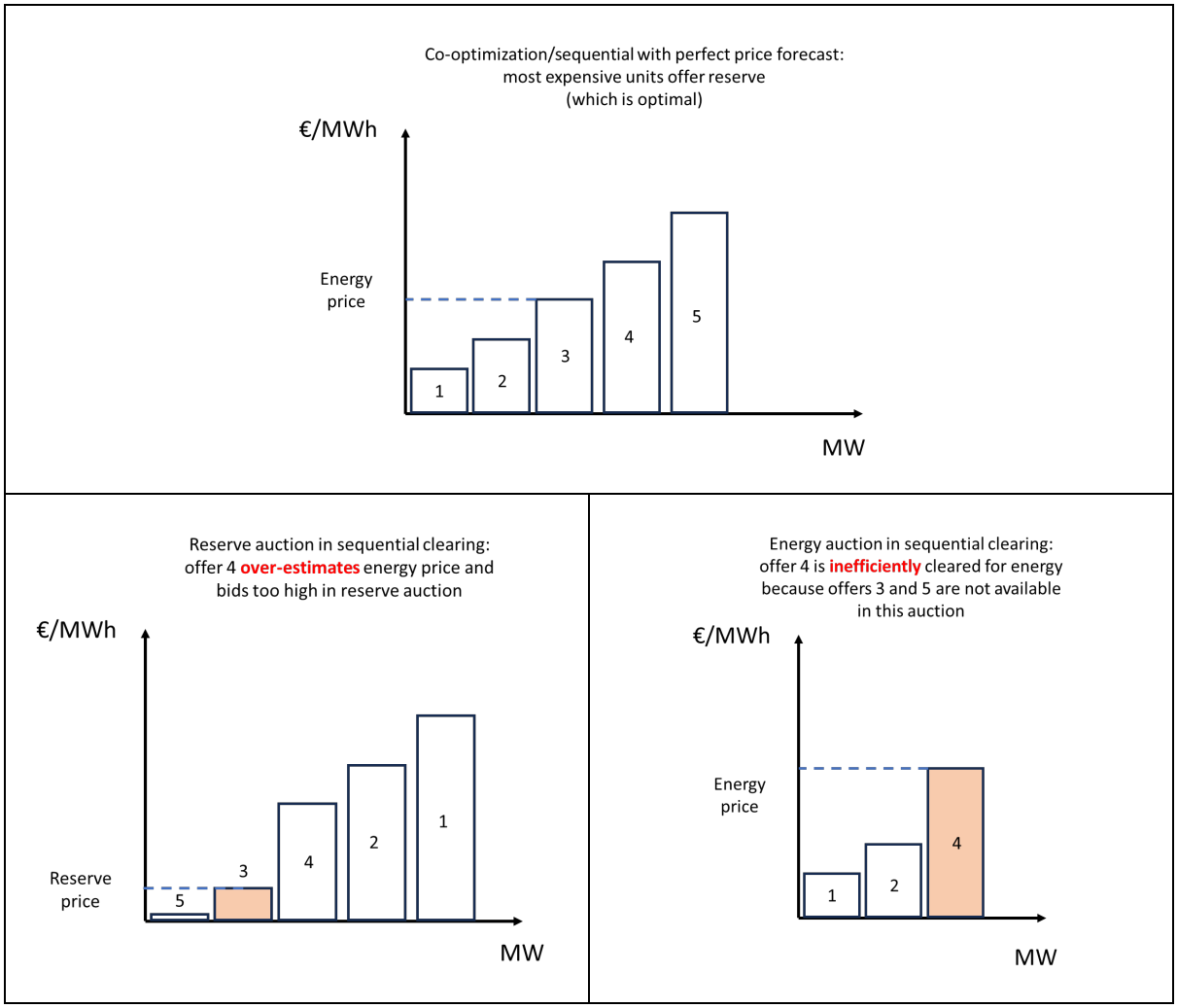
Energy and ancillary services are interdependent because they are mutually exclusive. Indeed, booking a certain amount of generation capacity for ancillary services means that this same capacity cannot be used to cover anticipated energy needs, and vice versa. In such settings with interdependencies among traded products, the separate auctioning of interdependent goods is possible, but it puts a great burden of anticipation of prices on market participants.

Where energy and ancillary services markets are operated separately, market participants are required to choose ex ante in which market(s) to enter. To do so, they consider the opportunity cost of forgoing revenues in the other markets. Opportunity cost estimations involve information imperfections, leading to significant forecast errors. This results in a reduction of allocative efficiency and distorts price signals which affects the clearing price of the whole market.

[1] Non-convexity is a mathematical property that corresponds to “jumps” in operating constraints or costs. For instance, the fact that a unit is on or off is a non-convexity. This notion is revisited repeatedly throughout the report.

Co-optimization recognizes the interdependencies between energy and ancillary services and seeks to optimize their provision simultaneously. It captures the synergies between different products and services, with the aim of facilitating more efficient and cost-effective operation of the system.

In practice, market participants enter linked, mutually exclusive bids into both the energy market and the ancillary services markets and a single clearing process allocates resources where they are most valuable. The intended outcome is to eliminate the risk associated with opportunity cost forecasts, reduce total system costs and ensure efficient price signals. The risk of inefficient allocation of resources between energy and ancillary services is illustrated diagrammatically in the following figure.



Relevance of co-optimization in the context of the evolving energy system in Great Britain and REMA

Integrating increasing amounts of renewable energy, including up to 50GW of offshore wind[2] by 2030, will pose challenges for managing the electricity system. Greater flexibility across the entire energy system will be required to cope with variable supply and to reach Net Zero at a reasonable cost for consumers. Well-designed markets creating efficient investment and dispatch signals (where and when to produce and use electricity) will be key for an economical and secure energy transition.

In this context, the Department of Energy Security and Net Zero (DESNZ) is undertaking a Review of the Electricity Market Arrangements (REMA) to explore what changes to the GB market arrangements are needed to deliver a cost-effective transition to the future larger, cleaner and more decentralized electricity system.

REMA acknowledges that ensuring system operability is crucial for the efficient and safe functioning of the electricity system and anticipates that the need for ancillary services is likely to grow in response to a greater proportion of variable renewables on the system along with changing patterns of demand[3]. One of the key options assessed in REMA for ensuring efficient operability is co-optimization of ancillary services with the wholesale market. This option is considered as part of broader wholesale market changes which could involve central dispatch (dispatch controlled by the System Operator), and possibly some level of network constraint consideration in wholesale electricity price formation (e.g., locational pricing).

[2] BEIS (2022). British Energy Security Strategy.

[3] BEIS (2022). Review of Electricity Market Arrangements

As the transition to Net Zero progresses, the case for implementation of co-optimization is likely to be enhanced. This is because the scale and impact of opportunity cost forecast errors is likely to grow further, as a result of:

- An acceleration of variable generation deployment, contributing to more volatile - and therefore more difficult to predict – prices;
- An increasing need for ancillary services to manage growing system operability challenges, due to an energy mix with higher share of non-dispatchable and non-synchronous resources;
- The presence of numerous interdependent ancillary services (with different market clearing intervals);
- Growing complexity due to a more interconnected and decentralized system.

Economic Foundations and Pricing

The trading of energy, access to transmission networks and ancillary services in electricity markets is a complex process, with diverse designs that are encountered worldwide. These design choices reflect objectives that can at times be in tension with each other, including transparency, incentive compatibility, economic efficiency, system security, and a host of other desirable objectives. Despite the diversity of design choices that are encountered in international electricity markets, closed-gate auctions based on marginal pricing principles are commonly used for trading electricity.

The maximization of economic welfare is the principal objective of these electricity market auctions. This is justified by a fundamental result in economics, which states that in well-behaved markets (which are technically defined as convex markets with perfect competition), the socially optimal allocation of resources is also consistent with the selfish profit maximization goals of agents. This means that, if agents bid truthfully

into these auctions, then the auctioneer should be able to find a price signal that induces them to respond in a way that is simultaneously in their own selfish best interest while also being in the best interest of the entire market.

The premise stated above about the existence of a clearing price that can coordinate agents in ideal settings (under perfect competition and convexity) actually generalizes to multi-product auctions where one trades not only energy but also interdependent ancillary services. This is reassuring, because it offers a roadmap on how to set up markets which trade energy and ancillary services simultaneously. However, real-world electricity markets contain non-convexities. Non-convexity can be intuitively translated as bulkiness in system operation: on-off decisions of firing up generation units, fixed costs related to starting them up, fixed costs related to keeping them online are a few examples of non-convexity. In the presence of such constraints of the “all or nothing” type, an economic equilibrium (i.e. a solution that all participants are perfectly happy with) may not exist.

In order to cope with this challenge, pricing approaches in pay-as-clear markets can be divided into two categories: those that employ side payments (as typically the case in the US markets) and those that do not (as typically the case in European markets), each with their relative merits and disadvantages. Defining suitable “pricing rules” requires careful consideration, as it inevitably implies some trade-offs between procurement costs in the short, medium and long-term, overall economic efficiency, bidding behavior (e.g. gaming risks) and computational tractability.

Bidding Product Design

Bidding product design is connected with key market design decisions around unit or portfolio bidding, as well as decisions on central or self-scheduling and dispatch. Choices on bidding language affect the level of flexibility that market participants have in communicating their capabilities to the market, but also the computational

complexity of the resulting market clearing models, including the technical feasibility of co-optimization.

There are two main approaches to bidding product design. We refer to them as the "European approach" and "US approach".

In the European approach, standardized products (simple bids) that are not asset-specific are used. This facilitates the aggregation of individual assets into portfolios. Such an approach provides high flexibility to traders (despite adverse effects on market monitoring), especially those with large and diverse portfolios, enabling them to quickly adapt to changing circumstances and choose (at the time of delivery) which assets of their portfolio are most efficient to use.

Moreover, according to certain researchers[4], portfolios with simple bids better enable the participation of new technologies and decentralized resources compared to the more prescriptive language of unit bidding with complex bids, where a central clearing algorithm and associated bidding information needs to be updated to account for new types of resources or upgraded to deal with a significant increase in the number of participating resources.

The European business rules for pricing do not involve side payments, which reduces concerns related to discriminatory settlements. However, they are intrinsically more challenging from a computational standpoint, which may limit the scope for detailed bidding products.

In theory, co-optimization of energy and reserves with simple bids and portfolio bidding is achievable. However, given that there is no such precedent internationally, R&D would be required to overcome technical limitations.

[4] Ahlqvist V., Holmberg P., Tangerås T., (2022). A survey comparing centralized and decentralized electricity markets. Energy Strategy Reviews, Volume 40.

In the US approach, so-called “multi-part” bids are used, which feature specific assets’ details to facilitate co-optimization of energy and ancillary services under a central dispatch unit commitment model. The design is supplemented by a series of rules to invite market agents to adequately follow the unit-based central dispatch paradigm. Co-optimization in a design that features multi-part bids is eased, as it is straightforward to model the linkages between energy and ancillary services provision (e.g., substitutability). Such an approach is in principle efficient thanks to high bid expressiveness.

However, as it is a more prescriptive bidding language, it can be seen as inflexible to the requirements of new technologies, hindering innovation. It may also be considered as potentially distortive in terms of settlement because of its use of side payments. Market monitoring is eased with unit bidding and multi-part bids, as a more granular view on the different cost components and operational constraints of the participating units (compared to portfolios with standardized products) can be obtained.

Both the European and US approach have implications in terms of the computational complexity of the underlying market clearing algorithm. The computational complexity in the US model relates to the large number of resources (unit bidding) and detailed network representation (nodal pricing) while the computational complexity in European markets results from the simultaneous search for bid matchings and compatible prices (price-based conditions that must be enforced). Including ancillary services adds non-trivial computational complexity in both paradigms. The integration of more resources (smaller units) and more services introduces increasing computational challenges on market clearing algorithms under both the European and US market design.

The fact that unit-based systems can account for precise network models enables locational pricing. There is a variety of benefits that result from locational pricing, including the lifting of INC-DEC gaming opportunities, better locational investment signals (especially for co-locating future renewable capacity closer to load centers),

more secure operation near real time, better utilization of available network capacity, more efficient day-ahead commitment decisions, and easier cross-border sharing of flexible resources near real time.

Locational Considerations

The introduction of transmission constraints in the energy market through locational pricing is contemplated in REMA. Co-optimization may be introduced under a locational (zonal or nodal) pricing regime, irrespective of whether the grid model applicable to energy and to balancing capacity is the same or not. When network constraints are considered in price formation, co-optimization allocates transmission capacity to the products (energy, reserve and response) that create the most value (i.e. maximize social welfare). In this process, it is critical that the transmission system operator performs efficient dimensioning of balancing capacity, in order not to restrict opportunities for trade in the energy market unnecessarily.

Allocating transmission capacity across energy and balancing capacity implies additional conceptual and computational challenges, compared to the allocation of transmission capacity for the trading of energy alone. This is because the actual activation patterns of balancing capacity is unknown at the time when the balancing capacity is auctioned off. In other words, the auction results must guarantee that the balancing capacity that is procured across transmission constraints can be delivered, irrespective of how this capacity is activated in real time. Technical solutions have been developed to ensure that a strict “reserve deliverability requirement” (so called “deterministic requirement”) is respected.

Alternatively, the “reserve deliverability” (i.e. the challenge of trading balancing capacity on a network with transmission constraints while ensuring that the network will be able to support the activations of energy from this balancing capacity in real time) may not be strictly applied, and may for example either rely on statistical

approaches (e.g. it is sufficiently probable that the procured ancillary services will be deliverable) or rely on contingency constraints (i.e. focus more specifically on the deliverability of balancing energy in case of significant contingencies).

Real-Time Co-optimization of Energy and Reserve through Reserve Scarcity Pricing

Scarcity pricing is informally defined as the process by which short-term energy prices escalate above the marginal cost of the marginal unit, i.e. the last unit in the merit order to produce power in the market. Scarcity pricing can typically occur under stressed system conditions. These prices are valuable for ensuring a long-term equilibrium in an energy market, since they allow generators to recover inframarginal rents and thus attract investment.

The Balancing Mechanism (BM) can be seen as the GB equivalent of a US-style real-time market. A key difference between the two approaches is that in GB there is no real-time market for balancing capacity, while US-style markets feature real-time co-optimization of balancing capacity and energy. The paradigm in US market operations is to acknowledge that headroom is valuable in real time, and to pay for it. For this reason, US markets conduct real-time markets as multi-product auctions for energy, transmission and reserves.

Real-time co-optimization of energy and balancing capacity in US-style markets is often achieved through Reserve Scarcity Pricing based on Operating Reserve Demand Curves (ORDCs). The technical complexity of implementing reserve scarcity pricing (telemetry, computation of ORDC adders) is limited and there are precedents of detailed implementations, e.g. in ERCOT. Reserve Scarcity Pricing based on ORDCs can be particularly relevant for systems with increased penetration of low marginal cost renewable energy resources, which exert a downward trend on wholesale market prices, possibly creating a “missing money” problem for investment

in flexible and dispatchable capacity. A disciplined implementation of scarcity pricing also provides incentives for participation in Balancing Markets.

Reserve Scarcity Pricing based on ORDCs can co-exist with, and does not replace, explicit co-optimization of energy and ancillary services in the day-ahead timeframe, which focuses primarily on elimination of opportunity cost forecast errors and allocative efficiency. Although the mechanism can be implemented in the day-ahead market, the true value of scarcity pricing through ORDC is in implementing it in the real-time balancing market.

Reserve scarcity pricing can and does co-exist with Capacity Remuneration Mechanisms (CRMs). The general effect of scarcity pricing is to reduce the scope of a CRM. This is because reserve scarcity pricing tends to suppress missing money. This is a healthy effect: if the energy market can take care of missing money, there is no double-payment for investment costs through capacity markets.

Considering the GB context - already involving a high share of low marginal cost renewable energy resources, which is expected to increase materially as the country transitions to a Net Zero power system by 2035 - there could be merit in assessing the case for introducing Reserve Scarcity Pricing based on ORDCs.

The GB market appears to be missing a real-time market for reserve capacity. This is common in European markets, and without precedent in US designs. It corresponds to a missing market, and the ultimate effect is that it interferes with price formation in the day-ahead reserve market. The introduction of a real-time market for reserve capacity can be achieved through the implementation of co-optimization of real-time energy and reserves, however this is not necessary. An alternative is to implement a real-time market for reserve capacity through an implicit approximation of co-optimization.

When considering the future evolution of the GB market, an important market design question is whether there is a desire to generate scarcity prices as a result of an ORDC, or as a result of internalizing inframarginal rents in balancing price bids. The former option can be complemented with an ex-ante mitigation of bids that are clearly above marginal cost. Given how the scarcity pricing mechanism is designed, this still allows balancing prices to rise above the marginal cost of the marginal unit. On the other hand, the latter (internalizing rents in bids) presents the market monitor with a difficult dilemma: during periods of scarcity, it becomes unclear whether these inframarginal rents result in a healthy recovery of fixed investments costs, or an exercise of market power.

1. Introduction

1.1 Takeaways

→ Currently in GB, energy and ancillary services markets are operated separately, meaning market participants are required to choose ex ante in which market(s) to enter. To do so, they consider the opportunity cost of forgoing revenues in the other markets. Opportunity cost estimations involve information imperfections, leading to significant forecast errors. This results in a reduction of allocative efficiency and distorts long-term investment signals.

→ Co-optimization recognizes the interdependencies between energy and ancillary services, and seeks to optimize their provision simultaneously. It aims to capture the synergies between these components, leading to more efficient and cost-effective operation of the system.

→ The transition to a Net Zero system will exacerbate the scale and impact of opportunity cost forecast errors, as it will entail an acceleration of variable generation deployment, including an ambitious target of up to 50 GW of offshore wind installed by 2030. Well-designed markets creating efficient investment and dispatch signals will be key for an economical and secure energy transition. Without co-optimization, the scale and impact of allocative inefficiency along with the distortion of investment signals (as a result of sub-optimal short-term wholesale energy and ancillary services prices) will grow as the energy transition progresses.

1.2 Scope and structure of this report

This report explores a set of market design questions related to the topic of co-optimization of energy and ancillary services. It provides background information on technical and economic aspects of different design options and qualitatively assesses their potential benefits and drawbacks. This document evaluates different models in "steady state". The detailed implications of departing from the current market design (e.g. governance, regulatory, implementation costs, etc.) require a more detailed investigation which is dependent on the precise design that would be favored.

Chapter 1 describes the evolving energy landscape in GB, presents the concept of co-optimization and its key drivers. It then sets out the theoretical basis on which co-optimized clearing delivers superior economic efficiency compared to sequential clearing.

Chapter 2 outlines a set of foundational notions in economics, such as market welfare, market equilibrium, equilibrium prices and the associated mathematical programming formulations that make these definitions precise. It then generalizes this discussion to auctions with not only one product but instead multiple products that are auctioned off simultaneously. The chapter is completed with a discussion on how to cope with products that are not "well-behaved", in particular products that feature a take-it-or-leave-it attribute.

Bidding products matter in the context of co-optimization, because co-optimization introduces the need of generalizing existing energy-only market models (by auctioning additional reserve products, the acceptance rules of which interact with the acceptance rules of energy products), and the choice of bidding product definitions affects both the ability of market participants to communicate their flexibility to the market, as well as the computational complexity of the resulting market clearing models.

Chapter 3 focuses on two major features of bidding product design, that of unit versus portfolio bidding, and that of multi-part bids versus simple bids. The two aspects are interrelated. Throughout the chapter we compare them as they are represented in the European versus US market design. Specifically, US market design is founded on unit-based bidding with multi-part bids. Instead, the vast majority of European markets are based on portfolios represented in the market through products that aim at being simpler.

Chapter 4 provides an overview of the most important design elements to take into account when considering the co-optimized procurement of energy and ancillary services in the presence of thermal transmission network constraints. The focus is on internal GB constraints; cross-border transmission capacity allocation is not within the scope of this report. It then discusses how it can be ensured that any pattern of reserve activation is “feasible” in terms of congestion in real time and what are the implications of different approaches in terms of allocative efficiency and overall procurement costs (including redispatch cost considerations).

Finally, chapter 5 outlines the general idea of scarcity pricing as an implicit form of co-optimization and how it compares to explicit co-optimization. It presents the implications of scarcity pricing for balancing market design, and specifically assesses various alternatives proposed for implementing scarcity pricing in European balancing markets with a focus on documenting their relative merits and weaknesses. Various parameters pertaining to the calibration of Operating Reserve Demand Curves (ORDCs) that influence the magnitude and frequency of scarcity signals are discussed. The interaction of scarcity pricing with capacity mechanisms is then taken on. The chapter is concluded with a comparison of alternative design options in view of the international experience with the implementation of scarcity pricing.

1.3 The evolving energy landscape in GB

The UK has achieved remarkable progress towards its power sector decarbonization targets, reducing emissions by 68% since 2010. It has moved away from coal generation, while delivering a fivefold increase of renewables, making them its leading source of energy generation[5]. However, meeting its commitment to deliver a fully decarbonized power sector by 2035 will require an even faster scale-up of low carbon technologies. To this end, in the British Energy Security Strategy, the government set out its ambition for up to 50GW of offshore wind by 2030. It also anticipates a fivefold increase in the deployment of solar by 2035[6].

The decarbonisation of the electricity system is leading to changes in three key areas[7]:

- *More variable resources:* this change refers to the increase in weather-dependent technologies, such as solar and wind, replacing traditional thermal generators (e.g. coal and gas units) that provide "firm" power. This increases the need for being able to balance the system in the face of uncertainty.
- *More asynchronous resources:* Historically, the power system relied on the inertia inherent in large and centralized generation plants to keep it stable. As synchronous resources, like coal and gas generators, are replaced by inverter-based resources (e.g. wind, solar, HVDC interconnectors), inertia levels fall and alternative means of stability are required[8].

[5] DESNZ (2023). Digest of Energy Statistics.

[6] BEIS (2022). British Energy Security Strategy.

[7] ESO (2022). Operability Strategy Report.

[8] For example, see the ESO's NOA Stability Pathfinder: https://www.nationalgrideso.com/industry-information/balancing-services/pathfinders/noa-stability-pathfinder

- *More dispersed resources:* It includes new generation locating at network extremities and further away from demand centers such as offshore wind in Northern Scotland and in South West England. It also refers to the increase in generation, storage and demand-side response on the distribution networks, resulting in bi-directional flows.

Integrating safely the increasing amount of renewable energy will pose great challenges for managing the electricity system. Greater flexibility across the entire energy system will be required to cope with variable supply and reach Net Zero at a reasonable cost for consumers. Investing in flexibility has the potential to deliver material net savings of up to £16.7bn per annum across all scenarios analyzed in 2050, according to a report commissioned by the Carbon Trust[9]. Well-designed markets creating efficient investment and dispatch signals (where and when to produce and use electricity) will be key for an economical and secure energy transition.

In this context, the Department of Energy Security and Net Zero (DESNZ) is undertaking a Review of the Electricity Market Arrangements (REMA) to explore what changes are needed to the GB market arrangements to deliver a cost-effective transition to the future larger, cleaner and more decentralized electricity system.

REMA acknowledges that ensuring system operability is crucial for the efficient and safe functioning of the electricity system. It also anticipates that the need for ancillary services is likely to grow in response to a greater proportion of variable renewables and to changing patterns of demand[10]. One of the key options assessed in REMA for ensuring efficient operability is co-optimization of ancillary services with the wholesale market. This option is considered as part of broader wholesale market changes which involve central dispatch (dispatch controlled by the System Operator), and possibly

[9] Carbon Trust (2021). Flexibility in Great Britain.

[10] BEIS (2022). Review of Electricity Market Arrangements.

some level of network constraints consideration in wholesale electricity price formation (e.g., locational pricing).

1.4 Scope of co-optimization

The term co-optimization in electricity markets covers a fairly broad scope. Our focus in this report will be directed towards the co-optimization of the three major products and services that are traded in wholesale electricity markets, namely energy, balancing capacity and transmission capacity. Moreover, our investigation will be centered on both real-time as well as day-ahead markets. In this section we explain in further details the meaning of each of the terms “energy”, “balancing capacity” and “transmission capacity” in the context of the GB market.

Energy is traded across different timescales in a self-dispatch market, from years ahead up to real time. It is traded in the form of physical over-the-counter transactions or via power exchanges, and is thus not centrally managed by the System Operator. In central dispatch markets it is typically traded at day-ahead and physically within-day. In this report, with the term “energy” we refer to electricity traded at the day-ahead stage and/or close to real time through closed-gate auctions.

The term “ancillary services” is commonly used to describe a set of services procured and activated by the System Operator to ensure that electricity can reach the end customer when and where it is required, in a safe manner, within acceptable quality standards. These include frequency response, reserve, stability, voltage and restoration, each of them meeting a different system need as presented in Table 1.

Table 1: Ancillary Services and System Needs

Ancillary Service	System Need
Frequency Response	Maintain frequency of the network at (or very close to) 50Hz.
Reserve	Accommodate unforeseen changes in demand or generation. Reserve power is not delivered as rapidly as frequency response but sees a greater volume of electricity delivered.
Stability	Ensure sufficient levels of inertia and robustness to withstand disturbances.
Voltage	Keep voltage within operational limits by modifying reactive power intake and injection. Indeed, adding reactive power or absorbing it has the effect of either increasing or decreasing voltage on the system. The GB network runs at 400, 275 and 132 kilovolts (kV) and must stay close to these figures at all times.
Restoration	Re-energize the network in the event of a partial or total grid shut down.

Of the different services listed above in Table 1, this report focuses on reserve and response. There could theoretically be value in co-optimizing the procurement of stability and voltage services with energy. However, considering the nature of these services (e.g., highly locational in the case of voltage), achieving such a goal would be technically challenging from a computational perspective (breaking new ground internationally) and may not be optimal from an economic one[11].

The GB market features a range of frequency response and reserve services, listed in Table 2.

[11] The case for the co-optimized procurement of frequency regulation services and energy, while considering changes in system inertia, is assessed by Imperial College in the COEF NIA project. Information about the project can be found on the ENA website.

Table 2: Frequency Response and Reserve Services in GB

Requirement	Services
Frequency Response	Dynamic Containment; Dynamic Moderation; Dynamic Regulation
	Mandatory Frequency Response; Dynamic Firm Frequency Response*
	Static Firm Frequency Response*; Static Recovery**
Reserve	Fast Reserve*; Quick Reserve**
	Short-Term Operating Reserve*; Slow Reserve**
	Bid/Offer Acceptances in the Balancing Mechanism; Balancing Reserve**
	Demand Flexibility Service
	Other Bespoke and Legacy Services[12]

** Phasing Out*

*** Under Development*

A subset of the frequency response and reserve services is bought "firm", meaning that providers receive (or pay) an availability payment for being available to deliver the service(s) during the contracted delivery window(s). Availability prices for most of the frequency response and reserve services are determined through pay-as-clear auctions held at the day-ahead stage. To balance the system in real time, the ESO can either take actions via the Balancing Mechanism or activate the pre-procured services[13].

[12] For example: Super SEL, Max Gen, Spin Gen. Thermal constraint management can sometimes interact with ancillary services – for example reserve capacity can be utilized by ESO to help manage constraints in real time.

[13] Frequency Response Services are activated automatically in response to frequency disturbances.

The terms:

- “frequency response and reserve capacity”
- “balancing capacity”
- “reserve(s)”
- “ancillary services”

are used interchangeably in this report when referring to co-optimization of energy and ancillary services. Despite the fact that there are important detailed distinctions in terms of how reserve versus response is treated in international electricity market designs, we only deep-dive on these detailed aspects in the relevant chapters of the report, in order not to distract the reader from the big picture.

Transmission, in the context of our analysis, carries the standard meaning that is encountered in all international markets. Although there are variations in the degree of details used for representing networks in international market design, and the extent to which these network models are faithful to the true underlying physics of electrical power flow, the network abstraction carries the same meaning in all markets.

Specifically, all market models consider a list of critical network elements. Then, what is traded in markets is access to the capacity of these network elements. The trade of energy and balancing capacity occupies space in these network elements. Therefore, the trade of energy or balancing capacity between different locations requires the procurement of access rights on these network elements. Ultimately, the combination of a price on energy at an asset’s present location and the price paid for the right to access the network in order to trade with different locations implies a market

equilibrium where energy and reserves at different locations trade at different prices per standard economic theory[14].

Having defined the three major products and services that our analysis is concerned with, it is also opportune to clarify the precise meaning of co-optimization in our investigation. In the context of general optimization theory, co-optimization refers to the act of jointly optimizing interdependent processes. In the specific case of electricity markets, it refers to the joint optimization of energy, balancing capacity (in the broad sense of including reserve and response), and transmission since these products and services all interact with each other and cannot be considered in isolation. Furthermore, an additional qualifier of which stage is taken into account in market operation (e.g. day-ahead markets or real-time markets) is also important for the discussion. We analyze pairwise interactions of these products and services in particular market timeframes in specific chapters of the report. For instance, in certain parts, we focus on the interaction of energy and reserve in real time (e.g. chapter 5), the interaction of reserve products between each other in day-ahead or earlier forward markets[15] (e.g. section 3.3 and 3.4), the interaction of energy, transmission and reserve in day-ahead markets (section 2.4), the interaction of transmission and reserve in forward markets (chapter 4), and so on. We do not discuss all feasible combinations of pairwise interactions in all possible market timeframes, but rather isolate specific aspects in which we believe the ESO can benefit from proven international experience. The structure of the analysis is clarified in Table 3.

[14] Samuelson, Paul A. "Spatial price equilibrium and linear programming." *The American economic review* 42.3 (1952): 283-303.

[15] Note that reference to “forward markets” in this report includes the day-ahead market. From a standpoint of economic theory, the spot market is the real-time market, therefore any market that precedes it and is indexed against this market, including the day-ahead market, is a forward market.

Table 3: A tabular description of the interacting products and services analyzed in the report in the context of co-optimization, as well as the corresponding market timeframes. An X indicates that the corresponding product/service/timeframe is relevant to a specific chapter of the report.

	Products/services			Time scales	
	Energy	Transmission	Balancing capacity	Day-ahead	Real-time
Pricing (Chapter 2)	X	X	X	X	
Bidding product design (Chapter 3)	X		X	X	
Locational considerations (Chapter 4)		X	X	X	
Reserve scarcity pricing (Chapter 5)	X		X		X

Note that specific GB products are not listed explicitly as columns in Table 3, but can rather be attributed to specific combinations of columns. In particular, “reserve” in GB terminology can be understood as manual frequency restoration reserve in European terminology. Thus, its reservation falls under the category of “balancing capacity” and “day-ahead” as far as its commitment is concerned in the day-ahead market, and under the category of “energy” and “real-time” as far as its activation is concerned. It is important to underline that selling 1 MW of this service in the day-ahead market is accompanied by an obligation to bid at least 1 MW of this capacity in the real-time balancing energy auction. Note that chapter 5 argues that there should also be a “real-time balancing capacity” dimension to this product as well, and develops the precise reasoning for this argument.

On the other hand, “response” in GB terminology can be understood as automatic frequency restoration or frequency containment reserve in European terminology, or

automatic generation control in US terminology. This type of service often has a capacity component only (at least as far as frequency containment reserve is concerned). This should be understood in the sense that the corresponding capacity is booked in day-ahead or earlier forward markets, and the reservation is honored in real-time economic dispatch models. In other words, these are slices of capacity that are reserved for automatic control actions related to very short time scales, without activation in these capacity slices being auctioned off or traded in real-time energy activation auctions (with some exceptions, e.g. aFRR in Europe), and without this capacity being released in real time for access by other services (e.g. manual frequency restoration). Thus, in the nomenclature of the table, these services would fall under the “day-ahead” and “balancing capacity” columns.

1.5 The theoretical case for co-optimization and key drivers

The resources that can be used to deliver ancillary services are, to a large extent, the same as the resources needed for the procurement of energy in the wholesale energy market. An optimal allocation of these same resources to the competing markets requires in principle to “co-optimize” the procurement; that is to jointly consider the value of the resources for both types of procurement in order to allocate them where they are the most valuable, while satisfying all applicable operational or economic constraints.

Co-optimization recognizes the interdependencies between energy and ancillary services, and seeks to optimize their provision simultaneously. It aims to capture the synergies between these components, leading to more efficient and cost-effective operation of the power grid.

In practice, market participants enter linked mutually exclusive bids into both the energy market and the ancillary services markets. Then, a single clearing process

allocates resources where they are the most valuable. This eliminates the risk associated with opportunity cost forecasts, reducing total system costs and ensuring efficient long-term price signals.

Currently in GB, energy and ancillary services markets are operated separately, meaning that market participants are required to choose ex ante in which market(s) to enter. To do so, they consider the opportunity cost of forgoing revenues in the other markets. Opportunity cost estimations involve information imperfections, leading to significant forecast errors. This results in a reduction of allocative efficiency and distorts long-term investment signals.

A combination of factors observed in the GB market drives opportunity cost forecast errors, including:

- High penetration of variable renewables, contributing to more volatile - and therefore more difficult to predict – prices;
- Increasing need for ancillary services to manage growing system operability challenges, as a result of an energy mix with higher share of non-dispatchable and non-synchronous resources;
- Numerous interdependent ancillary services (with different market clearing intervals);
- Growing complexity as a result of a more interconnected and more decentralized system.

The transition to a Net Zero system will exacerbate the scale and impact of opportunity cost forecast errors, as it will entail an acceleration of variable generation deployment, including GB's target of up to 50 GW of offshore wind installed by 2030. Therefore, in absence of co-optimization, allocative efficiency and long-term investment signals could be compromised.

2. Economic Foundations & Pricing

2.1 Takeaways

→ The maximization of economic welfare is the principal objective of electricity market auctions. In markets with perfect competition, the socially optimal allocation of resources is also consistent with the selfish profit maximization goals of agents.

→ In the presence of constraints of the type “all or nothing” an economic equilibrium (i.e. a solution that all participants are perfectly happy with) may not exist.

→ Pricing approaches in pay-as-clear markets can be divided into two categories: those that employ side payments (as is typically the case in the US markets) and those that do not (as is typically the case in European markets), each with their relative merits and disadvantages.

→ Defining suitable “pricing rules” requires careful consideration, as it inevitably implies some trade-offs between (1) procurement costs, in the short, medium and long-term (2) overall economic efficiency, (3) bidding behavior (e.g. gaming risks) and (4) computational tractability.

→ The separate auctioning of interdependent goods is possible, but it puts a great burden of price estimation on traders, and is therefore likely to lead to inefficiencies due to inaccurate estimates of opportunity costs. This principle applies to the separation in the trading of energy, transmission capacity, and reserve in electricity markets, and is a reason why many electricity markets trade these three interdependent products

and services (energy, transmission access and reserve) through co-optimization.

→ Consistent market models across timescales are crucial for limiting gaming opportunities and ensuring the back-propagation of real-time products and services to forward financial markets. INC-DEC gaming is an example of market manipulation opportunities that emerge from using day-ahead market models that are inconsistent with real-time mechanisms.

2.2 Overview

This chapter analyzes foundational notions in economics. In section 2.3 we define market welfare, market equilibrium, equilibrium prices, and the associated mathematical programming formulations that make these definitions precise. Then, section 2.4 focuses on generalizing this discussion to auctions with multiple products that are auctioned off simultaneously. Finally, section 2.5 concentrates on coping with markets that include products considered as not "well-behaved", in particular market products that feature a take-it-or-leave-it attribute.

Note that the discussion throughout this chapter is focused on market models where the objective is to maximize welfare. An alternative objective, that has been analyzed in certain scientific publications, is the minimization of procurement cost[16]. Payment

[16] Hao, S., Angelidis, G. A., Singh, H., & Papalexopoulos, A. D. (1998). Consumer payment minimization in power pool auctions. IEEE Transactions on Power Systems, 13(3), 986-991.

cost minimization can be preferred from the point of view of system operators in the context of ancillary services markets, since system operators are the ones procuring them, often with regulatory incentives to reduce procurement cost. Such a design can distort truthful bidding incentives[17], can exacerbate efficiency losses in the case of strategic bidding behaviour[18], and typically leads to computationally hard market clearing models[19]. For all of the above reasons, we focus our discussion on market models that maximize welfare, and consider models that aim at procurement cost minimization as being out of scope.

Zhao, F., Luh, P. B., Yan, J. H., Stern, G. A., & Chang, S.-C. (2008). Payment cost minimization auction for deregulated electricity markets with transmission capacity constraints. IEEE Transactions on Power Systems, 23(2), 532-544.

Zhao, F., Luh, P. B., Yan, J. H., Stern, G. A., & Chang, S.-C. (2010). Bid cost minimization versus payment cost minimization: A game theoretic study of electricity auctions. IEEE Transactions on Power Systems, 25(1), 181-194.

Litvinov, E., Zhao, F., & Zheng, T. (2009). Alternative auction objectives and pricing schemes in short-term electricity markets. 2009 IEEE Power and Energy Society General Meeting: IEEE.

[17] Litvinov, E., Zhao, F., & Zheng, T. (2009). Alternative auction objectives and pricing schemes in short-term electricity markets. 2009 IEEE Power and Energy Society General Meeting: IEEE.

Zhao, F., Luh, P. B., Yan, J. H., Stern, G. A., & Chang, S.-C. (2010). Bid cost minimization versus payment cost minimization: A game theoretic study of electricity auctions. IEEE Transactions on Power Systems, 25(1), 181-194.

[18] Zhao, F., Luh, P. B., Yan, J. H., Stern, G. A., & Chang, S.-C. (2010). Bid cost minimization versus payment cost minimization: A game theoretic study of electricity auctions. IEEE Transactions on Power Systems, 25(1), 181-194.

[19] Luh, P. B., Blankson, W. E., Chen, Y., Yan, J. H., Stern, G. A., Chang, S.-C., & Zhao, F. (2006). Payment cost minimization auction for deregulated electricity markets using surrogate optimization. *IEEE Transactions on Power systems, 21*(2), 568-578.

Fernandez-Blanco, R., Arroyo, J. M., & Alguacil, N. (2011). A unified bilevel programming framework for price-based market clearing under marginal pricing. *IEEE Transactions on Power Systems, 27*(1), 517-525.

2.3 Key notions

The co-optimization of energy and ancillary services generalizes the standard single-product market clearing model by introducing several interacting products (i.e. energy, reserve and response). These products depend on each other due to the fact that they occupy the same constrained resource, i.e. the finite generation capacity of a given asset. Understanding the interaction between these products can be greatly facilitated by analyzing the single-product market clearing model first, formally defining the notions of welfare and competitive equilibrium, and relating these notions to the mathematical programming formulation of the basic single-product market clearing model.

The mathematical programming formulation has a useful “twin” representation in the space of so-called dual variables, which is the space where indicators of economic significance are defined (market prices, agent profits, and so on). Drawing the connections between the primal and dual formulation allows us to then generalize our analysis of the competitive market from a single product (that of energy) to multiple interacting products (in our case, that of energy, reserve and response).

