Cost factors involved in computing centers that tend to motivate the centralization as opposed to the decentralization of computing services are evaluated, and a cost-minimization solution is presented.

Proposed and evaluated is a strategy for linking large regional service centers that perform standard production services with satellite centers that perform local personalized services.

Emphasized are techniques, including user waiting, for evaluating the two characteristic service types.

Centralization or dispersion of computing facilities

by D. N. Streeter

The growing and changing role of computer usage makes necessary a frequent reexamination of the distribution and arrangement of computing facilities within an organization. Such examinations often show that a system configuration that was eminently sensible several years ago is unsatisfactory today. Having made that discovery, management is confronted with the problems of determining a more appropriate system arrangement, and then persuading the parties involved – users, operations personnel, and management – that this change is desirable.

A chapter of a forthcoming book to be published by John Wiley and Sons, Inc., New York, New York, this paper examines several of the relationships and tendencies that influence centralization-dispersion judgments. Some guidelines for determining an appropriate distribution of computing resources are developed and illustrated by example. The goal is to devise a strategy for centralization versus decentralization, and to develop a methodology for decision. The general conclusion of this examination is that at least some of the computer services in a large, widely dispersed organization can be provided economically out of relatively few centers. This material is published here to expedite its availability.

The tendency toward centralization

Quite understandably, the use of computers in many organizations has grown, not in accordance with an overall plan, but as the sum total of the plans of several, or many, more or less autonomous computer centers. At some point, it becomes clear that a review of the consequences of this fragmentation is in

order—that substantial benefits might result from a greater centralization of the data processing function. The term "centralization" may refer to equipment and operations, or may be limited to a merger of some of the following ancillary functions:

- Strategic planning.
- Personnel selection and education.
- Measurement and evaluation.
- Systems support.
- Consulting and applications support.
- Charging, accounting, and control.

When operational consolidation is indicated, the concentration of systems in fewer—but larger—centers may be required. Alternatively, consolidation may be accomplished by interconnecting the existing centers in a suitably designed network.

advantages of centralization

This section concentrates on reasons in favor of physically consolidating general-purpose computing centers. (The book, of which this paper is a chapter, extends the analysis to include computer networks.) Let us first consider some of the advantages of centralization.

Economies of scale are possible with adequate processing volume. The larger and more cost-effective systems required may result in reduced cost per computation (provided the larger systems can be obtained).

Other economies are possible through reductions in record storage duplication and program preparation and maintenance. Site preparation and protection costs may similarly be reduced, since fewer sites are involved.

Fuller utilization of processing capability may result from the assignment of priorities over a larger and more diverse population of users and offer better opportunities for around-the-clock utilization. For example, engineering and research demands tend to peak on the first shift, whereas manufacturing and administrative demands frequently peak during the off-shifts. Operation costs for the third and fourth shifts are reduced, relative to benefits realized, in a large multisystem installation.

Certain personnel efficiencies may be possible by concentrating skilled programmers and technicians at a central site, thus making more effective use of their talents. A larger operation may

appeal more to highly qualified computer specialists because of broader career opportunities.

Improved quality of services is the result of reduced mean and variance of turnaround time in larger centers, as we shall demonstrate later in this paper. Also, a greater variety of services and programs can be offered to users of larger computing centers. As previously indicated, a larger and more expert pool of consultant services is available. There is less disruption to a computer user who transfers from one location to another when both sites use the same computer facility.

Integration of other functions, such as many administrative and technical services, may be considered for consolidation after a data communication network and common computational procedures are in place. Increasingly, there are functional and managerial advantages in centralizing a company's data base.

Some of the advantages of centralization just cited are difficult to quantify. It is possible, however, to evaluate several major considerations, at least to a first approximation.

economies of scale

In most cases, the primary motivation for consolidation of computing resources is to realize economies of scale. It has long been noted that—up to a point—larger units tend to be more efficient in producing and distributing goods and services than smaller units. For example, operations research procedures have been developed to determine an appropriate number and location of manufacturing plants[1] or warehouses,[2] taking into account their economies of scale. I say "appropriate" since optimal-determining algorithms have not been developed for most practical situations, although heuristic procedures have been found that apparently yield near-optimal solutions.

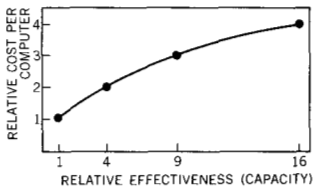
In the case of computing equipment, an economy of scale has been observed, wherein system effectiveness (E) is a quadratic function of system cost (C). This quadratic effect of computer system scale may be expressed as follows:

$$E = KC^2 \qquad (1)$$

where K is a constant of proportionality between the selected measure of effectiveness (performance, speed, throughput, etc.) and cost.

Today, the relationship cannot be so simply conceived. For example, throughput increases may be more directly related to larger and faster storage and data channels. On the other hand, personnel costs, which show very steep economies of scale,[3] are absorbing a larger fraction of the total equipment operation costs. The user may select the parameters that he believes most

Figure 1 Quadratic effect of scale



aptly characterize his system. The observed effect of scale may be somewhat greater or somewhat less than quadratic. For the analysis in this paper, we assume a quadratic effect as shown in Figure 1 as an example expression. Illustrative of the quadratic relationships is a four-fold increase in effectiveness for a doubling of cost. A review of this subject—including several attempts at corroboration with third-generation price-performance data—is given by Sharpe.[4]

Economy of scale is most obvious with respect to equipment costs, but it also obtains for other components of the total operational cost including floor space costs, number of operational and support personnel, and number of software packages to be maintained.

