Why Not the Best Allocation Scheme for Computing Resources in a University Environment? A Review of a Proposal

Julian Feldman

TECHNICAL REPORT #105

Department of Information and Computer Science, University of California, Irvine, CA 92717.

owners of fraction contracts will be able to use more than their fraction of ports and/or processor (p. 5).

SUMMARY

There is a growing movement towards share or subscription pricing of computing services. The Triangle Universities Computation Center has been the leader of the movement. Now other schools and military organizations have adopted the TUCC strategy with local variations. Einard Stefferud, a Huntington Beach consultant, has been involved in a number of these adoptions.

After several aborted proposals and an experiment in the Winter and Spring of 1975, the UCI Computing Facility offered a subscription pricing scheme under which a user or group of users would purchase a job slot on the PDP-10 or SIGMA 7 for $1,000 per month on an annual contract. The $1,000 per month charge covered all connect time, KCS, and I/O charges for the job running the contract slot. Subscription users paid for other services, e.g., disc storage, lines printed, at regular rates. Subscription pricing led to significant increases in computer usage. However, the 75-76 pricing scheme did not impose enough contraints on users. As a result the subscription users "wasted" resources and consumed an inordinate share of the resources. This led to the 76-77 and 77-78 contracts which provide a discount for annual contracts with a guaranteed monthly minimum.

The peak loading problem encountered in 75-76 and the growing demand for instructional computing have led to a reconsideration of subscription pricing schemes at UCI. This reconsideration has produced the most interesting proposal of its kind which has been made on campus. The essence of this proposal is that the users be offered the opportunity to purchase fractional shares of either the SIGMA 7 or DEC system 10. The minimum share would be 10%. The purchaser of an X% fraction of a system would be entitled to X% of the capacity of the computing system.

This review makes four criticismsof the proposed allocation scheme: (1) Other solutions to the problems are not considered. (2) The extent of the problem, e.g., the additional computing services required, is not specified. (3) The effects of increased utilization of the computing systems are not specified. (4) The proposed allocation scheme (a) requires the determination of capacity -- an extremely difficult task, (b) places the computing facility in a conflict situation, (c) does not contain an estimate of implemenation costs, and (d) represents a major change in the pricing system.

Two alternative recommendations are made. One is to modify the fractional share proposal to bring it closer to the TUCC scheme thereby obtaining the benefits of their experience. The second alternative is to increase utilization by lowering prices to University users.

In 1975-76, the University of California, Irvine, experienced a sharp increase in computer usage as a result of the introduction of a version of a subscription pricing scheme. For the years immediately preceding 75-76, connect time increased about 20% per year. In 75-76, connect time increased 50% and KCS went up 100% without any increase in enrollment or any other substantial change in demand factors aside from the change of pricing policy. As a result of the 75-76 increase in usage, there were some periods during which the computing systems were heavily loaded, i.e., some users could not obtain access to interactive ports and the users who were on experienced degradations in response. In addition, instructional users were forecasting increases in computing requirements.

So that he might take appropriate action on these problems--peak loading and forecasted increased in demand for IUC--the campus administrator responsible for computing, the Vice Chancellor for Academic Affairs, appointed a Resource Allocation Committee (RAC) to advise him on these two problems.

THE RAC RECOMMENDATIONS

In order to permit this paper to be relatively self contained, the RAC proposal is summarized in this section. The page numbers refer to the RAC report "Summary of Recommendations of the Resource Allocation Committee" dated 27 April 1977.

New Contract Form

A user or group of users can purchase fractions ofmachine capacity on the SIGMA 7 or DEC system 10. The minimum fraction is 10% of the capacity of the machine, and not more than 80% of the capacity of the machine will be sold on such contracts (p. 3).

An X% fraction entitles the owners of the fraction to a minimum of X% of the ports and X% of the processor. When usage falls below 80% of capacity, owners of fraction contracts will be able to use more than their fraction of ports and/or processor (p. 5).

New Procedure For Allocating IUC Funds

The Instructional Use of Computing (IUC) funds will be spent in the following manner: The PDP-11 would continue to be financed off the top of the IUC allocation (p. 10). This will cost about 10% of the IUC funds. The bulk of the IUC funds (80%) would be used to buy fractions of the SIGMA 7 and DEC System 10. Ten percent (10%) of the IUC monies would be used as a strategic reserve to buy additional required capacity for IUC users during peak periods (p. 4).

