THE ALLOCATION OF REAL-TIME COMPUTING WITHIN A MULTIPLE-USER ORGANIZATION

Jeffrey H. Moore

ABSTRACT

The allocation of computer time among competing uses within an organization has received considerable attention in the literature. Most approaches focus upon the use of decentralized mechanisms for effecting resource allocation of computer time within the non-market setting of an organization. Under the simplifying assumption that the only resource to be allocated is computer time, this paper investigates a decentralized mechanism, based upon bidding, for the optimal real-time allocation of computer time within a multiple-user organization. A simulation model is then developed to generate alternative decision rules for the cases in which no known analytical rules exist. A general principle for simulation methodology, called "reducto ad credibilis," is proposed for restricting the class of simulations.

INTRODUCTION

The spectacular growth in the use of time-sharing services by organizations has been accompanied by a variety of pricing schemes to assist the computer center management in allocating these services. However, the problem of effectively allocating a time-sharing system (TSS) within an organization in real-time has to date received little attention. This paper considers a decentralized allocation scheme which greatly minimizes implementation and operational costs both from the standpoint of TSS operation and the user interference.

The analysis considers an organization which operates a TSS whose operating cost is independent of its utilization. For convenience the organizational payoff function to be optimized is assumed to be the sum of its members individual quadratic payoffs from use of the TSS. Each user's utility function for the session is assumed to be a member of a simple quadratic family of utility functions. It is assumed that the maximum expected response time by the TSS to a request is a published statistic and is guaranteed to users of the system, at least on the average. Furthermore, it is assumed that the utilization of the TSS will be limited so as to prevent violation of the guarantee. This is accomplished by requiring every customer, defined as a person desiring to become a user of the TSS throughout a session, to submit a bid at the beginning of his use of the TSS. Since it is assumed that requests submitted by all users during a session are independent and share a common service time distribution, the allocation assigned to a customer in response to a bid is the rate parameter of his (Poisson) request arrival process, less a fractional amount for the incremental TSS overhead his presence introduces. It is further assumed that all customers of a session are not simultaneously present to participate in the bidding and that the only resource in short supply is CPU processing time. Finally it is assumed that once an allocation is made by the TSS, the allocation decision must be honored. That is, it is assumed that once the customer submits a bid and receives his allocation, he will not be required to engage in re-bidding even in the light of bids from new customers. However, the TSS will not know in advance what a customer will bid and so there is uncertainty in the allocation problem, as represented by a (known) bid distribution.

THE MODEL

Define Λ, to be the parameter of the Poisson request arrival process which for the given known service-time distribution yields an expected request response time, W, equal to the guarantee and define, $\lambda \leq \Lambda$, to be the aggregate arrival rate allocated to all current users of the session. The ratio, λ/Λ, will be called the congestion, c, of the TSS. The complement of congestion will be called the reserve, x, of the TSS:

$$c = \lambda/\Lambda \qquad 0 \leq c \leq 1,$$
$$x = 1 - c \qquad 0 \leq x \leq 1.$$

The fraction of x allocated to the i-th customer will be called his assignment, a_i. The assignment uniquely determines the request arrival rate of cutomer i, $\lambda_i = a_i \Lambda x$. The congestion introduced to an uncongested TSS by a single utility-maximizing user will be called the stress, k, on the TSS. It will be assumed throughout that $k << 1$. That is, each customer is assumed to be atomistic.

All customers are assumed to have utility functions of the same quadratic family, but each is allowed a different scale factor, defined for each customer, i, as his bid, b_i. A simple bivariate quadratic function for customer, i, will be used and is given

by:

$$U_i(x_i,a_i) = b_i(2kx_ia_i - a_i^2), \qquad (1)$$

where x_i is the reserve of the TSS when customer i arrives, a_i is the assignment he receives in response to his bid, b_i, and k is the stress of the TSS. An extended discussion is given in [1], justifying this utility function as appropriate for the class of quadratic payoff functions. The goal of the computerized speculator is to:

$$\max V_n = \max_{a=(a_1,a_2,a_3,\ldots,a_n)} E\left[\sum_{j=1}^{n} U_j(x_j,a_j)\right],$$

subject to

$x_n = X$, an initial reserve

$0 \leq a_j \leq x_j \qquad j = 1,2,3,\ldots,n$

$x_{j-1} = px_j - a_j \qquad j = 2,3,\ldots,n$

$b_j > m > 0 \qquad j = 1,2,3,\ldots,n$

where b_j is assumed to be from the view of the TSS an independent identically distributed random variable, $b_j \in (m,M], M < \infty$, with distribution function, β, and mean $\bar{b}$. 1 - p is the proportion of the reserve dedicated to the overhead necessary to accommodate the presence of a user on the TSS and is assumed to be independent of his assignment.

