Service system structure

John A. Buzacott*

Schulich School of Business, York University, Toronto, Ontario M3J 1P3, Canada

Received 22 April 1998; accepted 15 November 1999

Abstract

The structure of a service system should be matched to the requirements of customers. It is necessary to trade off the ability to handle the variety and complexity of the needs of different customers with the speed and efficiency of performing the required tasks. This paper develops a categorization of service system structures based on an analysis of their relative performance and how this performance is affected by the nature of the tasks that have to be performed. The categorization has the same property as the Hayes-Wheelwright categorization of manufacturing systems in that firms should aim at being near the diagonal of the matrix relating customer requirements and structural alternatives. The service system categorization also provides useful insight into the organization of manufacturing firms. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Service systems; Structure; Hierarchies; Requisite variety

1. Introduction

In developing and managing a service operation one of the critical decision areas is how to structure operations, i.e., how to divide up the tasks and assign them to people or machines. While there could be constraints of various types that restrict the set of available choices, there is usually considerable discretion. As compared to manufacturing operations, constraints which require a specific machine or type of machine to perform the task are usually less prevalent, so the range of choice is usually larger. Furthermore, there may be fewer constraints on the sequence in which operations are performed. Another key difference to manufacturing is that service operations are very frequently in direct contact with the customer. Given these differences, the ideas developed to understand how to divide up the tasks in manufacturing can rarely be applied to service operations without considerable adaptation.

Given a particular service situation, it would be desirable to establish some general guidelines on what sort of structure is likely to be most appropriate. So the goal of this paper is to develop a framework for thinking about the appropriate structure for a service system. Such a framework should help managers describe, diagnose, improve and design service operations. The framework should help managers in understanding the broad options that are available to them. Desirably it gives them a mental model of what is involved in making decisions about structure and what the feasible alternatives are likely to be. Also, given a particular service operation, the framework should be such that it helps managers diagnose the reasons behind
their experience of the effectiveness of the operation. It should be a guide for improving the system and suggest issues to look at more closely in order to make a particular structure perform effectively. It should identify those aspects of the operations that must be changed in order to make their selected structure appropriate. Lastly the framework should help designers choose the best structure for the system.

2. Methods of classifying service systems

There have been a variety of taxonomies proposed for classifying service systems. An excellent overview is contained in [1]. They describe one, two and three-dimensional approaches. For one dimension they describe the classification due to Chase [2] based on the extent of customer contact in service creation, while for two dimensions they describe the following:

- degree of tangibility vs. degree of technology,
- degree of tangibility vs. extent to which service provided to people or their possessions [3],
- degree of customization vs. extent to which service providers exercise judgement [3],
- customers either dominate or are dominating vs. participation is either active or passive.

Note that none of the above taxonomies directly relate to the structure of the service process.

Shostak [4] proposed a taxonomy that does consider the service process. She uses the two dimensions: degree of complexity of the service delivery structure, and the degree of divergence allowed at each process step. Another taxonomy that considers the service process is due to Wemerlov [5]. He proposes using the two dimensions: degree of complexity (standardized or customized service) and degree of customer contact (none, indirect, direct) customer service worker interaction, direct-customer/service worker interaction). Another approach is the service design matrix of Chase and Aquilano [6] that relates sales opportunity and production efficiency to service delivery options, where service delivery options are ranked in terms of the richness of information transfer. Production efficiency typically declines as the richness of information transfer increases, but on the other hand sales opportunity increases. This matrix enables production and marketing aspects to be traded off.

The best-known classification of service situations is probably that due to Schmenner [7]. Schmenner uses the two dimensions: degree of labor intensity and degree of customization or interaction. Within this he categorizes service situations into the four types: mass service (high labor intensity, low customization), service factories (low labor intensity, low customization), service shops (low labor intensity, high customization) and professional services (high labor intensity, high customization).

Another classification that has begun to attract interest is that due to Silvestro et al. [8]. Their classification has the two dimensions: number of customers processed per day and a factor related to balance between people and equipment, contact time, customization, discretion, front office/back office balance, and relative emphasis on process versus product. With this categorization they use empirical evidence to suggest that firms typically cluster near the diagonal. Professional service firms have low customers per day and high scores on the factor (e.g., high contact time, etc.), service shops are at a medium level on both dimensions and mass service has a high daily volume and a low score on the factor (i.e., low customization, etc.).

While all these taxonomies can provide useful insights on the issues to be considered in designing service systems, most of them do not have any underlying theory to justify the relevance and appropriateness of the classification. While different service processes are placed at different points in the two-dimensional classification, the classification provides no basis for saying whether one position is “better” than another or whether a particular position for a given firm is indeed the appropriate position given the processes used. The major exception is perhaps Chase’s service design matrix where it is possible to begin to construct an argument based on the amount and cost of information processing.