We then introduce a dent in the storyline by considering the effect of non-standard products (such as block orders) to an otherwise well-behaved market model. We find that these non-convex products create difficulties of a fundamental nature. In particular, they might make it impossible to coordinate agents through a market price. We describe how this effect is handled in different market designs, and in particular the European market design of paradoxically rejected orders and the variety of US designs for dealing with the problem through side-payments. This also sets the stage for the discussion of transmission constraints that is developed in chapter 4.

2.3.1 Primal formulation

Theory

We focus on auction-based power markets settled under paid-as-cleared[20] principles, where a set of purchase orders and a set of supply orders are provided to a clearing algorithm. Each order is composed of a price/quantity pair. In the introduction to this chapter, we explained why the clearing algorithm maximizes the so-called "social welfare", which refers to the overall "well-being" or "prosperity" that is being generated in the market.

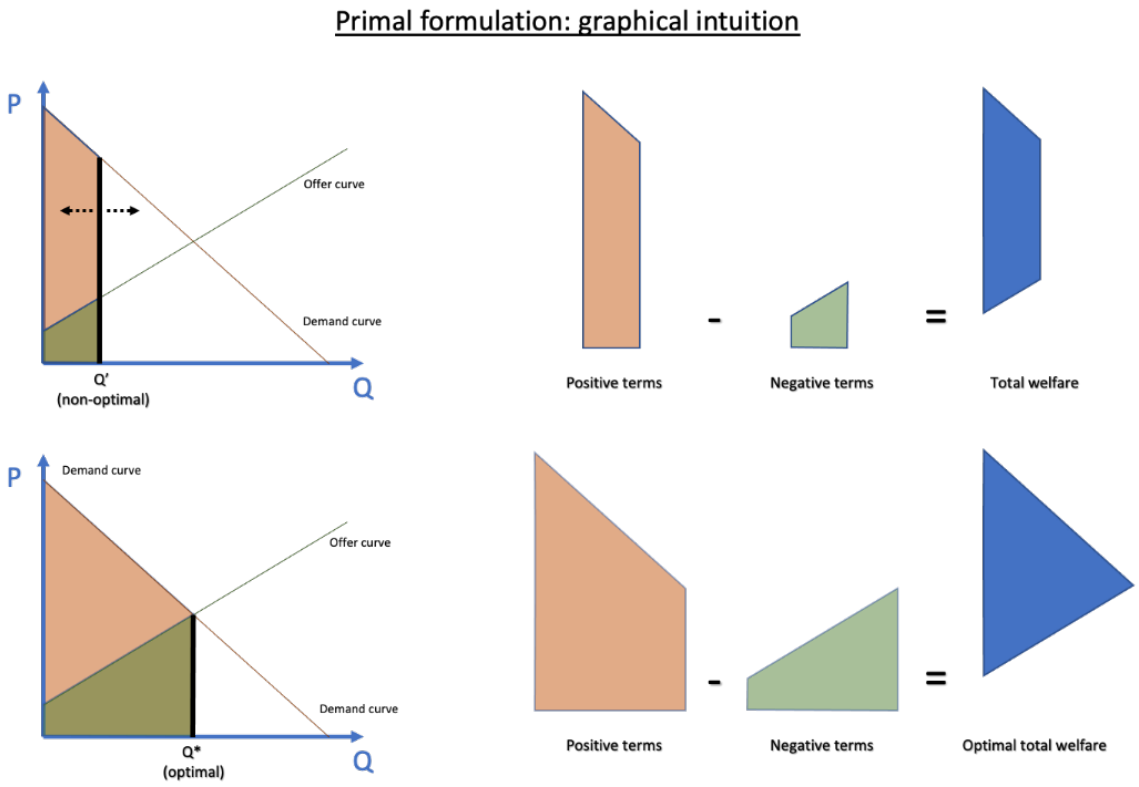
The welfare can be decomposed into the positive contributions of (accepted) buy orders and the negative contributions of (accepted) sell orders. Market clearing algorithms typically maximize such welfare and therewith determine the optimal volumes of products to be cleared. In optimization jargon, we name it the "primal formulation" of the welfare maximization problem. The primal formulation maximizes the social welfare as the "objective function" and uses exclusively quantities as variables. An important constraint of this problem is that the sum of all accepted buy volume of a particular product equals the sum of all accepted sell volume of this product.

Let us illustrate this notion graphically with a simplified example, shown in Figure 1, composed of an offer curve and a demand curve for a single product (e.g. energy). We assume that the offer and demand curves are continuous and monotonic (we will revisit this assumption later, when we discuss about "exotic" market products such as block orders). The primal formulation of the market clearing problem consists of finding the cleared volume Q* such that the area below the demand curve minus the area below the offer curve is maximal.

[20] Under a paid-as-cleared design, all trades for the same product are uniformly settled at an identical price, as opposed to non-uniform pricing schemes such as e.g. paid-as-bid.

Given the constraint that the accepted sell volume must equal the accepted buy volume, a welfare of zero is found when no volume is cleared. Starting from this solution, an increase of the cleared volume increases the objective function, i.e. welfare, as long as the demand curve is above the supply curve. This is because accepting a sell order that is priced below a purchase order generates positive welfare. The optimal solution (i.e. the solution that maximizes the objective function) is thus found at the intersection of the two curves. This solution is indeed optimal because it identifies all "profitable trades", i.e. it matches any supply that is priced below the demand, and which thereby generates value. Although there is no clearing price explicitly considered in this model, the optimal solution is such that there exists a clearing price that satisfies all cleared orders. This price is in fact the intersection of the supply and demand curves at the vertical axis (which is the axis of market prices).

Figure 1: A graphical intuition of the primal formulation is the maximization of the integral under the demand curve (orange) minus the integral under the supply curve (green) with the constraint that total demand equals total supply. We thus move a vertical line (representing the clearing volume) over the horizontal axis, and find that the largest difference between these two areas (blue) is to be found at the curves' intersection.



Technical Insights

To illustrate the concepts, we describe the primal formulation in mathematical programming terminology. In its simplest form, the primal formulation of the market clearing problem solves the following optimization model:

$$\max_{p,d} \sum_{l \in L} MB_l \cdot d_l - \sum_{g \in G} MC_g \cdot p_g$$

$$(\lambda): \sum_{l \in L} d_l - \sum_{g \in G} p_g = 0$$

$$(s_g): p_g \leq P_g, g \in G$$

$$(s_l): d_l \leq D_l, l \in L$$

$$p, d \geq 0$$

The optimization model above consists of the following “parts”:

a. The sets of market agents in the model: are the producers (indicated by the set G) and the consumers (indicated by the set L)

b. The decision variables to be optimized: are the amount of production matched from the set of sellers/producers (indicated by decision variables p) and the amount of demand matched from the set of buyers/loads (indicated by decision variables d). Decision variables of a mathematical program are usually listed under the maximization operator in the first line. The optimal quantities of these decision variables are what the platform decides, with the goal of maximizing the objective function of the market clearing model while respecting its constraints.

c. The parameters of the model: are the fixed values of technical and economic parameters which are submitted to the market clearing platform. These parameters are not decided by the platform but rather communicated by the

market participants. They include the price offers of sellers (denoted as marginal cost MC_g, in €/MWh, for supplier g), the price offers of buyers (denoted as marginal benefit MB_l, in €/MWh, for supplier l), the quantity offered of suppliers (denoted P_g, in MWh, for supplier g), and the quantity offered by buyers (denoted D_l, in MWh, for buyer l). The interpretation of these offers is as follows: when a seller is bidding a certain quantity at a certain price, it is asking to be paid at least this price in order to supply up to its bid quantity. When a buyer is bidding a certain quantity at a certain price, it is asking to pay at most the bid price for buying up to but no more than its bid quantity.

d. The objective function of the model: is what the platform aims to maximize. In our case, this is the market welfare defined as the difference between the benefit of consumers (first term in the objective function) and the cost of producers (second term in the objective function).

e. The constraints of the model: express the technical and economic requirements that must be satisfied by the solution of the market clearing platform. The first requirement is that the total demand must equal the total supply, which is given by the second line of the above mathematical program. The second requirement is to respect the boundaries that market participants have set on the quantity of the product that they are willing to trade. These are the third line (for producer quantity limits) and the fourth line (for consumer quantity limits). The last line of the model, corresponding to the last requirement, indicates that the traded quantities (both produced and consumed) must be non-negative.

Example

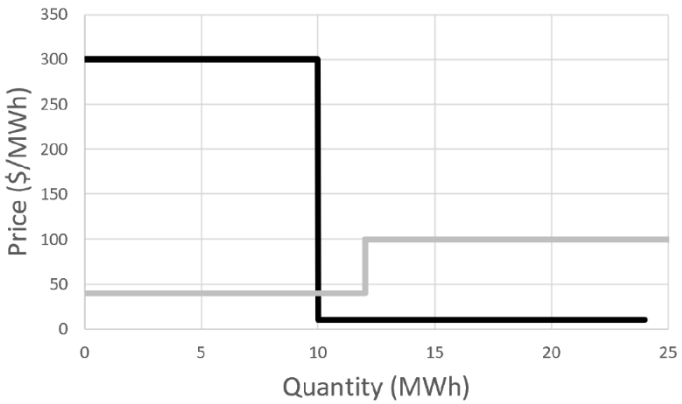
In order to better grasp the notions explained in this section, we provide the following example[21]:

Consider the order book that is represented in Table 4, and graphically in Figure 2. For the moment, the last column of the table, i.e. the minimum acceptance, is ignored. This feature will be introduced later for our discussion on non-convex market models.

Table 4: Order book corresponding to the running example of chapter 2. Source: (Papavasiliou A. , Optimization models in electricity markets, 2023).

Bid	Quantity (MWh)	Price ($/MWh)	Min. acceptance (MWh)
A (buy)	10	300	0
B (buy)	14	10	0
C (sell)	12	40	11
D (sell)	13	100	0

Figure 2: Graphical representation of the order book used in the running example of chapter 2. Source: (Papavasiliou A. , Optimization models in electricity markets, 2023).



The sets defined in this example are as follows:

- The set of buy orders is {A, B}.
- The set of sell orders is {C, D}.

The decisions being optimized are the matchings of sell orders (corresponding to variables p) and the matchings of buy orders (corresponding to variables d). The parameters of the model are the bid quantities and prices that are presented respectively in the second and third columns of Table 4. The constraints of the model are the requirements that orders must be matched at non-negative quantities that do not exceed the bid quantities of the different orders. The objective, which is economic welfare, is the total consumer benefit minus the total producer cost. The model can be solved graphically by finding the point of intersection of the inverse supply and inverse demand function of the system, as depicted in Figure 2. The inverse supply function, or marginal cost curve of the system, encodes the incremental cost at which we can source the q^{th} MWh of energy from the system. The inverse demand function, or marginal benefit function of the system, quantifies the incremental benefit of making the q^{th} MWh of energy available to the consumer population. The marginal cost curve is created by stacking supply orders in order of increasing marginal cost. The marginal benefit curve is created by stacking demand orders in order of decreasing marginal benefit. The graphical solution of matching orders optimally by intersecting the two curves encodes the "greedy" matching strategy whereby we continue matching sellers and buyers until the incremental benefit of matching orders fails to exceed the incremental cost of the matching. The most promising orders (from both the supply and demand side) are matched first, at the far left of the curve, whereas the least promising orders show up in the far right. The optimal matching for the case of our example occurs at 10 MWh. It corresponds to matching buy order A fully, matching 10

[21] Madani, M., Ruiz, C., Siddiqui, S., & Van Vyve, M. (2018). *Convex hull, IP and European electricity pricing in a european power exchanges setting with efficient computation of convex hull prices.* Baltimore, MD: arXiv.

Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

out of the 12 MWh of supply order C, and not matching any of the other orders that are available in the order book.

Economic welfare is the economic value generated by these transactions. It can be expressed equivalently as **consumer benefit** minus **producer cost**, or the sum of **producer surplus** and **consumer surplus**. We focus here on the first interpretation, since it corresponds to the primal formulation. In this case, consumer benefit is the benefit of buy order A, which is 10 MWh x 300 €/MWh = 3000 €, minus production cost, which is 10 MWh x 40 €/MWh = 400 €. The welfare thus amounts to 3000-400=2600 €.

Algorithmics

Before advancing to the discussion of the dual formulation, we remark briefly on algorithmics. The primal formulation presented in this section is a so-called linear program. Linear programs are optimization problems that feature an objective function and constraints that are only linear functions of the decision variables. This means that the objective function doubles whenever we double all decisions, and the constraints are either equalities or inequalities that are also linear functions of decisions. For this specific application of optimization programs to electricity market clearing, welfare is a linear function of decisions, because, if we double consumption, we double consumer satisfaction (for a constant valuation) and, if we double production, we double cost (for a constant marginal cost). The technology that is used for solving linear programs started being developed seriously around the 1940s to 1950s. Nowadays, with today's hardware and algorithmics, we are able to solve linear programs with up to millions of variables and constraints within reasonable run times. This is a dramatic improvement relative to a few decades ago, when we were able to only tackle problems with thousands or tens of thousands of variables and constraints. Note that we need this kind of scalability because market clearing models with thousands of orders, dozens of trading zones, and dozens of time steps, can escalate to the order of magnitude that starts pushing commercial solvers to their limits. There

are various algorithmic backbones for solving these problems. We mention a couple of options here:

- Interior point solvers that start from feasible market matches and then work their way towards the boundary of the feasible space;
- Simplex-based methods that hop from one corner of the feasible space to another, always making sure that the next step of the algorithm furnishes a welfare that is at least as good as the previous one;
- Dual simplex based methods apply the idea of the simplex method to the so-called dual problem, which is discussed in the next section.

2.3.2 Dual formulation

Theory

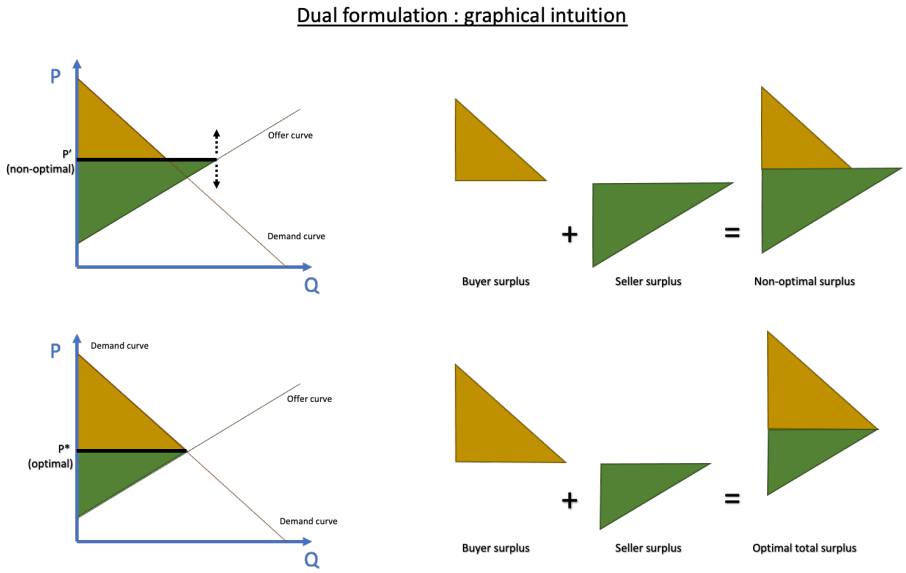
Along with this “primal formulation”, there exists another formulation that uses the same input data – although structured differently. We call this alternative the “dual formulation”.

In this variant, the variables are the clearing prices (as opposed to the primal formulation, where the variables are the cleared volumes). This alternative uses the notion of “surplus”, which can be understood as the positive profit of an individual order. On one hand, for supply orders, the surplus is positive when the clearing price is above the limit price of the order. It is defined as the volume of the order multiplied by the difference between the clearing price and the order price. On the other hand, for a demand order, the surplus is positive when the clearing price is below the price of the order. It is equal to the volume of the order multiplied by the difference between the order price and the clearing price.

Graphically (see Figure 3), the dual formulation consists in finding a clearing price P* such that the sum of (1) the area between the offer curve and the clearing price and

(2) the area between the clearing price and the demand curve is minimal. As it is the case in the primal formulation, the optimal solution is found at the intersection of the two curves. Indeed, as it can be seen in the figure, for an arbitrarily high clearing price, sell orders have large surplus while buy orders have low surplus. Similarly, for an arbitrarily low clearing price, buy orders have large surplus while sell orders have low surplus. Although this may not appear as being intuitive at first sight, the market equilibrium is to be found where the sum of all the surplus is minimal

Figure 3: A graphical intuition of the dual formulation is to move an horizontal line (representing the clearing price) along the y-axis. For a given clearing price level, we sum the profits of all "in the money" bids, which are represented by two positive areas (1) between our horizontal line and the offer curve below the axis – referred to as "supply surplus" (green) and (2) between our horizontal line and the demand curve above the axis – referred to as "demand surplus" (yellow). We find that the smallest area is obtained at the curves' intersection.



Technical Insights

From a mathematical point of view, this dual variable is defined in the so-called space of dual variables. There is one dual variable per constraint of the primal problem, and it corresponds to how much we can improve the primal objective (market welfare) if we are given an extra unit of the corresponding constraint. For instance, λ in the previous section is the dual variable of the market clearing constraint, which is the potential improvement that we could achieve in market welfare if one unit of demand is satisfied for free by having 1 MWh appearing out of nowhere in the right-hand side of the constraint. The overall dual formulation of the basic energy-only market clearing model of the previous section can be expressed as follows:

$$\min_{s,\lambda} \sum_{g \in G} P_g \cdot s_g + \sum_{l \in L} D_l \cdot s_l$$

$$(p_g): s_g \geq \lambda - MC_g, g \in G$$

$$(d_l): s_l \geq MB_l - \lambda, l \in L$$

$$s \geq 0$$

The procedure by which the dual linear program is constructed is based on general linear programming theory[22] but we decided to rather focus on the interpretation of the model in this report. The interpretation is exactly matching the graphical interpretation mentioned earlier where the goal is to find a market clearing price such that the surplus, expressed in the objective function, is minimized.

[22] More details on this procedure can be found in Table 2.1 of the following reference:

Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

Example

Let us now apply these abstract notions to the case of our running example given in Table 4. First, remember that, whereas the primal problem is deciding on the primal decisions, production p and demand d, the dual problem is deciding on the market price λ and the surplus per unit produced/consumed of producers and consumers s. The surplus per unit produced/consumed is expressed in €/MWh, i.e. the profit margin per MWh of the order that is accepted. If the optimal solution of the primal problem is the intersection of the inverse supply and demand functions in the horizontal axis, the optimal solution of the dual problem is the intersection in the vertical axis, i.e. the market price is therefore $\lambda = 40$ €/MWh. The surplus, given this price, is $s_A = 260$ €/MWh for order A (since agent A is willing to pay 300 €/MWh but only ends up paying 40 €/MWh) and $s_C = 0$ €/MWh for order C (since agent C is asking for 40 €/MWh and ends up being paid 40 €/MWh). The surplus of all other agents is 0 €/MWh, since they are not matched in the optimal solution.

Notice the connection between the objective functions of the primal and the dual problem. They are in fact equivalent, and two different ways of looking at the same economic indicator, i.e. market welfare. The primal view expresses welfare as consumer benefit minus production cost, which is the total “size of the pie”, i.e. how much economic value is generated by economic trade. The dual view splits this pie into two pieces, producer surplus and consumer surplus, that go respectively to the producers and consumers. It is important to note that this split of the pie depends on the market clearing price. In our model, producer surplus amounts to 0 €, whereas consumer surplus amounts to 2600 €. The equivalence of the primal and dual formulations corresponds to a more general result coming from the theory of linear programming named **strong duality.** This result states that the solution of the primal and dual problem must be equal in linear programs.

Note that the dual formulation is also a linear program in our example. This is also a specific instance of a more general result in linear programming: the dual of a linear

program is a linear program. Thus, if the primal linear program is easy to solve, its dual linear program (which is equivalent to the primal due to strong duality) is also easy to solve.

2.3.3 Primal Dual Formulation

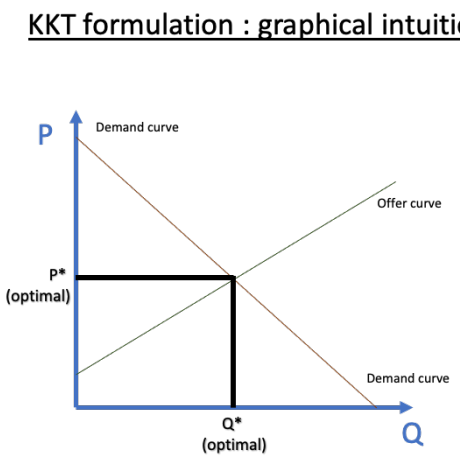
Theory

There exists a third formulation, which – in contrast to the primal and dual formulations – does not include any objective function to be optimized, but solely a set of constraints. Solving this set of constraints (which are technically more complex to state) directly identifies the desired solution, known as the "Karush Kuhn Tucker (KKT) formulation".

The KKT formulation uses the same input data, constraints and variables as in the primal and dual formulations combined. It thus determines both "volumes" and "prices" simultaneously. Graphically (see

Figure 4), this model can be seen as one that identifies the intersection between the offer and the demand curves.

Figure 4: The graphical intuition of solving the KKT equations is to identify the intersection of the offer and demand curves (i.e. the point that respects all constraints), without calculating/optimizing any area (i.e. without the need for any objective function).



This intersection comprises all necessary variables. It provides an equilibrium solution in the sense that it contains no remaining arbitrage opportunity and that the orders' acceptances are fully coherent with the market clearing prices. Specifically, all orders that are "in the money"[23] are fully accepted and orders that are "out of the money"[24] are fully rejected; partially accepted orders are necessarily "at the money" and thereby set the market clearing price. This also happens to be an optimal allocation because it maximizes the total value generated by the market (i.e. the welfare). All market participants are therefore happy with the market results (have no incentive to move away from their cleared position) since all the requirements expressed by their orders are satisfied.

In summary, the primal formulation (i.e. the welfare maximization problem) is a method to clear the market volumes that does not consider clearing prices, while the dual formulation is a method that calculates optimal clearing prices that does not decide which orders are to be accepted or rejected. Combining both methods into the KKT model boils down to finding the intersection between the offer and demand curves.

[23] Orders that are "in the money" are supply orders at a price below the market clearing price and demand orders at a price above the market clearing price.

[24] Orders that are "out of the money" are supply orders priced above the market clearing price and demand orders priced below the market clearing price.

Technical Insights

We return to the running mathematical formulation of the basic energy-only market model that has been discussed throughout this section. The KKT conditions for this market clearing model can be expressed as follows:

$$\sum_{l \in L} d_l - \sum_{g \in G} p_g = 0$$

$$0 \leq s_g \perp P_g - p_g \geq 0, g \in G$$

$$0 \leq s_l \perp D_l - d_l \geq 0, l \in L$$

$$0 \leq p_g \perp s_g \geq \lambda - MC_g, g \in G$$

$$0 \leq d_l \perp s_l \geq MB_l - \lambda, l \in L$$

Mathematically, the condition $0 \leq a \perp b \geq 0$ is equivalent to the three following conditions: $a \geq 0$, $b \geq 0$, and $a \cdot b = 0$. Notice that the system of KKT conditions has no explicit objective function that is being optimized. Furthermore, note that the system is formulated in both the primal and the dual variables of the formulations presented earlier. The idea is that if we can find primal and dual variables that simultaneously satisfy the full set of conditions, then the primal-dual pair is optimal for both the primal and dual formulations above. The KKT formulation, therefore, simultaneously seeks both order matchings as well as economic indicators (order matches, market clearing prices and surplus). As an example, let us focus on order A and check if the third condition is satisfied. It was explained earlier that the surplus for order A is $s_l = 260$ €/MWh, and that $d_l = D_l = 10$ MWh. We indeed confirm that since the surplus of this order is positive, the order must be consuming exactly its bid quantity, i.e. $s_l \cdot (D_l - d_l) = 0$.

The KKT conditions encode conditions of market equilibrium. We can understand conditions of market equilibrium by focusing on so-called piecewise simple orders, which are price-quantity pairs offered for a single time period. Three possible equilibrium behaviors can occur in the context of such simple orders:

- *In the money orders* are orders with a bid price that is better than the market price, i.e. sell orders where the seller is asking to be paid less than the market price or buy orders where the buyer is willing to pay more than the market price. For such orders, naturally the agent wants to trade at its maximum quantity, i.e. a seller wants to sell its full output and a buyer wants to buy its full requested quantity. This is because agents are making a positive profit margin for every MWh of energy traded. Interestingly, the KKT conditions guarantee exactly this: if the market price is favorable, then in the money agents are guaranteed to trade their maximum quantity.
- *Out of the money orders* are orders with a bid price that is worse than the market price, i.e. sell orders where the seller is asking to be paid more than the market price or buy orders where the buyer is willing to pay less than the market price. From a selfish point of view, these orders are wanting to trade a zero quantity. This is because every MWh of energy traded would result in financial losses. The KKT conditions enforce exactly this implication.
- *At the money orders* are orders with a bid price exactly equal to the market clearing price. For sell orders, the seller is paid by the market exactly what it is asking to be paid to produce, and for buy orders the buyer is paying to the market exactly what it is willing to pay. For such orders, from a selfish point of view, the agents are indifferent about the amount of energy traded, because they anyway make a zero profit margin. This possibility of asking at the money agent to produce/consume anything between zero and their bid quantity is exactly implied by the KKT conditions.

The above three business pricing rules are equivalent to so-called **quantity adjustment**, i.e. the fact that, given a price, agents decide on quantity so as to

maximize their own profit. In addition to enforcing these pricing business rules, the KKT conditions enforce one more condition, referred to as **price adjustment**. This condition states that prices go up or down until the market clears, i.e. until supply exactly equals demand. In fact, these quantity and price adjustments represent the entire set of KKT conditions, and are equivalent to the notion of a competitive market equilibrium. In economic theory, a **competitive market equilibrium**[25] is a combination of trades and market clearing prices, (i) such that no agent who is accepting price as a given is willing to deviate from its traded quantity, and (ii) such that the market clears, i.e. supply exactly equals demand. Note that the "competitive" qualifier refers to the fact that agents take price as a given, and are not large or strategic/clever enough to manipulate their decisions so as to influence the price away from competitive levels and in their favor.

Example

Let us now apply these abstract notions to the case of our running example. In this example, we have the following classification of orders, given that the market clearing price is 40 €/MWh, as argued previously:

- Order A is in the money: it is willing to pay 300 €/MWh, but only pays 40 €/MWh. It buys its entire asked quantity, which is to its advantage, because it makes a profit margin of 260 €/MWh for every MWh of energy traded.
- Order C is at the money: it is asking to be paid 40 €/MWh and it is indeed paid 40 €/MWh. It is therefore indifferent between producing 10 MWh or any other amount, since it earns a zero profit margin for every MWh traded.
- Orders B and D are out of the money: For instance, order B is asking to be paid 100 €/MWh, but the market is only paying 40 €/MWh, so the order prefers not

[25] Stoft, S. (2002). *Power system economics: designing markets for electricity.* Piscataway: IEEE press.

Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

to produce. Similarly, order D is willing to pay 10 €/MWh, so the market clearing price of 40 €/MWh is too high for this order. Therefore, the order is better off not buying anything which is indeed the case in the optimal solution of the market clearing platform.

Notice that primal formulation, dual formulation, and primal-dual formulation are equivalent, which carries a very important implication: **the socially optimal allocation of resources is also consistent with the selfish profit maximization goals of agents**. This is a realization of the "invisible hand" of Adam Smith to the case of single-product electricity markets, which can be summarized as follows: maximum satisfaction is guaranteed for market participants in platforms with convex/well-behaved market products.

Technical Insights

The model of section 2.3.3 (i.e. the KKT formulation of the market clearing problem) is very important for ensuring that platforms are optimal both for selfish agents as well as for the ensemble of agents, and for providing economic interpretations to the dual variables that were introduced in section 2.3.2. On the other hand, KKT formulations are computationally intractable. Therefore, they cannot be handled efficiently by commercial algorithms at the scale of realistic auction instances, because they entail equality constraints which feature products of decision variables. This breaks not only the property of linearity, but crucially also the property of convexity. Breaking convexity often leads to high computational difficulty. In practice, we prefer to work with linear models such as those of section 2.3.1 for the sake of computational tractability. The theory of section 2.3.3 is rather used for deriving economic interpretations.

2.3.4 Finding a competitive market equilibrium in the real world

Non-Convex models

The theory described above concludes that a market equilibrium is found at the intersection between offer and demand curves. Such a theory relies on the key assumption that offer and demand bids are continuous and monotonic. This is however not always the case in our practical context: specifically, orders may be subject to various forms of "all or nothing constraints" that encode underlying technical or economic constraints (e.g. startup costs, minimum load, etc.). In their simplest forms, such products go under the name of block bids, and postulate that the market either accepts an order fully at its asking price or a better price, or not at all.

In the presence of constraints of the type "all or nothing", offer and/or demand curves may be non-monotonic. Consequently, from a graphical perspective, identifying the intersection between the offer and demand curve is not always possible. Economically, this means that a "perfect equilibrium" (i.e. a solution that all participants are perfectly happy with) may not exist.

To illustrate our point, let us revisit our running example, and consider this time the last column of Table 4, i.e. we introduce a minimum acceptance ratio of 11/12 for order C. This means that, if order C is accepted, then it must produce at least 11/12 of its entire bid quantity, i.e. at least 11 MWh.

The model can be formulated mathematically as follows:

$$\max_{p,d,u} \sum_{g \in G} MC_g \cdot p_g - \sum_{l \in L} MB_l \cdot d_l$$

$$\sum_{l \in L} d_l - \sum_{g \in G} p_g = 0$$

$$p_g \leq P_g, g \in \{D\}$$

$$11 \cdot u_C \leq p_C \leq 12 \cdot u_C$$

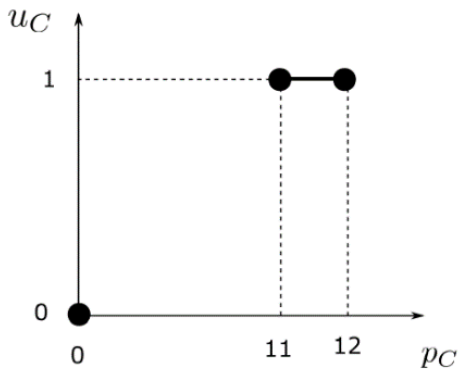
$$d_l \leq D_l, l \in \{A, B\}$$

$$p, d \geq 0$$

$$u_C \in \{0,1\}$$

Note the introduction of a new variable, **binary variable**, i.e. it is forced to equal either zero or 1. The minimum acceptance ratio of order C is enforced in the third constraint of the model. If $u_C = 0$, then the order is forced to produce zero. If $u_C = 1$, the production of the order is allowed to range between 11 and 12 MWh. Although the model correctly captures the requirement of the minimum acceptance ratio, it introduces a very serious complication to our otherwise rosy storyline so far: the feasible set of order C becomes non-convex. This is illustrated in Figure 5.

Figure 5: The feasible set of order C is non-convex. The horizontal axis is the set of feasible production levels. The vertical axis is the set of feasible commitment decisions. The figure is saying that if we choose to not accept the order, then the production of the unit is exactly equal to zero, whereas if we accept the order then the order can produce any amount of energy between 11 MWh and 12 MWh.



Algorithmics

Apart from the fact that the introduction of binary variables introduces formidable computational complications, it also has profound implications on market clearing and market design. We comment first on the computational aspect, and then provide a more extensive discussion on the market design implications.

From a computational standpoint, the market clearing problem described above is no longer convex. It does, however, belong to a specific class of non-convex models, named **mixed integer linear programs**. These are mathematical programs that feature linear constraints and objective function, but where some (though not all) of the variables are binary (i.e. zero-one, or take-it-or-leave-it). This is, in general, a computationally hard class of problems, but one for which our research community has developed many special-purpose methods, which allow us to solve instances of very large scale (and certainly of a scale that is compatible with day-ahead electricity auctions that cover the entire European continent). The backbone of these methods is the so-called **branch and bound algorithm**, which is a form of implicit clever enumeration of possible assignments of values to 0-1 variables, such that certain combinations are never even explored because they are guaranteed to not furnish promising results. Even though models of realistic scale cannot be solved to perfect accuracy in acceptable run times (e.g. multiple minutes to few hours), we can however reach solutions with very solid guarantees (e.g. within 1% to 0.1% of optimal for power system scheduling problems of realistic scale). The EUPHEMIA algorithm itself is based on a branch and bound algorithmic backbone, enhanced with various problem-specific attack strategies.

The more severe complication caused by the introduction of binary variables (i.e. non-convex or “complicated” take-it-or-leave-it market offers) is the fact that there may no

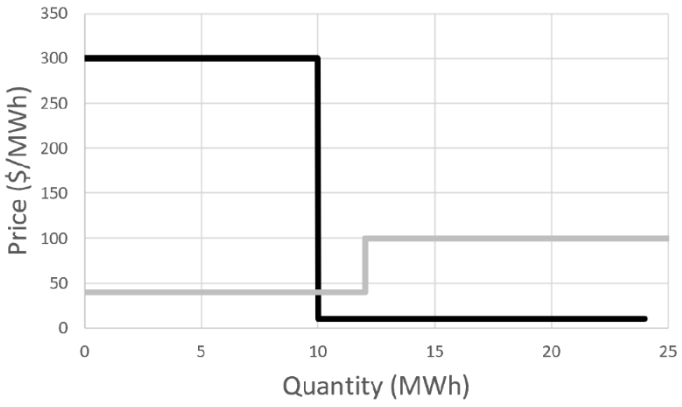
longer exist a market clearing price[26]. This may seem bizarre at first but can be understood by the fact that agents no longer respond smoothly to smooth changes in price, due to the non-smooth behavior of the binary take-it-or-leave-it decisions.

Example

Repeat of Table 4: Order book corresponding to the running example of chapter 2. Source: (Papavasiliou A. , Optimization models in electricity markets, 2023).

Bid	Quantity (MWh)	Price ($/MWh)	Min. acceptance (MWh)
A (buy)	10	300	0
B (buy)	14	10	0
C (sell)	12	40	11
D (sell)	13	100	0

Repeat of Figure 2: Graphical representation of the order book used in the running example of chapter 2. Source: (Papavasiliou A. , Optimization models in electricity markets, 2023).



[26] Stoft, S. (2002). Power system economics: designing markets for electricity. Piscataway: IEEE press.

Let us illustrate the inexistence problem specifically in our running example of Table 4, which is repeated above (along with Figure 2) for the convenient reference of the reader:

- For a price below 40 €/MWh, the market does not clear: the supply that is made available is 0 (because not even order C is willing to produce, let alone order D which is more expensive), whereas the demand is at least 10 MWh (since order A is in the money and therefore wants to consume its maximum bid quantity, and if the price is even lower, i.e. below 10 €/MWh, then even order B wishes to consume, thereby exacerbating the undersupply problem).
- For a price above 40 €/MWh, the market also does not clear: the supply that is made available is at least 12 MWh (because order C wishes to produce fully, and even order D would wish to produce if the price is above 100 €/MWh), whereas the demand is no greater than 10 MWh (since order B certainly is not willing to consume at these high prices). We have a problem of oversupply.
- For a price exactly equal to 40 MWh, which was the equilibrium price before the introduction of the minimum acceptance ratio in C, the market still does not clear. The supply cannot drop below 11 MWh, because of the minimum acceptance ratio of order C, but it will also never exceed 10 MWh, since again order B is not willing to consume at this price. Thus, the market is facing a problem of oversupply.

How the market design should cope with this issue is an ongoing debate in the area of power system economics over the past couple of decades. Numerous opinions/approaches/views have been published on the topic, and the adoption of proposals in practice is also quite diverse. We highlight some of the adopted approaches, and comment on their perceived merits and weaknesses, in section 2.5.

2.4 Multi-product auctions

In a co-optimization context, where there are multiple inter-related products to optimize, using a formal mathematical model enables us to maintain full coherence between all the co-optimized products. Optimization models are (under some conditions – which we discuss below) fairly easy to resolve (unlike models inspired by graphical representations which become quickly intractable in practice).

The idea of multi-product auctions is not unique to electricity markets. Multi-product auctioning occurs, for example, in charity auctions where buyers place bids for multiple products simultaneously. When the gate of the auction closes, the auctioneer determines a price for each product, and allocates each of the products to the respective winning bidder. Another form of multi-product auctioning, where billion-dollar trades take place, are auctions for spectrum licenses[27]. Electricity markets are yet another important arena of applications. The common element shared by all these applications is that the auctioned items are interdependent, and therefore there is value from trading them jointly. This is not to say that separate auctioning cannot take place, but it puts a great burden of price estimation on traders, and is therefore likely to lead to inefficiencies due to inaccurate estimates of opportunity costs.

2.4.1 Multi-product auctions of energy and transmission

Before discussing in detail the joint auctioning of energy and balancing capacity, we first comment on the joint trading of energy and transmission, which already takes place in European markets. Although it may not be directly apparent at first, the existing institution of zonal pricing in Europe or nodal pricing in various US and other

[27] Milgrom, P. (2004). Putting auction theory to work. Cambridge University Press.

worldwide markets is a form of multi-product auctioning, where the multiple products being auctioned off are:

- Energy (one different energy product per location and per time period[28])
- Transmission capacity (rights to use each critical network element at each direction) per time period

Separate auctioning of transmission rights in some original market designs implied that agents would bid explicitly for the right to use certain network elements. This then allowed them to trade with counterparties in different locations, by buying rights on the network to get their power over from their point of supply to the point of their counterparty's demand. The complexities of power flow and how it impacts the explicit trading process are part of the nodal/zonal debate, and are not elaborated on further here. However, one major challenge in such explicit auctions remains the same: agents are asked to anticipate the value of energy in different locations of the system, when deciding on whether to trade or not with a certain location, and therefore how much they should value access to the network in an explicit auction for transmission rights.

The implicit auctioning of transmission capacity which takes place in the existing pan-European day-ahead energy market executes this process implicitly: the auction engine matches trades between different locations in a way that maximizes the value of using the network. In other words, trades are matched between different locations

[28] Note that even a 24-hour day-ahead auction with different energy prices per time period on a copper plate market (with no transmission constraints) is already a multi-product auction, where intertemporal constraints such as ramp rates or min up/down times are a pertinent concern. The interdependency here of energy products in different hours amounts to the fact that certain operational constraints of generators link time periods to each other, e.g. ramp rates or minimum up and down times. In such contexts, it is dangerous to ignore time linkages, because if we do, then we might end up with commitment and dispatch instructions that are technically infeasible or unnecessarily expensive for generators. In order to avoid putting traders in the position of internalizing these constraints and costs, we rather ask them to submit offers that describe the constraints and costs of these resources, rely on a market clearing engine to internalize these constraints and costs, and dispatch assets in a feasible way while pricing the dispatch instructions consistently with the costs and constraints of asset owners.

in a way that generators get the best value for their assets by trading with parties who value them the most and are situated in parts of the system where power can be physically delivered.