Figure 2 shows the quadratic economies of scale on the cost of executing 16 units of computational workload in a given time on various computer sizes, ranging from a single 16-unit-capacity computer to 16 single-unit computers.

Analytically, this is expressed as $K_0 N^{1/2}$, where N is the number of installations and K_0 is the constant of proportionality between costs and multiplicity of installations. The analysis assumes a single computer system per installation.

duplication of data base maintenance

Duplication of data base maintenance is another efficiency consideration that increasingly encourages the consolidation of computing services. In the trend toward data-base systems, references are often made by many application programs to a common information pool or data base. Thus, multiple-installation data-base systems, frequently maintain multiple copies of the data base and must transmit modifications of the data base reciprocally.

In the limiting case, where all installations communicate directly with one another—as in a fully connected network—the total number of interinstallation *communication links* is given by the following formula:

$$L = \frac{N(N-1)}{2} \tag{2}$$

Here, N is the number of installations, and L is the number of interconnecting links. Each file modification must be transmitted to the $N-1$ other installations. Therefore, the total volume of interinstallation communication increases linearly with N.

Similarly, the volume of *traffic per link* decreases with increasing number of interconnected installations N according to the following formula:

$$\text{Traffic per link} = \frac{\text{total traffic}}{\text{total links}} = \frac{N-1}{N(N-1)/2} = \frac{2}{N} \qquad (3)$$

Economies of scale similarly apply to communication charges. Therefore, if we assume the quadratic relation (1) to apply to the cost of *interinstallation traffic*, we obtain the following relationship:

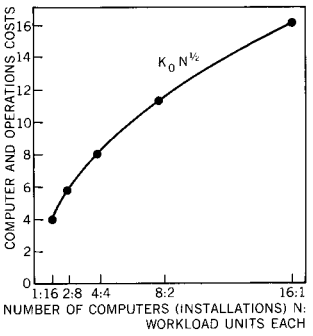
$$\text{Relative cost per link} = K_1 \sqrt{2/N} \qquad (4)$$

Where K_1 is a constant of proportionality. The *total cost of the interinstallation communications* is given as follows:

$$\begin{aligned}\text{Total linkage cost} &= \text{relative cost per link} \times \text{total links} \\ &= K_1 \sqrt{2/N}\,[N(N-1)/2] \\ &= K_1 \sqrt{N/2}\,(N-1) \qquad (5)\end{aligned}$$

The relations in Equations (2) through (5) are shown as functions of N in Figure 3.

Figure 2 Quadratic economies of scale for assumed 16 units of workload



Thus the quadratic effects of relative costs of computing equipment and operations and interinstallation communication costs, respectively, as a function of the number of installations, tend to motivate consolidation. If these were the only factors involved, the motivation to centralize the computing resources of an organization down to a single large installation would be very strong indeed. In the following section, we consider some of the countervailing arguments.

The tendency toward decentralization

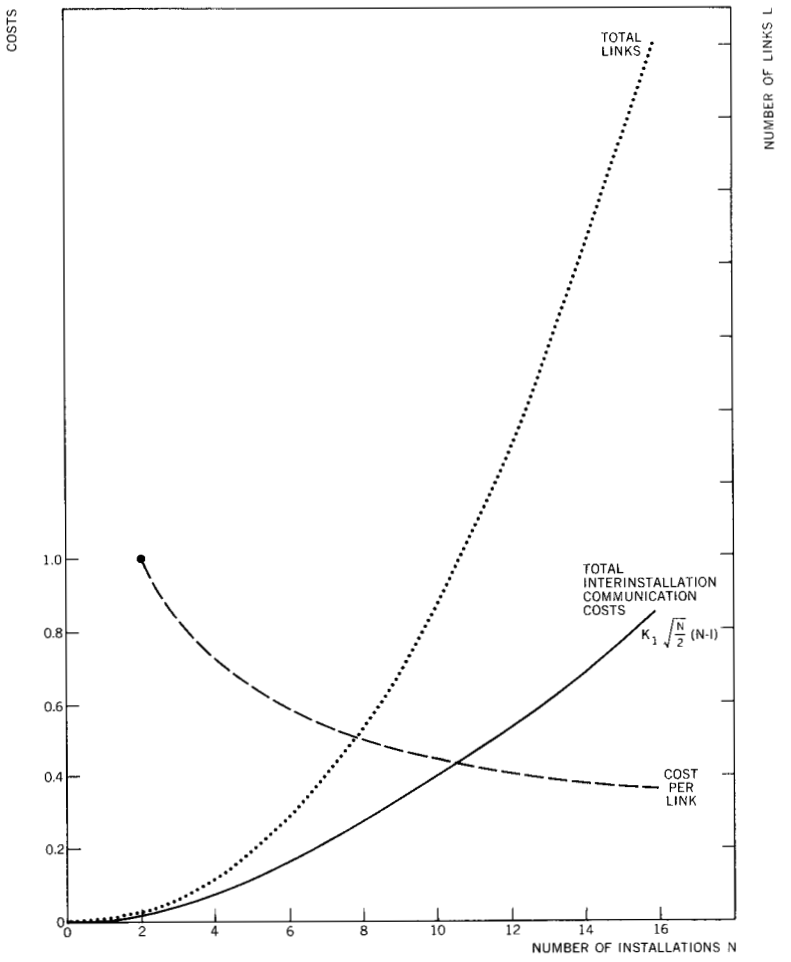
Let us begin by reviewing some of the arguments for decentralized computational operations. The following ideas are often advanced as advantages of decentralization.

Greater interest and motivation at local levels combined with greater knowledge of local conditions, are often said to produce information of higher quality and value. This is believed to be advantageous even though the unit processing costs may be higher.

