9 Demand Fulfilment and ATP

Christoph Kilger and Herbert Meyr

The planning process that determines how the actual customer demand is fulfilled is called *demand fulfillment*. The demand fulfillment process calculates the first promise date for customer orders and—thus—strongly influences the order lead-time and the on time delivery.[1] In today's competitive markets it is important to generate fast and reliable order promises in order to retain customers and increase market share. This holds particularly true in an e-business environment: Orders are entered on-line in the e-business front end, and the customer expects to receive a reliable due date within a short time period.

Further, e-business solutions have to support on-line inquiries where the customer requests a reliable due date without committing the order.

The fast generation of reliable order promises gets more complex as

- The number of products increases
- Products are configured during the ordering process
- The average product life cycles get shorter
- The number of customers increases
- Flexible pricing policies are being introduced
- Demand variations increase and get less predictable.

[1] In the following, we use the terms *order promising* and *order quoting* synonymously, as well as the terms *promise* and *quote*.

C. Kilger
Ernst & Young GmbH, Wirtschaftsprüfungsgesellschaft, Heinrich-Böcking-Straße 6–8,
66121 Saarbrücken, Germany
e-mail: christoph.kilger@de.ey.com

H. Meyr (✉)
Department of Supply Chain Management (580C), University of Hohenheim, 70593 Stuttgart, Germany
e-mail: H.Meyr@uni-hohenheim.de

H. Stadtler, C. Kilger and H. Meyr (eds.), *Supply Chain Management and Advanced Planning*, Springer Texts in Business and Economics,
DOI 10.1007/978-3-642-55309-7_9, © Springer-Verlag Berlin Heidelberg 2015

The traditional approach of order promising is to search for inventory and to quote orders against it; if there is no inventory available, orders are quoted against the production lead-time. This procedure may result in non-feasible quotes, because a quote against the supply lead-time may violate other constraints, e.g. available capacity or material supply.

Modern demand fulfillment solutions based on the planning capabilities of APS employ more sophisticated order promising procedures, in order

1. to improve the on time delivery by generating reliable quotes,
2. to reduce the number of missed business opportunities by searching more effectively for a feasible quote and
3. to increase revenue and profitability by increasing the average sales price.

In the following section, the principles of APS-based demand fulfillment solutions are described and the basic notion of ATP (available-to-promise) is introduced. Sections 9.2 and 9.3 show how ATP can be structured with respect to the product and the time dimension, whereas Sect. 9.4 introduces the customer dimension and the concept of allocation planning, resulting in allocated ATP (AATP). Finally, Sect. 9.5 illustrates the AATP-based order promising process by means of examples.

An early reference describing the concept of ATP and the improvement of the customer service level by ATP based on the master production schedule is Schwendinger (1979). In Ball et al. (2004) and Pibernik (2005) comprehensive overviews of ATP related work are given. Fleischmann and Meyr (2003) investigate the theoretical foundations of demand fulfillment and ATP, classify the planning tasks related to ATP, and discuss the generation of ATP and order promising strategies based on linear and mixed integer programming models. The practical application of ATP concepts in concrete APS is, for example, described in Chap. 26 for OM Partners, in Dickersbach (2009) and Fleischmann and Geier (2012) for SAP/APO and in i2 Technologies Inc (2000) for i2 Technologies that has been acquired by JDA in 2010.

9.1 Available-to-Promise (ATP)

The main target of the demand fulfillment process is to generate fast *and* reliable order promises to the customer and shield production and purchasing against infeasibility. The quality of the order promises is measured by the *on time delivery* KPI as introduced in Chap. 2. Using the traditional approach—quoting orders against inventory and supply lead-time—often will result in order promises that are not feasible, decreasing the on time delivery.

Figure 9.1 illustrates this by means of a simple example. Consider a material constrained industry like the high-tech industry, and let us assume that a specific component has a standard lead-time of 2 weeks. There are receipts from the suppliers scheduled for the next 2 weeks and—as we assume a material constrained industry—no additional supply will be available for the next 2 weeks. The volume of new customer orders that need to be promised for week 1 and week 2 is exceeding the volume of the scheduled receipts. Figure 9.1a illustrates this situation. Standard

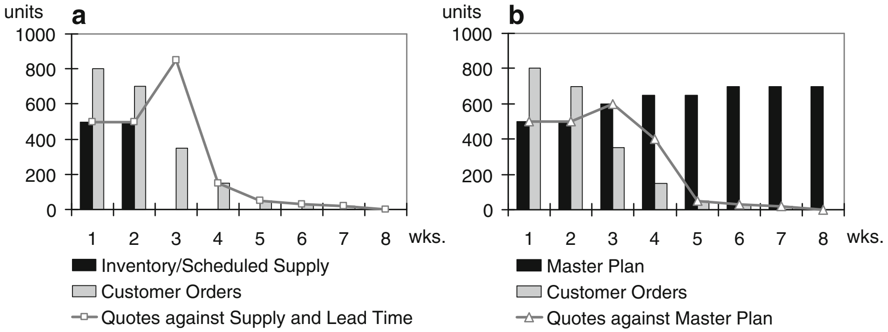

Fig. 9.1 Demand wave beyond the standard lead-time. (**a**) Order quoting against lead-time. (**b**) Order quoting against master plan

MRP logic is to schedule all new orders against the scheduled receipts and—if not all orders can be satisfied by that—against the standard lead-time (in our example 2 weeks). In other words, MRP assumes infinite supply beyond the standard lead-time and creates supply recommendations based on the order backlog. The gray line in Fig. 9.1a shows the quotes created by the MRP logic. In week 3 (i.e. after the standard lead-time) all orders are scheduled that cannot be quoted against the scheduled receipts. It is quite clear that the fulfillment of this "demand wave" will not be feasible, as the available supply will most probably not increase by 100 % from 1 week to the next week.