In this particular case, multi-product auctioning is already taking place, and the transmission products are the rights to use MWs of network elements. Their value is internalized in the different price of electricity in different locations. A two-node network with a price of 30 €/MWh in one location and 40 €/MWh in the other location has an underlying transmission right for the line connecting the two locations with an economic value of 10 €/MWh. This transmission right is needed in order for entities in the two locations to trade, since their trade uses up space on the network. This transmission right can be traded in forward markets for the sake of risk management, and in order to avoid the risk of variations in the cost of transporting power from one location to another. This transportation cost is typically not an explicit economic cost, but rather an opportunity cost, in the sense that a MW of transmission capacity that is used for matching two trades in different locations could be used for matching more valuable trades. Instead of asking traders to anticipate this opportunity cost, the multi-product auction simply asks them to submit their technical and economic information to the market. Then, the market matches orders while accounting for the inherent limitations of the network.

2.4.2 Multi-product auctions of energy and balancing capacity

Balancing capacity is an essential element of reliable and secure system operation. It is headroom provided to the system operator for balancing out unforeseen uncertainties that emerge in real time, such as forecast errors in renewable supply or load, as well as generation or transmission component failures. Due to the instantaneous requirement of balancing supply and demand in electric power systems,

the transmission system operator would be unable to run the system without balancing capacity services.

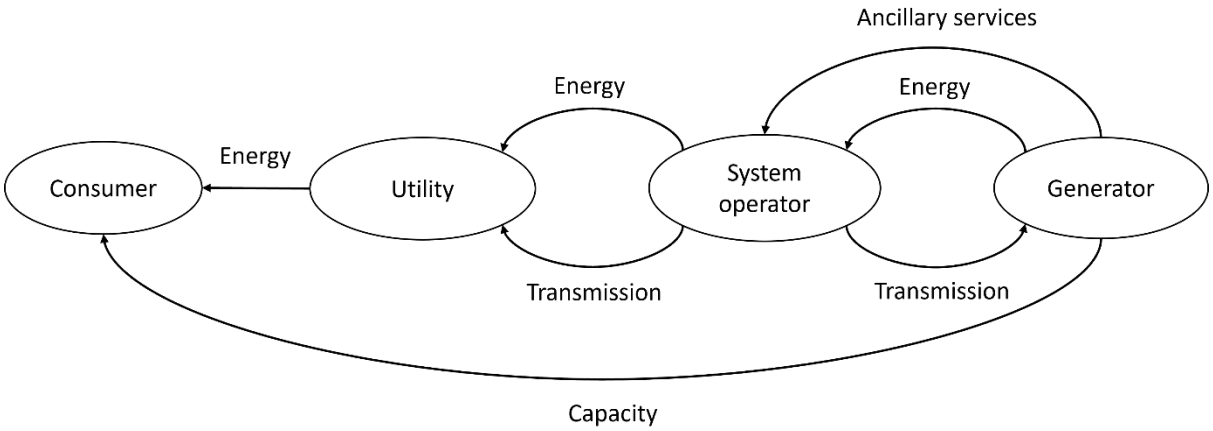
Balancing capacity interacts inherently with other products and services that are traded in electricity markets, which is why many international power markets trade balancing capacity jointly with energy and network access in multi-product auctions. The interdependency of energy and balancing capacity resides in the fact that MWs used for balancing capacity preclude that same capacity from being used for producing energy. The interdependency of transmission network access and balancing capacity is represented by the fact that network access used for unscheduled last-minute flows of balancing energy cannot be used for routine scheduled energy transactions. Due to these interdependencies, it is common practice in several international practice to trade balancing capacity jointly with energy and network access in multi-product auctions.

Balancing capacity can be thought of as a promise that generators sell to system operators for making their capacity shown up in real time. Balancing capacity is traded in MW, but can also be brought to comparable units as energy (MW-h): one MW-h of balancing capacity is the promise to make 1 MW of generation capacity available for an entire hour. The sale of 1 MW of balancing capacity from a generator to the system operator translates operationally to a promise of that generator to bid at least 1 MW of generation capacity to the real-time balancing energy market. This implies that the system operator can count on being able to access and dispatch this capacity in real time for balancing the system, as it sees fit. This is merely a lower bound: the generating unit owner can choose to bid more than 1 MW if it wishes to, the latter being referred to as “free bids” in certain European markets.

The blueprint of a typical electricity market is indicated in Figure 6. Circles in this figure indicate market agents, and arrows indicate traded products/services. Balancing capacity falls under the broader category of “ancillary services” in this figure. The blueprint is not representative of every international market, but does correspond to

the skeleton of many markets that are encountered worldwide. Note that the buyer of balancing capacity in this figure is the system operator, which procures balancing capacity as a public good on behalf of all market participants. Note also that two arrows are leaving the “Generator” agent simultaneously, thus the generator must evaluate the profitability of its alternative value streams when deciding how to allocate its capacity to each of these two markets.

Figure 6: Diagram of traded products/services and market agents in a typical electricity market. Source: (Papavasiliou A. , Optimization models in electricity markets, 2023).



Theory

We now move the discussion of our running example to the focus of this report, i.e. the joint auctioning of energy and balancing capacity. Let us consider again the market products that are described in Table 4, where we ignore the minimum acceptance ratio. Suppose that, in addition to trading energy, the sellers of the system (bids C and D, that are assumed to own generation assets) are also asked to provide balancing capacity. The complication introduced by this requirement is that the balancing capacity offered by sellers is actually headroom capacity that cannot be used for selling energy. This is exactly the interdependency between energy and balancing capacity that multi-product auctions are designed to account for.

Suppose that the system operator is willing to pay 1000 €/MWh for R MW of balancing capacity. And suppose, furthermore, that bids C and D are sufficiently flexible that their full bid quantity of MWh can also be dispatched fast enough to also qualify as reserve[29]. How would one measure economic value, or welfare, in this multiproduct auction setting? This is the sum of the welfare generated by the energy trades, as in section 2.3, plus the value of traded balancing capacity. The multi-product auction model can thus be described as follows, where **red font** marks the newly added elements relative to the energy-only model that is introduced in section 2.3:

$$\max_{p,d,r,rd} \sum_{l \in L} MB_l \cdot d_l - \sum_{g \in G} MC_g \cdot p_g + VR \cdot dr$$

$$(\lambda): \sum_{l \in L} d_l - \sum_{g \in G} p_g = 0$$

$$(\lambda R): dr - \sum_{g \in G} r_g = 0$$

$$(s_g): p_g + r_g \leq P_g, g \in G$$

$$(s_l): d_l \leq D_l, l \in L$$

$$(s_{TSO}): dr \leq R$$

$$p, d, r, dr \geq 0$$

The parameter VR in the objective function of this multi-product auction model is the valuation of the system operator for reserve, i.e. the 1000 €/MWh. The variable r is the supply of reserve, and dr is the demand of the system operator for reserve. The

[29] Generalizations of these assumptions, i.e. more complex demand curves for reserve as well as ramp limits or other technical reasons for limiting the amount of balancing capacity that can be made available by generation asset owners are discussed in further detail in chapter 5.

newly added second constraint states that the demand for balancing capacity should equal the supply of reserve, and the dual multiplier of this constraint λR (which is presented in the parenthesis to the left of the constraint) is the equilibrium price of reserve, in the sense that this is the price at which the TSO procures reserve from resources that can offer it. The same discussion about primal formulations, dual formulations, and KKT conditions that is developed in section 2.3 continues to hold in this section. Specifically, the dual variable λR is now interpreted as an equilibrium price for the balancing capacity product of the multi-product auction (while the dual multiplier λ retains its meaning as the equilibrium price of energy). The computation of λ and λR is now taking place jointly, in an integrated calculation that internalizes the competing incentives of agents and allocates the available generation capacity to its most valuable use. We demonstrate this point shortly, by expanding on our simple illustrative example.

It is important to note that generators do not bid an explicit opportunity cost for reserve in this auction. Instead, the capacity of the generators is allocated optimally between energy and reserves. In contrast to burning fuel for producing energy, there is no other cost for providing reserve in this stylized model than opportunity cost, but this is already handled intrinsically by the multi-product auction.

The interdependency of energy and reserve, which necessitates the multi-product auction in the first place, is captured in the red part of the modified third constraint. The essence of the constraint is that available MWh of sell offers from bids C and D can now be allocated either to energy, or reserve, but not both.

The fourth constraint of the model remains the same, as nothing has changed from the demand side of the market. On the other hand, we have added the fifth constraint, which expresses the fact that the system operator is willing to buy up to R MW of reserve, but no more. Its dual variable, s_{TSO}, can be interpreted as the economic surplus of the TSO per MW of balancing capacity procured for the hour in question. The last constraint imposes that all primal variables are non-negative.

Example

Let us now illustrate the generalization of the concepts presented in section 2.3 by considering a simple instance of this running example. Suppose that the TSO is asking for 16 MW of balancing capacity for one hour. Thus, the TSO participates explicitly in the multi-product auction as a bidder, by submitting a price-quantity bid of 16 MW at 1000 €/MWh, meaning that it is willing to pay no more than 1000 €/MWh for securing up to 16 MW of balancing capacity. Then the optimal solution is to cover 13 MW of this demand from the expensive sell offer (bid D), and the other 3 MW from bid C. This leaves only 9 MWh available for covering the buy bid A. Thus, order A is matched for only 9 MWh, not 10 MWh as was the case in the energy-only auction. Why is this optimal? Consider the contrary case, where order A is fully matched. Then only 12 MW of the TSO demand would be satisfied. For every extra 300 €/MWh of energy that we gain from matching order A, we lose 1000 €/MWh of value by not matching TSO demand for reserve. This tradeoff is exactly captured by the objective function of the multi-product auction model, which generalizes the notion of economic welfare to the welfare that is generated, in monetary terms, by the joint trading of the two interdependent products, balancing capacity and energy.

Let us now consider the equilibrium prices in this running example. The equilibrium price of energy is equal to 300 €/MWh. This is because buy order A is at the money, in the sense that it is asked to consume a positive amount of energy, but one which is below its maximum asked quantity. The only way that we can convince order A to do this voluntarily is by setting the price of energy at its valuation, which then makes the order indifferent about buying zero, a positive amount, or its maximum requested amount of energy.

For the equilibrium price of reserve, the key is to observe that sell order C is asked to split its capacity between energy and reserve, and to ask ourselves what price for balancing capacity would induce order C to do this voluntarily. The answer is a price that results in equal profit margins from the energy and balancing capacity markets. If

the profit margins in the energy market exceed those in the balancing capacity market, then order C will decide to allocate its available capacity exclusively to the energy market and zero to the balancing capacity market (but this is suboptimal), whereas if the profit margins in the balancing capacity market exceed those in the energy market then order C will decide to allocate its available capacity exclusively to the balancing capacity market and zero in the energy market (which is also suboptimal). The previous considerations allow us to infer that the price of balancing capacity is:

$$\lambda R = \lambda - MC_C.$$

This no-arbitrage condition is actually an explicit KKT condition of the multi-product auction model, which means that it is guaranteed to be satisfied by the market clearing engine. The price of reserve, based on this reasoning, is specifically equal to 260 €/MWh, i.e. the difference between the energy price of 300 €/MWh (which is set by buy order A which is partially accepted) minus the 40 €/MWh (the marginal cost of order C, which is splitting its capacity between both the energy and the reserve market).

2.4.3 Impact of imperfect price estimation

We discussed in section 2.3 that the primal-dual solution of the single-product auction model is equivalent to a competitive market equilibrium. This means that the prices and dispatch instructions are bestowed with the very important property that they induce self-interested agents to react voluntarily to the dispatch signal that is sent by the market. This important property continues to hold in the multi-product setting. However, it is now crucial to be careful about keeping track of what it means for agents to maximize their selfish profit when they have access to multiple markets. We already explained how it works for agents A and C in the previous paragraphs, a similar sanity check can be performed for the other orders in the market:

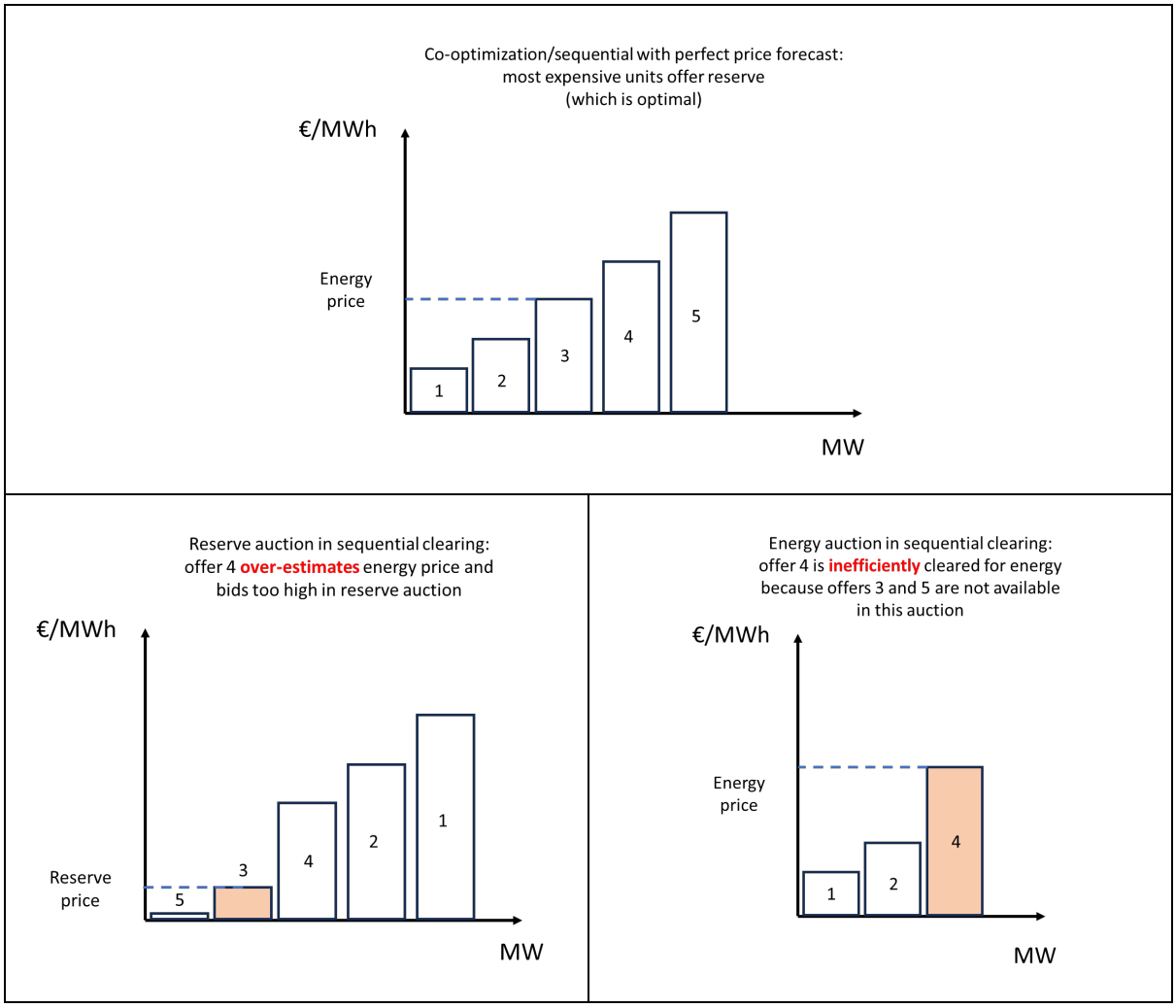
- At an energy price of 300 €/MWh, agent B is not willing to buy energy, which is also the optimal solution of the multi-product market clearing engine.

- At an energy price of 300 €/MWh and a balancing capacity price of 260 €/MWh, agent D is willing to sell balancing capacity (at a positive profit margin of 260 €/MWh), but not energy (since it would do so at a loss), which is also the optimal solution of the multi-product market clearing engine.

Alternatively, one could have set this market up by clearing energy and reserves separately, for instance by trading reserves first and then energy. However, this would require agents to anticipate the price of energy, so that they would be able to bid appropriately in the balancing capacity market. Furthermore, if their expectations of energy prices would be faulty, this would possibly lead to a misallocation of resources in the energy and reserves market. This issue is conceptually shown in Figure 7 and described in a detailed example in appendix A. In the example of Figure 7, we have five units that are numbered in order of increasing marginal cost. The top part of the figure presents the outcome of co-optimization, whereby the most expensive units (orders 4 and 5) are held back for covering reserves, while the cheapest units (orders 1, 2 and 3) are used for covering the energy demand of the system. It can be proven mathematically[30] that this is the optimal outcome of the co-optimization model. The energy price in this setting is determined by order 3.

[30] Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

Figure 7: A conceptual illustration of the efficiency losses of sequential market clearing relative to co-optimization.



The lower part of Figure 7 presents a possible misallocation of resources in the case of sequential clearing. In the left part, we present the merit order of the balancing market, where agents tend to order themselves according to opportunity cost. But opportunity cost depends on agents' estimate of the energy price (whence the origin of the problem). Thus, if an agent overestimates energy prices (as is the case for agent 4 in this example), then it will tend to bid an inefficiently high price in the reserve auction. In this illustration, order 4 which overestimates the energy price ends up bidding too high in the reserve market, and this creates an inefficient reshuffling of the

merit order in the reserve market, where orders 3 and 5 are cleared for reserve, instead of orders 4 and 5. And since order 3 is no longer available for participation in the energy market, the energy market then clears orders 1, 2, and 4 for energy, which results (in the case of this specific illustrative example) in unnecessarily high costs and an unnecessarily high energy price.

2.5 Pricing rules: overview of practice in the US and in Europe

In this section we expand on the problem of coping with non-convexities and the possibility of inexistence of a market clearing price that is described in section 2.3. Before advancing to a detailed discussion on proposed market designs for coping with non-convexities, we discuss the issue with fixed costs and how it connects to market clearing prices, revenue shortfalls, and the need for side payments. This issue is already discussed in section 2.3.3 (the reader is also referred to the example in Table 4 and

Figure 5), but we repeat some of the underlying problems here. Consider a market with identical units that have a capacity of 200 MW, a fixed running cost of 1000 € and a marginal cost of 5 €/MWh. Suppose that this market faces a demand of 360 MW. The optimal way to dispatch the system is by committing three of the identical units and running one of them at part-load, i.e. 60 MW out of the 100 MW that it has available. The difficult question is what the price should be in this market. For any price below 5 €/MWh there will be an undersupply because none of the units would be able to recover their fixed costs at this price. But 5 €/MWh would not cut it either, because if any of the units were to produce then they would not cover their fixed cost. We can crank up the price in order to contribute towards covering the fixed cost of the units, but there is a jump that occurs at 10 €/MWh. At this price, units can cover their fixed cost, but for such a favorable price they are only willing to offer their entire capacity,

and nothing less. So, none of the units in this market can be induced to produce at partial loading, i.e. at 60 MW.

There are two market design philosophies for coping with this issue. One is to have agents internalize their fixed cost in their bids, and insist on receiving simple bids in the market. What is interesting about this example (and can be shown mathematically) is that there is no pure strategy Nash equilibrium by which agents can internalize their fixed costs in their offers, i.e. they would need to randomize their offers in order for the market to settle at a Nash equilibrium. An alternative is to ask generators to submit their fixed and marginal costs separately in multi-part bids, so that the auctioneer can distinguish between the two, and to use side payments that are intended to compensate units for any fixed costs that are not compensated by the market clearing price. These payments, referred to as side payments, are contingent on the agents following the market schedule, so that agents face the incentive of abiding to the instructions of the market. For instance, the auctioneer could clear the market at 5 €/MWh, and pay the units that are asked to operate another 1000 € each in order to cover their fixed cost. Note that the side payment actually depends on the market clearing price. We revisit this point later in this section.

We classify the general approach towards coping with non-convex market clearing models between approaches with and without side payments[31], and further expand on variations of the methods that rely on side payments. Note that the discussion that follows generalizes beyond energy-only markets to multi-product auctions with energy and reserves, and the following analysis is therefore directly pertinent.

[31] The distinction is also referred to sometimes in the European market design jargon as uniform versus non-uniform pricing. The terminology is somewhat misleading, because so-called non-uniform pricing approaches also feature a uniform price, albeit with side payments. The fundamental distinction is relying on side payments (the US approach, referred to in some circles as non-uniform pricing) versus relying on market schedules that are to some extent incompatible with agent incentives (the European approach, referred to in some circles as uniform pricing).

- In methods that rely on side payments, the idea is to maintain the matching of market orders that maximizes welfare, and to rely on **side payments** in order to induce agents to follow the market schedule voluntarily (since these side payments are contingent on following the market schedule). We explore three specific variants of this approach. All variants rely on an overall process that is depicted graphically in Figure 8, whereby a "primal" problem is first solved for matching orders, followed by a "dual" problem that is used for computing market clearing prices. The three variants that we discuss below all use the same "primal" model, which aims at matching orders to maximize welfare. The point at which they differ is the dual pricing part of Figure 8:
    - **Convex hull pricing**: In this approach, prices are computed by considering the convex hull of the feasible region of the different bidding products. The convex hull of the bidding products is the closest "nicely behaved" approximation of the true market product. An approximation of this approach is adopted in MISO.
    - **Integer programming (IP) pricing**: in this approach, uniform prices are generated by fixing the integer variables of the model, and solving for market clearing prices while pretending that market agents have already decided to fix their binary decisions to their optimal values. This approach is adopted in CAISO.
    - **Linear programming relaxation**: in this approach, uniform prices are computed by using the linear programming relaxation of the model. In other words, we pretend that take-it-or-leave-it decisions are not truly take-it-or-leave-it, but can instead be fractional. This approach is adopted in PJM.
- In methods that do not rely on side payments, the idea is to choose order matchings such that no side payments are required, while also tolerating paradoxically rejected block orders. This is the approach that is adopted in the European market. The pricing model is a mixed integer quadratic program subject to complementarity constraints. The problem is mixed integer because

it is matching both continuous orders, as well as orders with binary attributes (e.g. of the block order type). The complementarity constraints encode the pricing business rules of the auction (e.g. the acceptance rules for simple orders that may be either “in the money”, “at the money” or “out of the money”, as explained in section 2.3, or the fact that block orders can be paradoxically rejected but not paradoxically accepted).

The overall ranking of the different pricing methods along a number of dimensions is presented in Table 5. The different methods are ranked according to six criteria in the table:

- **Computational tractability**: This criterion determines how easy it is to compute market clearing prices from an algorithmic standpoint.
- **Efficiency**: This criterion determines to what extent the design in question selects the most efficient matching of market orders.
- **Gaming**: This criterion determines to what extent market agents are incentivized to stray away from truthful bidding in order to extract surplus from the market.
- **Fairness**: This criterion determines the extent to which each different pricing method is able to achieve non-discriminatory market clearing outcomes.
- **Budget imbalance**: this criterion determines the extent to which the market operator is exposed to a budget imbalance that it needs to cover through side payments.
- **Price interpretability**: This criterion describes the extent to which different pricing methods exhibit certain properties that are deemed desirable by different market stakeholders.

The scores that are assigned in Table 5 vary as follows:

- Unfavourable (--) is indicated in solid red
- Somewhat unfavourable (-) is indicated in opaque red

- Neutral (0) is indicated in opaque purple
- Somewhat favourable (+) is indicated in opaque green
- Favourable (++) is indicated in solid green

We proceed to justify the scoring of each pricing approach along each of the criteria that are listed in Table 5 in the remainder of this section.

Table 5: Alternative methods for pricing with non-convexities and their relative pros and cons.

Pricing approach	Geography	Computational tractability	Efficiency	Gaming	Fairness	Budget imbalance	Price interpretability
No side payments	EU day-ahead market	--	-	+	-	++	+
Convex hull pricing	MISO	-	++	0	-	+	-
Integer programming (IP) pricing	CAISO	++	++	0	-	-	+
Linear programming relaxation	PJM	+	++	0	-	0	+

Computational tractability

The approach that relies on no side payments results in mixed integer quadratic programs subject to complementarity constraints. This is generally a computationally intractable class of optimization problems. Although the EUPHEMIA algorithm includes numerous special-purpose algorithmic methods for coping with the inherent complexity of the problem, this inherent computational complexity imposes scalability challenges (e.g. with respect to 15-minute time frames, increase in the number of bidding zones, increase in the geographic scope of the algorithm, increase in the

number of bids due to a move from portfolios to unit-based bidding, etc.) and certain limits on bidding products (e.g. block orders that are either zero or one over a set of periods, as opposed to unit commitment models where zero-one variables can vary from hour to hour).

Convex hull pricing receives the second lowest score in this dimension. Characterizing convex hulls is inherently hard, but there are methods that rely on approximations of the convex hull of certain assets[32] as well as dual decomposition methods[33] that can be employed for approximating convex hull prices. However, primal methods can be sensitive to the introduction of minor changes in the feasible sets of market assets (e.g. ramp rates) while dual decomposition methods may struggle to scale in systems with many market areas.

Integer programming and the approach based on linear programming relaxations receive the best score in this dimension, because both amount to solving large-scale linear programs, which are known to scale well, even in systems with a large number of market areas and fine time resolution. The integer programming approach receives a slightly higher score in this dimension because it fixes binary variables, whereas the linear programming relaxation does not and therefore optimizes over a larger set when computing market clearing prices.

Efficiency

All approaches that rely on side payments receive a perfect score in this dimension, because all of these methods retain the order matching that maximizes welfare. Instead, the approach that does not rely on side payments scores somewhat poorly

[32] Andrianesis, P., Bertsimas, D., Caramanis, M. C., & Hogan, W. W. (2021). Computation of convex hull prices in electricity markets with non-convexities using Dantzig-Wolfe decomposition. *IEEE Transactions on Power Systems, 37*(4), 2578-2589.

[33] Stevens, N., & Papavasiliou, A. (2022). Application of the Level Method for Computing Locational Convex Hull Prices. *IEEE Transactions on Power Systems, 37*(5), 3958-3968.

along this dimension. The reason is that this approach is designed to prioritize price discovery over welfare, meaning that the optimal solution may be discarded if a price cannot be found that supports this matching. Note that the strength of this effect may be less severe than indicated in the table, however this is not possible to establish based on publicly available data, since the unconstrained optimal order matching of the pan-European day-ahead auction is not reported in public documentation.

Gaming

Gaming in designs that rely on side payments is discussed in section 3.4, where we also provide detailed bibliographical references that explore gaming opportunities both analytically as well as empirically. An important point, however, is that the gaming strategies described in section 3.4 rely on a two-stage calculation of side payments which is contingent on both day-ahead as well as real-time market outcomes, whereas one-shot auctions do not exhibit this feature. For this reason, we score these designs as neutral along this dimension. On the other hand, there is no substantial experience of gaming in designs that do not rely on side payments (at least in the context of perfect competition).

Fairness

All approaches receive a somewhat unfavourable score in terms of fairness. The approaches that do not rely on side payments may be deemed as being unfair because they discriminate against paradoxically rejected block orders (since they do not allow them to be matched, even if doing so would result in a positive profit for these orders). On the other hand, approaches that rely on side payments discriminate through the side payments themselves. Thus, all methods discriminate (and any attempt to argue about which form of discrimination is more “severe” could be considered as being subjective). This discrimination is not desirable, but it is an inescapable mathematical consequence of the fact that the underlying market model is non-convex, and thus a non-discriminatory market clearing price is not guaranteed to exist.

Budget imbalance

The approach that does not use side payments receives a perfect score in this dimension. By design, this approach leaves the market operator with a zero budget imbalance, since there are no side payments to finance. Arguably, this exact property has been one of the reasons why this specific design has been preferred in Europe, since market stakeholders that have been heavily involved in the design of the European day-ahead market have also been the most averse to being exposed to side payments.

Convex hull pricing receives the second-best score. It can be argued that the amount of side payments that should be paid to market participants should equal lost opportunity cost (though certain market operators prefer to remunerate make-whole payments instead[34]). Since convex hull pricing minimizes lost opportunity cost, it results in the second-best performance in terms of side payments. The linear programming approximation ranks third in this dimension, because it can be interpreted as an approximation of convex hull pricing, and as such comes fairly close to convex hull pricing in terms of minimizing lost opportunity cost. IP pricing ranks least favourably in this dimension, as observed both analytically[35] as well as in empirical case studies of the Central Western European System and through a dataset compiled by the Federal Energy Regulatory Commission[36].

[34] Schiro, D. A., Zheng, T., Zhao, F., & Litvinov, E. (2015). Convex hull pricing in electricity markets: Formulation, analysis, and implementation challenges. *IEEE Transactions on Power Systems, 31*(5), 4068-4075.

[35] Hogan, W. W., & Ring, B. J. (2003). *On minimum-uplift pricing for electricity markets.* Cambridge, MA: Harvard Electricity Policy Group.

[36] Stevens, N., Papavasiliou, A., & Smeers, A. (2024). The Many Advantages of Convex Hull Pricing for the European Electricity Auction. *Energy Economics, under review*.

Price interpretability

Convex hull pricing receives the worst score in this dimension. The reason is because this pricing method exhibits certain properties that are deemed as being undesirable by certain stakeholders[37]. Such properties include, for example, the fact that there may be price separation between market zones that are connected by a line that is not congested, or the fact that offline units may affect price formation.

Integer programming pricing and the pricing approach that does not rely on side payments exhibit the desirable property that they satisfy certain well-known and well-accepted pricing business rules that concern networks, e.g. the fact that in transportation networks power flows from low-price to high-price locations and prices are equal in neighbouring locations that are connected by uncongested lines. On the other hand, paradoxical rejections may be challenging to explain in the case of EU pricing, while the IP pricing approach fails to account for fixed costs altogether. Furthermore, although linear programming pricing can account for non-convex costs, it does so only approximately, therefore exactly explaining price formation can sometimes be non-obvious in this case too. Finally, all methods can produce prices that are challenging to explain in the case of intertemporal constraints such as ramp rate limits. Thus, none of the designs receives a perfect score in this dimension.

2.6 Integration of market and system operations

A particularly relevant aspect of implementing co-optimization of energy and reserves is the separation of roles and responsibilities between transmission system operators and power exchanges. We comment briefly first on US design, and then contrast it to

[37] Schiro, D. A., Zheng, T., Zhao, F., & Litvinov, E. (2015). Convex hull pricing in electricity markets: Formulation, analysis, and implementation challenges. *IEEE Transactions on Power Systems, 31*(5), 4068-4075.

the European design, as it relates specifically to co-optimization of energy and reserves.

Independent system operators (ISOs) are at the heart of US market operations, in the sense that ISOs are responsible for both system operation (at all time stages, from planning to real-time balancing) as well as market operation (at all time stages, from forward to real-time markets). A key advantage of this setup, as it relates specifically to co-optimization, is that there is no need for separating the trading of reserves from energy. All functions are operated under one roof, both in day-ahead markets, where energy and reserves are co-optimized in unit-based models, as well as in real time (where co-optimization of energy and reserves is either directly implemented or approximated through ex-post adders).

By contrast, the European design draws a clear line between the roles of power exchanges and those of system operators. An intuitive way to think of the separation of functions is that transmission system operators (TSOs) maintain those parts of system function that are required for ensuring secure and reliable system operation (essentially all real-time system operation, corrections to network congestions that go under the name of congestion management, and forward reserve markets), whereas the rest (essentially forward energy markets, e.g. day-ahead and intraday markets) are handled by power exchanges. This is not necessarily the only setup that is encountered in Europe, nevertheless it is quite dominant and widespread, including in major EU Member States that strongly influence the formation of European electricity market design policy.

Here lies a conundrum in implementing co-optimization of energy and reserves in forward markets: since forward reserve markets are currently handled by system operators, whereas current energy markets are handled by power exchanges, how can the two processes be merged into a single function, in the spirit of article 40 of the EBGL? There has been an extensive stakeholder debate and technical analysis

around this topic[38]. Attempts to implement article 40, while maintaining the separate functioning of power exchanges and TSOs, essentially morphed into proposals for so-called “uni-lateral bid linking” between the energy and reserve markets. These uni-lateral linking proposals essentially defeat the purpose and spirit of article 40, and are not actually implementations of co-optimization between energy and reserves. Indeed, they are rather attempts to override it and to maintain instead the current status quo of first clearing reserves and then energy. Instead, multi-lateral linking, as described in these technical documents and explained in section 3.4.4, maintains the spirit of co-optimization of energy and reserves, albeit with some redundancy. Specifically, the idea in multi-lateral linking is to allow agents to submit separate bids in reserve and energy markets, as they currently do. These bids should nevertheless be tagged with a so-called multilateral link, which connects the energy bid to the reserve bid and vice versa. A multi-product auction that clears both energy and reserves is then run. Here lies the redundancy: this auction would be run by both power exchanges and system operators, or by a super-entity that collects bid data from both entities.

This tight interaction of roles in the day-ahead time frame can induce institutional resistance. This is understandable, but can arguably be overcome through redundancy. Redundancy already exists in European market operations, as the day-ahead energy auctions are run in multiple servers over different countries of different power exchanges. Similarly, some degree of redundancy is being introduced by possibly overlapping roles of regional coordination centres and national transmission system operators. This redundancy is arguably a low price to pay for the sake of

---

[38] Co-Optimization of Energy and Balancing Capacity in the European Single Day-Ahead Coupling, N-SIDE technical report, 2022.

Implementation impact assessment for the methodology for a co-optimized allocation process of cross-zonal capacity for the exchange of balancing capacity or sharing of reserves, ENTSO-E technical report, 2021.

arriving to an improved market design which is better equipped for the future needs of the market. By contrast, attempts to package “unilateral multi-step bid linking” as implementations of the spirit of article 40 are somewhat contradictory, since the multiple steps essentially maintain the status quo of performing one function first before the other and thus keeping the clearing of energy separate from that of reserve.

2.7 Consistency of market models and products between time stages

Electricity system operations is one of the most advanced receding horizon optimization processes that is encountered in large engineering systems. The system operator is involved in a continuous process of adapting decisions such as the scheduling and dispatch of generation units to the revelation of system conditions such as the realization of load and renewable forecast errors along with the failure of system components. The balance as we approach real time is that degrees of freedom are continuously decreased (since some decisions have to be updated no later than a certain point in time, e.g. slow-moving units cannot be started up in the last minute, instead start up decisions for certain technologies need to be fixed hours in advance) while information about the prevailing conditions in real time continuously increases. Electricity markets are organized in a way that mimics this interplay between advance decision-making and real-time adjustments. Thus, we have real-time markets where the physical delivery and real-time pricing of electricity takes place, day-ahead and earlier forward markets where decisions are taken in anticipation of what will happen in real time.

Principles

The principle of designing forward markets is to allow for risk management, without interfering with the efficient working of the market in real time. This is accomplished through forward contracts, whereby electricity is traded in the day-ahead (or earlier) at

a pre-agreed price, and any deviations from the promised day-ahead quantities are settled at the real-time price. If agents stick to their day-ahead agreed quantities, then they are able to trade at the day-ahead price, which allows for risk management, but nothing prevents them from deviating from their day-ahead price if this is to their own financial favor and aligns with the needs of the system in real time. For instance, consider a generator with a marginal cost of 40 €/MWh that has traded energy in the day-ahead market at 50 €/MWh. And suppose that the system finds itself in tight conditions in real time with the balancing market price shooting up to 150 €/MWh in response. This real-time price signal gives an incentive to the agent to squeeze out an extra MW of power in real time if the extra stress imposed on its generator does not exceed a financial damage of 150 €/MWh, which is economically efficient, but is also in line with the agent's selfish profit maximization goals. Alternatively, the agent can ignore the high balancing price and carry on with producing its originally traded day-ahead quantity, in which case the agent safely ends up being paid the day-ahead price.

The example provided above corresponds to the so-called two-settlement system for trading energy: forward quantities are traded at forward prices, and deviations from forward positions are settled at the price of the real-time/balancing energy market. The two-settlement system can be generalized to back-to-back forward markets, where the idea is that each market produces a price signal for settling changes in position relative to the closing of the previous market for the same underlying product. The two-settlement system can also be generalized in terms of market products, and can specifically apply identically to multi-product auctions, where forward markets for energy, transmission and reserve are cleared simultaneously.

A very important goal to keep in mind when setting up a two-settlement system, or more generally a system of forward markets followed by real-time markets, is to stick to consistent product definitions. This means that whatever is traded in real time should also be traded in day-ahead and all other forward markets. Deviating from this principle has empirically been observed to lead to trouble. Such trouble includes

providing the opportunity to market players to extract surplus from the market by finding ways to exploit inconsistencies in product definitions, or adverse side-effects in price formation in forward markets, to mention a few symptoms.

This principle of consistent product definitions between the day-ahead and real time, combined with the principle of consistency between real time and system physics, is very powerful, because it largely dictates how forward markets should be set up. Real time must obey to system physics. These physics dictate how the problem can be decentralized, and therefore what products should be defined. Consistency between real-time and forward markets then implies what the corresponding forward products should be. Here is one interesting application of the principle: in a system without network constraints, energy balance has to hold in real time, meaning that the supply of energy should equal the demand for energy. Interpreting real-time imbalances as inelastic demand for real-time energy and balancing market offers as price-elastic demand (for downward balancing energy offers) or supply (for upward balancing energy offers) for real-time energy then suggests that balancing prices and imbalance settlement should be tightly linked, and should serve the role of an index for settling day-ahead and other forward energy markets. Thus, day-ahead energy markets should trade energy one day in advance, and deviations from these traded quantities should be settled at prevailing balancing energy prices.

Physical needs

Consistency of product definitions in this setting is crucial. Pretending that real-time balancing and imbalance settlement are disconnected from day-ahead energy prices, and designing day-ahead markets first without paying attention to keeping the design of the real-time markets consistent with the day-ahead design, can lead to fundamental design flaws that can be extremely challenging to undo. Paradoxically, European market design has followed the opposite of the recommended order in its evolution: day-ahead market design and day-ahead market coupling were taken on first, and only recently did we turn our attention to the design of real-time market

coupling, only to find out that some serious issues have emerged because the real time also has to obey physical system constraints, and replicating the day-ahead design in real time markets (as opposed to the other way around) provides no guarantees that real-time physics re obeyed. A good example of this predicament is MARI, the platform for trading the activating of manual frequency restoration reserve between European countries. Various national TSOs in Europe have rightfully expressed concerns about the fact that MARI can threaten domestic system security, because MARI has been designed according to the forward day-ahead market model, in which power flow constraints are aggregated into models that do not correspond at a satisfactory level to physical reality. This has triggered numerous derogations by certain Member States which are delaying their entry to the MARI platform until TSOs can figure out how to cope with this problem. One way to cope is so-called bid filtering[39], where the idea is to block bids that, if activated, can threaten network security. Although this is a workable solution given the problematic setup, it also defeats the point of market coupling to a certain extent, because it prevents flexible resources from actually making their flexibility available to the market. The fundamental problem here is not filtering in itself, this is merely a way to cope with the symptoms. The origin of the problem is that the day-ahead model was designed first, without necessarily adhering to actual real-time constraints, and when the real-time model was subsequently designed in a way that mimics the day-ahead model it failed to satisfy real-time physical constraints. The remedy would have been to design real-time markets first, in a way that respects actual physical constraints, and then step back in time to design forward markets that are consistent with the real-time markets.

