Workload Characterization

Daniel A. Menascé • *George Mason University* • *menasce@cs.gmu.edu*

Analyzing an e-commerce site's scalability requires a solid understanding of its workload characteristics, but this characterization must occur at multiple levels and time scales. My colleagues from Brazil's Federal University of Minas Gerais, Rice University, George Mason University, and I completed several workload characterization studies, including auction sites.[1–4] Here, I present the general method we adopted and a few examples of the workload characterizations we studied.

Approach

We can characterize workload at four levels: business, session, function, and protocol. We can view an e-business's workload in a multilayer hierarchical way. Each layer has many features:

- *Business layer.* We can examine a business's overall characteristics and how they affect the way users interact with the site in this layer. For an auction site, for example, we would be interested in the number of bids the winner placed, the point at which the winner placed his or her first bid, the closing price's evolution over the auction's life, and the percentage of auctions that have winners.
- *Session layer.* We define a session as the sequence of requests a customer makes during a single visit to an e-business site. Thus, the session layer deals with characteristics such as duration (measured in the number of requests per session), navigation patterns within a session, and the buy-to-visit ratio (the probability that a session will result in a sale). Customer behavior model graphs (CBMGs) and customer visit models (CVMs) are two examples of models used to capture the way customers invoke the various functions an e-business site offers.[4,5]
- *Function layer.* An e-business site offers many functions to customers. An auction site, for example, provides `browse`, `search`, `register`, `login`, `view bid`, `bid`, and `sell` functions.
- *HTTP request layer.* Customers interact with e-commerce sites through the HTTP protocol; characterization at this level deals with the workload's features in terms of HTTP requests.

Although performance analysts and capacity planners can use this hierarchy to get a comprehensive view of an e-business site's workload, analyzing how some workload features change at different time scales is crucial. What do we learn when we change from hours to minutes or from minutes to seconds? Let's examine each level in more detail.

Business-Level Characterization

We analyze the workload at the business level to determine which aspects are common to all sessions and which ones influence the business's profitability and the site's overall revenue throughput (measured here as the number of dollars generated per time unit).[5]

Consider Table 1, which shows summary statistics collected from Yahoo!'s auction site during a two-week period for five categories of items.[6] The table shows that proxy agents placed 55 percent of all bids and humans placed the remaining 45 percent. Ten percent of the auctions received at least one bid; approximately 1.4 percent of the auctions received at least 10. Only 1.4 percent of the 26,910 auctions had a winner.

Session-Layer Characterization

The number of sessions submitted to an e-business site is huge, so grouping, or *clustering*, similar sessions can help characterize the site's workload. We use two classes of clustering algorithms for this purpose: distance-based and fractal-based.[2,4] Distance-based algorithms (such as the *k*-means clustering algorithm) group sessions according to a defined distance between the sessions. If we use a CBMG to represent different sessions, an underly-

Table 1. Business-layer characterization for an auction site.

	Count	Percentage
Total number of auctions	26,910	
Total number of bids	19,381	
Bids placed by humans	8,798	45.4
Bids placed by proxy agents	10,583	54.6
Auctions with one or more bids	2,715	10.1
Auctions with 10 or more bids	375	1.4
Auctions with a winner	385	1.4

