Abstract

Mass Customization (MC) is not a mature business strategy and hence it is not clear that a single or small group of operational models are dominating. Companies tend to approach MC from either a mass production or a customization origin (Duray 2002) and this in itself gives reason to believe that several operational models will be observable. This paper reviews actual and theoretical fulfillment systems that enterprises could apply when offering a pre-engineered catalogue of customizable products and options. Issues considered are: How product flows are structured in relation to processes, inventories and decoupling point(s);
• Characteristics of the OF process that inhibit or facilitate fulfilment;
• The logic of how products are allocated to customers;
• Customer factors that influence OF process design and operation.
Diversity in the order fulfillment structures is expected and is found in the literature. The review has identified four structural forms that have been used in a Catalogue MC context:
• fulfilment from stock;
• fulfilment from a single fixed decoupling point;
• fulfilment from one of several fixed decoupling points;
• fulfilment from several locations, with floating decoupling points.
From the review it is apparent that producers are being imaginative in coping with the demands of high variety, high volume, customization and short lead times. These demands have encouraged the relationship between product, process and customer to be re-examined. Not only has this strengthened interest in commonality and postponement, but, as is reported in the paper, has led to the re-engineering of the order fulfilment process to create models with multiple fixed decoupling points and the floating decoupling point system.

Keywords: Order fulfilment,

Author Biographies

Philip Brabazon is a Research Fellow in the Operations Management Division on the Nottingham University Business School. His area of research is the Mass Customization of products and services and his interests include the development of quantitative and qualitative operational templates.
philip.brabazon@nottingham.ac.uk, +44 (0) 115 951 4011

Bart MacCarthy is Professor of Operations Management at Nottingham University Business School and Director of the Mass Customization Research Centre (MCRC). As well as Mass Customisation, his research interests include the analysis and design of operational systems in business and industry with particular emphasis on responsiveness and time compression across the extended enterprise. He has researched and consulted with a wide range of industries including textiles and clothing, automotive, engineering, aerospace, consumer products and food, as well as with firms in distribution and logistics. He is a Fellow of IEE, The Institute of
1. Introduction

Mass Customization (MC) is not a mature strategy and hence it is unlikely that a single or small group of operational models dominate. From an examination of cases MacCarthy et al (2003) distinguished five operational modes. One of these modes, Catalogue MC, is the focus of attention in this paper and is defined as the mode in which a customer order is fulfilled from a pre-engineered catalogue of variants produced using standard order fulfilment processes. In this Mode the engineering of products is not linked to orders, but completed before orders are received. Customers select from a pre-specified product/option range and the products are manufactured by the order fulfilment activities that are in place. Likewise the order fulfilment activities are designed and engineered ahead of an order being taken.

Even when limiting the focus to the Catalogue MC Mode there is no reason to believe that organisations are constrained to one model. Companies are approaching MC and the Catalogue mode either from a mass production or a customization origin (Duray 2002) and this in itself is reason to believe that several order fulfilment models will be observable. In some cases it is product proliferation that has motivated the uptake of Catalogue MC, where variety is required for market segments, such as a global product that needs to be differentiated for different markets (Feitzinger & Lee 1997), and in other cases it is customization for the end customer that is the motivation, such as computer servers (Swaminathan & Tayur 1998). The diversity of contexts is further reason to believe that a number of models are being applied.

2. Delineating the order fulfilment processes

Order fulfilment is not a universally used term as noted by Kritchanchai & MacCarthy (1999) who found ‘few sources in the literature discussing the details of the order fulfilment process explicitly’. There is no standard definition of order fulfilment and no common understanding of what activities it involves.