[39] Håberg, Martin, Hanna Bood, and Gerard Doorman. "Preventing internal congestion in an integrated European balancing activation optimization." Energies 12.3 (2019): 490.

N-SIDE, “Study – System balancing solutions with detailed grid data”, April 30, 2020, available online: https://www.statnett.no/contentassets/3b981e22e5d64179bb22ea9e5b46f515/2020-study---system-balancing-solutions-with-detailed-grid-data.pdf

Consistent products across timescales

There are other interesting examples of trouble that has occurred from violating the principle of consistency between market models and physics along with consistency between day-ahead and real time. INC-DEC gaming is a famous such example[40]. Originally observed in the early 2000s, INC-DEC gaming contributed among others to the collapse of the original California market in 2001 and its ramifications are still felt intensely in various European member states, including for instance the tumultuous launch of the Greek real-time market under the target model. If the MARI example described above was a manifestation of violating consistency between physics and market models, INC-DEC gaming is a symptom of violating the principle of consistency between day-ahead and real-time market models. The idea in INC-DEC gaming is to exploit redispatch markets, which are corrective markets that ask market participants to back off from their forward (e.g. day-ahead) positions, often through a pay-as-bid arrangement. These redispatch markets are exactly required because day-ahead markets that aggregate networks are de facto inconsistent with real-time operations. The weak link that can be exploited in this setting is that this creates an arbitrage opportunity for market participants who can trade energy at a certain day-ahead price, and effectively buy it back at a forced location-dependent payment (the payment of the redispatch market) which is unavoidable since the constraints of the network must be respected. For instance, a generator that sits behind a line with a very limited capacity is (wrongfully) allowed to sell at the day-ahead market in a zonal design. If the generator knows that the TSO is forced to pay it back for withdrawing its sale in the redispatch market, then the generator can pretend that its redispatch cost is very low, or even zero, or even negative at the redispatch phase. In which case we have the generator being paid the difference between the zonal price and the redispatch payment for essentially offering nothing to the system. Interestingly, the generator

[40] Alaywan, Z., Wu, T. & Papalexopoulos, A. (2004), Transitioning the California market from a zonal to a nodal framework: An operational perspective, in 'IEEE PES Power Systems Conference and Exposition', pp. 862–867.

does not even need to be in a position of market power to exploit this design flaw[41], it merely needs to be able to predict that the line constraint will be binding.

Scarcity

One final example of adverse side effects between design inconsistencies is discussed in section 5.4 of the report. The point discussed there pertains to scarcity pricing, and the fact that European market design has notably put in place forward markets for balancing capacity but forgotten to also implement a mechanism for trading that balancing capacity in real time. We thus have a strange setup whereby we set up a forward market that is indexed to an inexistent real-time market. This predictably creates trouble with allowing for a day-ahead or earlier forward signal for balancing capacity to emerge, since back-propagation of real-time prices to forward markets is no longer at play (though one can still have non-zero forward prices if there are inherent economic costs for providing balancing capacity, such as start up, or binding constraints in day-ahead market models, but back-propagation is not one of these drivers).

Implementations

US markets have largely respected the principle of consistency between real-time and day-ahead markets and products. The products are identical between the two time frames. The only change resides in the fact that real-time offers eliminate some of the things that can still be decided upon in day-ahead markets. For instance, real-time economic dispatch is only comprised of a marginal cost function, whereas day-ahead unit commitment also includes, as part of multi-part bids, information relating to min up/down times, ramp rates, start up costs, min load costs, and everything else that is needed to formulate a unit commitment model. Note that this is not an inconsistency

[41] Hirth, L. & Schlecht, I. (2018), Market-based redispatch in zonal electricity markets, Technical report.

between real time and day ahead. It is merely a reflection of the fact that the only thing that can still be managed in real time is the setpoint of the unit, with unit commitment decisions fixed at this stage, whereas the day-ahead market model can still decide on unit commitment and thus requires all this additional information that is included in the multi-part bid. The real-time bid thus includes a subset of the information of the day-ahead bid, and only that information which is still pertinent for making real-time adjustments. The products, on the other hand, are identical in day-ahead and real time: in both cases energy is traded at a given location and at a given time period. Consistency between day-ahead and real time is also reflected in the fact that many US markets implement locational pricing in both stages, as well as the fact that real-time markets for balancing capacity are implemented in all US markets.

Things are not quite the same in European design. A number of examples have been provided earlier, which discuss deviations between market models and system physics, or deviations between day-ahead models and real-time models. An important corollary of the previous discussion is that viewing balancing operations as a side service of secondary importance just because it trades corrective actions and thus lower volumes than the day-ahead market is a naïve design flaw. Real time is bound by physics, and since day-ahead design should follow real-time design it dictates price formation in day-ahead and other forward market stages. This suggests that viewing real-time balancing as something separate and disconnected from wholesale market operations and design is a recipe for trouble, and that real-time balancing markets should carefully be designed first before turning attention to day-ahead market designs, which should follow the same architecture.

The discussion provided above is summarized in Table 6. The first column of the table indicates the time frame of market operation, the second column enumerates the three major products/services that electricity markets trade, and the last two columns describe the layout of the European and US design. Note that the US design is fully consistent in trading energy, balancing capacity and transmission in both the day ahead and in real time. Minor deviations that exist to this standard market design are

not discussed here, but we mention two in passing: loads are priced on ex-post zonal prices in ERCOT, so transmission access is not precisely traded for loads there, and balancing capacity is not yet co-optimized in real time in ERCOT, but real-time balancing capacity prices are rather computed ex post and added on top of real-time energy prices, as co-optimization dictates.

The European design features question marks in the transmission lines of both the day-ahead and real-time models, because transmission constraints are aggregated in the market models of both these time stages (though not completely ignored). Balancing capacity is absent in real time. The day-ahead and real-time designs are thus not entirely consistent over time.

Table 6: Architecture of European and US markets in the day ahead and real time.

Time frame	Product/service	Europe	US
Day ahead	Energy	X	X
	Balancing capacity	X	X
	Transmission	?	X
Real time	Energy	X	X
	Balancing capacity		X
	Transmission	?	X

Despite the fact that the European design deviates from the principle of consistency between the day ahead and real time in certain cases, as also illustrated in Table 6, there are workarounds to align the two time stages. Chapter 5 lays out a strategy for including real-time balancing capacity prices that are included ex post in balancing energy prices, in an attempt to mimic a co-optimization of energy and balancing capacity in real time. On the other hand, there is also a fundamental tension between

day-ahead and real-time models which relates to the role of uncertainty: uncertainty is largely knocked out when arriving to real time, but it is very much present in the day-ahead model. This might be fine for firm trades; however, it can create challenges when trading option-like products such as balancing capacity. Concretely, selling balancing capacity across borders to a national system operation entitles the operator to activate this capacity in real time up to the traded level of capacity. However, this is not an obligation. This raises a complex issue of ensuring that day-ahead trades do not violate network limits. This is very much a real issue in any system, and to the best of our knowledge it simply has not been faced head on in US professional or academic literature. Chapter 4 focuses on precisely this challenge.

3. Bidding Product Design

3.1 Takeaways

→ There are two main approaches to bidding product design. We refer to them as the “European approach” and “US approach”. The US approach has already proven its compatibility in terms of the co-optimization of energy and balancing capacity through successful designs such as MISO, whereas the topic of bidding product design and co-optimization is still debated in Europe.

→ In the European approach, standardized products (simple bids) that are not asset-specific are used in the wholesale energy market. This approach facilitates the aggregation of individual assets into portfolios. Such an approach provides high flexibility to traders, while the absence of side payments reduces concerns related to discriminatory settlement. However, standardized products limit bidding expressiveness and complexify the modelling of actual economic and technical constraints. The European business rules for pricing are intrinsically more challenging from a computational standpoint, which may limit the scope for detailed bidding products.

→ In the US approach, so-called “multi-part” bids are used, which feature assets’ details to facilitate co-optimization of energy and ancillary services under a central dispatch unit commitment model. The design is supplemented by a series of rules to invite market agents to adequately follow the unit-based central dispatch paradigm. Such an approach is in principle efficient thanks to high bid expressiveness. However, as it is a

more prescriptive bidding language, it can be seen as inflexible to the requirements of new technologies, hindering innovation. It may also be considered as discriminatory in terms of settlement because of its use of side payments.

→ An additional benefit of unit-bidding and multi-part bids which extends beyond co-optimization is that it facilitates effective market monitoring. This is due to the fact that a more granular view on the different cost components and operational constraints of the participating units (compared to portfolios with standardized products) can be obtained.

→ The computational complexity in the US model relates to the large number of resources (unit bidding) and detailed network representation (nodal pricing) while the computational complexity in European markets results from the simultaneous search for bid matchings and compatible prices (price-based conditions that must be enforced). Including ancillary services adds non-trivial computational complexity in both paradigms.

→ The fact that unit-based systems can account for precise network models implies that the distinction between balancing and congestion management close to real time becomes irrelevant. There is a variety of benefits that result from this, including the lifting of numerous gaming opportunities, better locational investment signals (especially for co-locating future renewable capacity closer to load centers), more secure operation near real time, better utilization of available network capacity, more efficient day-ahead commitment decisions, and easier cross-border sharing of flexible resources near real time.

3.2 Overview

Bidding product design refers to the bidding language that market participants can use to describe their economic preferences and technical constraints. The bidding language should ideally strike a good balance between expressiveness and simplicity. Expressiveness means for instance the possibility for the market participant to accurately describe technical constraints such as generation units' minimum up and down times, or start-up costs. The topic is of utmost importance in a co-optimization setup where market participants should be able to express linkages or interdependencies in the provision of energy and ancillary services; for instance, that a unit is able to provide upward reserve only if it is up and already producing above its minimum power output level, or that a unit cannot provide energy if it is booked to provide reserve.

In this chapter, we outline and discuss two major aspects of bidding language design, that of unit versus portfolio bidding, and that of multi-part bids versus simple bids. The two aspects are interrelated. Throughout the section we compare them as they are implemented in European versus US market design. Specifically, US market design is founded on unit-based bidding with multi-part bids. Instead, much of the EU market is based on portfolios that are represented in the market through products that aim at being simpler. The scientific literature[42] provides a juxtaposition of the two designs.

[42] Herrero, I., Rodilla, P., & Battle, C. (2020). Evolving Bidding Formats and Pricing Schemes in USA and Europe Day-Ahead Electricity Markets. *Energies*, 1-21.

Chapter 3.2.4 of Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

3.3 Key notions

The definition of bidding products is a continuous process of evolution for all market designs, including unit-based as well as portfolio designs. Original debates about product design focused on how to integrate the technical and economic characteristics of thermal assets, especially their non-convex features, into the market clearing model. The introduction of balancing capacity to market clearing models in US designs over the past years has required significant modelling and algorithmic innovations, but has also resulted in significant economic savings[43]. The new frontier in bidding product design is dictated by the rise of distributed resources, storage and demand response. The new bidding products that are proposed or implemented in order to accommodate these resources are discussed later in the report. Similar to thermal assets, these assets can interact with both the energy market as well as the balancing capacity market.

Two important dimensions that affect the definition of bidding products include whether units are represented individually or in portfolios, and how the characteristics of resources are represented in the market clearing algorithm. Both dimensions affect the definition of bidding products as discussed below.

An important thing to bear in mind in the subsequent discussion is the institutional dimension of market design and bidding product definition. US market design relies on an integration of system and market operation, whereas these roles remain distinct in the European market where market operation in the day ahead and in intraday is handled by power exchanges, and real-time market operation along with system operation is handled by the transmission system operator. This also explains, to some

[43] A notable accomplishment in this front was the award of the prestigious Franz Edelman award (https://www.youtube.com/watch?v=w_MYQEMy0h0&t=121s) in 2011 to the Midcontinent Independent System Operator for the successful deployment of an energy-reserves co-optimization model in their market operations through use of branch and bound mathematical optimization algorithms. MISO estimated the savings from this evolution at $2.1-3 billion in the period from 2007 to 2010.

extent, the stark difference in structures that are encountered in different parts of the world. US markets are typically unit-based, and the independent system operator clears energy simultaneously with reserves (and transmission). European markets are often (though not in all cases) portfolio based, and reserves are cleared separately from energy, either before or after, in day-ahead or less frequently in forward reserve markets. The energy market which is operated by power exchanges has no specific association with individual physical assets. Balancing capacity markets can also be dissociated from individual physical assets in the day ahead. However, the system operator *is* concerned with individual physical assets, because it needs to ensure that, as real time approaches, the system can be operated securely. Which is why portfolio positions are eventually disaggregated into individual physical unit schedules as real time approaches. These paradigms are discussed and compared in further detail below.

Before advancing to a detailed discussion on unit versus portfolio bidding, we remark on the fact that bidding product definitions have evolved notably in both the unit-based and portfolio-based designs. Indeed, bidding product definitions are continuously adapting in order to keep up with the evolution and increasing complexity of power system resources. Evidence of this evolution in the case of unit-based systems is the increased detail by which multi-stage combined cycle units[44] are represented in unit-based systems. Similarly, evidence in the case of portfolio-based systems, which also reflects the increasing importance of storage resources is the introduction of looped block orders which were introduced, for instance, in the European power exchange in 2018. We comment on these evolutions and their interplay with the computational complexity of the underlying algorithms subsequently in the report[45].

---

[44] Papavasiliou, A., He, Y., & Svoboda, A. (2015). Self-Commitment of Combined Cycle Units under Electricity Price Uncertainty. *IEEE Transactions on Power Systems*, 1690-1701.

[45] The evolution in the use of different bidding products in the European market is publicly available, see for instance page 25-27 of the CACM Annual Report, https://www.nemo-committee.eu/assets/files/cacm-annual-report-2022.pdf.

3.3.1 Unit versus portfolio bidding

Unit bidding systems are systems in which individual physical assets are represented distinctly in the market clearing algorithm. Once considered an algorithmic dead-end, this is no longer the case since modern optimization technology allows us to represent resources with a fair amount of modelling details.

Typical day-ahead market clearing models span 24 to 72 hours, and aim at scheduling the system over the following day. In those systems where the horizon exceeds 24 hours, the intent is to account for border conditions, make sure that the state in which the system finds itself when entering the beginning of the following day is accounted for (which justifies representing some hours before the first hour of the day for which the system is scheduled) and that the system is not scheduled in a greedy way towards the end of the day in question (by accounting for hours beyond the end of the day).

Nodal market clearing requires knowing where a resource is physically located in the grid. Therefore, unit-based models are sufficient for allowing an implementation of nodal pricing. Although unit-based systems are not necessary for nodal pricing, aggregation in portfolio-based systems with nodal pricing would only be possible at the level of an individual transmission bus. Unit-based systems are thus “almost necessary” for implementing nodal pricings at least at the level of multi-MW-scale assets. On the other hand, one can imagine that future unit-based systems with nodal pricing are likely to aggregate kW-scale resources to equivalent portfolios that adhere to a MW-scale unit-based resource, as long as these resources are located at the same high-voltage system bus.

Portfolios are aggregations of resources which are represented through a unique market offer in an electricity exchange. The motivation of relying on portfolios in electricity market design is to pass over the complexity of detailed technical and economic constraints to asset owners, rather than relaying their management to the electricity market operator.

Since resources within a portfolio cannot be distinguished, it is important to disaggregate market schedules following market clearing. This is performed at the nomination stage, where the production schedules of individual assets are announced to the system operator. The system operator can then check that the proposed schedules respect system security and can ensure that ancillary services requirements can be covered. In case security is compromised from the proposed nominations, the system operator can ask for resources to be re-dispatched in advance of real time.

Portfolios do not necessarily imply a sequential clearing of energy and ancillary services, nevertheless the two co-exist in numerous European markets (such as GB, Belgium, France, Germany, Austria or the Netherlands). On the other hand, there also exist unit-based systems with a sequential clearing of energy and ancillary services (such as Greece, where the day-ahead energy market is cleared first, followed by an integrated scheduling process where reserve is scheduled and priced – with similar setups occurring in Italy and Cyprus). It is worth noting that energy and ancillary services are represented jointly in the Integrated Scheduling Process (ISP) of Greece and Cyprus, even if only ancillary services are traded at this stage. The motivation for doing so is to better account for the interdependencies of energy and ancillary services at the ISP stage, even if ISP schedules and trades ancillary services only.

3.3.2 Multi-part bids versus simple bids

Multi-part bids

Multi-part bids are closely associated to unit-based systems, because the philosophy of unit-based systems is to request from market participants the precise technical and economic information that is needed for representing individual resources. For example:

- Multi-part bids for thermal units include (i) minimum load costs, (ii) startup costs, which can be temperature-dependent, (iii) a non-decreasing multi-segment

merit order curve (e.g. consisting of 10 segments, as for instance in CAISO), (iv) minimum up/down times, (v) ramp rates, (vi) startup and shutdown profiles.

- Multi-part bids for more advanced thermal unit models, such as multi-stage combined cycle gas units[46] are defined based on the idea to represent the different states in which a multi-stage unit can operate (e.g. with various combinations of steam and heat turbines), with each configuration corresponding to its own technical and economic information, and with constraints and costs associated to the transition from one configuration to another.
- One can similarly envision multi-part bids for storage of pumped hydro resources. The required information for populating unit commitment models for such resources includes maximum pumping and production capacity and energy storage limits.

Market clearing in systems with multi-part bids is complicated since multi-part bids typically include non-convex technical constraints and costs. Non-convexity is a mathematical property that corresponds to “jumps” in operating constraints or costs. For instance, the fact that a unit is on or off is a non-convexity. Fixed costs that are incurred when a unit is fired up are also examples of non-convexity. Technical minima are also non-convexities, because once a unit is turned on it has to “jump” at least to its technical minimum in order to operate properly. The problem with non-convexities is that they imply that resources also “jump” when they respond to smooth changes in price signals. For instance, if price exceeds a certain threshold whereby a unit is able to fully recover its fixed startup cost plus variable cost, then the unit is willing to “jump” from not operating at all to operating at its technical maximum, which means that we can find ourselves moving from a situation where supply is not enough to cover demand (under-supply) to a situation where supply strictly exceeds demand (over-

[46] Papavasiliou, A., He, Y., & Svoboda, A. (2015). Self-Commitment of Combined Cycle Units under Electricity Price Uncertainty. *IEEE Transactions on Power Systems*, 1690-1701.

supply). This is a problem, because it means that we may not be able to find prices that clear the market, in contrast to convex/well-behaved market clearing models, in which resources can be regulated smoothly upward or downward through smooth changes in market clearing prices, to the point where supply exactly matches demand.

Various ways are proposed for coping with the inexistence of market clearing prices, as already touched upon in section 2.5, and different system operators in the US have opted for different approaches. Convex hull pricing[47] has been adopted (approximately) in MISO. The idea of the approach is to develop the closest possible approximation to a well-behaved model of a convex economy. The relative advantages and disadvantages of these alternative pricing methods are discussed in section 3.4. An approach based on fixing the integer variables of the market clearing model and re-running a well-behaved pricing problem, referred to as "integer programming (IP) pricing[48]" has been adopted in CAISO. It is also under consideration as a possible way forward for the evolution of the European day-ahead market clearing design. Finally, an approach based on a pricing run that replaces binary (on-off) variables with their continuous relaxation (i.e. pretending that on-off decisions can be fractional), referred to as the "linear programming relaxation" is adopted in PJM. It is worth noting that the existing European design is not immune to these challenges, since European power exchanges also feature complex products with non-convexities, as discussed further below. The approach adopted in the European market is based on the principle of tolerating the rejection of market orders that would have made a profit given the posted price (these are referred to as paradoxically rejected orders), but not tolerating orders that stand to make a loss at the posted price if they are accepted (that is, there are no paradoxically accepted orders).

---

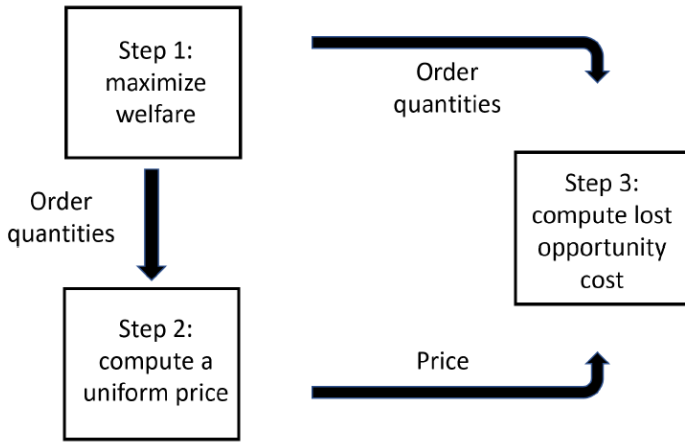
[47] Hogan, W. W., & Ring, B. J. (2003). *On minimum-uplift pricing for electricity markets.* Cambridge, MA: Harvard Electricity Policy Group.

[48] O'Neill, R. P., Sotkiewicz, P. M., Hobbs, B. F., Rothkopf, M. H., & Stewart, W. R. (2005). Efficient market-clearing prices in markets with nonconvexities. *European journal of operational research, 164*(1), 269-285.

Co-optimization in a design that features multi-part bids is a straightforward extension of the energy-only design. Specifically, the detailed technical and economic parameters that are required for populating a unit commitment model (such as ramp rate limitations or economic costs related to the provision of ancillary services) are included in the multi-part bid. The market clearing engine then solves a co-optimization model in order to decide on the matching of energy and ancillary services offers, and runs a separate pricing model in order to compute market clearing prices for energy and ancillary services. The process is depicted graphically in Figure 8. The inclusion of ancillary services in the model poses no fundamental complication. The market model is extended to a multi-product auction that simultaneously trades energy and ancillary services, and where the ancillary services prices are retrieved from the dual multiplier of the market clearing constraint that matches the total supply of ancillary services to its total demand. The generalizations to multiple ancillary services products are straightforward. The co-optimized model is guaranteed to furnish better prices for "higher-quality" ancillary services (i.e. with shorter full activation time), which is appropriate in terms of incentives and preventing price reversals[49].

[49] Oren, S. (2001). Design of Ancillary Service Markets. *Proceeding of the 34th Hawaii International Conference on Systems Sciences HICSS 34.* Maui, Hawaii: HICSS.

Figure 8: The two-step process that is used for clearing markets with non-convex constraints and costs. Step 1 matches orders, and step 2 computes prices. Source: (Papavasiliou A. , Optimization models in electricity markets, 2023).



Simple bids

As explained in section 2.6, the clearing of energy and balancing capacity in European markets is performed by different entities, with power exchanges handling the clearing of energy and transmission system operators the clearing of balancing capacity. The philosophy of exchanges that rely on simple bids is that the full complexity of detailed technical constraints and costs that are specific to particular technologies should be handled by asset owners, and that the market should instead clear standardized and fairly simple products. Over time, this original concept has evolved, and European power exchanges nowadays trade increasingly complex products, that aim at replicating the behaviour of unit commitment models. We discuss some of the products that exist[50], from simpler to more complex, and discuss the interplay between these energy-only products and ancillary services products in an eventual co-optimization setting.

[50] NEMO committee. (2020). *EUPHEMIA public description: single price coupling algorithm.*

Aggregated hourly orders correspond to standard increasing marginal cost curves. They can be specified either as linear interpolations between points, or as steps connecting consecutive points (the former resulting in quadratic programming market clearing models, the latter in linear programming clearing models). These are “well-behaved”/convex products, in the sense that if the market only consisted of such products, then it would be possible to always find prices that clear the market without any paradoxical order matches. Complex orders, which include load gradients and bids with minimum income conditions (MICs), are used for representing intertemporal dependencies. Between load gradients and MICs, load gradients are the easier to handle, in the sense that they capture ramp constraints between consecutive periods. On the other hand, minimum income conditions require that an order recuperate a minimum income (consisting of a fixed and variable component) if the order is accepted.

Block orders are take-it-or-leave-it, and thus non-convex. They are defined over a set of time periods, and are characterized by a price and a (possibly time-varying) quantity as well as a possible minimum acceptance ratio. These orders come closer to the physical reality of operating a power plant (which is either turned on or kept off for the following day) but also introduce non-convexity. Block orders can be synthesized into more complex products to capture exactly the logical interdependencies that unit commitment models can capture directly. For instance, linked block orders are pairs of parent and child block orders, where the child is only accepted if the parent is accepted, with various business rules governing the acceptance or rejection of parents and children. This type of product can capture the physical dependency whereby a unit can produce (child) only if it is started up in the first place (parent). This physical condition is directly modelled in a unit commitment model.

Similarly, exclusive groups are sets of block orders the acceptance ratio of which cannot exceed one, which means in simpler words that only one block order can be accepted. This captures the physical reality that, if a plant is to be operated, then it cannot operate simultaneously in multiple different trajectories, but can only follow one

single trajectory in a single day. Again, this operational reality can be modelled directly in a unit commitment model. Richer variants of these basic products exist[51], but we do not enter in further detail here and rather focus the remainder of this discussion on the possible interaction of these products with ancillary services.

One of the first efforts to extend European energy-only markets by introducing ancillary services products to them is documented in the professional literature[52]. The idea that is described in this report is to introduce ancillary services products which are tagged by a "multilateral link" to corresponding energy offers that are bid into the energy platform. The energy products can continue to exist as they do presently, while the ancillary services products can consist in their simplest form of a quantity of an ancillary service bid and the resource id of the associated energy offer. In particular, a separate price is not required as part of the ancillary services bid, because the co-optimization engine is, by design, endogenizing the opportunity cost of booking generation capacity for ancillary services and thereby foregoing proceeds from the energy market. The market clearing engine then translates these ancillary services bids into two types of constraints in the market clearing platform: one constraint that limits the quantity of ancillary services that can be offered per se, while another constraint requires that the matching of the ancillary service order and the matching of the multilaterally linked energy order cannot exceed the total quantity of the linked energy order. This captures the physical constraint that the total capacity of a unit cannot be double-booked in the energy and ancillary services markets.

The plethora of products that have emerged in the European day-ahead energy platform testifies that bidding product definition can become a very complex undertaking. This exercise is only expected to become more complex and challenging

[51] NEMO committee. (2020). *EUPHEMIA public description: single price coupling algorithm.*

[52] N-SIDE; AFRY. (2020). *CZC allocation with co-optimization.* Louvain la Neuve

as we introduce new ancillary services into the market, which make physical interactions even more complex.

3.4 Qualitative assessment of different design choices

3.4.1 Unit-based systems

The fact that unit-based systems can account for precise network models implies that the distinction between balancing and congestion management close to real time becomes irrelevant. There is a variety of benefits that result from this, which are well documented both in theory as well as in practice. Some merits include the lifting of numerous gaming opportunities, better locational investment signals (especially for co-locating future renewable capacity closer to load centres), more secure operation near real time, better utilization of available network capacity, more efficient day-ahead commitment decisions, and easier cross-border sharing of flexible resources near real time.

Another important advantage of unit bidding is that it allows for better market monitoring, since assets are offered at individual unit level. For instance, during the recent natural gas crisis of 2022/2023, there was a desire for certain EU Member States to mitigate the offers of peaking natural gas units in the electricity market, to palliate price spikes in the electricity market. This undertaking was fairly straightforward in unit-based EU Member States, since it was possible to directly mitigate the offers of natural gas units. By contrast, in Member States with portfolios it was rather challenging to associate the different segments of market offers to individual natural gas units. Insofar as market monitoring is concerned, it is also important to underline that there is a fundamentally different philosophy applied in European and US markets. Whereas US markets rely on ex ante market monitoring

(whereby offers are checked for consistency *before* being cleared in the market[53]), European markets rather rely on ex post market monitoring, which is governed by the Regulation on Wholesale Energy Market Integrity and Transparency (REMIT).

One disadvantage that is quoted in European unit-based systems is that domestic transmission system operators find it difficult to interface with the rest of the European market, which is predominantly a portfolio-based market with standardized exchange products.

Co-optimization of energy and ancillary services is straightforward in unit-based systems. The basic energy-only model (a security constrained day-ahead unit commitment or its real-time analogue, the security constrained economic dispatch model) can be enhanced to include ancillary services of multiple types in both directions (upward/downward). There are various appealing aspects to the unit-based model, insofar as the co-optimization of energy and ancillary services are concerned. (i) The model endogenously accounts for the complex interaction of energy and ancillary services (by introducing constraints which indicate that the capacity of unit must be allocated either to energy or to ancillary services, but not both). (ii) The model also accounts for the one-way substitutability of ancillary services products[54], whereby the allocation of resources accounts for the fact that technologies which are fast enough to offer ancillary services of shorter full activation time (e.g. Frequency

[53] US markets employ two types of market power mitigation measures (United States Federal Energy Regulatory Commission, 2014). One is structure-based mitigation (applied, for instance, in PJM and CAISO) and the other is mitigation based on conduct and impact. The idea of structure-based mitigation is to find the three largest counter-flow producers on a congested line, and to perform a check of removing these generators from the dispatch. If the system turns out to be infeasible, then the offers are mitigated, otherwise they are not. The idea in mitigation based on conduct and impact is to perform a two-step approach, where a conduct test is followed by an impact test. In the conduct test, it is checked whether an offer is exceeding its reference level by the minimum of $100 and 300% of its benchmark marginal cost. The impact test then checks whether a resource offer raises the clearing price by the minimum of $100 and 200%. If both checks are positive, then the offer is mitigated.

[54] Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

Containment Reserve) can also offer ancillary services of longer full activation time (e.g. Frequency Restoration Reserve and Replacement Reserve). (iii) Price formation in the model captures the complex interdependencies of one-way substitutability, and is therefore guaranteed to result in equilibrium market clearing prices. In other words, ancillary services of shorter full activation time are guaranteed to pay better. This is an important consideration, because adverse effects are avoided, such as the price reversals[55],[56] that were observed in some of the early contemplated designs in CAISO. (iv) More complex interactions between ancillary services and operational constraints such as ramping, and startup and shutdown profiles, can be captured straightforwardly and in a fair amount of detail in unit-based models[57].

3.4.2 Portfolios

Portfolio bidding has become a dominant paradigm in Europe. An advantage of portfolios is that they enable market participants to quickly adapt to changing circumstances (e.g. lower wind than forecasted) and choose at the time of delivery which assets of their portfolio are available and most efficient to use.

Portfolio-bidding also provides market participants (with diversity in their fleet) with greater flexibility to manage risk. For example, market participants often part-load a flexible plant to manage the risk of losing another plant in their portfolio due technical

---

[55] Oren, S. (2001). Design of Ancillary Service Markets. *Proceeding of the 34th Hawaii International Conference on Systems Sciences HICSS 34.* Maui, Hawaii: HICSS.

[56] Price reversals are phenomena that occur in sequential reserve market designs whereby markets for lower quality reserves exhibit higher prices. This can happen in sequential clearing arrangements because sequential market clearing fails to account for the one-way substitutability of production factors (fast versus slow technologies). Price reversals can be a serious problem, because they induce agents to pull out of the markets for higher quality reserves, and thus deplete the system operator from access to valuable reserve resources. Co-optimization models are, by design, ensured to prevent price reversals from occurring, because price reversals inherently violate economic equilibrium conditions, whereas co-optimization is guaranteed to respect economic equilibrium.

[57] Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

problems (i.e. unplanned outage). In this way, they reduce their exposure to imbalance penalties.

Moreover, portfolios with simple bids better enable the participation of new technologies and decentralized resources compared to unit bidding with complex bids, where a central clearing algorithm and associated bidding information may need to be updated to account for new types of resources or upgraded to deal with a significant increase in the number of participating resources[58].

Original inceptions of electricity exchanges featured only a subset of the currently increasing spectrum of products. Modern power exchanges are adopting increasingly complex products, some with pronounced non-convex take-it-or-leave-it features. These non-convex features reflect the technical reality that power units operate either on or off, and introduce pricing complications, as discussed in detail previously.

A summary of the discussion about portfolio and unit bidding of these two sections is provided in Table 7.

Table 7: Qualitative Analysis of Portfolio Bidding and Unit Bidding

	Portfolio Bidding	Unit Bidding
Definition	Portfolios are aggregations of resources which are represented through a unique market offer in an electricity exchange.	Unit bidding systems are systems in which individual physical assets are represented distinctly in the market clearing algorithm.
Motivation	Pass over the complexity of detailed technical and economic constraints to asset owners, and according to market participants,	Link bids to physical assets to better represent the economics of the power system and enable

[58] Ahlqvist V., Holmberg P., Tangerås T., (2022). A survey comparing centralized and decentralized electricity markets. Energy Strategy Reviews, Volume 40.

	allow for “flexible bidding strategies”.	an ex-post market monitoring process.
Co-optimization	In theory, co-optimization of energy and reserves with portfolio bidding could be technically achievable, however, further analysis would be required to overcome limitations.	Appealing for co-optimization by allowing a granular modelling of the linkages between energy and reserve provision (e.g substitutability).
Nodal/zonal market clearing	Aggregation in portfolio-based systems with nodal pricing would only be possible at the level of an individual transmission bus. Future unit-based systems with nodal pricing are likely to aggregate kW-scale resources to equivalent portfolios that adhere to a MW-scale unit-based resource, as long as these resources are located at the same high-voltage system bus. Zonal systems can operate under a regime of portfolio bidding and this design is implemented in the majority of the European markets.	Unit-based systems are “almost necessary” for implementing nodal pricing, at least at the level of multi-MW-scale assets. Zonal systems can operate under a regime of unit bidding and there are some European markets that have opted for such an arrangement, such as Spain.
Economic efficiency	Aggregation enabled by portfolio bidding leads to a less granular representation of the technical and economic characteristics of the units, which in turn can result in suboptimal allocations. However, portfolio bidding enables market participants to quickly adapt to changing circumstances (e.g. lower wind, plant trip, etc.) and use the least-cost resources they have available to deliver electricity. This	Improved economic efficiency thanks to a granular representation of the economics and technical constraints of the underlying units. However, it does not enable market participants to adapt to changing circumstances (e.g. lower wind, plant trip, etc.) and use the least-cost resources they have available to deliver electricity. This translates into higher risk per unit, which is

	lowers risk for market participants (that would have otherwise been passed onto consumers) and contributes to optimal allocation.	effectively passed onto consumers, and possibly suboptimal allocative efficiency.
Computational tractability	Algorithmically easy to handle, although the price-based requirements usually used in this context (e.g., the “no paradoxically accepted block conditions” in Europe) represent a computational burden.	Assuming some level of flexibility in the pricing requirements, modern optimization technology is able to cope with the complexity that comes with unit-bidding (i.e. represent resources with a fair amount of modelling detail).
Market Monitoring	Market monitoring (especially ex-ante), is more difficult with portfolio bidding compared to unit bidding, as the economic and technical characteristics are not detailed separately for each unit contained in the portfolio.	Easier, as the bid prices and technical assets should be directly related to the economic and technical characteristics of specific units.

3.4.3 Multi-part bids

Paradigms

As far as pricing methods for coping with non-convexities are concerned, there is a debate about the relative merits and disadvantages of alternative paradigms. Convex hull pricing is the closest possible mathematical approximation to a convex model of the economy, and as a result of this property it produces prices that come as close as possible to satisfying the ensemble of market participants (in the sense of minimizing total lost opportunity cost, which is the payment that market participants would need to receive in order to be indifferent between following the market schedule and self-scheduling their units). On the other hand, exact convex hull prices can be challenging to compute, and they can produce certain counter-intuitive phenomena, such as price

differentials between locations that are not congested[59]. This stems from the fact that, in an attempt to adjust prices in order to keep total lost opportunity cost low, one may introduce a lost opportunity cost for the network operator (reflected in the phenomenon above) in order to save more on lost opportunity cost of other market participants. Another pricing behavior of this approach that is deemed counter-intuitive is the fact that offline units can set the price. In contrast to convex hull pricing, integer programming pricing and linear relaxations are much easier to compute. But relative to integer programming pricing, the linear relaxation is better placed to account for fixed costs of units (even if not fully). Thus, integer programming pricing can exhibit fairly substantial lost opportunity cost in realistic instances[60]. On the other hand, integer programming pricing preserves the pricing business rules of a convex market for the "convex subset" of market participants. Thus, the pricing rules that prevail for simple bids and for the network operator continue to hold, even if the overall market model is non-convex. This means that familiar network pricing rules (such as "no congestion implies equal prices" or "power flows from cheaper to more expensive locations" which hold in transportation-based approximations of power flow models) continue to hold when more complex market products are introduced. This has been considered an important institutional advantage of this approach, which may explain to a certain extent why it is considered a front-runner in a possible evolution of the European market towards so-called non-uniform pricing. The aforementioned comments regarding the pros and cons of these alternative pricing methods continue to hold in a regime of energy and ancillary services co-optimization. Of course, the introduction of ancillary services makes computation even more challenging, and there

[59] Schiro, D. A., Zheng, T., Zhao, F., & Litvinov, E. (2015). Convex hull pricing in electricity markets: Formulation, analysis, and implementation challenges. *IEEE Transactions on Power Systems, 31*(5), 4068-4075.

[60] Stevens, N., Papavasiliou, A., & Smeers, A. (2024). The Many Advantages of Convex Hull Pricing for the European Electricity Auction. *Energy Economics, under review*.

has been a fair amount of effort in extending the characterization of convex hulls of thermal unit feasible sets to the case of ancillary services[61].

The appeal of the European approach to pricing which permits paradoxically rejected orders is that these PRBs are the only deviation from market equilibrium, which is institutionally considered as being acceptable. In addition, because of the absence of side payments, power exchanges are not financially exposed to remunerating such side payments. The approach has been criticized for reducing overall efficiency, since it prioritizes the discovery of a price that satisfies the business rules over the efficiency of the matching (contrast this to the fact that all alternative pricing approaches mentioned in the previous paragraph maintain the welfare-maximizing matching of orders). Moreover, the method is intrinsically computationally more complex than all of the alternatives that are discussed in the previous paragraph, since the matching of orders is not separated from the computation of market clearing prices (as in

Figure 4) but instead the matchings are computed simultaneously with the prices in an intrinsically complex mixed integer quadratic program with complementarity constraints. The introduction of ancillary services can be expected to make matters more challenging computationally, but a prototype of EUPHEMIA which handles fairly simple multilateral links of energy and ancillary services products has been successfully developed[62].