It is shown in [1] that if the assignment policy $\pi_n^* = (a_n^*, a_{n-1}^*, \ldots, a_2^*, a_1^*)$ maximizes the organizational payoff V_n for any $n \geq 1$, then an approximately optimal policy can be found without need for dynamic programming recursion from the infinite horizon policy:

$$\Pi^* = \lim_{n\to\infty} \Pi_n^*$$

Under this condition it is also shown in [1] that the optimal assignment by the TSS in response to a bid, b, from a customer is given by

$$a^* = \lim_{n\to\infty} a_n^* = kx\,\frac{b-kp\Phi}{b-k^2\Phi} \qquad (2)$$

where Φ is the fixed point of the integral

$$h(\phi) = \int_b \frac{b + (p-2k)p\phi}{b-k^2\phi}\, d\beta(b) \qquad (3)$$

That is, Φ is a point such that

$$\Phi = h(\Phi).$$

The fixed point, Φ, of $h(\phi)$ is unique under the restriction that for every bid, b, the following holds:

$$b > pk\Phi. \qquad (4)$$

Although no closed form solution for Φ exists, a computer program has been written to numerically find by recursion the fixed point of $h(\phi)$ for various k and p and for representative bid distributions: constant, uniform, exponential and normal. Surprisingly, for some parameter combinations the calculation of Φ is onerous, requiring up to a CPU hour or more of computation (FORTRAN H, IBM 360/67). Fortunately, the calculation need be done only once for a given TSS and a given β. The formulas for finding Φ and an easily found upper bound on the fixed point, which for a large TSS is tight, are reported on in [2].

Although the analysis thusfar appears to be a bit obtuse, its implementation on a given TSS is quite simple:

1. For a given TSS determine the parameters k and p.
2. Assess a reasonable representation for the bid distribution, β, of the user group.
3. Solve (3) once for the fixed point,
4. Install the optimal assignment function (2) as an interactive routine, automatically involked upon sign-on of a new customer to the TSS.
5. Evaluate (2) to yield the customer's assignment in response to his bid.

SIMULATION MODEL

The analytic model thusfar has assumed that users are allowed to arrive at random times during the session; no pre-planning or pre-bidding is required. However, the model implicitly assumes that once a customer becomes a user of the TSS that he will remain a user of the TSS until the end of the session. A more reasonable assumption would be that users come and go randomly. That is, a user would remain on the TSS for a session of randomly determined length. No closed form expression for a^* is known to exist for this case and simulation appears to be the only alternative under this more reasonable condition. Unfortunately it is unclear what is to be simulated. One cannot simulate the allocational behavior of the TSS without an assignment rule to govern its use. But the appropriate assignment rule to use is in functional form determined by parameters of TSS behavior. This simultaneity is a common problem in economics: optimal production and prices are a function of demand, while demand (and hence supply) is in turn a function of the prices. There is a clear analogy in the problem addressed here. It is all the more insidious, however, since not even the functional

form of optimal assignment rule is known, much less the value of its parameters.

In order to proceed, some form of heuristic must be adopted to identify plausible assignment rules so that their performance can be simulated for various customer arrival and departure rates. As a guide in identifying plausible rules, the following principle--called Reductio Ad Credibilis--is proposed for this and similar simulation studies in which decision rules are to be evaluated. The concept is quite simple: Any candidate decision rule must be rejected if it doesn't behave optimally under conditions where optimal behavior can be conclusively identified.

> Reductio Ad Credibilis: Any simulation model must produce responses consistent with known analytical results when the simulation model is reduced to conform with the assumptions of that analytic model.

Reductio Ad Credibilis means, literally, reduce to credibility and would appear to be a necessary, but not sufficient, validation condition of a simulation model. Note that Reductio Ad Credibilis is not a statement about optimality properties in general of a simulation model. A decision rule could conceivably be found for a given simulation model and criterion which on the average outperforms any single reductio-ad-credibilis decision rule. However, it would be serendipitous to find such a rule and there remains the problem of validating it, given that it fails to pass optimality tests in the particular cases where known analytical results are available.