In the context of manufacturing, it is intuitive to say that order fulfilment involves the handover of material to the customer. Beyond this it is less certain as to what should be treated as part of the OFP. If the goal of order fulfilment is to comply with the customer’s requirements, in particular the WHWW details (What product(s), How many, Where to deliver to, When to deliver) then a) the OFP is not involved directly with the customer to take the details of the order and b) for not only must OFP encompass some material processing/transportation activities but also some element of control logic. At one extreme the logic may be a simple rule for which product to take from the shelves and the activity be nothing more than handing it over to the customer, but at the other extreme the OFP may involve the triggering and sequencing of complex production and distribution processes. Therefore, while the details and scale of the OFP from one situation to another might differ greatly, in general terms the OFP encompasses the material processing activities concerned with complying with customer instructions and the control of these activities.

It is tempting to use the Customer Order Decoupling Point (CODP) which is ‘traditionally defined as the point in the manufacturing value chain where the product is linked to a specific customer order’ (Olhager 2003) to delineate the OFP. However, although activities upstream of the CODP will be controlled by forecasts, the state of these activities can have a bearing on the
future performance of the downstream activities. This is particularly relevant when customer orders are conditional on delivery dates promised during the sales negotiation. Although in some manufacturing systems the upstream and downstream activities could be insulated from each other, in general there are dependencies between them. If the customer’s WHWW requirements are to be fulfilled, the OFP must have good situational awareness of the system – i.e. a grasp of the current state of the material processing activities, how they got into this state and, more importantly, how they are going to develop over time. For this to be the case the OFP cannot be blind to the upstream activities and, consequently, it is not appropriate to use the CODP as an OFP boundary marker.

The process of Demand Management, as described by Vollmann et al (1997) perhaps provides a template for defining and describing OFP. To Vollmann et al demand management is a highly integrative activity that captures and co-ordinates demand on manufacturing capacity. They say ‘the basic concept of demand management is that there is a pipe of capacity which is filled in the short run with customer orders and the long run with forecasts; order entry is a process of consuming the forecast with actual orders’. To them it encompasses forecasting, order entry, order-delivery-date promising, customer order service, physical distribution and other customer-contact-related activities.

2.1. Interpretation of order fulfilment and scope of the review

In this review, order fulfilment is interpreted in the following way:

- The OFP receives and acts upon customer orders, which contain the WHWW details (What product(s), How many, Where to deliver to, When to deliver);
- The OFP requires an awareness of the current and future state of the material processing activities. It envisages a pipeline of real and planned products and links customers to either type of product;
- The activities upstream of the CODP are within the bounds of the OFP if downstream activities are dependent on their performance.
- Using this interpretation, literature that addresses the following issues with catalogue mass customization is relevant to the review:
  - How products flows are structured in relation to processes, inventories and decoupling point(s);
  - Characteristics of the OFP (pipeline) that inhibit or facilitate fulfilment;
  - The logic of how products are allocated to customers;
  - Customer factors that influence order fulfilment process design and operation.

3. Order fulfilment structures

The relative positions of processes and inventories are a fundamental aspect of order fulfilment models, as illustrated by Bucklin (1965) in his comparison of the speculation and postponement strategies. Compared to the speculation model, the stock of finished goods is not a component of the postponement model (Figure 3:1).
The review of literature has identified four structural forms that have been claimed to be used in a catalogue MC context:

- fulfilment from stock;
- fulfilment from a single fixed decoupling point;
- fulfilment from one of several fixed decoupling points;
- fulfilment from several locations, with floating decoupling points.

### 3.1. Fulfilment from stock

Product variety has been on the increase in many sectors (Cox & Alm 1998) and since fulfilment of customers from stock is still prevalent it is unsurprising that examples can be found that claim to have adapted this configuration to high variety / mass customization situations.

