Chapter 5

Bounds on Performance

5.1. Introduction

We begin this part of the book with a chapter devoted to the simplest useful approach to computer system analysis using queueing network models: *bounding analysis.* With very little computation it is possible to determine upper and lower bounds on system throughput and response time as functions of the system workload intensity (number or arrival rate of customers). We describe techniques to compute two classes of performance bounds: *asymptotic bounds* and *balanced system bounds.* Asymptotic bounds hold for a wider class of systems than do balanced system bounds. They also are simpler to compute. The offsetting advantage of balanced system bounds is that they are tighter, and thus provide more precise information than asymptotic bounds.

There are several characteristics of bounding techniques that make them interesting and useful:

- The development of these techniques provides valuable insight into the primary factors affecting the performance of computer systems. In particular, the critical influence of the system bottleneck is highlighted and quantified.
- The bounds can be computed quickly, even by hand. Bounding analysis therefore is suitable as a first cut modelling technique that can be used to eliminate inadequate alternatives at an early stage of a study.
- In many cases, a number of alternatives can be treated together, with a single bounding analysis providing useful information about them all.

In contrast to the bounding techniques discussed here, the more sophisticated analysis techniques presented in subsequent chapters require considerably more computation — to the point that it is infeasible to perform the analysis by hand.

Bounding techniques are most useful in *system sizing* studies. Such studies involve rather long-range planning, and consequently often are based on preliminary estimates of system characteristics. With such imprecision in knowledge of the system, quick bounding studies may be more appropriate than more detailed analyses leading to specific estimates of performance measures. System sizing studies typically involve consideration of a large number of candidate configurations. Often a single resource (such as the CPU) is the dominant concern, because the remainder of the system can be configured to match the power of this resource. Bounding analysis permits considering *as one alternative* a group of candidate configurations that have the same critical resource but differ with respect to the pattern of demands at the other service centers.

Bounding techniques also can be used to estimate the potential performance gain of alternative upgrades to existing systems. In Section 5.3 we indicate how graphs of the bounds can provide insight about the extent of service demand reduction required at the bottleneck center if it is to be possible to meet stated performance goals. (Service demand at a center can be reduced either by shifting some work away from the center or by substituting a faster device at the center.)

Our discussion of bounding analysis is restricted to the single class case. Multiple class generalizations exist, but they are not used widely. One reason for this is that bounding techniques are most useful for capacity studies of the bottleneck center, for which single class models suffice. Additionally, a major attraction of bounding techniques in practice is their simplicity, which would be lost if multiple classes were included in the models.

The models we consider in the remainder of this chapter can be described by the following parameters:

- K, the number of service centers;
- D_{max}, the largest service demand at any single center;
- D, the sum of the service demands at the centers;
- the *type* of the customer class (*batch*, *terminal*, or *transaction*);
- Z, the average think time (if the class is of terminal type).

For models with transaction type workloads, the throughput bounds indicate the maximum customer arrival rate that can be processed by the system, while the response time bounds reflect the largest and smallest possible response times that these customers could experience as a function of the system arrival rate. For models with batch or terminal type workloads, the bounds indicate the maximum and minimum possible system throughputs and response times as functions of the number of customers in the system. We refer to throughput upper and response time lower bounds as *optimistic* bounds (since they indicate the best possible performance), and we refer to throughput lower and response time upper bounds as *pessimistic* bounds (since they indicate the worst possible performance). While we treat only bounds on system throughput and response time in the following sections, the fundamental laws of Chapter 3 can be used to transform these into bounds on other performance measures, such as service center throughputs and utilizations.

5.2. Asymptotic Bounds

Asymptotic bounding analysis provides optimistic and pessimistic bounds on system throughput and response time in single class queueing networks. As their name suggests, they are derived by considering the (asymptotically) extreme conditions of light and heavy loads. The validity of the bounds depends on only a single assumption: that the service demand of a customer at a center does not depend on how many other customers currently are in the system, or at which service centers they are located.

The type of information provided by asymptotic bounds depends on whether the system workload is open (transaction type) or closed (batch or terminal type). We begin with the simpler case, that of transaction type workloads.

5.2.1. Transaction Workloads

For transaction workloads, the throughput bound indicates the maximum possible arrival rate of customers that the system can process successfully. If the arrival rate exceeds this bound, a backlog of unprocessed customers grows continually as jobs arrive. Thus, in the long run, an arriving job has to wait an indefinitely long time (since there may be any number of jobs already in queue when it arrives). In this case we say that the system is *saturated.* The throughput bound thus is the arrival rate that separates feasible processing from saturation.

The key to determining the throughput bound is the utilization law: $U_k = X_k S_k$ for each center k. If we denote the arrival rate to the system as λ, then $X_k = \lambda V_k$, and the utilization law can be rewritten as $U_k = \lambda D_k$, where D_k is the service demand at center k. To derive the throughput bound, we simply note that as long as all centers have unused capacity (i.e., have utilizations less than one), an increased arrival rate can be accommodated. However, when any of the centers becomes saturated (i.e., has utilization one), the entire system becomes saturated, since no increase in the arrival rate of customers can be handled successfully. Thus, the throughput bound is the smallest arrival rate λ_{sat} at which any center saturates. Clearly, the center that saturates at the lowest arrival rate is the *bottleneck* center — the center with the largest service demand. Let *max* be the index of the bottleneck center. Then:

$$U_{max}(\lambda) \ = \ \lambda \ D_{max} \ \leqslant \ 1$$

so:

$$\lambda_{sat} \ = \ \frac{1}{D_{max}}$$

Thus, for arrival rates greater than or equal to $1/D_{max}$ the system is saturated, while the system is capable of processing arrival rates less than $1/D_{max}$.