The master planning process (see Chap. 8) has the task to create a plan for the complete supply chain, including production and purchasing decisions. Thus, master planning generates a plan for future supply from internal and external sources (factories, suppliers) even beyond the already existing scheduled receipts.[2] The idea of APS-based demand fulfillment is to use the supply information of the master plan to create reliable order quotes. Figure 9.1b shows the master plan and the orders quoted against the master plan. For week 3 the master plan reflects the constraints of the suppliers, anticipating a slight increase of the supply volume that is considered to be feasible. As orders are quoted against the master plan the unrealistic assumption of infinite supply beyond the standard lead-time is obsolete—resulting in more reliable order promises.

In most APS—and also ERP systems—the supply information of the master plan that is used as the basis for order promising is called *available-to-promise (ATP)*. ATP represents the current and future availability of supply and capacity that can be used to accept new customer orders.

Figure 9.2 summarizes the role of master planning for demand fulfillment and ATP. The master planning process is based on the forecast, which reflects the capability of the market to create demand. During the master planning process

[2] For week 1 and 2 the master plan reflects the scheduled receipts.

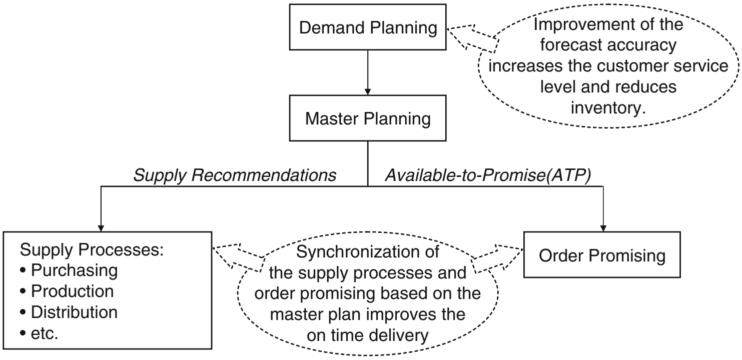
Demand Planning
Improvement of the forecast accuracy increases the customer service level and reduces inventory.
Master Planning
Supply Recommendations
Available-to-Promise(ATP)
Supply Processes:
• Purchasing
• Production
• Distribution
• etc.
Synchronization of the supply processes and order promising based on the master plan improves the on time delivery
Order Promising

Fig. 9.2 Master Planning as the common basis of supply processes and order promising

all material, capacity and time constraints of the supply chain are applied to the forecast, resulting in a feasible master plan. This plan is the common basis for the supply processes (supply recommendations for purchasing, production, distribution etc.) and the order promising process (based on the available-to-promise quantities). By that, supply processes are synchronized with order promising, resulting in reliable order quotes. As a consequence the on time delivery KPI is improved.

Please note that the on time delivery KPI is mainly influenced by the ability of the master planning model to reflect the reality in a sufficiently accurate way. ATP based on an accurate master planning model guarantees almost 100 % on time delivery which is only influenced by supply deviations on the supply side and unexpected capacity problems, e.g. in production, on the capacity side. Apart from on time delivery the delivery performance KPI plays an important role in demand fulfillment as it reflects how fast the supply chain is able to fulfill a customer order. Delivery performance in contrast to on time delivery depends mainly on the forecast accuracy and the ability of the supply chain to satisfy the forecast. The master planning process is responsible to create a feasible supply plan based on the forecast. If the forecast does not mirror future orders very well the probability is low that there is ATP available when a new customer order requests for it. In this case, the customer order receives a late, but reliable promise and the delivery performance is affected. If new customer orders come in as anticipated by the forecast and the master planning process was able to generate a feasible supply plan for the forecasted quantities, then supply matches the demand, the number of inventory turns increases and orders receive a reliable promise within a short lead-time.

As ATP is derived from the supply information of the master plan, it is structured according to the level of detail of the master plan. The typical dimensions for structuring ATP are product, time, supply location, sourcing type, customer, market, region, etc. For example, if the master plan is structured by product group, month, and sales region, the ATP originating from the master plan is represented on the same level of detail. To enable more detailed order promising decisions ATP can be

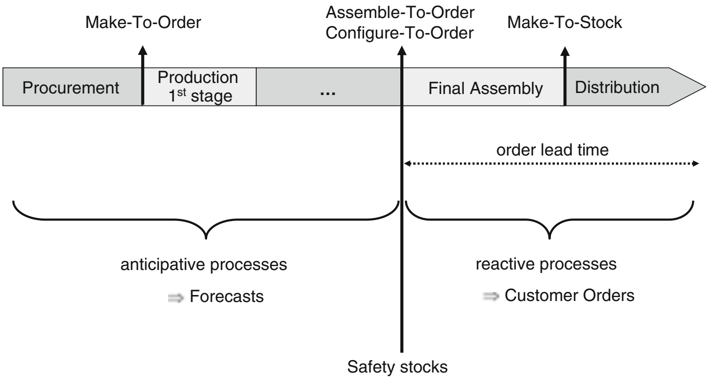

Fig. 9.3 Decoupling point in make-to-stock, assemble/configure-to-order, and make-to-order business environments (adapted from Fleischmann and Meyr 2003)

disaggregated to a more detailed level. In the following Sects. 9.2–9.4 we discuss the structuring of ATP along the most important dimensions: product, time and customer.

