Chapter 11

Quality of Service

Introduction

- 1960s: user perception of computer speed =
  - response time for mathematical computations, program compilations, or database searches
- Time-shared systems: more reasons for delays
  - contention for computational resources
- World wide web: more complex reasons for delays;
  - more graphics, network congestion, multiple sources of dropped connections
- All these concerns are usually discussed under the term *Quality of Service (QoS)*

- QoS stems from basic human values ► ► ►

Introduction

Basic human values:

- "Time is precious"
  - Lengthy or unexpected system response time can produce frustration, annoyance, and eventual anger
  - *which lead to frequent errors and low satisfaction*

► Introduction

Basic human values (cont.):

- "Harmful mistakes should be avoided"
  - This may sometimes means the pace of work must slow.
  - Speedy and quickly done work can result in users:
    - learning less
    - reading with lower comprehension
    - making more ill-considered decisions
    - committing more data-entry errors

    - *Stress can build in all these situations, especially if the damage is big.*

Introduction

Basic human values (cont.):

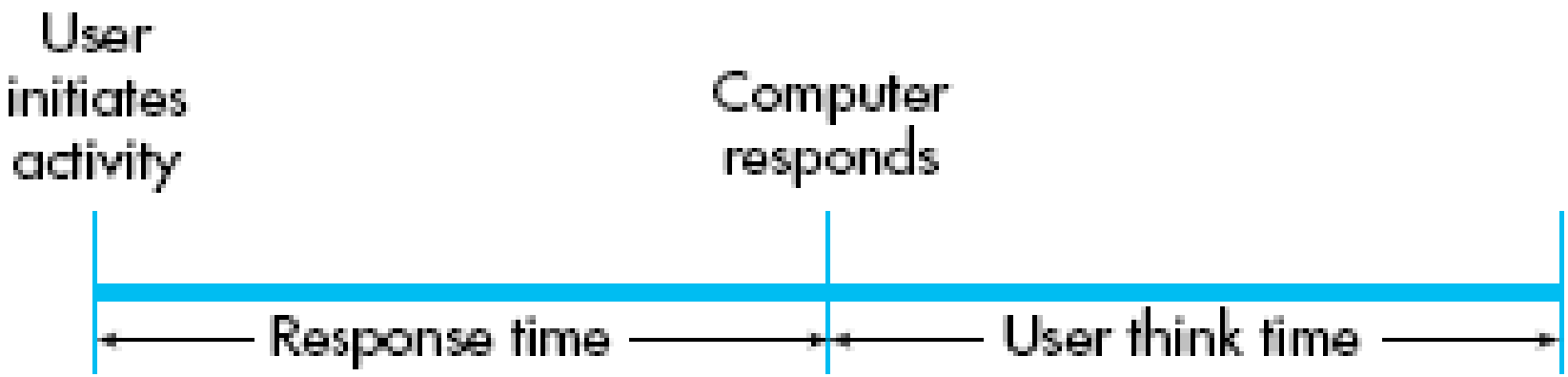
- "Reduce user frustration"
  - Frustration results in making mistakes and giving up working
  - Causes of frustration:
    - Long delays
    - Crashes that destroy data
    - Software bugs that produce incorrect results
    - Poor design that lead to user confusion
  - Network environments generate further frustrations:
    - Unreliable service providers
    - Dropped lines
    - Email spam, and viruses

Introduction

- Quality of service is mostly effected by decisions made by
  - Network designers and operators
  - Interface designers and builders
    - reduce byte count for web pages
    - reduce number of queries and access to the network
    - Users may have the opportunity to choose from fast or slow services and from viewing low-resolution versus high-resolution images
- For users the main concern for quality of service is computer *response time*.