There is a degree of uncertainty over whether to include these stock fulfilment models in the review and which papers qualify for inclusion. For example, the customization of printers by Hewlett Packard (Feitzinger & Lee 1997, Lee & Billington 1995) is heralded as mass customization but the end consumer is not involved in the process. The customization is required for the region in which the printer is to be sold and hence the study could be relabelled as solely a case-study in postponement. There is an argument that it illustrates postponed manufacture and that the label of mass customization was given to it before the strategy was more widely scrutinised. It is include it in the review, with two other examples, to show the diversity of approaches for coping with high product variety. The three examples are summarised below and in Figure 3:2:

- Hewlett-Packard printers are customized for each region by postponing some assembly and packaging activities. Standard unfinished units are shipped from a central facility to each region for completion (Feitzinger & Lee 1997, Lee & Billington 1995).
- In the context of the automotive sector Boyer & Leong (1996) study a structure in which multiple product types are supplied to many stock locations. They study the impact on the system of increasing the number of products that each plant can produce.
- Herer et al (2002) examines the method of transhipment for high variety of products, which is the ability to transfer stock between locations at the same echelon level. Transhipment is a form of physical postponement and as Herer puts it, creates the ability to transform a generic item (an item at any location) into a specific item (an item at a specific location) in a relatively short time.
A theme of the research into stock fulfilment structures is how to structure the processes that replenish the stock to cope with variety without suffering high costs. Hau Lee is one of the principal contributors in this area and he sees a key issue to be in how product design interacts with the process (e.g. Lee & Billington 1995, Lee 1996, Lee & Tang 1997, Lee & Tang 1998, Whang & Lee 1998). Whang & Lee (1998) present models to indicate the scale of benefit that postponement can bring through uncertainty reduction and reduced forecasting error. Lee & Tang (1997) use a model to study three approaches to delaying product differentiation, taking forward the models of Lee (1996). Lee & Tang (1998) study further the approach of operations reversal and put forward properties that an order fulfilment sequence should strive for when the major source of demand uncertainty lies in the option mix and the total demand for all options is fairly stable.

3.2. Fulfilment from a single fixed decoupling point

This structure takes the form of the postponement model described by Bucklin (1965, see Figure 3:1). Of the four types of OFS structures this is the format that tends to be associated with catalogue mass customization. In this structure the producer holds stocks of raw materials or part-finished products and once an order is received there are taken forward to be completed and delivered to the customer.

A standard classification of order fulfilment systems includes a set of fixed decoupling point structures: engineer-to-order (ETO); make-to-order (MTO); and assemble-to-order (ATO). Hill (1995) extended this by adding design-to-order and make-to-print. Recently, the category of configure-to-order (CTO) has been distinguished as a special case of assemble-to-order (Song & Zipkin 2003), in which components are partitioned into subsets from which customers make selections (e.g. a computer is configured by selecting a processor from several options, a monitor from several options, etc).
Many practicing mass customizers have one decoupling point and fit into the assemble-to-order or configure-to-order categories, though they can also perform some fabrication:

- Kotha (1995) describes the Japanese bicycle company, National Panasonic, who await each order before fabricating the frame and assembling the bicycle with components from stock;
- A series of articles describe how the UK company, RM, switched its computer supply business from a make-to-stock to an assemble-to-order fulfilment mode (Duffel 1999, Duffel & Street 1999).
- Orangebox is a UK company producing office furniture. Their products are modular and they produce high levels of variety in small batch sizes. Once an order is received they cut and sew the covers and assemble the product from components in stock (Tozer 2003).

3.3. Fulfilment from one of several fixed decoupling points

These structures have more than one decoupling point, i.e. there are two or more distinct stock holding locations among the production and delivery processes from which raw materials or part-finished products can be taken, allocated to a customer, finished and delivered. A customer need not be aware of which decoupling point is being used for their order.

Graman & Magazine (2002) study an OFP with two fixed decoupling points – one is mid-process and the other is the finished stock (Figure 3:3). They conclude that holding some items in a part-finished state and retaining some final processing capacity open to be used to fulfil orders can bring significant performance benefits, compared to a situation in which all orders are filled from stock.