9.2 Structuring of ATP by Product

In principal ATP can be represented on any stage of the supply chain, e.g. finished goods, components, or raw materials. The decision where to represent ATP best for a certain business is strongly linked with the location of the decoupling point (see Chap. 3) in that particular supply chain. The decoupling point separates the forecast-driven parts of a supply chain from the order-driven parts (Fleischmann and Meyr 2003). Typically, a safety stock is held at the decoupling point to account for forecast errors. Figure 9.3 shows the location of the decoupling point for make-to-stock (MTS), assemble/configure-to-order (ATO/CTO), and make-to-order (MTO) business environments. Note that decisions on the location of decoupling points are usually made as part of the long term, strategic planning (see Chaps. 1 and 4) because shorter customer order lead times for a downstream decoupling point have to be paid off by higher values and holding costs of decoupling point inventories.

9.2.1 Make-to-Stock

In an MTS environment (see Chap. 3) the standard way to represent ATP is on finished goods level (e.g. actual end products, articles, etc. that are to be sold or aggregated product groups). The supply and production processes in an MTS business are driven by the forecast—not by customer orders. Further, parts of the distribution processes can be forecast-driven (for example if products are to be transported to regional distribution centers, refer to Chap. 12). From there customer orders are served with a shorter lead time than from a central warehouse. The promise would be given under consideration of availability of finished goods ATP and transportation times. Examples for MTS industries are consumer packaged goods, food and beverages, and retail. In some MTS industries the decoupling point even moves with a seasonal pattern in the distribution network.[3]

9.2.2 Assemble/Configure-to-Order

In an ATO environment, all components are produced and/or procured driven by the forecast. Only final assembly is order-driven (see Fig. 9.3). Usually, there are some (or many) configuration options the customer can choose from (e.g. color, technical options, country specific options like power plug), and the actual configuration is determined only at order entry time. This is called *configure-to-order*. In an assemble/configure-to-order environment the forecast is created on finished products or product group level; the forecast is then transformed by master planning into a supply plan on component level. For this, the bills of materials of the finished products are exploded and lead-times and capacity usage are considered. If the master plan is represented on product group level specific *planning bills of materials* for the product groups are used that describe a *typical* representative product of that group. ATP is then represented on component level based on the planned material requirements on component level.

Upon customer order entry, the bill of materials of the (configured) product is exploded, component request dates are derived from the customer requested date, and component availability is checked for all ATP-relevant components. The latest availability date of all ATP-relevant components determines the quote for the complete order; all ATP consumptions are then synchronized according to the final quote, and lead times for assembly and transportation are added. This scheme is also called *multi-level ATP* (Dickersbach 2009), as ATP can be represented on multiple levels of the bill of materials.

For configurable products there exists no deterministic bill of materials representation for final products or product groups. Thus, the distribution of demand for the configuration options must be planned explicitly. For example, consider a color option with three possible values “red”, “blue” and “green”. The demand for the three options may be distributed as follows: “red” 60 %, “blue” 15 % and “green” 25 %. Based on this distribution of the demand for the configuration options and on the forecast on product group level, master planning provides a supply plan on component level, that is then represented as ATP. Consider the computer industry as an example (see Chap. 23 for further details). From a limited number of

[3] In the tire industry the decoupling point is usually located at the central DC. At the start of the winter tire business (in Western Europe usually in October), the demand for winter tires is at peak and exceeds the handling capacity of the central DC. Therefore the decoupling point for winter tire business is moved from the central DC to the regional DCs for that time period.

components—e.g. disk drives, processors, controllers, memory—a huge number of configurations can be made. An order consumes ATP from the base configuration of the computer (motherboard, housing, power supply, key board, etc.), and from all components that were configured by the customer, e.g. speed of processor, size of disk drive and memory.

9.2.3 Make-to-Order

MTO environments are similar to ATO, but the decoupling point is located further upstream. In an MTO environment procurement is driven by forecast, and production, final assembly and distribution are driven by customer orders (see Fig. 9.3). Finished products and components are either customer-specific or there are so many different variants that their demand cannot be forecasted with a high accuracy. Besides material availability, the required capacity is typically an important constraint for the fulfillment of customer orders. Thus, ATP in an MTO environment is representing (a) the availability of raw material (see description of multi-level ATP above) and (b) the availability of capacity. For this purpose, specific ATP sources are formed representing the capacity of a specific kind that is available for promising customer orders. The capacity ATP is either represented in the demand fulfillment module of the APS on an aggregated level (resource groups), or the production planning and scheduling module is used to generate a promise. In the first case, capacity is treated like a component, and the availability of that "capacity component" is checked as described above for ATO and CTO environments. In the second case, the customer order is forwarded to the production planning and scheduling module of the APS, is inserted into the current production plan, and the completion date of the order is returned to the customer as promised date. This concept is also called *capable-to-promise (CTP)*—see, e.g., Dickersbach (2009).[4]

With capable-to-promise, the production process is simulated for the new customer order. This simulation may involve all subordinate production levels (multi-level production). Both, material and capacity availability are checked, resulting in a highly accurate order promise. A further advantage of CTP is that planned production orders are created upon order promising directly, and have only to be changed later if orders have to be replanned (due to material or capacity shortages or additional demand with higher priority). In complex production environments with many levels in the bill of materials and complex capacity constraints including setup constraints, CTP does not lead to an optimal production plan and schedule. The reason is the order-by-order planning scheme applied by CTP, often leading to poor schedules.