Decentralization is also believed to permit tailoring to local requirements. The system standardization typically required for centralized processing may not be equally suitable for all divisions. With decentralization, special programs and services can be tailored to meet differing divisional needs.

Flexibility in coping with crises or changes in plan is more easily managed locally. When local management is in control of its computers, it may be able to take more immediate or preemptive action in reallocating resources than would a centralized service.

Figure 3 Relationships among costs and links for a range of network sizes



With regard to operations personnel in a decentralized data processing organization, there may be benefits from a feeling of identification with the mission of the functional division to which center employees belong, rather than to a service organization. Data processing personnel may also have more opportunities to communicate with, and transfer into, the line operations of the division, e.g., sales, research and development, or management.

In a decentralized data processing organization, certain communication costs, errors, and interruptions are avoided. These communications are those that occur between users and computers, rather than the interinstallation communications previously discussed. Arguments for retaining some decentralized computers include psychological and showcase effects, and hands-on requirements for educational and testing purposes.

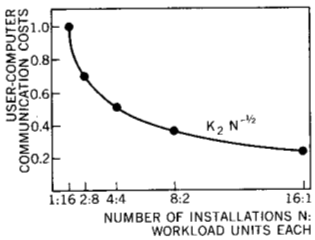
These considerations, as in those favoring centralization described earlier, involve many intangible factors that must be

considered in making the final decisions. Again, however, it is helpful to make approximate quantifications of the major factors so as to provide aid in the final, more comprehensive, decision-making process.

user-computer communication cost

The most obvious consequence of increased centralization is that the average distance between user and computer tends to increase. Data communication costs and problems increase commensurately. In this context, we continue the idealized model of providing 16 units of computational service under various options of physical consolidation. Again, consider the cases of 16 units of service provided in various ratios of installations (N) to units of workload.

Figure 4 User-computer communication costs



User-computer communication costs for such a distribution are shown in Figure 4. We may ignore the irregularities of subregion shape that result from some multiples and assume that users and computers are uniformly distributed geographically. Under these conditions, it is evident that the average user-computer distance varies inversely with the square root of the number of installations. The user-computer communication cost curve in Figure 4 assumes costs to be a linear function of distance.

cost of service interruption

The cost of service interruption provides an argument against putting all of an organization's computing capability into a single system or a single installation. The loss caused by a disruption of computing services varies according to the nature of the application, the duration of the interruption, the time of day, and the amount of warning provided. The trend, however, is that the cost of service interruptions is becoming a more dominant consideration as users become more dependent on computer services—especially with the growth of conversational and real-time usage.

For the first approximation analysis, assume that service interruptions at the various installations are mutually and statistically independent, and that the cost of service interruption is proportional to the probability that all systems are disabled. Therefore:

Cost of service interruption $= K_3 (P)^N$

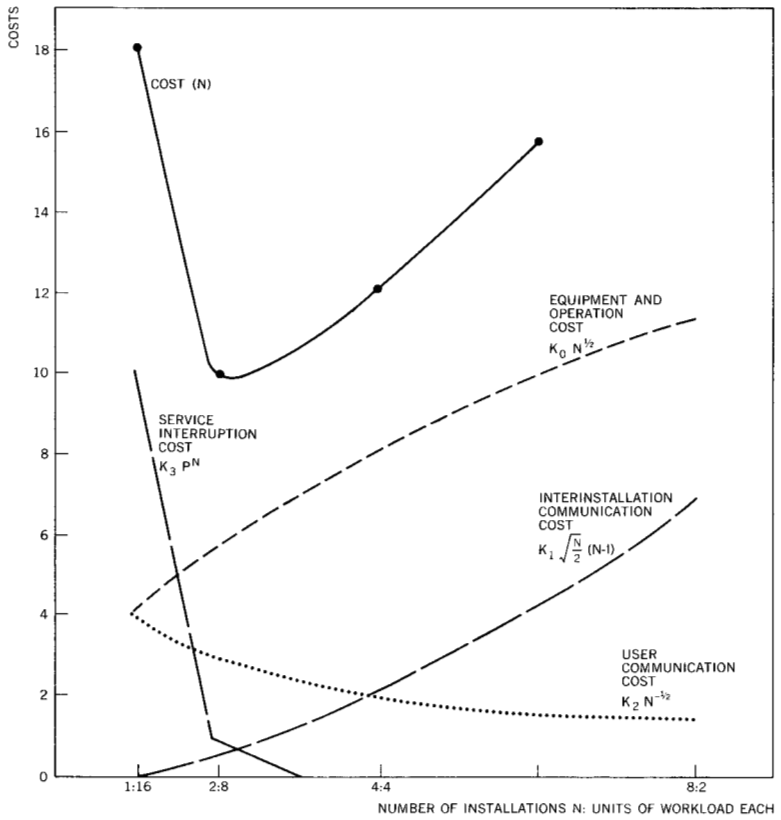
where P is the probability that the system will be disabled, and K_3 is a constant of proportionality.

centralization or dispersion costs

We now define a *cost of centralization for N installations, Cost* (N), as consisting of the sum of equipment and operations costs, interinstallation communications costs, user-computer communications costs, and cost of service interruptions as follows:

$$Cost\ (N) = K_0 N^{1/2} + K_1 \sqrt{\frac{N}{2}}\,(N-1) + K_2 N^{-1/2} + K_3 (P)^N \qquad (6)$$

Figure 5 Minimization of centralization or dispersion costs



A hypothetical case is given in Figure 5 that shows *Cost* (N) and its components for $K_0 = K_1 = 4$, $K_2 = 0.5$, and $K_3 = 100$. Also assumed is a service interruption probability $P = 0.1$. These values of K_0, K_1, K_2, and K_3 are typical of computing centers, and they have been chosen here for illustrative purposes. The analysis for an actual organization is based on analysis of costs at the existing level of centralization. In the hypothetical case (under the assumptions given), the optimal number of installations is seen to be $N = 2$.