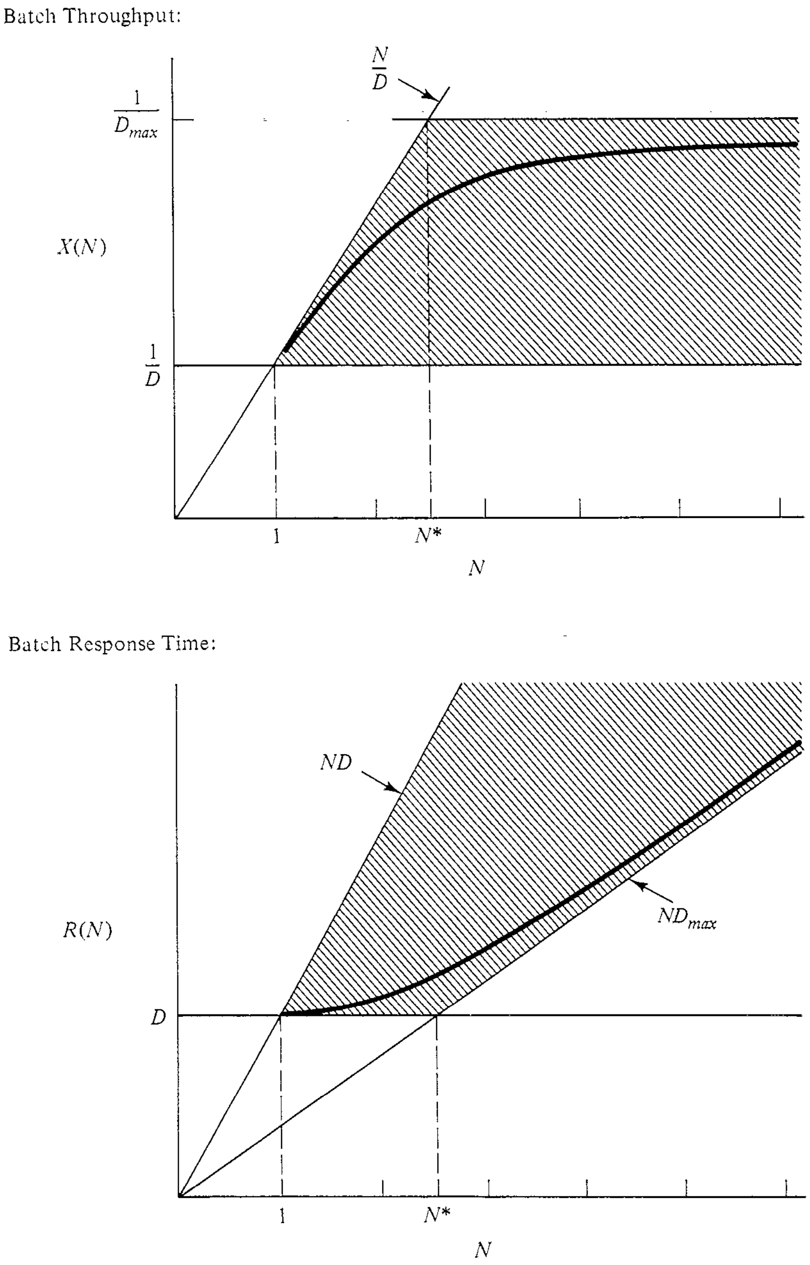
Asymptotic response time bounds indicate the largest and smallest possible response times experienced by customers when the system arrival rate is λ. Because the system is unstable if $\lambda > \lambda_{sat}$ we limit our investigation to the case where the arrival rate is less than the throughput bound. There are two extreme situations:

- In the best possible case, no customer ever interferes with any other, so that no queueing delays are experienced. In that case the system response time of each customer is simply the sum of its service demands, which we denote by D.
- In the worst possible case, n customers arrive together every n/λ time units (the system arrival rate is $\frac{n}{n/\lambda} = \lambda$). Customers at the end of the batch are forced to queue for customers at the front of the batch, and thus experience large response times. As the batch size n increases, more and more customers are waiting an increasingly long time. Thus, for any postulated pessimistic bound on response times for system arrival rate λ, it is possible to pick a batch size n sufficiently large that the bound is exceeded. We conclude that there is no pessimistic bound on response times, regardless of how small the arrival rate λ might be.

These results are somewhat unsatisfying. Fortunately, the throughput and response time bounds provide more information in the case of closed (batch and terminal) workload types.

5.2.2. Batch and Terminal Workloads

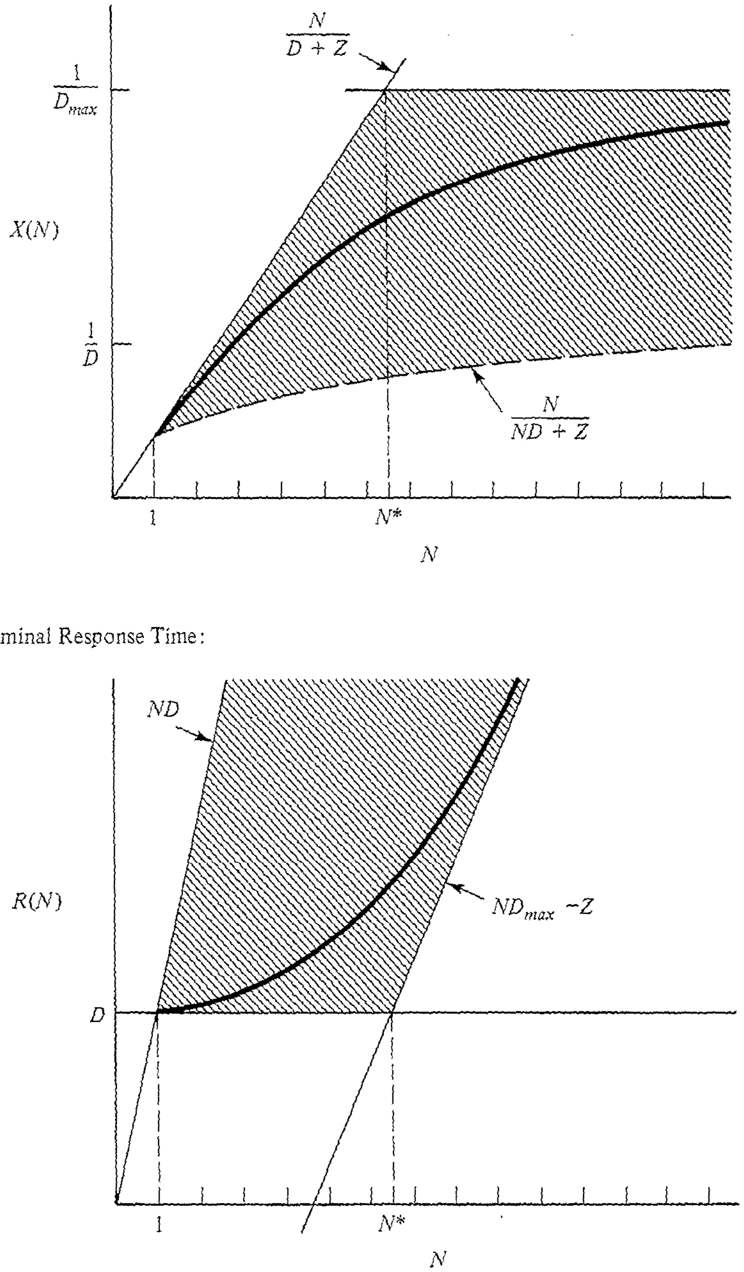
Figures 5.1a and 5.1b show the general form of the asymptotic bounds on throughput and response time for batch and terminal workloads, respectively. The bounds indicate that the precise values of the actual throughputs and response times must lie in the shaded portions of the figures. The general shapes and positions of these values are indicated by the curves in the figures.



Batch Throughput:

Batch Response Time:

Figure 5.1a — Asymptotic Bounds on Performance



Terminal Throughput:

Terminal Response Time:

Figure 5.1b — Asymptotic Bounds on Performance

To derive the bounds shown in the figures, we first consider the bounds on throughput, and then use Little's law to transform them into corresponding bounds on response time. Our analysis is stated in terms of terminal workloads. By taking the think time, Z, to be zero, we obtain results for batch workloads.

