Hierarchical computation of the resource allocation problem

Timothy Van Zandt*
Princeton University

September 28, 1994

Abstract

Some recent research on information processing in organizations has treated the agents who process information as endogenous. This paper discusses a sample of models in this area, which differ in their methodology but are unifed by the fact that they study the resource allocation problem. Computational constraints are related to the structure and returns to scale of hierarchies.

Author's address:

Timothy Van Zandt
Department of Economics
Princeton University
Princeton, NJ 08544
USA

Voice: 1-609-258-4050
Fax: 1-609-258-6419
Email: tvz@princeton.edu

*This research was supported in part by grant SBR-9223917 from the National Science Foundation. I thank the "Pôle d'Attraction Interuniversitaire" program of the Belgian Government which has supported this work under grant 26.

1 Information processing and the theory of organizations

In some of the recent models of information processing in organizations, the agents who process information are selected endogenously (as opposed, for example, to message exchange among a fixed set of agents). We shall give an overview of this research by discussing a small sample of models whose methodologies differ but that are all hierarchical and based on the resource allocation problem.[1]

Section 3 considers a team theory model by Geanakoplos and Milgrom (1991). This is a static model in which the bounded capacity to process information is represented by bounds on the amount of information each manager can use in decisions rules. Section 4 relates models of associative computation, such as Keren and Levhari (1979, 1983), Radner (1993), Van Zandt (1994c), and Bolton and Dewatripont (1994), to the resource allocation problem, and discusses some of their conclusions. Unlike team models, these models explicitly model the sequential nature of computation and computational delay, but they study the computation problem in isolation and take the cost of delay or value of throughput to be exogenous, rather than deriving it from a decision problem. Section 5 considers a model of real-time computation of the resource allocation problem by Van Zandt (1994a). The model combines a dynamic decision problem with parallel computation, as in Radner and Van Zandt (1992) and Van Zandt and Radner (1994).

2 The resource allocation problem

Consider the one-good resource allocation problem without externalities, framed as a cost minimization problem. Given a total quantity x_R of a resource, we choose an allocation $\langle x_1, \ldots, x_n \rangle$ of the resource to n shops in order to solve

$$(1) \qquad \begin{aligned} \min \quad & \textstyle\sum_{i=1}^{n} C_i(x_i) \\ \text{subj. to:} \quad & \textstyle\sum_{i=1}^{n} x_i = x_R \ . \end{aligned}$$

x_i is a transfer to shop i that could represent capital, some other input, or orders to be filled. The parameters in the problem are the shops' cost functions and the total resource vector. The size of the organization, for the purpose of characterizing returns to scale, is the number n of shops.

An important property of the resource allocation problem without externalities is that it can be decomposed into similar problems, whose solutions are independent. Call a group of shops a *division*. Given an allocation x_j to division j, let $C_j(x_j)$ be the division's minimized total shop costs. C_j

[1] For a more comprehensive survey, see Van Zandt (1994c).

is called the division's aggregate cost function. Allocating resources first to divisions (to minimize total aggregate costs of the divisions) and then suballocating resources within the divisions (to minimize the total costs of each division) solves the resource allocation problem.

When the cost functions are symmetric and quadratic,

$$C_i(x_i) \;=\; (x_i - \gamma_i)^2 \;,$$

the aggregate cost functions and the optimal decision rules are simple. Let θ_j and n_j be the set of shops and the number of shops, respectively, in division j. Let $\gamma_j = \sum_{i \in \theta_j} \gamma_j$; this is called division j's aggregate cost parameter. Then the aggregate cost for division j is

$$C_j(x_j) \;=\; \frac{1}{n_j}(x_j - \gamma_j)^2 \;.$$

If division j has subdivisions and if n_k is the number of shops in subdivision k, then the allocation to each subdivision k that minimizes the total costs of j's subdivisions is

$$x_k^* \;=\; \gamma_k + (n_k/n_j)(x_j - \gamma_j) \;. \tag{2}$$

In words: Subdivision k gets its ideal allocation γ_k (which would give it zero costs) plus its share n_k/n_j of the surplus resources $x_j - \gamma_j$.

3 Allocating resources in teams

There have been numerous team-theoretic models of the resource allocation problems (see Van Zandt (1994c) for references). However, in only one, Geanakoplos and Milgrom (1991), are the agents who make decision chosen endogenously. In that paper, the agents are managers in a hierarchy. Instead of using the hierarchy to aggregate information, managers learn about costs by observing exogenous signals that are correlated with the cost parameters. A manager's set of feasible cost signals characterizes the manager's capacity to process information. Communication between managers is limited to the recursive allocation of the resource down the hierarchy.

