1. Introduction

> "Apple Inc said it would delay for one month the international roll-out of its iPad tablet computer, due to heavy demand and swift sales after its launch in the United States. (...)
>
> The delays risk irking customers and creating headaches for retailers banking on the popular tablet computer to draw in store traffic. But Apple may reap higher margins from additional sales made at U.S. Apple stores and through its Website (...)."
>
> (Paul and Madway, 2010)

This quote illustrates an important decision many businesses face: Which customers to serve in times when supplies are low? In practice, many companies encounter this allocation problem on a regular basis, not just when planning the introduction of a new product.[1]

Taking a more general perspective, allocation decisions are ubiquitous in business environments. Examples include capital budgeting (Maier and van der Weide, 1976), marketing resource allocations (Mantrala et al., 1992), capacity allotment (Mallik and Harker, 2004) or inventory allocation decisions (Federgruen and Zipkin, 1984). To put it plainly, the problem of optimally matching the supply of scarce supplies with customer demand is at the heart of most supply chain management (SCM) initiatives.

Surprisingly, most of the existing SCM literature focuses primarily on suggesting appropriate production systems, planning structures or forecasting approaches to prevent any mismatch between supply and demand from occurring in the first place. Nevertheless, supply-demand imbalances materialize often in practice and need to be handled, as illustrated by the Apple example above. A main underlying reason is that supply adjustments are usually sluggish or even impossible to make while demands are often very volatile and difficult to forecast. In light of (nearly) fixed supplies, an obvious short-term solution to such shortage problems is to adjust prices to manage demand (Harris and Pinder, 1995). Yet, many companies are reluctant to raise prices even if supplies are tight.

Furuhata (2009, p. 8) summarized two key reasons why many firms prefer rationing over price adjustments in shortage situations:

1. Many business transactions are regulated by frame contracts with negotiated, fixed prices which cannot be adjusted in the short term. This leaves quantity allocations as the preferred alternative.

[1] An overview of approaches to select the initial markets for new product introductions is given in Geunes et al. (2005).

S. Vogel, *Demand Fulfillment in Multi-Stage Customer Hierarchies,* Produktion und Logistik,
DOI 10.1007/978-3-658-02864-0_1, © Springer Fachmedien Wiesbaden 2014

2. If price adjustments are implemented, large price increases may at times be necessary for products with rather price-inelastic demand to achieve a balance with given supply quantities. This lets customers often be more comfortable with quantity reductions than with price hikes.

Nevertheless, such allocation problems are addressed relatively seldom in the literature. De Véricourt et al. (2002) argued that there is often an implicit assumption that the efficiency of allocation decisions has only little overall effect and may be ignored in practice. As a result, with allocation decisions perceived to be more of an operational nature, companies should rather focus more on making proper strategic design choices.

However, this negligence constitutes a misconception. Doubtlessly, clever allocation decisions have beneficial short- and long-term effects. In the short-term, previous research has found that exploiting differences in customer profitability—or more generally: exploiting *customer heterogeneity*—indeed has the potential to contribute significantly to overall company profits. The corresponding decisions, i.e. deciding which customers to serve and how, are usually referred to as *demand fulfillment* (Fleischmann and Meyr, 2003). Recent literature contributions which stress the short-term benefits of exploiting customer heterogeneity in demand fulfillment include Ball et al. (2004), Meyr (2009), Tsai and Wang (2009) or Gao et al. (2012).

Furthermore, a clear focus on customer profitability is also beneficial in the long term, due to a straightforward argument: High service levels—even if overall stocks are low—tend to increase customer satisfaction, and in turn, customer loyalty. As most companies depend on repeat business, giving preferred service to those customers who contribute most to company profits is thus a reasonable decision. This link between customer satisfaction, loyalty and profitability has already been studied extensively in the literature (see, e.g., Reichheld (1993) or Anderson et al. (1994)).

To realize the benefits of customer orientation in demand planning and demand fulfillment, proper models, decision rules and software systems are required (see Quante et al., 2009b). Unfortunately, most of the existing contributions in that area focus on situations where data availability and information transparency are of no concern. The related modeling approaches build on the assumption that customer profitability can be managed from a central perspective and that comprehensive and detailed information is available to make appropriate demand fulfillment decisions.

More specifically, it is usually assumed in the existing literature that a central, omniscient planner makes these decisions. Accordingly, individual customers or customer segments can easily be sorted: Both simple ordinal rankings in the form of a priority index (cf. Pibernik, 2006) and a ratio scale (e.g. a profitability measure, see Meyr (2009)) prevail. In these approaches, there is an obvious sequence according to which customers shall be served if supplies are scarce. However, in larger organizations, especially in those which serve a geographically dispersed customer base, neither is a central administration feasible nor can such a sequence of the customer segments be established easily, usually due to information asymmetries.

Contrariwise, such larger organizations with a more comprehensive customer base are typically managed in a decentral manner, i.e. with the help of a multi-stage hierarchy. Individual key accounts or entire customer segments represent the leaf nodes of a hierarchical tree structure. These leaf nodes differ from each other in terms of profitability—or more generally, in terms of importance. Higher hierarchical levels correspond to larger, coarser segments. The resulting arborescent structure will be referred to in the following as a *heterogeneous customer hierarchy*.[2] An intuitive example in practice is a geography-based customer hierarchy where the leaf nodes represent customer segments in local sales districts and where higher levels correspond to, e.g., national sales organizations.

