

# Achieving Breakthrough Service Delivery Through Dynamic Asset Deployment Strategies

Morris A. Cohen  
Matsushita Professor of Manufacturing and Logistics  
The Wharton School  
University of Pennsylvania  
Founder and Chairman, MCA Solutions Inc.  
[Morris.Cohen@mcasolutions.com](mailto:Morris.Cohen@mcasolutions.com)

Narendra Agrawal  
Associate Professor  
Leavey School of Business  
Santa Clara University  
[nagrawal@scu.edu](mailto:nagrawal@scu.edu)

Vipul Agrawal  
Founder and EVP Products  
MCA Solutions Inc.  
[Vipul.Agrawal@mcasolutions.com](mailto:Vipul.Agrawal@mcasolutions.com)

20 September 2005

Forthcoming Interfaces

## INTRODUCTION

The last decade has witnessed a substantial shift in emphasis on the part of many firms from a focus on the products they produce to a concentration on their customers and the value that their customers derive from ownership and use of these products. This “customer-centric” perspective is correlated with the blurring of the line between products and services. Consider the examples of companies like Hewlett Packard and IBM. Historically, they have competed by positioning themselves as manufacturers of cutting edge, high-technology products. However, recent years have witnessed a shift in their competitive strategy towards providing technology solutions, of which, services is a very significant component. In fact, today, most firms would describe their output as a bundle of goods and services and are seeing significant growth in the proportion of their revenue derived from services. For example, in 2001, maintenance servicing was responsible for over \$5 billion of IBM’s revenues (the recent sale of the PC manufacturing division to the Lenovo group of China presents further evidence of their continued commitment towards a service oriented strategy and away from a product based competitive strategy). A recent survey by AMR, (Bijesse et al [2002]), puts this percentage at 24% with many traditional manufacturing firms at over 50% and climbing, as reported in a recent Wharton-Stanford conference on the after-sales service industry, (see <http://opim.wharton.upenn.edu/fd/forum/>). In fact, after sales service is recognized as an important source of revenue and profit, customer acquisition and retention, and competitive differentiation. Not surprisingly, therefore, companies are beginning to direct their focus on their *service supply chains*, which we define as the network of resources that includes the appropriate material (service parts), people (customer engineers, call center staff, repair depot staff, warehouse and transportation staff) and infrastructure, (for material movement and storage, repair, transportation, information systems, and communication). Unfortunately, the management of this service supply chain remains a challenge. This is surprising since it may be reasonable to expect that lessons learned from supply chain management for manufactured goods are directly applicable to the world of service delivery. However, the mechanisms required for designing, producing and delivering post-sales service in a cost-effective and competitive manner are quite different than those used to manufacture goods and to procure direct materials. It is, therefore, not surprising that significant assets must be dedicated towards service support.

Consider the case of Cisco Systems, a San Jose, California, based leading high technology manufacturer of computer, data storage, communication and related equipment. Cisco has long recognized the strategic importance of after sales service as a core element of their overall growth strategy. With revenues of about \$3.5 Billion in 2004, and gross margins of nearly 65%, the Global Product Services (GPS) division is expected to grow by about 13% next year (source: Cisco Systems Financial Statements, [www.cisco.com](http://www.cisco.com)).

To address the wide range of customer support requirements Cisco offers a gamut of services such as troubleshooting, hardware and software support, systems monitoring and management. It does so through warranty services, requiring only parts as well as service contracts, which require parts as well as field engineers. Examples of service contracts range from 2-hour to 10 day delivery for return-to-factory services. To deliver services for the resulting contracts, which number in the millions, Cisco partners with a multitude of external partners for its repair, manufacturing and logistics needs. Its service supply chain consists of a world wide network of field engineers, about 800 fulfillment depots, 18 repair centers, and 5 material return processing centers, which collectively deliver approximately 720,000 parts and repair approximately 450,000 parts annually.

The complexity of Cisco’s service environment has resulted in:

- Less than desirable customer service levels
- Inability to respond to frequent changes in the installed base and customer entitlements
- High investments in extremely slow moving service assets
- Complex business processes
- Need for significant manual oversight, and, ultimately,
- Lack of regional and global coordination.

Consequently, the challenge facing Cisco's service support function is to manage their resources and business processes in a manner that effectively supports this differentiated portfolio of customer support entitlements within its complex global network. While pressures to control operating costs and capital investments continue, customer expectations for service are rising.

Situations such as the one described above are not uncommon in many other industries that require the use of capital intensive assets in mission critical environments – e.g., high technology, aerospace and defense, telecommunications, automotive, etc. In all of these environments, the problem faced by the service organizations can be articulated in the form of the following two critical questions, which are the focus of this paper:

- a) Service Asset Management: How should service supply chain resources be optimally positioned and managed to support delivery of after-sales service?
- b) Service Demand Fulfillment: What is the most cost and service effective way to deliver such support?

As we shall see, it is especially difficult to answer these management questions for after-sales service supply chains as a consequence of their high level of complexity and risk. A key conclusion of our analysis will be that traditional modes of thinking, which are inspired by manufacturing and finished product distribution thinking (e.g., ERP and DRP), and which attempt to match service supply to demand by assigning enabling resources to specific service products in a static and separable fashion, are inefficient and ineffective. The essential requirement for competitive success in delivering a service-centric strategy is flexibility. Such flexibility must be based on a deep understanding of the mechanisms that come into play in a service supply chain to fulfill customer demands for service which generate demands for support assets and capacities. We shall introduce a collection of management policies that promote this flexibility and refer to it as Dynamic Asset Deployment (DAD). As compared to the static resource management policies utilized by many companies, we propose a hierarchical approach to asset management and deployment, in which different sets of strategic and tactical decisions are made in a dynamic fashion based on the most recently updated forecasts of upcoming supply and demand conditions. At each step, estimates of future uncertainties are used to position assets, also defined in a hierarchical manner, such that when service demand does occur, it can be satisfied in a cost effective and flexible manner. In this sense, our approach relies on a real options based view of resource management. We shall propose a framework that can be used to design such policies and describe the most significant capabilities that a firm must develop to compete profitably through them to effectively deliver services. Our research and experience in implementing such strategies at a number of companies in diverse industries leaves no doubt that firms that do not adopt such dynamic strategies are destined to mediocrity when it comes to competing on the basis of customer service. We will illustrate this potential by reviewing the results at Cisco and at other leading companies that have adopted the DAD framework.

## **SERVICE ASSET MANAGEMENT**

The key to understanding dynamic asset management and using it to support successful customer-centric strategies is the fact that after-sales service products purchased by customers are equivalent to “entitlements” for response to a support need within a specified time limit, at a given level of reliability and for a given price. In fact, a goal of a firm's service supply chain is to maximize the benefit their customers derive from ownership and use of the products they have purchased. Such service products cannot be produced in advance of their demand and thus they cannot be stored on the shelf. Rather they are produced at the time of consumption, which is typically triggered by a contingency, or a random “service event,” such as the failure of an installed product in the field. Such events are very difficult to predict. Fulfillment of service product demands, however, is enabled by physical assets such as spare parts inventory, repair depot capacity and field engineer contact hours. These assets must be deployed in advance of the occurrence of a service event if the response time entitlement standard is to be met, (e.g., responding within minutes or hours). These assets are utilized or consumed as the service response is produced to meet the specific requirements of a given service event.

Such asset deployment decisions must be made in a manner that allows firms to maximize the profit that it derives, both directly and indirectly, through provision of a portfolio of diverse service products to a heterogeneous group of customers with a geographically distributed installed equipment base which requires different levels of support. These policies must be designed in a manner that takes advantage of the unique inter-relationships between resources, decisions and information that characterize the service supply chain environment. This requires companies to develop a detailed understanding of 3 key issues:

- Key tradeoffs associated with service delivery.
- Strategic options for deploying various service delivery assets.
- Effect of asset deployment decisions and fulfillment policies on service delivery tradeoffs.

Each issue is discussed in detail in the following paragraphs.

## Tradeoffs for Service Delivery

The principal tradeoff facing service supply chain managers is between **revenue**, **cost** and **service performance**. Let us consider each of these metrics in turn before discussing tradeoffs among them and their implications for asset management.