In the TSS allocation problem to be examined here, one known analytical result exists, (2), under the no-departure-of-users assumption. A second analytical result can also be found under the assumption that a user departs instantaneously after receiving his assignment. Under this condition a^* can be shown to be given by:

$$a^* = \lim_{n\to\infty} a_n^* = kx \qquad (5)$$

Note that the optimal assignment in this case is independent of the bid submitted, so long as it is non-zero. This is because instantaneous departure implies no congestion, and hence no reason exists to limit access to the TSS. It can also be verified that, calling the optimal assignment for the condition in (2), $\underline{a}^*$ and calling the optimal assignment for the condition in (5), $\bar{a}^*$:

$$\bar{a}^* \geq \underline{a}^*$$

Furthermore, it would never be optimal to make an assignment outside these bounds. Therefore for the case of departures in general

$$\bar{a}^* \geq a^* \geq \underline{a}^*.$$

Using these analytical results, a simulation model was written to simulate the bidding scheme for allocating the TSS under both arrival and departure of customers. It was assumed that the TSS was an M/M/∞ system, implying that the TSS was an infinite server (no queuing of customers waiting for a TSS terminal) system with both session lengths (holding times) and customer inter-arrival times exponentially distributed. As a practical matter, however, it is, not relevant to assume an infinite number of customers would arrive during a session. The number of users was limited to 2/k after a few trial simulations revealed that this was the practical maximum number of users accommodatable by the TSS, i.e. the system became so "loaded" with users that assignments for customers beyond that number became very small, even for large bids.

A total of seven candidate assignment rules were tested, only three of which exhibited reductio ad credibilus (RAD). The others were considered plausible alternatives and were included for contrast. The rules tested were:

1. Use $\underline{a}^*$ in all cases
2. Use $\bar{a}^*$ in all cases
3. Alternate $\underline{a}^*$ and $\bar{a}^*$; use $\underline{a}^*$ for customer 1, $\bar{a}^*$ for customer 2, $\underline{a}^*$ for customer 3, $\bar{a}^*$ for customer 4, etc.
4. Use $\underline{a}^*$ for the first 2 customers, then use $\bar{a}^*$ for the 3rd customer, then $\underline{a}^*$ for the next 2 customers, etc.
5. Use $a = (\bar{a}^* + \underline{a}^*)/2$.
6. Use $\bar{a}^*$ if the previous system transition was caused by a customer departure; use $\underline{a}^*$ if the previous system transition was caused by a customer arrival.
7. Use a weighted average of $\underline{a}^*$ and $\bar{a}^*$: $a^* = P\bar{a}^* + (1-P)\underline{a}^*$, where P is the ratio of the departure rate to the arrival rate.

Note that Rule 1 exhibits RAD for the no-departure case and Rule 2 exhibits RAD for the instantaneous departure case. Rules 3, 4 and 5 take various combinations of $\underline{a}^*$ and $\bar{a}^*$. None of these rules exhibit RAD. Rule 6 applies the instantaneous departure rule, $\bar{a}^*$, if the last transition was a departure and uses the no-departure rule, $\underline{a}^*$, if the last transition was an arrival. Rule 7 explicitly uses the customer departure and arrival rates in a weighted average. It exhibits RAD whenever Rules 1 and 2 do, and it was hypothesized that Rule 7 would dominate both Rules 1 and 2 in all cases.

The simulation model was run for the following parameter values: k=.05, p=.99, b=100. And for four bid distributions: Uniform, Constant, Exponential and Normal. The fixed point, Φ, was calculated for each distribution and used to determine the minimum acceptable bid by (4). A bid from a customer which is below the minimum is

always rejected as being too low relative to other bids likely from future customers. The maximum number of simultaneous users of the TSS was set to 40 and P, the ratio of departure rate to arrival rate, was set to (1, .5, .25, .1, .05, .01, 0).

Tables 1-4 present the results of the simulations. For each bid distribution, assignment rule and P ratio, 40 simulation runs were made and the aggregate organizational payoff for all the customers was calculated for each run, under the assumption that the customer process was M/M/40. The aggregate payoff was averaged over each of the 40 runs to produce an entry in each of the tables. The computer time involved in producing Tables 1-4 was 6.2 CPU minutes (FORTRAN H, IBM 360/67).

DISCUSSION

As can be seen from the payoff maximizing rules for various combinations (each is marked with a *), no one rule is optimal for every case. There is a clear pattern, however, for high P, RAD rules are best and similarly for low P. Surprisingly, for intermediate values of P the simple average of the two RAD rules is best, dominating in all cases the more complicated averaging schemes. The hypothesis concerning Rule 7 was not confirmed. This suggests for intermediate values of P that a non-RAD rule (Rule 5) be used. This would seem to violate the RAD principle. However this could be easily rectified by:

Rule 8: If .5 > P > .05 Use Rule 5; otherwise, Use Rule 7.