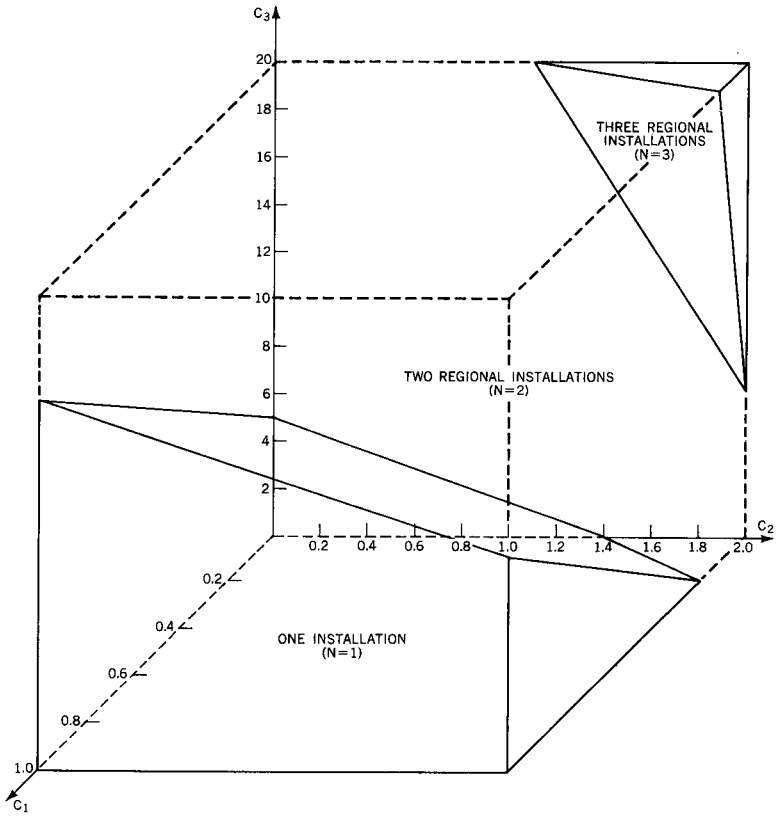
To examine the effect of the relative magnitude of K_0, K_1, K_2, K_3, on the optimal number of installations N, let us rewrite Equation 6 as follows:

$$Cost\ (N) = K_0 \left[N^{1/2} + \frac{K_1}{K_0} \sqrt{\frac{N}{2}}\ (N-1) + \frac{K_2}{K_0} N^{-1/2} + \frac{K_3}{K_0}\ (P)^N \right]$$

$$= K_0 \left[N^{1/2} + C_1 \sqrt{\frac{N}{2}}\ (N-1) + C_2 N^{-1/2} + C_3\ (P)^N \right]$$

Now we can study regions in a three-dimensional space of the parameters C_1, C_2, C_3, where

Figure 6 Relative cost space for centralization-dispersion decision



$C_1 = \frac{K_1}{K_0}$ = ratio of interinstallation communication to computing costs

$C_2 = \frac{K_2}{K_0}$ = ratio of user-communication to computing costs

$C_3 = \frac{K_3}{K_0}$ = ratio of service interruption to computing costs (sometimes referred to as availability insurance factor)

Figure 6 shows the volume of this C-parameter space where

$0 < C_1 < 1$, $0 < C_2 < 2$, and $0 < C_3 < 20$.

This range of parameters has been chosen to include the feasible relative costs of providing general-purpose, nonmilitary computing services over a region of about one thousand miles radius, given current U.S. computer, personnel, and communication costs.

Within this feasible space can be seen three regions. These are the regions in which one, two, and three installations, respectively, comprise the optimal solution to the cost equation.

Using the conditions and analysis given in this paper, then, one would base his decision to consolidate into one, two, or three regional installations on the following observations from Figure 6:

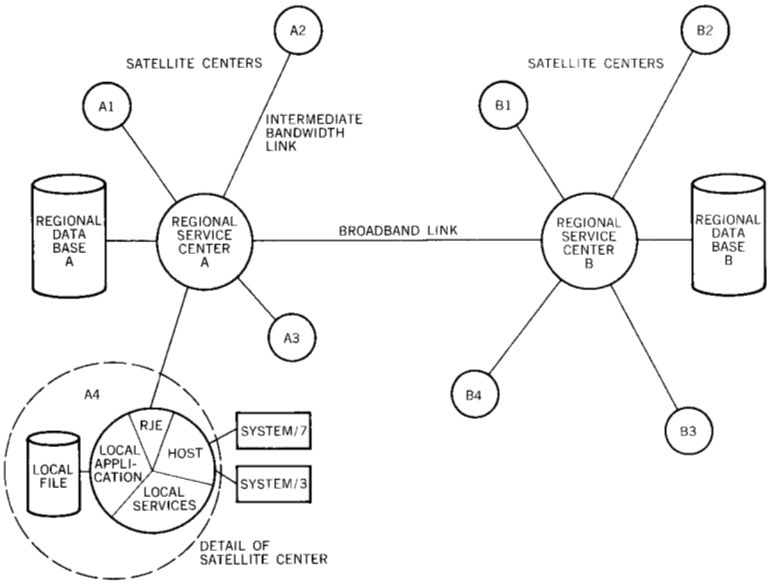
- In the vicinity the origin, equipment and operating costs dominate, thus suggesting the optimal strategy of a single installation ($N = 1$) for all cases.
- Higher values of C_1 (higher interinstallation communication costs) increase the desirability of single-installation operation ($N = 1$).
- Higher values of C_2 (higher user-computer communication costs) and/or higher values of C_3 (higher service interruption costs) increase the desirability of multiple installations.
- No economic basis is indicated for operating more than three general-purpose computing installations ($N = 3$) within an organization in a region about half the size of the United States, under the simplifying assumptions and limitations of this analysis. (The assumption that economies of scale still prevail at this level of centralization should be verified in each case).