Figure 3:3 Structure studied by Graman & Magazine (2002)

Swaminathan & Tayur (1998) go a step further and study an OFP with three fixed decoupling points – one preceding final assembly, one mid-assembly and finished stock (Figure 3:4). They develop a model to tackle a problem in which a producer offers a broad product range but in each time period orders are received for a fraction of variants only. They compare a vanilla box strategy (in which sub-sets of components are pre-assembled into a number of vanilla boxes, exploiting the inherent commonality in the product family) against MTS and ATO strategies (and mixes of the three) and find the vanilla box approach can be superior significantly. In exploring their model, they show how factors including capacity constraints, demand correlation, number of vanilla box types and breadth of product range alter the performance of each strategy. In a second study, Swaminathan & Tayur (1999) go on to develop models that take account of assembly precedence constraints, in particular the feasibility of a vanilla box in terms of whether it can be assembled.
3.4. Fulfilment from several process points, with floating decoupling points

The key feature of order fulfilment systems with this structure is that products can be allocated to orders at any point along the process, hence the coining of the term *floating decoupling point*. This structure is observed in the capital goods sector but is being adopted elsewhere including the automotive sector.

Manufacturers of complex goods such as machine tools have been facing the challenges of increased product diversity and shortening of delivery lead times. The requested delivery lead time is often less than the sum of purchasing, fabrication and assembly lead times. As a consequence such companies have been evolving their order fulfilment processes. In their study of three heavy manufacturing firms Raturi et al (1990) describe how firms have implemented a build-to-forecast (BTF) schedule in which they forecast end-product mix, create a master schedule of end-products and then release production orders before specific customer orders are received (Figure 3:5). In BTF there is no stopping point in the production process and buffer inventories are avoided. Customer orders are matched to items in any state of production that will meet the due date. Customer orders rarely match the end products being built hence orders are fulfilled by:

- changing products early in the process if the basic model is an appropriate one and the production plan can be altered to accommodate the actual order;
- reconfiguring an end product, with features removed and replaced as required. On occasion the changes are so extensive that a loss is incurred.

The development of information systems has led automotive fulfilment processes to evolve into this type of structure. The multi-mechanism system has been labeled *Virtual Build-to-Order* and Agrawal et al (2001) describe it as connecting customer ‘either via the internet or in dealer’s showrooms, to the vast, albeit far-flung, array of cars already in existence, including vehicles on dealer’s lots, in transit, on assembly line, and scheduled for production’, with the expectation that ‘customers are likely to find a vehicle with the colour and options they most want’. Holweg (2000) also describes the multiple fulfilment mechanisms by which a customer can receive a
vehicle: from the local dealer’s stock; by a transshipment from another dealer’s stock; by a vehicle
taken from a central stock holding centre; by a vehicle being submitted into the order bank as a
build-to-order product; or by a vehicle that is in, or scheduled for, production being allocated to
the customer, which may involve its specification being amended (summarized in Figure 3:6).

Figure 3:6 Structure studied by Holweg (2000)

This model has been studied by Brabazon and MacCarthy (2004a, b, 2005). This work has
identified reconfiguration flexibility as being a key capability of the system. The greater the
flexibility the greater the likelihood a customer can be matched to a product in the production
and distribution pipeline. Their simulation based studies have shown an unexpected result that
this fulfilment system has a propensity to cause average stock levels to rise even when production
and demand are harmonised.

4. Pipeline characteristics

In environments of high product variety and customization, the characteristic of flexibility is
picked up in the literature as being the key facilitator/inhibitor. Several sources consider
flexibility to be an enabler of mass customization (Fogliatto et al 2003, Da Silveira et al 2001,
Kakati 2002, Duffell & Street 1998) and the ability to be flexible is assumed within analyses of the
economic impact of mass customization (de Vaal 2000, Norman 2002). A wider range of
products and increased customization are identified by De Toni & Tonchia (1998) as two of five
motivations for flexibility, the others being: variability of demand (random or seasonal); shorter
life-cycles of the products and technologies; and shorter delivery times.