Clearly no optimality claims can (yet) be made about Rule 8, but the RAD principle is now preserved. In general any non-RAD simulation can be converted to a RAD simulation by a suitable choice of decision rules. This is clearly a second best scheme--analysis to yield the optimal results is best. It is proposed, however, that suitable convex combinations of RAD simulation rules can arbitrary closely approximate the optimal rule, in general.

For the TSS under simulation study a suitable RAD rule has been identified. Further simulation research will be, of course, needed to validate it or some other rule. The results thusfar are suggestive, however, that for the quadratic family of payoffs studied, a simple scheme, involving little operational cost, can be utilized for real-time TSS allocation.

BIBLIOGRAPHY

[1] J.H. Moore, Decentralized allocation a time shared computer in a quadratic team, Research Report LR-25, Center for Research in Management Science, University of California, Berkeley (1974)

[2] J.H. Moore, A bidding model for allocating time-sharing services in an organization, Proceedings of the Eighth International Systems Science Conference, Western Publishing Company, Los Angeles (1975)

TABLE 1

AVERAGE TOTAL UTILITY FOR UNIFORM BID DISTRIBUTION, k=.05, p=.99, n=40

P RATIO IS	1.00	0.50	0.25	0.10	0.05	0.01	0.0
RULE 1	6.58	6.29	5.81	4.98	4.49	3.64	3.36*
RULE 2	9.16*	8.38*	7.19	5.24	4.22	2.84	2.52
RULE 3	7.93	7.42	6.57	5.19	4.35	3.17	2.83
RULE 4	7.45	7.03	6.32	5.14	4.45	3.34	2.39
RULE 5	8.65	8.10	7.20*	5.66*	4.79*	3.48	3.12
RULE 6	8.21	7.50	6.57	5.15	4.49	3.57	3.01
RULE 7	9.16*	8.10	6.71	5.23	4.58	3.65*	3.36*

TABLE 2

AVERAGE TOTAL UTILITY FOR CONSTANT BID DISTRIBUTION, k=.05, p=.99, n=40

P RATIO IS	1.00	0.50	0.25	0.10	0.05	0.01	0.0
RULE 1	6.90	6.65	6.17	5.29	4.56	3.56*	3.29*
RULE 2	9.12*	8.47*	7.31	5.48	4.20	2.81	2.52
RULE 3	8.05	7.61	6.78	5.44	4.38	3.10	2.81
RULE 4	7.66	7.29	6.58	5.41	4.44	3.27	2.39
RULE 5	8.69	8.21	7.34*	5.86*	4.74*	3.37	3.06
RULE 6	8.32	7.71	6.78	5.45	4.53	3.50	2.98
RULE 7	9.12*	8.21	6.93	5.51	4.62	3.56*	3.29*

TABLE 3

AVERAGE TOTAL UTILITY FOR EXPONENTIAL BID DISTRIBUTION, k=.05, p=.99, n=40

P RATIO IS	1.00	0.50	0.25	0.10	0.05	0.01	0.0
RULE 1	6.89	6.62	6.11	5.25	4.55	3.55*	3.33*
RULE 2	9.22*	8.47*	7.31	5.36	4.18	2.76	2.52
RULE 3	8.09	7.61	6.78	5.35	4.37	3.06	2.82
RULE 4	7.68	7.28	6.56	5.32	4.43	3.25	2.40
RULE 5	8.76	8.23	7.34*	5.78*	4.73*	3.36	3.08
RULE 6	8.37	7.70	6.84	5.39	4.46	3.49	3.01
RULE 7	9.22*	8.23	6.90	5.46	4.62	3.55*	3.33*

TABLE 4

AVERAGE TOTAL UTILITY FOR NORMAL BID DISTRIBUTION, k=.05, p=.99, n=40

P RATIO IS	1.00	0.50	0.25	0.10	0.05	0.01	0.0
RULE 1	9.77	9.55	9.09	7.45	6.57	5.26*	4.68*
RULE 2	12.01*	11.09*	9.97	7.09	5.69	3.98	3.37
RULE 3	10.92	10.37	9.57	7.28	6.06	4.51	3.86
RULE 4	10.51	10.08	9.43	7.36	6.26	4.82	3.18
RULE 5	11.59	10.98	10.14*	7.73*	6.46	4.77	4.12
RULE 6	11.24	10.47	9.58	7.31	6.34	5.10	4.12
RULE 7	12.01*	10.98	9.78	7.60	6.60*	5.26*	4.68*