By contrast, the Hayes and Wheelwright [9] taxonomy in manufacturing has proved to be very useful as a diagnostic and design tool. It has two
dimensions, product or market variety, ranging from high to low, and type of production system, ranging from job shop, through batch production, flow line to continuous process. In general, the Hayes–Wheelwright model recommends that firms position their choice of process so that it is near the diagonal. The Hayes Wheelwright model is supported by evidence from field studies such as of Woodward [10] on the choice of process and the resulting performance of actual firms. The Hayes-Wheelwright model also indicates how a firm’s production process must evolve over the life cycle of its products.

What seems to be lacking in services is a taxonomy that is both descriptive and prescriptive in the way that the Hayes–Wheelwright taxonomy is for manufacturing. The purpose of this paper is to propose such a taxonomy for services. The basis for the taxonomy is a consideration of how different structures are able to cope with variability and disturbances. Disturbances or variability can arise either outside or inside the system. External variability can arise because it is not known what individual customers will want from the service providers until they are actually being served. Further, there can be changes in the level of demand, the requirements of different customers, the timing of demands, or the mix of demands. Internal variability can arise due to the nature of the tasks. As the cognitive demands of the task increase there can be increasing variability in the time to perform the task. But internal variability can also arise because of breakdowns of equipment, planned or unplanned absences of service personnel, and inadequate training of people.

The service design problem is to develop a system that enables customers’ demands to be met in an effective way. It is clear that as the complexity of the demands of customers increases, the complexity with which the service processes have to cope will increase. That is, the design of the system has to conform to Ashby’s law [11]: only variety can destroy variety. In other words, the service system has to be capable of coping with the range and complexity of customer requirements and the uncertainty about the timing and level of customer demands. Also, the service system has to be tolerant of its internal variability and uncertainty. So the approach used in this paper is to look at the match between the variability in demands from customers and the inherent capability in variability handling of the way in which the processes that meet customer demands are structured.

A system that is structured to provide effective handling of high levels of variability is likely to be inherently less efficient and more costly than a system that only has to cope with limited variability. So it is always necessary to make tradeoffs in system design between the effectiveness in coping with variability and the efficiency, speed and cost of providing the service. Different service providers can make different choices of where to operate and their choice will influence their market share and attractiveness to customers. However, it is likely that certain choices are likely to survive in the market better than others.

3. Structuring processes

In order to meet the demands of a given customer for service certain tasks will have to be performed. Not all these tasks may be evident before the customer arrives; indeed some of them will only become apparent as the service process proceeds. For example, diagnosis of the needed repairs on a piece of equipment is necessary before the actual repair can be carried out, and even after the initial diagnosis and repair have been performed other problems might surface which require further tasks to be done. So the service process might have to recognize such possibilities and it should be structured appropriately to deal with such situations. Rather than providing the internal capability to deal with variation, the service process can also be designed and marketed to exclude customers whose needs do not fit within a given envelope. For example, Starbucks Coffee shops makes it well known that they specialize on coffee and perhaps cakes and biscuits to accompany the coffee and so exclude customers who want a burger. Such exclusion (or isolation) is an effective means of limiting the variability that has to be handled by the service provider. This then leads to one of the key dimensions in the taxonomy: the degree of variability and uncertainty in customer demand. Along this
dimension there are two aspects that need to be considered. The first is the range of customer demands with which the system is designed to cope, and the second is the extent to which the processing requirements of a given customer are known at the time when they make their initial contact with the service provider. The second aspect becomes increasingly important as the range of customer demands increases. When all customers require identical service then the requirements of the system design are less complex.

Given the nature of the customer demand and requirements and hence the tasks that might be required then an appropriate structure will have to be chosen for the service system. Two possible approaches can be used, one is to start with the desired attributes of the individual jobs, and another is to start with the overall pattern of relationships that defines the structure of the system. It is useful to begin by discussing the options available for individual jobs, and then look at the overall patterns that these options imply.

The relevant aspects of individual jobs are perhaps best described using the approach suggested by Rolfe [12]. She uses two attributes: technical complexity and discretion/prescription. Technical complexity means the complexity of the procedures, the knowledge required, and the number of steps or range and variety of tasks. Discretion relates to the judgement that the individual must exercise. This will be determined by the degree of uncertainty and the prescribed range over which discretion can be exercised.

If no discretion is required, i.e., tasks to be done are the same for all customers, then jobs could be designed to have either a low technical complexity or a high technical complexity. Achieving a low technical complexity will require giving each worker a small number of tasks. This in turn implies that a number of workers will be needed in order to perform all the tasks required by a given customer, and one possible way of structuring the system consistent with this is as a flow line or series system. If jobs can have a high technical complexity then it might be possible to give to a worker the responsibility of doing all tasks required by a customer. This would then result in the system having a parallel structure with any of the workers capable of doing all tasks for a customer and each customer’s service being provided by only one worker. Different customers would deal with different workers.