**Service revenue** is derived from the sale of service “products” designed to support customers’ requirements for uninterrupted use of their tangible products. Such revenue can be captured through performance based service contracts and/or the direct sale of time (customer engineer, repair depot, etc.) and material (replacement parts). Many firms also address their customers’ expectation for service through provision of product warranties that act as a pre-paid service contract operating for a fixed period of time at the beginning of the ownership cycle. As a consequence, the revenue model for service should include the initial sale price of the product, the terms of the warranty, the design and price of the service support contracts sold after the warranty expires, the prices for non-contract sale of time and material, and the impact of customer satisfaction derived from delivery of service on repeat buy behavior. Many firms find it advantageous to reduce the initial product sale price in order to capture the long-term revenue stream associated with delivering service (i.e., the famous “razor blade” strategy where the product is distributed free of charge in order to capture the revenue stream of support products from a captive user market)<sup>1</sup>. Others focus on the design of their products and infrastructure in order to promote serviceability and to reduce the likelihood that customers will seek other sources of service in the after-market.<sup>2</sup> In all cases, it is clear that the revenue a service provider can expect is a function of the specific service products designed and delivered to its customers.

**Service costs** are driven by the wide range of long and short-term decisions that service supply chain managers make concerning deployment and control of resources. Operationally, these decisions typically are made in a manner that meets specific service targets while satisfying a variety of budget and resource constraints.<sup>3</sup> Cost components include the direct costs of matching demand for service resources. These include the cost of parts consumption, fault diagnosis, material handling, transportation, and repair. There are also a variety of capital and

---

<sup>1</sup> See “Profiting from Spare Parts” by Gallagher et al. [2005] for a description of the profit opportunities associated with the sale of service parts to support customer service.

<sup>2</sup> See Cohen and Whang [1997] for an economic analysis of the strategic tradeoff between product price and service quality. Cohen et al [2000] describes the celebrated case of Saturn and the automobile after-market. The Saturn case illustrates how high quality service drives revenue through retention of market share of after-market demand for post-warranty services as well as vehicle repeat buy behavior.

<sup>3</sup> Service targets are required since the computation of actual shortage, downtime and delay costs to the service provider is not feasible. See Cohen and Pierskalla [1979] and Cohen et al [2003] for empirical studies drawn from the blood distribution and semiconductor equipment industries, respectively, that demonstrate that managers typically under-estimate such costs.

fixed costs associated with the service business process. The most visible of these are the investment tied up in parts inventory and the fixed cost of warehousing and repair facilities. There are also considerable infrastructure, training and human resource fixed costs, which are associated with the service support function.

**Service performance** is directly related to product availability or “up-time” and is the best surrogate for shortage costs. From a customer perspective, any measure of service should be directly related to the time delay between recognition of need for support service (e.g., an unscheduled maintenance due to product failure in the field) to the time of restoration of their product to its full operating condition in the field. From a service supply chain provider perspective, there are a variety of measures associated with availability of resources required to meet such customer needs. The principal internal metric is *part fill rate*, i.e., the fraction of demand for parts that is fulfilled by the stock available at the site receiving the demand. The primary measure used to capture the customer’s perception of service quality is related to *product availability*, which is directly related to the delay times associated with matching supply with demand throughout the service supply chain. Both classes of service metrics (fill rate based and availability based) are functions of the underlying risk structure of future demand and the positioning and management of resources throughout the service supply chain. As we shall see, the specifics of how one computes such metrics and uses them will have a major impact on the design and implementation of effective decision support systems.

The tradeoff between these metrics can be best expressed using the concept of an *efficient frontier curve*. Recall that a wide range of resources are required to fulfill service demands in a manner consistent with the range of performance entitlements associated with the purchased service product(s). However, since the underlying demand processes are fundamentally random, it is critical to note that the timing, location, extent and consequences of a service demand cannot be forecast with certainty. Since the response times specified by the service product contracts typically are much shorter than the lead times for moving or acquiring material, and the timing of the demands is uncertain, it is necessary to position many of these assets prior to the fulfillment of the service demand, i.e., prior to the occurrence of the “service event”.<sup>4</sup> In general, the greater the promised level of service performance, the larger the required investment in such assets, which increases the total costs incurred by the service provider. For example, if no (or very little) inventory of parts is deployed, the financial cost of these assets is small. However, the response time to address each service event will be long since the required parts will have to be ordered from a backup location, or repaired. While the exact calculation of the relationship between service performance and costs is non-trivial, it can, nonetheless be performed using advanced analytical methodologies<sup>5</sup>. The nature of this relationship can be depicted using the *efficient frontier* curve as shown in Figure 1. Note that the curve rises steeply: the costs increases disproportionately as the promised service performance level increases. Consequently, the revenue model must incorporate this relationship.

The primary service delivery challenge can now be more precisely stated as determining a way to deliver service products that meet pre-determined commitments for performance (in terms of cost and speed) in the most effective manner *as defined by the cost/service efficient frontier*. To be inside the frontier suggests inefficiency. To be on the frontier suggests that increasing the level of service for a service product can only be achieved at a higher cost. Balancing the tradeoffs among revenue, cost and service is challenging because of escalating service expectations, complexity of the service supply chain, and, as mentioned before, the high degree of uncertainty associated with service events. The DAD approach we are proposing, in fact, helps companies not only to negotiate their way on to the efficient frontier, but also pushes the efficient frontier downwards, as shown in Figure 1. This is precisely what allows firms to achieve quantum improvements in their service delivery performance and profitability.

---

<sup>4</sup> In a recent meeting, a major semiconductor fab operator mandated its equipment suppliers to make parts required to repair a down machine on a fab line available within 15 minutes. The justification for this requirement was that having a line down costs the fab operator about \$1 million per day.

<sup>5</sup> See Cohen et al [1990] and Feeney and Sherbrook [1966] for two classic multi-echelon model formulations for high technology and aerospace and defense respectively.

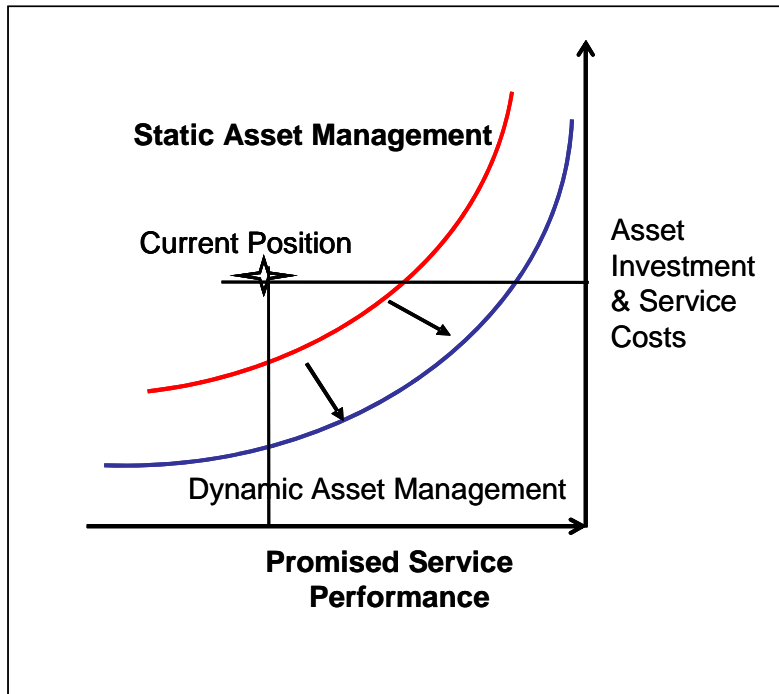


Figure 1. Key Tradeoffs for the Service Provider

### Service Asset Deployment Options

In order to better understand how the revenue/cost/performance tradeoff noted above manifests itself in the service supply chain, we introduce a framework that considers a two dimensional hierarchy, based on the supply chain geography (*geographical hierarchy*) and the underlying product design architecture (*product hierarchy*). This joint geography/product hierarchy can be used to describe how the necessary resources are deployed and managed. In particular, for the particular case of after-sales service we will answer the question “where” (to deploy) in terms of the geographical hierarchy and “what” (to deploy) in terms of the product hierarchy.

