Appendix G
Response Time

William Stallings

Supplement to
Operating Systems, Eighth Edition
Pearson 2014
http://williamstallings.com/OperatingSystems/

Response time is the time it takes a system to react to a given input. In an interactive transaction, it may be defined as the time between the last keystroke by the user and the beginning of the display of a result by the computer. For different types of applications, a slightly different definition is needed. In general, it is the time it takes for the system to respond to a request to perform a particular task.

Ideally, one would like the response time for any application to be short. However, it is almost invariably the case that shorter response time imposes greater cost. This cost comes from two sources:

- **Computer processing power:** The faster the processor, the shorter the response time. Of course, increased processing power means increased cost.
- **Competing requirements:** Providing rapid response time to some processes may penalize other processes.

Thus the value of a given level of response time must be assessed versus the cost of achieving that response time.

Table G.1, from [MART88] lists six general ranges of response times. Design difficulties are faced when a response time of less than 1 second is required. A requirement for a subsecond response time is generated by a system that controls or in some other way interacts with an ongoing external activity, such as an assembly line. Here the requirement is straightforward. When we consider human-computer interaction, such as in a data entry application, then we are in the realm of conversational response time. In this case, there is still a requirement for a short response time, but the acceptable length of time may be difficult to assess.

Table G.1 Response Time Ranges

Greater than 15 seconds

This rules out conversational interaction. For certain types of applications, certain types of users may be content to sit at a terminal for more than 15 seconds waiting for the answer to a single simple inquiry. However, for a busy person, captivity for more than 15 seconds seems intolerable. If such delays will occur, the system should be designed so that the user can turn to other activities and request the response at some later time.

Greater than 4 seconds

These are generally too long for a conversation requiring the operator to retain information in short-term memory (the operator's memory, not the computer's!). Such delays would be very inhibiting in problem-solving activity and frustrating in data entry activity. However, after a major closure, such as the end of a transaction, delays from 4 to 15 seconds can be tolerated.

2 to 4 seconds

A delay longer than 2 seconds can be inhibiting to terminal operations demanding a high level of concentration. A wait of 2 to 4 seconds at a terminal can seem surprisingly long when the user is absorbed and emotionally committed to complete what he or she is doing. Again, a delay in this range may be acceptable after a minor closure has occurred.

Less than 2 seconds

When the terminal user has to remember information throughout several responses, the response time must be short. The more detailed the information remembered, the greater the need for responses of less than 2 seconds. For elaborate terminal activities, 2 seconds represents an important response-time limit.

Subsecond response time

Certain types of thought-intensive work, especially with graphics applications, require very short response times to maintain the user's interest and attention for long periods of time.

Decisecond response time

A response to pressing a key and seeing the character displayed on the screen or clicking a screen object with a mouse needs to be almost instantaneous—less than 0.1 second after the action. Interaction with a mouse requires extremely fast interaction if the designer is to avoid the use of alien syntax (one with commands, mnemonics, punctuation, etc.).

That rapid response time is the key to productivity in interactive applications has been confirmed in a number of studies [SHNE84; THAD81; GUYN88]. These studies show that when a computer and a user interact at a pace that ensures that neither has to wait on the other, productivity increases significantly, the cost of the work done on the computer therefore drops, and quality tends to improve. It used to be widely accepted that a relatively slow response, up to 2 seconds, was acceptable for most interactive applications because the person was thinking about the next task. However, it now appears that productivity increases as rapid response times are achieved.

The results reported on response time are based on an analysis of online transactions. A transaction consists of a user command from a terminal and the system's reply. It is the fundamental unit of work for online system users. It can be divided into two time sequences:

- **User response time:** The time span between the moment a user receives a complete reply to one command and enters the next command. People often refer to this as think time.
- **System response time:** The time span between the moment the user enters a command and the moment a complete response is displayed on the terminal.

As an example of the effect of reduced system response time, Figure G.1 shows the results of a study carried out on engineers using a computer-aided design graphics program for the design of integrated circuit chips and boards [SMIT83]. Each transaction consists of a command by the engineer that alters in some way the graphic image being displayed on the screen.

