Análise e Modelagem de Desempenho de Sistemas de Computação

Virgilio A. F. Almeida

2º Semestre de 2005
Revisão de Limites da Performane
Bounding Analysis - Class 3
Bounds on Performance

• Let us derive some simple bounds on system throughput
  • e.g., How high or low can throughput be as a function of the number of customers in the model (N)?

• The key is bottleneck analysis, that means: identifying which component of the system will saturate first and what system throughput will be at that point.
Asymptotic Bounds

- Let us consider a system model with $K$ components (e.g., processors, IO devices, etc). The service demands are $D_1, D_2 \ldots D_K$. A hard upper bound on throughput is given by $1/D_{\text{max}}$, where $D_{\text{max}}$ is the service demand of the bottleneck component.
  - $U_i = D_i \times X_0$
  - $U_i \leq 100\% \rightarrow D_i \times X_0 \leq 100\% \rightarrow X_0 \leq 1/D_i$
  - $D_{\text{max}} = \max\{D_1, D_2 \ldots D_K\}$
  - $X_0 \leq 1/D_{\text{max}}$
Asymptotic Bounds

- Let us assume now that there are no delays in the model, execution time equals service demand
  - \( D = \sum D_i \)
  - By Little’s Law \( X(N) = N \times RU \)
  - \( X(1) = 1/D \)
Asymptotic Bounds

- As $N$ increases from 1 to 2... The **best** thing that can happen is that the customer never runs into each other in the components. The **worst** thing that can happen is they always run into each other.
A C/S System

- Consider that a transaction $t$ in a C/S system uses 5 msec of CPU at the browser, 10 msec of CPU at the server, and reads ten 2048-byte blocks from the server’s disk.
- The average seek time at the disk is 9 msec, the average latency is 4.17 msec and the transfer rate is 20 MB/sec.
- Consider that the client and server are connected by a 10 Mbps Ethernet and that a request going from the client to the server takes a full packet (1518 bytes) and the reply requires 7 packets.
C/S System Response Time

• What is the minimum response time?
C/S System Response Time

- What is the minimum response time?

\[ R_r \geq D_{\text{client}} + D_{\text{network}} + D_{\text{server}} \]

- we are ignoring all waiting times.
Calculating the Service Demands (1)

- At the client
  - $D_{\text{client}} = D_{\text{ccpu}} = 5$ msec

- At the network
  - $D_{\text{network}} = \frac{(m_1 + 7m_2)}{B}$
    - $m_1 = m_2 = 1518$ bytes
    - $B = 10$ Mbps
  - $D_{\text{network}} = 0.0097$ msec
Calculating the Server's Service Demands (2)

- \( D_{server} = D_{cpu} + D_{disk} \)
- \( D_{disk} = 10 \times S_{disk} \)
  
  \[
  S_{disk} = \text{AvgSeek} + \text{AvgLatency} + \text{TransferTime} \\
  = 0.009 + 0.0047 + \frac{2048}{20,000,000} \\
  = 0.0133 \text{ sec}
  \]

  \[
  D_{disk} = 10 \times S_{disk} = 0.133 \text{ sec}
  \]

- \( D_{cpu} = 0.010 \text{ sec} \)
- \( D_{server} = 0.143 \text{ sec} \)
A C/S System: solution

- What is the minimum response time?

\[ R_r \geq D_{client} + D_{network} + D_{server} \]

\[ R_r \geq 0.005 + 0.0097 + 0.143 = 0.158 \text{ sec.} \]
Client/Server Response-Time Law

How can we model this system to calculate response time?

N – no. of client processes
Z - average time between client requests (think time)
D - servicedemand at server
N_{server} – avg no of requests at server
N_{client} – avg no of clients withou outstanding requests ("thinking")
R_0 – mean response time
X_0 - system throughput
How to model the C/S system?
How to model the C/S system?

\[ N_{\text{client}} + N_{\text{server}} = n \]
\[ N_{\text{client}} = X_0 \times Z \]
\[ N_{\text{server}} = X_0 \times R_0 \]
\[ R_0 = \frac{n}{X_0} - Z \]
How to understand response time bounds

\[ n \gg 1 \text{ load very high} \]
\[ \text{Server always busy } X_0 = 1/D \]

\[ n = 1 \text{ no waiting} \]
\[ R_0 = D \]

\[ n_{\text{saturation}} = 1 + Z/D \]
Assignment

• For a message processing server, composed of a processor and 2 disks, each transaction was found to make 7 Ios to disk A and 8 Ios to disk B and a total of 16 visits to the processor. The service time per visit to the two disks and processor were 20, 30 and 10 milisec, respectively. For this processor, answer the following:

• 1) What is the bottleneck device?
• 2) What is the minimum average response time?
• 3) What is the maximum possible disk A utilization for the current server configuration?
• 4) What is the maximum possible throughput of this system?
• 5) What changes in the processor speed would recommend to achieve a response time of 10 sec with 25 process-clients. Would you need a faster disk A or disk B?