The **product hierarchy** can be described by breaking down the product into major modules, sub-modules and parts (see Figure 2). In the network representation shown, each node represents a particular item (finished product, module, sub-module, etc.) and the item membership specifics are captured by the arcs connecting an item to its parent(s) (the “goes into” items) and to its children (the “includes” items). The information required to specify a product hierarchy typically is encoded in the Bill of Material (BOM), which describes the recipe for producing the product. The “Service BOM” represents the current hierarchical structure used to maintain the product as it is configured in the field and therefore incorporates engineering changes and part supersession. Note that a Service BOM is not necessarily the same as the production BOM which is used for manufacturing functions. For example, it is not uncommon for the product, as configured in the field, to be quite different from the version assembled in the factory, in terms of parts installed at various levels in the Bill of Material. Also, the usage rate of a part in the field for support purposes due to repair/service/upgrade actions may be different than the usage rate associated with assembly of the product during the production phase. Consider the example shown in Table 1, which indicates that in the repair of a router, the expected number of mother boards consumed is 0.1 (i.e. a replacement rate of 0.1). Similarly in the repair of the mother board, part # REA-048, which occurs 3 times in the assembly is replaced at a rate of 1.6 units per repair. Note that for certain parts whose replacement rate is zero, we can conclude that the part

has never been observed to be replaced (at this location, for this customer, etc.), or by design it is not replaced and instead a higher indenture level component which contains that part is replaced. Finally we note that it is not uncommon for one version of a part to be superceded by a more recent (re-designed) version, as a result of an engineering change order. As a result different versions of different parts will be observed to be present in the product as it is used in the field.

INDENTURE	PART NUMBERS/NAMES		BOM-QTY	VERSION	REPLACEMENT RATE
	FRU	SRU			
1	Router			1	
2		Mother Board	1	1	0.1
3			REA-048	3	2
2		Power Supply	1	1	0.1
3			145D0111-3	1	2
3			REA-048	3	1
2		Comm. Unit	1	1	0
3			RLA27SM4-00	2	1
3			PLRA6560010	3	1
3			PLRA6560010	2	2
3			MS51989-105-10	1	1
3			145D0111-3	8	3
2		Interface Unit	4	1	0
3			145D0121-3	1	1
2		Display Unit	1	1	1
3			145DS011-10	6	1
3			145D0101-5	1	1

**Table 1: Service Bill of Material**

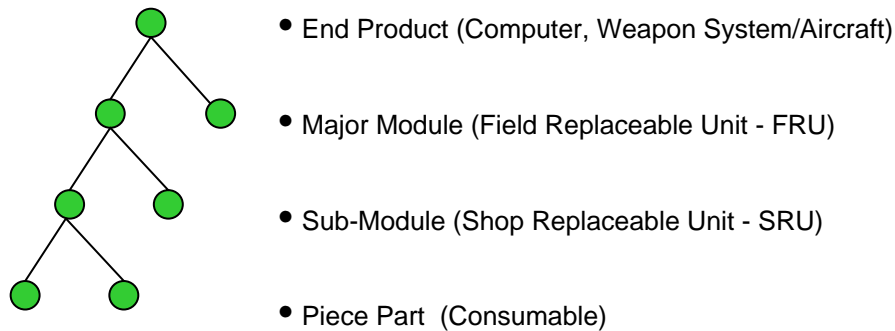


Figure 2. Product Hierarchy

A **geographical hierarchy** organizes stocking locations (nodes) by echelon to capture material and information flows associated with demand fulfillment in the service supply chain, (see Figure 3). At the lowest echelon, we position the individual forward locations, which could also include customer sites. At the “top” of the hierarchy we include the central stocking sites that act as emergency backup and/or replenishment sites for the downstream child locations. In between, the network might include additional field or regional stocking locations. Both emergency (customer) demand and replenishment demand flow “up” through the spatial hierarchy. The availability of material to meet these demands is determined by the inventory stocking policy for each part/location combination. This deployment of resources, along with the rules and procedures for matching supply with demand, drives the performance metrics related to lead times and customer service. Reverse material flows also occur in the spatial hierarchy to support repair of failed items. Figure 3 illustrates the basic elements of the material flow pattern found in a typical service supply chain.

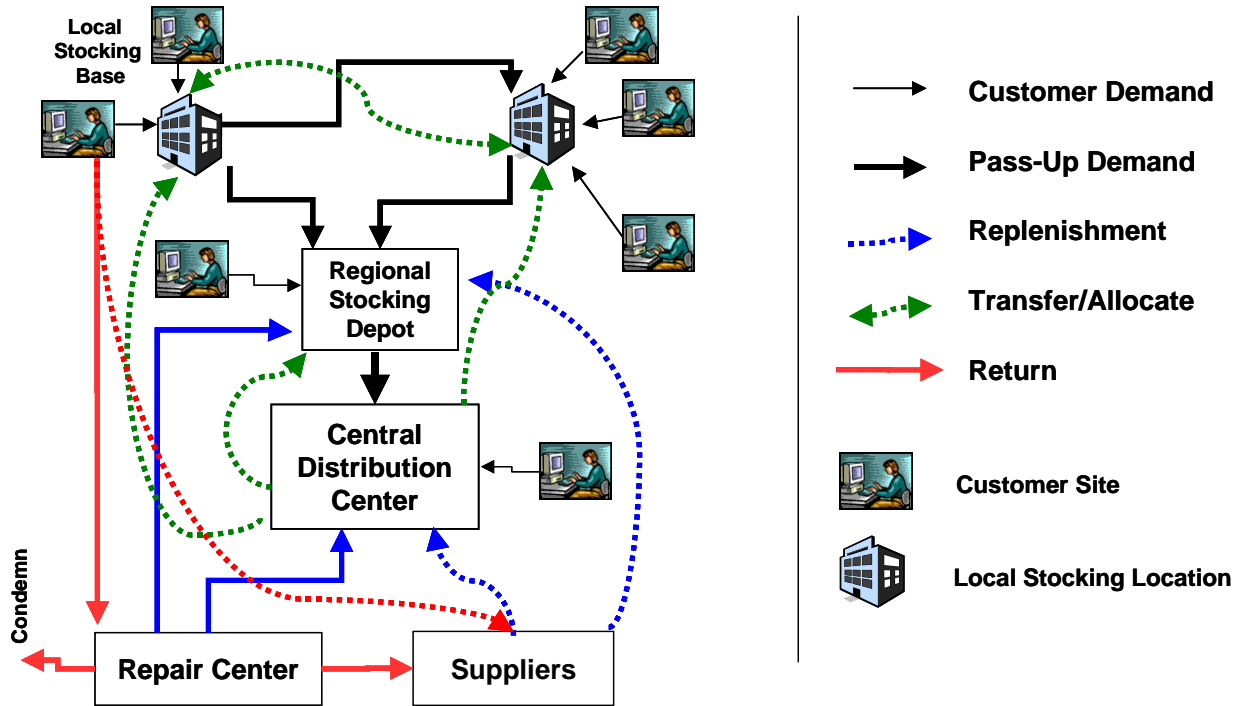


Figure 3. Multi-Echelon Service Supply Chain Material Flows

The interplay between these two hierarchies essentially describes a firm’s service supply chain asset management strategy. Consider Figure 4, which illustrates how these hierarchies can interact. The fastest way to meet customer response time targets is to replace failed products with spare (standby) capacity units that are positioned at the customers’ forward field locations. This is, of course the most expensive way to meet the need to restore a customer. However, depending upon the criticality of these products to the customer’s mission, this may in fact be the appropriate strategy. For example, the routers produced by Cisco Systems for use in computer servers that support key financial transactions at a major financial institution incur high cost for the customer in case of failure. Consequently, a standby spare may be located very close to the customer’s location (perhaps at the customer’s location) so that downtime is minimal. Of course, the customer may be charged a premium for such service. On the other hand, the most economical way to meet the service demand is to identify and replace the specific components of the product that have failed and to do so at a central location. This can require time for extensive diagnosis and material movement. Repair in this case involves replacement or refurbishment of the failed components. This approach will of course be the slowest. A defect in specific components on a router circuit board for example could cause the router to fail. Replacement of these components, however, will be much cheaper than replacing an entire board assembly or the router itself.