If different customers have different requirements, and so different customers would require different tasks, then the system has to be structured to deal with this. One approach is to offer customers a menu so they choose the menu items that are closest to their needs. Each menu item defines a set of tasks that must be performed in order to deliver the service. The service provider is not called upon to decide what service to deliver to the customer, the choice of the item from the menu determines the service. Different ways of designing the jobs are then possible.

(i) Specialization: One possibility is to have a worker specialize on the set of tasks required for a specific menu item so there would be different workers for different menu items. The worker is not then called upon to exercise any significant discretion.

(ii) Parallel: Another possibility is to have any worker able to do all tasks for any customer, i.e., a parallel structure. This then requires the worker to exercise sufficient discretion to relate the customer’s menu choice to the appropriate set of tasks.

(iii) Series: If the different menu items require essentially the same types of steps, although the specific details differ between different menu items, then a series structure can be used with multiple workers required in order to perform all required tasks for a given customer. Again each worker has to exercise sufficient discretion to do the tasks appropriate for the customer’s menu choice.

Note that as well as greater discretion, parallel would require greater technical complexity than the specialized structure.

If customer requirements are not known when the customer arrives and the worker takes on part of the task of determining what their requirements are, then the worker has to exercise some discretion and the structure will have to be able to cope with this uncertainty and its resolution. One way is to have workers who are totally general purpose, i.e., capable of performing all required tasks irrespective of the customer’s requirements. This would then require either series or parallel structures where in
both alternatives there would be substantial technical complexity and significant discretion in the worker's tasks.

Because of the complexity of general purpose approaches, it might be desirable to simplify the set of tasks required of a worker. This can be done in several ways. One is to separate diagnosis from the tasks consequent on the diagnosis so a worker either does just diagnosis or just the tasks consequent on a particular diagnosis. Indeed, the complexity of diagnosis might be such that diagnosis tasks are allocated to a number of different workers. This means that the diagnosis performed by a given worker sometimes simplifies to making one choice out of a very small range of possible choices. Often, diagnosis proceeds hierarchically (or as a sieve) with the first worker performing some tests, and, if they do not identify the action required, then a second worker would perform more tests. The complexity of the diagnosis can either increase with successive steps or decrease with successive steps. If it increases, then the resulting structure is bottom-up, while if it decreases then it is top-down.

As an example, suppose a customer wants to trade a convertible debenture through a telephone discount brokerage. First they will speak to an agent who will verify identity and find out what the customer wants. They will then be transferred to a representative who will recognize, unlike shares, that the debenture cannot be traded on the computer system. Lastly they, or the representative, will speak to the bond desk who will carry out the trade. On the other hand, if the same customer wants to do the same trade with a full service broker, the customer will speak to their broker, discuss the trade, and the broker will then probably arrange the trade and call the customer back when it is complete. The discount broker is using a bottom-up diagnosis structure while the full service broker is using a top-down structure.

There is a further aspect involved in comparing top-down and bottom-up hierarchies. The bottom-up structure has to be established prior to the arrival of a customer. It requires rules to be set up in advance limiting the discretion available at each diagnostic step. However, for top-down hierarchies, there is no essential need to define in advance how each customer requirements are to be met. Part of the role of the customer's initial contact (the top of the hierarchy) is to define the process steps required by the customer and assign them to workers. So clearly this structure will be the one most capable of responding to variety in customer needs.

4. Models for comparing alternative systems

It is necessary to rank the alternative structures in terms of their ability to respond to variety. In a menu system, it is fairly clear that the specialized system has the greatest tolerance for variety. But it is possible to develop simple mathematical models that provide insight into the relative capability of the different structural alternatives. These models are described elsewhere in detail [13] but it is useful to summarize their main results so that the implications can be shown.

4.1. Basis of comparison

In comparing alternative systems it will be assumed that customers or jobs arrive according to a stationary process. The mean arrival rate is denoted by $\lambda$ and the squared coefficient of variation of the time between arrivals is $C_\lambda^2$.

The performance measure that will be used is the average time required for providing service to a customer, or, equivalently, using Little's law, the average number of jobs or customers in process (WIP). Given WIP, it would also be possible to determine the costs incurred in providing service and compare systems on the basis of cost. Rather than using costs, comparisons between systems will be made assuming that all systems use the same number of workers or facilities. The differences between system structures arise from the way in which tasks are assigned to the workers.