There is a considerable body of flexibility research but the breadth of the topic is vast with the
topic being approached at one end as a concept, and at the other it is examined in the context of
a specific situation. The scale of concern ranges from the flexibility of a sector down to the
flexibility of a machine or fixtures and the concept also has a temporal property – flexibility over
a short or long time horizon. So wide ranging is the topic that within the large volume of
literature there is little that focuses on the flexibility of mass customization systems specifically.
This is not to say that flexibility has not been of interest in mass customization research,
however, while it is not difficult to argue that for many of the studies referred to above flexibility
is important, flexibility has not been the focus.

Several studies have been identified that assess flexibility and are relevant to catalogue mass
customization.

Bradley & Blossom (2001) estimate the change in cost and the improvement in delivery lead
time that would be achieved by an assembly process if it were to accept a higher proportion of
customer orders. The study is in the automotive sector and the order fulfilment process under
consideration resembles a floating decoupling point system. The study does not look at how
customer orders are matched to vehicles in the pipeline, but recognizes this is an area that needs
attention. Their supposition is that flexibility can be increased in the assembly line by adding production capacity (people or equipment) so that a fluctuating mix of products can be produced. Thus the products made on the assembly line can be those that the customers want, when they want them, rather than units selected for attainment of maximum efficiency. By simulating a generic automotive system they estimate, in the worst case, cost would rise by around 0.017% at a level of 70% make-to-order (a significant reason being that direct labour accounts for only 6% of costs typically) and delivery lead times would reduced by around 70%.

Bukchin et al (2002) develop a heuristic for designing assembly lines for mixed model operations. They assume the model-mix is determined ahead of time and stable (say for a year ahead) but the sequence of launching products to the line must be determined by actual short range demand patterns and customer orders. Their approach assumes a model mix for which the combined workload is balanced for the duration of the entire shift and not on the basis of station cycle times (as was the case for single model assembly).

Boyer & Leong (1996) develop a model for evaluating the benefit of increasing levels of flexibility. Their context is the automotive sector and the point of interest is the ability for a number of plants to produce more than one product line. Without flexibility, unused capacity in one plant cannot be used to fulfil demand that exceeds the capacity of another. They find that opening up a fraction of the feasible cross-links between products and plants brings substantial gains in overall performance, even with a throughput loss of 20% due to changeovers.

To counter supply chain effects, the Quantity Flexibility (QF) contract has become popular (Tsay & Lovejoy 1999). It attaches a degree of commitment to the forecasts by installing constraints on the buyer’s ability to revise them over time. The extent of revision flexibility is defined in percentages that vary as a function of the number of periods away from delivery. The QF contract formalizes the reality that a single lead time alone is an inadequate representation of many supply relationships, as evinced by the ability of buyers to negotiate quantity changes even within quoted lead times. The model indicates that inventory is a consequence of disparities in flexibility. In particular, inventory is the cost incurred in overcoming the inflexibility of a supplier to meet a customer’s desire for flexible response and they coin the term flexibility amplification. All else being equal, increasing a supply chain participant’s input flexibility reduces its costs, and all else equal, promising more output flexibility comes at the expense of greater inventory costs.

4.1. Fulfilment logic

The issue of how to link orders to products or production capacity is an aspect of Demand Management and rules such as assigning orders the ‘earliest available’ and ‘latest available’ production slot have been examined (e.g. Guerrero 1991). The production seat system (Tamura & Fujita 1995, Tamura et al 1997, Tsubone & Kobayashi 2002) is a demand management system for producers of a variety of complex products in mixed or small lots, developed for the purpose of shortening delivery lead times. It deliberately creates capacity slots of different dimensions in recognition of differences in product manufacturing requirements and the sales team can which slots are free when negotiating with the customer.