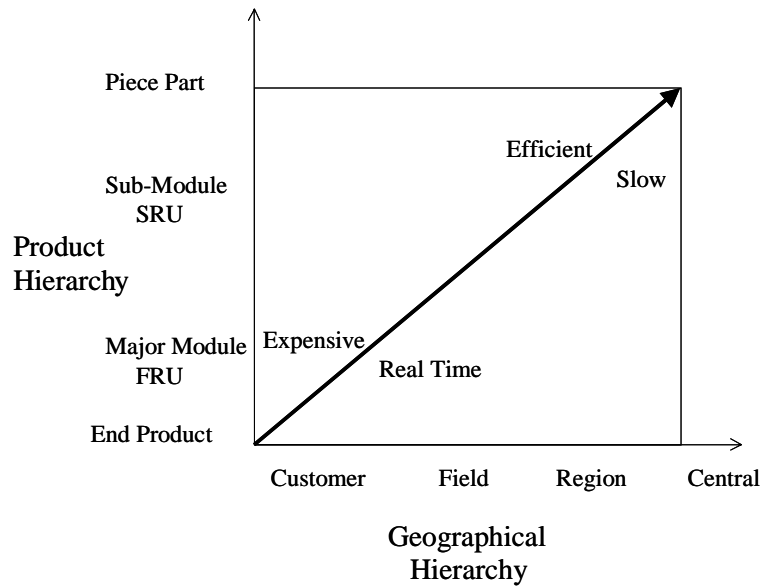


Figure 4. Tradeoffs Across the Geographical and Product Hierarchies

### Service Asset Decisions

The joint geography/product hierarchies of Figure 4 provide a framework for understanding choices that a firm can make for positioning and managing the resources required to fulfill service demand entitlements. The geography hierarchy is concerned with the question of “where” such resources should be deployed. The product hierarchy focuses on the question of “what” should be deployed. These decisions must be made for all resource/location combinations. In typical environments, such as Cisco’s, there are millions of such decisions to be made. It is important to note that these deployment decisions are highly inter-related in the sense that an investment in a resource at one location can and will influence investment decisions for many other item/location combinations that are connected via the joint hierarchy structure. For example, positioning resources of Field Replaceable Units (FRU’s) at forward locations can decrease the emergency demands for piece parts that are experienced at higher echelon locations. Similarly, investing in additional safety stock at a central depot reduces the effective lead-time for replenishment at the “child” locations connected to it. This lead-time reduction will, in turn, affect the stocking requirements at the child locations. Alternatively, such decisions are often constrained by the budgets allocated to the service organization. Consequently, if a particular asset is assigned to a specific location, it affects what can be assigned to other locations. Thus, when the available budget for asset investment is fixed, the service levels that can be offered to customers at various locations are interrelated; a high level of service to one customer may imply a lower level of service to another.

In addition to the periodic deployment of assets, service providers must also manage the flow of such assets over time to replenish stock levels and to adjust deployments in anticipation of future service event requirements. This includes tactical decisions concerning replenishment (purchase and repair) of assets consumed and re-deployment of assets within the service supply chain (allocation of incoming assets, transshipment of excess assets, etc.). Optimization of the full set of asset management decisions noted here also must include consideration of factors such as budget, cash flow and service constraints. Finally, we have observed that management of service operations requires planning with respect to a variety of performance objectives, including maximization of product uptime, minimization of total system costs and minimization of cash flow requirements associated with repair and purchasing.

This service asset management problem is complicated further by the fact that many of the drivers of resource requirements (e.g., product utilization rates and deployments, evolving part failure rates, etc.) are non-stationary, i.e., they change over time. Thus the third dimension of resource deployment is concerned with the answer to the “when” question. Finally, as we have already noted, demand for the enabling assets are triggered by service events that are highly uncertain. As a consequence, solving the overall asset management problem requires a probabilistic, dynamic representation of its environment.

Given the complexity of the asset management problem, it is necessary to decompose it into a collection of inter-related decision problems. Figure 5 illustrates a *chronological planning hierarchy* that mirrors the decomposition of managerial decision making that we have observed in many service supply chain environments. Each component corresponds to a different length of time, over which managerial tradeoffs and objectives must be considered as the relevant decisions are made (i.e., planning horizon). At the longest planning horizon we have *Service Business Design*, where decisions that determine specification of the overall service strategy are made. Although not the focus of our discussion in this paper, such decisions can include design of the products being supported, the design of the “service products” that are offered to customers in the after sales market, and the design of the infrastructure used to deliver these service products. The planning horizon for this set of decisions is typically measured in months or years. As we proceed to the next level of decision making, *Strategic Asset Positioning*, the length of the planning horizon decreases. At this level, management is concerned with the strategic positioning of its material, human and knowledge resources, in anticipation of the need to meet customer service demands in a manner consistent with the response and cost entitlements as set out in the warranty and service agreements in force. These strategic resource deployment decisions give rise to a challenging optimization problem that must be solved periodically (monthly, weekly), if the DAD strategy is to be implemented in a cost effective manner. At a nearer-in planning horizon (weekly, daily), we group the re-deployment decisions, *Tactical Asset Deployment*, that are associated with re-positioning resources, usually within relevant lead times. At this level we consider material flow decisions such as new buys, repairs, replenishment, allocation and transshipment. Finally, of course, are the decisions that must be taken after the realization of the service event which lead to fulfillment of the service demand. These decisions, grouped as *Service Demand Fulfillment*, are triggered by actual product failures, and other potential supply chain disruptions. These activities may also be triggered by planned preventive maintenance. Ultimately, customer satisfaction and operating profitability are determined by the efficiency with which this last step is performed. A more detailed discussion of service demand fulfillment is presented in the next section.

In practice, the frequency with which each step in the above hierarchy is performed varies from one organization to another. Obviously, this depends upon the firm’s competitive strategy. It is also a function of the specific capabilities resident within the organization. As an example, at Cisco Systems, the strategic asset positioning decisions are made on a daily basis, while KLA-Tencor, a semiconductor equipment manufacturer, reviews them monthly for their global networks. In each case, strategic asset decisions are reviewed in the tactical asset deployment step on a more frequent basis, e.g. every shift or daily. This capability for tactical adjustment is supported by a high level of visibility into changes in various supply and demand conditions on an almost real time basis.