4.2. Series vs. parallel

Suppose that when the work is organized in series there are $m$ workers and each worker has the same amount of work. Let $\tau$ be the average time required by each worker to do the required work for a customer and let $C_\tau^2$ be the squared coefficient
of variation (scv) of this time. Now compare this with the parallel system with \( m \) workers in parallel each of whom does all the required tasks for a given customer. Assuming that there is no loss of performance if a worker does all the required tasks, then the time a worker requires to do all the tasks will have mean \( m \tau \) and squared coefficient of variation \( C_S^2/m \). Now when customers arrive at a rate \( \lambda \) with the squared coefficient of variation of time between arrivals being \( C_a^2 \), then they will have to queue for service at each worker. In the parallel case assume that the customers are allocated to workers at random. This means that the mean arrival rate at service at each worker. In the parallel case assume

\[
\text{arrivals being}
\]

\[
\text{with the parallel system with}
\]

\[
\text{of variation (scv) of this time. Now compare this}
\]

\[
20
\]

\[
J.A. Buzacott
\]

\[
C
\]

\[
\text{have mean}
\]

\[
\text{the time a worker requires to do all the tasks will}
\]

\[
\text{mance if a worker does all the required tasks, then}
\]

\[
1
\]

\[
\text{if a worker does all the required tasks for a given}
\]

\[
2\text{(1}
\]

\[
\text{variation of the interarrival time is}
\]

\[
C
\]

\[
2\text{(1}
\]

\[
C
\]

\[
\text{coefficient of variation of service time is the same}
\]

\[
2S
\]

\[
\text{for both types and equals}
\]

\[
C
\]

\[
\text{Then in the specialized case, one server will be dedicated to type}
\]

\[
2
\]

\[
\text{while the type 2 server will have utilization given by}
\]

\[
\text{Then the queue length at server}
\]

\[
2(1 - \rho)WIP_{\text{SER}} = C_a^2 + (2m - 1)C_S^2,
\]

\[
\text{Series:}
\]

\[
2(1 - \rho)WIP_{\text{PAR}} = m(1 - 1/m + C_a^2/m + C_S^2/m)
\]

\[
= m - 1 + C_a^2 + C_S^2.
\]

\[
\text{Parallel:}
\]

\[
\text{It is then easy to show that the WIP will be less for series if}
\]

\[
C_S^2 < 1/2.
\]

\[
\text{For other values of } \rho \text{ it is possible to show that the requirement for series to}
\]

\[
\text{be better than parallel is that both}
\]

\[
C_a^2 \text{ and } C_S^2 \text{ be small, e.g., for } \rho \text{ small the condition becomes}
\]

\[
C_a^2 + C_S^2 < 1.
\]

\[
\text{Series is only appropriate if there is}
\]

\[
\text{very little variability in the time required by a worker to perform the required tasks and the}
\]

\[
\text{variability of the time between arrivals is also fairly}
\]

\[
\text{small.}
\]

\[
\text{Performance of parallel systems can be improved by using other rules to allocate customers to servers. For example, allocating them cyclically: customer 1 to server 1, customer 2 to server 2, customer 3 to server 1, etc. will ensure that the parallel system is always better than the series system. But such an allocation requires resources to be used for such a control. Alternatively, if there is a single queue and if the customer at the head of the queue goes to the next free server then the performance of the parallel system is further improved.}
\]

\[
4.3. \text{Parallel vs. specialized}
\]

\[
\text{A specialized system is only meaningful once different customers require different service. Suppose there are two types of customers. Let}
\]

\[
f_1 \text{ be the fraction of type 1 and } f_2 \text{ the fraction of type 2 (} f_1 + f_2 = 1 \text{) and let the total arrival rate of all}
\]

\[
\text{customers be } \lambda. \text{ Assume for simplicity that arrivals are Poisson. Suppose the time to serve a type 1 customer has mean } \tau_1 \text{ while the time to serve a type 2 customer has mean } \tau_2. \text{ Suppose the squared coefficient of variation of service time is the same for both types and equals } C_S^2. \text{ Then in the specialized case, one server will be dedicated to type 1 while the other server will be dedicated to type 2. The utilization of the type 1 server } \rho_1 = f_1 \tau_1, \text{ while the type 2 server will have utilization given by } \rho_2 = f_2 \tau_2. \text{ Then the queue length at server}
\]

\[
2(1 - \rho)WIP_{\text{SER}} = \rho_1^2(1 + C_S^2).
\]

\[
\text{In the parallel case there are two parallel servers that serve either type of customer and work is allocated to the two servers so that their utilization is identical, i.e. } \rho = (\rho_1 + \rho_2)/2.
\]

\[
\text{The variance of the service time of an arbitrary customer will be given by}
\]

\[
\text{var } S = (f_1 \tau_1^2 + f_2 \tau_2^2)C_S^2 + f_1 f_2 (\tau_1 - \tau_2)^2.
\]

\[
\text{It is then possible to determine the queue length as a function of the rule used for allocating customers to the parallel servers. To simplify the illustration, assume } \rho_1 = \rho_2 = \rho, \text{ i.e., } f_1 \tau_1 = f_2 \tau_2. \text{ Then the total WIP with specialized servers is}
\]