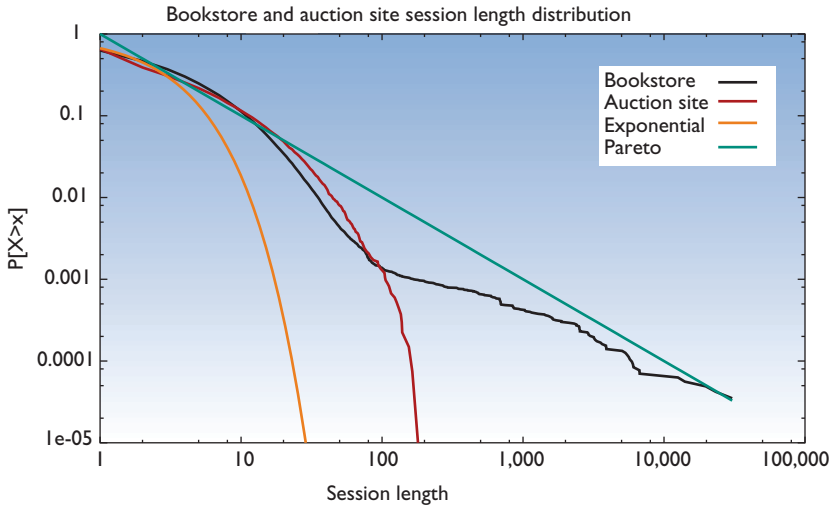

Figure 1. Log-log graph of the session length distribution's tail measured in the number of requests to execute the site's functions. Data for an online bookstore and an auction site appear along with the tail of a Pareto (a heavy-tailed distribution) and an exponential distribution.

ing matrix indicates each session's probabilities of transition between states. We can then use the Euclidean distance between the matrices representing two separate CBMGs for the clustering process.[4]

Distance-based clustering has its limitations — for example, all clusters must have regular geometrical shapes. Fractal-based clustering offers an improvement in quality: it forms clusters with any arbitrary shape.[2] Fractal-based clustering allocates each point to the cluster whose fractal dimension changes the least by the inclusion of that point. A more thorough discussion of a data set's fractal dimension appears elsewhere.[2]

Two important aspects of session-layer workload are session length distribution and a multiple time scale analysis of the number of sessions initiated per time unit. Figure 1 shows a log-log plot of the session length distribution's tail (the probability that the length exceeds a given value) for an online bookstore and an auction site. Here, we measure session length as the number of e-business functions invoked per session. Figure 1 also shows the tail of a Pareto and an exponential distribution; Pareto has a heavy tail and therefore its tail is a straight line in a log-log scale. The bookstore has a much heavier tail than the online auction site, mainly due to long, robot-initiated sessions.

Figure 2 shows the number of sessions initiated per day and per hour over a 16-day period on an online

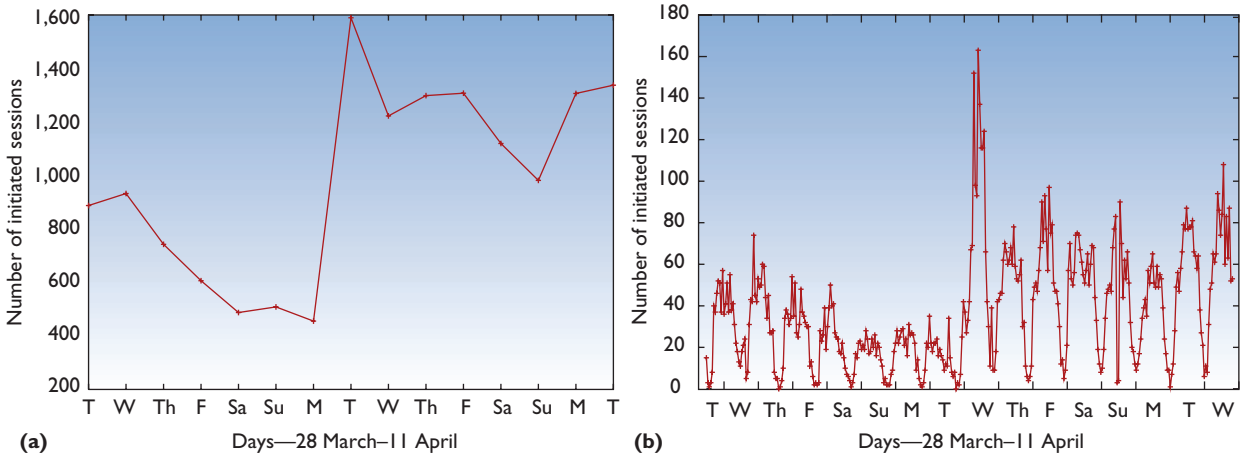

Figure 2. Multiple time-scale analysis of the number of sessions initiated. The figure shows (a) the number of initiated sessions per day and (b) the number of initiated sessions per hour for an online auction site over a 16-day period.

