Quantifying Performance Models
Bounds on Performance

Dr. John Burns

Dept. of Computing
ITT Dublin

November 6, 2017

Bounds on Performance I

- Upper bounds on throughput and lower bounds on response time can be obtained by considering the service demands only (i.e., without solving any underlying model).
- This type of bounding analysis can be quite useful since it provides the analyst with the best possible performance one could hope from a system.
- The bounding behavior of a computer system is determined by its bottleneck resource.
- The bottleneck of a system is that resource with the highest utilization (or, equivalently, the resource with the largest service demand).

Motivating Example I

Consider again the database server of the previous example and the service demands for the CPU and the three disks computed in the example.

Example

- The service demands were computed to be: $D_{CPU} = 0.092$ sec/tx, $D_{disk1} = 0.079$ sec/tx, $D_{disk2} = 0.108$ sec/tx, and $D_{disk3} = 0.142$ sec/tx.
- Correspondingly, the utilization of these devices are 35%, 30%, 41%, and 54%, respectively (the table of that example). What is the maximum throughput of the database server?

Motivating Example I

- Using the Service Demand Law, it follows that
  $U_{CPU} = D_{CPU} \times X_0 = 0.092 \times X_0$,
  $U_{disk1} = D_{disk1} \times X_0 = 0.079 \times X_0$,
  $U_{disk2} = D_{disk2} \times X_0 = 0.108 \times X_0$, and
  $U_{disk3} = D_{disk3} \times X_0 = 0.142 \times X_0$.
- Since the service demands are constant (i.e., load-independent), they do not vary with the number of concurrent transactions in execution.

Motivating Example II

- The service demands do not include any queuing time, only the total service required by a transaction at the device. Therefore, as the load (i.e., as the throughput, X_0) increases on the database server, each of the device utilizations also increases linearly as a function of their individual D_i's. See Fig. 1.
- As indicated in the figure, the utilization of disk 3 will reach 100% before any other resource, because the utilization of this disk is always greater than that of other resources.
- That is, disk3 is the system's bottleneck.
- When the system load increases to a point where disk 3's utilization reaches 100%, the throughput cannot be increased any further.

Motivating Example III

- Since $X_0 = U_{disk3}/D_{disk3}$, $X0 \leq 1/D_{disk3}$.
- Therefore, the maximum throughput, $X_0^{max} = 1/D_{disk3} = 1/0.142 = 7.04$ tps.