As is usual in team theory, the managers can compute any functions of their information. However, an important simplifying assumption is that managers do not draw inferences from the allocations they receive, even though these allocations reveal some of their superiors' information. The optimal decision rule for manager j, once she has received her allocation x_j from her immediate superior and has chosen and processed her information, is given by (2), except that γ_j and the γ_k's are replaced by their expected values $\hat{\gamma}_j^j$ and $\hat{\gamma}_k^j$, respectively, conditional on j's information.

There is never a bounded optimal firm size in this model, without additional restrictions on the hierarchical structure. One can always merge

two firms by making one a subsidiary of the other, in a way that reduces total expected costs but does not change the managerial costs. Specifically, the root manager of one firm can be made a subordinate of a manager in the other firm. Even without acquiring information about the subsidiary, the supervisor of the subsidiary can make advantageous transfers to the subsidiary based only on information about her other subordinates' costs.

After merging hierarchies in this way, the informational integration of the organization is one-sided. Decisions are made at one level without any information about a subtree. Here is an example with increasing returns that does not have this asymmetry. The hierarchies in the example can be balanced (managers in the same tier have the same span) and the decisions by each division use information about all the division's shops. The example also illustrates how decentralized processing can reduce expected costs in this model.

Assume that all aggregate cost parameters are available to the managers. Assume also that each manager can read three numbers, after reading his allocation. One manager cannot solve the resource allocation problem exactly; this would require being able to read n numbers. However, r managers, organized in any hierarchy such that each has a span of three, can solve without error the resource allocation problem for $2r+1$ shops,[2] because each manager can read the aggregate cost parameters of her three subordinates.

Geanakoplos and Milgrom (1991) show that there is an optimal firm size if the shops' cost parameters are IID, if only disaggregate information is available, and if hierarchies must be balanced. In large balanced hierarchies, the root manager either has many subordinates or has subordinates that are large divisions; either way, if the root manager can only observe the cost parameters for a few shops, then the root manager cannot effectively allocate resources to his subordinates and the firm should be divided.

The assumption that aggregate cost parameters are exogenously available to managers is overly optimistic; however, the assumption that only disaggregate information is available is too restrictive, given that this model does not permit aggregation of information by the hierarchy. Section 5 discusses a model in which both the aggregation of information and the disaggregation of the allocations is performed by the hierarchy.

4 Associative computation

We now turn to models in which the computation of the decision rules is more explicit than in team theory. The models in this section characterize the optimal parallel algorithms for associative computation. They are

[2] In any tree with r interior nodes that all have s children, there are a total of rs children, $r-1$ of which are interior nodes. Hence, there are $r(s-1)+1$ terminal nodes.

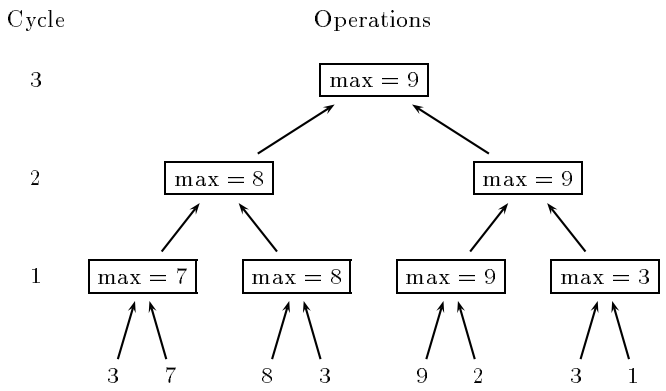

FIGURE 1. Associative computation by a PRAM via a balanced binary tree (Gibbons and Rytter (1988, Figure 1.2)).

related to the resource allocation problem because the aggregation of cost functions is an associative operation (for example, with quadratic costs, aggregation involves summing the γ_i's) and because the disaggregation of allocations has the structure of associative computation in reverse.

These models engage in an exercise that is common in theoretical computer science: to specify a computational model and then to characterize the optimal algorithm for computing some problem. What can economists hope to learn? We can use a model of parallel computation in which the processors represent managers or clerks in an organization. We can use the characterization of the optimal algorithm to measure the resource cost and delay of computing a decision problem. We can also interpret the communication flows between processors as communication flows in an organization, which we know are closely related to an organization's internal structure.

A simple and widely-used model of parallel processing in computer science is the PRAM, which stands for Parallel Random Access Machine. A Random Access Machine (RAM) is similar in power to a Turing machine. It has an infinite memory and its capabilities are described by the elementary operations it can perform and the time each operation takes. An elementary operation is some manipulation of its memory, such as reading two numbers, summing them, and storing the answer in another memory register. A PRAM is a network of RAM's that share the same global memory. Because memory is shared, communication between the RAM's is trivial, or even a vacuous concept.