[61] Morales-España, G., Gentile, C., & Ramos, A. (2015). Tight MIP formulations of the power-based unit commitment problem. *ORSpectrum, 37*(4), 929–950.

Morales-España, G., Latorre, J. M., & Ramos, A. (2013). Tight and compact MILP formulation for the thermal unit commitment problem. *IEEE transactions on power systems, 28*(4), 4897–4908.

[62] N-SIDE. (2022). *Co-Optimization of Energy and Balancing Capacity in the European Single Day-Ahead Coupling.* Louvain la Neuve, Belgium: N-SIDE.

Gaming

No mechanism is immune to gaming, and multi-part designs are also susceptible. The way in which multi-part bids can be gamed depends on how market prices are precisely computed and, relatedly, how side payments are set up in order to induce agents to follow market dispatch instructions voluntarily. There is a fair amount of theoretical work on the topic[63], but there are also very interesting empirical observations about precise strategies for gaming multi-part bids[64].

Gaming case in the CAISO and MISO markets

In 2013, the United States Federal Energy Regulatory Commission issued a detailed report about the manipulation of uplift payments in the CAISO and MISO markets by JP Morgan. The gaming of the CAISO and MISO markets took place between September 2010 and November 2012. Following the ex-post investigation of the FERC, JP Morgan accepted to pay a civil penalty of $285 million and disgorge unjust profits of $125 million. The bids concerned gas-fired power plants, of which 4000 MW were located in Southern California, and

---

[63] Fabra, N., von der Fehr, R.-H., & Harbord, D. (2006). Designing electricity auctions. The RAND Journal of Economics, 37(1), 23-46.

Sioshansi, R., & Nicholson, E. (2011). Towards equilibrium offers in unit commitment auctions with nonconvex costs. Journal of Regulatory Economics, 40(1), 41-61.

Liberopoulos, G., & Andrianesis, P. (2006). Critical review of pricing schemes in markets with non-convex costs. Operations Research, 64(1), 17-31.

Wang, G. U. (2012). On Nash equilibria in duopolistic power markets subject to make-whole uplift. 51st IEEE Conference on Decision and Control (CDC) (pp. 472-477). IEEE.

Oren, S. S., & Ross, A. M. (2005). Can we prevent the gaming of ramp constraints? Decision Support Systems, 40(3-4), 461-471.

[64] FERC. (2013). *Make-Whole Payments and Related Bidding Strategies, Docket Nos. IN11-8-000, IN13-5-000.* Washington, DC: Federal Energy Regulatory Commission.

another 545 MW in Michigan. The report describes a number of strategies for manipulating bid cost recovery. These strategies (referred to as strategies A-H in the FERC report) rely largely on the manipulation of uplifts in the two-settlement system of CAISO and MISO. The principle of the two-settlement system employed in CAISO and MISO is to issue scheduling instructions in the day-ahead market, followed by corrective adjustments in the real-time market. The scheduling decisions of the real-time market constitute a subset of the decisions that take place in the day-ahead market. For example, dispatch decisions can still be adjusted in real time, whereas unit commitment decisions are often finalized in the day-ahead timeframe. Market participants have the option of (i) bidding their resources into the market, (ii) self-committing, or (iii) self-scheduling. Units that are bid into the market are guaranteed make-whole payments if the market commitment and dispatch instructions that are issued to the units result in costs that cannot be compensated by market prices. On the other hand, units that **self-commit** fix their unit commitment schedule (but not their dispatch schedule), and are not guaranteed recovery of their fixed costs. Units that **self-dispatch** fix both their commitment schedule and also their production schedule. Similarly to units that self-commit, units that self-dispatch are not guaranteed recovery of their fixed costs through uplifts. For those units that bid into the market, gaming opportunities emerge when the units are allowed to change their real-time offers drastically relative to their day-ahead offers. Such drastic changes in real-time offers relative to day-ahead offers result in apparent unrecovered costs or binding constraints, that sometimes imply uplifts. Many of the gaming strategies employed by JP Morgan involved some kind of interaction in the two-settlement system, where:

- units are forced to be online in the day-ahead market clearing (either by presenting themselves as being very attractive economically, or by self-scheduling their production, or by scheduling the provision of ancillary services which requires these units to be turned on),

- and these same units alter their bids in the real-time market to create dispatch schedules which cannot be fully compensated by prevailing market prices in the aggregate.

Co-optimization

Multi-part bids can accommodate co-optimization straightforwardly. As discussed above with regards to unit bidding, economic and technical parameters related to reserves can be included in a multi-part offer. The idea is to include any additional parameters that are needed for populating a unit commitment model. The entire theory, discussed above, about pricing in models with non-convexities (e.g. convex hull pricing, linear relaxations, and integer programming pricing) can be generalized to a multi-product auction model which simultaneously trades energy and ancillary services. The incentives of agents now need to be aligned across the energy and the ancillary services market.

Computationally, the undertaking is fairly straightforward and is not expected to introduce unbearable computational challenges. Empirically, however, we can report that the introduction of ancillary services in realistic cases can cause a sharp increase in the run time of branch and bound commercial solvers (since, even though the problem size does not increase dramatically, the scheduling decisions themselves become much less obvious than when scheduling energy alone, which can pose challenges to implicit enumeration algorithms such as branch and bound).

Moreover, one specific pricing method, that of convex hull pricing, might struggle relative to others in this context, because convex hulls in energy and ancillary services models become more challenging to characterize. An alternative to computing convex hull prices by characterizing the convex hull of the primal feasible set is to work directly

on the dual space through Lagrangian decomposition methods[65], nevertheless the dual space now also includes ancillary services variables, and therefore one can expect a non-trivial effect of ancillary services on the run time of the algorithm.

3.4.4 Simple bids

If electricity markets could operate adequately with simple bids, this would be an excellent way forward: the clearing of electricity markets with simple bids is computationally trivial, and the market model is guaranteed to have equilibrium prices. The problem is that simple bids fail to account for the technical reality of power generating units. Therefore, representing a power generating unit or portfolio of units through simple bids can produce a technically infeasible schedule. As a simple example, consider a thermal unit with a minimum power generation of 80 MW and a maximum of 100 MW which is offered into the market as a simple bid of 100 MW at a certain price. Suppose that this bid is cleared partially at 70 MW. Then, the unit cannot honour its position in the market, and is forced to deviate from its schedule.

In this regime, one also inherits the fact that simple bids do not precisely populate a model of the operation of a specific asset, therefore a fair amount of non-trivial reverse engineering is required by traders who try to match the true physical constraints and costs of their assets to the format of the bidding language.

Complexity of representing technical reality

To illustrate the point of reverse engineering, consider a thermal asset with a minimum up time of 2 hours, which is required to operate in a market horizon of three hours, and suppose that the asset has a minimum power generation equal to its maximum output of 100 MW, a startup cost of 10000 €, and a marginal cost of 50 €/MWh. In a

[65] Stevens, N., Papavasiliou, A., & Smeers, A. (2024). The Many Advantages of Convex Hull Pricing for the European Electricity Auction. *Energy Economics, under review*.

unit commitment model, the multi-part bid consists of the following vector of information: (minimum power output, maximum power output, minimum up time, startup cost, marginal cost). This is all the information that is needed for populating a unit commitment model, and the market clearing algorithm will consider all possible trajectories that the asset could follow given these parameters from which the most efficient one for the market and the asset owner will be picked. By contrast, in a European power exchange, one way to offer the asset would be to construct the following exclusive group of block orders:

- Block 1: (100 MW, 100 MW, 0 MW) at 100 €/MWh
- Block 2: (0 MW, 100 MW, 100 MW) at 100 €/MWh
- Block 3: (100 MW, 100 MW, 100 MW) at 83.33 €/MWh

The point of the example here is that the asset owner has to pre-calculate all the possible ways in which an asset can operate contiguously over at least two hours in a three-hour interval, as well as the corresponding total cost of each of these trajectories. If this reverse-engineering exercise is workable in this example's setting, in the context of a 48-period horizon (in a 24-hour market horizon with 30-minute market clearing intervals as in GB), the number of combinations can explode. The appealing aspect of a unit commitment model is that the "hard work" of scanning over all these possible trajectories is performed endogenously by the algorithm, and does not need to be defined as an input by a human/trader.

Bidding language

On a related point, it is not clear that an appropriate bidding language even exists for certain products. For instance, before the introduction of loop blocks, it was not apparent how one could trade storage resources in the day-ahead market. Of course, one can always innovate in terms of defining new exchange products that are specifically designed to tailor the needs and constraints of specific physical assets. But this process takes time for research and development, prototyping, and stakeholder agreement. This is also the case for unit commitment models based on

multi-part bids, which face a similar critique about the fact that introducing models for specific assets in the market clearing model is a time-consuming and laborious process. There is no easy way around this challenge, and designs inspired by simple bids do not necessarily hold an advantage with respect to this challenge relative to designs based on multi-part bids. However, a portfolio-based design may somehow "hide" the problem as the trader can cope with the flexibility of its portfolio to account for the elements which are not precisely modelled/modellable by the available bids. This may create a competitive advantage for larger market participants. In order to ensure a level playing field for flexible resources such as storage and demand response in the future, it is imperative that bidding language innovations keep up with the rapid integration of these resources.

Concerns over the expressiveness of the bidding language carry over to the issue of introducing ancillary services in exchange models. We commented above on how ancillary services can be introduced in models of portfolio-based markets, with the notion of multi-lateral bid linking between energy and reserves[66]. This idea of multi-lateral linking can be merged with the idea of enumerating trajectories for feasible unit commitment sequences through groups of exclusive block orders. The idea could then be to introduce reserves and energy in portfolio-based markets with simple bids by multilaterally linking reserve bids to energy bids that belong to mutually exclusive groups of block orders, with each block order corresponding to a different feasible trajectory of a resource over the day. This is an easily achievable starting point of a path towards introducing reserves into these models, but one could reasonably flag some limitations of this approach (e.g. how this could scale up to aggregations of resources into portfolios, the high number of combinations that would be needed for expressing a fair amount of possible trajectories, and so on). The point being made

[66] This notion was first articulated in (N-SIDE; AFRY, 2020).

here is that a path forward is possible, but that numerous issues related to bidding product definition would need to be resolved.

Table 8 summarizes the outcomes of the previous discussion about differences between multi-part and simple bids.

Table 8: Qualitative Analysis of Simple and Multi-Part Bids

	Simple Bids	Multi-Part Bids
Definition	Price-Volume Pairs, possibly enhanced with simple indivisibility requirements but without detailed description of techno-economic conditions.	Request from market participants the precise technical and economic information that is needed for representing individual resources. For thermal units include (i) minimum load costs, (ii) startup costs, which can be temperature-dependent, (iii) a non-decreasing multi-segment merit order curve (e.g. consisting of 10 segments, as for instance in CAISO), (iv) minimum up/down times, (v) ramp rates, (vi) startup and shutdown profiles.
Motivation	Easier to understand bidding products and more tractable computations depending on the pricing rules used. Enables aggregation of resources and portfolio bidding.	Improved economic efficiency by representing directly the actual economic and technical characteristics of the assets.
Co-optimization	In theory, co-optimization of energy and reserves with portfolio bidding could be technically achievable, however, further analysis would be required to overcome limitations.	Co-optimization in a design that features multi-part bids is an extension of the energy-only design: it is straightforward to model the linkages between energy and ancillary services provision (e.g., substitutability).

Pricing	Used in Europe in combination with strict linear pricing (no side payments), though pricing rules using side payments are considered.	In principle require pricing with side payments, to enable efficient allocation while avoiding losses for market participants, and for computational tractability.
Economic efficiency & Feasibility	Representing a power generating unit or portfolio of units though simple bids can produce a technically infeasible schedule. Also, reduced bid expressiveness (i.e. actual cost structure and the technical constraints of the assets) could lead to efficiency losses.	More efficient from an economic point of view as the actual cost structure and the technical constraints are better represented.
Computational tractability	Computationally more tractable compared to multi-part bids.	Market clearing in systems with multi-part bids is complicated by the fact that multi-part bids typically include non-convex technical constraints and costs.
Market Monitoring	Less granular information on the specific economic and technical characteristics of the underlying asset(s).	More granular information on the specific economic and technical characteristics of the underlying asset(s).

4. Locational Considerations

4.1 Takeaways

→ Where network constraints are considered in price formation (e.g. locational markets), co-optimization allocates transmission capacity to the products (energy, reserve and response) that create the most value (i.e. maximize social welfare).

→ Allocating transmission capacity across energy, reserve and response implies additional conceptual and computational challenges, compared to the allocation of transmission capacity implicitly derived only from energy products. This is because the actual activation patterns of the reserve and response products are unknown at the time of the auction. The auction results must guarantee that balancing capacity procured across transmission constraints can be delivered, irrespective of how these products are activated in real time.

→ Technical solutions have been developed to ensure that a strict “reserve deliverability requirement” (so called “deterministic requirement”) is respected.

→ Alternatively, a “reserve deliverability requirement” may not be strictly applied. Where this the case, system operators often rely on statistical approaches (i.e. sufficiently probable that the procured ancillary services will be deliverable) or focus on specific scenarios (e.g. deliverability of ancillary services in case of significant contingencies).

4.2 Overview

This chapter provides an overview of the most important design elements to take into account when considering the co-optimized procurement of energy and ancillary services in the presence of thermal transmission network constraints[67]. Referring to Table 3, this section focuses on the interaction of reserves and transmission in forward markets. It introduces the concept of reserve deliverability, i.e. how it can be ensured that any pattern of reserve activation is "feasible" in terms of congestion in real time. As soon as ancillary services are procured under transmission constraints (whether under "co-optimization" or separately from energy), how to address reserve deliverability becomes an important aspect.

It then discusses what are the implications of different approaches in terms of allocative efficiency and overall procurement costs (including redispatch costs).

For illustrative purposes, this chapter presents examples with two price zones (simplified granularity) separated by one ATC-based transmission constraint (simplified grid flow modelling) which applies equivalently to both energy and reserves (simplified consistency). Also, co-optimization is illustrated with a single balancing capacity product for upward delivery of energy. While this is a simplified example, the observations presented below are generic in nature and can remain applicable to more sophisticated setups.

[67] The focus is on internal GB constraints; cross-border transmission capacity allocation is not within scope.

4.3 Key notions

There are different approaches to consider transmission constraints in co-optimized markets. They are influenced by the broader applicable market design context, including:

- **Granularity:** This attribute relates to the level of granularity that is used for representing the transmission grid in price formation. One approach consists of not reflecting transmission constraints at all in price formation ("copper plate model"). When transmission constraints are not considered, the market clears at a single price. Another approach is to consider each physical connection node and transmission line of the grid separately, which results in each node having a different price (i.e. nodal models). The models in-between these two approaches only consider the most important (structural) transmission constraints, and aggregate the physical nodes that are geographically close to each other into zones within which all nodes are priced equally, i.e. "zonal models".

- **Consistency regarding granularity:** Grid models may also differ between different trading timeframes or traded products. For example, fewer transmission constraints may be considered in ancillary services markets compared to those taken into account in the wholesale energy market.

- **Grid flow modelling:** The way the market model represents electricity flows is particularly important in Alternating Current (AC) meshed networks. Broadly speaking, there exist two main approaches for representing networks in zonal models: Available Transmission Capacity (ATC) and Flow-based. Under ATC-based models, electrical flows between price zones are independent from each other and limited by ATC values. Under flow-based models, the inter-dependencies between the flows of meshed AC grids are modelled in a way

that is (in principle) closer to the actual physical reality as they intend to approximate Kirchhoff's laws.

A key notion that we introduce below is **"reserve deliverability"**. Reserve deliverability relates to the challenge of trading reserves on a network with transmission constraints while ensuring that the network will be able to support (some or all) activations of energy from these reserves in real time.

Insofar as reserve deliverability is concerned, it is worth reflecting on why this issue is a lively topic of ongoing discussion at this juncture in time. An important driver in terms of timing is the fact that the coupling of European balancing capacity markets across national borders is the next step of evolution of European electricity markets, and the timelines are dictated by EU legislation. Day-ahead energy market coupling is largely considered complete in Europe, and although many adaptations to the basic design are being pursued, the attention is now shifting to real-time energy/balancing market coupling (e.g. through the PICASSO and MARI platforms) as well as day-ahead coupling of balancing capacity. It is precisely the issue of day-ahead coupling of balancing capacity across national borders that motivates the issue of deliverability of balancing capacity.

There exist many different flavours of reserve deliverability. We focus primarily on the "deterministic requirement", where any set of possible activations of procured reserve is expected to be feasible in real time (in the sense that any possible activation pattern must respect all modelled network constraints independently of any stochastic/probabilistic considerations). To do so, we need to focus on preparing the system for the least favourable realization of uncertain outcomes. In this sense, the requirement is based on the notion of *robust optimization*[68]. The problem of reserve

[68] As opposed to (1) enumerating some scenarios – as is the case in the context of the below-mentioned US two-stage models – and without (2) acknowledging a certain level of "non-availability" due to transmission constraints in real-time as in the ALPACA approach mentioned below.

deliverability under the “deterministic requirement” is meaningful in markets with multiple reserve products and multiple zones where the procuring entity wants to secure that the capacity procured across various geographical zones will respect the grid constraints for whatever energy activation pattern of this capacity. This is notably the approach followed by ENTSO-E in its “methodology for a co-optimized allocation process of cross-zonal capacity for the exchange of balancing capacity or sharing of reserves in accordance with article 40 of EBGL”. This deterministic requirement towards reserve deliverability is arguably driven by the aggregate representation of reserve exchange in European market clearing models. Concretely, and as we explain below, in US systems the treatment of reserve deliverability benefits from the representation of individual units and detailed networks in the market clearing model. This makes it possible to represent outages of critical generating units and individual network elements as explicit constraints in market clearing models, with the explicit demand within the model that the system should be dispatched in a way that it is able to serve load even in the case of outages of single components. Such a granular representation of units and networks is not possible/preferred in European market design. This may explain, in some level, why the deterministic requirement has been used instead of explicit security constraints as a method for accounting for reserve deliverability in European market clearing formulations.

A very different approach is contemplated in the ALPACA project, a (voluntary) initiative establishing a cooperation for the procurement of automatic Frequency Restoration Reserve (aFRR) between Austria, the Czech Republic and Germany (with Hungary, Croatia, the Netherlands and Slovenia being observers). In this project – which focuses exclusively on the exchange of aFRR (and therefore does not consider co-optimization with energy) – the available transmission capacity is not explicitly allocated to balancing capacity. Instead, the cross-zonal reserve procurement mechanism is based on a probabilistic approach (in accordance with Article 33(6) of EBGL) which accepts that reserves may be procured from other zones even though there is no certainty that the procured reserve will effectively be deliverable in real-time (for example because the transmission grid is fully utilized for energy trades). The

procurement mechanism instead relies on forecasting techniques to assess "Maximum Exchange Limits", i.e. levels of transmission capacity that are available with a sufficiently high probability such that cross-zonal procurement – even if not fully reliable – nonetheless reduces the total balancing procurement costs. Note that there is no precedent for the application of such a methodology.

For other parts of the world, the issue of reserve deliverability is discussed in the literature that concerns US markets[69] (in particular ISO New England). In the context of these markets, the concern is for system operators to be able to respond to specific contingencies of system components. This is a widely adopted approach for dealing with reserve deliverability in US market clearing models. Such models can be cast as two-stage optimization programs, where reserve is committed in the first stage, a contingency occurs, and the system then responds in the second stage by activating reserve that has been committed in the first stage to respond to the contingency, while ensuring that the system is not overloaded. A key difference between these contingency-constrained formulations and the deterministic requirement is that the contingency-constrained formulations account for a finite number of alternative uncertainty realizations which relate to component failures, whereas the deterministic requirement accounts in principle for infinitely many uncertainty realizations which relate to any possible level of activation of transmission system operator demand. The latter turns out to be harder to handle computationally, and practically viable approximations of this requirement have been proposed by N-SIDE.

A detailed analysis of the most appropriate "reserve deliverability" implementation to be applied in the GB context should be carried within the broader assessment of the

---

[69] Zheng, T., & Litvinov, E. (2008). Contingency-based zonal reserve modeling and pricing in a co-optimized energy and reserve market. *IEEE transactions on Power Systems, 23*(2), 277-286.

Chen, Y., Gribik, P., & Gardner, J. (2014). Incorporating post zonal reserve deployment transmission constraints into energy and ancillary service co-optimization. *IEEE Transactions on Power Systems, 29*(2), 537-549.

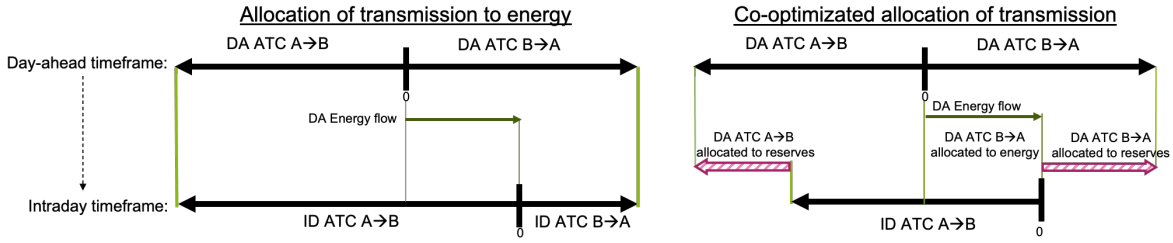
transmission model to be taken into consideration in energy and ancillary services markets. In this document, we assume a deterministic requirement and focus on key impacts of transmission capacity allocation under co-optimization.

4.4 Transmission and balancing capacity

When transmission is allocated to energy, for example in the day-ahead market, a firm flow is expected to follow in a given direction at the time of delivery. Consequently, for subsequent markets, the total available transmission capacity in that direction is reduced by the volume of the firm allocated flow for this direction. In addition, because the flow is firm, it can also be “netted” such that the transmission capacity in the opposite direction can be increased by the same flow volume for subsequent markets. The “total volume of transmission capacity”, i.e. the size of the grid model domain, remains fixed (unless it is reassessed); and the “working point” moves within this domain.

The situation is fundamentally different for balancing capacity. Firstly, transmission can be allocated simultaneously in multiple directions (for example to different products). Secondly, it is unknown at the time of the allocation if the allocated transmission will effectively be used in the balancing timeframe. This means that the transmission allocated to balancing capacity cannot be “netted”. Consequently, the total volume of transmission that is available in subsequent markets is reduced, i.e. withheld as headroom for possible balancing activations. This is schematically represented in Figure 9

Figure 9: Co-optimized allocation of transmission across energy and reserves.



The allocation of transmission capacity to energy implies a firm flow in one direction (left). This flow can be "netted", such that the sum of the transmission capacities in both directions remains equal after the allocation. For the direction opposite to the allocation, more transmission capacity becomes available for subsequent timeframes due to this netting. Under co-optimization (right), transmission capacity can be allocated simultaneously in all directions to different products. The transmission allocated to balancing capacity cannot be netted because it is unknown at the time of allocation whether the transmission will effectively be used or not. Consequently, a bandwidth of the total transmission capacity is "withheld" and hence not available for subsequent markets.

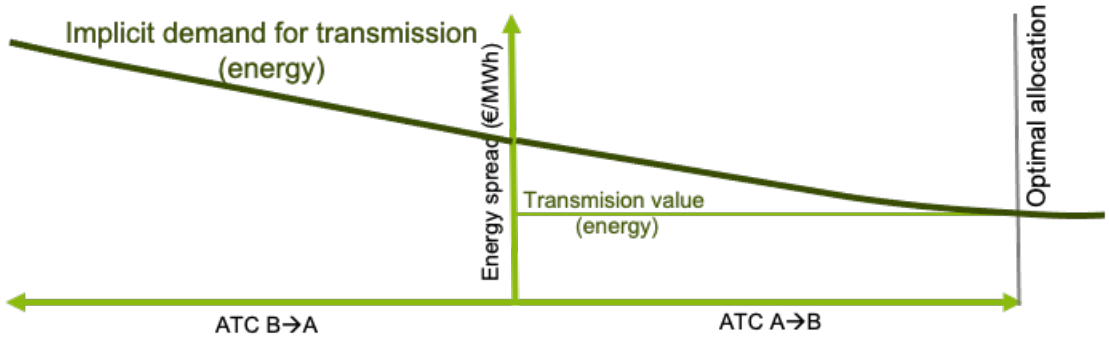
4.5 Optimal allocation of transmission across energy and ancillary services

4.5.1 Incremental value of cross-zonal-capacity

Under market coupling, transmission pricing is such that, when transmission is not scarce, then energy flows freely across the different price zones and no price difference is observed, i.e. the "value of transmission" is zero. When there is insufficient transmission available to allow prices to equalize, then the "value of transmission" is precisely this price difference (see Figure 10). The transmission

operator collects as a congestion revenue this price difference multiplied by the (necessarily congested) cross-zonal flow. Interestingly, the demand for transmission is not explicitly expressed by the market participants. Instead, it is implicitly derived from the offer and demand for energy in the neighbouring energy markets. This is why market coupling is often referred to as an "implicit auction" or "implicit allocation of transmission capacity".

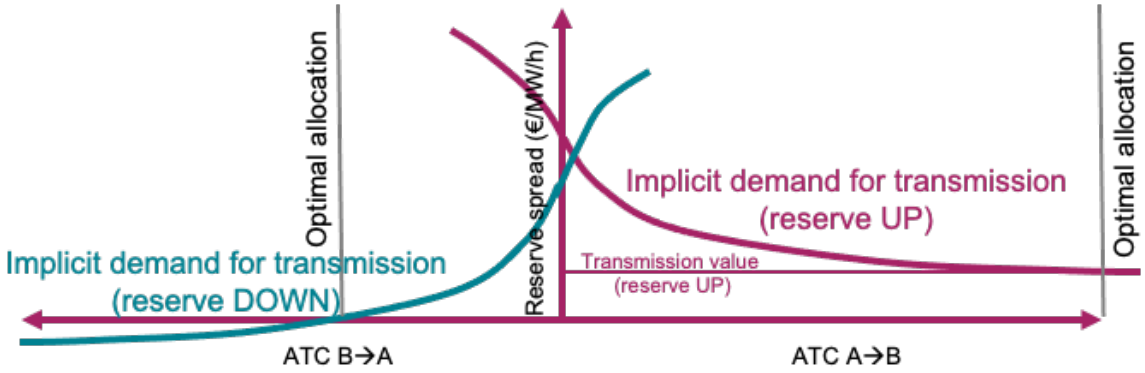
Figure 10: Illustration of implicit allocation for energy.



The same principles apply when balancing capacity is traded implicitly over transmission constraints. However, there is typically more than one product to be allocated. For example, there may be an upward and a downward reserve product, as illustrated in Figure 11. This implies that a given cross-zonal link can be allocated simultaneously in both directions. This is different compared to energy because trading energy necessarily results in a single netted energy flow – whereas different balancing capacity products cannot net out with each other[70].

[70] This also relates to the fact that balancing reserve does not imply a firm energy flow.

Figure 11: Illustration of implicit allocation for upward and downward balancing capacity.



More generally, multiple products can compete for transmission capacity, possibly in the same direction, under co-optimization. The market clearing process therefore needs to apportion the allocation of transmission capacity amongst multiple products.

4.5.2 Optimal split of cross-zonal capacity

A consequence of the increasing value/volumes in reserve markets is the emerging debate on whether to introduce welfare optimization clearing algorithm to split the allocation of cross-zonal capacity. The following section sets out the economic theory as to why this is desirable. A concrete illustrative example is first used in order to demonstrate the principle of the "optimal capacity split" problem[71].

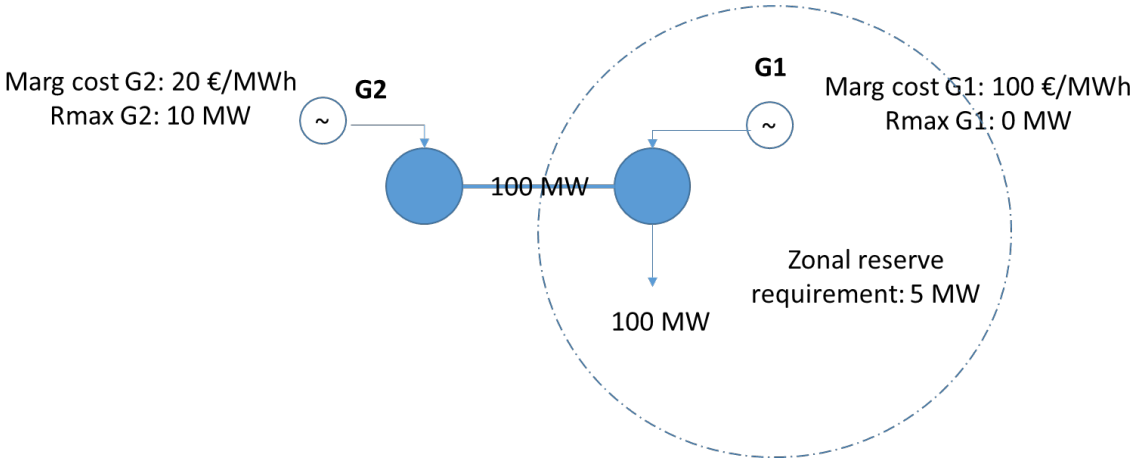
Example 4.1: *the optimal split of cross-zonal capacity equalizes the marginal value of energy and balancing capacity*.

Consider the system of Figure 12. The system consists of two zones that are connected through a limited capacity of 100 MW. The right-hand zone has an inelastic

[71] Explanatory Document to all TSOs' proposal for a methodology for a co-optimized allocation process of cross zonal capacity for the exchange of balancing capacity or sharing of reserves in accordance with Article 40 of Commission Regulation (EU) 2017/2195 of 23 November 2017 establishing a guideline on electricity balancing.

energy demand of 100 MW, and includes a local power generating unit (assumed to be “expensive” and unable to provide reserve capacity). This zone also requires 5 MW of reserve. “Cheap” generation is located in the left-hand zone, which has no load or balancing capacity requirements, but can provide cheap generation and balancing capacity to the right-hand zone.

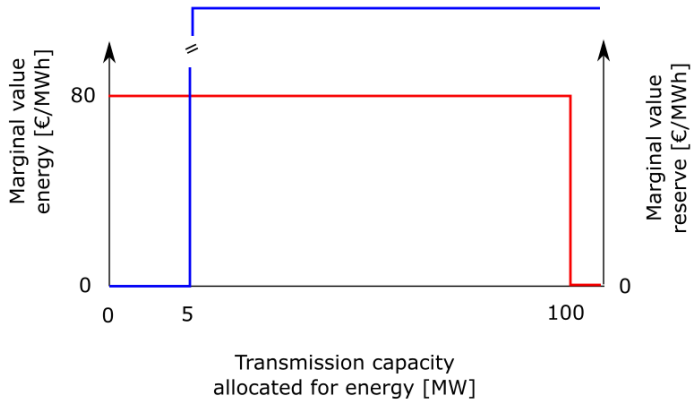
Figure 12: A two-zone system which demonstrates the optimal capacity split problem.



The question that we attempt to address in this example is how the available cross-zonal capacity should be used. Should it be reserved for carrying balancing capacity, or allocated to transmitting cheap energy from the left-hand zone to the right-hand zone?

In the example of Figure 12, a co-optimization of energy, balancing capacity (and the cross-zonal capacity) would result in an allocation of 5 MW of transmission line capacity for balancing capacity, and the remaining 95 MW of the line for transporting 95 MW of cheap generation from G2 to the right-hand zone. An additional 5 MW of power are produced by the local and expensive generation in the right zone. The solution can be understood in terms of the optimal capacity split problem, as indicated in Figure 13, whereby the optimal capacity allocation is such that the marginal value from the sharing of energy intersects with the marginal value for the sharing of reserve.

Figure 13: The optimal split of cross-zonal capacity occurs at the point where the incremental value of cross-zonal capacity for transportation of energy equals the incremental value of cross-zonal capacity for enabling the trade of reserve.



The value of the available transmission capacity for the sharing of energy, up to 100 MW of capacity, amounts to 80 €/MWh (every additional unit of transmission capacity allows the replacement of 1 MW of cheap generation from the left zone with 1 MW of expensive generation in the right zone). The value of the available transmission capacity for the sharing of balancing capacity, up to 5 MW of capacity, is equal to the cost of not meeting the balancing capacity requirement of the right zone. The two curves intersect at the optimal capacity allocation, namely 5 MW. The energy price in the left zone amounts to 20 €/MWh and the energy price in the right zone amounts to 100 € (since the units are not capacity-constrained), while the reserve price in the right zone amounts to 80 €/MWh.

A welfare optimization clearing algorithm will split the allocation of transmission capacity between these products optimally from an economic perspective, and identify the solution which generates the highest economic value. In other words, transmission is allocated to the products that value transmission the most.

This notion is graphically represented in

Figure 14. In this example, it is valuable to allocate transmission in the A→B direction to both energy (upper part of the picture) and upward reserve (middle part of the

picture). The optimal split of the A→B transmission capacity is found at the intersection of the implicit demand curves for transmission of energy and upward reserve. With such a split, the value of the transmission is the same for both products while any other split creates less total welfare. Indeed, allocating more transmission to balancing capacity (i.e. shifting the allocation split to the left) creates additional value on the balancing capacity market, but it reduces more than this value on the energy market because the marginal increment in welfare on the balancing capacity market is lower (i.e. lower transmission price) compared to the marginal decrement in welfare on the energy market (i.e. higher transmission price).

Figure 14: Illustration of implicit allocation of multiple products.

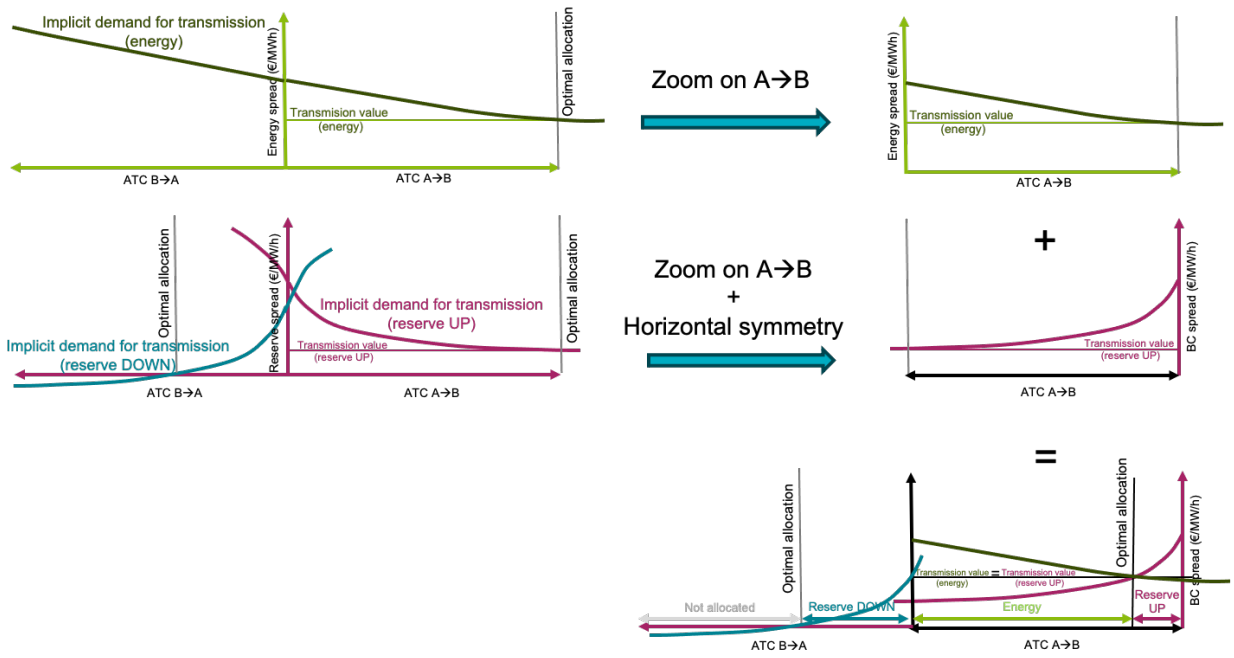
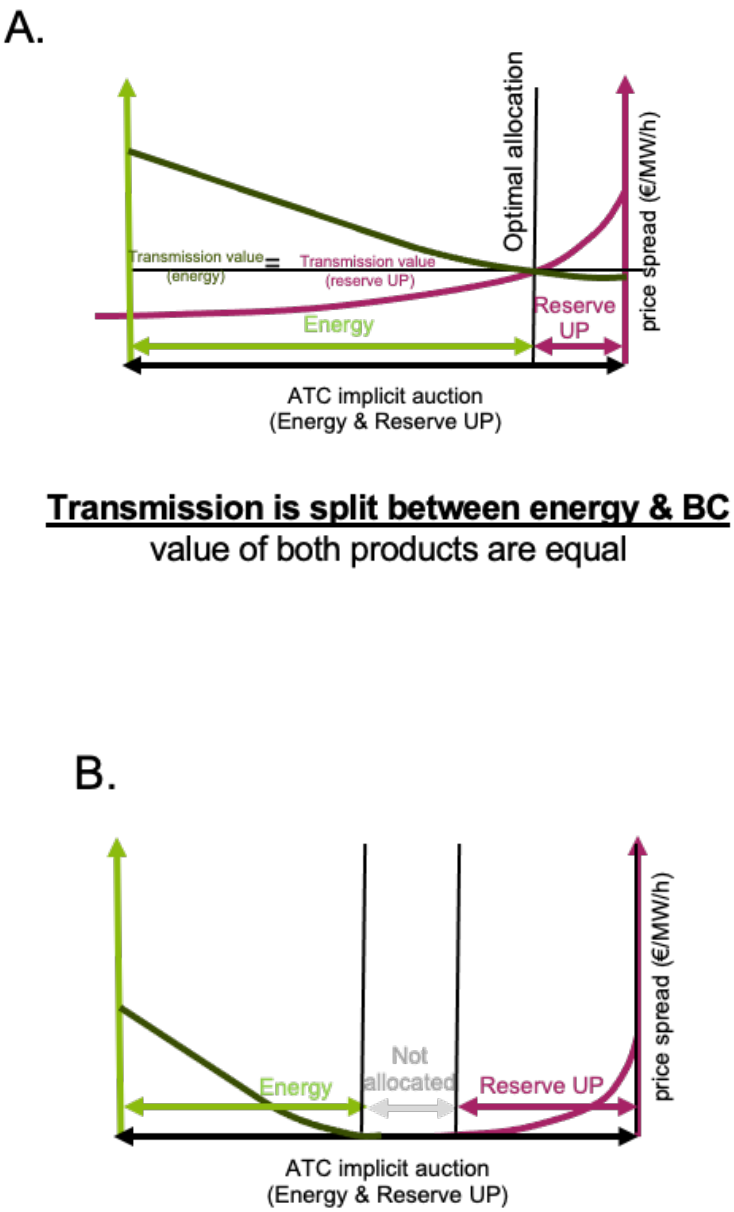


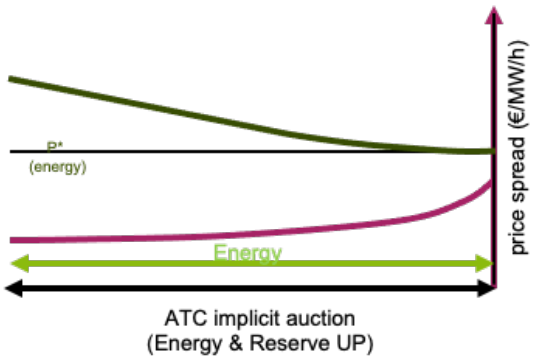
Figure 15 illustrates various other possible allocation patterns.