The computing services purchased with IUC funds will be allocated as follows: Five percent (5%) will be used to provide a minimum allocation to every UCI student and faculty member who asks for it (pp. 10-11). Thirty percent (30%) of the IUC funds will be allocated to schools on the basis of the previous year's usage. Thirty-five percent (35%) will be allocated to schools on the basis of enrollments in classes planning to utilize computing. 30% will be allocated to the Dean of Special Programs for development and contingencies (pp. 10-11).

The IUC funds will be spent in accordance with the following rules: IUC credits will be allocated one quarter at a time. Each quarter's allocation will be divided into weekly increments. If a user exhausts his/her weekly increment, the user must petition for more monies, presumably an advance on next week's increment. Only 50% of the week's balance is carried over to the next week (pp. 6-7).

Instructional users will be charged for services used on the basis of the Computing Facility's standard price list except when utilization of the IUC fraction falls below 60% at which point prices charged to IUC users woul go down on a sliding scale to zero when IUC usage falls below 40% of the fraction (p. 6).

To insure full utilization of the IUC fraction, the IUC allocation committee can create more credits at any time (p. 9).

"During anytime when more than 80% of the IUC ports on a machine are in use, there will be an automatic one hour time limit. After 50 minutes and after 55 minutes, the user receives warnings and is then logged off at 60 minutes. He is unable to get back on for five minutes, which assures that someone else will have a chance to use the port. If the port is still free after five minutes, he may log back on (pp. 8-9)."

A CRITIQUE OF THE RAC RECOMMENDATIONS

One. One criticism of the RAC recommendations is that the RAC did not consider responses to the problem of overloading other than the proposal of a new allocation scheme.

The RAC focussed its efforts on examining and analyzing alternative allocation schemes to the exclusion of hardware, software, and support alternatives. Given the recent history of the campus and the pressure in universities at this time to economize, it is not surprizing that the RAC's efforts were focussed on developing an allocation scheme. However, a more complete analysis of the problem must consider at least the other responses to the problem of an overloaded computing system.

The classic response to an overloaded computing system is to obtain more hardware. This is not always the least expensive way to solve the problem, but it has a high probability of solving the problem.

A second response is the acquisition or development of higher performance software. The history of computing is full of examples where software improvements have made dramatic changes in the capacity of a computing system. Some of the software improvements come from better algorithms and some come from specialization. But experience clearly indicates that software can make a dramatic difference. It is, of course, much safer to buy ready-made software than embark on a software development project.

A third response is the provision of additional staff support to educate and help users. Many users are naive. They don't know all of the software/hardware alternatives available to them, nor do they know the costs of alternatives. This naivete is as true for users who write their own programs as it is for users who use library programs.

I do not want to argue that one or more of these responses is more cost effective than the RAC recommendations. I merely want to point out that RAC did not consider a range of alternative responses.

Two. A second criticism is that the nature of the problem has not been clearly identified. The RAC does not make an estimate of the amount of computing services which are required, i.e., the present shortfall, nor do they make any estimate of how much additional computing services the present software/hardware system can provide, nor do they make any estimate of how much additional computing services the proposed allocation scheme can generate. As long as the benefits received from the proposed allocation scheme will be greater than the costs, the scheme should be considered. It should be implemented only if the return is at least as great as that which would be obtained from a like investment in any other activity. However, if the additional computing services which are required are much greater than the hardware/software system can provide or which the allocation scheme can provide, the promulgation of the allocation scheme at this time will only cloud the solution of the larger problem.

Three. The third criticism relates to the notion of increasing the utilization of the computing system by making use of so called "idle" time. Unless the system is very lightly utilized, one cannot increase the level of

(This observation was reported by a group of ICS seniors who did an analysis of the impacts of contract accounts in the Spring of 1975 under the direction of Professor Rob Kling.)

One can seriously question the limits to which utilization can (should) utilization without adversely affecting both the old and the new users. As the utilization of a system increases, the turnaround time and the length of the queues associated with the system will increase, i.e., the quality of service will deteriorate. Maybe the responsible decision makers are willing to trade off deterioration of service for increased level of utilization. However, the costs and benefits involved should be made explicit. And the RAC proposal does not make them explicit.