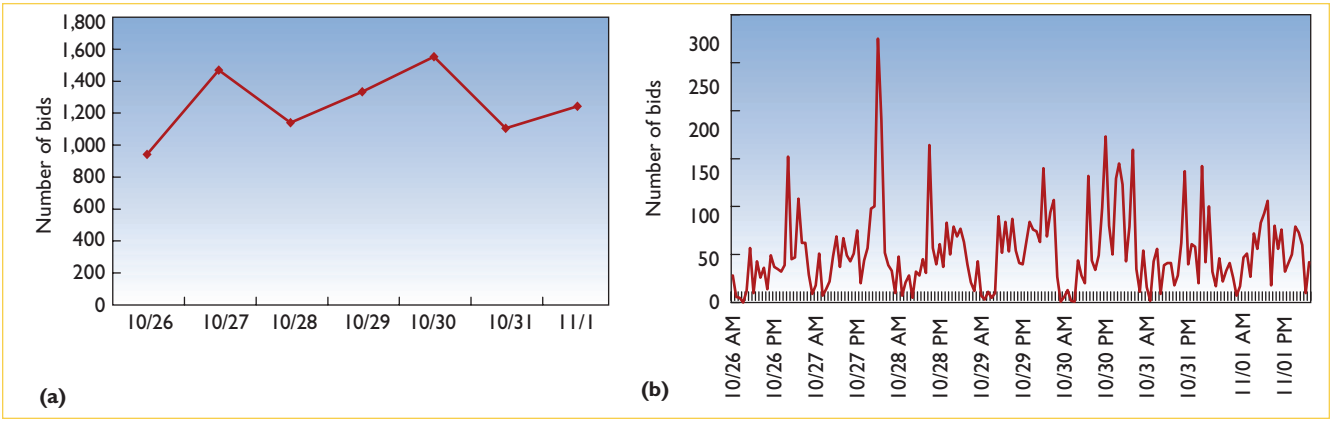

Figure 3. Multiple time-scale analysis of the number of bids on an online auction site over a seven-day period. The figure shows (a) the number of bids placed by all customers on a one-day time scale and (b) the same data on a one-hour time scale.