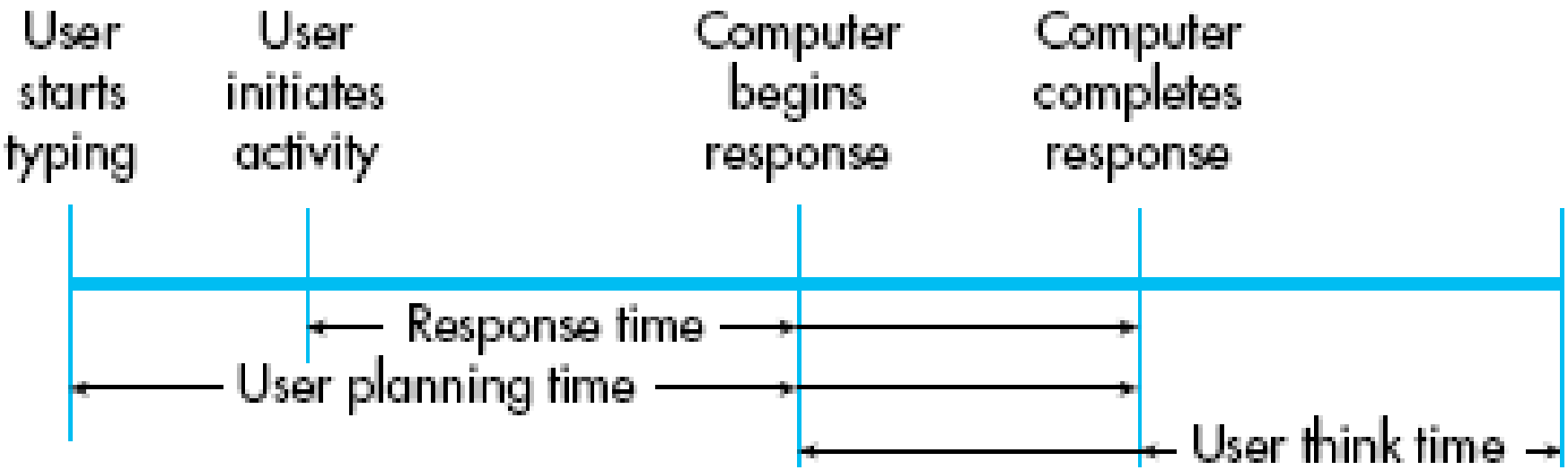
Models of response-time impacts

- Simple model of response time
  - Users (1) initiate, (2) wait for response, (3) watch results, (4) think for a while, and initiate again
    - Response time (?)
    - Think time (?)



Models of response-time impacts

- More realistic model of response time
  - People will use whatever time they have to plan ahead



Models of response-time impacts

- Overall majority of users prefer rapid interactions, however, overall *productivity* depends on
  - interaction speed
  - error rates
  - ease of recovery from errors
- Lengthy response times (>15 seconds) are harmful to productivity
  - increasing error rates and decreasing satisfaction
- Rapid response times (1 second or less) are preferable, but can increase errors for complex tasks if the user does not spent sufficient time to think.

- The high cost of providing rapid response times and the loss from increased errors must be evaluated in the choice of an optimum pace

Models of response-time impacts

- Display Rate
  - Alphanumeric displays: The speed in characters per second at which characters appear for the user to read. e.g., 120cps for mobile devices
  - World Wide Web Applications: Bytes/Sec. e.g, 56Kbs for modems
    - Display rate may be limited by network transmission speed or server performance
- Reading textual information from a screen is a challenging cognitive task
  - Users relax when the screen fills instantly
  - It is useful to display text first, leaving space for the graphical elements

► Models of response-time impacts

Limitations of short-term and working memory

- Magic number **7±2** (George Miller, 1956)
  - The average person can rapidly recognize seven chunks of information at a time
  - This information can be held for 15 to 30 seconds in short-term memory
  - Size of the chunks depends on the person' s familiarity with the material
- *Short-term memory* and *working memory* are used in conjunction for processing information and problem solving
  - Short-term memory processes perceptual input
  - Working memory generates and implements solutions
- People learn to cope with complex problems by developing higher-level concepts using several lower-level concepts brought together into a single *chunk*
- Short term and working memory are highly volatile
  - Disruptions cause loss of memory
  - Delays require that memory be refreshed
  - Visual distractions, noisy environments, and anxiety interfere with cognitive processing

