INTRODUCTION

The use of prices as a mechanism for allocating resources is generally well understood. Nearly two hundred years have elapsed since Adam Smith, in The Wealth of Nations, discussed the functioning of the "invisible hand" in a market economy, but the principles which he enunciated have not been altered or invalidated by ensuing generations of economists. In the United States today, markets are the dominant economic form, and the price system is used to allocate nearly all the product of our private sector, over 75% of gross national product. (Governments also participate in markets, at least to the extent of obtaining resources.) To be sure, the price mechanism does not always work as well in real markets as in theory—a defect shared by other allocative mechanisms—and certain categories of goods and services continue to be allocated by means other than prices. Among these goods are most of those produced by the several levels of governments and nonprofit organizations such as universities. In addition, non-price allocation techniques are frequently used in instances when market allocation violates social canons of equity, for example during periods of rationing in wartime.

Prices are not the dominant allocative mechanism for computer time, and our purpose in this paper is to examine whether they should generally be used. We will need first to establish the general conditions under which pricing will be efficacious, and then to determine whether the allocation of computer time satisfies those conditions. Finally, we will examine some of the allocative techniques used instead of prices, to see whether non-price methods can be expected to work as well as prices on this particular problem.

Why use prices at all?

Any economic system must solve the problem of how to use scarce resources. Most resources can be used to produce many goods; most products are useful to many consumers. Due to the scarcity of resources, however, the system is closed: resources used by one producer are not available to any other and goods consumed by one person reduce the total consumption possibilities of all others. Some determination must therefore be made of the preferences of different economic units for the same economic resources or products.

The price system is the vehicle by which economic units express their preferences in a market context. It is axiomatic that the preferences which underlie prices cannot be measured other than through the behavior of buyers. In other words, if a consumer buys a shirt for $10, it is possible to infer only that he derives at least $10 worth of satisfaction from that shirt. It is not generally possible to measure that satisfaction in any way other than as $10 worth.

If preferences are expressed predominantly by prices, then some other desiderata can be obtained within market resource allocation. One such desideratum is that those consumers who place the greatest value on a good are the ones who obtain it (assuming an equitable distribution of income). In a properly functioning market, prices
will be bid up (or down) to the point at which the available supply is allocated to the consumers willing to pay the highest prices.

Implicit in this approach is that the value to society of the goods is the maximum price paid for the goods by any consumer. Prices normally cover costs of production (including profits), but in some instances particular allocations of resources impose costs (or benefits) on third parties other than the producer and consumer. For example, the location of a junkyard may lower surrounding property values, thus imposing costs on persons other than the owner of the junkyard and his customers. The social value of the junkyard will be less than its private value (to the owner) unless the owner is required to charge prices high enough to pay compensation to surrounding property owners. Prices high enough to compensate for external costs are efficient (in the technical sense that no other allocation of resources will produce output with a greater net value). In general, external costs can be included in a price-system allocation when the third parties can enter into the market determination of price. Typically, the goods whose production imposes unrecognized or uncompensated external costs, will be underpriced and over-supplied compared to the amount of production which is socially optimal.

Not only does a price system establish a priority of users, it also establishes a priority of wants. The producer willing to pay the highest price for a resource will be the one whose product is valued most highly by the members of society (adjusting for differences in the productivity of resources). Those wants to which the members of society give priority will be able to obtain the resources necessary for their satisfaction—and they will be satisfied only to the extent that no other want receives greater priority.

Finally, a price system can solve the problem of allocating resources dynamically. It does so by transmitting information about consumers' demands to producers, and offering consumers information about the cost of satisfying different wants. If consumers express great demand for resources away from less profitable locations, the producer an incentive to increase production, bidding resources away from less profitable allocations. Over a period of time, labor and capital will be drawn to produce a particular goods only if demand for the goods is high enough to cover the cost of the additional resources.

In summary, a price system can solve the problems of distributing goods and services among consumers, allocating resources among producers, and conveying information to determine the flow of resources among different allocations over time. If prices are to solve these problems, they must be able to respond in certain ways. When demand rises or falls, prices must rise or fall sufficiently to ration the available supply. Over time, changes in supply must affect the price of any goods; that is, constancy of price over time is neither necessary nor desirable for efficient resource allocation. And since the role of price is to allocate resources, the price at any point in time need not bear any relation to the cost of production at that time. This point should be stressed: prices are a rationing device, not a mechanism for recovering cost. If demand for a good is low, its price may well fall below cost, transmitting information to the producer that demand is inadequate. Unless price is permitted to fall below cost, the proper information about demand may never be obtained, and the allocation of resources can never adjust properly to the unprofitability of that good.