The following section outlines several refinements to the *Cost* (N) model and indicates the general effect of these modifications on our conclusions. Before proceeding, however, let it again be noted that, even with the subsequent refinements, our study represents only a first approximation analysis of some of the factors. Clearly, other subjective factors must also be weighed in any particular case. It should also be understood that a reduction in the number of general-purpose centralized installations does not preclude the existence of many limited function local facilities. Perhaps the following hypothetical example of a satellite-central network might clarify this distinction.

example

A given company has an autonomous, general-purpose computing installation at each of ten locations. Each installation is responsible for satisfying the computing needs of all employees at its particular location. The following strategy for regional integration of computing services is proposed. (1) Concentrate widely used, general-purpose services and data bases at one, two, or three regional service centers as indicated in the foregoing analysis. The centers are connected by broadband links as shown in Figure 7. Standard production services, such as batch, time sharing (TSO), information management (IMS), and text processing (ATS) may be provided from such centers. (2) Maintain processing capability and file residence at each of the original locations as required to satisfy the following needs:

Figure 7 Regionalization of computing services



- To send and receive work to the large regional centers via Remote Job Entry (RJE). These locations are thus satellite to one or more of the large regional centers.

- To provide services and major applications used only at one site or which are in a state of development. Also to provide host support for local intelligent terminals and controllers.

- To provide host support for local intelligent terminals and controllers.

Such a division of computing service responsibility permits an organization to enjoy many of the advantages of centralization, while at the same time avoiding many of the disadvantages that sometimes follow complete consolidation.

Centralization-dispersion effects on service quality

Previous sections have concentrated on costs of centralization or dispersion of computing facilities. Let us now examine some effects on the quality of the services provided, using system turnaround time as the indicator of performance. The effect of scaling on system turnaround time is well known in queuing theory but otherwise not widely appreciated.

The scaling effect in queuing can be described briefly as follows. Assume customers arrive randomly at a service facility with an average *arrival rate* of λ customers per hour. Also, the system

has the capacity to service customers at an average *service rate* of μ customers per hour (with $\mu > \lambda$, otherwise the waiting line grows indefinitely). Under these conditions a waiting line develops, and an average waiting time T_W is experienced by customers before they receive service.

queuing model

The scaling effect is such that doubling both the service rate ($\mu \rightarrow 2\mu$) and the arrival rate ($\lambda \rightarrow 2\lambda$) results in reducing the waiting time by half. This effect is a factor to be considered in studies of potential centralization, if actual computer service systems behave as does this simple queuing model of turnaround time. Therefore, let us examine the model and its underlying assumptions, which are those generally made in elementary queuing theory.[5, 6]

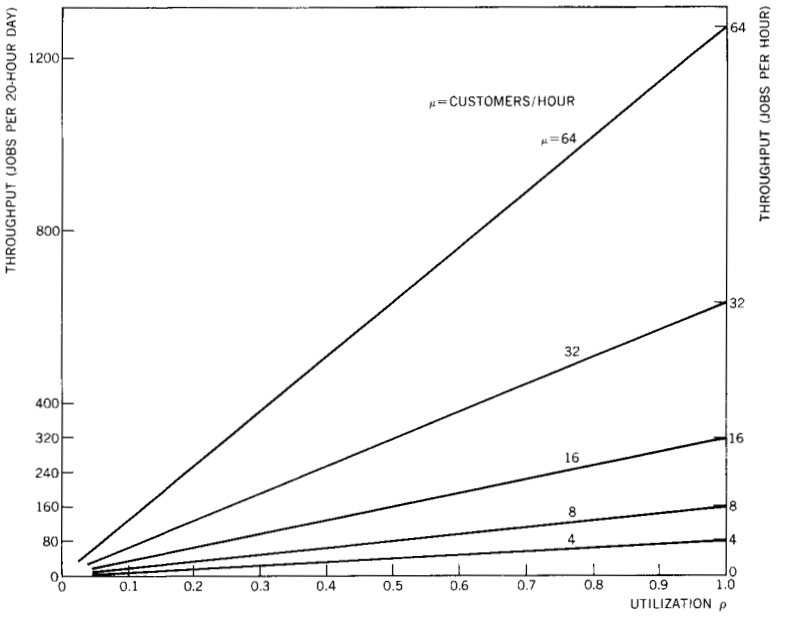
The arrival traffic is assumed to be Poissonian, which means that the arrival of a customer is independent of past events and has no influence on future arrivals. Also, simultaneous arrivals are freak events with a negligably small chance of happening. This assumption rules out situations where customers arrive either with periodic regularity, or simultaneously in large batches. It is generally agreed that this assumption describes quite well the situation in most scientific computing facilities in which jobs arrive from many customers acting independently. Because of their lack of dependence, arrivals with Poisson distribution are often called random inputs.

The traffic load does not change with time. Although this assumption does not preclude our analyzing a system with time-varying loading, it does mean that we are performing a steady-state analysis. Therefore, we cannot obtain a movie of transient effects after a change in arrival rate. However, we can obtain a picture of the system status during a time interval that is long enough for the arrival rate to be considered a stationary quantity.