[4] Note that capable-to-promise can also be applied to ATO environments.

Please note that the consumption of ATP does not mean that a certain supply represented by ATP is reserved for a certain customer order. ATP is a concept that allows a customer order to enter the planning sphere of a supply chain to a certain date (promise date) so that it can be delivered on time. The detailed material and capacity assignment for a customer order is only done in detailed scheduling and execution to keep the flexibility for optimization.

9.3 Structuring of ATP by Time

ATP is maintained in discrete time buckets. As ATP is derived from the master plan, the ATP time buckets correspond to the time buckets of the master plan. Please note that the master plan time buckets might be different from the time buckets used by demand planning. Usually the master planning and ATP time buckets are more granular than the demand planning time buckets. For example the forecast could be structured in weekly or monthly buckets whereas master plan and ATP could be structured in daily or weekly buckets. Orders are quoted by consuming ATP from a particular time bucket.

The time granularity of the master plan is usually a compromise between the needed level of detail to offer accurate promises and the performance of an APS. The higher the level of detail the more exact a master plan has to be calculated and the more time buckets have to be searched for ATP to generate a promise. An approach to combine the generation of detailed promise dates and the achievement of a high performance is to split the time horizon: in the near term horizon ATP is represented in detailed time buckets (e.g. days or weeks); in the mid term horizon, ATP time buckets are more coarse (e.g. months). The near term horizon is often called *allocation planning horizon*. The concept of allocation planning is described in the next section.

9.4 Structuring of ATP by Customer

A supply chain (or a part of a supply chain) operates either in *supply constrained mode* or in *demand constrained mode*. If material and/or capacity are bottlenecks, then there is "open" demand that cannot be fulfilled. The supply chain supplies less finished goods than the customers request and operates in *supply constrained mode*. If demand is the bottleneck, then all demand may be matched with supply. The supply chain operates in *demand constrained mode*. It is the task of the master planning process to anticipate the operating mode of the supply chain, and to provide good decision support to take appropriate counter measures in advance. In the following we describe both operating modes of supply chains in more detail and explain the impact on the structuring of ATP along the customer dimension.

9.4.1 Demand Constrained Mode

In demand constrained mode the supply chain is able to generate "excess" supply that is not requested and will—most probably—not be consumed by customers. In demand constrained mode, the master plan must help to identify sources of excess supply. The usage of the corresponding supply chain components might then be reduced in order to save costs, or additional demand has to be generated by promotional activities or other additional sales measures.

The capability of a supply chain to produce excess supply is an indicator for inefficiencies in the supply chain (refer to Chap. 15). A supply chain is working more profitable if it is "operated on the edge" (Sharma 1997) by removing all inefficiencies, e.g. excess capacity, excess assets and excess expenses. Thus, on the long term a demand-constrained supply chain should move toward supply constrained mode. This can be achieved either by generating additional demand or by reducing the ability of the supply chain to generate excess supply (see Chap. 6).

In demand constrained mode there are no specific considerations for structuring ATP as all demand can be fulfilled by the supply chain.

9.4.2 Supply Constrained Mode

In supply constrained mode, not all customer demand can be fulfilled. Master planning must support the decision how to generate additional supply and also how to allocate supply to demand. If orders are promised on a *first-come-first-served policy*, all orders are treated the same without taking the profitability of the order, the importance of the customer and the fact whether the order was forecasted or not into account. As a consequence the profitability of the business, the relationship to the customers and the performance of the supply chain may be jeopardized.

A good example of how business can be optimized by using more sophisticated order promising policies than first-come-first-served is given by the revenue management activities of international premium airlines (Smith et al. 1992). Premium airlines keep a specific fraction of the business and the first class seats open even if more economy customers are requesting seats than the total number of economy seats. For each flight, some of the business class and first class seats are *allocated* to the business and first class passengers based on the forecasted passenger numbers for that flight. Only a short time before the flight departs the allocations are released and passengers are "upgraded" to the next higher class. By that, airlines achieve a higher average sales price for the available seats and strengthen the relationship to their important customers, the business class and first class passengers.

Talluri and van Ryzin (2004) classify revenue management in price-based and quantity-based approaches. Price-based approaches try to gain higher revenues by varying the sales prices over time, thus actively influencing demand. This is commonly practiced by budget airlines and in retail (see also Elmaghraby and Keskinocak 2003), but also important for promotions planning within demand planning modules of APS (see Chap. 7). Quantity-based approaches, however, segment customers into several groups showing different buying behavior, strategic importance and/or average profits. Thus they try to exploit the customers' different willingness to pay or various profit margins, as it has been illustrated in the premium airline example above.[5]

APS also apply these quantity-based ideas by allocating ATP quantities to customer groups or sales channels in order to optimize the overall business performance. A classification scheme is defined that is used to segment and prioritize customer orders. Typically, the order classes are structured in a hierarchy. The ATP quantities are allocated to the order classes according to predefined business rules (also called *allocation rules*, ref. to p. 187). These allocations represent the right to consume ATP. The principal connection between allocations and ATP is straightforward: When an order is entered, the order promising process checks the allocations for the corresponding order class. If allocated ATP is available, ATP can be consumed and the order is quoted accordingly. Otherwise, the order promising process searches for other options to satisfy the order, e.g. by checking ATP in earlier time buckets, by consuming ATP from other order classes (if that is allowed by the business rules defined) or by looking for ATP on alternate products.

