Incentive-Compatible Priority Pricing and Transfer Analysis in Database Services

Yong J. Kim*

〈Abstract〉

A primary concern of physical database design has been efficient retrieval and update of a record because predictable performance of a DBMS is indispensable to time-critical missions. To maintain such phenomenal performance, database managers often spends more than or as much as the goal of an organization can warrant. The motivation of this research stems from the fact that even predictable performance of a physical database can be hampered by stochastic query processing time, physical configurations of a database, and random arrival processes of queries. They all together affect the overall performance of a DBMS. In particular, if there are queuing delays due to limited capacity or during on-peak congestion, this paper suggests to prioritize database services. A surprising finding of this paper is that such a transition from a non-priority system to a corresponding priority-based system can be Pareto-improving in the sense that no users in the system will be worse off after the transition. Thus prioritizing database services can be a viable option for efficient database management.

* Department of Management Information Systems, Kon-Kuk University, Seoul, Korea

1. Introduction and Motivation of the Research[1)]

Traditionally, physical database design has concerned efficient retrieval time of a record. Miscellaneous B-tree or hashing schemes are such examples to support time-critical missions by providing predictable retrieval times. In this paper, we address a managerial aspect of a database management system (the system or DBMS henceforth) plagued by queuing delays. Consider a database server (It may be a pay-per-service like 114 service or an intranet service which should answer queries on real-ime basis) with the goal of maximizing its value for the organization that owns it. Owing to the physical layout of the database or the nature of individual queries, users of the system suffer queuing delays: when queries are submitted to the system, owners of the queries wait until subsequent retrieval/update operations are rendered by the database server. Such queuing delays are also further compounded by stochastic arrivals of queries entailing heterogeneous processing requirements.

Because there is no market for the queuing delays (a negative externality), the system will be operated in a suboptimal state in which more queries tend to be tendered than otherwise. Thus, control of delay costs is an important issue in a DBMS. We call the aggregate value of the queries submitted to the DBMS as the *system value*, and denote it as $V(\cdot)$. Following Mendelson and Whang (1990), we assume the following:

- There are N heterogeneous query classes: the result of a class-i query produces a value but requires a processing time governed by a distribution associated with class i. We will use service time and processing time interchangeably. We also identify a class-i query with a class-i user, meaning that class-i query belongs to a class-i user.
- The system $V(\cdot)$ values is an aggregation of the values garnered by individual queries across all classes, and the system delay cost is an aggregation of the delay costs experienced by individual queries.
- There is a benevolent system manager who wants to maximize the net system value -- the system value minus the total delay cost.
- Each individual user knows his own delay cost per unit time, the distribution of his query processing time, the value of his query, the expected queuing delay, and the access charge. The system manager knows the expected queuing delays and access charges, but knows only *aggregate* statistics about users, queries: delay costs, service time distributions, and query valuation of user classes.

The basic model of the DBMS server as a queuing system is illustrated in <Figure 1>. In this general model, λ, $c^{(j)}$, ν, and $ST(\lambda)$ denote the mean arrival rate, j-th moment of the query processing time distribution, delay cost per unit time (in monetary terms), and the average sojourn time of a query submitted to the DBMS which is modeled as an $M/G/1$ system.

1) I owe my genuine gratitude to the Institute of Economics and Management at Kon-Kuk University, which supported this research under 1996 settlement fund for new faculty members.