Can computer time be priced?

Two aspects of the question of pricing computer time must be discussed. The first is whether prices can allocate computer time of a fixed quantity and productivity (as determined by machine configurations at any time). The second is the extent to which pricing can allocate computer time over periods in which demand patterns and machine productivity both change.

As noted above, the conditions for pricing to work are fairly simple: users must be unable to obtain any scarce resource at a zero price; social and private direct benefits must be identical, so that neither benefits nor costs are incurred except by the buyer and seller; and prices must be free to vary without regard to cost of production. (These are rigorous requirements; in fact, pricing will usually work if these requirements are satisfied loosely.) Applied to computers, the first condition requires that computer time be made available to any user willing (and able) to pay the price. The price may be stated in dollars, as at computer service bureaus, or it may be stated in terms of some flat money. It is essential, however, that the user feels that he is truly paying a price.
If the price is stated in dollars, then the user's budget must be limited. If the price is simply an accounting of computer time, then the user's overall access to the computer must be limited. Only if there is a budget constraint will the user have an incentive to evaluate the benefits of computer time relative to its cost to him and other users, for otherwise the cost will be zero.

The second condition requires that all costs and benefits be incurred within the market. We have noted elsewhere that a user of computer time imposes costs on all other users by increasing their turnaround times. The same point has been expressed by other authors, including Marchand and Greenberger (implicitly). The external costs imposed by one user on all others are no less important a category of cost than the direct use of the computer, and efficient allocation will result only if each user pays a price high enough (in principle) to compensate all other users for the reduction in quality of their service. This situation is by no means unique to the allocation of computer time. Every rush-hour commuter imposes external costs on all other users in the form of longer commuting times, and every attendee at a Broadway opening night imposes costs on those unable to obtain tickets in the form of reductions in their satisfactions. One difference between these two examples is that social policy has dictated a zero price for highway commuters, thus prohibiting the use of prices to eliminate congestion; whereas show ticket prices can fluctuate widely, and in fact a premium is charged for many opening nights. The problem of planning to eliminate or reduce congestion can be attacked through queuing theory (as by Kleinrock and Greenberger among many others), but it appears perfectly feasible to eliminate congestion to the extent desired by users by letting them bid up prices. Equilibrium will be reached where each user is paying a price which is greater than the total value of the decrements in service to all other affected users. Note that these external costs will arise only for a heavily-used computer on which each job seriously affects all other turnaround times. For a batch processor, such "capacity" use could occur for any turnaround time, or set of turnaround times. For a time-sharing system, the related notion of access time (to a console or to the central processing unit) could be used.

The third requirement for pricing to be effective is that prices be free to fluctuate without regard to the cost of production. In general, a good should be supplied to any consumer willing to pay the cost of the additional resources needed to produce it. In the case of computer time, the additional resources necessary for production are virtually free (at least within eight-hour blocks) and in any case are inexpensive relative to the amortization cost of the computer. Nonetheless, price is not a cost-recovery mechanism, and it is proper at any point in time to disregard completely the capital cost of a computer in setting price. If a computer is utilized at less than capacity, the external costs imposed by an additional user will normally be low, and the price of computer time should correspondingly be low so as not to discourage use of an essentially free facility. The extent to which amortized cost exceeds revenue is an operating loss, and should be treated in the same way as, for example, a loss experienced by a realtor on an apartment house with some unrented units.

We have been using the term "capacity" to mean the ability to process some maximum number of jobs (per time period) of given quality of service, or turnaround time. (There is, of course, no reason why the turnaround time used to estimate machine capacity should be that desired by users, or why all users should prefer the same turnaround times.) We conclude from our discussion that pricing is a feasible allocative device for a computer with fixed capacity in the short run, when demand is stable. The proper price will cover the marginal cost of operating the computer to satisfy the demand of the additional user. If the computer is so heavily utilized that each user imposes external costs on others by lengthening their turnaround times, the price must be high enough to cover these external costs, as well as other costs of using the computer.