The time buckets for the allocations and the actual ATP quantities may differ. Allocations must be carefully controlled and regularly adjusted by human planners (as described for the "airline" example above). Otherwise the order lead time for some order classes will deteriorate while the ATP buckets for other order classes remain full as they are not consumed as anticipated. Thus, it is helpful to provide allocations in a larger granularity, e.g. weeks or months than the actual ATP quantities, in order to support the manual control and adjustment processes. Furthermore, the two levels of granularity for allocations and actual ATP quantities provide the opportunity to implement a two-phased order promising process: In step 1, customers receive the allocation time bucket (e.g. a week) as delivery date. In step 2, this initial promise is detailed down to the actual delivery day depending on the actual consumption of ATP. A two-step order promising approach keeps a certain degree of flexibility until the actual delivery day is promised to the customer in step 2.

The allocation of ATP to order classes can be exploited to increase the revenue and profitability of the business. For example, the average selling price may be increased by allocating supply to customers that are willing to pay premium prices, instead of giving supply away to any customer on a first-come-first-served basis. Traditional ATP mechanisms without allocation rules have to break commitments that have been given to other customers in order to be able to quote an order of a key customer or an order with a higher margin. It is obvious that this business policy has a negative impact on the on time delivery and deteriorates the relationship to other customers.

[5] For an overview on the relations between revenue management, demand fulfillment and ATP, inventory management/rationing, and pricing the reader is referred to Quante et al. (2009).

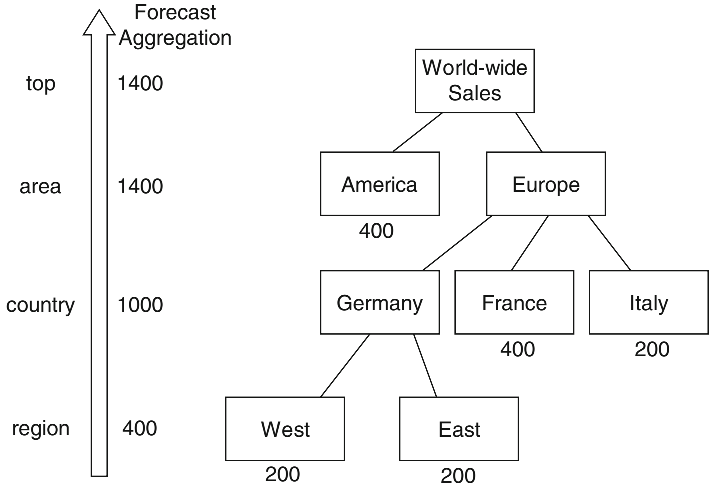

Fig. 9.4 Sales forecast aggregated along the customer hierarchy

9.4.3 The Customer Hierarchy and Allocation Rules

In order to allocate supply to customers a *model of the customer structure* and a *forecast of the future customer demand* is required. The model of the customer structure should be aligned with the geographic dimension in demand planning (see Chap. 7), as demand planning is structuring the forecast in terms of the geographic dimension. Hence, the customer structure forms a hierarchy similar to the geographic dimension in demand planning. Figure 9.4 shows an example of a customer hierarchy.

In the first step the forecast quantities for each customer (or customer group, resp.) are aggregated to the root of the hierarchy. This number gives the total forecast for that specific product (or product group). The total forecast is transferred to master planning, and master planning checks whether it is feasible to fulfill the total forecast considering the supply constraints. In our example, the total forecast is 1,400, and we assume that master planning can confirm only 1,200 to be feasible.

In the second step the total feasible quantity according to the master plan is allocated from the top down to the leaves of the customer hierarchy. This allocation process for our example is visualized in Fig. 9.5 (the quantities in parentheses indicate the original forecast for this customer group). The allocation of the master plan quantities to the nodes of the customer hierarchy is controlled by allocation rules. In our example we have used three different allocation rules:

- *Rank based*: U.S. customers receive a higher priority (rank 1) compared to customers in Europe (rank 2). Thus, the available quantity for the U.S. and European customers is allocated to the U.S. first up to the original forecast for that area. A rank-based allocation policy may be helpful to support sales to a specific market, e.g. if the development of that market is in an early stage.

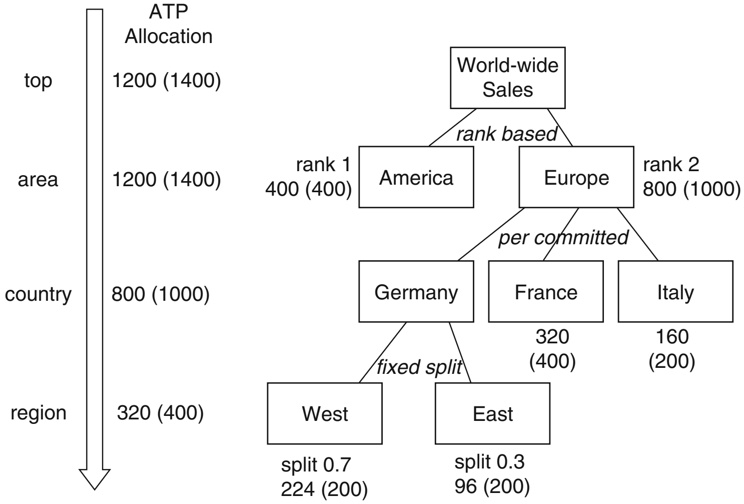

Fig. 9.5 Allocation of ATP in the customer hierarchy

- *Per committed*: The available quantity is allocated to the nodes of the customer hierarchy according to the forecast the customers have committed to. In our example Germany and France have forecasted 400 each, and Italy has forecasted 200, making 1,000 in total. However, for this group of customers, only 800 is available. The quantity of 800 is split according to the fraction of the original forecast, i.e. Germany and France receive 40 % each (320), and Italy receives 20 % (160). The per committed allocation policy is well suited if each customer group shall get a fair share allocation according to what has been forecasted by that customer group.[6]
- *Fixed split*: The fixed split allocation policy applies predefined split factors to distribute the feasible quantity to the customer groups. In our example, the customers in the Western part of Germany receive 70 % of the available quantity, the customers in East Germany 30 %. Please note that the resulting quantities are independent of the individual forecast of the customer groups. (But it does depend on the total forecast of these customer groups.)