\[
\rho^2((1 + C_S^2)/(1 - \rho)). \text{ If customers are assigned to the parallel servers at random then the total WIP will be given by}
\]

\[
(1 - \rho)WIP_{\text{PAR}} = \rho^2(1 + (C_S^2 + (f_1 - f_2)^2/4f_1 f_2)).
\]

\[
\text{Note that the specialized system and the parallel system have the same performance when } f_1 = 1/2, \text{ but otherwise the specialized system is always better. However, if an alternative rule for allocating customers to the parallel servers is used, such as cyclic or allocate to first free server, then it is easy to}
\]
show that the parallel system will be better than specialized when \( f_1 \) is neither close to zero or one, i.e., when \( \tau_1 \) and \( \tau_2 \) are not too dissimilar. That is, parallel systems can, with fairly minimal control, outperform the specialized system when the service requirements of the different customer types are not too different.

### 4.4. Specialized vs. Hierarchical

It is clear that the performance of a specialized system would deteriorate if the different types of customers have difficulties in identifying the server appropriate to their needs. If the allocation of customers to servers is incorrect a server will end up having to serve a mix of different types of customers. As the allocation errors increase, the specialized servers would end up with a mix of work comparable to that which they would have if a parallel allocation were used.

There are several possible approaches to deal with this depending on how difficult it is for a server to identify what service a customer requires. If diagnosis requires essentially zero time (instantaneous diagnosis) then one approach is to have all customers queue in front of servers specialized to short duration service. As soon as the customer reaches a server and the server identifies that they require long service then they are immediately diverted to a server specialized to long duration service. If, however, the diagnosis requires a finite time, then one server will specialize in diagnosis and allocate customers to either a server specialized to short duration service or to a server specialized to long duration service as appropriate (separate diagnosis).

#### 4.4.1. Instantaneous Diagnosis

The simplest case is that where there are two servers. Suppose short service requires a time \( S_S \) while long service requires a time \( S_L \) and the fraction of jobs requiring short service and long service is \( f_S \) and \( f_L \) respectively \( (f_S + f_L = 1) \). Assume that the squared coefficient of variation of short- and long-duration services is the same and equals \( C^2 \). Also assume that the squared coefficient of variation of arrivals to the system is \( C^2_a \). Suppose the work load on the two servers is balanced so that the service time at the first server, the short-duration server, will consist of a mixture of zero duration services corresponding to the identification of long duration service requirements and short duration services. Hence \( C^2_{SS} \), the squared coefficient of variation of service at the short duration server will be given by

\[
C^2_{SS} = (f_L + C^2_3)/f_S. 
\]

At high traffic the scv of departures from this server will be \( C^2_{SS} \) and a fraction \( f_L \) will be diverted to the long duration server. Using the results in the appendix for the splitting of customer streams the scv of arrivals at the long-duration server will be given by \( 1 - f_L + f_L C^2_{SS} = (f_S^2 + f_L^2 + f_L C^2_3)/f_S \). Hence, as \( \rho \to 1 \) the total queue length in the system will be given by

\[
2(1 - \rho)WIP_{ID} = C^2_a + (f_S + C^2_3)/f_S + (f_S^2 + f_L^2 + f_L C^2_3)/f_S + C^2_L. 
\]

Note that since \( f_L < f_S \), it will always be preferable to have the short duration server do the instantaneous diagnosis. Also note that as \( f_L \) increases the queue length increases, i.e., it is least when \( f_L = 0 \).

Now compare with two parallel servers who do both short and long services. In this case \( C^2_{SP} \), the scv of the service time at a server, will be given by

\[
C^2_{SP} = [(f_S - f_L)^2 + C^2_3]/4f_Sf_L. 
\]

Assuming cyclic allocation of arriving jobs to servers the total queue length as \( \rho \to 1 \) will be given by

\[
2(1 - \rho)WIP_{PAR} = C^2_a + 2((f_S - f_L)^2 + C^2_3)/4f_Sf_L. 
\]

It can be seen that this will decrease as \( f_L \) increases from 0 to \( 1/2 \). For \( f_L \) less than about 0.2, the system with two parallel servers will be worse than the system with specialized servers and instantaneous diagnosis.

#### 4.4.2. Separate Diagnosis

For simplicity assume that there are three servers. All customers require diagnosis which takes a time of \( d \). A fraction \( f_{LL} \) require long service of duration \( S_L \), while the remaining fraction \( f_S \) require short-duration service of duration \( S_S \). Assume that these service times are deterministic and that \( d = f_L S_L = f_S S_S \) so the utilization of all three servers
will be identical and equal to \( \rho \). Then the scv of arrivals at the short-duration server will be \( f_1 \), while the scv of arrivals at the long-duration server will be \( f_s \). Hence it follows that the total queue length in the system as \( \rho \to 1 \) will be given by

\[
2(1 - \rho)WIP_{SD} = C_s^2 + 1.
\]

Now, in the alternative parallel system consisting of three servers, all service, including diagnosis, is undertaken by the same server. So \( C_{SP}^2 \), the scv of the service time will be given by

\[
C_{SP}^2 = f_s f_1 [(s_L + d) - (s_L + d)]^2 / (d + f_s s_L + f_1 s_L)^2,
\]

which simplifies to \( C_{SP}^2 = (f_s - f_1)^2 / (f_1 s_L) \). Hence it follows that if customers are assigned to the parallel servers cyclically that the arrangement with separate diagnosis will be better if \( (f_s - f_1)^2 > 3f_s f_1 \), i.e., if \( f_1 < 0.173 \), or \( s_L/s_s > 4.73 \).