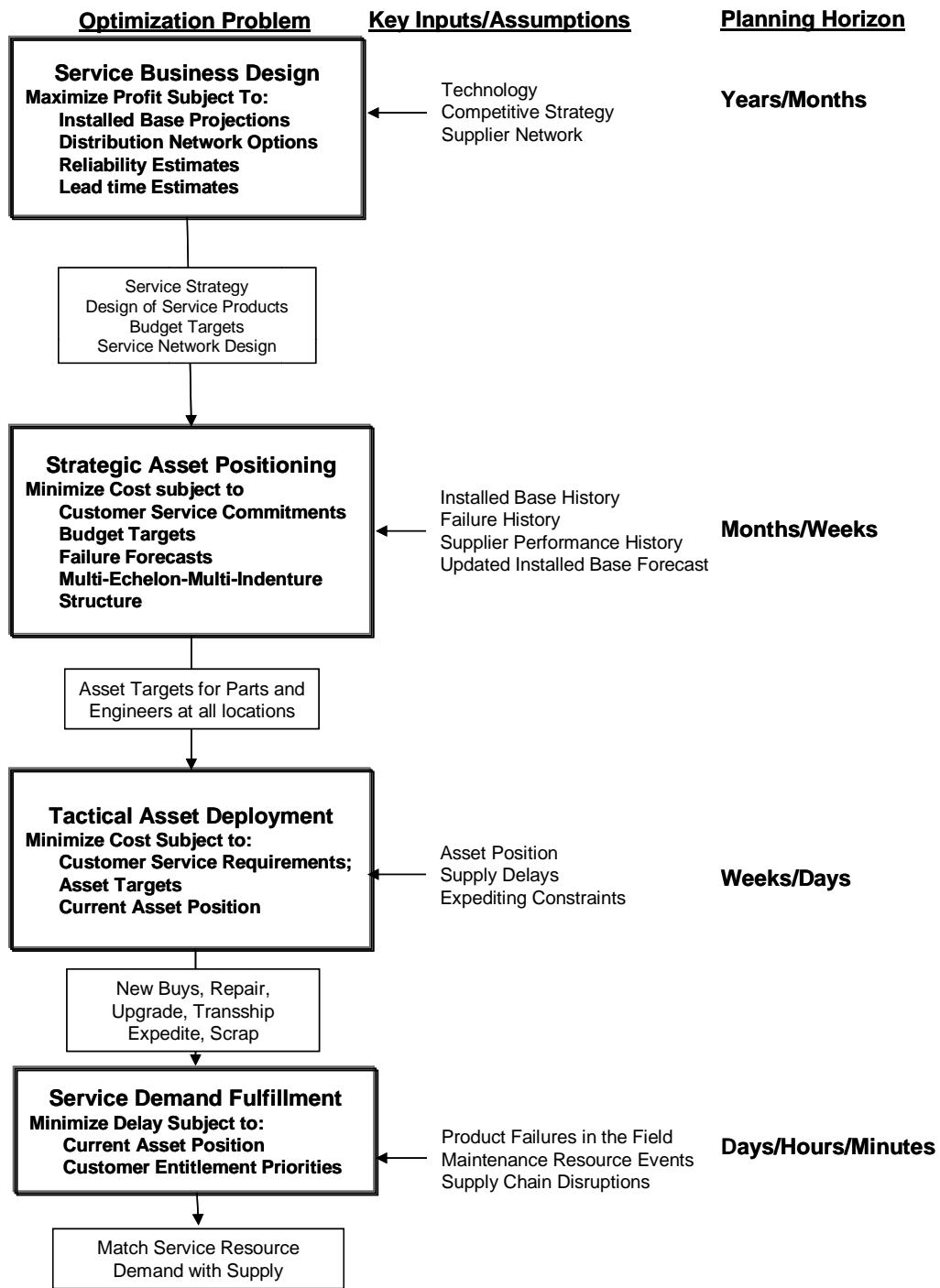


Figure 5. Chronological Planning Hierarchy

It is important to note that all of the resource decisions described in the Service Business Design, Strategic Asset Positioning and Tactical Asset Deployment phases must be made prior to the occurrence of a particular service event whose fulfillment will require use of those resources. Clearly, the efficiency of the actual fulfillment will depend upon how well the service supply chain resources and assets have been positioned in anticipation. This is analogous to the purchase of a real option; a decision made prior to the exercise of that option. Contingencies that determine how and where the option to deliver service is exercised include occurrence of events, such as a product failure, or a maintenance event that cannot be predicted with certainty. Hence the decisions included in the Service Business Design, Strategic Asset Positioning and Tactical Asset Deployment phases are made on the basis of estimates of future resource requirements along with visibility of all of the events that affect supply and demand of such resources that have occurred throughout the service supply chain prior to the occurrence of the service event in question. In the DAD approach, we maintain that demand uncertainty cannot be eliminated through forecasting, and hence, trade-offs must be evaluated on the basis of future risk assessments captured by estimates of the demand probability distribution relevant to specific customer products and locations at particular future points in time. The decisions made at all pre-event planning levels thus constitute an exercise in risk management.

## **SERVICE DEMAND FULFILLMENT**

The “last mile” of decision making in the planning horizon hierarchy concerns fulfilling the service demands. We can view this as a problem of event management which is distinguished from all other planning/management problems discussed so far by the fact that it is concerned with decisions that are made after service event based demands for resources have been realized. At this level, managers control the real-time reaction of the service supply chain to shortages and excesses. Such responses ultimately determine how the service support process meets the strategic goals of customers. This is where the service product is actually “produced.” Intelligent decision making here can improve the performance of the system by allowing managers to make the best use of current and projected resource deployments throughout the service supply chain. These choices will, in turn, act to mitigate the risks of service fulfillment delay through a mismatch in resource supply and demand.

### **Actions to Fulfill Service Product Demands**

When the supply of the enabling assets is sufficient to meet the demands generated by service events, the system can fulfill demands from available resources. When demand exceeds supply at a location, however, managers of the fulfillment process have a wide range of actions that are available to them as they react, in real time, to the occurrence of a resource shortage. The following is a summary of some of the principle actions that an asset manager can consider in this situation:

- *Demand transfer within the geographical hierarchy through vertical and lateral pooling:* When material is not available at a stocking location the demand is transferred to an alternative location that acts as a backup site. While this can decrease the delay in responding to demand at forward locations, and is a way to enable risk pooling, it increases the time required to fulfill the demand at the alternate site that served as the backup location.
- *Demand delay for lower priority customers:* Lower priority customers can be made to wait, even when there is supply of assets, as a way to ration scarce resources and ensure availability for higher priority customers. If such policies are not implemented, lower priority customers get to free-ride off of the stock maintained for higher priority customers. The waiting customers can then be served when the stocking location receives a shipment from a supply source (either as a new buy replenishment, internal transshipment of excess or receipt of a repaired item from a depot).
- *Demand substitution from one product to another:* The repair or restoration of a customer’s failed end product can be achieved in multiple ways. As noted, the fastest way is to use Field Replaceable Units

(FRU's) at the most forward location. In the event that required resources and/or repair capacity is unavailable, the requirement for repair can then be based on diagnosis and repair of an item at a lower level in the multi-indentured bill of material product hierarchy. Often, this transference of demand to a different item in the product hierarchy is associated with a shift to a more central and distant location that has the required capabilities and material. Depending upon the priority of the customer or the service event, the substitution can be made in either direction in the product hierarchy. An alternate approach is to substitute a superior item (this is different from substitution within the product hierarchy). If an alternate part that can meet the requirements of the demanded part is available, then the alternative can be issued instead of the item that was originally requested. Typically the substituted item is more expensive and of superior performance quality (e.g., use of a 60 GB disk drive to replace a failed 20 GB disk drive). Clearly, substitution enables the demand to be met quickly, albeit at a higher cost.

- *Dynamic pricing and incentives* to modify demands, which have caused the shortage: Use real time incentives and penalties to match supply and demand, e.g., provide a side payment to customers willing to wait for service fulfillment. This is analogous to overbooking management for an airline flight. This approach can align incentives to promote first-best decisions that act to optimize total supply chain performance. However, the contractual terms may be difficult to verify and enforce.

The heart of the DAD philosophy is not simply figuring out a way to implement such actions – for these actions have been observed at many companies – but to develop asset and resource management strategies that are based on the assumption that such contingent actions will be taken. In other words, it is not enough for the service supply chain to be able to react to mismatches between supply and demand; one must *plan* to be responsive and manage assets accordingly. This is precisely what sets our DAD philosophy apart from static policies commonly observed at many companies.

For example, consider the case where two customers, who are served from the same forward location, have purchased service contracts that require the use of the same spare part. To avoid potential conflict in case of part shortages, the service provider may simply maintain two separate stockpiles of the same part, one for each customer. The inventory policy for each stockpile may be commensurate with the service level agreed to in the contract. While this approach is very easy to implement, and avoids potential conflict, it can be shown that it is much more cost effective for the service provider to realize risk pooling gains by combining the two stockpiles and serving both customers out of the same pool of inventory, (see Cohen, Deshpande and Donahue [2003] and Deshpande, Donahue and Cohen [2003] for an analysis of this problem in the context of military spare parts procurement at the U.S. Defense Logistics Agency). The problem, of course, is that unless managed appropriately, the customer who has purchased the “cheaper” service contract gets to free-ride off of the higher service level contract<sup>6</sup>. This is precisely where appropriately designed rules to ration inventory can avoid such problems while simultaneously accessing the benefits due to risk pooling. The optimization methodology we describe later on explicitly plans for this possibility while determining optimal asset deployment strategies.

Similarly, consider the case where the service delivery system allows for the substitution of a more expensive or higher quality part for a cheaper one as a way to meet supply demand mismatches. Again, while many companies may be practicing such demand substitution in an ad hoc manner, it can be shown that asset deployment decisions resulting under the DAD methodology, which anticipates the possibility of demand substitution, are significantly more cost effective. In fact, this idea has been practiced for a long time in the context of yield management in the hospitality, airline and rental car<sup>7</sup> industries. The DAD strategy attempts to implement similar approaches in the context of service delivery.