Utilisation vs. throughput I

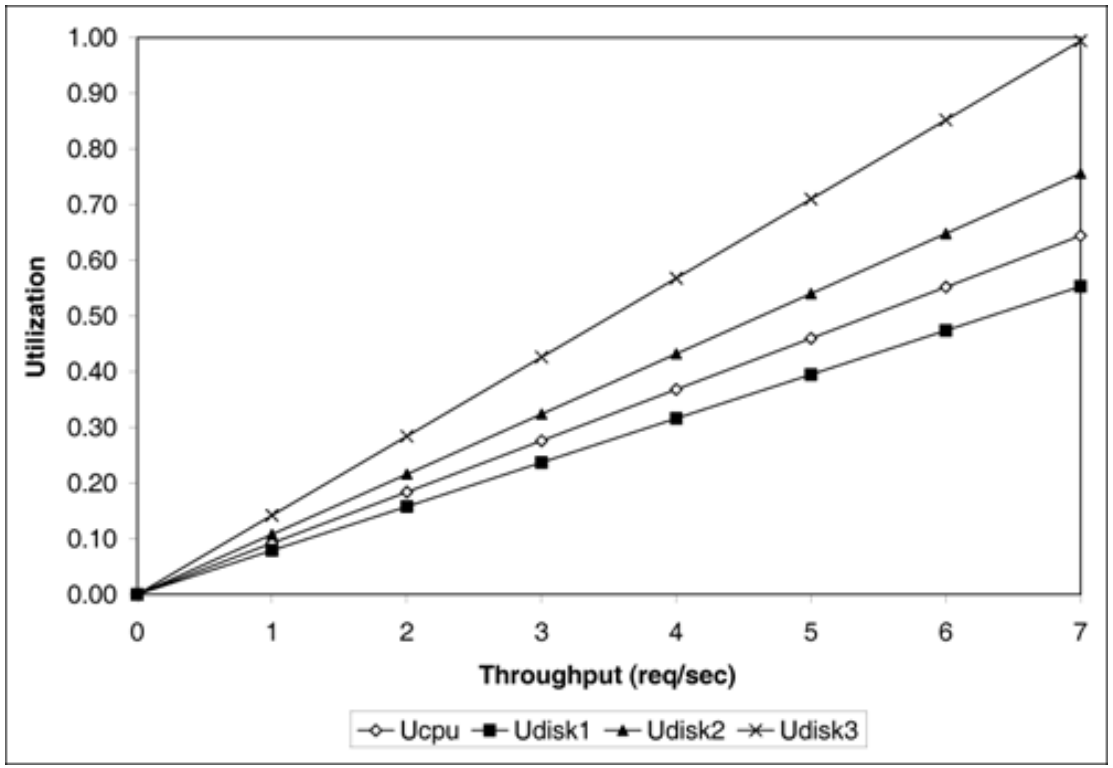

Figure 1: Utilization vs. throughput

Bounds on Performance I

This example demonstrates that:

$$X_0 = \frac{U_i}{D_i} \leq \frac{1}{D_i} \tag{1}$$

for all resources i. The resource with the largest service demand will have the highest utilization and is, therefore, the system's bottleneck. This bottleneck device yields the lowest (upper bound) value for the ratio $1/D_i$. Therefore

$$X_0 \leq \frac{1}{max(D_i)} \tag{2}$$

- This relationship is known as the upper asymptotic bound on throughput under heavy load conditions.

Bounds on Performance II

- Now consider Little's Law applied to the same database server and let N be the number of concurrent transactions in execution.
- Via Little's Law, $N = RX_0$.
- But, for a system with K resources, the response time R is at least equal to the sum of service demands $\sum_{i=1}^{K} D_i$ when there is no queuing. Thus,

$$N = RX_0 \geq (\sum_{i=1}^{K} D_i)X_0 \tag{3}$$

which can be rewritten as:

$$X_0 \leq \frac{N}{\sum_{i=1}^{K} D_i} \tag{4}$$

Bounds on Performance III

This relationship is known as the upper asymptotic bound on throughput under light load conditions.

Upper Asymptote

Combining Eqs. (2) and (4), the upper asymptotic bounds are:

$$X_0 \leq min[\frac{1}{max D_i}, \frac{N}{\sum_{i=1}^{K} D_i}] \quad (5)$$

Bounds on Performance I

- To illustrate these bounds, consider the same database server in previous examples . Consider the two lines (i.e., from Eq. (5)) that bound its throughput as shown in Fig. 2.
- The line that corresponds to the light load bound is the line $N/0.421$ (solid line with solid diamonds).
- The horizontal line at 7.04 tps (solid line with unfilled diamonds) is the heavy load bound for this case.
- The actual throughput curve is shown in Fig. 2 as the dotted line with solid diamonds and lies below the two bounding lines.
- Consider now that the bottleneck resource, disk 3, is upgraded in such a way that its service demand is halved (i.e., by replacing it with a new disk that is twice as fast).

Bounds on Performance II

- Then, the sum of the service demands becomes $0.35 (= 0.092 + 0.079 + 0.108 + 0.071)$ sec / tx.
- The maximum service demand is now that of disk 2, the new bottleneck, and the new heavy load bound (i.e., the inverse of the maximum service demand) is now $9.26 (= 1/0.108)$ tps.
- The solid lines with triangles show the bounds on throughput for the upgraded system.
- The actual throughput line (dashed line with triangles) is also shown.