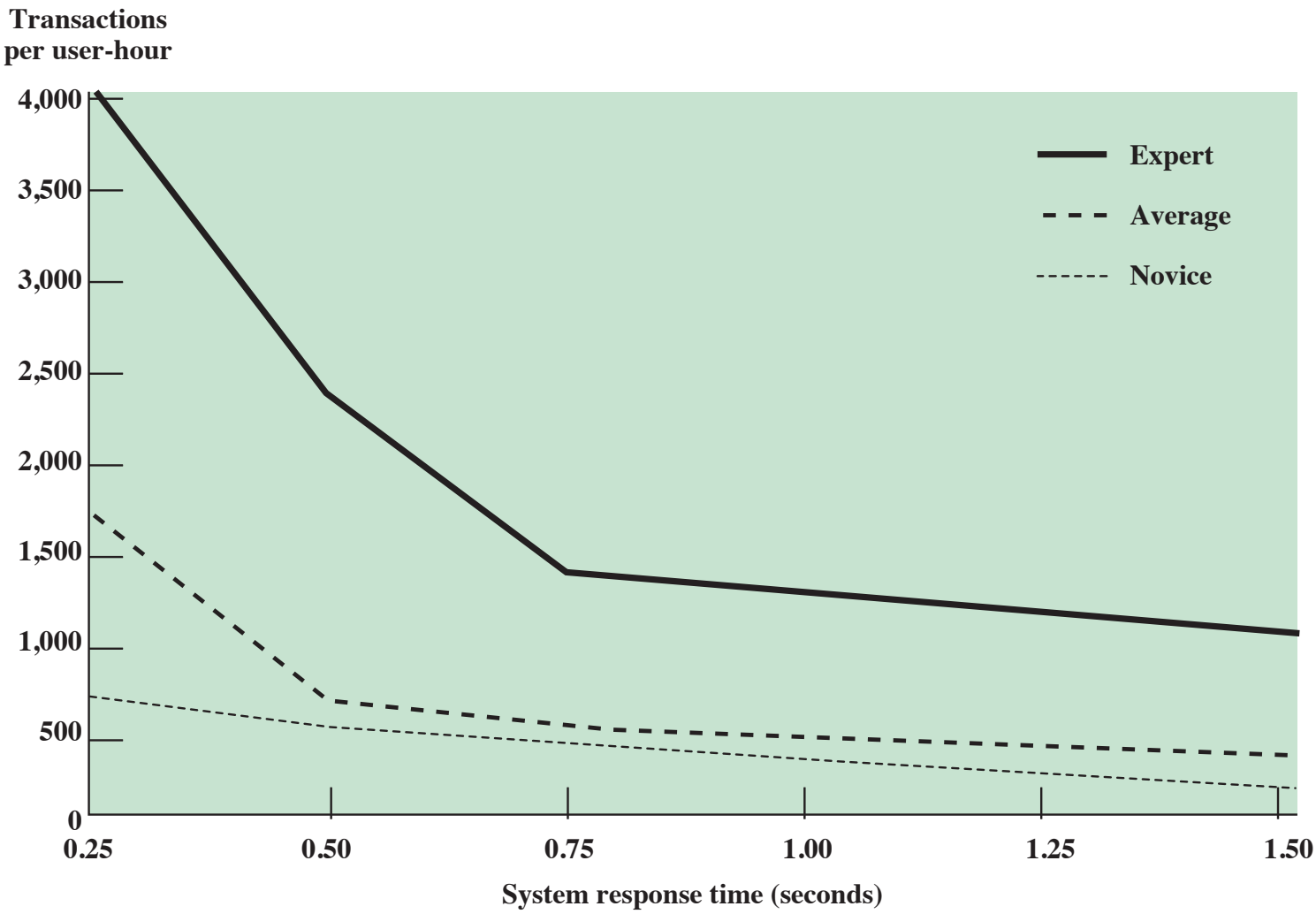

Figure G.1 Response Time Results for High-Function Graphics

The results show that the rate of transactions increases as system response time falls and rises dramatically once system response time falls below 1 second. What is happening is that as the system response time falls, so does the user response time. This has to do with the effects of short-term memory and human attention span.

Another area where response time has become critical is the use of the World Wide Web, either over the Internet or over a corporate intranet. The time it takes for a typical Web page to come up on the user's screen varies greatly. Response times can be gauged based on the level of user involvement in the session; in particular, systems with vary fast response times tend to command more user attention. In a study by Sevcik [SEVC96, SEVC02], illustrated in Figure G.2, Web systems with a 3-second or better response time maintain a high level of user attention. With a response time of between 3 and 10 seconds, some user concentration is lost, and response times above 10 seconds discourage the user, who may simply abort the session. Other studies of Web response time generally confirm these findings ([BHAT01].

