Re-Evaluating the Evaluation — Does Performance Still Matter?

Christopher Stewart and Kai Shen
Department of Computer Science, University of Rochester

Is your contribution better than the alternatives? This simple-yet-daunting question has killed more research proposals than The Plague. The answer, in the systems community, has historically involved a direct comparison of raw performance. Faster response time, higher throughput, and lower CPI have been good heuristics to arrive at the question's hard-to-quantify intent: enhancing end-user experience. But times have changed, in particular, the average-case performance of today's Internet services has already exceeded the demands of end-users. Research proposals that further improve the raw performance will not significantly affect end-users' perceived quality of service.

It may sound wild and crazy, but raw performance should no longer be used to evaluate server system research. Instead, research proposals should focus on other aspects of the end-user experience. In particular, we discuss the importance of performance dependability, *i.e.*, performance that consistently matches user expectations.

Today's Average-Case Performance Is Satisfactory End-users interact with Internet server systems by issuing queries and waiting for responses. While end-users (humans) can perceive subtle performance differences, they report qualitative differences only when the response time crosses certain thresholds [3]. For instance, end-users can discern the difference between a response time of 1 and 2 seconds, but they may categorize the performance of both similarly. Studies in the field of human-computer interactions have characterized the response time thresholds that mark qualitative performance differences (*e.g.*, around 4 seconds [3] and 5 seconds [2]). Figure 1 compares such thresholds to the actual average request response time at 15 commercial Internet services [4] that serve static and dynamic content under moderate (Mother's Day) and heavy (holiday season) workloads. *The average-case performance of today's commercial Internet services exceeds basic end-user demands, even under heavy workloads.*

The $95^{\mathbf{th}}$ Percentile: Performance is Still Inconsistent Despite excellent average-case performance, response times still vary significantly from request-to-request and over time. For instance, the 95^{th} percentile of hourly average response times was 30%–182% larger than the average-case response times for the commercial dynamic-content sites referenced above [4]. Such inconsistency violates the end user's expectations, and, as a result, degrades their perception of the service quality [3]. The effect is disproportionate; a few abnormally slow responses can ruin the user's perception of an entire service [1]. Further, end users that frequently query a service (*e.g.*, repeat shoppers) are more likely to eventually encounter such inordinate delays which degrades the perceived service quality to those who matter the most.

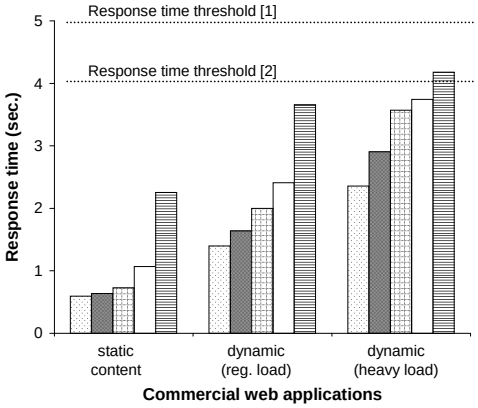

Figure 1: Average request response time of real-world Internet services (data source: [4]).

We contend that research proposals should be evaluated by their ability to produce consistent and dependable performance. For example:

- Merge sort can be better than quick sort.
- Request scheduling should strive for uniformity rather than maximum per-request performance.
- System maintenance should be handled by long-running background processes, rather than intensive end-of-the-day batch jobs.

In conclusion, raw performance doesn't matter anymore for Internet services, but the consistency and dependability of performance does.

References

[1] N. Bhatti, A. Bouch, and A. Kuchinsky. Integrating user-perceived quality into web server design. *Computer Networks*, 33, June 2000.

[2] A. Bouch, A. Kuchinsky, and N. Bhatti. Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service. In *Proc. of the Conf. on Human Factors in Computing Systems*, pages 297–304, Apr. 2000.

[3] R. Miller. Response time in man-computer conversational transactions. In *Proc. of the AFIPS Fall Joint Computer Conf.*, pages 267–277, Dec. 1968.

[4] `www.websitepulse.com`. Website monitoring and web server monitoring, 2007.