Models of response-time impacts

- When using an interactive computer system users may formulate plans and have to wait for execution time of each step

- If there is an unexpected result (error), or long delay, then users may forget part of the plan or be forced to review the plan continually

► Models of response-time impacts

- For a given user and task, there is a preferred response time

Long response time	**Short response time**
■Lead to wasted effort and more errors, because the solution plan must be reviewed repeatedly	■May generate a faster pace in which solution plans are prepared hastily and incompletely
■Causes uneasiness because the penalty for error increases	■The user may pick up the pace of interface and fail to fully comprehend the presented materials

Models of response-time impacts

- A related issue is:
  - Performance in paced vs. unpaced tasks
- The car speed limit analogy:
  - More speed more accidents
  - Progress indicators result in higher satisfaction and shorter perceived elapsed time

Models of response-time impacts

- Rapid task performance, low error rates, and high satisfaction can come from:
  - Users have adequate knowledge of the objects and actions necessary for the problem-solving task
  - The solution plan can be carried out without delays
  - Distractions are eliminated
  - There is feedback about progress toward solution
  - Errors can be avoided or handled easily

Models of response-time impacts

- Other conjectures in choosing the optimum interaction speed
  - Novices may exhibit better performance with slower response time
  - Novices prefer to work at slower speeds
  - With little penalty for an error, users prefer to work more quickly
  - When the task is familiar and easily comprehended, users prefer more rapid action
  - If users have experienced rapid performance previously, they will expect and demand it in future situations

Expectations and attitudes

- How long will users wait for the computer to respond before they become annoyed?

- Related design issues may clarify the question of acceptable response time
  - E.g. how long before hearing a dial-tone
- Two-second limit (Miller, 1968) appropriate for many tasks
- But users have adapted a working style and expectation based on responses within a fraction of a second. e.g., key typed, wheel turn, …
- In other situations, users are accustomed to longer response times. e.g., traffic light

Expectations and attitudes

Factors influencing acceptable response time:

1. People have established expectations based on their past experience for a given task.
   - What would be your reaction when the system response is:
     - Almost as you expected
     - Later than expected
     - Sooner than expected
     - Very much sooner than expected
   - Response-time choke
     - A system is slowed down when the load is light and potential performance high
     - Makes the response time more uniform over time and across users, avoiding expectations that can't always be met
   - Rapid start-up
     - tradeoff between start-up vs. usage

► Expectations and attitudes

Factors influencing acceptable response time: (cont)

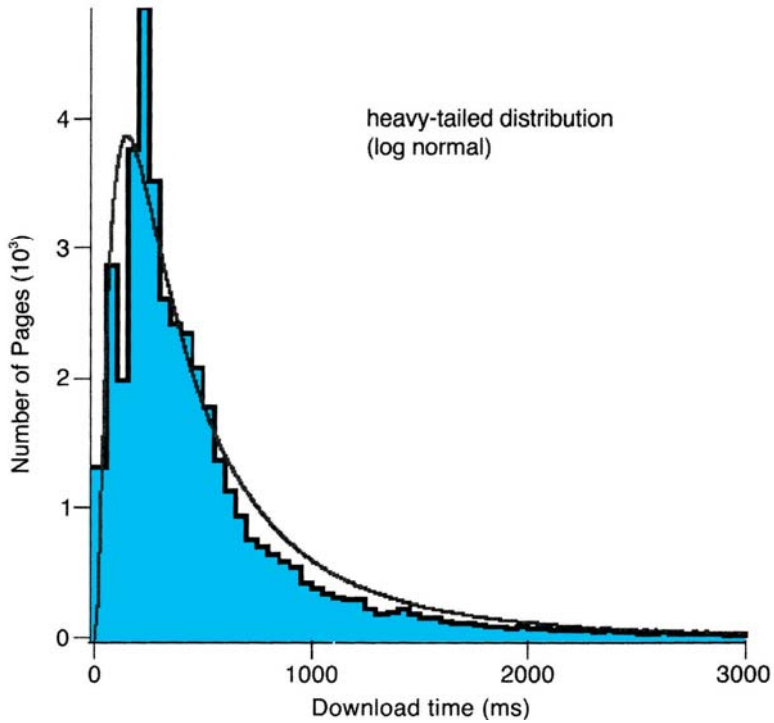
2. The individual tolerance for delays
   - Novice users maybe willing to wait much longer
   - There are large variations in what individual consider acceptable waiting time
     - These variations are influenced by many factors: personality, age, mood, …