In such multi-stage customer hierarchies, rationing decisions are particularly challenging as lower-level planners typically enjoy an information advantage over higher-level planners. From a technical point of view, modern business warehousing (BW) software can easily deal with vast amounts of data. They are capable of handling a multitude of customer segments and an enormous number of different products. If such solutions are available and fully integrated into all business processes and all legal entities, the problem of decentral allocation and demand fulfillment decisions under decentral information is less material. However, in many practical cases, the initial and the running costs associated with such BW systems are particularly high. Furthermore, many companies still prefer to make operational decisions such as demand fulfillment on a decentral basis. As a result, integrated BW systems for sales and demand fulfillment decisions which have a global scope are rather an exception in practice.

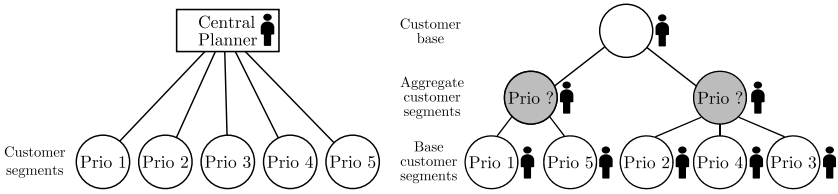
Therefore, in the absence of integrated BW systems, the ranking of the customer segments at the leaf nodes in such a customer hierarchy is no longer immediately obvious for a planner at a higher hierarchical level. Consider the following Figure 1.1 for illustration: The situation at the left-hand side, in Figure 1.1a, corresponds to the traditional demand fulfillment approach where a single planner can immediately observe all customer segments in the entire customer base of the company. The planner may thus easily order the priorities or profitabilities of all these (heterogeneous) customer segments. For simplicity, such a situation will be referred to in this thesis as a *flat partitioning* of customer segments.[3] In contrast, the right-hand side representation, Figure 1.1b, gives a *multi-stage customer hierarchy*:[4] Here, the planner at the top may only observe the demands and aggregate profitabilities of the immediate next lower level directly. Detailed data at the level of the *base customer segments* at the leaf nodes is hidden. Instead, this lower-level data can only be observed by the planners at the two intermediate nodes at the middle level (highlighted in gray in Figure 1.1b).

Therefore, once supplies are scarce, the planner at the top level has to make a first allocation based on some aggregate measure of priority. In a second step, the allocation from the intermediate level to the leaf-node level will be performed in a decentral manner.

[2] A more formal definition will follow in Section 3.3.

[3] This term first appeared in Christou and Ponis (2009).

[4] It is convenient to refer to this situation as a three-level customer hierarchy, i.e. to count the uppermost node, representing the entire customer base, as a separate level. Following this logic, a flat partitioning always corresponds to a two-level customer hierarchy.



(a) Focus in the existing literature: flat partitioning of the customer segments with central control

(b) Focus in this thesis: multi-stage customer hierarchy with decentral control

Figure 1.1. – Perspectives on customer heterogeneity in demand fulfillment

Note, however, that in this situation, the ranking of the aggregate customer segments is not obvious for the planner at the root node.[5] In Figure 1.1b, a simple ordinal priority ranking has been specified for the leaf nodes. The use of a simple average of the priority figures at the leaf nodes will yield an identical aggregate figure of 'priority 3' both for the left and the right subtree. This is barely a satisfactory decision basis for the planner at the root node. As will be shown later (see Chapter 4), the use of (demand-weighted) average values of the leaf-node level priority figures will lead to inferior allocation decisions at the higher hierarchy levels. Put differently, it is not obvious which aggregate priority measure should be used at the intermediate level. In general, the more hierarchy levels exist, the more difficult these allocation decisions are, especially if they have to be made using decentral schemes and only with partial or even with inaccurate information.

These first few remarks already indicate that decentral quota allocation and demand fulfillment decisions constitute a challenging, though not yet addressed problem for many large firms with a broad and dispersed customer portfolio. It is the objective of this thesis to provide an overview of the problem area and to present a number of initial research results to help making better demand fulfillment decisions in such multi-stage customer hierarchies. More concretely, the key problem addressed in this thesis can be refined as summarized below.

Situation. A company manufactures its products to stock at a central facility based on forecast demand. It uses a hierarchically structured, multi-level salesforce to serve its customers in many different markets. However, not all markets are equally attractive; in particular, individual customer segments differ in terms of their profitability. Due to this dispersed customer segmentation, protected product quotas need to be determined to ensure high service levels for the most important customer segments during shortage situations. This quota reservation process has to be done in a decentral manner, due to decentral information in the hierarchical sales organization.

[5] This is indicated by the question marks in Figure 1.1b.

Definition 1. *The demand fulfillment problem in multi-stage customer hierarchies (for short: DMC problem) is to maximize overall firm profits by serving the most profitable customer segments with priority.*

The DMC problem consists of two aspects: Finding a good allocation of product quotas based on forecast demand (**allocation planning**) and determining the best fulfillment approach for arriving orders in a real-time process (**consumption planning**).[6]

The major difference to existing research is that both decisions need to be made by local decision makers in the hierarchy based on aggregate and decentral information. Compared to most practical settings where multiple products need to be considered, the discussion in this thesis focuses on a simplified version of the DMC problem by considering only a single product.

To characterize the DMC problem more closely, to narrow down the key research questions and to specify the approach taken in this thesis, the remaining sections of this introduction have been organized as follows: Section 1.1 contains a summary of a case study from the oil industry which illustrates the DMC problem in a practical setting. Then, in Section 1.2, a detailed characterization of the key aspects of the DMC problem will be given. The resulting research questions as well as the aspired contributions of this thesis will be stated in Section 1.3. Finally, Section 1.4 will give an overview of the approach chosen in this thesis to address the DMC problem.

1.1. Motivating Case Study

This DMC problem is likely to occur frequently in practice, but has only rarely been discussed in the academic literature. Besides the dissertation by Fischer (2001), the problem has primarily been reported in a case study by Roitsch and Meyr (2008). This case study will be summarized in the following as a motivating example for the DMC problem.

