

A Greedy Algorithm for E- Systems Management

L. Lawrence Ho

For those electronic merchants with limited resources but limitless ambition for profits, providing quality-of-service/experience differentially to customers in proportion to their buying potential or “worth” seems a logical way to maximize profits. While on-line users forever demand fast response and secure transactions and shopping bargains, electronic merchants and service providers are as relentless in managing their finite resources to maximize profits. For electronic merchants, being able to gratify the needs of “clickers” with most money and urge to spend is therefore a reasonable guide to profits. More, this “greed-based” principle can also guide the monitoring, control, and management of the performance of electronic commerce infrastructures. Effective performance management of electronic commerce includes two key factors. First, customers should be prioritized according to their previous buying patterns and the current likelihood of buying. Second, infrastructure resources should be assigned to customers according to their priorities.

A typical electronic commerce infrastructure (Figure 1) includes its access router, Ethernet switch, load balancer, and web and application and database servers, all working in concert to serve the customers’ browsing and purchasing needs and transactions. The load balancer diverts HTTP requests to web or application servers according to their load conditions. The various application/web/database servers process the HTTP requests, run the CGI programs and other applications, and fulfill the database transactions. Therefore, the key infrastructure resources for serving on-line customers include the throughput of the router and load balancer, and the processing power and IO throughput and available memory of the various servers. Under light load condition, when customers are not many and their requests not overtaxing the available resources, the quality-of-service parameter (i.e., latency) for all customers are within acceptable bounds. Only when resources are oversubscribed then response time is proportionally and aversely degraded. With limited resources comes the needs to actively manage and control the performance and resources of the infrastructure on a per customer basis, before service anomalies and service outages overtake the E-commerce infrastructures.

Figure 1 here

Merchants' infrastructure-related resources (e.g., network resources and server processing power) become limited when (a) the number of on-line users and their resource demands exceed certain physical limits, or (b) the merchants' infrastructures are anomalous and faulty. Often, scenario "a" leads to scenario "b," if it is not corrected in time. In both cases, the merchants must choose between two mutually exclusive choices. First, they can choose to maintain and indeed guarantee the quality-of-service of their on-line users delimited by their now limited resources. This effectively means that only a fraction of the users can be adequately served while the rest are denied services, effectively instituting admission control. Second, merchants can choose to serve all on-line users equally, at the expense of quality-of-service for all. This leads to uniformly degraded quality-of-service/experience for all on-line users. In the first case, a fraction of the users are happy while the rest are not. In the second case, everybody fumes and whines, and some leave for other providers. The second case, if followed to its limits, will lead to service denial to all, since the performance of an overloaded system will spiral to non-performance if load is not limited. Thus, admission control must even be instituted in the second case. If admission control must be instituted to protect electronic commerce infrastructures and guarantee acceptable quality-of-service to on-line users, what factors determine the admission policies?

Here is where the "greed" principle of maximizing merchants' profits comes in. Admission policy for overload protection can be based on

- The past purchasing history of users, effectively granting higher priorities to those who spent more money in the past, and
- How close at present the customers are from completing real transactions.

The first implies that the more a customer has spent in the past, the more likely he will be served by the limited resources of the E-commerce sites. The second implies that those who are close to completing real transactions should be guaranteed some level of service. Both imply that those customers who are most likely to generate revenues for the electronic merchants are given proportionally higher priorities in services. This approach effectively

- Maximizes the profits of electronic merchants given the physical limits of their infrastructure resources,
- Protects electronic commerce sites from performance degradation and service failure/outages by instituting admission control, and
- Guarantees response time and the corresponding quality-of-service/experience to the admitted users.

The two inputs of these admission policies—customers' purchasing histories and their current location and status within the E-commerce site—must be available in real time for making admission policy decisions, which is also done in real time. Since both inputs are stored and updated on the merchants' sites, they can be uploaded to an admission control module in real time for decision making.

The admission control module can be an add-on in the load balancer. The load balancer, in addition to analyzing the resource consumption of the servers (i.e., loads, CPU, IO, memory), also processes the customer identities and their admission policies/inputs, and maps these customer requests to a set of priority queues for switching and load balancing, and overload protection.

All these various components for E-commerce performance management are currently available piecemeal. The load balancer is available, so are the billing infrastructures that capture and summarize the purchasing histories of E-customers, so are the tracking software that records status and locations of E-customers in real time. Putting all these together into a powerful platform and framework for E-commerce performance management is forthcoming. When that happens, the greed of electronic merchants can be satisfied, and their infrastructures will be protected and their performance optimized. As for the on-line users: don't expect the best services if you are not a regular customer or a big spender.

So it is that greed guides admission which optimizes performance which maximizes profits.

L. Lawrence Ho

Dr. Lawrence Ho is a Principal Investigator in the Networking Research Laboratory of Bell Labs Research of Lucent Technologies. He received his Ph.D. from Yale University in December 1996. He was awarded

the 1996 William D. Carey Annual Science Award by the American Association for the Advancement of Science (AAAS). Since joining Bell Labs in the late 1996, he has been performing research and building software in monitoring and control of data (IP), transaction-oriented, wireless, and electronic commerce networks and services, spanning algorithms, implementation, and experimental studies. Dr. Ho has authored more than 20 technical publications, and has served on the technical program committees of many international conferences. He is also an editor of the Journal of Network and Systems Management, to which he regularly contributes a column called Toolbox.

Figure 1. Typical E-Commerce Infrastructure.