The service times vary according to the exponential probability distribution. This assumption is similar to the assumption of Poisson arrivals in that it requires the service time to fluctuate randomly from customer to customer. The assumption is also one of convenience, for it leads to great simplification in the form of the solution. If computer service times are, in fact, distributed according to some other density function, a suitable correction can often be applied after the analysis based on the random-service-time assumption. In our simple case, in which we seek only the mean of the waiting times, the results depend only on the mean and variance of service times. Therefore, the exponential service time assumption need not hold.

Proceeding now on these assumptions, consider again the case of a single queue with Poisson arrivals and a single server with

Figure 8 Relationships among utilization, service rate, and throughput



exponential service times. The queue is serviced on a first-come-first-served basis. If the queue is not empty, the server finishes service of one customer and immediately begins to service the next. Under these conditions,

$$\text{Service rate } \mu = \frac{1}{T_S}$$

and

$$\text{Utilization } \rho = \frac{\lambda}{\mu}$$

If, for example,

$\lambda = 3.2$ customers per hour

and

$\mu = 4.0$ customers per hour

then

$$\rho = \frac{3.2}{4.0} = 0.8$$

Figure 8 shows the relation between ρ, μ, and throughput.

The mean leangth of the waiting line is as follows:

$$L_W = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1 - \rho}$$

Figure 9 Effect of utilization on waiting line length

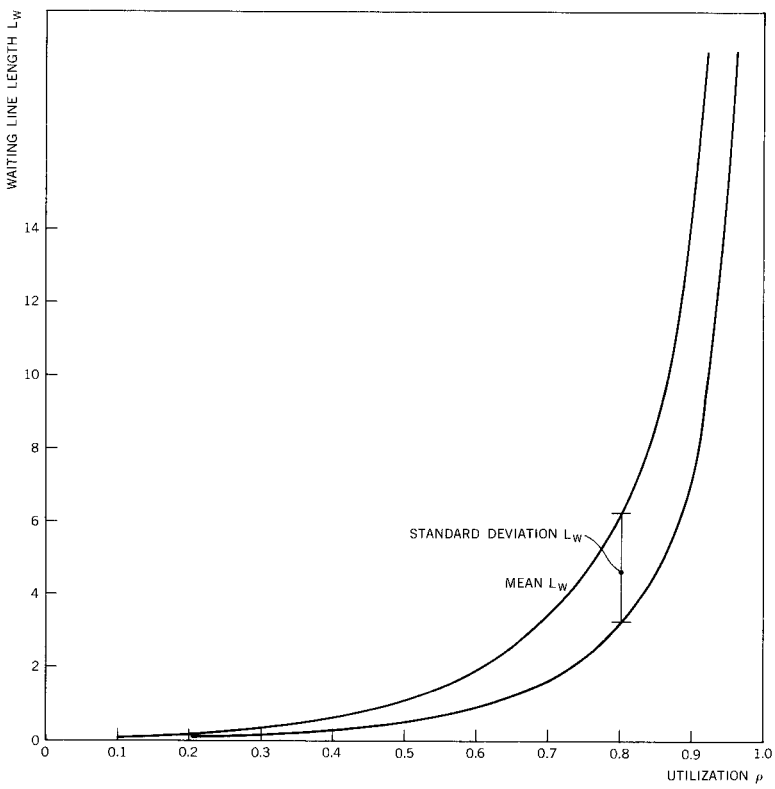


Figure 9 shows the effect of utilization on line length. The mean and variance of line length increase sharply as utilization approaches one hundred percent. The mean waiting time is

$$T_W = \frac{L_W}{\lambda} = \frac{\lambda}{\mu(\mu - \lambda)}$$

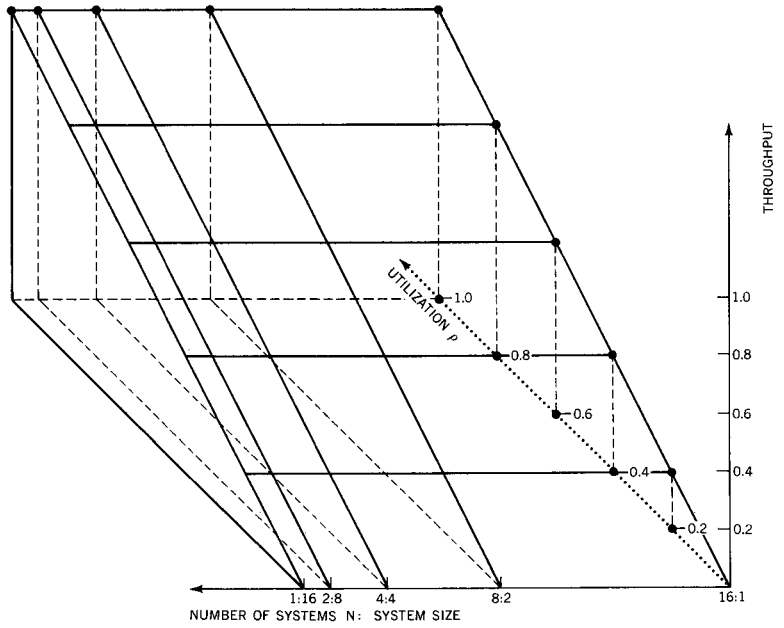
Now suppose we use a larger capacity service system, i.e., the service rate is $m\mu$ customers per second. Substituting these rates in the Equations 7 and 8 we obtain the following new waiting line length:

$$(L_W)_m = \frac{m^2\lambda^2}{m\mu(m\mu - m\lambda)} = \frac{\lambda^2}{\mu(\mu - \lambda)} = L_W$$