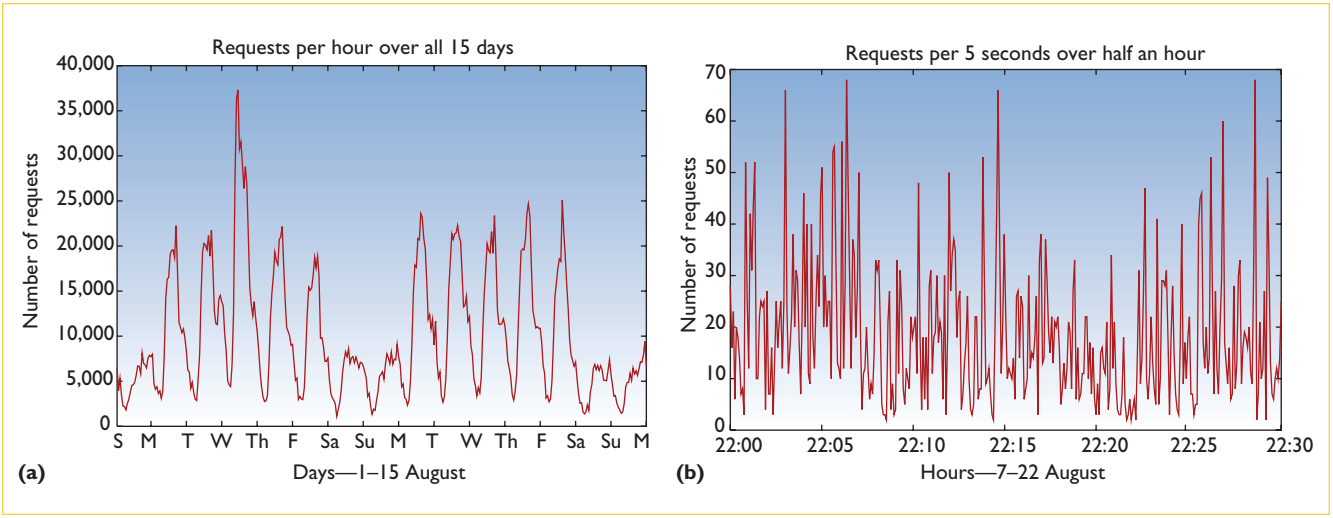

Figure 4. Multiple time-scale plots of the number of arriving HTTP requests. The figure shows times scales of (a) one hour and (b) five seconds.

auction site that sells domain names. The graph of Figure 2a clearly shows daily changes in activity, but fails to capture the spikes that occur at particular hours of the day.[1] Knowledge about traffic spikes is important for capacity planning — the process of predicting when site resources must be upgraded to provide adequate quality of service under increased workload-intensity levels.[5]

Function-Layer Characterization

Workload characterization at the function layer includes a breakdown of the number of e-business functions customers invoke during their sessions as well as a global analysis across all users and several time scales.

Figure 3 shows a multiple time-scale analysis of the number of bids submitted to the Yahoo! auction site over a seven-day period for five categories of items.[6] Figure 3a shows the number of bids placed by all customers at a one-day time scale; Figure 3b shows the same data but on a one-hour time scale. A multiscale analysis shows a twofold increase in bid-submission traffic toward the end of the day.

Request-Layer Characterization

Plotting the number of HTTP requests at several time scales — such as one hour and five seconds — to conduct a visual inspection is a first step to understanding the process by which requests arrive at the site.[1] Figure 4 shows a visual inspection of the number of requests arriving per time slot on different time scales; even to a novice, an apparently strong dependence shows long sequences of increased or decreased volume (trends), particularly at intermediate time scales.

We can perform a multiple time-scale quantification of the arrival process's dependence on HTTP requests by drawing a variance time plot (VTP) for different time scales. The VTP plot is a log-log plot of the sample variance against the time

scale; it helps detect and quantify self-similarity.[1]

Final Remarks

Workload characterization is one of the most important steps in the process of planning an e-business site's capacity.[5] This step's output is a workload model that captures the workload's essential features. In addition to their role in capacity planning, we can use good and representative workload models to design synthetic workload generators that mimic the workload's behavior. We can then use these generators as benchmarks to compare competing systems and architectures.

References

1. D.A. Menascé et al., "A Hierarchical and Multiscale Approach to Analyze E-Business Workloads," *Performance Evaluation*, vol. 54, no. 1, 2003, pp. 33–57.
2. D.A. Menascé et al., "Characterizing E-Business Workloads Using Fractal Methods," *J. Web Eng.*, vol. 1, no. 1, 2002, pp. 74–92.
3. D.A. Menascé et al., "In Search of Invariants for E-Business Workloads," *Proc. 2000 ACM Conf. E-Commerce*, ACM Press, 2000, pp. 56–65.
4. D.A. Menascé et al., "A Methodology for Workload Characterization for E-Commerce Servers," *Proc. 1999 ACM Conf. Electronic Commerce*, ACM Press, 1999, pp. 119–128.
5. D.A. Menascé and V.A.F. Almeida, *Scaling for E-Business: Technologies, Models, Performance, and Capacity Planning*, Prentice Hall, 2000.
6. D.A. Menascé and V. Akula, *Towards Workload Characterization of Auction Sites*, tech. report, E-Center for E-Business, George Mason Univ., July 2003.

Daniel A. Menascé is a professor of computer science, the co-director of the E-Center for E-Business, and the director of the MS in E-Commerce program at George Mason University. He received a PhD in computer science from UCLA and published the books *Capacity Planning for Web Services* and *Scaling for E-Business* (Prentice Hall, 2002 and 2000). He is a fellow of the ACM and a recipient of the A.A. Michelson Award from the Computer Measurement Group.