Roitsch and Meyr (2008) give a detailed illustration of the hierarchical planning procedure in the oil and refining industry. This industry is instructive for at least three reasons:

1. Many end products have **applications in different markets**, e.g. diesel oil may propel combustion engines in the automotive and industrials industry, but can also be used as a fuel in domestic and industrial heating applications.

2. Sales prices—and therefore also **margins—depend on** a number of **external factors**, including the current oil price, exchange rates and the level of local taxes.

3. For many refinery and base chemical products, **several types of customers** exist: Firms often commit to long-term delivery contracts which typically stipulate

[6] A more comprehensive characterization of these two planning problems will be given in Section 2.4.

contractual penalties on service level infringements. Selling to these contractual customers guarantees a certain minimum utilization of the company's production assets, but may at times be less profitable than serving casual customers when prices on the spot market are higher.

The company covered in the case study by Roitsch and Meyr (2008) is active in a number of national and regional markets and serves a very heterogeneous customer base through several business units. It employs a local salesforce to handle the individual sub-markets and relies on higher-level sales managers to administrate larger geographic areas, sales channels and entire business units. The local salesforce agents possess an intimate knowledge of their market segments. They are thus ideally positioned to prepare the necessary forecasts of the demand and of the price level in their respective markets. This input data is then aggregated within the sales hierarchy by summing demand volumes and by calculating weighted average prices to obtain figures at higher planning levels. The aggregated figures are primarily used to plan production and distribution operations. An example for this aggregation across customer segments and regional markets is shown in Figure 1.2 for the business unit 'Commercial'.

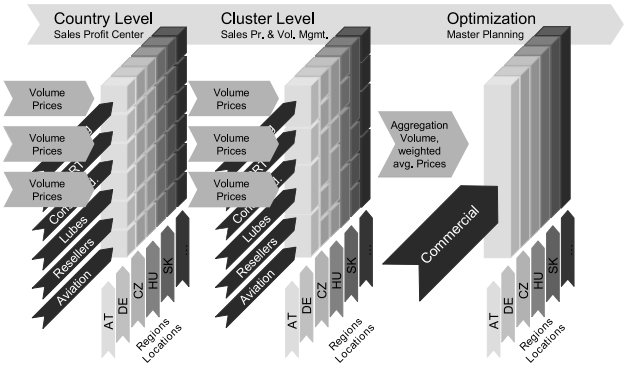


Figure 1.2. – Hierarchical aggregation in the oil industry case study (Roitsch and Meyr, 2008, Fig. 21.5, p. 411)

As reported by Roitsch and Meyr (2008), the manufacturing capacity in the refineries constitutes a major bottleneck in the planning process. Furthermore, production is subject to a significant lead time, predominantly due to crude oil procurement lead times (6–8 weeks). These constraints are handled in a *master planning* step which coordinates procurement, manufacturing and distribution decisions across the production and sales network.[7] Master planning across the refinery network is repeated on a quarterly basis, with monthly updates.

The resulting output figures from this master planning step are communicated within the entire organization and all stakeholders typically work towards achieving the estab-

[7] A more comprehensive definition of master planning will be given in Section 2.3.

lished monthly targets. However, during the production lead time, market conditions often change, but also forecasts typically become more accurate. Therefore, the actual demands per market segment will differ from the original forecasts which have been used in the planning phase. Given that production volumes are fixed due to the long procurement lead times and due to the high capacity utilization of the refinery network, the available product quantities often need to be allocated. As part of this process, a further profit optimization is usually possible during each month, i.e. between two master planning runs. More precisely, the disaggregation of aggregate production quantities to individual (end) products can still be adjusted, typically along three dimensions:

- **Within a product family**, i.e. between product variants,
- **between business units** which sell essentially the same product for different applications, and lastly,
- **between different customer segments** within a particular market, e.g. between contract and spot-market customers.

To exploit these intra-month optimization potentials, a simple approach is to increase the frequency of master planning. As some decisions can no longer be altered within a month, the corresponding decision variables in the master planning model need to be kept fixed. Unfortunately, the size, complexity and computational requirements of the employed master planning models generally do not allow for an interval length between two planning iterations which is less than one month.

An alternative is therefore the introduction of a dedicated *allocation planning* step which supports the limited number of still adjustable decisions. Allocation planning can be run more frequently than master planning.

Roitsch and Meyr (2008) described a rule-based allocation planning system to disaggregate the master planning results per business unit in a step-wise manner along the different sales channels, geographical locations and along the time dimension. This allocation planning system observes minimum supply requirements and allows for margin management by giving preference to customer segments with particularly high sales prices. This system-supported disaggregation (illustrated in Figure 1.3) then serves as the basis for further manual consultation-based adaptations. The individual sales managers in the customer hierarchy fine-tune the final quota allocations via a series of mutual negotiations.

Once actual orders arrive, individual customer requests can be confirmed against the resulting quotas. In contrast to a first-come-first-served approach where all customers compete for a common stock of scarce product quantities, this allocation planning step ensures that all customer segments receive a predefined level of customer service which depends on the priority of the segment.

However, as has become obvious in the introductory remarks above and as will be further refined in the course of this thesis, this allocation planning step is not without its difficulties, especially due to the rule-based decentral decisions in the customer hierarchy. Yet, the allocation planning approach described in this case study is already of an advanced nature, compared to common practice which prevails in most other companies.