---

<sup>6</sup> For example, if the available inventory is allocated to customers following a first-come-first-serve rule, in case of shortage situations, the low service customer can get access to the inventory that should have been reserved for the high service customer, simply because their demand occurred first. This denies service to the high service customer.

<sup>7</sup> Recall the offer made to you for an *inexpensive* upgrade to a larger sized car, or a better hotel room upon checking in. Typically, upon refusal by the customer, the upgrade is offered gratis. This is done because the smaller car, or

Thus, if post-event actions are not linked to pre-event asset management decisions, when service events do occur, the resources needed to meet service product entitlement targets are not at the right place at the right time. Many adjustments have to be made in order to support the customer's current needs for support. This in turn leads to high levels of emergency transport, expediting and shortage penalty costs.

On the other hand, recourse-based asset planning, which explicitly accounts for the uncertainty associated with service events and the value of contingency actions, is fundamental to the DAD strategy. Optimization of asset management in the context of a DAD process leads to more robust deployments and more cost-effective ways of delivering the service promised to customers. In this sense, our approach greatly resembles the well-known real-options framework that we alluded to earlier. Our observations from the field confirm that the dominant mode of thinking prevalent in most companies is based on static, deterministic forecasts that are more appropriate in finished goods DRP or MRP environments.

The DAD strategy differs from current practices in other important ways as well. For instance, an important factor that should influence asset management decision-making is the wide range in attributes of the items being managed, e.g., unit cost, expected demand rate and lead times. A common approach to dealing with this issue is to create item classification groups and to use group membership as the driver of rule-based asset deployment decisions. The approach that we recommend as a part of the DAD strategy involves the development of a solution to a constrained optimization problem whose formulation captures the tradeoffs, interactions and constraints that have been noted as being relevant to service support. In a subsequent section we will describe how a state-of-the-art software system has been designed and implemented to generate solutions to variants of the problem that are consistent with the DAD framework.

Other traditional approaches to solving the asset management problem are based on myopic heuristics, such as looking at only one product, one customer or one location at a time. In some cases the problem is decomposed into separable segments such as a central depot problem or a field location level problem that are solved in an independent manner. In many instances the product's service bill of material is ignored. The multiple (geography and product) hierarchies introduced suggest that there is a high degree of interaction across locations and items and hence managing assets in a way that ignores such interdependencies results in lower service and higher costs. The DAD methodology we have developed avoids such problems by explicitly considering such interactions in a comprehensive manner. In fact, understanding these interactions is the key to enable risk pooling across the product and geographical hierarchies, which is critical to realize efficiencies in the service supply chain. This issue represents a fundamental point of difference between supply chains for finished goods and service supply chains. In the former case, separate supply chain structures might be best suited for different products. For example, as noted in Fisher (1997), short life cycle products are best served by responsive supply chains while long life cycle products require physically efficient supply chains. However, in service supply chains, since the same underlying asset may be needed to satisfy demand for multiple customers, it is most effective to engineer as much commonality of resources and assets as possible. In other words, the use of a single service supply chain, with appropriate allocation rules, is more cost effective in delivering differentiated services than having multiple supply chains, each targeted for a different customer type.

We can summarize the distinction between static and dynamic (DAD) management practices by referring to Figure 6. As companies move to more advanced and comprehensive policies for asset management decision making, they have the opportunity to enhance the performance of their service demand fulfillment processes by improving the quality of decision making at both the strategic and tactical levels. At the Strategy level (denoted on the x-axis in Figure 6) we identify static methods that include specification of target service levels based on simple part classification rules or single location/single item models. The dynamic methods for Strategy incorporate tradeoffs across either multiple locations or indenture levels, where the most advanced approaches consider simultaneous interactions across both geography and product hierarchies. Tactical asset decision making, (denoted on the y-axis in Figure 6), can be broken down into two categories. At the basic/static level firms make unplanned expediting

---

the lower rate room, was unavailable, and the company had strategically planned to offer the better option as a substitute.

decisions to shift resources in a reactive manner, based on MRP/DRP logic, (i.e. using point forecasts and stationary forecasts). At the optimal/dynamic level, tactical decisions are based on assessment of cost and risk tradeoffs that respond to changing conditions in the environment.

If these two levels of asset management are optimized and effectively integrated, the firm will position assets based on long term strategic planning goals in a manner that meets service performance objectives at minimum cost or inventory investment. Firms will then be able to generate maximum service from their service resource investments as long as their asset re-deployment decisions are made in a manner that is consistent with the risk based forecasts and decisions generated at the strategic level. The result is a position in the upper right (Dynamic) quadrant. We have found, however, that most firms are positioned in the bottom left (Static) quadrant because they do not use appropriate methods at both the strategic and tactical levels or adequately integrate these levels of decision making.

A firm positioned at the top left quadrant will try to rebalance inventory to get the maximum service value, but given the lack of strategic optimization, they will often find that their asset investments are in the wrong items at the wrong locations. Very few companies find themselves in this position. Those that do attempt to manage tactical re-deployment in a risk based manner with sub-optimal strategic planning asset positioning (typically based on the use of basic planning methods), will achieve limited service benefits and/or experience high costs. Firms located in the bottom right quadrant do not adequately integrate strategic and tactical service asset management decision making. Such firms may position the right items at the right place, but will fail to reposition those assets in the short run based on recognition of short-term demand changes and anticipated supply shortages.

There is a natural progression from the lower left to the upper right quadrants as firms move from static asset management processes to DAD processes. As we shall see, companies who have done so have improved service, lowered investment requirements and lowered operating costs.

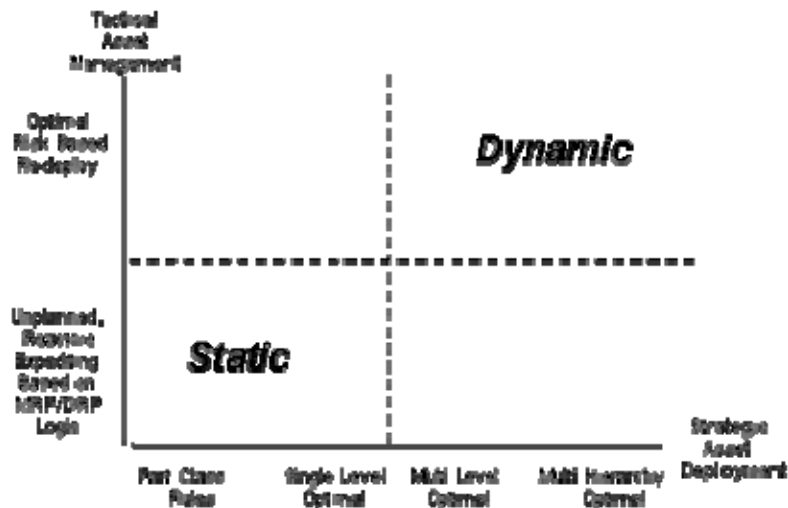


Figure 6. Service Asset Management

**PRACTICAL IMPLICATIONS OF DYNAMIC ASSET DEPLOYMENT:**

While conceptualization of a customer-centric service strategy is a necessary step, it is not sufficient for actualization of breakthrough service delivery performance. It is our observation that many service asset managers understand the need for flexibility and are constantly struggling to accommodate the high levels of risk and the

complexities of their environment. The dynamic model for delivering customer service introduced here, we believe, is in fact, a true reflection of how such service process managers conceptualize their efforts to satisfy customer service needs. What has been lacking, however, are (1) effective decision support tools, (2) appropriate information technology to enable them to deliver the higher levels of service and efficiency that are required to support their company's service-centric service strategy goals and (3) a comprehensive approach to service product design and overall service business strategy development.

In this section, we consider what it takes for firms to transform service supply chain strategies into actionable plans. We will, in particular, consider the question of how firms can effectively deliver differentiated service products to support their service-centric strategy goals in a dynamic, optimized, option based manner. In order to illustrate how this can be done we describe a new generation of models and algorithms that support the delivery of service products in a dynamic and optimal fashion. These models and algorithms, which support implementation of the risk management approach introduced in this paper, are based on the following three steps:

- a) Probabilistic Forecasting
- b) Optimized Resource Deployment (Strategy)
- c) Optimized Resource Re-Deployment and Material Management (Tactics)

We now consider each of these steps and examine how they act to support a dynamic asset management approach.