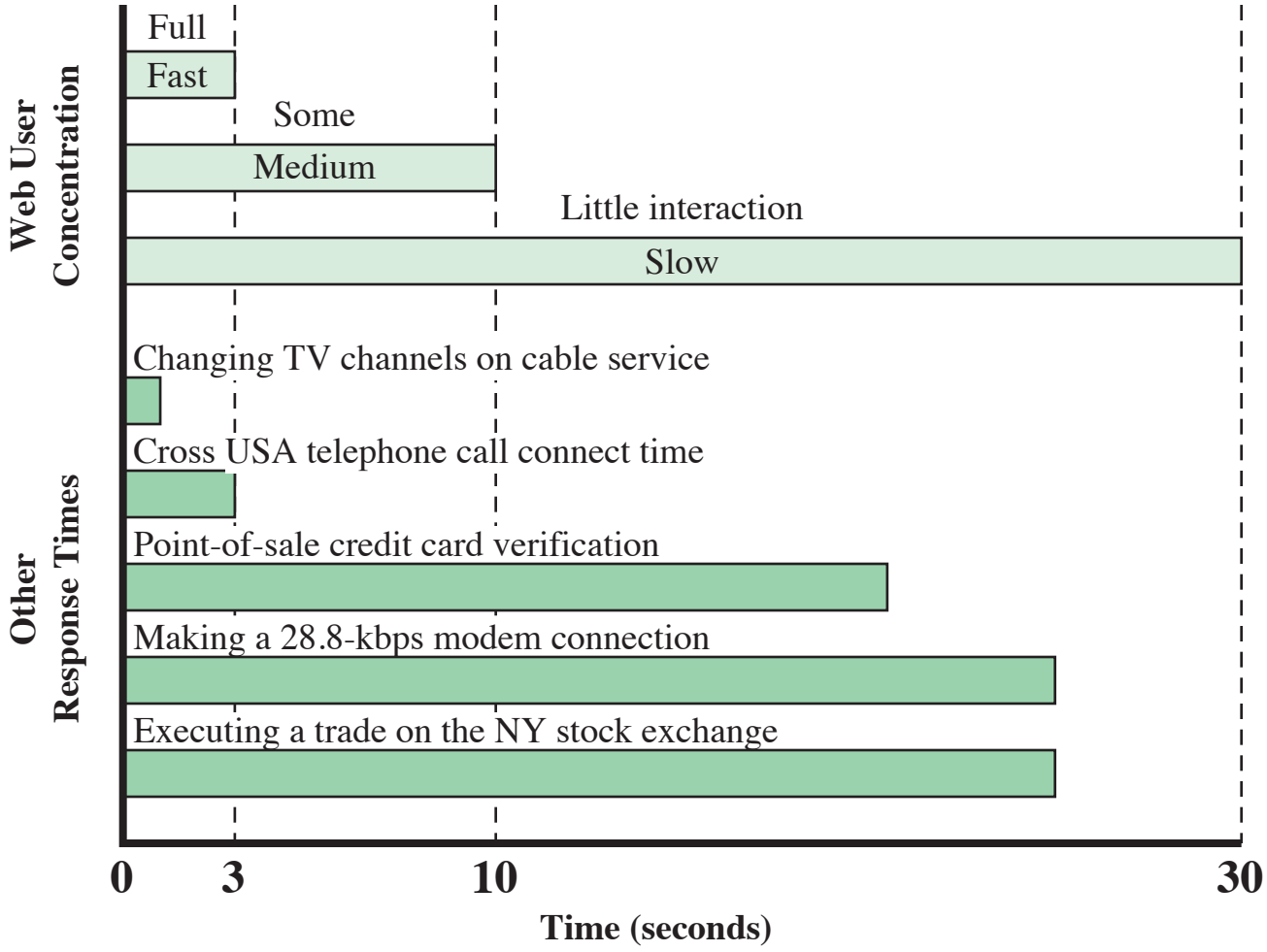

Figure G.2 Response Time Requirements

References

BHAT01 Bhatti, N.; Bouch, A.; and Kuchinsky, A. "Integrated User-Perceived Quality into Web Server Design." Proceedings, 9th International World Wide Web Conference, May 2000.

GUYN88 Guynes, J. "Impact of System Response Time on State Anxiety." *Communications of the ACM,* March 1988.

MART88 Martin, J. *Principles of Data Communication.* Englewood Cliffs, NJ: Prentice Hall, 1988.

SELV99 Selvidge, P. "How Long is Too Long to Wait for a Webpage to Load." *Usability News*, Wichita State University, July 1999.

SEVC96 Sevcik, P. "Designing a High-Performance Web Site." *Business Communications Review*, March 1996.

SEVC02 Sevcik, P. "Understanding How Users View Application Performance." *Business Communications Review*, July 2002.

SHNE84 Shneiderman, B. "Response Time and Display Rate in Human Performance with Computers." *ACM Computing Surveys*, September 1984.

SMIT83 Smith, D. "Faster Is Better: A Business Case for Subsecond Response Time." *Computerworld*, April 18, 1983.

THAD81 Thadhani, A. "Interactive User Productivity." *IBM Systems Journal*, No. 1, 1981.