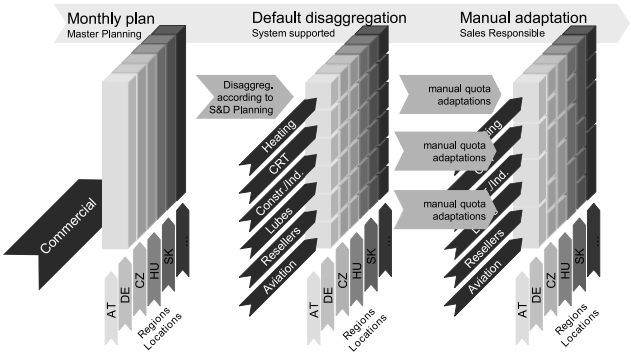


Figure 1.3. – Hierarchical disaggregation in the oil industry case study (Roitsch and Meyr, 2008, Fig. 21.6, p. 413)

The following section will give a more detailed overview of the DMC problem and will highlight its general characteristics.

1.2. The Demand Fulfillment Problem in Multi-Stage Customer Hierarchies

The motivating example described in the previous section exhibits three important aspects which are typical of the DMC problem in general settings. In the following, each of these aspects will be characterized in more detail.

1. Customers are **heterogeneous** and differ in terms of their importance to the firm.
2. A **hierarchical** organization with distributed decision-making is used to serve the heterogeneous customer segments. For example, demand forecasts per customer segment are made on a decentral basis. Overall requirements are determined via aggregation within the customer hierarchy.
3. As production is subject to a significant lead time, shortage situations may occur which require a rationing process. Product quotas per customer segment need to be determined by an **iterative disaggregation** process in the customer hierarchy.

1.2.1. Heterogeneous Customer Segments

Not all customers of a firm carry the same importance and not all orders are equally attractive. This heterogeneity is often rooted in one or several of the following—not necessarily mutually exclusive—dimensions.

Profit margin There is usually a large variation in the difference between revenues and direct costs to serve a particular customer, often driven by the costs which may

still be influenced at the time when an order is accepted. Different margins arise naturally for individualized products which are manufactured or assembled based on customer specifications. But even in the case of standardized products with uniform sales prices, margin differences often exist. For example, these can arise from additional transportation costs, from local taxes or exchange-rate fluctuations which are absorbed by the manufacturer.

Strategic importance In some instances, the immediate financial contribution of a particular customer may appear to be low, but preferential treatment is still advisable. In business markets, this is often true for long-term, partnership-based relationships (Ball et al., 2004) and for developing customers where promising future profit potential is likely though not yet readily quantifiable. When selling to consumers, opinion leaders who influence purchase decisions of potentially many other consumers—often more profitable ones—can also be of particular strategic importance. Such customers may warrant preferential customer service and have to be rated above their face value contribution (Mulhern, 1999).

Type of contractual obligations As indicated in the oil industry case study in the previous section, some customers who require reliable supplies often commit to long-term delivery contracts with predetermined service levels. Companies often aim at maintaining a certain share of long-term, contractually-bound customers, but might as well serve spot market customers in an opportunistic manner (Kleijn and Dekker, 1999).

Criticality of use Customers often require a certain product with different degrees of urgency. For example, if the delivered product quantities are solely used to raise the inventory position, this type of usage is less urgent than if the product will instead immediately be used in a production process. If price discrimination is not enforceable for customers with urgent needs, at least customer retention may be increased by serving the orders of those customers with highly critical usages first. Dekker et al. (1998) presented a model from spare parts management where one item can have different uses, a critical and a non-critical one. More concretely, Möllering and Thonemann (2008) described the case of a telecommunications component which can either be used in an antenna or in a network computer. Breakdown of the component in the network computer simultaneously takes 30 antennas off-line and is thus significantly more severe. Ha (1997) presented the case of a common component which is used by the customer in the production of several end products with different values.

Recency of the last stockout To prevent annoying and ultimately losing a customer due to frequent stockouts, the recency of the last stockout should often be explicitly considered in the order acceptance decision. Examples where this aspect has been considered can be found in Weisenborn and McCright (1999) or in Fischer (2001).

In theory, it may be more appropriate to jointly consider several of these dimensions to determine the overall importance of a customer or of a particular order. For example, Korpela et al. (2002) combined the dimensions profit margin, strategic importance and the historic sales volume into an index of overall customer importance. Niinistö (2010) determined the overall importance of backordered customer requests by combining information on the current customer waiting time, a measure of the order urgency and a criterion for customer importance. In both examples, the Analytical Hierarchy Process (AHP) (see Saaty, 1980) was used to construct a single-dimensional priority index. It will be argued later that most of the above criteria either directly or indirectly imply a profitability-based segmentation. A one-dimensional criterion like profitability allows making unambiguous comparisons between individual customers.

In practice, however, firms rarely manage at the level of individual customers—except for particularly important, usually large customers known as key accounts. Rather, companies focus on larger *customer segments* which consist of a number of individual customers which are perceived to be relatively similar with respect to one of the above criteria.[8] This perspective will also be employed in the remainder of this thesis.

1.2.2. Multi-Stage Customer Hierarchies

It is usually impracticable to cluster the entire customer base of a company only by using one criterion such as customer profitability. Many firms rely on additional, rather straightforward criteria to define more aggregate, higher-level customer segments. Such aggregate customer segments are often defined using one or both of the following two criteria:

- Geography: Individual customer segments are additionally grouped by continent, country, sales region or ZIP code (Kilger and Meyr, 2008).
- Sales channel: Higher-level customer segments are characterized by the path through which orders are received, e.g. wholesale, retail or direct marketplaces such as e-commerce (Frazier, 1999).