We begin with the heavy load (many customer) situation. As the number of customers in the system (N) becomes large, the utilizations of all centers grow, but clearly no utilization can exceed one. From the utilization law we have for each center k that:

$$U_k(N) \; = \; X(N)\, D_k \; \leqslant \; 1$$

Each center limits the maximum possible throughput that the system can achieve. Since the bottleneck center (*max*) is the first to saturate, it restricts system throughput most severely. We conclude that:

$$X(N) \; \leqslant \; \frac{1}{D_{max}}$$

Intuitively this is clear, because if each customer requires on average D_{max} time units of service at the bottleneck center, then in the long run customers certainly cannot be completed any faster than one every D_{max} time units.

Next consider the light load (few customers) situation. At the extreme, a single customer alone in the system attains a throughput of $1\,/\,(D+Z)$, since each interaction consists of a period of service (of average length $D \; = \; \sum_{k=1}^{K} D_k$) and a think time (of average length Z). As more customers are added to the system there are two bounding situations:

- The smallest possible throughput occurs when each additional customer is forced to queue behind all other customers already in the system. In this case, with N customers in the system, $(N-1)D$ time units are spent queued behind other customers, D time units are spent in service, and Z time units are spent thinking, so that the throughput *of each customer* is $1/(ND + Z)$. Thus, system throughput is $N/(ND + Z)$.
- The largest possible throughput occurs when each additional customer is not delayed at all by any other customers in the system. In this case no time is spent queueing, D time units are spent in service, and Z time units are spent thinking. Thus, the throughput *of each customer* is $1/(D+Z)$, and system throughput is $N/(D+Z)$.

The above observations can be summarized as the asymptotic bounds on system throughput:

$$\frac{N}{ND + Z} \; \leqslant \; X(N) \; \leqslant \; \min\,(\frac{1}{D_{max}}\,,\,\frac{N}{D+Z})$$

Note that the optimistic bound consists of two components, the first of which applies under heavy load and the second of which applies under light load. As illustrated by Figure 5.1, there is a particular population size N^* such that for all N less than N^* the light load optimistic bound applies, while for all N larger than N^* the heavy load bound applies. This crossover point occurs where the values of the two bounds are equal:

$$N^* = \frac{D+Z}{D_{max}}$$

We can obtain bounds on response time $R(N)$ by transforming our throughput bounds using Little's law. We begin by rewriting the previous equation:

$$\frac{N}{ND + Z} \leqslant \frac{N}{R(N) + Z} \leqslant \min\left(\frac{1}{D_{max}}, \frac{N}{D+Z}\right)$$

Inverting each component to express the bounds on $R(N)$ yields:

$$\max\left(D_{max}, \frac{D+Z}{N}\right) \leqslant \frac{R(N) + Z}{N} \leqslant \frac{ND + Z}{N}$$

or:

$$\max(D, ND_{max} - Z) \leqslant R(N) \leqslant ND$$

5.2.3. Summary of Asymptotic Bounds

Table 5.1 summarizes the asymptotic bounds. Algorithm 5.1 indicates the steps by which the asymptotic bounds can be calculated for batch and terminal workloads. (The calculations for transaction workloads are trivial.) Note that all bounds are straight lines with the exception of the pessimistic throughput bound for terminal workloads. Consequently, once D and D_{max} are known, calculation of the asymptotic bounds expressed as functions of the number of customers in the network takes only a few arithmetic operations. The amount of computation is independent of both the number of centers in the model and the range of customer populations of interest.

5.3. Using Asymptotic Bounds

In this section we present three applications of asymptotic bounds: a case study in which asymptotic bounds proved useful, an assessment of the effect of alleviating a bottleneck, and an example of modification analysis.

	workload type	**bounds**
X	batch	$\frac{1}{D} \leqslant X(N) \leqslant \min(\frac{N}{D}, \frac{1}{D_{max}})$
	terminal	$\frac{N}{ND + Z} \leqslant X(N)$ $\leqslant \min(\frac{N}{D+Z}, \frac{1}{D_{max}})$
	transaction	$X(\lambda) \leqslant 1 / D_{max}$
R	batch	$\max(D, ND_{max}) \leqslant R(N) \leqslant ND$
	terminal	$\max(D, ND_{max} - Z) \leqslant R(N) \leqslant ND$
	transaction	$D \leqslant R(\lambda)$

Table 5.1 – Summary of Asymptotic Bounds

5.3.1. Case Study

Asymptotic bound analysis was enlightening in the case study introduced in Section 2.6. (That section may be reviewed for additional background.)

An insurance company had twenty geographically distributed sites based on IBM 3790s that were providing unacceptable response times. The company decided to enter a three year selection, acquisition, and conversion cycle, but an interim upgrade was required. IBM 8130s and 8140s both were capable of executing the existing applications software, and consequently were considered for use during the three year transition period. After discussions with the vendor, the company believed that the use of 8130s would result in performance improving by a factor of 1.5 to 2 over the 3790s, while the use of 8140s would lead to performance improving by a factor of 2 to 3.5. (No precise statement of the significance of the "performance improvement factor" was formulated.)

A modelling study was initiated to determine those sites at which the less expensive 8130 system would suffice. It was known that the 8130 and 8140 systems both included a disk that was substantially faster than that of the 3790. With respect to CPU speed, the 8130 processor was slightly slower than the 3790, while the 8140 was approximately 1.5 times

(Steps are presented assuming a terminal workload; to treat a batch workload, set Z to zero.)

1. Calculate $D = \sum_{k=1}^{K} D_k$ and $D_{max} = \max_k D_k$.
2. Calculate the intersection point of the components of the optimistic bounds:

$$N^* = \frac{D+Z}{D_{max}}$$

3. Bounds on throughput pass through the points:

   optimistic bound :

   $(0\ ,\ 0)$ and $(1\ ,\ \frac{1}{D+Z})$ for $N \leqslant N^*$

   $(0\ ,\ \frac{1}{D_{max}})$ and $(1\ ,\ \frac{1}{D_{max}})$ for $N \geqslant N^*$

   pessimistic bound :

   This bound is not linear in N, and so must be calculated for each population of interest using the equation in Table 5.1.