3. Task complexity
   - In simple repetitive tasks, users want to perform rapidly

Expectations and attitudes

- Some tasks place high demands on rapid system performance
  - e.g., User-controlled 3D animations, simulators, VoIP telephony

- The range of response time is highly varied across web sites
  - As response times increase, users find web-page content less interesting and lower in quality
  - It may affect a company's image



► Expectations and attitudes

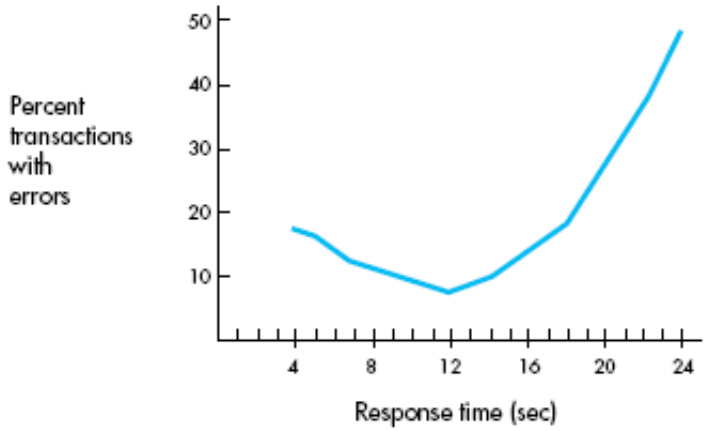
- In summary, three conjectures emerge:
  1. Individual differences are large and users are adaptive. They will work faster as they gain experience and will change their working strategies as response time change. It may be useful to allow people to set their own pace of interaction (e.g., in games)
  2. For repetitive task, user prefer and will work more rapidly with short response times.
  3. For complex tasks, users can adapt to working with slow response time with no loss of productivity, but their dissatisfaction increases as response time lengthen.

User productivity

- Shorter response times usually lead to higher productivity
  - but at longer response times, users can find ways to do concurrent processing to reduce effort and time
- Nature of the task has a strong influence on whether changes in response time alter user productivity
- Repetitive tasks
  - Shorter response time means users responds more quickly
    - decisions may not be optimal, but penalty for a poor choice is small
  - Goodman and Spence (1981) – reduced response time lead to more productivity
  - Teal and Rudnecky (1992) – slower response time lead to more accuracy

User productivity

- Complex problem solving tasks
  - Users will adapt their work style to the response time
  - Grossberg, Wiesen, and Yntema (1976) – the time to solution was invariant with respect to response time
  - Barber and Lucas (1983) – error rates were lowest as 12 sec response time, but productivity increased linearly with reduction in response time.



- Summary
  - Users pick up the pace of the interface, and they constantly prefer a faster pace
  - Error rates with shorter response time increase in complex tasks.
  - Each task appears to have an optimal pace for lowest errors

Skipped sections

- The following sections have been skipped
  - 11.5 Variability in response time
  - 11.6 Frustrating experiences