The efficient algorithms for associative computation with a PRAM is illustrated in Figure 1. Each cycle, the data or previous results are distributed in pairs to processors, and each processor computes the maximum (for example) of the two assigned numbers. The answer is obtained in $\lceil \log_2 n \rceil$ cycles. There are $n - 1$ operations, which is the resource cost of the computation if managers are paid on hourly basis, and is an approximate measure of the

cost in other cases. The algorithm can be represented by a balanced binary tree, but the nodes represent operations, rather than processors. Hence, we cannot interpret the tree in Figure 1 as the hierarchical structure of the organization. This illustrates a deficiency of the PRAM for studying organizations. Because all the processors have equal access to raw data and partial results, the assignment of processors to operations is indeterminate, and the flow of information is not explicit.

To learn more about organizational structure, we need a model of computation with local memory. This means that each manager has her own memory, and managers communicate by sending messages through a communication network. We can then keep track of information flows, and also include a cost or delay for sending, reading and transmitting messages.

Radner (1993) introduced a model with local memory in which there is one additional processing cycle for each message. This communication cost implies that parallelization, which reduces delay, also increases the cost of computational resources. Radner (1993) characterizes the algorithms that are efficient with respect to delay and resources. Keren and Levhari (1979, 1983) can also be interpreted as the Radner model of associative computation, with the exogenous restriction that the communication networks be balanced hierarchies. However, although Radner (1993) does indeed find that the efficient networks are hierarchical, they are not balanced and there is even skip-level reporting, which means that a manager can have subordinates with varying ranks. In fact, in the efficient hierarchies, all managers process the same amount of raw data.

The above research presumes that there is a single problem to compute or least that managers are hourly workers and each problem can be computed by a different set of managers. Several variations have been studied for when problems arrive periodically. Van Zandt (1994c) shows that the Keren and Levhari (1979, 1983) hierarchies become more uniform (the spans of the tiers vary less) when each problem must be computed by the same hierarchy. Radner (1993) studied the periodic processing with salaried managers, and proposed a class of balanced hierarchies that do well. However, Van Zandt (1994b) shows that the efficient networks are not even hierarchical and the processing of each problem closely resembles the efficient processing of a single problem. Bolton and Dewatripont (1994) study the importance of throughput when each problem must be computed by the same network, using a generalized computation model, and find that the throughput criterion can lead to more regular hierarchies.

Having measured the resource cost and delay of associative computation, we can study the returns to scale of the resource allocation problem if we exogenously specify the "cost" of delay (which will increase with problem size). This exercise is studied in Keren and Levhari (1983) and Radner (1993). However, it is more accurate to derive the cost of delay implicitly

from the computation problem, and this we do in the next section.

5 Allocating resources in real time

The cost of computational delay is that it increases the lag of information upon which decisions are based. This cost is best captured in a dynamic decision problem with real-time computation. Radner and Van Zandt (1992) and Van Zandt and Radner (1994) study the real-time computation of predictions of the sum of a collection of stochastic processes. The real-time computation of resource allocations is studied in Van Zandt (1994a), upon which this section is based.

In the dynamic problem, resources are allocated each period. Each shop's cost parameter is a discrete-time stochastic process $\{\gamma_{it}\}$ that we assume to be AR(1), with

$$\gamma_{it} = \beta\gamma_{i,t-1} + W_{it} \ .$$

W_{it} is IID over i and t, and has mean zero. Recall that $E[\gamma_{it}|\gamma_{i,t-d}] = \beta^d\gamma_{i,t-d}$.

We use the PRAM as our computational model. The lack of communication costs in the PRAM is actually an advantage for this exercise, because it will make clear that the benefits of decentralization are due to computational delay rather than communication costs.

We study a family of hierarchical decision procedures. The hierarchical structure refers to the structure of the algorithm, rather than the communication between managers. The nodes of the hierarchy are "offices" rather than managers, and the organizational chart connects components of computation decision making rather than actual administrators. Keep in mind that the offices and their computing resources are formed endogenously and are independent of the shops, which merely supply data and receive allocations.