Bounds on throughput I

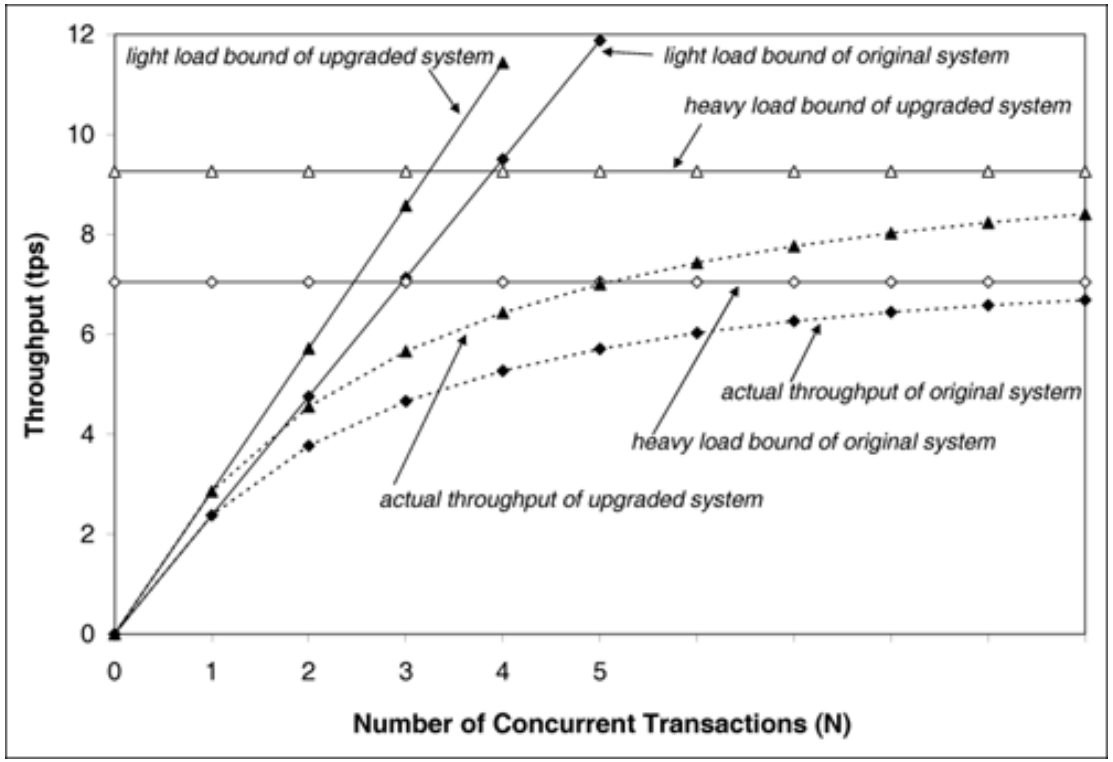

Figure 2: Bounds on throughput

Bottleneck I

- Note that when the bottleneck resource was upgraded by a factor of two, the maximum throughput improved only by 32% (from 7.04 tps to 9.26 tps).
- This occurred because the upgrade to disk 3 was excessive. Disk 2 became the new bottleneck.
- It would have been sufficient to upgrade disk 3 by a factor of 1.32 ($= 0.142/0.108$) instead of by a factor of 2 to make its service demand equal to that of disk 2.
- By using simple bottleneck analysis and performance bounds in this manner, performance can be improved for the least amount of cost.

Bounds on Performance I

- Consider the same database server.
- Let the service demand at the CPU be fixed at 0.092 sec / tx.
- What should be the values of the service demands of the three disks to obtain the maximum possible throughput, while maintaining constant the sum of the service demands at the three disks?
- Note that this is a load balancing problem (i.e., the goal is to maximize the throughput by simply shifting the load among the three disks).
- As demonstrated, the maximum service demand determines the maximum throughput.

Bounds on Performance II

- In this example, since the CPU is not the bottleneck, the maximum throughput is obtained when the service demands on all three disks is the same and equal to the average of the three original values.
- This is the balanced disk solution.
- In other words, the optimal solution occurs when $D_{disk1} = D_{disk2} = D_{disk3} = (0.079+0.108+0.142)/3 = 0.1097$ sec.
- In this case, the maximum throughput is 9.12 $(= 1/0.1097)$ tps.
- Therefore, the maximum throughput can be expanded to increase 29.5% (i.e., from 7.04 tps to 9.12 tps) simply by balancing the load on the three existing disks.

Bounds on Performance III

- To be convinced that the balanced disk solution is the optimal solution, assume that all disks have a service demand equal to D seconds.
- Now, increase the service demand of one of them by ϵ seconds, for $\epsilon > 0$.
- Since the sum of the service demands is to be kept constant, the service demand of at least one other disk has to be reduced in such a way that the sum remains the same.
- The disk that had its service demand increased will now have the largest service demand and becomes the bottleneck.
- The new maximum throughput would be $1/(D + \epsilon) < 1/D$.
- Thus, by increasing the service demand on one of the disks the maximum throughput decreases.