In addition to these allocation rules a portion of the available quantity can be retained at every level of the customer hierarchy. These retained quantities are consumed based on a first-come-first-served policy. Retained ATP can be used

[6] In allocation situations (supply constrained supply chain) the per committed allocation policy may lead to a so-called *shortage gaming* behavior, as planners are motivated to forecast higher quantities than actually needed in order to increase their allocations. It is necessary to establish incentive systems to prevent shortage gaming. Otherwise this behavior may induce a bullwhip effect into the supply chain. Shortage gaming and the bullwhip effect are described in more detail in Chap. 1.

to account for potential variations of the actual demand related to the forecasted demand. For example, if 25 % of the total quantity available for European customers is retained at the customer group Europe, 200 would be available on a first-come-first-served basis for all European customers, and only 600 would be allocated to German, French and Italian customers as defined by the corresponding allocation rules. The retainment of ATP can be interpreted as a *virtual* safety stock on an aggregate level, as it helps to balance deviations between forecast and actual demand.

The allocations are the basis for generating order quotes. Thus, the allocations are an important information for the sales force before making commitments to their customers. Further, the APS keeps track of the consumptions due to already quoted orders. The total allocated quantities and the already consumed quantities give a good indication whether the order volume matches the forecast. If orders and forecast do not match, some allocations are being consumed too fast, whereas others remain unconsumed. This can be sent as an early warning to the supply chain that the market behaves differently than forecasted—and an appropriate action can be taken. For example, sales can setup a sales push initiative to generate additional demand to consume the planned ATP.

9.4.4 Allocation Planning

The process that assigns the overall ATP quantities received from master planning to the nodes of the customer hierarchy is called *allocation planning*. Allocation planning is executed directly after a new master plan has been created—which normally takes place once a week. Thus, once a week the adjusted forecast is transformed into ATP by master planning and allocated to the customer hierarchy.

In addition to that, the allocations are updated on a daily basis in order to reflect changes in the constraints of the supply chain. For example, if the supplier of some key component announces a delay of a scheduled delivery this may impact the capability of the supply chain to fulfill orders and—because of that—should be reflected in the ATP as soon as the information is available in the APS. Please note, if ATP is short and if additional capacity and raw material might be available, which have not been needed in the last master planning run and thus have not been exhausted, a new run has to be triggered in order to generate a new ATP picture reflecting the current supply capabilities of the supply chain.

The planning horizon of allocation planning cannot be longer than the planning horizon of master planning, as no ATP is available beyond the master planning horizon. The master planning process covers usually 6–12 months. However, in many cases it is not necessary to maintain allocations over 6 months or more. For example, in the computer industry, 90 % of the orders are placed 3 weeks prior to the customer requested delivery date. Thus, dependent on the lead-time from order entry date to the customer requested delivery date a shorter planning horizon for allocation planning can be chosen compared to master planning. In the computer industry, for example, a 3-months horizon for allocation planning is sufficient.

9.5 Order Promising

Order promising is the core of the demand fulfillment process. The goal is to create reliable promises for the customer orders. The quality of the order promising process is measured by the on time delivery and the delivery performance.

The *on time delivery* KPI is described in detail in Chap. 2; it measures the percentage of the orders that are fulfilled as promised (based on the first promise given). Thus, to achieve a high on time delivery it is important to generate reliable promises. A promise is *reliable* if the supply chain is able to fulfill the order as promised, i.e. if the customer receives the promised product in the promised quantity at the promised date. A supply chain that is able to consistently generate reliable promises over a long time period gets a competitive advantage over supply chains with a lower on time delivery.

There are multiple execution modes to promise customer orders (Ball et al. 2004; Pibernik 2005):

- *On-line ("real-time") order promising:* A new customer order is promised during the order entry transaction. After the new order is booked, the promised date and quantity are transferred to the customer immediately.
- *Batch order promising:* Customer orders are entered into the sales transaction system without generating a promise. At certain periods (e.g. once per day) a batch order promising is triggered and all new orders receive a promise. For instance, these promises could be generated by a production planning module.
- *Hybrid order promising:* Each new customer order is temporarily promised at order entry time. In addition to that a batch order promising run is triggered regularly in order to detail the promises (cf. Ball et al. 2004, Sect. 2.3 describing an example of the Dell Computer Corporation and the two-phase approach of Sect. 9.4.2) or to improve the promises of all customer orders in total (re-promising). Please note that hybrid order promising schemes can lead to changes and delays of customer order promises.

On-line order promising offers advantages in responsiveness and performance toward the customer. On the other hand, on-line order promising prevents that promises can be reviewed by order management before they reach the customer. In the remainder of this chapter we focus on on-line order promising schemes.