- Case A splits the entire transmission capacity into several products. As already explained above, this implies that the price difference across this transmission constraint – the value of the transmission – is the same for all products for which transmission is allocated.
- In Case B, all the transmission capacity is only partly allocated, which implies that the price differences across all products is zero (because transmission is not scarce).
- In Case C, all the transmission capacity is allocated to energy because the energy price difference – even if it becomes narrower due to the exchange between the two zones – remains higher than the price difference of upward balancing reserve.
- Case D is particularly interesting because it combines different notions explained above. In this example, the optimal split implies that energy flows against the energy price difference – hence exacerbating it – until it equalizes with the large balancing reserve price difference. Although this may appear as counterintuitive at first glance, this is in effect economically rational. Indeed, flowing energy “in the wrong direction” destroys some value/welfare. However, it also frees up additional transmission capacity due to netting, which can be used to reduce the price spread in the reserve market. Given that this later price difference is higher, the gains obtained in the balancing capacity market are larger than the welfare losses in the energy market, which is why this is optimal.

Figure 15: Various possible allocation patterns and transmission capacity split.

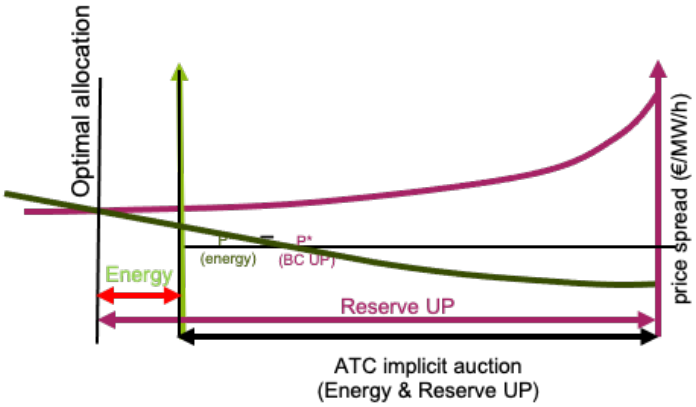


C.



Transmission is entirely allocated to energy
value of energy > value of BC

D.



More than the ATC is allocated to BC
Energy flows « in the wrong direction »
because this frees up CZC for the more valuable BC

4.5.3 Implementation considerations

The optimal split of cross-zonal capacity described so far is achieved endogenously by a well-defined market clearing algorithm that co-optimizes energy, network access, and balancing capacity. The basic formulation generalizes the energy-transmission co-optimization model. Specifically:

- The energy-transmission co-optimization model accepts energy offers from generators, and a network model that is defined by the system operator. The energy-transmission-reserves co-optimization model accepts energy and balancing capacity offers from generators, and the same network model by the system operator.

- In the energy-transmission co-optimization model, generators are paid for selling energy. The energy price is the dual multiplier of an energy balance constraint. In the energy-transmission-reserves co-optimization model the generators are paid for selling energy and balancing capacity. The balancing capacity price is the dual multiplier of a reserve balance constraint.

- Cross-zonal capacity in the energy-transmission co-optimization model corresponds to rights for accessing the network in order to transport energy from energy producers to energy consumers, and these rights need to be bought by the parties that are trading energy. Cross-zonal capacity in the energy-transmission-reserves co-optimization model corresponds to rights for accessing the network in order to transport energy from energy producers to energy consumers, and balancing capacity from generators to the transmission system operator. These rights are required for producers to trade energy with consumers in different locations, and for producers to be authorized to sell balancing capacity to different parts of the network from where they are located.

- The energy-transmission co-optimization model is formulated by using an energy flow variable to represent the use of the network for trading energy between different locations. The energy-transmission-reserves co-optimization model additionally introduces a "balancing capacity flow" variable in the market model.

If the energy-transmission-reserves co-optimization model is properly formulated, the optimal allocation of cross-zonal capacity that is described above is performed endogenously by the market clearing algorithm. In fact, and referring to the principles that are drawn out in chapter 2, the equalizing of the marginal value of cross-zonal capacity for trading energy and balancing capacity is part of the KKT conditions of the market model, i.e. the market clearing solution is guaranteed to satisfy such a fundamental economic property.

The modeling enhancement of a radial market model is fairly straightforward, and loses no generality, in the sense that it captures perfectly the deterministic requirement which dictates that every MW of balancing capacity that is traded between neighboring locations must occupy a corresponding MW of space in the line that is linking the locations. The market clearing model is therefore not burdened significantly. Additional variables are introduced for representing balancing capacity offers, and additional constraints for representing the market clearing conditions in balancing capacity. The representation of the deterministic requirement in a meshed network cannot be captured precisely in a computationally acceptable way, but it can still be approximated in a way that is “as conservative as necessary and not more than that”.

The practical takeaway is that market models which perform effectively in an energy-transmission co-optimization regime stand a very good chance of also coping computationally with the introduction of balancing capacity. There are important caveats to this statement that relate to the specific representation of resources (whether these are unit-based or portfolios, and in each case what specific constraints/bidding products are modeled and at what level of detail), but the

imposition of the deterministic requirement on the network itself is not a fundamental show-stopper from an algorithmic standpoint.

What may be tricker to handle institutionally is the possible ramifications of such a design on market clearing prices. Overall, the system stands to gain from the exchange of balancing capacity over the network. But the new competitive equilibrium that emerges, although welfare-enhancing, is not guaranteed to make *all* market participants better off. In particular, there may be locations/regions which end up exporting balancing capacity that they had in abundance prior to introducing the cross-zonal exchange of balancing capacity. Such regions may experience an increase in balancing capacity prices relative to a regime where balancing capacity is not exchanged. And due to the intricate connections between energy and balancing capacity at market equilibrium (see chapter 5 for a more detailed discussion on the subject), these regions may also face increased energy prices. Speculating on the potential implications of such an evolution on the GB system is probably meaningless, and a detailed market analysis study would instead be a more meaningful way forward for better assessing the impact of such an evolution in market design.

A way of paraphrasing the phenomenon described above is that in a market that co-optimizes energy, transmission and balancing capacity jointly, the transmission system operator would need to reserve capacity for wholesale energy trading based on its assessment of balancing capacity needs. This places a burden on the transmission system operator to perform efficient dimensioning of balancing capacity, in order not to restrict opportunities for trade in the energy market unnecessarily.

4.6 Discussion

The introduction of transmission constraints in the energy market through locational pricing is contemplated in REMA. Co-optimization may be introduced under a zonal or nodal pricing regime, irrespective of whether the grid model applicable to energy and to balancing capacity is the same or not. Although nodal markets clearly outperform

zonal designs in terms of accurately representing the physics of power flow, and are therefore better positioned to ensure the deliverability of energy, it is interesting to point out that existing US nodal markets do not impose a deliverability requirement in the form of the deterministic requirement described in this chapter. This truly appears to be an idiosyncratic yet reasonable requirement of European electricity markets, to the best of our knowledge.

The key difference between energy and balancing capacity products is that – when balancing capacity is exchanged across different price zones, it is not known at the time of the allocation if the transmission capacity will be activated or not. It is therefore not possible to net the trades of balancing capacity in opposite directions, as it is the case for energy.

An important consequence thereof is referred to in this document as the “reserve deliverability” requirement, which relates to the desired of guaranteeing that the balancing energy from the procured reserves will be supported by the grid. Such a requirement can take various forms that have significant conceptual, economic, and computational impacts. In the case of a “loose” deliverability requirement, the cost to procure reserves may turn out to be useless at the time needed because the grid can ultimately not support the activation of the procured reserves. In case the “reserve deliverability” requirement implies some firm guarantees (e.g. deterministic requirement), then some transmission capacity needs to be “withheld” from the energy market and converted into some sort of “headroom” that is kept out of the energy market in order to secure the delivery of procured reserves in real time. Moreover, such headroom could be “inflated” thanks to “energy flows against the price difference” if this is economically justified (cfr. Case D in Figure 15).

5. Real-Time Co-optimization of Energy and Reserve through Reserve Scarcity Pricing

This chapter focuses on reserve scarcity pricing through operating reserve demand curves, which is a real-time mechanism for stimulating investment in tight systems that experience scarcity in flexible generation capacity. Referring to Table 3, this chapter concentrates on the interaction of reserves and energy in real-time markets.

5.1 Takeaways

→ Scarcity pricing is informally defined as the process by which short-term energy prices escalate above the marginal cost of the marginal unit, i.e. the last unit in the merit order to produce power in the market. Scarcity pricing can typically occur under stressed system conditions.

→ Scarcity prices are valuable for ensuring a long-term equilibrium in an energy market, since they allow generators to recover inframarginal rents. These inframarginal rents provide investors with the reassurance that their capital investment in power generation capacity can be recovered, which invites new capacity build in the market.

→ The Balancing Mechanism (BM) can be seen as the GB equivalent of a US-style real-time market. A key difference between the two approaches is that in GB there is no real-time market for balancing capacity, while US-style markets feature real-time co-optimization of balancing capacity and energy.

→ The paradigm in US market operations is to acknowledge that headroom is also valuable in real time, and to pay for it. For this reason, US markets conduct real-time markets as multi-product auctions for energy, transmission and reserves.

→ Real-time co-optimization of energy and balancing capacity in US-style markets is often achieved through Reserve Scarcity Pricing based on Operating Reserve Demand Curves (ORDCs). This approach, known as "implicit co-optimization", does not rely on explicit joint auctioning of two co-optimized products (i.e. balancing capacity and energy), but rather on an administrative computation of scarcity adders, which can be applied to: (i) the balancing energy utilization prices, (ii) imbalance settlement (cash out price) and (iii) settlement of real-time reserve. A disciplined implementation of scarcity pricing based on ORDC provides incentives for participation in Balancing Markets.

→ Reserve Scarcity Pricing based on ORDCs can co-exist with, and does not replace, explicit co-optimization of energy and ancillary services in the day-ahead timeframe, which focuses primarily on elimination of opportunity cost forecast errors and allocative efficiency. Although the mechanism can be implemented in the day-ahead market, the true value of scarcity pricing through ORDC is in implementing it in the real-time balancing market.

→ Reserve Scarcity Pricing based on ORDCs can be particularly relevant for systems with increased penetration of low marginal cost renewable energy resources, which exert a downward trend on wholesale market prices, possibly creating a "missing money" problem for investment in flexible and dispatchable capacity.

→ Reserve scarcity pricing can and does co-exist with capacity remuneration mechanisms (CRMs). The general effect of scarcity pricing is to reduce the scope of a CRM. This is because reserve scarcity pricing tends to suppress missing money. This is a healthy effect: if the energy market can take care of missing money, there is no double-payment for investment costs through capacity markets.

→ Considering the GB context - already involving a high share of low marginal cost renewable energy resources, which is expected to increase materially as the country transitions to a Net Zero power system by 2035 - there could be merit in assessing the case for introducing Reserve Scarcity Pricing based on ORDCs.

→ The GB market appears to be missing a real-time market for reserve capacity. This is common in European markets, and without precedent in US designs. It corresponds to a missing market, and the ultimate effect is that it interferes with price formation in the day-ahead reserve market. The introduction of a real-time market for reserve capacity can be achieved through the implementation of co-optimization of real-time energy and reserves, however this is not necessary. An alternative is to implement a real-time market for reserve capacity through an implicit approximation of co-optimization.

→ When considering the future evolution of the GB market, an important market design question is whether there is a desire to generate scarcity prices as a result of an ORDC, or as a result of internalizing inframarginal rents in balancing price bids. The former option can be complemented with an ex-ante mitigation of bids that are clearly above marginal cost. Given how the scarcity pricing mechanism is designed, this still allows balancing prices to rise above the marginal cost of the marginal unit. On

the other hand, the latter (internalizing rents in bids) presents the market monitor with a difficult dilemma: during periods of scarcity, it becomes unclear whether these inframarginal rents result in a healthy recovery of fixed investments costs, or an exercise of market power.

→ The technical complexity of implementing reserve scarcity pricing (telemetry, computation of ORDC adders) is limited and there are precedents of detailed implementations, e.g. in ERCOT.

→ Care is needed when implementing a reserve scarcity pricing mechanism. An inconsistent introduction of reserve scarcity adders in imbalance settlement alone (without implementing a real-time market for reserve and adjusting balancing energy settlement accordingly) can induce self-dispatch of flexible assets (NIV chasing) and strip the system operator from access to much-needed flexibility in real time.

5.2 Overview

The large-scale integration of low-marginal-cost renewable resources is exerting a downward pressure on energy prices due to the merit order effect, while increasing the requirements for reserves due to the largely uncontrollable and unpredictable fluctuation of renewable resources such as wind and solar power. This shift in value streams motivates a reform in market design which places an increasing emphasis on the accurate valuation of reserve.

The accurate valuation of reserve can influence energy prices in the market: in the presence of a real-time market for reserve[72], higher prices for reserves (due to an increase in the value of reserves) invite higher profit margins in the energy market if units are to be indifferent between splitting their capacity in both of these markets. This equilibrium effect on prices can occur despite the downward pressure exerted in the energy market by the reduced marginal costs of solar and wind energy.

Scarcity pricing based on operating reserve demand curves (ORDCs) is precisely a step towards this evolution in market design, and focuses on the real-time market for energy and reserve. Scarcity pricing based on ORDC amounts to the introduction of adders in real-time balancing market prices when the system runs tight. These adders reflect the real-time value of reserve, since they are equal to the willingness of the transmission system operator to pay for balancing capacity in real time, and are activated during periods of stress in the system, when reserve capacity runs low. As explained later in the chapter, these adders uplift the real-time balancing energy price above the marginal cost of the marginal unit. They send a signal to flexible units that they have greater value and revenue opportunities if they can respond by being available and/or activated in real time, and penalizes them when they promise to be available but fail to do so.

Scarcity pricing can serve as a non-disruptive no-regret measure which can facilitate the valuation of reserve and balancing energy in future markets with deep levels of

[72] Electricity markets typically trade three major products and services: energy, access to networks, and ancillary services. A disciplined market design relies on consistency between the day-ahead market model and the real-time market model. Although this consistency is respected in various international designs, it is interesting to note that the European market design has omitted to introduce trade in one of these three foundational pillars: that of reserve. This may have been workable in systems where reserves have had a secondary role, but is likely to become increasingly challenging to cope with in systems where reserves have a major role in system operation, which is increasingly the case in systems with deeper penetration of renewable resources. Putting in place a real-time market for reserve facilitates that the pay-for-performance goals of scarcity pricing, prevents inefficient arbitrage and self-balancing of flexible resources and rather induces them to participate voluntarily in the balancing market, and results in a back-propagation of forward reserve prices based on their valuation through operating reserve demand curves in real time.

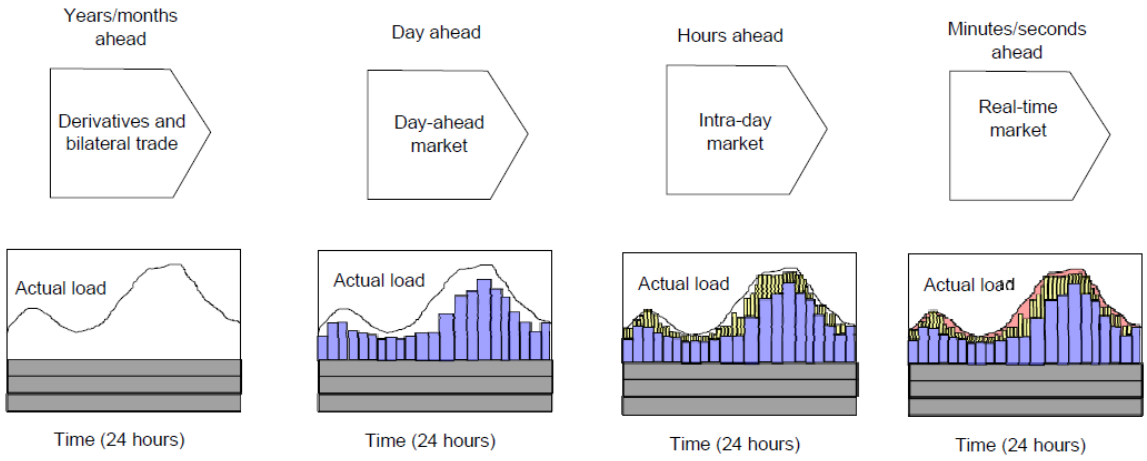
renewable energy integration. Despite its appeal, there are various aspects of the design that need to be approached with care, including the precise effect of the mechanism on balancing market settlement and imbalance charges, the introduction of a real-time market for reserve, the calibration of ORDCs, the interaction of the mechanism with capacity markets, and numerous other aspects. This chapter provides an account of the practical and academic experience that is accumulating in various US and European markets, and how it can relate to the GB context.

5.3 Key notions

5.3.1 Balancing market definitions

Balancing markets are the last stage of system operation leading up to real time. At this stage of operation, the transmission system operator ensures that the supply of electricity exactly matches the consumption of electricity. This is realized through centralized actions whereby the system operator mobilizes flexible resources, dispatchable in real time, that are used for cancelling out any real-time deviations of resources from their setpoints, forecast errors in demand or renewable supply, failures of system components, and other realizations of uncertainty. The timeline of events leading up to real-time operation is illustrated graphically in Figure 16.

Figure 16: Sequence of events leading up to the balancing market, which is the last stage in this diagram, and essentially corresponds to a real-time market for energy. Source: (Papavasiliou A. , Optimization models in electricity markets, 2023).



Apart from the operational function of the balancing market, it is also important to draw the analogies to its economic role. The balancing market is essentially a real-time market for energy in Europe. In the US it has the expanded role of a real-time multiproduct auction for energy and reserve. As a real-time market for energy, it sets the reference price for energy, in the sense that all forward trades are (or should be, since this is the moment in time when true physical constraints are revealed to be binding, i.e. when true physical scarcity is revealed in the system) indexed against the balancing market price.

The resources that are able to be dispatched upward or downward in real time by the system operator are referred to as **balancing service providers**, a terminology which is identical in the EU and GB market (as indicated in appendix D). From an economic perspective, these correspond to price-responsive suppliers of real-time energy. Balancing service providers can show up in the balancing market without an obligation to do so, in which case they are referred to as **free bids**, or by virtue of the fact that they have sold **reserve** to the system operator. If a BSP sells 1 MW reserve to the system operator in advance of real time, this is functionally equivalent to an obligation of the BSP to bid *at least* 1 MW of balancing energy in the balancing market.

Balancing service providers are meant to be mobilized in order to counteract/neutralize **imbalances**, which are unanticipated deviations from unit setpoints, forecast errors, and any other sources of deviations from a perfect equality between supply and demand of electricity. Imbalances threaten system security because they cause a deviation in system frequency, and electricity is unique as a supply chain in the sense that it cannot support sustained deviations between supply and demand. **Reserves** are exactly the necessary headroom that the system operator needs in order to ensure that resources which are flexible enough to be adjusted in real time can do so, in order to neutralize system imbalance. If reserves are the promise to be available for activation, balancing energy is the precise activation of reserves. There are diverse definitions of reserves, and the British taxonomy is presented in appendix D. But a main source of differentiation between reserves is the speed by which different reserves should respond, referred to as **full activation time** (abbreviated FAT). This response time ranges from instantaneous to a few minutes, with shorter FATs corresponding to higher-quality reserve products.

Since balancing markets are effectively real-time markets for energy, they have an underlying price. This is true for US as well as European markets. This is the **balancing price**[73], and is the true spot price (in the sense of real-time price) of energy. This is also the price paid to BSPs for their activation in real time.

Balancing responsible parties (abbreviated BRPs) are entities that control aggregations of both flexible as well as inflexible assets, and that are tasked with maintaining a balance in their perimeter. Deviations from said balance result in **imbalance charges**. From an economic standpoint, imbalances of BRPs correspond

[73] Throughout this section, we assume that there is a uniform balancing price in the balancing market, which is typical practice in US and various EU markets. The extension of the analysis to a pay-as-bid design for balancing market clearing is left for future analysis, nevertheless one can expect that many of the results carry over if we assume that agents can anticipate the marginal unit in the system and thus adapt their bidding behavior accordingly.

to a price-inelastic demand for real-time energy. The law of one price[74] strongly suggests that imbalance charges should closely track (if not exactly equal) the balancing price, because deviations from this principle would introduce arbitrage opportunities, incentives for inefficient operation, reduced liquidity in the balancing market (due to e.g. **self-dispatching**, which is a practice whereby flexible resources belonging to a BRP perimeter are not offered to the balancing market, i.e. the system operator, but instead activated by the BRPs themselves), non-truthful bidding of balancing resources in the balancing market, and other adverse side effects. Nevertheless, it is common practice in numerous systems for imbalance charges to include adders on top of the balancing price in order to penalize BRPs for failure to perfectly balance their portfolios.

Even if reserve is a promise to the system operator for keeping headroom available, this does not mean that it only has meaning in advance of real time. The paradigm in US market operations is to acknowledge that headroom is also valuable in real time, and to pay for it. For this reason, US markets conduct real-time markets as multi-product auctions for energy, transmission and reserves.

5.3.2 General idea of reserve scarcity pricing

Scarcity pricing is informally defined as the process by which short-term energy prices escalate above the marginal cost of the marginal unit, i.e. the last unit in the merit order to produce power in the market. Scarcity pricing can typically occur under stressed system conditions, where price-responsive consumers who value power at a valuation which exceeds the marginal cost of the most expensive unit in the system are partially curtailed, thereby setting the equilibrium market price at a level that

[74] Jevons, W. S. (1879). *The theory of political economy.* London: Macmillan and Company.

exceeds the marginal cost of the most expensive peaking units in the market[75]. Scarcity prices are valuable for ensuring a long-term equilibrium in an energy market, since they allow all generators, including peaking units, to recover inframarginal rents that contribute towards covering long-term investment costs. These inframarginal rents provide investors with the reassurance that their capital investment in power generation capacity can be recovered, which invites new capacity build in the market.

Scarcity pricing, as described above, relies on the workings of the energy market alone. Under conditions of risk neutrality and perfect competition, the mechanism ensures that the energy market induces not only an investment in the optimal amount of capacity but also an investment in the optimal mix of technologies[76]. The mathematical explanation of the workings of scarcity pricing is provided in appendix C.

In practice, however, this optimal outcome is not guaranteed to materialize. An important blocking point towards this idealized outcome is risk aversion combined with the fact that many consumers are not willing to be curtailed and those that might be willing cannot easily express when they are willing to be curtailed. In terms of economic theory, the demand side is inelastic and does not have a clear path towards expressing its true valuation for power. Specifically, existing markets do not allow the majority of (retail) consumers to express their true valuation for power. However, even if they did, until a significant amount of storage (e.g. in the form of EVs) becomes available in the market, demand remains largely inelastic. The latter implies that energy prices can behave in a highly volatile way: when the system is able to cover demand, short-term equilibrium prices under conditions of perfect competition do not exceed the marginal

[75] Alternatively, scarcity prices can occur when peaking units bid above their marginal cost. This is not necessarily healthy practice, since it relies on having units recuperate their investment cost through the exercise of market power under stressed system conditions. We note, however, that STOR can be (and has been) bid up to the VOLL of 6000 GBP/MWh in the GB market. Such behavior has triggered the intervention of Ofgem, which has attempted to put in place stricter conditions for participating in STOR.

[76] Boiteux, M. (1960). Peak-load pricing. *The Journal of Business, 33*(2), 157–179.

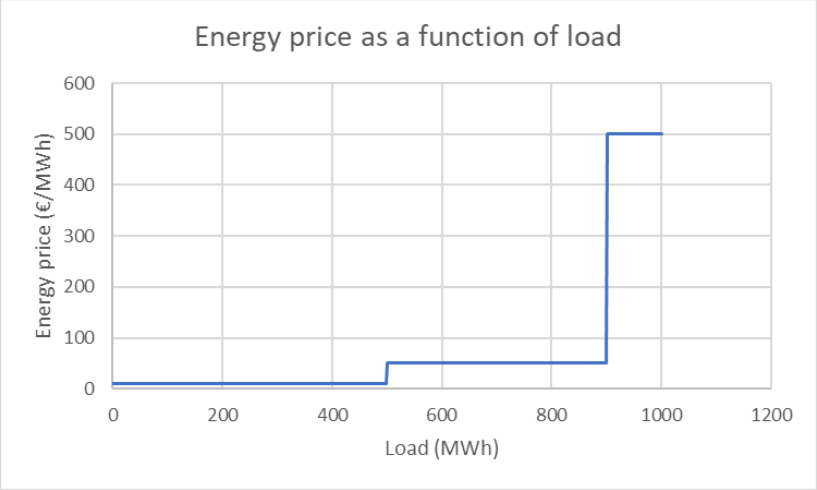
cost of the marginal unit (e.g. a couple hundred GBP per MWh) whereas they spike to VOLL when the system is forced to resort to involuntary load shedding (e.g. a few thousand GBP per MWh). From the point of view of risk-averse investors, such volatile energy prices make for a risky investment environment, since these investors rely on very few hours of involuntary load shedding with administratively determined prices for recuperating their annual investment costs.

Example 5.1*: Scarcity in a market that only trades energy*. Consider a market with the following offers:

- An energy demand offer DA for 1000 MWh at 500 €/MWh
- An energy supply offer GA for 500 MWh at 10 €/MWh
- An energy supply offer GB for 400 MWh at 50 €/MWh

The energy equilibrium price in this simple market is 500 €/MWh and is set by the demand offer DA which is at the money. In fact, the equilibrium price as a function of the quantity bid of offer DA exhibits a “volatile” behavior: it is 10 €/MWh for demanded quantities that do not exceed 500 MWh, it jumps to 50 €/MWh for demanded quantities that range between 500 and 900 MWh, and it spikes to 500 €/MWh for demanded quantities above 900 MWh. This dependency is depicted in Figure 17.

Figure 17: Energy price as a function of load in example 5.1.



■

An important goal of introducing scarcity pricing through operating reserve demand curves is to exactly overcome the fact that energy prices can be volatile in energy-only markets where there is limited demand-side elasticity. The workings of scarcity pricing with operating reserve demand curves are explained mathematically in appendix C. The intuition can be summarized as follows. Operating reserve demand curves introduce price elasticity in the reserve market. Since peaking generators split their total capacity between the energy and reserve markets under tight system conditions[77], it must be the case that the reserve price (which is the profit margin of peaking generators in the reserve market) must equal the energy price net of the marginal cost of the marginal unit (which is the profit margin of peaking generators in the energy market). Since the energy price is pegged onto the reserve price through this no-arbitrage condition, the smooth behavior of the reserve price which is due to the price-elasticity of the ORDC translates to a smooth behavior of the energy price, even if the demand side of the energy market may be inelastic. The intuition is illustrated through the following example[78].

Example 5.2*: Scarcity in a market that trades energy and reserves*. An important objective of having more accurate and formalised pricing for reserve is to ensure revenue sufficiency for flexible assets which are increasingly valuable. The way in which this is accomplished is demonstrated in the following example. Consider a market based on example 5.1:

---

[77] This can be formally proven to be the case in basic energy-reserve co-optimization models. The intuition of this is that the optimal allocation of reserve holds back those units with highest marginal cost for providing reserve. The system becomes tight when the amount of expensive capacity (i.e. generating capacity with high marginal cost) starts becoming too low to cover reserve requirements. The marginal unit in conditions of scarcity is the most expensive unit (i.e. the one with highest marginal cost) among those providing energy and the least expensive (i.e. the one with least marginal cost) among those providing reserve.

[78] As indicated previously in the report, in the examples that follow, we use MWh as a unit of measurement of quantities for both energy as well as reserve bids. The interpretation is as follows: 1 MWh of energy is the amount of energy produced by 1 MW of power generation capacity that runs for 1 hour. 1 MWh of reserve is the booking of 1 MW of power generation capacity (which therefore cannot be used for generating energy) for 1 hour.

- An energy demand offer DA for 800 MWh at 500 €/MWh
- An energy supply offer GA for 500 MWh at 10 €/MWh
- An energy supply offer GB for 400 MWh at 50 €/MWh
- Generator GA can also offer reserve. Thus, it submits a reserve bid (for up to 500 MWh) which is linked to the energy bid of GA. As we discuss in section 2.4.2, there is no explicit cost submitted for the reserve bid, because the simultaneous clearing of energy and reserve implies that the opportunity cost of the reserve bid is internally handled by the market clearing algorithm in an optimal way.
- Generator GB can also offer reserve. Thus, it submits a reserve bid (for up to 400 MWh) which is linked to the energy bid of GB.
- The TSO submits a demand for 200 MWh of reserve at a valuation of 100 €/MWh.

As explained in example 5.1, the energy-only market would produce an equilibrium price of 50 €/MWh, despite the fact that there is only 100 MW of capacity still available, and the system is thus approaching a condition of scarcity. The market clearing outcome of the energy-reserves co-optimization produces an equilibrium energy price of 150 €/MWh, which is determined by the TSO reserve demand which is only partially served. Therefore, GA uses its entire capacity for producing energy, GB produces 300 MWh of energy and uses its remaining 100 MWh for partially covering the TSO demand, and the energy price is equal to 150 €/MWh since this is the only price that keeps GB indifferent about splitting its capacity voluntarily between the energy and reserve markets. The noteworthy effect of the co-optimization of energy and reserves is that a scarcity price emerges (i.e. a price that exceeds the marginal cost of the most expensive unit in the market) even if the energy demand of the market is fully served, merely by virtue of the price elasticity of the TSO demand for reserve. This scarcity price, which emerges when the system starts running tight (i.e. for demand levels that are at 600 MWh or more) can generate inframarginal rent which can be used for enabling GB to cover its investment cost, even if demand is never curtailed in the

system, and without the need for generator GB to internalize its investment cost in its energy bid.

■

5.3.3 Explicit versus implicit co-optimization

Scarcity pricing based on ORDC, as described in section 5.3.2, is the real-time analogue of the day-ahead co-optimization of energy and reserves. What are referred to as GA and GB in the previous examples correspond to upward balancing energy offers. The energy demand corresponds to the system imbalance, and can be price-inelastic (as is typically the case in balancing markets that operate very close to real time, e.g. PICASSO in Europe).

Thus, although examples 5.1 and 5.2 can be limited in scope to the day-ahead market time frame, the true value of scarcity pricing through ORDC is in implementing these mechanisms in the real-time balancing market. This is because true scarcity is revealed in the system in real time, and what occurs in the day ahead may dilute the true stress that the system can experience in real time. One might then question to what extent the mechanism is implementable in an EU context, since the EU balancing markets (essentially the upcoming pan-European balancing platforms TERRE, MARI and PICASSO) are energy-only platforms which only trade energy, and which do not co-optimize energy and reserves. Similar considerations may be pertinent in the GB context.

Real-time co-optimization is sufficient, but not necessary, for implementing scarcity pricing[79]. Specifically, the original rollout of the ORDC mechanism in Texas[80] in 2015

[79] Papavasiliou, A., Cartuyvels, J., Bertrand, G., & Marien, A. (2023). Implementation of scarcity pricing without co-optimization in European energy-only balancing markets. *Utilities Policy*, forthcoming.

[80] Electric Reliability Council of Texas. (2014). *Purpose of ORDC, methodology for implementing ORDC, settlement impacts of ORDC.* Austin, TX: ERCOT market training.

was in a market with an energy-only real-time optimization model. The idea of implementing scarcity pricing on energy-only real-time platforms relies on the idea of *implicit co-optimization*: even if we do not co-optimize energy and reserves in the real-time economic dispatch model, the optimal solution of the co-optimized model (including dual information, i.e. prices) can be inferred by the energy-only solution under sufficiently simple settings. Some elements that constitute "sufficiently simple settings" include (i) no inter-temporal coupling, i.e. a balancing platform that only accounts for the current imbalance settlement period, (ii) no binary variables, i.e. unit commitment variables are fixed in advance of balancing, (iii) a single reserve product, (iv) no limits on the reserve variables themselves apart from power generation capacity, e.g. no ramp constraints. These are restrictive assumptions, but the idea of the ERCOT design was to make these assumptions anyway and compute reserve prices after the fact, that approximate the equilibrium prices that a co-optimization model would produce, even if these prices were not perfectly in equilibrium with the dispatch instructions of the energy-only balancing platform. A similar design has been proposed for the Belgian market[81].

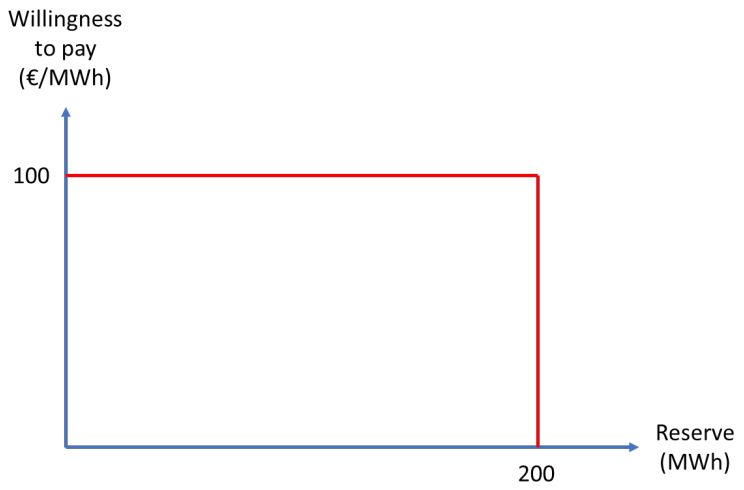
Example 5.3*: Implementing scarcity pricing based on ORDC in an energy-only platform*. Consider the system of example 5.2. Assume that the market operates as an energy-only balancing platform, but suppose that we would still like to implement scarcity pricing based on ORDC, even if we are not co-optimizing energy and reserves in our balancing platform. The idea for achieving this is the following:

- The upward balancing offer of GA is fully matched
- The upward balancing offer of GB is matched up to 300 MWh
- The imbalance (which is the demand DA) is fully matched

[81] Belgian Commission for Electricity and Gas Regulation. (2021). *Study on the implementation of a scarcity pricing mechanism in Belgium.* Brussels, Belgium: CREG.

The system telemetry then records that the total leftover reserve capacity in the system is 100 MWh, i.e. the headroom of GB which has been used for 300 out of its 400 MWh. This corresponds to leftover reserve supply. In order to compute reserve prices, this leftover reserve supply needs to be matched with the TSO demand for reserve, which is depicted in Figure 18. The crossing of the reserve supply and demand curves occurs at 100 MWh of reserve, at a clearing price of 100 €/MWh. The energy price is computed as the marginal cost of the marginal unit plus the price of reserve, i.e. 100+50=150 €/MWh. Equivalently, the energy price in the implicit co-optimization of reserve is the price of the energy-only platform plus the reserve price (which is also referred to as a **scarcity adder** or **ORDC adder**).

Figure 18: The TSO demand for reserve in example 5.3.



■

To summarize, the implementation of an implicit co-optimization of energy and reserves proceeds as follows:

- We clear the energy-only platform.
- We measure the leftover flexible capacity through telemetry (this can be done after the fact, at the stage of settlement / billing).
- We trace this leftover reserve capacity on the ORDC, and this gives us the reserve price of the market.

- We compute the corrected balancing price as the energy price of the energy-only platform plus the ORDC adder.

5.3.4 Operating reserve demand curves

An important design aspect of a scarcity pricing mechanism is the shape of the operating reserve demand curve. The calibration of the operating reserve demand curve has been connected to loss of load probability as a function of the amount of reserve that the system carries, as well as loss of load probability, according to the following formula[82]:

$$VR(x) = (VOLL - MC) \cdot LOLP(x).$$

Here, the notation is as follows:

- $VOLL$: this is the estimated value of lost load. In Great Britain, the value has been lifted from 3000 GBP/MWh to 6000 GBP/MWh following P305[83].
- MC: this is the estimated marginal cost of the marginal dispatchable resource in the system in real time. This also explains why scarcity adders recede when the energy market price is sufficiently high to signal scarcity on its own.
- $LOLP(x)$: this is the loss of load probability as a function of available reserve x in the system in real time. LOLP is depicted graphically in Figure 19. The shaded area is the probability that the system fails to serve demand when it is

[82] Hogan, W. W. (2013). Electricity scarcity pricing through operating reserves. Economics of Energy and Environmental Policy, 2(2), 65-86.

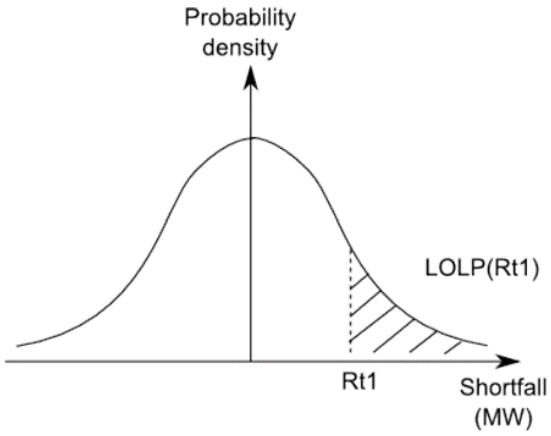
Papavasiliou, A., & Smeers, Y. (2017). Remuneration of Flexibility using Operating Reserve Demand Curves: A Case Study of Belgium. The Energy Journal, 38.

Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

[83] Department for Business, Energy and Industrial Strategy. (2022). *Digest of UK Energy Statistics.* London: BEIS.

carrying a given amount of reserve. This is the area of the probability density function of capacity shortfall for x MW or more. This probability density function is estimated based on historical measurements, e.g., through historical records of system imbalance. The shaded realizations of imbalance in the figure correspond exactly to those circumstances when the system fails to serve demand.

Figure 19: Loss of load probability as a function of available reserve. Source: (Papavasiliou & Smeers, 2017).



It is interesting to note that an explicit co-optimization of energy and reserve implies an ORDC that varies at every imbalance settlement period, as the anticipated marginal cost of the marginal unit in the system varies. In a setting of implicit co-optimization, the computation of real-time reserve prices simply amounts to telemetering the amount of available reserve capacity in the system, and plugging the result into the formula above. In practical implementations of reserve scarcity pricing, such as the Texas design[84], the LOLP function is dynamically adjusted, depending on the season and 4-hour interval of the day that we find ourselves in. Such a dynamic adjustment of the

[84] Electric Reliability Council of Texas. (2014). *Purpose of ORDC, methodology for implementing ORDC, settlement impacts of ORDC.* Austin, TX: ERCOT market training.