These additional aggregate segments lead to a hierarchical structure. Consider the criterion 'geography'. Such a hierarchy often arises naturally because many firms rely on geography-based multi-echelon inventory systems. As stressed by Graves (1996), such systems are very cost-effective for firms with a geographically dispersed demand, with economies of scale in production and with market-driven service requirements. In a resulting geography-based customer hierarchy, larger regions (e.g. continents, countries) at the higher levels are further subdivided into a number of smaller sales districts and territories at lower levels (see Kilger and Meyr, 2008). The customers in each of the smaller

[8] This process is known as *clustering*. Its objective is to find a partitioning of objects—here customers—in a manner that objects within one cluster are as homogeneous as possible while different clusters are as heterogeneous as possible (Gordon, 1987, p. 119). Meyr (2008) demonstrated the application of basic clustering methods to form customer segments for allocation planning. More comprehensive clustering methods were discussed in Fraley and Raftery (1998) and Kaufman and Rousseeuw (2005).

sales territories can again be clustered according to the heterogeneity criteria discussed in the prior paragraph. Customers in these lowest segments are homogeneous with respect to profitability (e.g. have the same transportation costs, same taxes or the same currency). In the more aggregate, higher-level customer segments, the level of customer heterogeneity is usually significant.

The management of these hierarchical structures usually relies on a decentralized organization. Multiple planners at different hierarchical levels can exploit their close proximity to the customer with a bottom-up forecasting approach (see Cox, 1989). Such a *sales-force composite* approach is firmly rooted in many industrial markets and rather difficult to change (Weinstein, 1987, p. 452–453). A major reason is that a single planner usually cannot handle the complexity of the tasks involved and cannot manage the significant information requirements. Furthermore, decentral, participative structures are also advantageous from an organizational perspective. As frequently stressed in the budgeting literature, these have the effect of increasing employee motivation (e.g. see Kanodia, 1993).

For planning at the higher levels, the detailed information gathered by the salesforce at the lower hierarchy levels needs to be aggregated to ensure manageability of data at the higher levels. While demand forecasts can easily be summed, prices and margins are usually aggregated by calculating demand-weighted averages (e.g. see Vollmann et al., 2005, p. 41). However, since local planners usually have an information advantage compared to their superior managers, this hierarchical forecast aggregation entails the risk that the former may misrepresent their forecast reports and thus game the system. They may exaggerate their reports in an attempt to obtain better allocations if their compensation is linked to actual sales.[9]

1.2.3. Iterative Disaggregation and Allocation in Customer Hierarchies

The resulting aggregated forecasts then drive production and manufacturing processes. The production plan is usually frozen during a significant lead time before the products become available for distribution and sale. The minimum length of this frozen period is determined by the sum of the cumulative procurement, manufacturing and potentially distribution lead times (see e.g. Chung and Krajewski, 1984). During this frozen period, actual customer demand may change and corresponding forecasts usually improve in accuracy, resulting in updated aggregate forecasts. Shortage situations result if the (updated) aggregate demand turns out to be larger than the actual supply replenishments.

Not only demand-, but also supply-side factors may be responsible for shortages. For example, shortages can result if unanticipated disturbances in the forecast-driven processes of the supply side occur. Today's lean supply chains with limited (raw) material inventories are particularly vulnerable to disruptions in the procurement processes, e.g. due to natural disasters. Further disturbances may originate from the production and as-

[9] Such manipulations due to asymmetric information endowments are at the heart of *principal-agent problems*. A more thorough discussion of this area in the context of the DMC problem will be provided in Section 3.4.

sembly processes if capacities turn out to be less than anticipated (e.g. due to unplanned stoppages, repairs, strikes) or if actual production outputs are less than planned, e.g. due to yield losses or quality rejects.

A consequence of shortages is that not all customer segments can be served as originally planned. A simple solution consists of serving the customer requests based on a first-come-first-served rule until running out of supply. Obviously, this approach neither caters to the hierarchical nature of the customer organization nor can it exploit any differences in profitability between individual orders (and customers). Hence, most firms prefer to employ a rationing scheme to disaggregate the available supplies within the customer hierarchy and to prevent an intense competition for the scarce product quantities. In practice, such a disaggregation is performed in an *iterative manner*. Product quotas are broken down one hierarchical level at a time rather than having a central planner determine allotments per customer segment at the bottom of the customer hierarchy. In practice, the actual disaggregation process often involves several rounds of negotiations at multiple levels in the sales organization.

Such a decentral and iterative approach has many advantages:

- It is an effective way to avoid having to gather large amounts of data, possibly for many different products, at a central entity.
- In addition, decentralization of the allocation process is advantageous in the presence of constraint processing capacities and has been shown to be generally faster (see e.g. Radner, 1992; van Zandt, 2003).
- Furthermore, a decentral approach respects the organizational structure and existing decision paths in the company.
- Moreover, relying on a decentral organization for the quota allocation process may reduce the tendency of decentral sales agents to bias their forecasts. Compared to a distant, central planner, direct supervisors usually have better means to monitor and verify the validity of the reports of their direct subordinates.

Some decision support for the DMC problem is provided by modern Advanced Planning Systems (APS). Usually, a number of standard rules are provided which help determining the break-down in a planning hierarchy in an APS. For an illustration, consider Figure 1.4, which has been taken from Kilger and Meyr (2008). A simple geography-based customer hierarchy is represented. Its lowest hierarchy levels contain a sufficient level of customer heterogeneity, e.g. in terms of profitability due to regionally different transportation costs. On the left side, demand forecasts are aggregated in a bottom-up manner. Due to capacity constraints in the production system, not all requests can be fulfilled and a product allocation is required. The iterative quota disaggregation is depicted on the right hand side. The resulting allotments per leaf node may be consumed to serve the incoming customer orders per leaf-node customer segment.