In the context of floating decoupling point systems, Brabazon & MacCarthy (2004 b) study how search rules alter the likelihood of finding a match for a customer. Their abstract models identify the ratio of product variety to pipeline length as being an indicator of fulfilment performance of a system.

In their study of using vanilla boxes in the fulfilment process, Swaminathan and Tayur (1998) make the conjecture that it could be cost effective for the producer to supply the customer with a product that has superior grades of component(s) or even includes redundant components that
the customer may not be made aware of, if the consequence of not doing so is to lose the sale. Giving customers substitutions when there are shortages is not a new idea, but the point to take from their conjectures is that these approaches provide an MC enterprise with options in the face of capacity and responsiveness constraints.

5. Customer factors

Differences across customers are, of course, the prime reason for the growth in product variety. However, customer differences can be expected to create other forms of ‘service’ variety within the order fulfilment system, such as variety in lead time and price.

Price and lead-time are interrelated. Price is connected to value (e.g. Meredith et al. 1994) and it is well understood that value tends to decay over time (e.g. Lindsay & Feigenbaum 1984). However, the rate of decay is not uniform across customer groups and for some customers, delivery earlier than an agreed date is undesirable.

Methodologies for exploiting customer differences are now emerging under the banner of yield management (also known as revenue management) and its proponents see many opportunities for exploiting its principles (Marmorstein et al. 2003) but the research in the xTO sector is scarce. Tang & Tang (2002) study time-based policies on pricing and lead-time for a build-to-order and direct sales manufacturer of products whose value is decreasing rapidly, such as is the case with high technology components. Although Chen (2001) is not focusing on high variety systems, his work is relevant. He proposes customers be given the opportunity to select from a menu of price and lead-time combinations, with greater price discounts on offer for longer delivery times. By reducing the proportion of customers who demand immediate fulfilment, his model shows how safety stock/inventory along a supply system can be reduced.

Customers can be expected to differ in their attitudes toward specification compromise as well as to delivery time. In an ATO context Iravani et al. (2003) use simulation to find that customer tolerances to substitutions have an impact on the stock policy of an ATO system. They divide customers into four groups that differ in regard to which components are key and non-key, and which substitutions they are prepared to accept (e.g. m prepared to accept item B instead of A and 1-m lost, and if B is also unavailable n will take C and m-n will also be lost). In their system, customers must get their key items or acceptable substitutions for their key items, but will still purchase if a non-key item is unavailable and cannot be substituted. They use several overlapping measures:

- Fully satisfied – customers getting the exact match for key and non-key components;
- Key satisfied – customer who get all of their key items, but some or all of their non-key items are substituted (note Fully satisfied is a subset of this category);
- Substitution satisfied – customers who accept a substitution for at least one of their key items.

A similar approach to segmenting customers is used by Brabazon & MacCarthy (2005). They assume every customer is seeking a target specification but that customers will treat each feature as being critical or non-critical; a critical feature is one for which the customer must receive their target option; but the customer will tolerate an alternative option for a non-critical feature. The proportion of each customer type is varied, revealing the sensitivity of fulfilment metrics to the mix.
6. Discussion

Diversity in the order fulfilment structures was expected and is reported in the literature. Product variety has been increasing (Cox & Alm, 1998) and mass customization is not a mature operations model, hence diversity in operations models can be expected. What is apparent is that producers are being imaginative in coping with the demands of high variety, customization and short lead times. These demands have encouraged the relationship between product, process and customer to be re-examined. Not only has this strengthened interest in commonality and postponement, but, as shown here, has led to the re-engineering of the order fulfillment process to create models with multiple fixed decoupling points and the floating decoupling point system.

A second observation is that there are many avenues worthy of research. Market conditions and technology are driving the re-engineering of the order fulfillment process but there remains the question as to how these structures and their control logic perform and under what circumstances they offer benefits, in particular where there are customer differences, not just in terms of time and cost trade-offs that are exploited in revenue management but in terms of product features.

References