Both models can be modified to reflect different relative utilization of the specialized servers. The general conclusion from both modes of diagnosis is that as the difference in the service times of the different customer types increases the attractiveness of diagnosis and separation of the customer types increases.

4.5. Top-down vs. bottom-up diagnosis

As the complexity of diagnosis and the associated service increases it is necessary to consider how to structure the system. Typically, the process of diagnosis can be considered to involve a sequence of tests. One server can perform all tests, or it may be split between a number of servers. If it is split, then the task of each server is simpler, and so they do not need to have the same level of skill as in the situation where all steps in the sequence are carried out by one server. If the tests performed by one server are simple, then it is often the case that for certain outcomes of the test the server will also carry out a service task, while for other outcomes the job will be passed on to another server for further diagnosis and service. By “bottom-up” diagnosis is meant the situation where the diagnosis steps are split among a number of servers, with typically each step becoming more complex, while by “top-down” is meant the situation where one server can do all diagnostic steps.

Suppose there are three possible outcomes of the diagnostic tests, outcomes 1, 2, 3, and outcome \( i \) will require a service of type \( i \), performed by server \( i \). There are two tests that can be sequenced so that test 1 has two possible outcomes, service of type 1 is required or test 2 is required. Test 2 has the two outcomes, service of type 2 is required or service of type 3 is required. Then compare the two different structures, “bottom-up” where there are two test servers, one specializing in the first test, the other specializing in the second test, and “top-down” where there are two test servers, each of which does both tests. Suppose that the time to perform the first test is \( D_1 \), the time to perform the second test is \( D_2 \), and the probability that the second test is needed is \( p \). Assume that in the “bottom-up” alternative the two test servers are equally loaded, i.e., \( E[D_1] = pE[D_2] \), and assume that in the “top-down” alternative arriving customers are allocated cyclically to the two test servers. This means that the utilization of all test servers in both alternatives will be identical and equal to \( \rho \). Suppose that the scv of \( D_1 \) is the same as the scv of \( D_2 \), i.e., var \( D_1/D_1^2 = var D_2/D_2^2 = C_s^2 \). Let \( p' \) be the probability that the second test reveals that service of type 2 is required and for simplicity assume that the various types of service are deterministic and result in the utilization of the servers all being equal to \( \rho \). Let \( C_a^2 \) be the scv of arrivals at the system.

**Bottom-up**: With bottom-up the scv of the service time at either test server will be \( C_s^2 \). Arrivals at server 1 will have scv of \( 1 - (1 - p) + (1 - p)C_s^2 \), arrivals at the second test server will have scv of \( 1 - p + pC_s^2 \), arrivals at server 2 have scv of \( 1 - p' + p'C_s^2 \), and arrivals at server 3 have scv of \( 1 - (1 - p') + (1 - p')C_s^2 \). Note that the total of the scv’s of arrivals at all servers is then \( C_a^2 + 2 + 2C_s^2 \). Then the total queue length in the system as \( \rho \to 1 \), will be given by

\[
2(1 - \rho)WIP_B = C_a^2 + 2 + 4C_s^2.
\]

**Top-down**: The scv of the service time at one of the parallel test servers \( C_f^2 \) will be given by

\[
C_f^2 = (E[D_1])^2 + p(E[D_2])^2)C_s^2 + p(1-p)(E[D_2])^2/(E[D_1] + pE[D_2])^2.
\]
The scv of the stream of customers going from one test server to server 1 will be \( 1 - (1 - p) + (1 - p)C_p^2 \), while to server 2 it will be \( 1 - p(1 - p') + p(1 - p')C_p^2 \) and to server 3 it will be \( 1 - pp' + pp'C_p^2 \). Assuming that when the streams from the two test servers merge the asymptotic results for merging streams can be used, the arrivals at server 1 will have the same scv as the stream coming from a given server. It follows that the total queue length in this system as \( \rho \rightarrow 1 \) will be given by

\[
2(1 - \rho) \text{WIP}_T = C_a^2 + 2C_p^2 + 2 + 2C_r^2.
\]

**Comparison:** It follows that \( \text{WIP}_T < \text{WIP}_S \) if \( C_p^2 < C_r^2 \). This condition reduces to \( 1/p - 1 < 1/(3 - 1/p) \).