In the long run, when demand and supply conditions are free to vary, the pricing problem becomes more complex. We have described elsewhere what we envision as the typical time pattern of efficient prices for computer use—a pattern in which prices initially are low, to encourage use of a large, fast, new computer, and eventually rise reflecting use at or near the capacity of the computer. Throughout this pricing cycle, minor investments may be made, adding to the capacity of the computer, and eventually a new main frame is procured. At each point in time, prices are set in accordance with the principles stated above. If the computer is an efficient investment, over-re-
covery of cost during periods of capacity use will just offset early losses. Of course, there is no requirement that cost recovery be exactly equal to operating loss, any more than any investment is required to yield benefits exactly equal to costs. Typically, some computers will prove to be efficient investments and others will not.

In this context, it is worth noting that we have previously expressed our view that idle time is not an indication of inefficient use of a computer, since capacity cannot be defined without reference to the quality of service. If users place a high value on short turnaround times, idle time will be valuable to users as a guarantee of high-quality (i.e., fast) service. In this case, users should be willing to pay a price for idle time at least equal to the value of that time to the excluded users. Our conclusion followed from explicit consideration of the quality of service as a variable to be set by the computer center. The same conclusion is reached from queuing-theoretic considerations by Araoz and Malmgren.5

"Some idle capacity can now be seen as a device of efficient production in many cases, and the 'excess capacity' which is often observed . . . may sometimes, if not often, arise out of a seeking of efficiency rather than restriction. This will be true insofar as there are any rigidities (or what an economist might call indivisibilities with respect to time) . . . which might give rise to a queuing process." (p. 209)

Do alternative techniques work?

We will discuss three types of non-price allocation mechanisms: average-costing, overhead charges, and various priority plans. Average-costing is used widely as an internal cost-recovery device, and is imposed externally in many cases by federal government accounting regulations. Its basis is that a facility should recover its costs in any time period, but while that basis is not unreasonable over the lifetime of a computer (or any other capital goods), serious misallocations result when the time period is chosen as a year or even less. We have discussed elsewhere the various abuses that may arise under average-costing, so they will only be listed here: (1) computer centers are unable to provide service to additional users at marginal (social) cost; (2) the resulting pattern of charges encourages use during peak periods and discourages use when the computer is idle; (3) over-investment is likely to result if contractors are guaranteed recovery of average cost; and (4) under federal government auditing regulations, funding of computers by foundations and other non-users is discouraged. Average-cost charging is a popular internal accounting device due to its ease of administration, but the incentives it offers are unlikely to promote efficient resource allocation.

When a facility is used widely and the cost of its services is difficult to impute to individual users, the facility is frequently called an "overhead expense" and its cost is then allocated to users on an arbitrary basis. When computers are treated as overhead, the full costs (including amortization) of the computer center are included in the firm's general overhead pool which is imputed to individual projects on a basis such as total labor costs, total man-hours, or total operating costs of each project. Some universities have used this method for recovering the costs of their computer centers.

It should be obvious, however, that overhead charges can offer the proper incentives to neither the user of the computer nor the administration concerned with supplying computer time. Each user will prefer to substitute computer time for other resources, thus reducing his basis for overhead charges, and the overall effect must be to increase the demand for computer time. (In addition, overhead charges will discriminate against projects which are not computer-intensive, thus creating an inequitable set of charges.) If each user substitutes the same ratio of computer time for direct charges, the result will be to leave the pattern of charges unchanged, but to bias upward total use of the computer. The supplier of computer time will then be misled into overestimating the demand for the computer, resulting in overinvestment in subsequent computer facilities. Finally, the overhead rates that the firm must charge (to sponsors, if it is a contractor, or to purchasers, if it is a producer) will be inflated by the misallocation of resources and overinvestment, leading eventually to declining revenues and reduced profits in the case of a firm.

Priority mechanisms have received wide attention in the literature on managerial and operations-research problems. In contrast, they have been virtually ignored by economists. The reason for the disinterest of economists appears to be that
priorities are simply a surrogate set of prices that may in some instances work as well as a true price mechanism but will almost never be superior. For their part, operations analysts seem unaware that priorities are a form of prices; thus, Kleinrock discusses "bribes" which are merely prices, and Greenberger tries to minimize the cost of delay, a cost which can never be known except in terms of the price users would pay to avoid the delay.