In addition, increasing the utilization by load leveling will mean changing the nature of courses and the work habits of students. Course assignments using the computing systems will have to be spaced more evenly over the quarter and major end-of-quarter assignments will have to be modified. Students will have to do more of their computing at night and on weekends. While this may only be an inconvenience for students living on campus, it can be a major problem for commuter students without personal transportation. (This observation was reported by a group of ICS seniors who did an analysis of the impacts of contract accounts in the Spring of 1975 under the direction of Professor Rob Kling.)

One can seriously question the limits to which utilization can (should) be pushed. One hundred percent (100%) utilization is clearly too much to expect. What is a reasonable standard? The University's classroom space occupancy standard is 2/3 of the stations 3/4 of the time (the 70 hours between 8:00 AM and 10:00 PM on weekdays). That's 35 hours per week. If each of UCI's 150 computer ports was occupied for 35 hours per week, the Facility would provide 5,250 connect hours per week or 52,500 hours per quarter or 157,500 hours per three-quarter academic year. But the Facility is already providing 180,000 connect hours per year. If RAC were arguing space utilization, they would be arguing for more resources. But the

Resource Allocation Committee is arguing computer utilization, and it is arguing for more intensive utilization.

The notion of staged release of IUC funds has some merit and deserves a trial and implementation. However, the total proposed effort to reduce idle time and spread the load will probably result in excessive hardship.

Four. The fourth criticism relates more directly to the proposed allocation scheme. One problem with the RAC recommendation is that it requires that the capacity of the computing system be measured. Several references are made in the report to the determination of the capacity of the campus computing systems: Each fraction is to be 10% or more capacity; fraction shareholders can get more than their fraction if usage is under 80% of capacity; internal prices to IUC users will be discounted if IUC utilization is less than 60% of its fraction.

The measurement of the capacity of a modern computing system is a very difficult task. For example, one can count the number of ports (into which terminals can be connected). However, the number of ports is not necessarily the number of users which a system can handle simultaneously because the number of users which a system can handle simultaneously depends on the demands of the users (e.g., CPU, CORE, I/O) and the response standards of the computing environment. Similar examples can be provided for CPU resources (it is not possible to obtain 100% CPU utilization for an extended period in a typical time sharing system in an average environment) and for core (one may well be prepared to sacrifice core utilization to improve CPU or I/O utilization). In fact scheduling the use of a multi-user computing system to accomplish a set of goals (so called policy driven scheduling) is only now becoming available (e.g., IBM's OS/VS2) and the problem is far from being understood.

A second problem with the allocation scheme proposed by RAC is that it places the Computing Facility in conflict situation. If the Facility sells 50% of its services on the basis of fractional shares, and has to sell 50% in a conventional fashion, the Facility may find itself catering to the conventional user (at the expense of the fractional share user) in order to balance its budget. If IUC users want 50% of "capacity" for 50% of the cost of running the Facility, it is unlikely that the Facility could bring in the other 50% of its costs with only 50% of "capacity".

A third problem is the cost of implementation. While the RAC strategy of ignoring implementation costs has something to be said for it, one cannot ignore these costs forever. I submit that the short-run and long run cost of implementing the RAC proposal are significant. I believe that there are major technical problems in implementating the proposed allocation schemes. While I hesitate to say what the cost would be, I think that the initial cost would be at least $60,000 ($30,000 per system for people and machine time). However, the greater cost will be the continuing maintenance of the software necessary to implement the scheme. Everytime there is a new release of the operating system the software may have to be redone. And when there are problems with the operating system scheduler, we may have trouble getting help from the vendors because of our bastard system. A continuing cost of $30,000 per year is probably a low estimate.

A fourth problem results from the size of the change. The RAC proposal is a big change. And big changes are invariably upsetting. One of the arguments for incremental changes is that they are less upsetting. Vendors and users can forecast the impact of modest changes in the price level. But the effect of drastic changes in the price structure are much harder to forecast.

TWO ALTERNATIVES TO THE RAC RECOMMENDATIONS

Modification of the RAC recommendations. If the RAC recommendations on fractional shares are accepted, I recommend that the recommendations be modified to take advantage of the experience of the Triangle Universities Computation Center (TUCC). TUCC is a non-profit corporation, established in 1965 by Duke, University of North Carolina, and North Carolina State University to provide computing services. Now there is a fourth partner--the North Carolina Educational Computing Service (NCECS)--which provides computer services to educational institutions other than the three founding institutions of TUCC. Each of the four partners is the equivalent of a RAC fractional share holder. At TUCC, the users and management don't worry about whether each of the four fractional share holders is getting 25% of capacity. The concern is whether each gets an equal share of the measured resources on any given day. If one share holder is getting less than the others, than that shareholders priority is increased until the usage is equal.