That is, if \( p < \frac{1}{3} \) it is impossible for \( \text{WIP}_T < \text{WIP}_S \) while for \( p > \frac{1}{3} \) the critical \( C_r^2 \) decreases, so when \( p = 1 \), the critical \( C_r^2 = 0 \). When \( p = \frac{1}{3} \), then the critical \( C_r^2 = 1 \). Typically a high \( C_r^2 \) is associated with complex decision making, so the conclusion is that “top-down” is preferable when it is likely that both test steps will be needed and when decision making is complex. “Bottom-up” is preferable when one test can usually verify the problem, or when decision making is relatively simple.

It is possible to construct other models of this comparison. For example compare a situation where in “bottom-up” the service of type \( i \), \( i = 1, 2 \), is carried out in conjunction with the test step, while in “top-down” there are two servers, one of which does both test steps and the other does both services of types 1 and 2. Again, it is found that if a single test step is usually enough “bottom-up” is better, but if most customers require both test steps then “top-down” is better. For example, equipment repair usually uses a “bottom-up” structure. Many sources of failures are easy to identify and easy to fix, so can be done by less skilled people. They will then pass on the harder to find and harder to fix problems to more skilled people. By contrast, professional services such as law tend to use a “top-down” structure where a customer sees a senior lawyer first. There are usually many possible diagnoses that the senior lawyer considers. Once diagnosis has been made then the investigation and preparation of documents and briefs can be passed on to a junior.

5. **Tradeoffs and diagonals**

The insights from the above models enable the development of a matrix to describe the structure of service processes. The matrix has two dimensions (see Fig. 1). The first describes the variability and complexity of the tasks that have to be done in order to meet customer requirements. This complexity is determined by the two aspects, the complexity of diagnosis, and the variability of the actual service task. The second dimension relates to how the system is structured in order to cope with this complexity and variability. The simplest structures are those where all customers are treated in the same way, so essentially no diagnosis is needed and the service facilities or people have simple tasks, i.e., the service system has a series structure. Next, in order of complexity are those structures which also treat all customers the same way, but, because of the variability associated with individual tasks or the arrival of customers, it is preferable to use a parallel structure for the service system. The next step in complexity of structure occurs when the differences in service required by different customer types becomes substantial, but it is easy for customers to recognize what type of service they need and choose the appropriate server. Facilities then specialize by customer type, and the time required for diagnosis is minimal. The next level of structural complexity arises when customers can no longer determine the type of service they need and the differences in service requirements between different diagnoses become significant enough that service delivery requires specialization. If diagnosis and service are usually simple and relatively rarely is more complex diagnosis and service required then bottom-up structures are appropriate. If, however, many customers require most diagnostic steps, then a top-down structure is preferable.

Representing these choices by a two-dimensional matrix gives the requirement that the best process choice lies on the diagonal. Choices to the right of the diagonal imply that the structure is too simple and as a result will not be effective in meeting customer needs or it will have a poor response. Choices to the left of the diagonal use a more complex structure than necessary so are likely to incur high costs. Further, when the opportunity for
a service is first recognized it is likely to be structured so that it is well down the diagonal, i.e., a complex structure is chosen for service delivery. With time (and increases in demand, simplification of processes and education of customers) movement up the diagonal will occur (see Fig. 2).

5.1. Examples of placement of processes

To illustrate the placement of typical processes on the matrix consider the following examples.

Law office: Typically a top-down structure, however, some of the store front law offices advertising a small range of functions such as wills and property conveyancing are moving to a bottom-up structure.

Repair hotline: Computer hardware and software manufacturers typically use a bottom-up structure. Many customer problems are frequent and can be dealt with by front-line agents with fairly limited breadth of training. More complex problems can be then passed on to technical specialists. The increasing use of voice response systems so that customers can do some self diagnosis of the problem (and at least separate out customers according to product line) indicates a movement towards a system with customer self diagnosis and specialization.

Bank branch: Traditionally, most bank branches were structured so that they had different counters for clients just depositing or cashing cheques for those clients purchasing CD’s or other investments, i.e., specialization with customer diagnosis. Many bank branches have moved to have multifunction tellers who do all functions and have invested in providing them with the computer support and
training needed to do this. This represents a move-
ment towards parallel structures, enabled by the
computer technology that simplifies the task of
selling CDs.

**Fast food outlets:** The typical way of organizing
fast food outlets is McDonalds with parallel gen-
eral purpose queues. Given different customers
vary in the complexity (and number of items) of
their order, such a structure seems appropriate.
However, some fast food outlets have moved
towards more of a series structure with a single
initial queue and several subsequent queues in
series. This results in separation of the order
taking task from the task of delivering the order
to the customer and the task of getting drinks
(indeed in some this task is transferred to the
customer). Such structures seem to be more com-
mon when orders are made up for each customer
(rather than having a stock of product ready to
dispense). Probably this means that the time re-
quired for individual order preparation shows little
variability.