**a) Forecasting:**

As we have noted, current approaches to service supply chain decision support are often based upon manufacturing (ERP), or finished product logistics (DRP), thought processes. In these environments, it is customary to generate forecasts of future demand and to use such forecasts as an input to resource deployment decisions to determine master production scheduling and finished product distribution. In the after-sales environment, which is characterized by many, very low demand items and dispersion of demand over multiple customer locations, the treatment of an estimate of future demand as a point value (deterministic number) can be quite misleading. While uncertainty is reduced as the planning horizon compresses (i.e., as we move from Business Model to Strategy to Tactics decisions), it is not eliminated. Since, in many cases, demand is low (e.g., the expected usage of a part at a location is 0.1 units per month), additional assumptions must be made to convert fractional forecasts into discrete values and to schedule the arrival of these rounded quantities.

The principle requirement to support a real option model is to have accurate estimates of future risks. Parameters describing probability distributions of item/location demand must be generated. These forecasts are influenced by both historical (time series) and causal factors. In effect what is needed is a "blended" forecast that takes a weighted average of time series and causal parameter estimates. Factors such as mean time between demand, local installed population, and projected end product utilization are used to come up with the causal component.

**b) Optimized Resource Deployment (Strategy)**

Service asset management is concerned with calculation of optimal resource deployment plans, where optimization refers to maximization of service performance subject to budget, risk and multiple hierarchy interaction constraints. The underlying decision problem at the Strategy level can be formulated as a constrained optimization problem where the decisions include the location, quantity and capabilities of the resources deployed (e.g., target inventory stocking levels for every part/location). The objective of this optimization is to maximize service, as it relates to customer satisfaction (e.g., availability of the products) or to minimize relevant costs (inventory investment, cash flow, etc.). Constraints include the interactions in the product/geography hierarchies as well as limits on budgets, cash flows and support capacities, probabilistic resource requirement forecasts, service agreement entitlements and weight/volume related constraints. Constraints imposed due to engineering changes and the life cycle of the installed base are also relevant.

**c) Optimized Resource Re-Deployment and Material Management (Tactics)**

The decision problem at the Tactics level involves managing the flow of materials within lead times. The specific decisions include new buys, repairs, allocations and excess stock transshipments. Other (human resource) asset re-deployments can also be considered at the Tactical level. The objective of this optimization is to minimize the total cost associated with new buys, repairs, transportation, shortages or travel, or, to minimize the risk of shortage and delay. The constraints include availability of materials, service level targets, inventory thresholds, availability of repaired materials, material flow constraints for normal and emergency shipments, depot capacities, and exception management thresholds, if any. The constrained optimization problem to be analyzed here trades off costs (including transportation, ordering and handling) against the service impact (again as measured by impact on the underlying risk of shortage and delay). Advanced filtering and triggering technology can be used to limit the demands on service management for recommended action review.

Analysis of the Tactical problem can also be used to quantify the risk profile for each part/location based on accurate information concerning the item's inventory position (on-hand, in repair, defectives on route, confirmed new buy orders, etc.). The basic idea is to use the probability distribution forecasts for item demand as an input and to compare projected material flows against projected demand scenarios. Accurate current system status can be retrieved from the underlying operational (transaction) systems. A typical output is illustrated in Figure 7.

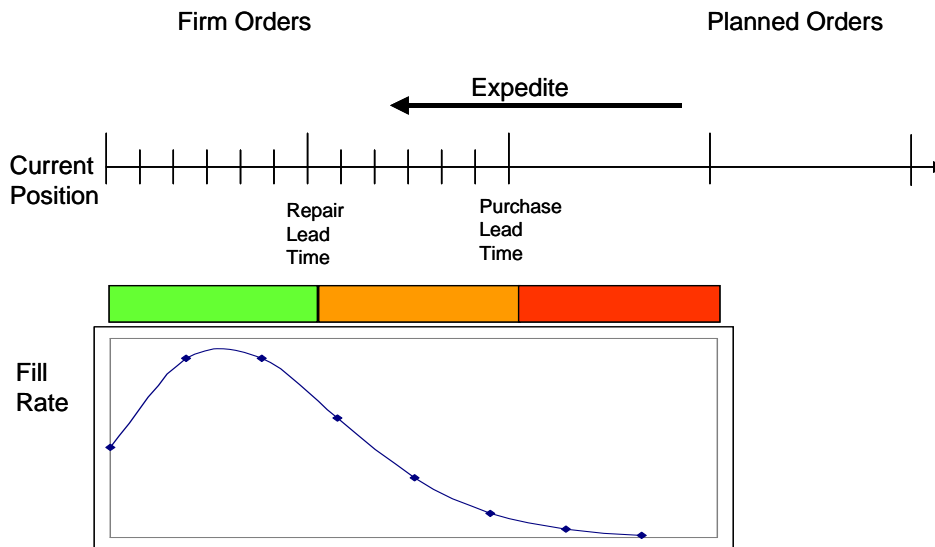


Figure 7. Shortage Probability Risk Assessment

MCA Solutions, Inc. ([www.mcasolutions.com](http://www.mcasolutions.com)), a company specializing in service supply chain decision support tools, has developed a web enabled, commercial software platform that incorporates these steps. Their product, the Service Planning and Optimization (SPO™) suite, has been successfully implemented in key high technology and aerospace and defense industries where service support is mission critical. SPO acts as a decision support software system that is linked to the underlying, transaction software systems that firms use to manage their service delivery processes, i.e., Enterprise Resource Planning (ERP), Supply Chain Management (SCM), Customer Relationship

Management (CRM) and Product Life Cycle Management (PLM). Figure 8 illustrates the information/decision flows between such systems and SPO.

Note that the optimization problems corresponding to the steps in the hierarchy illustrated in Figure 5 can be formulated as non-linear, integer, combinatorial, stochastic, non-stationary problems. These formulations are extremely difficult to solve, computationally as well as analytically. However, there is a rich history of academic literature on various approaches to solve variants of such problems in the area of multi-echelon, multi-indenture inventory modeling. The optimization algorithms needed for solving such problems are based on extensions to methodologies that have been developed for this class of problems in military and high technology applications (see Cohen et al[1990] and [1999] for selected applications at IBM and Teradyne respectively). The specific contribution made by the approach used by SPO is threefold. First, it represents, to the best of our knowledge, the first attempt to significantly expand the scope of the problems that have been formulated in an integrated way. In particular, the specific steps that are mathematically optimized are the strategic asset positioning, tactical asset deployment and the service demand fulfillment steps of figure 5 (the service business design problem is not solved in this system). Second, SPO has developed heuristic solutions to these problems that are excellent approximations to the optimal solutions, as validated through simulations and rigorous testing. These are also extremely effective, fast and highly scalable, which makes them practical for commercial use. Finally, it represents one of the few examples of solutions to the service supply chain design and management problem that have been commercially implemented, empirically tested, and significantly altered existing practice to yield comprehensive and compelling benefits to companies.

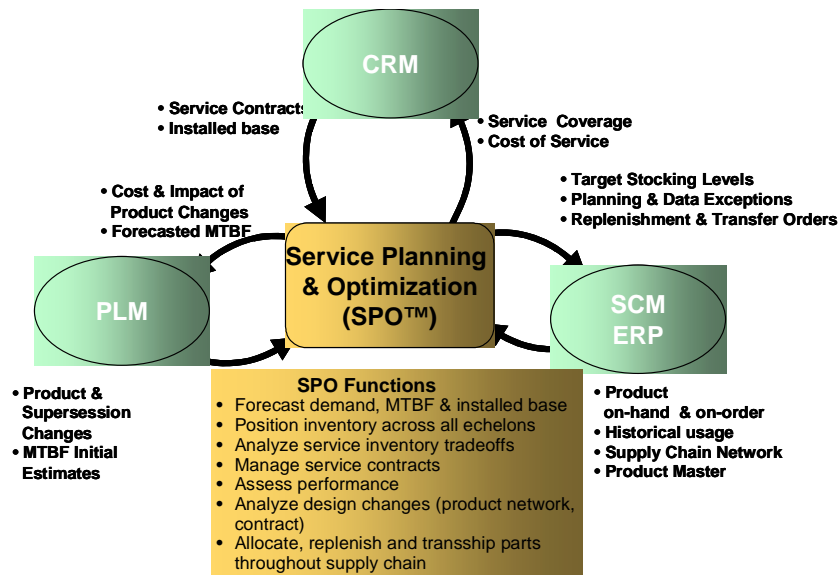


Figure 8. Service Planning and Optimization System Linkages

## CONCLUSION

In this paper we propose that firms adopt the principles of dynamic asset deployment to design their service products and to manage the deployment and utilization of enabling resources in the processes used to fulfill service product demands. The cornerstones of effective and profitable service delivery are service asset management, and service demand fulfillment. It is crucial for companies to recognize the interactions between these two levels of decision-

making. Specifically we propose the following recommendations for developing a DAD enabled customer-centric service strategy:

1. Recognize that a service product represents a commitment to customers that the product they have purchased will provide a guaranteed level of value creation. Accordingly, design and deliver such service products based on metrics directly related to customer satisfaction (e.g., equipment up-time) and link service product prices and revenues to achieved performance.
2. Include feedback derived from actual after sales product field experience to product design and improvement processes. The overall goal is to integrate service factors into the overall product architecture (“Design for Serviceability”), and to respond to performance issues through part engineering change. To do so requires a deep understanding of how customers derive value from products and designing services and service delivery processes that maximize customer value generation (Womack and Jones, 2005).
3. Optimize service asset management decisions that determine the capacity, location and capabilities of the resources that are consumed and/or utilized in the course of fulfilling service demands. Optimization requires that these decisions be made in a manner that explicitly accounts for the tradeoffs in cost and service quality that are driven by the dynamics and uncertainties associated with service demand and service resource supply processes. We propose, in particular, that firms adopt a risk management perspective supported by the use of appropriate solution algorithms. Such algorithms must consider the complex interactions among decisions throughout the service supply chain, as well as the sources and mechanisms that introduce risk.
4. Use integrated decision support tools to support efficient service product demand fulfillment. In particular, maintain customer service quality entitlements by linking asset management decisions and process execution systems to a common database driven by real time visibility of the underlying service related transactions. The fulfillment process must also take into account the potentially competing priorities for resources driven by service product commitments to customers.
5. Invest in processes and information technology capabilities that can enable the collection, analysis and dissemination of relevant information in a timely and collaborative manner. Recent advances in RFID (radio frequency identification device) technologies and remote sensing and diagnosis capabilities hold great promise in this dimension.
6. Design the service supply chain network to be consistent with the firm’s service strategy. This includes consideration of outsourcing of the infrastructure needed to deliver service through use of 3<sup>rd</sup> and 4<sup>th</sup> party service and logistics providers. In general, determine the appropriate mix of providers that can be integrated to support flexible and efficient response mechanisms.

Making good on these recommendations will require companies to adopt a wholly new paradigm for service supply chain management. The approach proposed here is analogous to treating service delivery as a real option, i.e., resource investments are required to “purchase” the option to deliver service and subsequently random contingencies occur that determine how the option will be “exercised” as the service requirement is fulfilled.

The concept of flexibility is not new! In fact, it has been a key *mantra* followed by many in the area of supply chain management for production materials. Indeed, managers of service supply chains have always recognized that they are engaged in a high stakes gamble requiring decision making in a complex and risky environment. The dynamic asset deployment approach that we are proposing here is a way to introduce the concepts of *flexibility* and *planned responsiveness* formally into the area of service delivery. Up until recently the analytical tools, data visibility and information systems integration required to implement such flexibility have not been available. The DAD strategy and software tools to support it, which are described in this paper, can help service supply chain managers to achieve the goal of supply chain flexibility. Companies cannot afford to neglect the potential of this approach in today’s

hyper-competitive, customer-centric world where service is often the key competitive differentiator. Companies that have embarked upon this path have met with resounding successes in a very short time, as evidenced from the results observed at Cisco Systems, where such strategies were implemented.

The new system that Cisco implemented over a 6 month period included the decision support system, SPO™, described earlier in the paper. The system's multi-echelon, probabilistic mathematical modeling and solution capabilities are used to recommend a "dynamic sparing" rule for setting target service levels for every part-location combination (more than 75 million in this case). Because Cisco operates in a highly dynamic environment characterized by tens of thousands of service contract transactions (new or changes) and thousands of service event fulfillment transactions (material flows) per day, the asset deployment decisions are re-optimized daily. Clearly, crucial to this capability is the need for an IT infrastructure that can provide real time visibility into the relevant data throughout the service network and solution algorithms that are extremely fast and responsive to changes in the environment.

As quoted by Mr. Jim Reily, Vice President, Technical Support, Cisco Systems "the results of implementing such a dynamic, service differentiated asset management solution have been dramatic for Cisco. Inventory investments have been reduced by nearly 21%. This has been accompanied by a simultaneous increase in service level from 94% to 97%. Working capital requirements have been further optimized by realizing a reduction in the purchase of new parts by almost \$65M. From the customers' point of view, this has resulted in improved productivity, maximized return on network investments, lower operating expenses and increased operating efficiencies."

## References

1. J. Bijesse, M. McCluskey and L. Sodano, "Service Lifecycle Management (Part 1): The Approaches and Technologies to Build Sustainable Competitive Advantage for Service," AMR Research Report, August, 2002.
2. M. A. Cohen and W. Pierskalla, "Target Inventory Levels for a Hospital Blood Bank or a Decentralized Regional Blood Banking System." *Transfusion*, Vol. 19, No. 4, July-August 1979, pp. 444-454.
3. M. A. Cohen, A. Tekerian, P. Kamesam, H. Lee and P. Kleindorfer, "OPTIMIZER: IBM's Multi-Echelon Inventory System for Managing Service Logistics." *Interfaces*, Vol. 20, No. 1, January-February 1990, pp. 65-82.
4. M. A. Cohen and S. Whang "Competing in Product and Service: A Product Life-Cycle Model." Special Issue of *Management Science* on Frontier Research in Manufacturing and Logistics, Vol. 43, No. 4, April 1997, pp. 535-545.
5. M. A. Cohen, Y. Wang and Y-S. Zheng, "Identifying Opportunities for Improving Teradyne's Service Parts Logistics System." *Interfaces*, Vol. 29, No. 4, July-August 1999, pp.1-18.
6. M. A. Cohen, H. Lee, C. Cull and D. Willen, "Supply Chain Innovation: Delivering Values in After Sales Service," *Sloan Management Review*, Vol. 41, No. 4, Summer 2000, pp. 93-101.
7. M. A. Cohen, J. Ren, T. Ho and C. Terwiesch, "Measuring Imputed Cost in the Semiconductor Equipment Supply Chain," *Management Science*, Volume 49, No.12, 2003
8. M. A. Cohen, V. Deshpande and K. Donohue, "An Empirical Study of Service Differentiation for Weapon System Service Parts." *Operations Research*, Vol. 51, No. 4, 2003.
9. M. J. Dennis and A. Kambil, "Service Management: Building Profits After the Sale," *Supply Chain Management Review*, Jan.-Feb. 2003, 42-49.
10. V. Deshpande, V., M.A. Cohen and K. Donohue, "A Threshold Inventory Rationing Policy for Service Differentiated Demand Classes," *Management Science*, Vol. 49, No. 6, 2003.
11. G.J. Feeney, and C.C. Sherbrooke, "The (s-1, s) Inventory Policy under Compound Poisson Demand," *Management Science*, Vol. 12, 1966, pp. 391-411.
12. Fisher, M.L., "What Is the Right Supply Chain For Your Product?" *Harvard Business Review*, Vol. 75, No. 2, 1997, 105-116.
13. T. Gallagher, M. Mitchke, and M. Rogers, "Profiting from Spare Parts", McKinsey Quarterly, 2 March, 2005.
14. E. Liang, "Optimizing the "Other Supply Chain": Software Breakthroughs in the Service Supply Chain Market," White Paper, Battery Ventures, San Mateo, CA, <http://www.battery.com/>, Jan 2002.
15. R. Wise and P. Baumgartner, "Go Downstream: The New Profit Imperative in Manufacturing," *Harvard Business Review*, Sept.-Oct. 1999, 133-141.
16. Womack, J.P. and D.T. Jones, "Lean Consumption," *Harvard Business Review*, March 2005,