9.5.1 ATP Search Procedure

The general ATP-based order promising process works as follows: First, the order promising process searches for ATP according to a set of search rules that are described below. If ATP is found, it is reduced accordingly and a quote for the order is generated. If ATP can only be found for a portion of the ordered quantity and partial fulfillment of the order is allowed, a quote is generated for the partial order quantity. If no ATP can be found, no quote is generated, and the order must be either rejected or confirmed manually at the end of the allocation planning horizon.

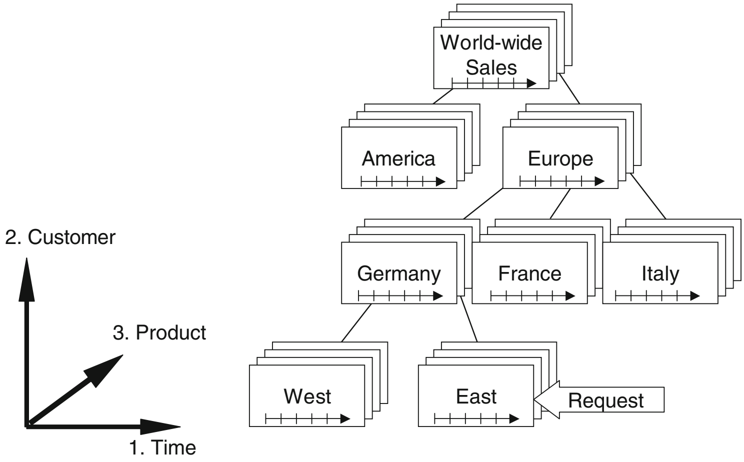

Fig. 9.6 Three dimensions of ATP search paths

Note that if no ATP can be found for an order, the supply chain will not be able to fulfill the order within the allocation planning horizon.

In principal ATP can be searched along all dimensions used to structure ATP (see Sect. 9.1). In the following, we describe the ATP search procedure based on examples where ATP is structured by time, customer and product. Figure 9.6 illustrates these three dimensions of the ATP search paths. The following *search rules* are applied (for simplicity we assume that the ATP is on finished goods level; the search rules are similar for ATP on product group level and component level):

1. The leaf node in the customer hierarchy, to which the customer belongs, the product being requested by the order and the time bucket containing the customer requested date are determined. The ATP at this point is consumed—if available.
2. If ATP is not sufficient, then the time dimension is searched back in time for additional ATP (still at the leaf node in the customer hierarchy and at the product requested by the order); all ATP found up to a predefined number of time buckets back in time is consumed. Note that if ATP is consumed from time buckets earlier than the time bucket containing the customer requested date, the order is pre-built, and inventory is created.
3. If ATP is still not sufficient, steps 1 and 2 are repeated for the next higher node (parent node) in the customer hierarchy (searching for retained ATP quantities), then for the next higher and so on up to the root of the customer hierarchy.
4. If ATP is still not sufficient, steps 1–3 are repeated for all alternate products that may substitute the original product requested by the order.
5. If ATP is still not sufficient, steps 1–4 are repeated, but instead of searching backward in time, ATP is searched forward in time, up to a predefined number of time buckets. Note that by searching ATP forward in time, the order will be promised late.

The set of search rules described above is only one example of an ATP search strategy. In fact, an ATP search strategy may consist of any meaningful combination and sequence of the following search rule types:

- *Search for Product Availability:* This is the standard ATP search for a product including future receipts and constraints.
- *Search for Allocated ATP:* ATP is searched for along the customer dimension.
- *Search for Forecasted Quantities:* The creation of a quote for an order is based on forecasted quantities. The forecasted quantities in general are not customer specific.
- *Search for Component Availability:* For complex production processes and bill of materials structures a multi-level ATP search for component availability is performed.
- *Capable-to-Promise:* ATP is dynamically generated by invoking the production planning and scheduling module.
- *Perform Substitution:* If no ATP can be found for a given product in a given location this type of rule allows to search for (a) the same product in another location, or (b) another product in the same location, or (c) another product in other locations. This so-called *rule-based ATP* search requires the maintenance of lists of alternate products and/or locations and a rule to define the sequence in which the product and/or location substitutions are to take place.

In the following, we illustrate the ATP search procedure by means of a simple example.

9.5.2 ATP Consumption by Example

Let us assume an order is received for 300 units from a customer in East Germany, with a customer requested date in week 4. The ATP situation for East Germany is depicted in Fig. 9.7. First, the ATP is checked for the customer group East Germany for week 4, then for week 3 and for week 2. (We assume that the ATP search procedure is allowed to consume ATP 2 weeks back in time.) The ATP that is found along that search path is 10 in week 4, 60 in week 3 and 50 in week 2, 120 in total (see Fig. 9.7).

As the ATP search procedure may not consume ATP from a time bucket that is more than 2 weeks prior to the customer requested date, 180 units of the requested quantity is still open after the first step. In the second step, ATP is searched along the customer dimension. We assume for this example that there is ATP in the next higher node in the customer hierarchy, i.e. Germany as shown in Fig. 9.8, but no ATP in the next higher nodes, i.e. Europe and World-wide Sales. From the ATP allocated at Germany, another 120 units can be consumed in weeks 4, 3 and 2, resulting in a total promised quantity of 240. Sixty units are still open, as the requested quantity is 300 units.

In the next step, the ATP search algorithm looks for alternate products as shown in Fig. 9.9. The alternates are sorted by priority. First, the alternate with the highest priority is considered, and the same steps are applied as for the original product, i.e.

