Maximum Computing Power and Cost Factors in the Centralization Problem

A.E. Oldehoeft
Iowa State University*
and
M.H. Halstead
Purdue University†

A simple analysis of some computer-economic factors involved in comparing multimachine installations versus large single machine installations is given, and a mathematical model is derived to assist policy decisions.

Key Words and Phrases: centralization, decentralization, economics of computers, computer management, economies of scale

CR Categories: 2.41

* Department of Computer Science, Ames, IA 50010. † Department of Computer Sciences, Mathematical Sciences Building, Lafayette, IN 47907.

Introduction

Decisions on centralizing or decentralizing computing facilities require the analysis of many factors ranging from the structure of the institution to actual computer costs and capabilities. Perlman [6] discusses these factors and cites advantages for both types of operation. This article is concerned with only one aspect of the economics, namely, the computing cost-power relationship between a consolidated operation and several smaller decentralized operations.

In studies of the history of computing technology, Knight [3, 4, 5] displays technology curves which relate cost ($ rental/seconds) to computing power (conglomerate measure of operations/second). Over time, equal cost will purchase increasing performance. Thus, a periodic review of rented computer equipment is a sound business policy.

In addition to increases in the level of technology, one can expect for any given level, a return to scale approximated by Grosch's Law (computing power P is proportional to the square of the cost C). For the time period 1944–62, Knight [3, 4] demonstrates this cost-power relationship $C = kP^{\alpha}$ by computing $\alpha = .519$ for scientific computation and $\alpha = .459$ for commercial computation. For the time period 1963–66, Knight [5] computes $\alpha = .322$ for scientific computation and $\alpha = .404$ for commercial computation.

Solomon [7, 8] computes values of α for the IBM 360 line using monthly rentals and execution times for various kernels (programs or program segments). For matrix multiplication, $\alpha = .494$; for square root computations, $\alpha = .478$; for field scanning, $\alpha = .682$; and for Arbuckle's instruction mix, $\alpha = .507$. Arbuckle's mix represents a composite of a number of scientific and engineering applications [2]. The variation between values computed by Knight and Solomon can be partially attributed to differences in estimates of monthly rental and methods of measuring computing power.

It is possible to demonstrate completely different relationships if the measure of computing power is based on a highly restricted set of attributes. One such relation, $C = kP^2$, was proposed by Adams [1]. This is derived as a relation between monthly rentals and the reciprocal of access times (rather than an overall measure of computing power) and the empirical data crosses levels of technology. Such relations have to be discounted since studies have demonstrated that computing power cannot be adequately measured using one-attribute characteristics.[1]

It is debatable whether the relation $C = kP^\alpha$ is a natural law or an artifact of the computer industry's pricing policy. There is, however, strong evidence that α is close to the value $\frac{1}{2}$.

Bounds on Computing Power and Cost Factors

We are concerned with the economic feasibility of installing n separate computers with a specified distribution of computing power or installing one centralized computer to serve all n sites. The computing power requirement at the ith site is P_i and the total computing power requirement is $P = \sum P_i$. It is assumed that the computers will be selected from the same level of technology so that $C = kP^\alpha$ with k and α constant. Also, down time must not be critical (i.e. backup computing power is not necessary).

Denote $\beta_i = P_i/P$ so that $\beta_i > 0$, $\sum \beta_i = 1$, and the ith site requires computing power of $\beta_i P$. The centralized site requires computing power of θP where $\theta \geq 1$ may include overhead requirements for extra software (i.e. more than one system on the computer), additional hardware for diverse applications, and terminal service to $n - 1$ sites. Denote the cost of the ith computer by C_i and the cost of the centralized computer by C. The decision to consolidate computing power depends on the cost ratio $R = \sum C_i/C$. If $R > 1$, it is more economical to consolidate; if $R = 1$, the economic factor is irrelevant; and if $R < 1$, it is more economical to decentralize. The computing power-cost relation yields:

$$C = k\theta^\alpha P^\alpha.$$

$$C_i = kP_i^\alpha = k\beta_i^\alpha P^\alpha.$$

$$R = (\sum C_i)/C = (\sum \beta_i^\alpha)/\theta^\alpha.$$

Subject to $\beta_i > 0$ and $\sum \beta_i = 1$, $\max(\sum \beta_i^\alpha)$ occurs at $\beta_i = 1/n$ for a fixed number, n, of decentralized sites and for a given $0 < \alpha < 1$. Using $\min(\theta) = 1$ (i.e. assuming no overhead), we have $R < n^{1-\alpha}$. This upper bound on R denotes the maximum penalty one can pay for failure to consolidate. If there is no overhead demand on the computing power ($\theta = 1$) of the consolidated site and the alternative is to equally distribute the power over n computers, consolidation can result in a maximum reduction in cost. For example, if $\alpha = .5$, three computers, each of computing power $P/3$ could cost as much as $\sqrt{3}$ times one computer of computing power P. The bound $n^{1-\alpha}$ is directly applicable if an institution installs n computers of equal computing power and each is used for the same job mix. The maximum cost factor will be less than $n^{1-\alpha}$ under either of two conditions:

(1) the decentralized computing powers are not equally distributed; or

(2) the application of each computer is homogeneous with respect to hardware and software and the total application is heterogeneous with respect to hardware and software (in which case, $\theta > 1$).