As illustrated, most APS use rather simple allocation rules for the iterative allocation to the lower-level customer segments. These simple rules, e.g. based on fixed priorities

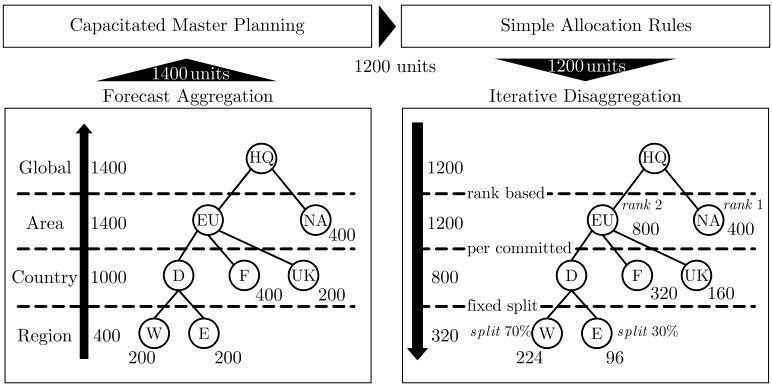


Figure 1.4. – Forecast aggregation and quota disaggregation in APS (cf. Kilger and Meyr, 2008, pp. 191–192)

(*rank-based*), based on previous forecasts (*per-committed*), or based on fixed splitting percentages (*fixed split*), are a means to restrict competition for the scarce product quantities only to the customers within each customer segment. Unfortunately, these allocation rules do not explicitly observe the level of heterogeneity which exists within the customer hierarchy in terms of different levels of profitability. It is therefore questionable whether the use of these rules in customer hierarchies is an appropriate means to maximize overall profits for a company. Furthermore, it is an open question which of the illustrated rules should actually be used under which circumstances.

Some of these aspects will be addressed in this thesis. The following section gives a more concise summary of the key problem areas which will be investigated. It will also state the key research questions addressed in this thesis.

1.3. Problem Focus and Research Questions

Based on the case study (Section 1.1) and the more detailed characterization of the DMC problem setting (Section 1.2), five problem areas can be distinguished which have a major influence on the quality of the demand fulfillment decisions in a multi-stage customer hierarchy. These problem areas can be identified by following the temporal sequence of the steps in the DMC problem.

1. **Demand forecast errors:** As future demand is unknown, forecasts regarding the demand level (quantities per time period) and regarding future prices (or margins) per customer segment are required. While forecast accuracy largely depends on the forecasting technique employed and on the predictability (i.e. variability) of the underlying time series, other factors may also have an impact. In a customer hi-

erarchy, *hierarchical forecasting* according to the so-called pyramid principle (Muir, 1979) is often used, and the choice of the level at which a forecast is being made has a decisive impact on the resulting errors.[10]

2. **Forecast misrepresentation and aggregation bias:** Generally, aggregation is associated with a loss of detailed information. In decentral settings, additional errors and distortions may be introduced due to the large number of individuals involved in forecasting and in forecast aggregation. Because planners in the lower parts of the customer hierarchy typically enjoy an information advantage over their superiors, incentives may exist for the former to hide or bias their specific knowledge in a pursuit of selfish interests. Such incentives are typically induced by flawed compensation schemes.

   Note that the forecast misrepresentation problem is not specific to customer hierarchies with distributed decision-making. It is also present in centrally administered situations with a flat partitioning of the customer segments. The central planner usually has to rely on reported forecasts, e.g. from the sales managers who are responsible for each customer segment.

3. **Supply forecast errors**, i.e. the **level of shortage:** Due to constraints in the production system, satisfying all expected customer demands may not always be feasible. Furthermore, no planning and no production system is perfect, and actual supply arrival schedules will often deviate from previous plans. Hence, in practice, some uncertainty remains both regarding the timing and the magnitude of replenishment supplies.

4. **Disaggregation and allocation errors:** In an iterative disaggregation process with distributed decisions, decentral planners will only rarely be in possession of all necessary information to make accurate quota assignments. This is one of the major reasons why simple rule-based approaches prevail in practice. In the presence of decentral information, additional errors may be introduced by agency behavior as planners may intentionally bias allocation decisions to pursue selfish objectives.

5. **Quota consumption rules:** Once quotas per customer segments have been determined, these may be consumed only by the incoming orders of the same segment. This base case is referred to as a *dedicated consumption*. However, many companies also employ more enhanced consumption rules to soften these strict reservations per customer segment. For example, if the reservation for an important customer segment has been depleted, additional orders of that segment may often be confirmed by allowing the consumption of the available quotas of other, less important customer segments. Naturally, there is a close interdependence between the allocation planning and the consumption planning problem.

The superposition of these five problems is ultimately responsible for any differences between a demand fulfillment outcome determined by an omniscient, central planner

[10] This aspect is covered in more detail in Section 2.2.5.

with perfect information and by a result obtained in practical settings with distributed, imperfect decisions.

The following paragraph will give an overview of several assumptions to further characterize and limit the exact problem setting studied in this thesis.

Problem Focus: This thesis focuses on the demand fulfillment problem in heterogeneous, multi-stage customer hierarchies. Customer heterogeneity is primarily considered in terms of different levels of profitability. Overall, it is assumed that both the individual customer segments as well as their hierarchical structure are given. They are not subject to adjustments in the short run. Customer segmentation can therefore be thought of being determined as part of more long-range planning processes. In the following, it will be convenient to assume a geography-based customer hierarchy as an intuitive example. Such a type of customer hierarchy is likely to derive from an existing (multi-echelon) inventory system which is employed by the company to serve its customers, and the design of inventory systems is a typical output of rather long-term planning processes.[11]

To simplify both the analysis and the simulation experiments, a make-to-stock (MTS) production environment will be assumed in which a single good is made in a single facility. This factory has an associated central inventory from which all orders are served. Note that allocation problems in customer hierarchies also exist with respect to demand fulfillment in make-to-order and assemble-to-order production environments. In the former case, primarily production capacities, in the latter case also raw materials and subassemblies need to be allocated to the different, hierarchically structured customer segments. Obviously, the different types of resources which are jointly required to fulfill a particular customer order render the overall fulfillment problem more complex. As will become clear, not even for MTS—the conceptually simplest setting for demand fulfillment—have any approaches been reported yet which tackle the problem of multi-level customer hierarchies. This rationalizes the limitation of this thesis to MTS situations. However, the ideas developed in the following can be seen as a starting point for extensions to other production environments.