LOLP in the ORDC formula has also been proposed in the Belgian scarcity pricing design[85], and dynamic LOLPs are also apparently employed in Great Britain[86].

The use of an ORDC function that depends on VOLL and LOLP is often adopted in practice, but it is not the only option for estimating an ORDC. Instead, stepped ORDCs can also be used, as is the case in the examples that will be presented in section 5.4. Note that an extreme version of a stepped ORDC is a perfectly inelastic reserve requirement (e.g., the 140 MW aFRR requirement in Belgium). In this sense, *all* systems have ORDCs, even if some are more price-elastic than others.

Wider ORDCs imply higher scarcity prices. Some nuanced design choices that can affect the shape of the ORDC, and thus the level of scarcity prices, include the following: (i) the assumed/estimated VOLL, (ii) whether the measured reserve is the leftover reserve in the system before or after the system clears an imbalance, (iii) in the case of multiple reserve products, the degree to which 15-minute imbalances are the result of perfectly correlated or independent imbalance increments. A comprehensive study of how these nuanced choices can affect the behavior of scarcity pricing in a case study of the Belgian market is documented[87]. Alternative case studies on the sensitivity of scarcity adders in real systems include the case of Texas[88] and Illinois[89].

---

[85] Papavasiliou, A., Smeers, Y., & de Maere d'Aertrycke, G. (2019). *Study on the general design of a mechanism for the remuneration of reserves in scarcity situations.* Louvain la Neuve, Belgium: https://ap-rg.eu/wp-content/uploads/2020/07/CREGReportFinal.pdf.

[86] OFGEM. (2022). *Annual Report on the Operation of the CM 2020/21 and 2021/22.* London, UK: OFGEM.

[87] Cartuyvels, J., & Papavasiliou, A. (2023). Calibration of Operating Reserve Demand Curves Using a System Operation Simulator. *IEEE Transactions on Power Systems*.

[88] Zarnikau, J., Zhu, S., Woo, C. K., & Tsai, C. (2020). Texas's operating reserve demand curve's generation investment incentive. *Energy Policy, 137*, 111143.

[89] Zhou, Z., & Botterud, A. (2014). Dynamic scheduling of operating reserves in co-optimized electricity markets with wind power. *IEEE Transactions on Power Systems, 29*(1), 160-171.

Example 5.4*: An ORDC based on LOLP and VOLL*. Suppose that we have estimated, from historical records of system imbalances or activated balancing energy or some other indicator of capacity shortfall[90], that the mean of the capacity shortfall in the system has a mean of 0 MWh and a standard deviation of 60 MWh. Suppose, furthermore, that the capacity shortfall is assumed to obey a normal distribution. Finally, we assume that the VOLL of the market is equal to 6000 €/MWh, and that the marginal cost of the marginal unit that is currently online in the system is 100 €/MWh. The leftover reserve in the system after clearing the imbalance is 100 MWh. The scarcity price is thus computed as:

$$VR(100) = (VOLL - MC) \cdot LOLP(100) = (6000 - 100) \cdot LOLP(100)$$

where

$$LOLP(100) = \mathbb{P}[Imb > 100] = 1 - \mathbb{P}[Imb \leq 100] = 1 - \Phi_{0,60}(100) = 4.78\%.$$

Here, $\Phi_{\mu,\sigma}(\cdot)$ is the cumulative distribution function of a normal random variable with a mean μ and a standard deviation σ. Thus, the scarcity price equals

$$VR(100) = (6000 - 100) \cdot LOLP(100) = 5900 \cdot 0.0478 = 281.96 \frac{€}{\text{MWh}}.$$

5.3.5 Interaction of reserve scarcity pricing with Capacity Remuneration Mechanisms

Reserve scarcity pricing can and does co-exist with capacity remuneration mechanisms (CRMs). Configurations of CRMs and scarcity pricing in various EU Member States are depicted in Table 9.

[90] Note that this capacity shortfall should correspond to the activation time of the reserve in question. Typical settings that are investigated in EU market design include 15-minute imbalance settlement periods, and mFRR products.

Table 9: A cartography of ORDCs and CRMs in the European market. Columns in the table indicate concerned Member States. Green boxes correspond to Member States where the respective mechanism is already in place, while orange boxes correspond to Member States where the rollout of the respective mechanism is contemplated. Source: (Papavasiliou & Mou, 2023).

The capacity market in GB relies on four-year-ahead auctions, as well as one-year-ahead top-up auctions. New builds engage in long-term contracts, and are only eligible for the four-year-ahead auctions. The contracts run up to 15 years[91]. The CRM demand is calibrated to achieve a LOLE target of three hours per year. Existing capacity is eligible only for one-year auctions. If there is a system stress event (there are methodologies to define this), entities that are awarded in the capacity auction have to prove that their capacity was available in the market. Specifically, generation capacity owners must prove that they were not only available, but that the metered volume was traded with other market participants or with the ESO in the balancing market. The GB CRM is not complemented by reliability options.

The general effect of scarcity pricing is to reduce the scope of a CRM[92]. This is because reserve scarcity pricing tends to suppress missing money. Supply bids in the capacity market are essentially internalizing this missing money. Thus, reducing the amount of missing money through scarcity pricing exerts a downward pressure on supply bids in the capacity auction. This in turn exerts a downward pressure on capacity prices. This is a healthy effect: if the energy market can take care of missing money, there is no double-payment for investment costs through capacity markets. Double payment has been used in the public debate for supporting the argument that

[91] OFGEM. (2022). *Annual Report on the Operation of the CM 2020/21 and 2021/22.* London, UK: OFGEM.

[92] Mou, Y., Papavasiliou, A., Hartz, K., Dusolt, A., & Redl, C. (2023). An analysis of shortage pricing and capacity remuneration mechanisms on the pan-European common energy market. *Energy Policy*.

scarcity pricing cannot coexist with CRMs[93], for which there appears to exist no evidence in a setting of perfect competition. One can also demonstrate analytically[94] that a "canonical" CRM[95] results in a capacity market price of zero, and thus implies that the capacity market does not interfere with the functioning of the energy market. This raises questions about the calibration of CRM demand curves, since it is common practice in various designs (including the GB design) for CRM demand curves to reach zero valuation strictly above the level (e.g. 100+x%) of target capacity[96].

An interesting question that emerges with respect to the interaction between CRMs and reserve scarcity pricing relates to the interplay between reliability options and scarcity prices. Although the GB CRM does not involve reliability options, we comment briefly on this point for the sake of completeness. The value of reliability options is factored into the supply bids of participants in the CRM auctions. Scarcity adders increase the effective real-time price of energy, thus the payoff of the reliability options. At the same time, the adders apply in full force to those resources that are not cleared in the CRM, since those resources do not forego their rents during scarcity periods through reliability options.

[93] Papavasiliou, A., Cartuyvels, J., Bertrand, G., & Marien, A. (2023). Implementation of scarcity pricing without co-optimization in European energy-only balancing markets. *Utilities Policy*, forthcoming.

[94] Mou, Y., Papavasiliou, A., Hartz, K., Dusolt, A., & Redl, C. (2023). An analysis of shortage pricing and capacity remuneration mechanisms on the pan-European common energy market. *Energy Policy*, under review.

[95] We refer to a "canonical" CRM as one which is consistent with the reliability standards of the TSO. In particular, this is a CRM that is dimensioned such that the valuation for additional capacity becomes equal to zero at the level where capacity ensures the target LOLE.

[96] Papavasiliou, A. (2021). *Overview of EU Capacity Remuneration Mechanisms.* Louvain la Neuve, Belgium: UCLouvain.

5.3.6 Points of attention in the GB design

Ofgem launched the Electricity Balancing Significant Code Review (EBSCR) in August 2012, in order to address concerns that there were insufficient signals for the market to balance[97]. BSC modification P305 was subsequently approved by Ofgem to be implemented on November 5, 2015, followed by a second phase of changes that came into effect in November 2018. Part of the balancing market modifications includes the introduction of a single balancing price. This is consistent with the law of one price[98], and it is argued in section 5.4 that an alignment between imbalance settlement and balancing prices may be further considered. The goal of such an alignment is to prevent inefficient self-dispatching, and thus to allow the balancing market to clear more efficiently and improve price discovery.

There are notable scarcity signals in the existing GB market. This is more so given the P305 reforms, which include changes that allow the marginal activated unit to set the cash out price. This has resulted in stronger balancing price signals (even if balancing prices have decreased on average). An important market design question is whether there is a desire to generate scarcity prices as a result of an ORDC, or as a result of internalizing inframarginal rents in balancing price bids. The former option can be complemented with an ex-ante mitigation of bids that are clearly above marginal cost. Given how the scarcity pricing mechanism is designed, this still allows balancing prices to rise above the marginal cost of the marginal unit. On the other hand, the latter

---

[97] OFGEM. (2023). *Analysis of the first phase of the Electricity Balancing Significant Code Review.* London, UK: OFGEM. Retrieved from https://www.ofgem.gov.uk/sites/default/files/docs/2018/08/analysis_of_the_first_phase_of_the_electricity_balancing_significant_code_review_as_final_

[98] Jevons, W. S. (1879). *The theory of political economy.* London: Macmillan and Company.

Cramton, P. C., & Stoft, S. (2006). *The convergence of market designs for adequate generating capacity with special attention to the CAISO's resource adequacy problem.* Cambridge, MA: MIT Center for Energy and Environmental Policy Research.

(internalizing rents in bids) presents the market monitor with a difficult dilemma: during periods of scarcity, it becomes unclear whether these inframarginal rents result in a healthy recovery of fixed investments costs, or an exercise of market power.

An important promise of scarcity pricing based on ORDC is the back-propagation of real-time prices to forward energy and reserve markets. Virtual trading is allowed in the GB energy market[99], and this is conducive towards back-propagating energy prices to forward markets. On the other hand, the GB market does appear to be missing a real-time market for reserve capacity. This is common in European markets, and without precedent in US designs[100]. It corresponds to a missing market, and the ultimate effect is that it interferes with price formation in the day-ahead reserve market, because we put in place forward markets for services that we are not trading in real time. Is the introduction of a real-time market for reserve synonymous with co-optimization of energy and reserve in real time? No. The introduction of a real-time market for reserve capacity *can* be achieved through the implementation of co-optimization of real-time energy and reserves, however this is *not necessary*. An alternative is to implement a real-time market for reserve capacity through an implicit approximation of co-optimization, as described in section 5.3.3.

Additional aspects of reserve scarcity design that may be worth investigating in the context of the GB design include (i) multiple ORDCs, i.e. adders for multiple reserve

[99] OFGEM. (2023). *Analysis of the first phase of the Electricity Balancing Significant Code Review.* London, UK: OFGEM. Retrieved from https://www.ofgem.gov.uk/sites/default/files/docs/2018/08/analysis_of_the_first_phase_of_the_electricity_balancing_significant_code_review_as_final_

[100] Papavasiliou, A. (2020). Scarcity pricing and the missing European market for real-time reserve capacity. *The Electricity Journal, 33*(10), 106863.

products, with a consideration of the effect of one-way substitutability[101], and (ii) the application of scarcity pricing on a network[102].

The GB market has had in place a reserve scarcity pricing function in the utilization of Short-Term Operating Reserve (STOR)[103] (see also Table 2). STOR is a service that provides additional power from generation or demand reduction. Slow reserve is intended to be the new version of STOR.

Original inception. STOR used to be contracted about 18 months ahead with a fixed utilization price which was not responsive to real-time conditions. This however resulted in low utilization prices, even in periods of scarcity.

Reserve scarcity pricing (RSP) function. Due to the fact that the original design was fixing utilization prices in advance, which were fairly low even in true periods of scarcity, as reported by Elexon[104], the STOR mechanism was backed up by a reserve scarcity pricing (RSP) function, which was put in place in 2015. The idea of the reserve scarcity pricing function is to compute a scarcity adder, per standard international practices, based on value of lost load. In a phased implementation of the mechanism, a VOLL of £3000/MWh was used, which was subsequently escalated to £6000/MWh. Also per international standard practices, the RSP also depends on an estimate of the

---

[101] Hogan, W. W. (2013). Electricity scarcity pricing through operating reserves. Economics of Energy and Environmental Policy, 2(2), 65-86.

Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

[102] N-SIDE. (2023). *Svk project on scarcity pricing.* https://www.svk.se/siteassets/om-oss/rapporter/2023/report-scarcity_pricing-phase-1.pdf : Papavasiliou, Anthony.

[103] ESO, “Short-term operating reserve (STOR)”, available online: https://www.nationalgrideso.com/industry-information/balancing-services/reserve-services/short-term-operating-reserve-stor

[104] https://www.elexon.co.uk/article/bsc-insight-why-the-reserve-scarcity-price-is-being-reviewed/#:~:text=The%20Reserve%20Scarcity%20Price%20represents,the%20Value%20of%20Lost%20Load

loss of load probability, which is a function of excess capacity in the system. The computation of this excess capacity, once static, subsequently evolved to a dynamic computation which depends on system conditions. Indicatively, it starts producing non-zero adders at an excess capacity of 1800 MW or less, where excess capacity is the difference between available capacity and forecast demand. The loss of load probability was updated and published one day ahead, eight hours ahead, four hours ahead, and an hour ahead of the settlement period in question. The result of this computation affected both the STOR units that would be activated for the settlement period in question, as well as the system marginal price for those periods where the activated STOR units were marginal. As Elexon reports, the RSP produces non-negligible results somewhat rarely during the period of its implementation.

Current status. The procurement of ancillary services has moved to the day ahead timeframe. Moreover, in the current arrangement, the contracting sets the availability price, not the utilization price. Crucially, in the current arrangement flexibility providers can bid for utilization as they see fit close to real time, thus the original motivation for including the RSP no longer stands.

5.4 Implementation considerations

The implementation of scarcity pricing in Europe has lagged behind relative to US markets. Since the disciplined implementation of the approach affects not only balancing market pricing but also imbalance settlement and requires the effective introduction of a new real-time market (for balancing capacity), the proposal has raised various concerns regarding desirability, actual need for implementation, and legal

feasibility. Some of the arguments that have been raised are documented in the professional[105] and academic[106] literature.

The question of whether the mechanism is desirable or needed relates to the extent to which the system operator considers that its existing arrangements are capable of attracting sufficient investment in technologies that are future-proof, i.e. able to support the future needs of a system that is dominated by renewables. A confusion that sometimes emerges in this debate is whether the scarcity pricing mechanism is put in place for dealing with security or adequacy. Quoting Stoft[107]: "*Security is the system's ability to withstand sudden disturbances, while adequacy is the property of having enough capacity to remain secure almost all of the time.*".

Although the two issues are not the same, they are interdependent, and attracting flexible resources to the market often contributes to resolving at least some part of the adequacy problem. The questions regarding desirability are accentuated by the recent energy crises. Given the recent escalation in energy and natural gas prices (see also the discussion in section 0), it may be argued that scarcity pricing is not at the top of the concerns of market stakeholders at the moment. An appealing aspect of a properly designed mechanism is that scarcity prices recede when the energy market is, on its own right, able to signal scarcity, and take care of sparking new investment. It is also

---

[105] CREG, 2021. Study on the implementation of a scarcity pricing mechanism in Belgium, Brussels: Commission for Electricity and Gas Regulation.

ELIA, 2020. Final Report on Elia's Findings Regarding the Design of a Scarcity Pricing Mechanism for Implementation in Belgium. ELIA, Brussels.

[106] A. Papavasiliou, J. Cartuyvels, G. Bertrand, A. Marien, Implementation of scarcity pricing without co-optimization in European energy-only balancing markets, Utilities Policy, vol. 81, 101488, April 2023

A. Papavasiliou, Scarcity Pricing and the Missing European Market for Real-Time Reserve Capacity, The Electricity Journal, vol. 33, no. 10, September 2020

[107] Stoft, S. (2002). Power system economics: designing markets for electricity. Piscataway: IEEE press.

worth reflecting about how coherent it is to have capacity mechanisms in place while also putting in place measures that mute the scarcity signals of the energy market.

The question of whether the disciplined implementation of scarcity pricing is in line with EU legislation is somewhat specific to the European electricity market. The issue of legal compatibility with EBGL is debated extensively between the Belgian regulator and system operator[108]. However, the debate is somewhat specific to European legislation, and therefore probably not of general interest. It is nevertheless important to underline that the basis for the legal debate is whether the scarcity prices produced by the RSP can be applicable both to BSPs as well as to BRPs. In Belgium, it was claimed that the latter is possible, but the former is not.

One more interesting point of debate is whether the implementation of scarcity pricing (and, in fact, the alignment of imbalance settlement with the prices of the balancing market) can threaten system security. Concretely, concerns have been expressed by ELIA (the Belgian TSO) that certain balancing market designs and imbalance settlement arrangements can induce the Belgian system to respond to the needs of much larger balancing areas, such as Germany. However, in doing so, the activations of Belgian balancing service providers can cause congestion within the Belgian system. This is an issue that extends beyond scarcity pricing, and relates to imbalance settlement in general. It will emerge in the future in the context of the implementation of imbalance settlement in the presence of pan-European balancing platforms such as MARI and PICASSO, as well as transmission-distribution coordination.

An additional implementation concern that has been raised in the context of scarcity pricing is whether the mechanism can be implemented by one country alone, or

---

[108] CREG, 2021. Study on the implementation of a scarcity pricing mechanism in Belgium, Brussels: Commission for Electricity and Gas Regulation.

ELIA, 2020. Final Report on Elia's Findings Regarding the Design of a Scarcity Pricing Mechanism for Implementation in Belgium. ELIA, Brussels.

whether it should be introduced simultaneously by interacting neighbouring countries. A related concern, in the case of unilateral implementation, is whether the country implementing the mechanism on its own ends up subsidizing the adequacy needs of neighbouring countries. Another related concern in case of unilateral implementation is whether the country which executes the mechanism introduces an unfair competitive advantage for its domestic BSPs and induces them to "dump" their resources below cost. As explained in the academic and professional literature cited above, the mechanism can be implemented unilaterally, and the disciplined implementation of the mechanism does not grant an unfair competitive advantage to domestic BSPs.

One issue that is not discussed in the aforementioned references, but may be relevant in terms of practical implementation, is the possible delay in computing the adder. Following a request by the Belgian regulator, the Belgian system operator implemented an ex-post computation[109] of scarcity adders in 2018. The scarcity adder was computed based on the so-called Available Reserve Capacity (ARC) of the Belgian system, and scarcity adders were posted online one day after the fact. The ex-post computation of the adders is not a problem if the mechanism follows a disciplined implementation, because the mechanism is specifically designed in order to render agents indifferent between showing up in real time for providing standby balancing capacity or activating it. In either case, they receive the same profit margin, and they are thus rewarded equally for being able to relieve the system in either of the two modes. Thus, even if this reward is computed after the fact, it is designed to be compatible with the profit maximization incentives of agents. This was also the spirit in which the original Texas mechanism was implemented. On the other hand, for implementations of the mechanism that deviate from the disciplined approach, one of

[109] ELIA, 2018. Study Report on Scarcity Pricing in the Context of the 2018 Discretionary Incentives. Brussels: Belgian transmission system operator.

A. Papavasiliou, Scarcity Pricing and the Missing European Market for Real-Time Reserve Capacity, The Electricity Journal, vol. 33, no. 10, September 2020

the adverse side effects is that agents are given an incentive to internalize the (forecasted) adder, therefore delays in estimating or computing it could affect their behavior. But such internalization of adders is anyway defeating the purpose of the mechanism, and should be avoided by design, as discussed also in section 5.5.

Delays in computing adders are natural. For instance, there is a difficulty in forecasting the settlement price in the current GB system. This information can be relevant for market participants, especially in a regime of imbalance settlement with multiple interacting balancing energy products (e.g. aFRR and mFRR), where some of these products (specifically aFRR) are settled at a time resolution which is more granular (i.e. one minute or four seconds) than that of imbalance settlement (i.e. fifteen minutes). This misalignment in time granularity of imbalance settlement and balancing products raises interesting tradeoffs between security and efficiency of real-time operations, and extends beyond scarcity pricing to the broader issue of imbalance settlement, thus extending beyond the scope of the current report.

The remainder of chapter 5 focuses on the issue of how the mechanism can be implemented in a disciplined fashion, and in particular what it implies for the remuneration of BSPs and BRPs. Many alternatives have been debated in the professional community. Their rationale is discussed, and their properties are summarized, with appropriate references to technical literature being provided along the way.

5.5 Qualitative assessment of different design choices

Apart from the IT aspects of implementing an implicit co-optimization of energy and reserves in real time (telemetry, computation of ORDC adders), which are fairly

straightforward to overcome[110], there are important market design implications from a disciplined implementation of scarcity pricing. The specific design dilemmas can be summarized in the form of the following question: when implementing scarcity pricing based on ORDC in an implicit co-optimization setting, where do the computed scarcity adders apply?

5.5.1 Disciplined approximation of co-optimization

The disciplined implementation of scarcity pricing based on ORDC requires the application of these adders to all of the below[111]:

- As add-ons to the balancing price of the energy-only platform
- As add-ons to the imbalance charge
- As prices for settling a real-time market for reserve

The economic rationale for applying the ORDC adders to all of the above entities is based on (i) the law of one price[112] which has been stated in economic theory[113] since 1879, (ii) the no-arbitrage conditions that characterize a market equilibrium, and (iii) the fact that trading a product or service forward (reserve, in the case of our analysis) requires that a real-time market exist for said product or service.

In what follows, we discuss alternatives to the disciplined implementation of scarcity pricing based on ORDC that have been debated among market stakeholders. We

---

[110] Electric Reliability Council of Texas. (2014). *Purpose of ORDC, methodology for implementing ORDC, settlement impacts of ORDC.* Austin, TX: ERCOT market training.

[111] Papavasiliou, A. (2020). Scarcity pricing and the missing European market for real-time reserve capacity. *The Electricity Journal, 33*(10), 106863.

[112] Note that the introduction of a single price for upward/downward activation in the balancing market which has been instituted in the GB market (OFGEM, 2023), (PSR, 2023) is consistent with the law of one price.

[113] Jevons, W. S. (1879). *The theory of political economy.* London: Macmillan and Company.

discuss these alternatives using our running example, and highlight some of the side-effects of the alternative design proposals that have emerged.

Example 5.5*: Disciplined application of scarcity pricing based on ORDC*. The introduction of scarcity pricing based on ORDC in example 5.3 implies the settlements that are indicated in Table 10. Note that the last row of the "balancing energy" column is guaranteed to exactly cancel out the last row of the "imbalance" column, since one is the payment to the suppliers of real-time energy, whereas the other is the payment from consumers of real-time energy. The last row of the "reserve" column is guaranteed to be equal to zero, since payments from the TSO for buying reserve equal payments to BSPs for offering reserve. Note furthermore how GB receives revenues from both the energy market and the reserve market. Compared to an energy-only design, this flexible supplier receives an additional 100 €/MWh (the price of reserve) for its entire capacity, whether it is offered in the energy market or the reserve market. This enhances the incentive of flexible suppliers to roll out new flexible capacity, which can be rewarded under scarcity conditions, when these resources deliver by producing or being on standby.

Table 10: Settlement table for the disciplined approximation of scarcity pricing.

Agent	Balancing energy	Imbalance	Reserve	Total
GA	150 [€/MWh] x 500 [MWh] = 75000 €	N/A	100 [€/MWh] x 0 [MWh] = 0 €	75000 €
GB	150 [€/MWh] x 300 [MWh] = 45000 €	N/A	100 [€/MWh] x 100 [MWh] = 10000 €	55000 €
D	N/A	150 [€/MWh] x (-800) [MWh] = -120000 €	N/A	- 120000 €
TSO	N/A	N/A	100 [€/MWh] x (-100)[MWh] = -10000 €	- 10000 €
Total	120000 €	-120000 €	0 €	

Table 10 is to be contrasted with the settlement in an energy-only design, the payoffs of which are presented in Table 11.

Table 11: Settlement table for an energy-only design.

Agent	Balancing energy	Imbalance	Reserve	Total
GA	50 [€/MWh] x 500 [MWh] = 25000 €	N/A	N/A	75000 €
GB	50 [€/MWh] x 300 [MWh] = 15000 €	N/A	N/A	15000 €
D	N/A	50 [€/MWh] x (-800) [MWh] = -40000 €	N/A	- 40000 €
TSO	N/A	N/A	N/A	N/A
Total	40000 €	-40000 €	N/A	

■

One important appeal of the disciplined approximation of scarcity pricing is that (i) it ensures that flexible resources voluntarily bid their flexible capacity into the balancing market, as opposed to taking matters in their own hands by self-dispatching their units. This is a very important virtue of the proposed design, because, as increasing amounts of renewable resources are being integrated in the system, the TSO can use all the flexibility it can get its hands on, and designs that induce this flexibility to not reveal and make itself available to the market should be avoided. Moreover, (ii) this reserve capacity is bid into the balancing market at its true marginal cost. Finally, (iii) the design gives rise to a forward reserve price which is driven by the value of reserve capacity, as quantified by the ORDC. This, too, is an important virtue of the mechanism, since it results in a robust investment signal that indicates that flexible capacity is welcome in the market. The **back-propagation** of real-time prices to forward markets is an

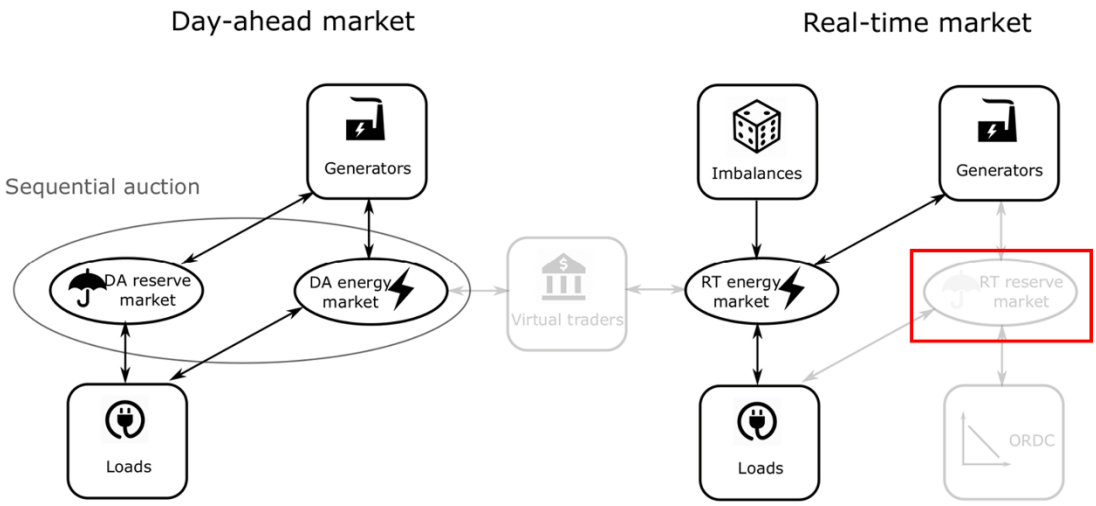
essential part of this process. And virtual trading, which is a notable aspect of the GB market design[114], further enables such back-propagation to take place.

This is to be contrasted to the case of an energy-only design. The energy-only design (i) maintains the appealing aspect of having generators bid their entire flexible capacity voluntarily into the balancing market at its true marginal cost. However, (ii) this design fails to generate a forward reserve price. This is hardly surprising. Forward markets in a risk-neutral environment track the expected real-time price of the underlying commodity or service. If there is no market for trading the underlying commodity or service, then the forward price that emerges for said commodity or service is zero. Put another way, the existing EU balancing market design features a **missing market**, one for real-time reserve. The point is illustrated in Figure 20. Given that modern electricity markets trade three major products and services (energy, reserve, and transmission access), this is a quite remarkable oversight in EU balancing market design[115], which has also been inherited in other worldwide markets. In practice, electricity markets *do* produce forward reserve prices. This can be due to a variety of other reasons (e.g., fixed costs related to the provision of reserve, such as the commitment of units that need to be brought online in order to make reserve capacity available, or even dominant positions in reserve markets), but back-propagation of a real-time value as quantified by an ORDC is not one of those reasons. This is in stark contrast to US market design, which properly trades all three products and services (energy, reserves and transmission access) in both the day ahead as well as real time.

[114] Department for Business, Energy and Industrial Strategy. (2022). *Digest of UK Energy Statistics.* London: BEIS.

[115] Papavasiliou, A. (2020). Scarcity pricing and the missing European market for real-time reserve capacity. *The Electricity Journal, 33*(10), 106863.

Figure 20: The European and GB market design features a missing market in real time: that of reserve/balancing capacity. A disciplined implementation of scarcity pricing based on operating reserve demand curves restores this missing market (red box). Source: presentation[116] at Isaac Newton Institute, Cambridge University.



5.5.2 Adder on imbalance settlement only

Certain market stakeholders have argued in favor of limiting the application of scarcity adders to BRPs only[117]. For instance, adders on imbalance settlement do exist already in the Belgian market. Scarcity pricing would simply amount to topping up the existing imbalance settlement scheme with an extra adder, one related to scarcity. The settlement is illustrated in the following example.

---

[116] https://ap-rg.eu/wp-content/uploads/2020/07/INI2019-2.pdf

Papavasiliou, Anthony, Yves Smeers, and Gauthier deMaere d'Aertrycke. "Market design considerations for scarcity pricing: A stochastic equilibrium framework." *The Energy Journal* 42.5 (2021): 195-220.

[117] Giesbertz, P. (2021). The power market design column. Retrieved from The scarcity of scarcity pricing: https://www.linkedin.com/pulse/power-market- design-column-scarcity-pricing-paul-giesbertz/

ELIA. (2021). Final report on ELIA's findings regarding the design of a scarcity pricing mechanism for implementation in Belgium. Brussels, Belgium: ELIA.

Example 5.6*: Adder on imbalance settlement only*. The settlement under the design that applies adders on imbalance settlement only is presented in Table 12. Note that reserve is no longer traded in this design. Moreover, the imbalance charges exceed the payments that are collected by BSPs for upwards activation.

Table 12: Settlement table for the design which applies an adder on imbalance settlement only.

Agent	Balancing energy	Imbalance	Reserve	Total
GA	50 [€/MWh] x 500 [MWh] = 25000 €	N/A	N/A	25000 €
GB	50 [€/MWh] x 300 [MWh] = 15000 €	N/A	N/A	15000 €
D	N/A	150 [€/MWh] x (-800) [MWh] = -120000 €	N/A	- 120000 €
TSO	N/A	N/A	N/A	N/A
Total	40000 €	-120000 €	N/A	

■

An important drawback of this mechanism is that it induces certain BSPs (especially the ones with lower marginal cost) to take their chances with collecting scarcity adders in imbalance settlement by self-dispatching their flexible units, since this pays better than bidding and being activated in the balancing market, which does not feature a scarcity adder. This is not good since it strips the balancing market from flexibility that exists in the system due to distorted market design. As we are increasingly integrating renewable resources in power grids, this is a move in the wrong direction, because now, more than ever, the TSO requires these balancing resources. This process also results in inefficient dispatch, and it obviates price discovery. Interestingly, this design *does* imply an opportunity cost for resources that self-dispatch, thus it generates a forward reserve price signal, however the process by which this reserve price signal emerges is exactly the inefficient self-dispatch of flexible resources with low marginal

cost. This back-propagation of reserve opportunity costs is weaker than the expected value of the scarcity adder, as determined by an ORDC[118].

5.5.3 Adder on imbalance settlement and the balancing price

An alternative proposal that has recently emerged in the market design debate over scarcity pricing is the application of scarcity adders on both balancing prices and imbalance settlement, but without putting in place a market for reserve[119]. The design is illustrated in the following example.

Example 5.7*: Adder on balancing prices and imbalance settlement*. The application of this design to our running example is presented in Table 13. The only thing that changes relative to Table 10 is that the "reserve" column is now entirely dropped.

Table 13: Settlement table for the design which applies an adder on imbalance settlement only.

Agent	Balancing energy	Imbalance	Reserve	Total
GA	150 [€/MWh] x 500 [MWh] = 75000 €	N/A	N/A	75000 €
GB	150 [€/MWh] x 300 [MWh] = 45000 €	N/A	N/A	45000 €
D	N/A	150 [€/MWh] x (-800) [MWh] = -120000 €	N/A	- 120000 €
TSO	N/A	N/A	N/A	N/A
Total	120000 €	-120000 €	N/A	

[118] Papavasiliou, A., & Bertrand, G. (2021). Market design options for scarcity pricing in European balancing markets. *IEEE Transactions on Power Systems, 36*(5), 4410-4419.

[119] Cartuyvels, J., Bertrand, G., & Papavasiliou, A. (2023). Market Equilibria in Cross-Border Balancing Platforms. *IEEE Transactions on Power Systems*, under review.

This design induces agents to fully internalize the scarcity adder in their balancing market bid. The merit order of the balancing market is essentially depressed by the amount of the anticipated scarcity adder, because the design is otherwise identical to the energy-only design, with the exception that the energy price is uplifted by the scarcity adder. This is fine in terms of efficient dispatch. Indeed, assuming that all agents in the balancing market correctly anticipate the scarcity adder, they simply understate their marginal cost by this adder and are otherwise efficiently dispatched, as they would be in the energy-only balancing market. On the other hand, this optimistically assumes that all agents correctly anticipate the adder, and ultimately what happens is that the true marginal costs are concealed from the balancing market. Moreover, the design fails to produce an opportunity cost for selling reserve forward, thus it generates no forward reserve price signal, as in the case of the energy-only market design.

5.5.4 Comparative overview of alternative designs

The previous analysis is summarized in Table 14. The second, third and fourth column describe the features of the alternative designs. The last three columns score these designs on the basis of how well they are able to (i) back-propagate reserve prices to forward markets, (ii) induce flexible resources to actually present themselves in balancing markets, and (iii) induce flexible resources to bid their marginal cost truthfully to the balancing platform. We find that the design which is based on a disciplined approximation of scarcity pricing is the only design which achieves all these goals simultaneously. The analytical arguments on which these derivations are based are presented in a number of scientific publications[120].

[120] Cartuyvels, J., Bertrand, G., & Papavasiliou, A. (2023). Market Equilibria in Cross-Border Balancing Platforms. *IEEE Transactions on Power Systems*, under review.

Table 14: Comparative overview of features and strengths-weaknesses of alternative market design proposals for implementing scarcity pricing based on ORDC.

	Adder on balancing price?	Adder on imbalance settlement?	Real-time market for reserve?	Back-propagation of real-time value of reserve?	Bid flexibility in balancing market?	Truthful bidding
Energy-only balancing market	No	No	No	No	Yes	Yes
Disciplined approximation of co-optimization	Yes	Yes	Yes	Yes	Yes	Yes
Adder on imbalance settlement only	No	Yes	No	Weak	Not always	Yes
Adder on the balancing price and imbalance settlement	Yes	Yes	No	No	Yes	No

Papavasiliou, A. (2020). Scarcity pricing and the missing European market for real-time reserve capacity. *The Electricity Journal, 33*(10), 106863.

Papavasiliou, A., & Bertrand, G. (2021). Market design options for scarcity pricing in European balancing markets. *IEEE Transactions on Power Systems, 36*(5), 4410-4419.

Papavasiliou, A. (2021). *Analytical Derivation of Optimal BSP / BRP Balancing Market Strategies.* https://ap-rg.eu/wp-content/uploads/2022/04/Analytical.pdf: Appendix to "Market Design Options for Scarcity Pricing in European Balancing Markets".

5.5.5 Risk neutrality, perfect competition, and circuit breakers

The analysis of the alternative designs in this section is based on assumptions of risk neutrality and perfect competition. We comment briefly on how one might attempt to generalize these analyses and the institutional relevance of these assumptions.

The first thing to point out before embarking on detailed comments is that one can always raise questions about the ability of models to perfectly represent reality. The answer is simple, they never do perfectly represent reality. But they are useful in gaining insights about disciplined designs, as well as designs that are fundamentally flawed due to the fact that they violate economic first principles. If a design fails under ideal conditions, this is a sign that one should not expect said design to overcome its failures in more complex and realistic settings. On the other hand, if a design satisfies certain desiderate under ideal conditions, this is not a guarantee that the design is robust to non-ideal conditions (and probably it would not be as no design can be expected to achieve such an impossible standard). Therefore, the relevant issue to focus on here is whether to choose between a design that is already failing under ideal conditions, or one which is at least satisfying certain key objectives under ideal settings, without any guarantees for non-ideal settings.

Market Power

To make this discussion less abstract, let us focus first on the issue of market power. One concern that has been flagged about scarcity pricing is whether it is susceptible to market power. It is, as is an energy-only market without an operating reserve demand curve. A more interesting elaboration of this question is whether scarcity pricing is “more” exposed to market power, because it creates a tendency for prices to shoot up under tight conditions. The intuition in this argument is clear, and the concern is valid, but this immediately invites the question of what we are comparing this regime against. If one is allowed to introduce market power into a model, then the

picture can become extremely ugly for many commonly encountered designs. For instance, Cournot competition or a dominant firm model in an energy-only real-time market with a largely inelastic demand can yield prices that can easily shoot up to very high values, depending on the assumptions that prevail regarding the demand function of the market. The matter of the fact is that the academic literature on the topic of market power in scarcity pricing based on ORDC is scarce if at all existent. But a warning sign here is that the results of game theoretic models of strategic interaction can vary widely depending on the specific assumptions that one adopts regarding the strategic interactions of agents (e.g. simultaneous or sequential interactions, games in prices or in quantities, dominant firms or oligopolies, etc.), thus it can sometimes be challenging to extract some prescriptive policy messages from such analyses.

The theoretical discussion provided above has some practical implications. Since any design is susceptible to market power, but some designs perform better under ideal conditions than others, it could make sense to prefer these designs combined with measures for bringing the market closer to ideal conditions as a reasonable way forward in market design. Concretely, in systems with a reasonable amount of thermal resources with well-known cost structures and technical constraints, some form of ex ante market power mitigation in real time could be a very reasonable way to cope with market power. Why, for instance, would we allow a flexible natural gas plant to bid 1000 €/MWh in the real-time energy market, or withhold its capacity altogether, if telemetry indicates that the plant is capable of producing during the real-time market interval in question, and the marginal cost of the plant during the period in question is known to be in the order of 150 €/MWh? Ex ante market power mitigation suggests we don’t allow this behavior, and rather limit the boundaries within which the plant can mark up its bid (no economic withholding), and also require it to offer its capacity if it is technically capable to do so (no technical withholding).