Bounds on Performance IV

- Similarly, suppose that the service demand of one of the disks is decreased.
- Then, the service demand of at least one of the other disks will have to increase so that the sum remains constant.
- The service demand of the disk that has the largest increase limits the throughput.
- Let $D + \delta$, for $\delta > 0$, be the service demand for the disk with the new largest demand.
- Then, the maximum throughput is now equal to $1/(D + \delta) < 1/D$.
- Either way, the maximum throughput decreases as one departs from the balanced case.

Bounds on Performance V

- Said differently, the natural (and obvious) rule of thumb is to keep all devices equally utilized.
- Now consider a lower bound on the response time. According to Little's Law, the response time R is related to the throughput as $R = N/X_0$.
- By replacing X_0 by its upper bound given in Eq. (5), the following lower bounds for the response time can be obtained.

$$R = \frac{N}{X_0} \geq \frac{N}{min[\frac{1}{max\{D_i\}}, \frac{N}{\sum_{i=1}^{K} D_i}]} = max[N \times max\{D_i\}, \sum_{i=1}^{K} D_i] \tag{6}$$

Response Time I

Example

Consider the same database server as before. What is the lower bound on response time? The sum of the service demands is $0.421 (= 0.092 + 0.079 + 0.108 + 0.142)$ and the maximum service demand is 0.142 sec. Therefore, the response time bounds are given by

$$R \geq max[0.142 \times N, 0.421] \tag{7}$$

These bounds are illustrated in Fig. 21, which also shows the actual response time curve. As seen, as the load on the system increases, the actual response time approaches the heavy load response time bound quickly.

Bounds on response time R I

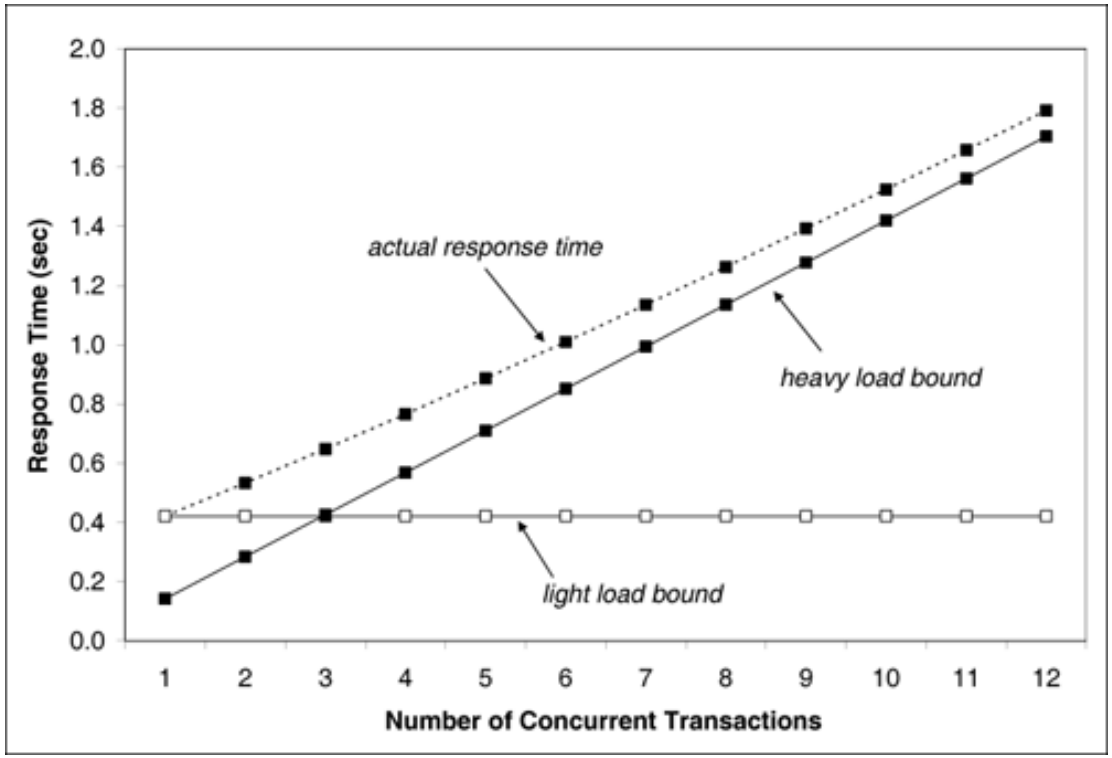

Figure 3: Bounds on R