Holding cost constant, we can compute the maximum allowable consolidated computing power θP for a given alternative decentralized distribution of computing power. In order to consolidate and not exceed the cost of decentralization, we set the cost ratio $R = 1$. This allows θ to assume a maximum value of

$$\theta_{\max} = (\sum \beta_i^\alpha)^{1/\alpha}.$$

The relationship between θ and R is clear. If $\theta > \theta_{\max}$, then $R < 1$; if $\theta < \theta_{\max}$, then $R > 1$. This value of $\theta_{\max}$ depends on n, α, and the distribution of the β_i. For fixed n and fixed $0 < \alpha < 1$, we have $1 \leq \theta_{\max} \leq n^{1/\alpha - 1}$. The upper bound is attained in the case of maximum decentralization where $\beta_i = 1/n$ for all i. $\theta_{\max}$ approaches the lower bound as the distribution of the β_i approaches a total consolidation of computing power, where $\beta_i = 1$ and $\beta_j = 0$ for $j \neq i$. The intermediate values of $\theta_{\max}$ depend on the manner in which the distribution of the β_i is allowed to vary on the surface $\sum \beta_i = 1$, $\beta_i > 0$. As a sample path, consider the set of points $(1 - (n - 1)(r/n), r/n, ..., r/n)$ where r serves as a measure of decentralization from $(1, 0, ..., 0)$. The semilogarithmic graph of the corresponding

$$\theta_{\max} = [(n - 1)(r/n)^\alpha + (1 - (n - 1)(r/n))^\alpha]^{1/\alpha}$$

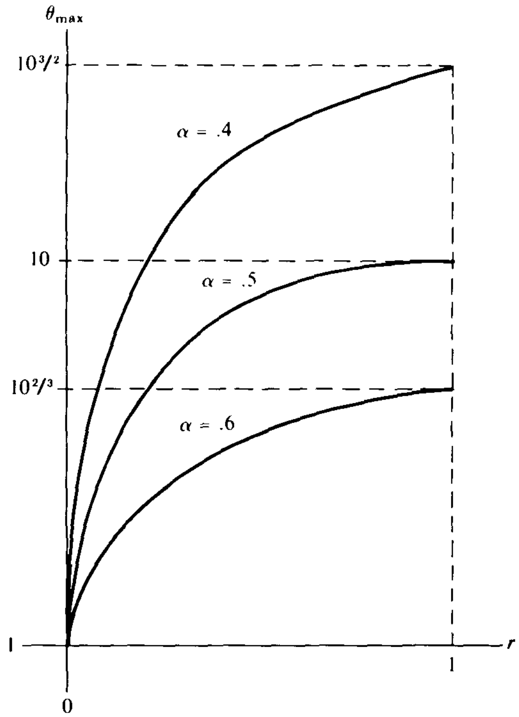
appears in Figure 1 for $n = 10$ and several values of α.

In general, a fixed amount of money will purchase much more consolidated computing power θP than the total decentralized power P, even for the conservative estimate $\alpha = .6$. The more favorable conditions for consolidation occur for larger values of r since a larger amount of computing power is available to absorb overhead requirements. The case for consolidation is less clear for small r since $\theta_{\max}$, the allowable value for θ, is also small. In terms of raw computing power and costs, decentralization is more likely to be feasible in the case of small separate applications where the bulk of computing power is concentrated at one site and the applications are sufficiently diverse.

[1] See Knight [3, p. IV-2] and the discussion in [2].

Received July 1970; revised May 1971

Fig. 1. θ_{max}. (balancing overhead power factor) vs. r (effective decentralization) for $n = 10$.



References

1. Adams, C.W. Grosch's law repealed. *Datamation* (July 1962), 38–39.
2. Arbuckle, R.A. Computer analysis and throughput evaluation. *Computers and Automation* (Jan. 1966), 12–15, 19.
3. Knight, K.E. A study of technological innovation—the evolution of digital computers. Doctoral diss., Carnegie Institute of Technology, Nov. 1963.
4. Knight, K.E. Changes in computer performance. *Datamation* (Sept. 1966), 40–54.
5. Knight, K.E. Evolving computer performance 1963–1967. *Datamation* (Jan. 1968), 31–35.
6. Perlman, J. Centralization vs. decentralization. *Datamation* (Sept. 1965), 24–29.
7. Solomon, M.B. Economies of scale and the IBM System/360. *Comm. ACM 9*, 6 (June 1966), 435–440.
8. Solomon, M.B. Correction to economies of scale and the IBM System/360. *Comm. ACM 10*, 2 (Feb. 1967), 93.

Corrigendum

Computer Systems

In "Dynamic Microprogramming: Processor Organization and Programming" by Allen B. Tucker and Michael J. Flynn, *Comm. ACM 14*, 4 (Apr. 1971), 240–250, the following corrections should be made.

Page 249, replace the 360 code in Example 2(b) by the following:

```
        LR   9,11
LOOP    C    11,0(8)
        BC   8,DONE
        C    10,0(8)
        BC   8,FOUND
        LA   8,8(8)
        B    LOOP
FOUND   L    9,4(8)
```

Page 250, add the following sentence to the end of the second complete paragraph. "The value of 4 $*$ n is initially stored in register R7."

Page 250, replace the 360 code in Example 3(a) by the following.

```
LOOP    ST   11,0(10,8)
        CR   7,10
        BC   8,DONE
        LA   10,4(10)
        AR   12,11
        ST   12,0(10,8)
        CR   7,10
        BC   8,DONE
        LA   10,4(10)
        AR   11,12
        B    LOOP
DONE
```

Note that these changes will have a slight but insignificant effect on the 360 timing figures given in Examples 2(d) and 3(c). The conclusions of that section, however, remain intact.

The authors are indebted to Mr. Steven S. Muchnick of Cornell University for bringing these errors to their attention.