These ex ante measures do not prohibit scarcity pricing. Indeed, adders are computed based on the telemetry of available flexible capacity in real time, which is a physical quantity that is independent of the economic offers of plants in the balancing interval

in question. It implies that non-zero adders can emerge even if units are not allowed to mark up their energy bids. Instead, these adders emerge because real-time reserve is low, which increases the loss of load probability in real time, which in turn introduces a non-zero adder in the balancing market which also becomes the price of the real-time market for reserve, and all of this happens without units marking up their price offers. Contrast this to an energy-only design which, absent sufficient elasticity on the demand side, hopes that strategic agents mark up their bids at a "reasonable" level for recovering investment costs. Microeconomic models, coupled by empirical evidence[121], show that this does not happen.

Ex ante market power mitigation may be considered as an administrative measure which "goes against the spirit of the market", but it is prevalent, even in situations which are not entirely obvious, and in designs which claim to be more "market-based". Cost-based redispatch in congestion management, for instance, is a form of ex ante market power mitigation which is used for limiting INC-DEC gaming, to some extent, in zonal market design. It makes sense for the market monitor to exploit the information that it has at its disposal for ensuring orderly conduct in electricity markets. Things are expected to become hairier in future energy systems with storage resources, demand response, and portfolios of distributed resources, where the monitoring of the technical and economic constraints of these assets will likely be far more difficult if not impossible. However, these will not be the only resources in the market. It is therefore unclear why one would not exploit available information for the rest of the resources in the market whenever the time arrives for these resources to achieve a significantly deeper penetration in the market. Scarcity pricing based on ORDC is anyway a form of helping wheels for passing through the current market to such an ideal regime of high price elasticity. In such a future the whole point of scarcity pricing based on ORDC is likely to be diminished, if not entirely obsolete.

[121] S. Borenstein, J. Bushnell, F. Wolak, "Diagnosing market power in California's restructured wholesale electricity market", NBER working paper series, working paper 7868, September 2000.

Risk aversion

In contrast to the analysis of market power in scarcity pricing based on ORDC, for which the academic literature is relatively scarce if at all present, risk aversion has in fact been analyzed in the context of scarcity pricing based on ORDC[122]. The main phenomenon of the back-propagation of the value of reserve still holds in this generalized setting, since agents continue to face the tradeoff (in a disciplined implementation of scarcity pricing) between collecting reserve payments in a day-ahead or other forward market versus buying back that position in real time at the price set by an ORDC. In this generalized analysis, the same economic mechanisms are at play as in the risk-neutral case, namely the results are driven by the no-arbitrage conditions between energy and reserve markets and between day-ahead and real-time markets. The complication introduced by risk aversion is that real-time outcomes are now weighed against a risk-neutral measure, as opposed to the "physical"/"statistical" measure. In simple words, this means that high prices are not just averaged by their actual probability of occurrence, but instead a pessimistic view is adopted by risk-averse agents, who then demand a risk premium above the statistical average of real-time prices for trading products in the forward market. This is typical of market equilibria in settings with risk-averse agents, and this analysis is no exception.

Circuit breakers

Despite the goal of scarcity pricing to attract investment in generation, especially flexible generation, recent experience during the 2021-2022 European natural gas and electricity market crisis, which was triggered by the recovery of the economy from COVID and the Russia-Ukraine war, has reaffirmed that high energy prices are largely unacceptable and produce strong political backlash. This triggered an entire debate

[122] A. Papavasiliou, Y. Smeers, G. de Maere d'Aertrycke, "Market Design Considerations for Scarcity Pricing: A Stochastic Equilibrium Framework", The Energy Journal, vol. 42, no. 5, pp. 195-220, 2021.

about revisiting market design fundamentals[123] and re-opened issues that have been settled over decades of theoretical analysis and empirical evidence, with certain strange reforms being advocated for even at the highest level of EU governance. Fortunately, many of the bizarre and panicked proposals that threatened the fabric of European electricity market design were withdrawn and did not see their way through to implementation. One interesting topic that opened up, nevertheless, was a discussion on circuit breakers. Such a discussion was also raised during the 2021 crisis in the Texas electricity market, where the debate was specifically focused on circuit breakers for scarcity pricing[124].

The idea of circuit breakers is to cap prices when annual profits exceed a certain threshold. More specifically, circuit breakers have been inspired in other financial markets, and have also been introduced recently in the Texas market after the 2021 crisis. The idea of a circuit breaker is to allow the market to clear at marginal prices even during tight conditions, up to the point where payments to producers secure a multiple of their fixed long-term investment cost (e.g. 315,000 $/MW). The design that was discussed during the European energy crisis foresaw that once the circuit breakers would be activated, generator offers would be capped, and offers which would exceed this cap would be paid their asking price, in a paid as bid fashion, if cleared, without however setting the clearing price of the auction.

The balance that one attempts to strike with circuit breakers is, among others, to (i) not mute demand response, during extended periods of system stress, by sustaining high prices for at least a certain amount of time, (ii) while protecting consumers from exhaustingly high prices during extended periods of scarcity, (iii) to mobilize producers to deliver energy during such stressed periods, (iv) to reduce regulatory uncertainty

---

123 https://www.euractiv.com/section/electricity/opinion/the-greek-market-design-proposal-would-be-the-end-of-electricity-markets-as-we-know-them/

124 P. Cramton, “Lessons for Peru from the 2021 Texas electricity crisis”, March 17, 2021.

and missing money for producers, and (v) to trigger investment in renewable resources with low marginal costs that can displace conventional thermal technologies with unacceptably high variable costs.

The Texas circuit breaker design had a similar spirit, although the specifics relate in particular to scarcity pricing, and so do the proposals that have been set forth post-2021. Concretely, the Texas design was tailored to summer scarcity events, when electricity demand was expected to be high without a corresponding shortage in natural gas demand. The design therefore applied breakers on scarcity prices, which would be moved from 9000 $/MWh to the maximum of 2000 $/MWh or 50 times the fuel price index. Alas, the scarcity events that occurred in the Texas crisis in 2021 were winter shortage events, which also led to significant scarcity in the natural gas market and not only the electricity market, which extended for a fair amount of time. In such situations, the cascading shortage of natural gas markets which tightened electricity markets essentially voided the circuit breaker mechanism. Indeed, during these winter periods the price of natural gas shot up from its then-typical value of 4 $/MMBtu to 359.14 $/MMBtu, which essentially meant that the cap on scarcity prices was raised to 17,957 $/MWh, and was thus effectively irrelevant. Since the lessons learned from the 2021 Texas crisis, adaptations to the circuit breaker proposal have been set forth in the Texas market, for adapting the mechanism to a winter shortage.

References

European Commission. (2017, November). Commission Regulation (EU) 2017/2195 of 23 November 2017 establishing a guideline on electricity balancing (Text with EEA relevance.). Brussels: European Commission.

Electricity System Operator. (2023). *Markets Roadmap.* London: March.

Department for Business, Energy and Industrial Strategy. (2022). *Digest of UK Energy Statistics.* London: BEIS.

National Grid ESO. (2023). *Future Energy Scenarios.* London: NGESO.

Kraljic, D., Sobocan, B., Katanec, J., Logar, M., & Troha, M. (2022). Inertia constants for individual power plants. *18th International Conference on the European Energy Market (EEM)* (pp. 1-5). September: IEEE.

Papavasiliou, A. (2023). *Optimization models in electricity markets.* Cambridge, UK: Cambridge University Press.

ESO. (2023). *Assessment of investment policy and market design packages.* London: February 27.

Richstein, J. C., Lorenz, C., & Neuhoff, K. (2020). An auction story: How simple bids struggle with uncertainty. *Energy Economics*, 104784.

Ahlqvist, V., Holmberg, P., & Tangerås, T. P. (2018). *Central-versus self-dispatch in electricity markets.* Cambridge, UK: University of Cambridge, Faculty of Economics.

Stoft, S. (2002). *Power system economics: designing markets for electricity.* Piscataway: IEEE press.

Boiteux, M. (1960). Peak-load pricing. *The Journal of Business, 33*(2), 157–179.

Lete, Q., Smeers, Y., & Papavasiliou, A. (2023). Investment with Market-Based Redispatch. *Energy Journal*, under review.

Papavasiliou, A., Cartuyvels, J., Bertrand, G., & Marien, A. (2023). Implementation of scarcity pricing without co-optimization in European energy-only balancing markets. *Utilities Policy*, forthcoming.

Electric Reliability Council of Texas. (2014). *Purpose of ORDC, methodology for implementing ORDC, settlement impacts of ORDC.* Austin, TX: ERCOT market training.

Belgian Commission for Electricity and Gas Regulation. (2021). *Study on the implementation of a scarcity pricing mechanism in Belgium.* Brussels, Belgium: CREG.

Papavasiliou, A. (2020). Scarcity pricing and the missing European market for real-time reserve capacity. *The Electricity Journal, 33*(10), 106863.

Jevons, W. S. (1879). *The theory of political economy.* London: Macmillan and Company.

OFGEM. (2023). *Analysis of the first phase of the Electricity Balancing Significant Code Review.* London, UK: OFGEM. Retrieved from https://www.ofgem.gov.uk/sites/default/files/docs/2018/08/analysis_of_the_first_phase_of_the_electricity_balancing_significant_code_review_as_final_

PSR. (2023). *Market and regulatory analysis: Great Britain.* Rio de Janeiro, Brazil: April.

Giesbertz, P. (2021). *The power market design column*. Retrieved from The scarcity of scarcity pricing: https://www.linkedin.com/pulse/power-market- design-column-scarcity-pricing-paul-giesbertz/

ELIA. (2021). *Final report on ELIA's findings regarding the design of a scarcity pricing mechanism for implementation in Belgium.* Brussels, Belgium: ELIA.

Cartuyvels, J., Bertrand, G., & Papavasiliou, A. (2023). Market Equilibria in Cross-Border Balancing Platforms. *IEEE Transactions on Power Systems*, under review.

Papavasiliou, A. (2021). *Analytical Derivation of Optimal BSP / BRP Balancing Market Strategies.* https://ap-rg.eu/wp-content/uploads/2022/04/Analytical.pdf: Appendix to "Market Design Options for Scarcity Pricing in European Balancing Markets".

Hogan, W. W. (2013). Electricity scarcity pricing through operating reserves. *Economics of Energy and Environmental Policy, 2*(2), 65-86.

Papavasiliou, A., & Smeers, Y. (2017). Remuneration of Flexibility using Operating Reserve Demand Curves: A Case Study of Belgium. *The Energy Journal, 38*.

OFGEM. (2022). *Annual Report on the Operation of the CM 2020/21 and 2021/22.* London, UK: OFGEM.

Papavasiliou, A., & Bertrand, G. (2021). Market design options for scarcity pricing in European balancing markets. *IEEE Transactions on Power Systems, 36*(5), 4410-4419.

Papavasiliou, A., Smeers, Y., & de Maere d'Aertrycke, G. (2019). *Study on the general design of a mechanism for the remuneration of reserves in scarcity situations.* Louvain la Neuve, Belgium: https://ap-rg.eu/wp-content/uploads/2020/07/CREGReportFinal.pdf.

Cartuyvels, J., & Papavasiliou, A. (2023). Calibration of Operating Reserve Demand Curves Using a System Operation Simulator. *IEEE Transactions on Power Systems*.

Zarnikau, J., Zhu, S., Woo, C. K., & Tsai, C. (2020). Texas's operating reserve demand curve's generation investment incentive. *Energy Policy, 137*, 111143.

Zhou, Z., & Botterud, A. (2014). Dynamic scheduling of operating reserves in co-optimized electricity markets with wind power. *IEEE Transactions on Power Systems, 29*(1), 160-171.

Papavasiliou, A. (2021). *Overview of EU Capacity Remuneration Mechanisms.* Louvain la Neuve, Belgium: UCLouvain.

Papavasiliou, A., & Mou, Y. (2023). *Modeling Energy-Only Markets in the Presence of Shortage Pricing and Capacity Remuneration Mechanisms.* Athens, Greece: NTUA.

Mou, Y., Papavasiliou, A., Hartz, K., Dusolt, A., & Redl, C. (2023). An analysis of shortage pricing and capacity remuneration mechanisms on the pan-European common energy market. *Energy Policy*, under review.

NGESO. (2023, September 23). *Short-term operating reserve (STOR)*. Retrieved from https://www.nationalgrideso.com/industry-information/balancing-services/reserve-services/short-term-operating-reserve-stor

Cramton, P. C., & Stoft, S. (2006). *The convergence of market designs for adequate generating capacity with special attention to the CAISO's resource adequacy problem.* Cambridge, MA: MIT Center for Energy and Environmental Policy Research.

N-SIDE. (2023). *Svk project on scarcity pricing.* https://www.svk.se/siteassets/om-oss/rapporter/2023/report-scarcity_pricing-phase-1.pdf : Papavasiliou, Anthony.

NGESO. (2023, September 23). *Mandatory frequency response*. Retrieved from https://www.nationalgrideso.com/industry-information/balancing-services/frequency-response-services/mandatory-frequency-response

Zhou, Z., Levin, T., & Conzelmann, G. (2016). *Survey of US ancillary services markets.* Argonne National Lab.(ANL), Argonne, IL (United States).

N-SIDE. (2022). *Co-Optimization of Energy and Balancing Capacity in the European Single Day-Ahead Coupling.* Louvain la Neuve, Belgium: N-SIDE.

ENTSO-E. (2021). *Implementation impact assessment for the methodology for a co-optimised allocation process of cross-zonal capacity for the exchange of balancing capacity or sharing of reserves.* Brussels, Belgium: ENTSO-E.

Van den Bergh, K., Bruninx, K., & Delarue, E. (2018). Cross-border reserve markets: network constraints in cross-border reserve procurement. *Energy Policy, 113*, 193–205. doi:https://doi.org/10.1016/j.enpol.2017.10.053

Hogan, W. W., & Pope, S. L. (2019). *PJM Reserve Markets: Operating Reserve Demand Curve Enhancements.* Cambridge, MA: Harvard University Kennedy School of Government.

Papavasiliou, A., Bouso, A., Apelfröd, S., Wik, E., Gueuning, T., & Langer, Y. (2022). Multi-Area Reserve Dimensioning using Chance-Constrained Optimization. *IEEE Transactions on Power Systems, 37*(5), pp. 3982-3994.

Zheng, T., & Litvinov, E. (2008). Contingency-based zonal reserve modeling and pricing in a co-optimized energy and reserve market. *IEEE transactions on Power Systems, 23*(2), 277-286.

Chen, Y., Gribik, P., & Gardner, J. (2014). Incorporating post zonal reserve deployment transmission constraints into energy and ancillary service co-optimization. *IEEE Transactions on Power Systems, 29*(2), 537-549.

Vrakopoulou, M., Margellos, K., Lygeros, J., & Andersson, G. (2013). A probabilistic framework for reserve scheduling and N-1 security assessment of systems with high wind power penetration. *IEEE Transactions on Power Systems, 28*(4), 3885–3896.

Roald, L., Misra, S., Krause, T., & Andersson, G. (2016). Corrective control to handle forecast uncertainty: A chance constrained optimal power flow. *IEEE Transactions on Power Systems, 32*(2), 1626–1637.

N-SIDE; AFRY. (2020). *CZC allocation with co-optimization.* Louvain la Neuve, Belgium: November.

Nohadani, O., & Kartikey, S. (2018). Optimization under decision-dependent uncertainty. *SIAM Journal on Optimization, 28*(2), 1773-1795.

Bemporad, A., Filippi, C., & Torrisi, F. D. (2004). Inner and outer approximations of polytopes using boxes. *Computational Geometry, 27*(2), 151-178.

Caramanis, M., Ntakou, E., Hogan, W. W., Chakrabortty, A., & Schoene, J. (2016). Co-optimization of power and reserves in dynamic T&D power markets with nondispatchable renewable generation and distributed energy resources. *Proceedings of the IEEE, 104*(4).

Cho, J., & Papavasiliou, A. (2023). Exact Mixed-Integer Programming Approach for Chance-Constrained Multi-Area Reserve Sizing. *IEEE Transactions on Power Systems*, forthcoming.

Anselm Eicke, T. S. (2022). Fighting the wrong battle? A critical assessment of arguments against nodal electricity prices in the European debate. *Energy Policy*, 170:113220.

Harvey, S. H. (2010). Nodal and Zonal Congestion Management and the Exercise of Market Power. *Work. Pap. Harv. Univ*.

Bertsimas, D., Litvinov, E., Sun, X. A., Zhao, J., & Zheng, T. (2013). Adaptive Robust Optimization for the Security Constrained Unit Commitment Problem. *IEEE Transactions on Power Systems, 28*(1), 52-63.

Herrero, I., Rodilla, P., & Battle, C. (2020). Evolving Bidding Formats and Pricing Schemes in USA and Europe Day-Ahead Electricity Markets. *Energies*, 1-21.

United States Federal Energy Regulatory Commission. (2014). *Price formation in organized wholesale electricity markets.* Washington, DC: FERC.

Oren, S. (2005). *Market design and gaming in competitive electricity markets.* Berkeley, CA: UC Berkeley.

Oren, S. (2001). Design of Ancillary Service Markets. *Proceeding of the 34th Hawaii International Conference on Systems Sciences HICSS 34.* Maui, Hawaii: HICSS.

Papavasiliou, A., He, Y., & Svoboda, A. (2015). Self-Commitment of Combined Cycle Units under Electricity Price Uncertainty. *IEEE Transactions on Power Systems*, 1690-1701.

Hogan, W. W., & Ring, B. J. (2003). *On minimum-uplift pricing for electricity markets.* Cambridge, MA: Harvard Electricity Policy Group.

O'Neill, R. P., Sotkiewicz, P. M., Hobbs, B. F., Rothkopf, M. H., & Stewart, W. R. (2005). Efficient market-clearing prices in markets with nonconvexities. *European journal of operational research, 164*(1), 269-285.

NEMO committee. (2020). *EUPHEMIA public description: single price coupling algorithm.*

Schiro, D. A., Zheng, T., Zhao, F., & Litvinov, E. (2015). Convex hull pricing in electricity markets: Formulation, analysis, and implementation challenges. *IEEE Transactions on Power Systems, 31*(5), 4068-4075.

Stevens, N., & Papavasiliou, A. (2022). Application of the Level Method for Computing Locational Convex Hull Prices. *IEEE Transactions on Power Systems, 37*(5), 3958-3968.

Morales-España, G., Gentile, C., & Ramos, A. (2015). Tight MIP formulations of the power-based unit commitment problem. *ORSpectrum, 37*(4), 929–950.

Morales-España, G., Latorre, J. M., & Ramos, A. (2013). Tight and compact MILP formulation for the thermal unit commitment problem. *IEEE transactions on power systems, 28*(4), 4897–4908.

Fabra, N., von der Fehr, R.-H., & Harbord, D. (2006). Designing electricity auctions. *The RAND Journal of Economics, 37*(1), 23-46.

Sioshansi, R., & Nicholson, E. (2011). Towards equilibrium offers in unit commitment auctions with nonconvex costs. *Journal of Regulatory Economics, 40*(1), 41-61.

Liberopoulos, G., & Andrianesis, P. (2006). Critical review of pricing schemes in markets with non-convex costs. *Operations Research, 64*(1), 17-31.

Wang, G. U. (2012). On Nash equilibria in duopolistic power markets subject to make-whole uplift. *51st IEEE Conference on Decision and Control (CDC)* (pp. 472-477). IEEE.

FERC. (2013). *Make-Whole Payments and Related Bidding Strategies, Docket Nos. IN11-8-000, IN13-5-000.* Washington, DC: Federal Energy Regulatory Commission.

FERC. (2013). *Make-Whole Payments and Related Bidding Strategies, Docket Nos. IN11-8-000, IN13-5-000.* Washington DC: Federal Energy Regulatory Commission.

Oren, S. S., & Ross, A. M. (2005). Can we prevent the gaming of ramp constraints? *Decision Support Systems, 40*(3-4), 461-471.

Bertsekas, D. P., & Sandell, N. R. (1982). Estimates of the duality gap for large-scale separable nonconvex optimization problems. *21st IEEE conference on decision and control*, (pp. 782-785).

Madani, M., Ruiz, C., Siddiqui, S., & Van Vyve, M. (2018). *Convex hull, IP and European electricity pricing in a european power exchanges setting with efficient computation of convex hull prices.* Baltimore, MD: arXiv.

Hao, S., Angelidis, G. A., Singh, H., & Papalexopoulos, A. D. (1998). Consumer payment minimization in power pool auctions. *IEEE Transactions on Power Systems, 13*(3), 986-991.

Zhao, F., Luh, P. B., Yan, J. H., Stern, G. A., & Chang, S.-C. (2008). Payment cost minimization auction for deregulated electricity markets with transmission capacity constraints. *IEEE Transactions on Power Systems, 23*(2), 532-544.

Zhao, F., Luh, P. B., Yan, J. H., Stern, G. A., & Chang, S.-C. (2010). Bid cost minimization versus payment cost minimization: A game theoretic study of electricity auctions. *IEEE Transactions on Power Systems, 25*(1), 181-194.

Litvinov, E., Zhao, F., & Zheng, T. (2009). Alternative auction objectives and pricing schemes in short-term electricity markets. 2009 IEEE Power and Energy Society General Meeting: IEEE.

Luh, P. B., Blankson, W. E., Chen, Y., Yan, J. H., Stern, G. A., Chang, S.-C., & Zhao, F. (2006). Payment cost minimization auction for deregulated electricity

markets using surrogate optimization. *IEEE Transactions on Power systems, 21*(2), 568-578.

Fernandez-Blanco, R., Arroyo, J. M., & Alguacil, N. (2011). A unified bilevel programming framework for price-based market clearing under marginal pricing. *IEEE Transactions on Power Systems, 27*(1), 517-525.

Milgrom, P. (2004). *Putting auction theory to work.* Cambridge University Press.

Andrianesis, P., Bertsimas, D., Caramanis, M. C., & Hogan, W. W. (2021). Computation of convex hull prices in electricity markets with non-convexities using Dantzig-Wolfe decomposition. *IEEE Transactions on Power Systems, 37*(4), 2578-2589.

Stevens, N., Papavasiliou, A., & Smeers, A. (2024). The Many Advantages of Convex Hull Pricing for the European Electricity Auction. *Energy Economics, under review*.

ESO. (2022). *Operability Strategy Report.*

BEIS. (2022). *British Energy Security Strategy.* Retrieved from https://www.gov.uk/government/publications/british-energy-security-strategy/british-energy-security-strategy

DESNZ. (2023). *Digest of Energy Statistics.*

Carbon Trust. (2021). *Flexibility in Great Britain.*

Papavasiliou, A., Smeers, Y., & de Maere d'Aertrycke, G. (2021). Market Design Considerations for Scarcity Pricing: A Stochastic Equilibrium Framework. *The Energy Journal, 42*(5), 195-220.

Appendix A: Efficiency gains of co-optimization

In this appendix we provide a motivating example[125] which illustrates the efficiency losses that result from the artificial split of an inherently interdependent procedure, that of allocating energy and ancillary services. To illustrate our point, we consider three generators with the following technical-economic characteristics:

- Generator 1: Maximum power of 100 MW, unlimited ramp rate, marginal cost of 0 €/MWh
- Generator 2: Maximum power of 100 MW, ramp rate of 1 MW/minute, marginal cost of 10 €/MWh
- Generator 3: Maximum power of 100 MW, ramp rate of 5 MW/minute, marginal cost of 80 €/MWh

For a given period, there is an inelastic demand for energy of 100MWh and an inelastic demand for balancing capacity for a reserve product of 100 MW in the system. Suppose that the response time of the reserve product in question is 10 minutes.

The opportunity cost of the generators in the reserve market can be expressed as

$$\max\left(0, \lambda^* - MC_g\right)$$

This is what the generators would bid in the reserve market. For instance, suppose that the agents correctly anticipate that the equilibrium energy price will be 10 €/MWh. Then, the generators bid as follows:

- Generator 1 offers 100 MW at 10 €/MWh
- Generator 2 offers 10 MW at 0 €/MWh

[125] The example presented here is based on (Papavasiliou A. , 2023).

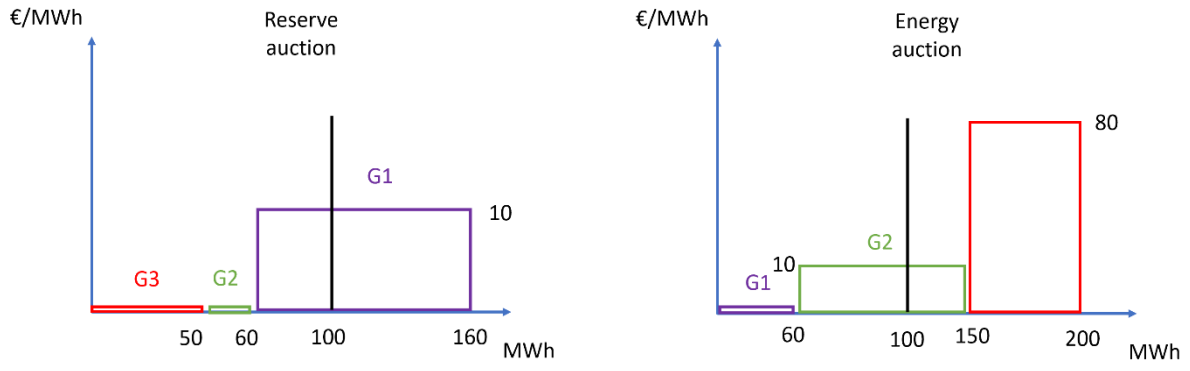
- Generator 3 offers 50 MW at 0 €/MWh

The reserve auction clears generator 1 for 40 MW, generator 2 for 10 MW, and generator 3 for 50 MW, as indicated in Figure 21. The clearing price is determined by generator 1, which is at the money. Thus, the reserve price becomes 10 €/MWh. Given these reserve allocations, generators enter the energy auction with the following offers:

- Generator 1 offers 60 MW at 0 €/MWh
- Generator 2 offers 90 MW at 10 €/MWh
- Generator 3 offers 50 MW at 80 €/MWh

The outcome of the energy auction is to match the generators in order of increasing marginal cost, as indicated in Figure 21. Thus, generator 1 is matched for 60 MW, generator 2 is matched for 40 MW, and generator 3 is not matched. The market price is set by generator 2, which is at the money, to 10 €/MWh.

Figure 21: Outcome of the sequential clearing of energy and reserves. Source: (Papavasiliou A. , Optimization models in electricity markets, 2023).



Note that this outcome is identical to that of co-optimization. Specifically, with an energy price and a reserve price both equal to 10 €/MWh, the aforementioned allocation is indeed profit maximizing for the generators.

To identify inefficiencies in sequential clearing, let us assume that generator 2 anticipates the energy price to be equal to 21 €/MWh instead of 10 €/MWh. With this assumption, the dispatch is perturbed in the case of sequential clearing, because generator 2 submits an opportunity cost of 11 €/MWh to the reserve auction, and is cleared after generator 1 in this auction. This is inefficient, because generator 1 should be held aside for energy to the greatest extent possible. Specifically, in this scenario, the reserve auction clears generator 1 for 50 MW, generator 2 for 0 MW, and generator 3 for 50 MW, with the reserve price being set by generator 2, which is at the money, to 11 €/MWh. The energy auction then clears generator 1 for 50 MW and generator 2 for 50 MW, at a price of 10 €/MWh. The total cost of this dispatch is (50 MWh * 10 €/MWh) = 500 €. Contrast this to the efficient sequential dispatch, which amounts to (40 MWh * 10 €/MWh) = 400 €, thus a 20% increase in cost.

The resulting dispatch of generators in energy and reserves is indicated in Table 15. Note that MWh as a unit of measurement of reserve should be interpreted as the reservation of 1 MW of headroom for 1 hour.

Table 15: Energy and reserves dispatch in the case of co-optimization and sequential clearing of energy and reserves.

Generator	Co-optimization	Sequential clearing
G1	Energy: 60 MWh Reserve: 40 MWh	Energy: 50 MWh Reserve: 50 MWh
G2	Energy: 40 MWh Reserve: 10 MWh	Energy: 50 MWh Reserve: 0 MWh
G3	Energy: 0 MWh Reserve: 50 MWh	Energy: 0 MWh Reserve: 50 MWh

The implied cost of each dispatch is broken down in Table 16. Note that the table is calculating the system cost, as opposed to the total settlements in the two markets. The latter is the same, thus the loss in welfare ends up being absorbed by the

generators in this example. The example below illustrates how efficient dispatch can be distorted in sequential market clearing, due to errors in anticipating the efficient energy price. Such errors are more likely to occur in an environment with multiple market clearing intervals (which are increasing in number as the temporal granularity of market clearing platforms increases), intertemporal interdependencies, and cross-product dependencies (especially in a market such as the one of GB, which introduces numerous interdependent ancillary services products). For instance, in a day-ahead market such as the EU common market with 96 15-minute intervals and FCR, upward/downward aFRR, upward/downward mFRR, and RR, we require the computation of 96*(1+2+2+1) = 576 prices. Perfect anticipation is out of the question and dispatch inefficiencies can thus be expected in sequential clearing. The GB market has fewer time intervals, given its 30-minute time resolution[126], but more ancillary services, thus a similar challenge is anticipated. This is an important reason why co-optimization can be expected to deliver superior performance in terms of economic efficiency.

Table 16: Cost breakdown for the co-optimization and sequential clearing approach in the case of incorrect anticipation of market clearing prices by generator 1.

Generator	Co-optimization	Sequential clearing
G1	(60 MWh) x (0 €/MWh) = 0 €	(50 MWh) x (0 €/MWh) = 0 €
G2	Energy: (40 MWh) x (10 €/MWh) = 400 €	(50 MWh) x (10 €/MWh) = 500 €
G3	Energy: (0 MWh) x (0 €/MWh) = 0 €	(0 MWh) x (80 €/MWh) = 0 €
Total	400 €	500 €

[126] The settlement period in Great Britain is 30 minutes long. However, the fsinest products in the day-ahead auctions are 1 hour long for the moment. Intraday auctions include 30-minute products.

Appendix B: Equivalence of co-optimization versus sequential clearing of energy and reserves

In this appendix we discuss the equivalence between sequential clearing and the co-optimization of energy and reserves.

Sequential clearing consists of a reserve auction followed by an energy auction. The reserve auction model can be expressed as follows:

$$\min_{r \geq 0} \sum_{g \in G} OC_g(r_g)$$

$$(\lambda R): R - \sum_{g \in G} r_g = 0$$

$$(\mu R_g): r_g \leq R_g, g \in G$$

Here, the notation is as follows:

- G: the set of generators in the market
- r_g: the amount of reserve provided by generator g
- R: reserve requirement
- $OC_g(\cdot)$: opportunity cost of generator g for offering reserve
- R_g: limit of reserve capacity that can be offered by generator g

The opportunity cost function of generator g is the sensitivity of generator profits to an increased commitment of reserve in the reserve market, i.e. the profit that is foregone from the energy market due to a commitment of reserve in the reserve market. Mathematically, the profit maximization problem of generator g can be expressed as follows:

$$OC_g(r_g) = \max_{p \geq 0}(\lambda - MC_g) \cdot p_g$$

$$(\mu_g): p_g + r_g \leq P_g$$

The notation for this model is as follows:

- p_g: the energy sold by generator g in the energy market
- MC_g: the marginal cost of generator g
- P_g: the nominal capacity of generator g

The full set of conditions which characterize the sequential market is as follows:

- Seq-A: The set of KKT conditions that correspond to the reserve market
- Seq-B: The set of KKT conditions that correspond to the profit maximization model of each generator in the energy market
- Seq-C: The energy market clearing condition

The set of conditions Seq-A can be summarizedby the following equations:

$$\sum_{g \in G} r_g = R$$

$$0 \leq \mu R_g \perp R_g - r_g \geq 0, g \in G$$

$$0 \leq r_g \perp \pi_g + \mu R_g - \lambda R \geq 0, g \in G$$

Here, π_g is a subgradient of the opportunity cost function OC_g. Convex analysis implies that this subgradient is the sensitivity of the profit to a change in reserve, i.e. $\pi_g = \mu_g$, where μ_g is the dual multiplier of the constraint of the energy market profit maximization problem.

The set of conditions Seq-B for all generators can be summarized as follows:

$$0 \leq \mu_g \perp P_g - p_g - r_g \geq 0, g \in G$$

$$0 \leq p_g \perp MC_g + \mu_g - \lambda \geq 0, g \in G$$

The energy market clearing condition Seq-C is described as follows:

$$\sum_{g \in G} p_g = D$$

Here, D is the total energy demand in the energy market.

Let us now consider the model that co-optimizes energy and reserves:

$$\min_{p,r \geq 0} \sum_{g \in G} MC_g \cdot p_g$$

$$(\lambda): D - \sum_{g \in G} p_g = 0$$

$$(\lambda R): R - \sum_{g \in G} r_g = 0$$

$$(\mu_g): p_g + r_g \leq P_g, g \in G$$

$$(\mu R_g): r_g \leq R_g, g \in G$$

The KKT conditions of the co-optimization model are described as follows:

$$\sum_{g \in G} p_g = D$$

$$\sum_{g \in G} r_g = R$$

$$0 \leq \mu_g \perp P_g - p_g - r_g \geq 0, g \in G$$

$$0 \leq \mu R_g \perp R_g - r_g \geq 0, g \in G$$

$$0 \leq p_g \perp MC_g + \mu_g - \lambda \geq 0, g \in G$$

$$0 \leq r_g \perp \mu_g + \mu R_g - \lambda R \geq 0, g \in G$$

These KKT conditions are identical to those of the sequential model, if one notes that $\pi_g = \mu_g$ in the sequential market model, which is indeed true, as argued before.

Appendix C: Mathematical description of scarcity pricing

An energy-only market can be described as follows.

$$\max_{p \geq 0, d \geq 0} V \cdot d - \sum_{g \in G} MC_g \cdot p_g$$

$$(\lambda): d - \sum_{g \in G} p_g = 0$$

$$(\nu): d \leq D$$

$$(\mu_g): p_g \leq P_g, g \in G$$

The notation used here follows the notation of appendix B. The additional notation is the following:

- d: demand of consumers that is actually served (decision variable)
- D: load of consumers (parameter)

The KKT conditions which characterize the optimal dispatch and supporting equilibrium prices of this energy-only market are described as follows:

$$\sum_{g \in G} p_g = d$$

$$0 \leq \mu_g \perp P_g - p_g \geq 0, g \in G$$

$$0 \leq p_g \perp MC_g + \mu_g - \lambda \geq 0, g \in G$$

$$0 \leq \nu \perp D - d \geq 0$$

$$0 \leq d \perp \lambda - V \geq 0$$

Scarcity in an energy-only market means that the system is so tight that it fails to fully cover load. Mathematically, this means that $d < D$. The second-to-last KKT condition above implies that, when this happens, then consumer surplus is 0: $\nu = 0$. For this to occur, the energy price must be equal to the valuation of consumers: $d > 0 \Rightarrow \lambda = V$. Thus, during scarcity periods the energy price becomes equal to the valuation of consumers. This is typically a very high price, which occurs rarely (3 hours per year, if the system adheres to its target reliability standard), which implies significant volatility in energy prices, and thus increased investment risk.

The co-optimization of energy and reserves with an operating reserve demand curve can be described as follows:

$$\max_{p \geq 0, d \geq 0} V \cdot d - \sum_{g \in G} MC_g \cdot p_g + \int_{x=0}^{dr} VR(x)dx$$

$$(\lambda): d - \sum_{g \in G} p_g = 0$$

$$(\lambda R): dR - \sum_{g \in G} r_g = 0$$

$$(\nu): d \leq D$$

$$(\mu_g): p_g + r_g \leq P_g, g \in G$$

The notation follows that of appendix B, with the addition of an operating reserve demand curve:

- $VR(\cdot)$: operating reserve demand curve, i.e. willingness to pay of the TSO for increments of reserve capacity
- dr: demand for reserve capacity

The KKT conditions of this model can be expressed as follows:

$$\sum_{g \in G} p_g = d$$

$$\sum_{g \in G} r_g = dr$$

$$0 \leq \mu_g \perp P_g - p_g \geq 0, g \in G$$

$$0 \leq p_g \perp MC_g + \mu_g - \lambda \geq 0, g \in G$$

$$0 \leq r_g \perp \mu_g - \lambda R \geq 0, g \in G$$

$$0 \leq \nu \perp D - d \geq 0$$

$$0 \leq d \perp \lambda - V \geq 0$$

$$0 \leq dr \perp -VR + \lambda R \geq 0$$

Note that, in this setting, even if demand is fully served ($d = D$), the system can be tight, in the sense that the leftover reserve may be low. In such a situation, the price for reserve is set by the operating reserve demand curve: $dr > 0 \Rightarrow \lambda R = VR$. If this occurs, it implies that a generator which offers reserve earns a profit margin from the reserve market which is equal to the price of reserve: $r_g > 0 \Rightarrow \mu_g = \lambda R$. If that same generator also decides to allocate energy to the energy market, then it must be indifferent between the profits that it achieves in both markets: $p_g > 0 \Rightarrow \mu_g = \lambda - MC_g = \lambda R$. Which implies that the energy price is no longer equal to the marginal cost of this peaking unit, but rather to the marginal cost of the unit plus the price of reserve. This is how scarcity pricing through operating reserve demand curves uplifts energy prices even in the absence of demand-side elasticity in the energy market, just by virtue of price elasticity in the reserve market combined with a no-arbitrage condition between the energy and the reserve market.

Appendix D: Correspondence of GB, EU and US terminology

In this appendix we provide a glossary that (approximately) maps terms that are used in GB, EU and US market design literature and industry practice, to better trace correspondences between international markets and operations. The mappings are not perfect, but should rather be understood as approximations, since ancillary services are rarely identical between different systems.

Table 17: A table of GB, EU and US power system operations and electricity market design terminology.

EU	GB	US
Automatic frequency restoration reserve (aFRR)	Frequency response / reserve	Frequency responsive reserve (not specified as automatic)
Balancing capacity	Balancing capacity	Reserve
Balancing energy	Balancing energy / bid offer acceptance (BOA)	Real-time energy
Balancing market	Balancing mechanism	Real-time energy market
Balancing price	Cash-out price	Real-time energy price
Balancing responsible party	BSC Party	N/A
Balancing service provider	Balancing service provider	Price-responsive real-time energy offer
Downward balancing offer	Bid	Real-time energy demand bid
Frequency containment reserve	Frequency response	Automatic generation control
Frequency restoration reserve	Reserve	Frequency responsive reserve

Imbalance	Imbalance / Net Imbalance Volume (NIV)	Imbalance
Imbalance charge	Imbalance charge	Uninstructed deviation penalties
Manual frequency restoration reserve (mFRR)	Reserve	Operating reserve / load following (in some systems)
Nomination	Physical Notification	N/A
Replacement reserve	STOR, Slow Reserve	Operating reserve (in some systems) / contingency reserve
Upward balancing offer	Offer	Real-time energy supply bid