The mean line length does not change as the service and arrival rates are scaled up proportionately. However, the waiting time varies inversely with the system capacity indicated by the service rate $m\mu$ as follows:

$$(T_W)_m = \frac{L_W}{m\lambda} = \frac{m\lambda}{m\mu(m\mu - m\lambda)} = \frac{1}{m}(T_W)$$

Figure 10 Throughput as a linear function of utilization

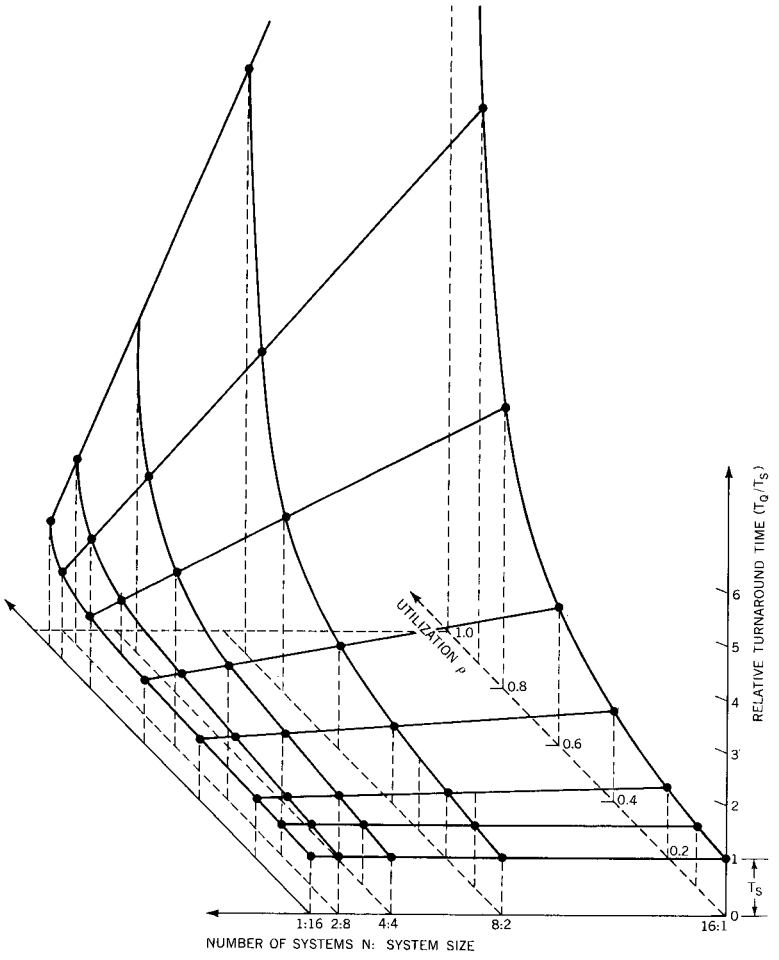


This can also be understood intuitively since, despite the fact the line length remains the same, it moves faster—i.e., in direct proportion to system capacity.

Now let us apply these results to the hypothetical situation used earlier in this paper that involves 16 units of computational workload to be executed on systems of various sizes. Figure 10 combines the relationships between system utilization and throughput illustrated in Figure 8 with the concept of system size and number of systems wherein each combination has the capacity of 16 units of work. Throughput is shown as a linear function of the utilization in each case.

Figure 11, however, shows the effect of system size on the relative turnaround time, T_Q/T_S, which is the ratio of the turnaround time $T_Q = T_W + T_S$ to the service time T_S. This effect is quite dramatic as utilization increases. One way of interpreting Figure 10 is that, if all processors are loaded to a certain utilization, say $\rho = 0.8$, the turnaround time is significantly greater when smaller systems are used. This penalty is a "cost of compartmentalization," extracted as a result of users being constrained in their access to resources. Another way of viewing this phenomenon is to draw contours of equal turnaround time on the surface of Figure 11. These contours are shown in a plan view of the surface in Figure 12. Note the increased utilization achievable in larger systems, at comparable system responsiveness.