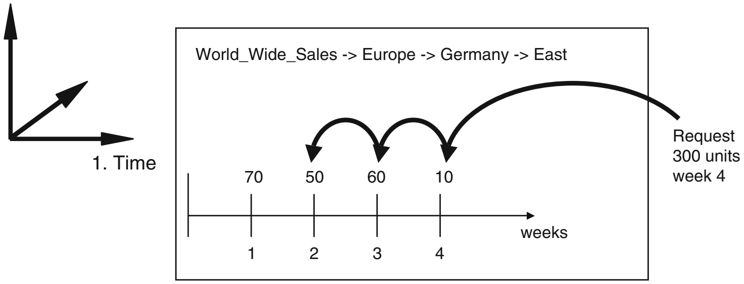

Fig. 9.7 Consumption of ATP along the time dimension

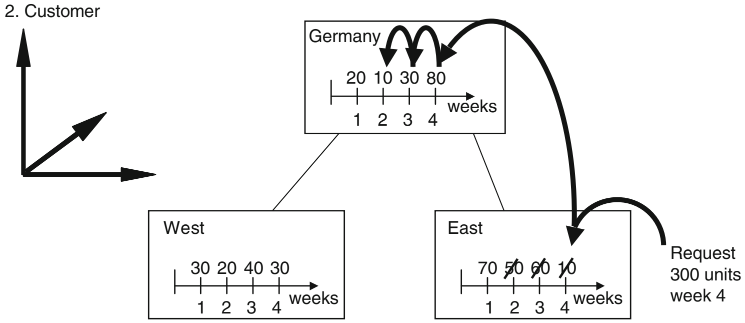

Fig. 9.8 Consumption of ATP along the customer dimension

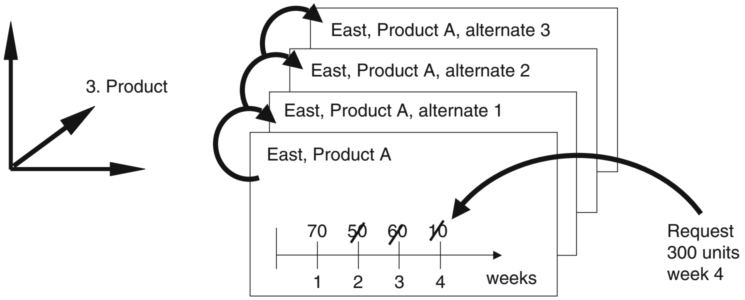
3. Product
East, Product A, alternate 3
East, Product A, alternate 2
East, Product A, alternate 1
East, Product A
70 50 60 10
1 2 3 4
weeks
Request
300 units
week 4

Fig. 9.9 Consumption of ATP along the product dimension

first search back in time and second search up the customer hierarchy. Then, these steps are applied to the alternate with the second highest priority and so on.

The reader who is interested in scientific research on optimization-based allocation planning and order promising is, for example, referred to Meyr (2009) and Quante (2009) for single-level and Vogel (2014) for multi-level customer hierarchies.

References

Ball, M. O., Chen, C.-Y., & Zhao, Z.-Y. (2004). Available-to-promise. In D. Simchi-Levi, S. D. Wu, & Z.-J. Shen (Eds.), *Handbook of quantitative supply chain analysis – Modeling in the e-business era* (Chapter 11, pp. 447–483). Boston: Kluwer Academic.

Dickersbach, J. (2009). *Supply chain management with SAP APO* (3rd ed.). Berlin: Springer.

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science*, *49*(10), 1287–1309.

Fleischmann, B., & Geier, S. (2012). Global available-to-promise (global ATP). In H. Stadtler, B. Fleischmann, M. Grunow, H. Meyr, & C. Sürie (Eds.), *Advanced planning in supply chains—Illustrating the concepts using an SAP APO case study* (Chapter 7, pp. 195–215). Berlin: Springer.

Fleischmann, B., & Meyr, H. (2003). Customer orientation in advanced planning systems. In H. Dyckhoff, R. Lackes, & J. Reese (Eds.) *Supply chain management and reverse logistics* (pp. 297–321). Berlin: Springer.

i2 Technologies Inc. (2000). *RHYTHM demand fulfillment concept manual, Part No. 4.2.1-DOC-ITM-DFCM-ALL*. Irving, TX: i2 Technologies Inc.

Meyr, H. (2009). Customer segmentation, allocation planning and order promising in make-to-stock production. *OR Spectrum, 31*(1), 229–256.

Pibernik, R. (2005). Advanced available-to-promise: Classification, selected methods and requirements for operations and inventory management. *International Journal of Production Economics, 93*, 239–252.

Quante, R. (2009). *Management of stochastic demand in make-to-stock manufacturing. Forschungsergebnisse der Wirtschaftsuniversität Wien* (Vol. 37). Frankfurt a.M. et al.: Peter Lang.

Quante, R., Meyr, H., & Fleischmann, M. (2009). Revenue management and demand fulfillment: Matching applications, models, and software. *OR Spectrum, 31*(1), 31–62.

Schwendinger, J. (1979). Master production scheduling's available-to-promise. In *APICS Conference Proceedings* (pp. 316–330).

Sharma, K. (1997). Operating on the edge. private conversation.

Smith, B., Leimkuhler, J., Darrow, R., & Samuels, J. (1992). Yield management at American Airlines. *Interfaces, 22*, 8–31.

Talluri, K. T., & van Ryzin, G. J. (2004). *The theory and practice of revenue management. International Series in Operations Research & Management Science* (Vol. 68). London: Kluwer Academic.

Vogel, S. (2014). *Demand fulfillment in multi–stage customer hierarchies. Produktion und Logistik*. Wiesbaden: Springer Gabler.