The Analysis of an Adaptive Workload Balancing Strategy in Computing System Resources Management

Hisao Kameda

Received January 1975; revised April 1975

An adaptive workload balancing scheme is analyzed. This scheme is intended to automatically dissolve the imbalance in resource utilization of computing systems and to bring about the proper balance between throughput and turnaround time. This is done by means of the "invisible hand" due to the bidding of priorities of resource use, with no contradiction between the flexible bid for each resource and the stable price for the entire service. The result of this analysis shows that this scheme would regulate those imbalances, and that it includes the load balancing policy of IBM OS/VS2-2 as a special case. Some considerations and methods on an actual implementation are also presented.

KEY WORDS: Bottleneck; coherent priority queueing; IBM OS/VS2 Release 2; invisible hand; microeconomic analysis; multiprogramming; operating system; priority pricing; resource utilization; system resources management; throughput; turnaround time; workload balancing.

1. INTRODUCTION

Among several aspects of operating system performance, we have often the problem of imbalance in resource utilization. For example, in a computing system, occasionally user programs concentrate on one of the resources

1 This paper is a revised version of material which appeared as Research Report RC5072 of IBM Thomas J. Watson Research Center, Yorktown Heights, New York.

2 Department of Computer Science, University of Toronto, Toronto, Canada. On leave from the University of Electro-Communications, Chofu, Tokyo, Japan.
in the system and this resource forms the bottleneck of the system while other resources are underutilized.

We can see that the notions of "bottleneck" and "resource" are also common to economics. Although there must be a great difference between the detailed aspects of both fields, we have the chance of benefiting by exploring the common aspects of computing systems and economic systems. One possible notion might be that of "invisible hands." Nearly 200 years ago, Adam Smith said that the price mechanism would provide invisible guiding hands to regulate the imbalance of resource utilization; i.e., if the demand for a resource becomes too high, its price goes up and the demand becomes lower, since not all can afford such a high price, and vice versa. In the actual economic situation, to apply this principle straightforwardly is sometimes too severe a strain upon such a resource as human labor. However, inside computing systems, we have few problems of this kind and can bring about such an economic environment. Thus, several investigators have discussed this idea of making use of the invisible hand.\(^{1-10}\)

In the case of computing systems, we can expect that such a price mechanism will regulate not only the balance of resource utilization but also the balance between turnaround time and throughput. Although, as is well known, better scheduling algorithms give the system a better relationship between turnaround time and throughput, still turnaround time and throughput do not often go hand in hand. As the throughput comes close to the capacity of the system (i.e., the average cost per job becomes cheaper), generally the average turnaround time of each job will increase. And the supply and demand relation mediated by the price mechanism would give the good balance between them. As for this kind of balancing, there have been several proposals and some are already in use.\(^{1-4}\)

As to the pricing of the priority of each resource in order to eliminate imbalance, we will have the following contradiction: On the one hand, the price of each resource must frequently be changed along with the fluctuation of resource congestion in order to be adaptive to the changing congestion. On the other hand, users normally want stable prices and frequently changing prices are not convenient to users; in particular, accounting errors are difficult to detect and claim in this case.

Here, let us consider a scheme which might avoid such a contradiction as stated above, and analyze the scheme to see whether it can regulate the imbalance of resource utilization. The main point of this scheme is as follows: Each user program bids the price for each resource; each resource gives to each user the priority of using it according to the offered price. In order to avoid the above-mentioned contradiction, we can provide an agent program for each user, and let the agent program explore the optimal strategy of each user and select the optimal set of bids adaptively within