4. Bounds on average response time pass through the points:

   optimistic bound :

   $(0\ ,\ D)$ and $(1\ ,\ D)$ for $N \leqslant N^*$

   $(0\ ,\ -Z)$ and $(1\ ,\ D_{max} - Z)$ for $N \geqslant N^*$

   pessimistic bound :

   $(0\ ,\ 0)$ and $(1\ ,\ D)$

Algorithm 5.1 — Closed Model Asymptotic Bounds

faster. Through a combination of this information, "live" measurements of existing 3790 systems, and benchmark experiments on two of the systems (3790 and 8140), the following service demands were determined:

	service demands, seconds	
system	**CPU**	**disk**
3790 (observed)	4.6	4.0
8130 (estimated)	5.1	1.9
8140 (estimated)	3.1	1.9

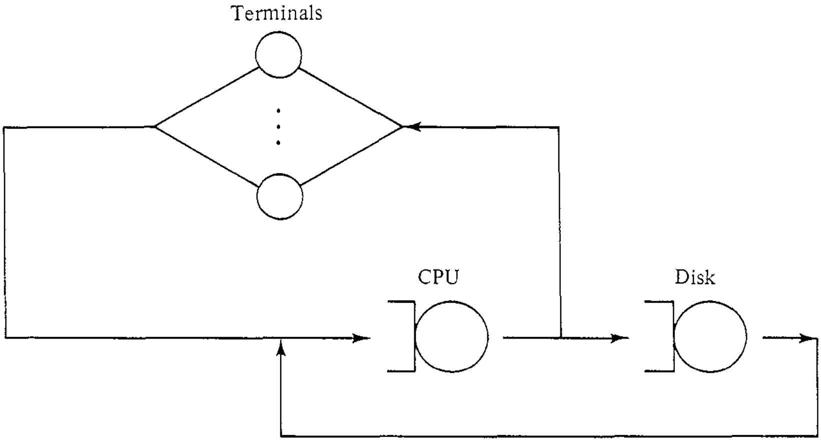


Figure 5.2 — Case Study Model

With the service demands established, a bounding model was used to assess the performance to be expected from each of the three systems. Figure 5.2 depicts the queueing network. (Although some sites had two physical disk drives, the disk controller did not permit them to be active simultaneously. For this reason, having only a single disk service center in the model is appropriate.) The parameters are:

- K, the number of service centers (2);
- D_{max}, the largest service demand (4.6 seconds for the 3790, 5.1 for the 8130, and 3.1 for the 8140);
- D, the sum of the service demands (8.6, 7.0, and 5.0, respectively);
- the type of the customer class (terminal);
- Z, the average think time (an estimate of 60 seconds was used).

Applying Algorithm 5.1 to the model of each of the three systems leads to the optimistic asymptotic bounds graphed in Figure 5.3. (The pessimistic bounds have been omitted for clarity.) These reveal that, at heavy loads, performance of the 8130 will be *inferior* to that of the 3790. This is a consequence of the fact that the 8130 has a slower CPU, which is the bottleneck device. Thus, rather than a performance gain of 1.5 to 2, a performance degradation could be expected in moving from 3790s to 8130s whenever the number of active terminals exceeded some threshold. Figure 5.3 indicates a performance gain in moving from 3790s to 8140s, although not the expected factor of two or more.

On the basis of the study, additional benchmark tests were done to re-assess the advisability of involving 8130s in the transition plan. These studies confirmed that the performance of 8130s would be worse than that of 3790s when the number of terminals was roughly fifteen or more,