Figure 11 Effect of system size on relative turnaround time

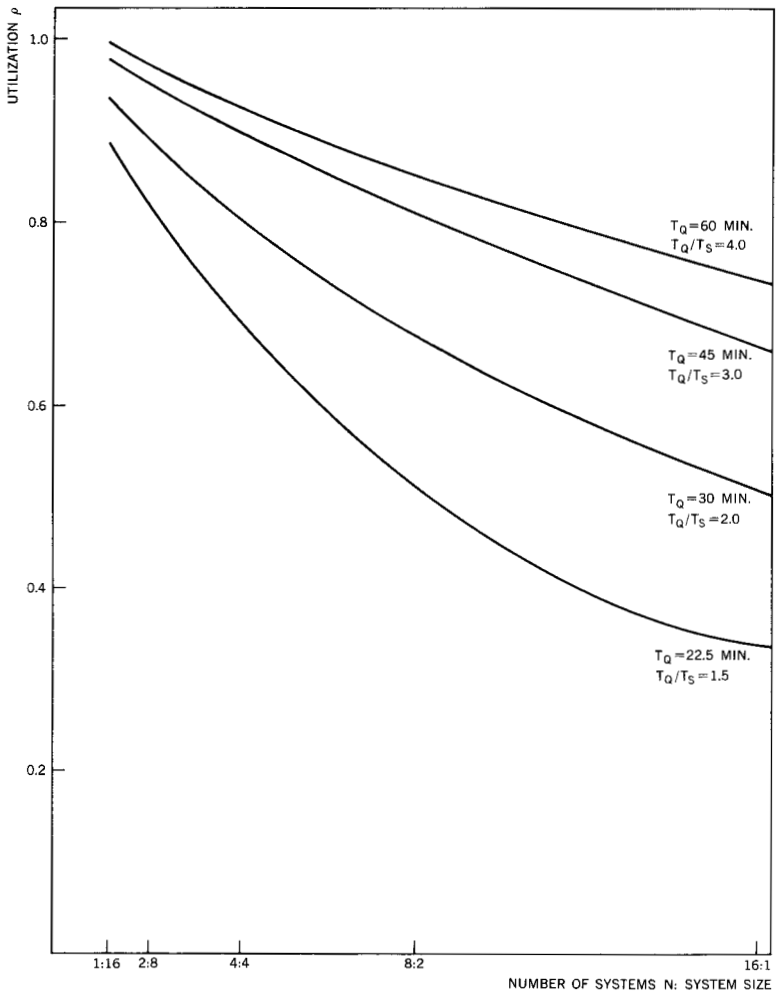


If a relative value is associated with the computer service as a function of turnaround time,[6] then the surface shown in Figure 13 can be obtained. The difference between the value obtainable at the optimal condition of centralization and the existing operating point can be viewed as a "cost of compartmentalization." This is a cost that results from barriers within the organization that restrict access to the total resources of the organization. This analysis permits the determination of a cost of the quality of service that can be used, in conjunction with the previously defined costs to guide decisions of centralization versus dispersion.

Concluding remarks

We have presented a number of considerations that are pertinent to the degree of centralization or dispersion of computing facilities within an organization. We have argued that costs can be

Figure 12 Contours of equal turnaround time

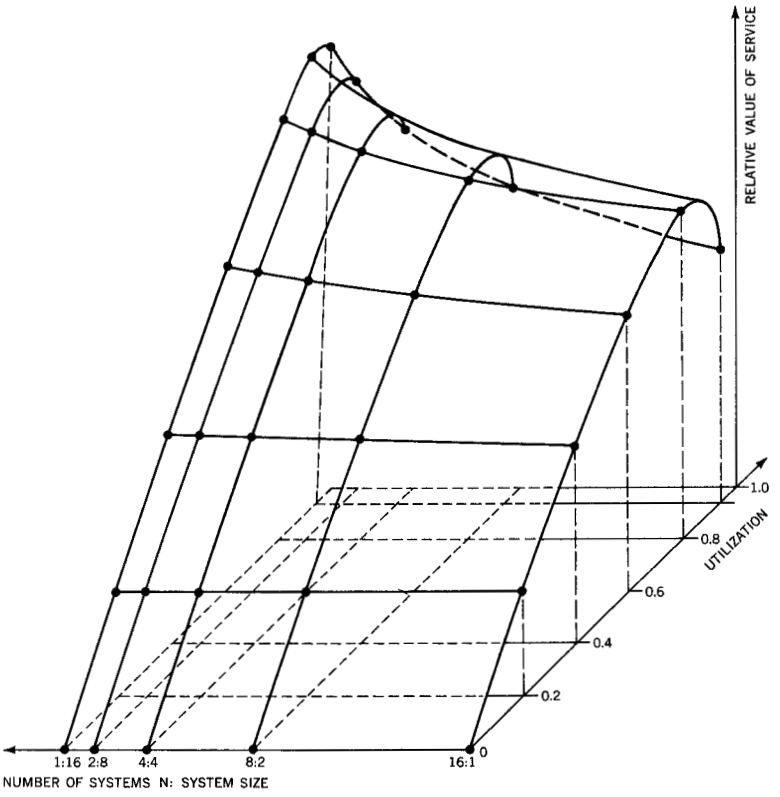


associated with the major factors involved so that effects of proposed changes in dispersion can be assessed quantitatively, if only approximately.

A dimensionless analysis has been used to demonstrate the economic feasibility of providing general production computing services over a large geographic region with relatively few large computer centers.

A strategy for regional integration of computing services has been proposed, the goal of which is to allow economies of scale and other advantages of centralization, without losing the variety and quality of services provided by a small local facility. The success of this strategy depends on making a distinction be-

Figure 13 Relative value of computer service



tween standard production services that can be adequately delivered from a large remote center, and locally anomalous personalized or evolving services that may still be better provided on site.

This division of workload permits the proper exploitation of two different types of operating systems. The first, as exemplified by OS/360, is well suited for production and supplying of standard, stable services and applications. The second type, as exemplified by CP/67 or VMS/370, provides diversity, protective application autonomy, and facilities for growth.

The effect of centralization on service responsiveness was determined and used in conjunction with the notion of value of responsive services to determine another cost—the "cost of compartmentalization." This is represented as the cost of barriers within an organization that prevent free access to the resources of that organization.

CITED REFERENCES

1. A. S. Manne, "Plant location under economies of scale-recentralization and computation," *Management Science* **11,** No. 2, 213–235 (November 1964).
2. E. Feldman, F. A. Lehrer, and T. L. Ray, "Warehouse location under continuous economies of scale," *Management Science* **12,** No. 9, 670–684 (May 1966).
3. L. L. Selwyn, "Competition and structure in the computer service industry," *Proceedings of the Second Annual ACM SIGCOSIM Symposium,* 48–56, Gaithersburg, Maryland (October 26, 1971).
4. W. F. Sharpe, *The Economics of Computers,* 314–322, Columbia University Press, New York, New York (1964).
5. *Analysis of Some Queuing Models in Real-Time Systems,* Form No. GF20-0007, IBM Corporation, Data Processing Division, White Plains, New York.
6. D. N. Streeter, "Cost-benefit evaluation of scientific computing services," *IBM Systems Journal* **11,** No. 3, 219–233 (1972).