With the TUCC strategy, there is no need to measure the capacity of the system for allocation purposes. Each shareholder can expand its fraction if jobs are available to run. (This does not solve the problem of big jobs or interactive users which get started and prevent other shareholders from getting on. TUCC has a big enough batch load to balance the inequities deriving from the interactive load.) The Computing Facility will not be placed in a potential conflict situation if all users work as part of a fractional share. Rather than have the Computing Facility sell both shares and retail services, the Computing Facility will sell only shares. Another unit analagous to the NCECS (maybe split off of the Computing Facility) will sell the services provided by the residual fraction.

Summary. The RAC proposal can be improved by having the Computing Facility sell only fractional shares and setting up a separate organization analogous to NCECS, to sell services to those users who do not want to buy shares.

A Simpler Alternative

Another way of increasing utilization of the computing systems is simply to lower prices until the desired level of utilization is reached (or conversely the quality of service becomes barely tolerable).

The RAC apparently considered this alternative but rejected it on the grounds that the Computing Facility might lose money. After rejecting the idea of a global price reduction, RAC proceeded to recommend a partitioning of the computing system which would enable the IUC managers to increase IUC utilization by effectively lowering prices to IUC users without(?) impacting Computing Facility earnings.

Why not increase utilization by lowering prices to everybody? If prices are lowered, the Computing Facility might sustain a loss, i.e., income obtained from services provided might not cover the costs incurred. But the services are being provided to University students, faculty, and staff. Isn't the loss just a paper transaction? Not if the income is used to control the quantity of computing services provided. Than it's not our money; it's his and hers. But let's assume that we can make an independent determination of how much service is to be provided? The campus has "capacity" Q and has to charge price P to clear the market (i.e., to use up quantity Q) and P*Q is less than cost (i.e., the Computing Facility has a loss). Is something wrong? Economists tell us that as long as the consumer's surplus is greater than the loss, the organization is making the right decision. Consumer's surplus is the value (benefit) which the consumer gets from a transaction above the price which he/she pays. If the consumer gets a good or service for less than he/she would be willing to pay for that good or service, the consumer is the beneficiary of consumer's surplus. Normally, consumer surplus belongs to the consumer; however, if the consumer is a member of the organization, the organization can use the consumer surplus in making the decisions about the quantity of computing services to be provided. Admittedly, the measure of consumer's surplus is difficult, but it is probably not more difficult than measuring the "capacity" of a computing system. But the essence of the argument here is that an allocation scheme should not be rejected simply because it leads to a paper loss for a cost center.

The question of recapturing the consumer's surplus for non-university users has a simple answer. The non-university user can be charged a surcharge in order to cover his/her parts of the loss.

Summary. The RAC proposal is essentially a scheme for lowering the price of computing services to instructional users and others who would purchase fractional shares. It would be far simpler and less discriminatory to lower prices to everyone with appropriate discounts for long term committments and appropriate surcharges for non-university users.

MORALS

Moral one: Allocation schemes are important tools in the management of computing resources. In designing new allocation schemes, at least two contraints should be considered. First, the allocation scheme is only one of the problem solving tools available to the manager of computing resources. The alternative of changing the allocation scheme must be compared to other alternatives (e.g., hardware changes, software changes, support level changes) to select the best tool for solving the problem at hand. Second, while allocation schemes are effective tools in modifying users' behavior, in increasing utilization, and in increasing the utility which a computing system can provide, they are not magic potions which can solve all problems.

Moral two: An internal computing center should set prices so as to clear the market. The price for computing services should be set so that users will consume all of the services the hardware/software system can produce subject to the constraints on quality of service discussed above.

Moral three: Whether this market clearing price is arrived at by subscription pricing or by a conventional price list (where users are charged for services consumed with appropriate discounts for slack periods and appropriate surcharges for peak periods) is not of great consequence.

Moral four: If a subscription pricing scheme is used, it is important to avoid possible problems in cost of implementation, in measuring capacity, with inadequate constraints, and between subscription and non-subscription users.

Moral five: The organization should be willing to have an internal computing center run losses as long as these losses do not exceed the consumer's surplus of in-house users.