In a similar manner, the restriction to a single product is justified as this constitutes the simplest case. Of course, real-world situations are more complex and are characterized by a number of additional effects,[12] but those will be suppressed in this first analysis of the DMC problem.

In the course of this thesis, a simplified version of the supply forecast errors will be studied. In line with the existing demand fulfillment literature, it will be assumed that the amount of available supply quantities is determined exogenously. Therefore, primarily

[11] This will be discussed further in Section 2.1.

[12] For example, many customers typically order a number of products simultaneously. Often, several of these items are complementary goods. The customer has no or only a significantly reduced utility if just one item or parts of the original order can be fulfilled. In this situation, the best allocation strategy if at least one item is in short supply remains an open question. Nevertheless, the allocation of complementary products may be addressed in follow-up research.

different levels of shortage will be studied. Uncertainties regarding the availability of the supply quantities will be disregarded.

Key Research Questions: The research described in this thesis applies to companies which pursue profit optimization as their overall corporate objective. Once supplies are short, the firm needs to decide which customers in the customer hierarchy to serve with priority. The focus lies on situations where this rationing process consists of a series of decentral allocation decisions from the top to the leaf nodes of the customer hierarchies.

Question 1: To what extent is it worthwhile to exploit profitability differences among customer segments in the quota reservation process in a customer hierarchy?

Allocation planning approaches which consider differences in customer profitability have long been used if the partitioning of the customer segments is flat.[13] A central planner can make allocation decisions which will lead to higher overall profits by reserving scarce quantities for more profitable segments. In the context of multi-stage customer hierarchies, primarily simple quantity-based rules are used (see Figure 1.4). Examples for profit-based approaches are rare and can primarily be found in the master planning step; for an example, see the oil industry case study in Section 1.1. A first objective is thus to establish to what extent overall profits can be improved by explicitly considering profitability differences in a customer hierarchy. Furthermore, to improve the design of planning systems which involve multi-stage customer hierarchies, it is helpful to establish whether profit-based allocation approaches are equally important at all stages of the customer hierarchy. Put differently, do particular allocation steps benefit more from a profit-oriented approach than others?

Question 2: How significant is the disadvantage of a decentralized profit-based allocation planning procedure over a centralized, full information approach?

Profit-based allocation planning for a flat partitioning of the customer segments coincides with the use of a central, fully informed planner. Should such an omniscient central planner also be available in a multi-stage customer hierarchy, the multi-level nature of the hierarchy may be disregarded and similar results as with a flat partitioning of the customer segments can be expected. However, this rarely reflects the reality in many companies where many restrictions need to be obeyed in the presence of multi-stage customer hierarchies. If not all information is available at a central entity and if decisions need to be made locally, does a profit-based allocation procedure continue to be an attractive choice?

[13] An overview will be given in Section 2.4.

Question 3: To which extent do the different environmental conditions affect the outcome of a profit-oriented, decentral allocation planning-based demand fulfillment approach in a customer hierarchy?

In addition to the impact of a decentralized scheme, several other factors are likely to affect the quality of profit-based allocation planning in a customer hierarchy. It is another goal of this thesis to establish guidelines which will mitigate any negative impacts on total profits from the above five problem areas (see pages 13–14). If mitigation is not possible, it will at least be assessed to what extent each problem area affects the solution quality of the DMC problem.

Regarding forecast misrepresentations and the resulting disaggregation errors, the primary focus will be on an investigation of possible ways to prevent such issues from occurring in the first place in multi-stage customer hierarchies. With respect to forecast errors, to the shortage level and to consumption rules, the objective will be to determine their quantitative impact on total profits.

Ultimately, the quantification of these effects will allow addressing a key question relevant for many practitioners: Under which worst-case conditions does the resulting performance of profit-based, decentral allocation planning schemes with subsequent consumption planning still constitute an improvement over trivial demand fulfillment policies such as first-come-first-served?

1.4. Overview of Approach

The presentation in the subsequent five chapters aims at giving answers to these three research questions. For a comprehensive discussion, the second and third chapter will be used to provide a background perspective on the planning problem and to characterize multi-stage customer hierarchies. In the fourth and fifth chapter of this thesis, the analytical and experimental results which have been obtained for this problem setting will be summarized and discussed. Ultimately, the three key research questions will be answered in the final chapter. The following paragraphs give a more detailed outline of the treatise.

Chapter 2 delineates the background of this thesis and of the DMC problem in the supply chain and operations management literature. Initially, the fundamental supply chain planning problems, the corresponding main planning tasks as well as their interrelations will be described. Afterwards, the focus shifts to those planning tasks which have a direct impact on the demand fulfillment decision. In particular, the key planning tasks of demand planning, master planning and demand fulfillment will be introduced by stating their objectives, by explaining major sub-tasks and by summarizing basic planning models.

- For **demand planning**, the major forecasting structures, processes and controlling instruments will be presented, including an overview of **hierarchical forecasting**.

Demand planning is essential to the demand fulfillment problem in hierarchies as its output is required both to plan actual production and to make the allocation decisions.