6. **Conclusions**

6.1. **Use for diagnosis and design**

It is clear from the examples that the classification
can be used for diagnosis and design. It sug-
gests directions for improving processes, e.g.,
through using information technology to reduce
task variability and hence permitting movement up
the diagonal. It also suggests directions in which
certain processes might evolve. For example, inter-
net tools might enable significant levels of self diag-
nosis of health problems by patients, and hence
a move of the medical system from its current
predominately bottom-up structure to a system
with specialized facilities selected by patients
(rather than by referral by primary care doctors).

6.2. **Implications for manufacturing**

While the classification was developed based on
service systems, it has relevance to manufacturing.
Generally in manufacturing the control activities are structured hierarchically, with the long-time horizon decisions made by people at the top of the hierarchy and the short-time horizon decisions made by people at the bottom of the hierarchy. The manufacturing process can be viewed as the “customer” whose problems create demands for service by the control system. From this perspective it can be seen that if short-time horizon decisions are also quick to deal with or “serve”, then manufacturing control is characterized by a “bottom-up” structure. Increasing automation of control can be viewed as a movement towards a specialized structure. But this could mean that the people involved in control are only left with difficult and complex decisions for which a “top-down” structure might be more appropriate. This probably implies “flat” organizations and enhanced knowledge required by those involved in control.

Another set of tasks most manufacturing firms have to deal with are those related to selling the product to its external customers. A “bottom-up” structure is appropriate if most orders are reasonably standard so can be taken and dealt with by an order desk or professional sales staff. A “top-down” structure (involving top management in the sales negotiation process) is appropriate if each product sold is fairly unique so there is a fair bit of discretion and a wide prescribed range covering the possible products that can be made. The “diagnosis” of the best way of meeting a customer’s requirements is then complex, and so the “top-down” structure is appropriate. Clearly, if a firm makes a significant number of “specials” (involving a higher level of expertise in sales and application engineering) as well as standard products (involving an order desk and a catalogue) then it is going to have problems in dealing with either efficiently. Such a situation is likely to be best resolved by segregation or specialization.

It is of interest to note that the firm’s control structure has to deal with at least two types of “customers”, the external customers to whom it sells products, and the production processes that generate demands for service. Clearly, if the nature of the tasks are similar in duration at the points which these two control systems share in common, or if they can be segregated, then control will be easier. For example, if both use a bottom-up structure with the same individual at the top of both hierarchies then this individual will have similar duration tasks from both sources. This would be typical of consumer product firms that use assembly lines to make the product. Both the sales and production control would be able to use a bottom up structure with top management at the top of both structures. It is clear that complexity of control (and probably diminished performance) will occur if an individual’s role in the two control systems results in the tasks for one control system being significantly different in duration to the tasks in the other. Another example would be a small firm making customized products using a job shop manufacturing organization might use a top-down structure for both production control and sales.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant OGP 0138270.

Appendix A. Queuing approximations

The following results from queueing theory are used as the building blocks for analyzing the different configurations.

A.1. Queue length and waiting time

Because there is no exact formula for the queue length or waiting time in a queue with general interarrival times and general service times (the G/G/1 queue) it is necessary to use an approximation. Suppose that the arrival rate of customers is \( \lambda \) and the squared coefficient of variation of interarrival times is \( C^2_s \), while the service time has mean \( s \) and squared coefficient of variation \( C^2_S \). Then the utilization of the server is \( \rho = \lambda s \). Under heavy traffic, i.e., as \( \rho \to 1 \), an appropriate approximation for the mean queue length \( E[L] \) is that due to Harrison and Nguyen [15].

\[
\lim_{\rho \to 1} 2(1-\rho)E[L] = C^2_s + C^2_S.
\]
Because the server utilization is close to 1, the squared coefficient of variation of departures $C_3^2$ will be equal to $C_5^2$.

A.2. Splitting and merging (see [14])

Consider a stream of customers where the time between customers has mean $1/\lambda$ and squared coefficient of variation $C^2$. Then if customers are selected from the stream with probability $p$, then the time between selected customers will have mean $1/(p\lambda)$ and squared coefficient of variation $1 - p + pC^2$.

Suppose there are two streams of customers such that the time between customers have means $1/\lambda_1$ and $1/\lambda_2$ and squared coefficient of variation $C_1^2$ and $C_2^2$, respectively. Then if the streams are merged the mean time between customers in the merged stream is $1/\lambda$, with $\lambda = \lambda_1 + \lambda_2$, and the squared coefficient of variation, $C$, can be approximated by

$$\hat{\lambda}C^2 = \lambda_1C_1^2 + \lambda_2C_2^2$$

References