In a two-tier hierarchy, there is a central office (the top tier) that allocates resources to the shops (the bottom tier) by gathering data each period and computing the statistically optimal allocation (2). Let d be the delay in performing this computation. Then the allocation in period t is based on data from period $t-d$. The expected value $\hat{\gamma}_i^R$ of the aggregate cost parameter is $\sum_{i=1}^n \beta^d\gamma_{i,t-d}$. The statistically optimal allocation is thus

$$x_{it} \ = \ \beta^d\gamma_{i,t-d} + (1/n)\left(x_{Rt} - \textstyle\sum_{h=1}^n \beta^d\gamma_{h,t-d}\right) \ . \tag{3}$$

If multiplication and addition each take one period, then this can be computed by a PRAM in $4 + \log_2 n$ periods and there are $3n + 1$ operations.

In a balanced three-tier hierarchy, the root is again called the central office, the nodes in the middle tier are called division offices, and the bottom

tier contains the shops. Let s_0 be the number of divisions, and let $s_1 = n/s_0$ be the size of each division. The divisions receive allocations from the center, and then compute the same decision rule as the center did in the two-tier hierarchy. Each division's delay is $d_1 = 4+\log_2 s_1$. A partial result of division j's computation is $\hat{\gamma}^j_{jt} = \beta^{d_1}\gamma_{j,t-d_1}$, which is j's conditional expectation of its aggregate cost at time t. The center also computes a similar allocation rule as in the two-tier hierarchy, but its informational inputs are the expected aggregate costs that have been computed by the divisions. The center's delay is $d_0 = 4 + \log_2 s_0$. Its allocation rule is

$$x_{jt} \;=\; \beta^{d_0}\hat{\gamma}^j_{j,t-d_0} + (1/s_0)\left(x_R - \textstyle\sum_{k\in\Theta_R}\hat{\gamma}^k_{k,t-d_0}\right) \;.$$

(Θ_R is the set of divisions.)

The benefit of decentralization is that the allocations within each division are based on information that is $\log_2 s_0$ periods more recent than in the two-tier hierarchy. (The center's delay in the two-tier hierarchy is $4+\log_2 n$, whereas the delay of each division in the three-tier hierarchy is $4+\log_2 s_1$.) The main cost of decentralization is that the number of operations increases from $3n+1$ for the two-tier hierarchy to

$$(3s_0+1) + s_0(3s_1+1) \;=\; 3n + 1 + 4s_0$$

for the three-tier hierarchy. There is also a small delay overhead of 4 periods for the center in the three-tier hierarchy; its cumulative delay is

$$d_1 + d_0 \;=\; 8 + \log_2 n$$

periods, compared to $4+\log_2 n$ periods in the two-tier hierarchy.

For both two-tier and three-tier hierarchies, unit costs as a function of n are eventually increasing even when the wage is zero, because the computational delay degrades the value of the information used to allocate resources. However, typically the unit costs are lower for the three-tier hierarchy, and the optimal firm size is higher for the three-tier hierarchy than for the two-tier hierarchy, because the suballocation within divisions is based on more recent data in the three-tier hierarchy.

When the managerial wage is zero, there is no bound on the size of organizations. However, the value of the root's resource allocation decreases with the size of the firm because of the cumulative delay. If the managerial wage is strictly positive, there is an optimal firm size that minimizes the per-unit costs. A decrease in β, which means that the environment changes more rapidly, causes the optimal firm size to fall. This is consistent with the common perception that firms have downsized in the last decade in order to respond more quickly to a rapidly changing environment.

References

Bolton, P. and Dewatripont, M. (1994). The firm as a communication network. London School of Economics and Université Libre de Bruxelles (forthcoming in the *Quarterly Journal of Economics*).

Geanakoplos, J. and Milgrom, P. (1991). A theory of hierarchies based on limited managerial attention. *Journal of the Japanese and International Economies*, *5*, 205–225.

Gibbons, A. and Rytter, W. (1988). *Efficient Parallel Algorithms*. Cambridge: Cambridge University Press.

Keren, M. and Levhari, D. (1979). The optimum span of control in a pure hierarchy. *Management Science*, *11*, 1162–1172.

Keren, M. and Levhari, D. (1983). The internal organization of the firm and the shape of average costs. *The Bell Journal of Economics*, *14*, 474–486.

Radner, R. (1993). The organization of decentralized information processing. *Econometrica*, *62*, 1109–1146.

Radner, R. and Van Zandt, T. (1992). Information processing in firms and returns to scale. *Annales d'Economie et de Statistique*, *25/26*, 265–298.

Van Zandt, T. (1994a). Real-time hierarchical resource allocation. Princeton University (in preparation).

Van Zandt, T. (1994b). The scheduling and organization of periodic associative computation. Princeton University.

Van Zandt, T. (1994c). The structure and returns to scale of organizations that process information with endogenous agents. Princeton University.

Van Zandt, T. and Radner, R. (1994). Information processing and returns to scale of a statistical decision problem. AT&T Bell Laboratories and Princeton University.