There are two distinct types of priority rules: one which determines the access pattern of a given set of users, and another which offers incentives to potential users in determining their demands for computer time. Some of the variants of the first type of rule are listed by Greenberger: first-come-first-served, the c/t rule (the next job served is that with highest waiting cost per service time), shortest-job-next, and round-robin scheduling. The difficulty with rules of this sort is that an implicit assumption must be made about the value placed on computer time by each user. For example, application of the c/t rule involves "judicious approximation"—in other words, arbitrary judgments about users' costs—unless all costs are equal and either constant or exponentially-discounted functions of time. In general, users will not value time equally, nor consider waiting equally costly, so such a rule will not allocate time properly. Moreover, the only way for the computer center to discover if all users value waiting and access equally is to ask them—that is, to establish a market in which preferences may be expressed. Priority rules of this type, therefore, are equivalent to prices under the special assumption that all users place the same values on computer time and experience the same costs for waiting. These conditions are, of course, most unlikely to be satisfied in practice, but under any other conditions priority rules will result in a misallocation relative to the one obtainable by pricing.

The second type of priority is the one which rations access to the computer according to the relative importance attached to the user's project. Such priorities are a form of artificial money, for a high priority in the absence of dollar prices has the same effect as a large project budget if prices are set in dollars. There are, however, two problems. First is the question of flexibility—are priorities reset regularly, as prices must be, reflecting changing conditions of user importance? If not, projects are likely to have either "too easy" or "too hard" a time obtaining access to the computer. Are priorities set on a sliding scale, reflecting the varying importance of successive quanta of computer time? The answer is usually that they are not variable, but that all users in a project receive the priority attached to that project.

Second, a priority system for computer access discourages efficient substitution of other resources for computer use. A project leader whose access to the computer depends on his project's priority will use computer time extensively if the priority is high, even if the computer is not the proper research tool. The computer administrator may react to pressures from projects by assigning more high priorities than the computer can accommodate. This situation is akin to a government's attempt to obtain resources by printing money: the result is to inflate all prices, or in this instance to downgrade the value of all priorities. There will be no incentive for any project to estimate the value of computer time, since the costs are not expressed in any systematic way and since trades of computer time for other resources generally cannot be made.

The one remaining advantage of priorities over prices is that they are inexpensive to administer. An efficient price system must provide periodic—if not continuous—variations in price in response to fluctuations in supply and demand. It might be expensive to "make a market," that is, to provide price information to suppliers and demanders. This is largely an empirical question, and well beyond the scope of this paper. One bit of evidence, however, indicates that the incremental cost of establishing queues of variable length with prices that fluctuate fairly often, given the existence of equipment for accounting for users' time, is of the order of no more than one per cent of the total cost of a batch processor. Even if the cost of pricing a time-sharing system were twice as great, it seems likely that the net gain in efficiency would be substantial. In any case, a price system which adjusted no more frequently than a priority system would be no more expensive.

**SUMMARY**

We have made several contentions which may appear radical in view of current costing procedures for computer systems. At the heart of our recommendations is the assertion that prices are not a mechanism for recovering cost, but are instead a device for allocating scarce resources and obtain-
ing the efficient level of investment over time. We believe that it is feasible to use prices to allocate computer time, whether the users are a firm's internal projects or its external customers. Prices are only now beginning to be used for batch processing, and it is too early to evaluate the difficulties in pricing time-sharing computers. But the conditions for pricing will be present in time-sharing systems—identification of output, costs to excluded users (either at the central processor or at the console), and scarcity of total resources—and so we expect prices to be a feasible allocative mechanism.

The role of prices is enhanced by the unsatisfactory nature of the alternatives. Average-cost pricing creates a perverse set of incentives for user and supplier. Overhead charges are even worse, for by themselves they establish no mechanism for allocating computer time, and they are likely to be inequitable. Priority rules are the least unsatisfactory alternative, but their desirable properties are the ones they share with prices. Priority systems are unlikely to obtain an allocation of computer time preferable to that of prices, or to cost less to administer.

REFERENCES

1 H KANTER A MOORE N SINGER
   The allocation of computer time at university computer centers
   Journal of Business 41 375 1968

2 M MARCHAND
   Priority pricing with respect to time-shared computers
   Working Paper No 247 Center for Research in Management
   Science Univ of California Berkeley

3 M GREENBERGER
   The priority problem and computer time sharing
   Management Science 12 888 1966

4 L KLEINROCK
   Optimum bribing for queue position
   Operations Research 15 304 1967

5 A B ARAOZ H B MALMGREN
   Congestion and idle capacity in our economy
   Review of Economic Studies 28 202 1961