- **Master planning** is relevant in demand fulfillment for two reasons: First, its output determines the amount of supply quantities which will be available to ultimately serve customer orders. Second, master planning typically also entails a first allocation decision—allotting the produced quantities to particular business units, as in the oil industry example, or to certain geographical regions.

- Lastly, the **demand fulfillment** planning task ultimately takes care of arriving customer orders. An overview of different types of fulfillment approaches and system types will be given and the comprehensive literature discussing demand fulfillment for heterogeneous customers in an MTS environment will be reviewed. As it turns out, the problem of customer segments organized in the form of multi-stage hierarchies has not yet been addressed in the demand fulfillment literature.

Given the lack of comprehensive research in this area, **Chapter 3** will provide an introduction to **multi-stage customer hierarchies**. It will start with an abstract perspective on hierarchies. After establishing a formal terminology, a classification of basic hierarchy types will follow. Two essential mathematical operators will be introduced which are typical of hierarchies—**aggregation** and **disaggregation**. Then, the focus will expand to also include hierarchies in an organizational context with decentral decisions, as encountered in the DMC problem. Employing the well-known framework by Schneeweiß (2003), a categorization of hierarchical relationships and of decentral decisions in organizational hierarchies will be given. Moreover, after discussing not only the challenges but also the benefits of hierarchical organizations in comparison to centrally administered organizations, it will become clear that the former organizational structures with distributed decision-making are necessary in many demand fulfillment settings.

After extending the definition and notation of formal hierarchies to customer hierarchies, the Schneeweiß framework will be employed to describe the key characteristics of customer hierarchies. An important aspect in this context relates to the information asymmetries between local sales agents and their superior sales managers. The principal-agent setting allows studying the **forecast misrepresentation** problem which may arise in this situation. A number of compensation and incentive schemes from the existing literature will be analyzed to assess whether they can mitigate the incentives to report biased forecasts in customer hierarchies.

The closing section of this chapter will take on customer heterogeneity in the form of customer segments which differ in terms of profitability. In a first step, an analogy between the measurement of income inequality in econometrics and the **measurement of customer heterogeneity** in multi-stage customer hierarchies will be established. In a second step, it will be shown that one measure, Theil's index T, is particularly well-suited to capture profitability-induced customer heterogeneity in multi-stage hierarchies.

The following two chapters aim at finding solutions to the DMC problem by considering a single-period problem, consisting of an allocation planning step and a subsequent order consumption phase. The presentation in **Chapter 4** will focus on the **allocation planning** step and will assume that a basic dedicated consumption policy is used. This means that orders from a particular customer segment will be served solely from the corresponding quota. In particular, it is not permitted to 'steal' from quotas which have been reserved for other segments. After presenting a structured overview of **quantity-based allocation rules** which do not exploit customer heterogeneity, several **profit-based allocation schemes** will be formalized. These schemes correspond to different degrees of central control and to different levels of data transparency in the customer hierarchy. It will be shown that an intuitive decentral profit-based allocation rule leads to quotas which may deviate significantly from optimal allocations under central control with full information.

To improve the quality of such decentral allocation decisions, a **novel profit-based allocation scheme** will be introduced. It is well-suited for distributed decision-making, and its superiority over the intuitive scheme stems from respecting the level of customer heterogeneity in any particular subtree of the customer hierarchy. A number of numerical experiments will confirm the superiority of the scheme over existing quantity- and profit-based rules. Moreover, it will be shown that the deviation from an optimal allocation scheme under central control and full information transparency is small.

Extensions to these initial findings will be provided in **Chapter 5**. The discussion of the DMC problem will be expanded to also include forecast errors and enhanced consumption planning policies.

In a first step, it will be established that the new scheme will continue to lead to superior quota allocations if **forecast errors** are introduced. An important aspect of forecast errors is that even allocations determined by a central, omniscient planner can turn out to be wrong if planning is subject to a lead time and if demand changes in the meantime.

In the presence of forecast errors, it is important to consider how the quotas are consumed. Therefore, further simulations will be presented to illustrate the interdependency between allocation and **consumption planning** decisions. Extending existing consumption rules from models for a flat partitioning of the customer segments and adapting them to the situation in multi-stage customer hierarchies, it will be shown how nested quotas can lead to better order acceptance decisions compared to a dedicated consumption. An important finding is that these enhanced consumption rules constitute an additional improvement lever to obtain better solutions to the overall DMC problem.

Moreover, the modeling setup with individual orders used in Chapter 5 allows analyzing specific worst-case conditions: When is an allocation planning-based demand fulfillment approach advantageous compared to a simple **first-come-first-served** order acceptance policy? For a given hierarchy size, the simulations will indicate lower threshold values of the three key environmental parameters shortage rate, forecast accuracy and level of customer heterogeneity. Once these thresholds are met, an application of the new

decentral profit-based allocation planning scheme is clearly worthwhile in a multi-stage customer hierarchy.

Finally, also the effect of retaining some supply quantities at higher hierarchical levels will be studied. This strategy constitutes an alternative to nested quotas. Such **virtual safety stocks** may provide a hedge against too high quota reservations for particular customer segments in the presence of forecast errors (overprotection). They may be consumed by several customer segments on a first-come-first-served (FCFS) basis if the original quotas for the segment have been depleted.

Chapter 6 concludes this thesis. First, a summary of the main discussion points of this thesis will be given. Then, the results of the numerical experiments will be summarized to present answers to the three key research questions in the form of managerial implications from this research. These answers establish guidelines how demand fulfillment decisions in multi-stage customer hierarchies can be improved with the help of the new profit-oriented, decentralized allocation planning approach. Lastly, the presentation closes with an outlook on directions for further research opportunities.