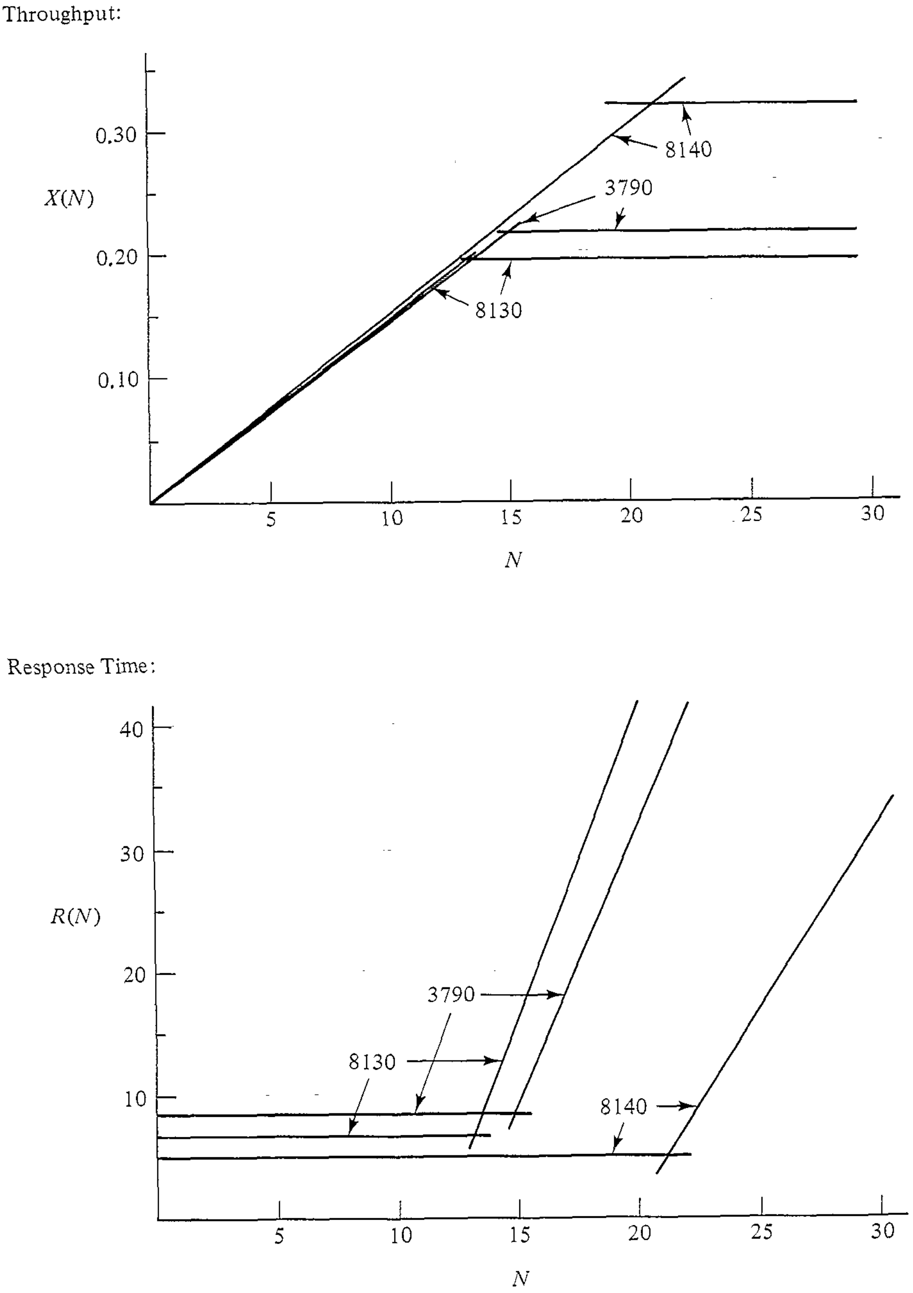


Figure 5.3 — Asymptotic Bounds in the Case Study

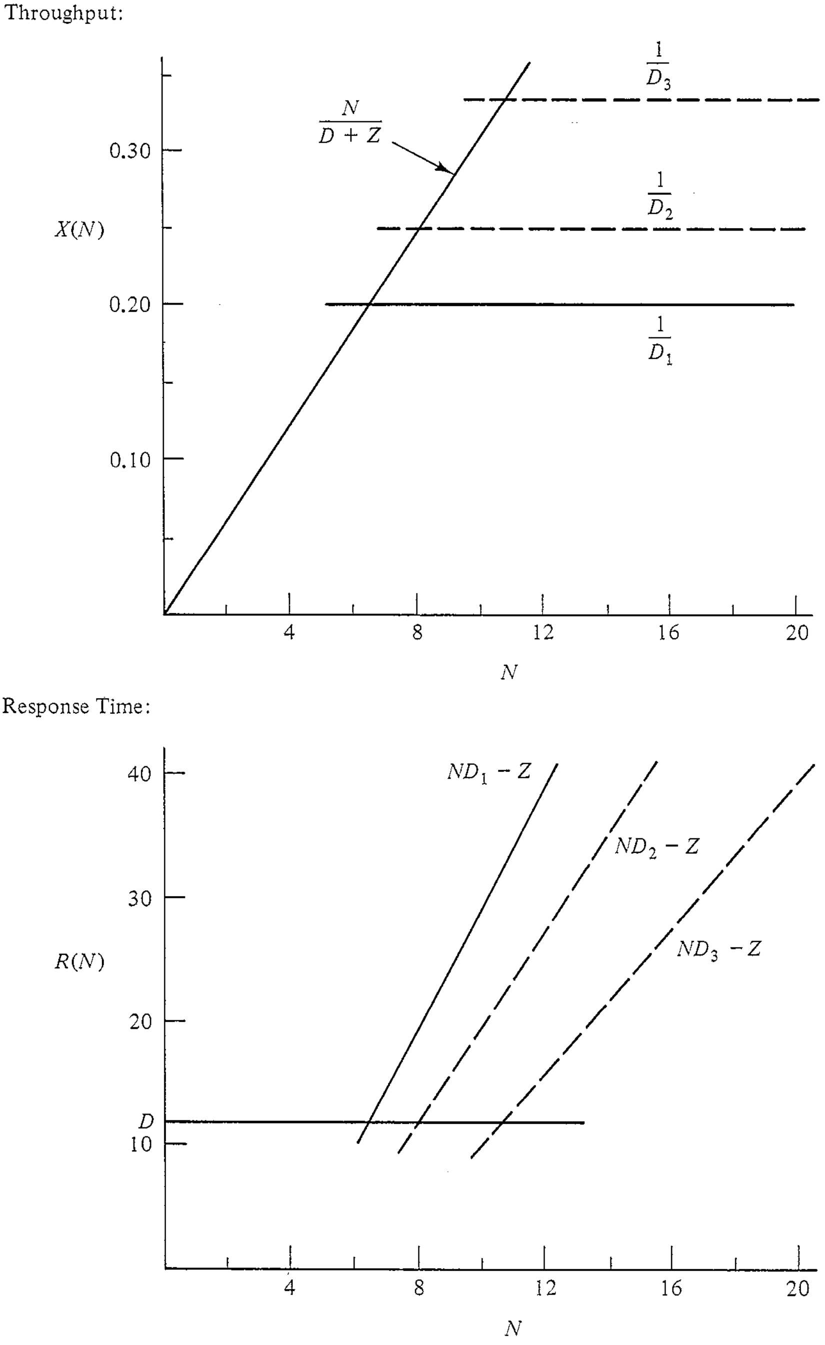
and that the performance gain of 8130s over 3790s at lighter loads would be negligible. Consequently, there was no performance reason to invest in 8130s for any sites. Eventually the company decided to install 8140s at all sites during the transition period. Without the simple modelling study, the company might have ordered 8130s without doing benchmark tests on them, with disappointing results. (A note of caution: the conclusions reached in this study would not necessarily hold in a context involving a different workload.)

5.3.2. Effect of Bottleneck Removal

So far we have been most concerned with the bottleneck center, which constrains throughput to be at most $1/D_{max}$. What happens if we alleviate that bottleneck, either by replacing the device with a faster one or by shifting some of the work to another device? In either case, D_{max} is reduced and so the throughput optimistic bound, $1/D_{max}$, increases. A limit to the extent of this improvement is imposed by the center with the second highest service demand originally. We call this center the *secondary bottleneck*, as contrasted with the *primary bottleneck.*

Consider a model with three service centers ($K=3$) and a terminal workload with average think time equal to 15 seconds ($Z=15$) and service demands of 5, 4, and 3 seconds at the centers ($D_1=5$, $D_2=4$, and $D_3=3$). Figure 5.4 shows the optimistic asymptotic bounds for this example, supplemented by lines indicating the heavy load optimistic bounds on performance corresponding to each center. Such a graph provides a visual representation of the extent of performance improvement possible by alleviating the primary bottleneck. As the load at the bottleneck center is reduced, the heavy load optimistic bound on throughput moves upwards, while the heavy load optimistic bound on average response time pivots downward (about the point $(0 , 0)$ for batch workloads and about the point $(0 , -Z)$ for terminal workloads). The light load asymptotes also change, but they are much less sensitive to the service demand at any single center than are the heavy load asymptotes.

An important lesson to be learned is the futility of improving any center but the bottleneck with respect to enhancing performance at heavy load. Reducing the service demand at centers other than the bottleneck improves only the light load asymptote, and the improvement usually is insignificant. Figure 5.5 compares the effects on the asymptotic bounds of independently doubling the speed (halving the service demand) at the primary and secondary bottlenecks for this example system. Observe that, at heavy load, performance gains only are evident when the demand at the primary bottleneck is reduced.



Throughput:

Response Time:

Figure 5.4 — Secondary and Tertiary Asymptotic Bounds

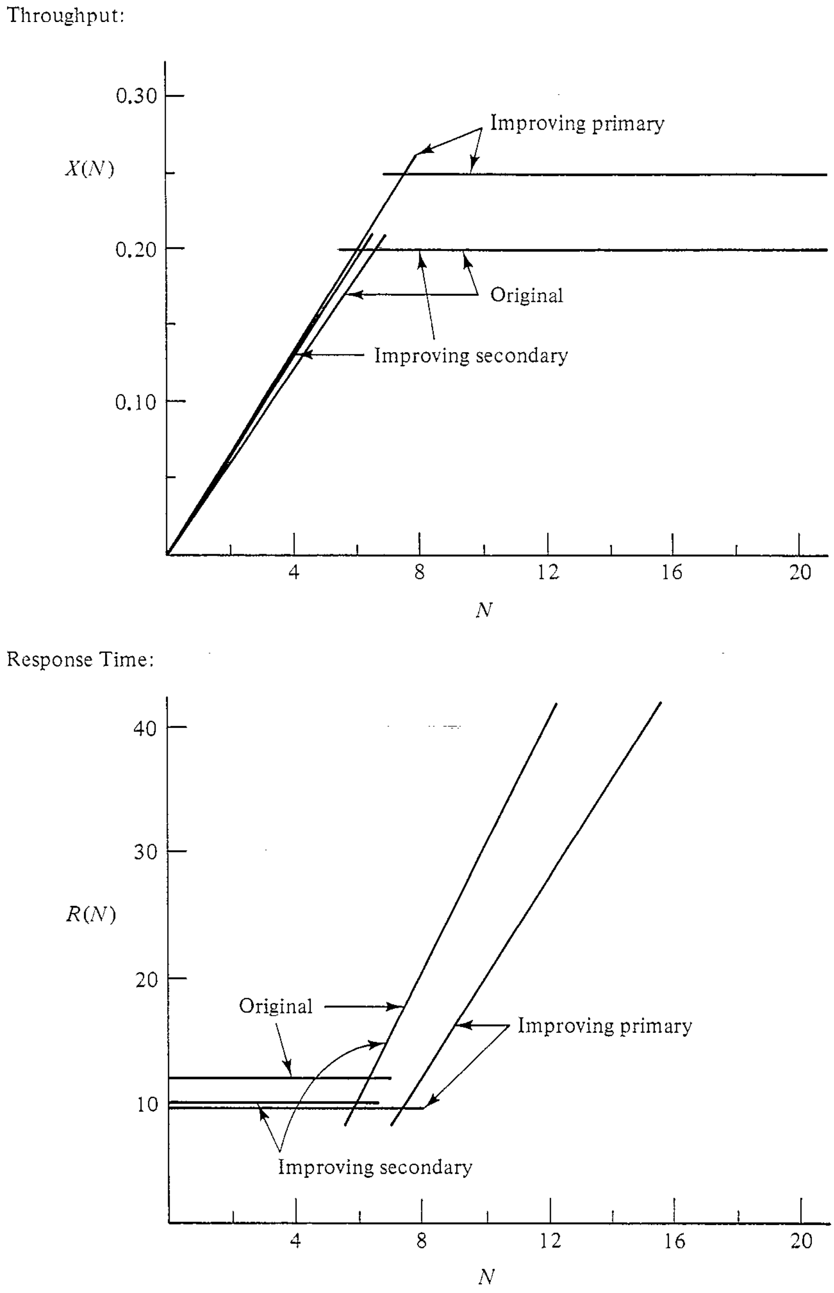


Figure 5.5 — Relative Effects of Reducing Various Service Demands

5.3.3. Modification Analysis Example

Here we examine the use of asymptotic bounds to assess the impact of modifications to an existing system. Consider a simplified interactive system for which the following measurements have been obtained:

T	= 900 seconds	length of the measurement interval
B_1	= 400 seconds	CPU busy
B_2	= 100 seconds	slow disk busy
B_3	= 600 seconds	fast disk busy
C	= 200 jobs	completed jobs
C_2	= 2,000	slow disk operations
C_3	= 20,000	fast disk operations
Z	= 15 seconds	think time

The service demands per job are $D_1=2.0$, $D_2=0.5$, and $D_3=3.0$. The visit counts to the disks are $V_2=10$ and $V_3=100$. The service times per visit to the disks are $S_2=.05$ and $S_3=.03$. We consider four improvements that can be made to the system. These are listed below, along with an indication of how each would be reflected in the parameters of the model:

1. Replace the CPU with one that is twice as fast. $D_1 \leftarrow 1$

2. Shift some files from the faster disk to the slower disk, balancing their demands. We consider only the primary effect, which is the change in disk speed, and ignore possible secondary effects such as the fact that the average size of blocks transferred may differ between the two disks. The new disk service demands are derived as follows. $V_2 + V_3 = 110$. Because $S_2=.05$ and $S_3=.03$, this is the same as:

$$\frac{V_2S_2}{.05} + \frac{V_3S_3}{.03} = 110$$

   Since we wish to have $D_2 = V_2S_2 = V_3S_3 = D_3$:

$$D_2 \left[\frac{1}{.05} + \frac{1}{.03} \right] = 110$$

   and $D_2 = D_3 = 2.06$. Dividing by the appropriate service times, we obtain the new visit counts: $V_2=41$ and $V_3=69$.

3. Add a second fast disk (center 4) to handle half the load of the busier existing disk. Once again, we consider only the primary effects of the change. $K \leftarrow 4$, $D_3 \leftarrow 1.5$, $D_4 \leftarrow 1.5$

4. The three changes made together: the faster CPU and a balanced load across two fast disks and one slow disk. Service demands become $D_1=1$, $D_2=1.27$, $D_3=1.27$, and $D_4=1.27$. These were derived in a manner similar to that employed above. We know that $V_2 + V_3 + V_4 = 110$. To ensure that $D_2 = D_3 = D_4$:

$$\frac{V_2 S_2}{.05} + \frac{V_3 S_3}{.03} + \frac{V_4 S_4}{.03} = 110$$

$$D_2 \left[\frac{1}{.05} + \frac{1}{.03} + \frac{1}{.03} \right] = 110$$

$$D_2 = D_3 = D_4 = \left[\frac{.0015}{.13} \right] 110 = 1.27$$

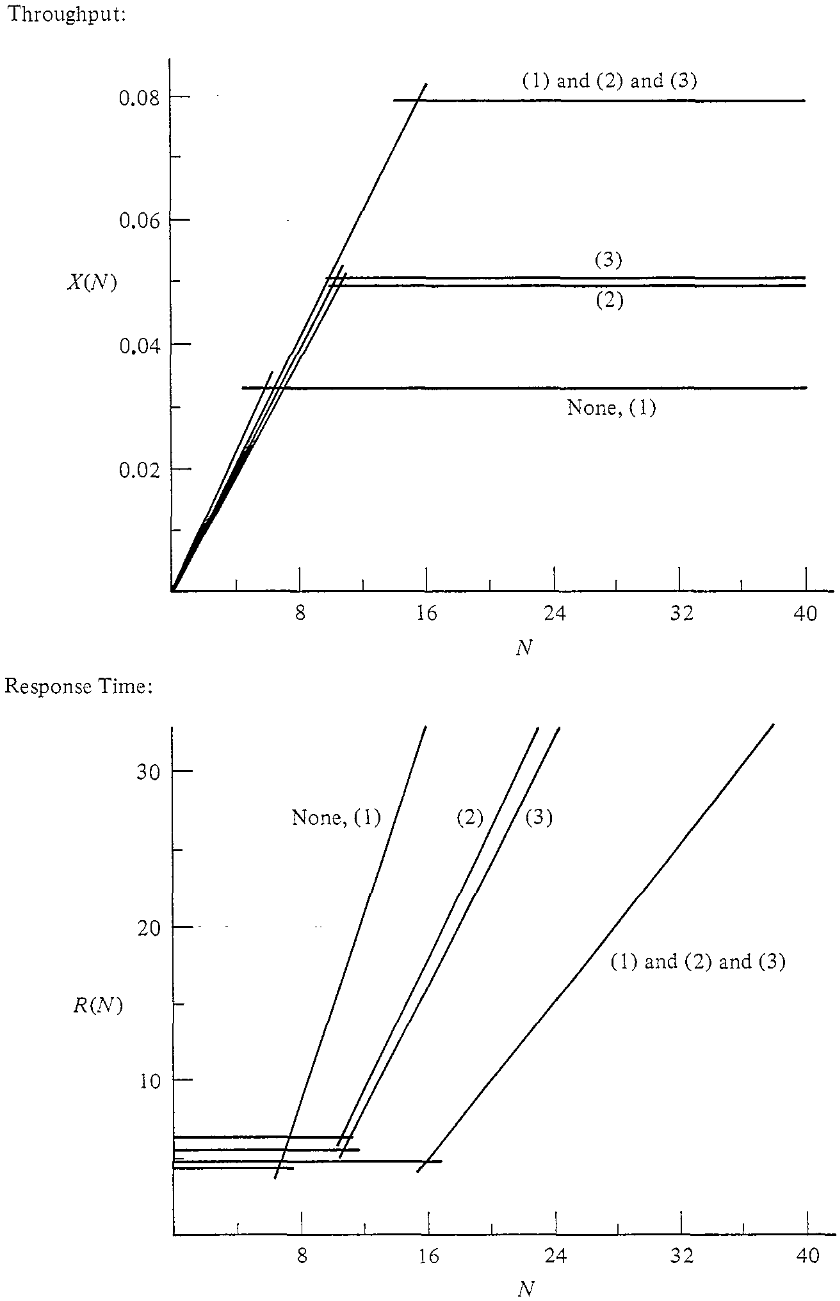
Figure 5.6 shows the optimistic asymptotic bounds for the original system (labelled "None"), for each modification individually (labelled "(1)", "(2)", and "(3)", respectively), and for the three in combination (labelled "(1) and (2) and (3)"). Intuitively, the first change might appear to be the most significant, yet Figure 5.6 shows that this is not true. Because the fast disk is the original bottleneck, changes 2 and 3 are considerably more influential. Note that change 2 yields almost as much improvement as change 3 although it requires no additional hardware. The combination of the three modifications yields truly significant results.

The modification analysis done in this section has involved only asymptotic bounds on performance. In Chapter 13 we will consider modification analysis once again, using more sophisticated techniques to evaluate our models.

5.4. Balanced System Bounds

With a modest amount of computation beyond that required for asymptotic bounds, tighter bounds can be obtained. These bounds are called *balanced system bounds* because they are based upon systems that are "balanced" in the sense that the service demand at every center is the same, i.e., $D_1=D_2=D_3= \ldots =D_K$. Figures 5.7a and 5.7b show the general form of balanced system bounds (together with the asymptotic bounds) for batch (5.7a) and terminal (5.7b) workloads.

We first establish some special properties of balanced systems. We then show how these properties can be exploited to determine bounds on performance that complement the asymptotic bounds and lead to more precise knowledge of system behavior. The derivation of balanced system bounds is shown for batch workloads only. The reader is asked to work through the derivation for transaction workloads in Exercise 5. Bounds


Throughput:

Response Time:

Figure 5.6 — Example of the Effects of Various Changes

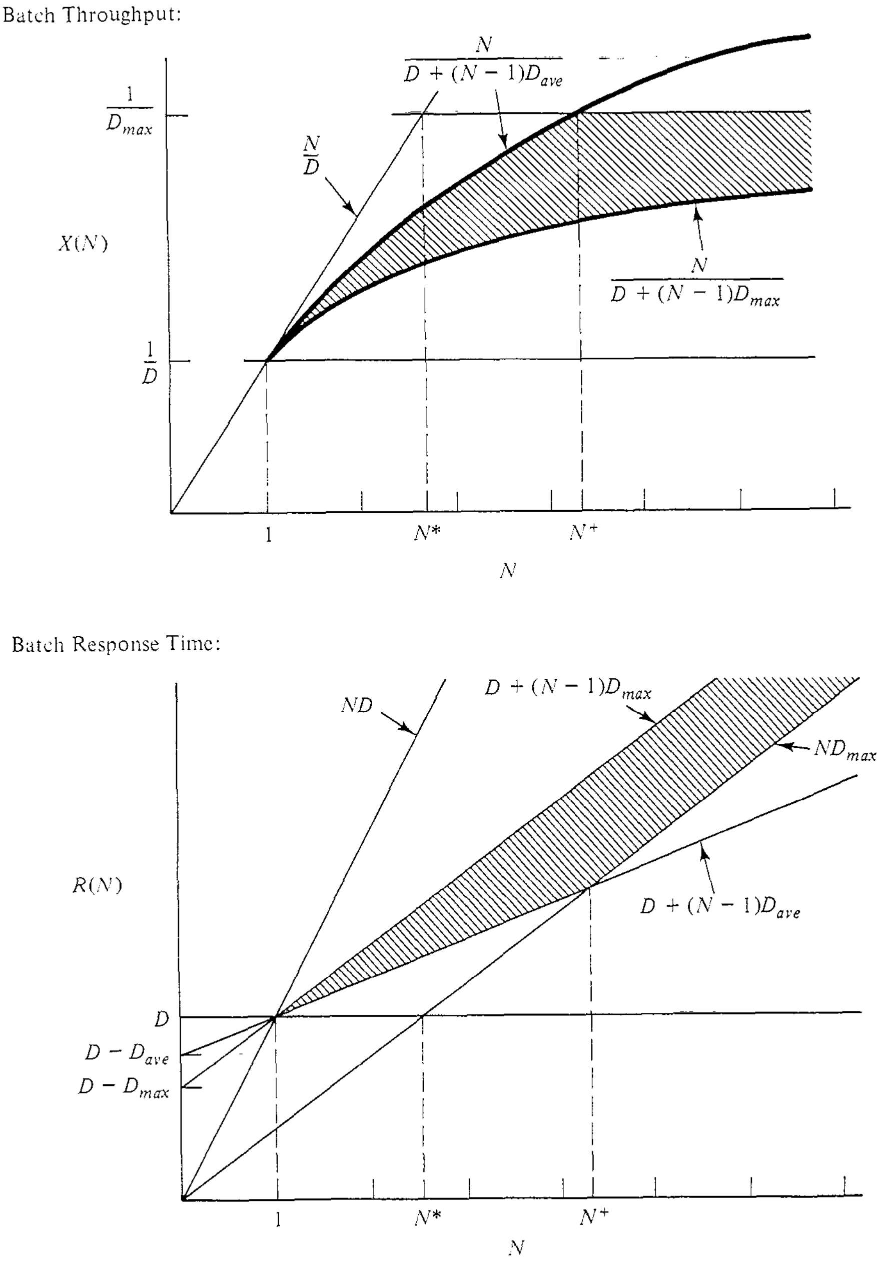
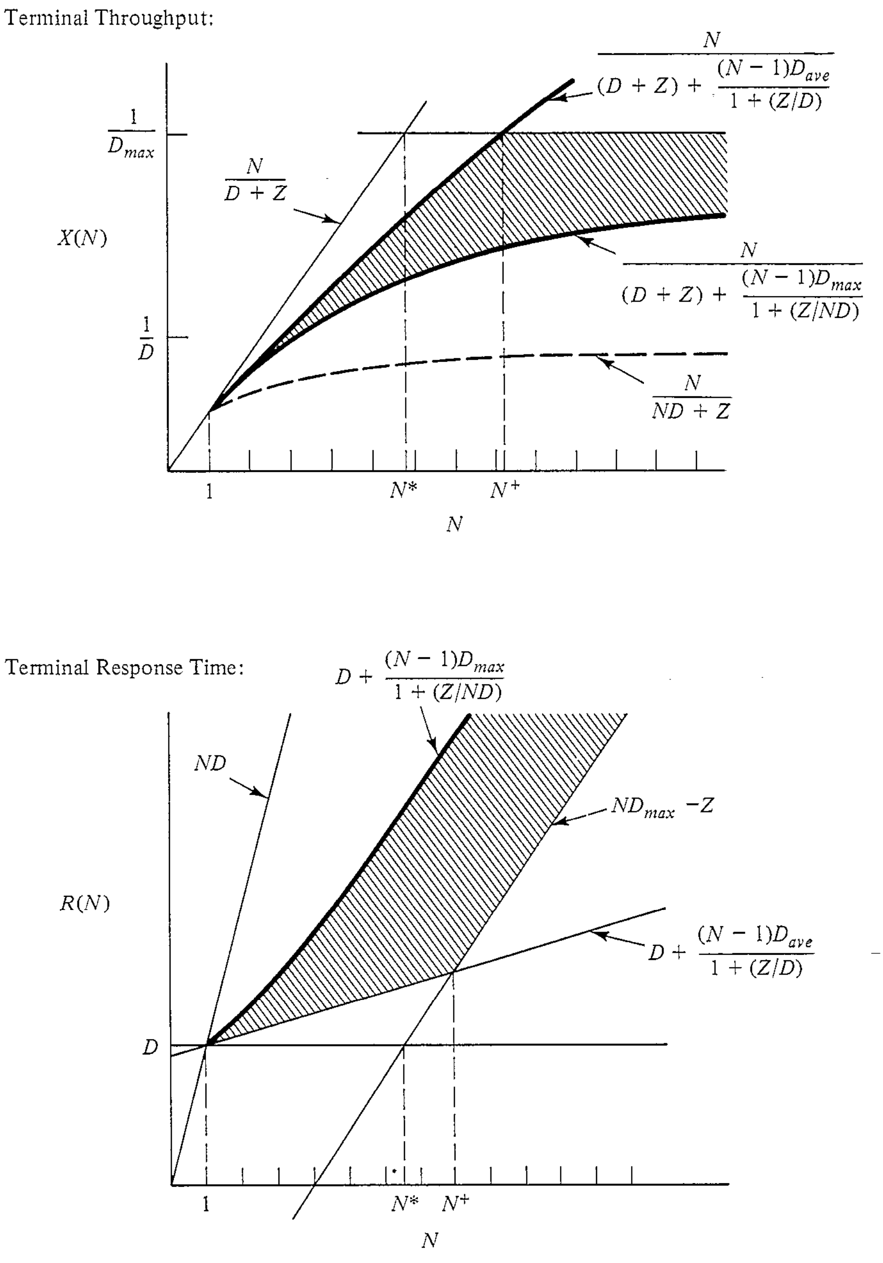


Figure 5.7a – Balanced System Bounds on Performance


Terminal Throughput:

Terminal Response Time:

Figure 5.7b – Balanced System Bounds on Performance

for each of batch, terminal, and transaction workload types are given in Table 5.2.

The analysis of balanced systems is a special case of the techniques to be presented in Chapter 6. Formally, this analysis requires that various assumptions be made about the system being modelled. (These assumptions will be described in Chapter 6.) This is in contrast to asymptotic bounds, which require only that the service demand of a customer at a center does not depend on how many other customers are currently in the system or at which centers they are located.

For balanced systems, the techniques to be presented in Chapter 6 have a particularly simple form. The utilization of every service center is given by:

$$U_k(N) = \frac{N}{N+K-1}$$

(We do not attempt to justify this now, either intuitively or formally.) By the utilization law, system throughput is then:

$$X(N) = \frac{U_k}{D_k} = \frac{N}{N+K-1} \times \frac{1}{D_k}$$

where D_k is the service demand at every center.

Let D_{max}, D_{ave}, and D_{min} denote respectively the maximum, average, and minimum of the service demands at the centers of the model we wish to evaluate. We bound the throughput of that system by the throughputs of two related balanced systems: one with service demand D_{min} at every center, and the other with service demand D_{max} at every center:

$$\frac{N}{N+K-1} \times \frac{1}{D_{max}} \leqslant X(N) \leqslant \frac{N}{N+K-1} \times \frac{1}{D_{min}}$$

These inequalities hold because, of all systems with K centers, N customers, and maximum service demand D_{max}, the one with the lowest throughput is the balanced system with demand D_{max} at each center. Similarly, of all systems with K centers, N customers, and minimum demand D_{min}, the one with the highest throughput is the balanced system with demand D_{min} at each center. Corresponding bounds on average response times are:

$$(N+K-1)\, D_{min} \leqslant R(N) \leqslant (N+K-1)\, D_{max}$$

Tighter balanced system bounds can be obtained by constraining not only the maximum service demand, D_{max}, but also the total demand, D (or equivalently, the average demand, D_{ave}). Of all systems with a given total service demand $D = \sum_{k=1}^{K} D_k$, the one with the highest throughput (and the lowest average response time) is the one in which all service demands are equal (i.e., $D_k = D/K$, $k = 1, \ldots, K$). This confirms our intuition that the increase in delay resulting from an increase in load is greater than the decrease in delay resulting from an equivalent decrease in load. Therefore, optimistic bounds are given by:

$$X(N) \leqslant \frac{N}{N+K-1} \times \frac{1}{D_{ave}} = \frac{N}{D + (N-1)\, D_{ave}}$$

and:

$$D + (N-1)\, D_{ave} \leqslant R(N)$$

Note that the optimistic balanced system bound intersects the heavy load component of the optimistic asymptotic bound (at a point that we will denote by N^{+}). Beyond this point, the balanced system bound is defined to coincide with the asymptotic bound.

Analogously, of all systems with total demand D and maximum demand D_{max}, the one with the lowest throughput has D/D_{max} centers with demand D_{max}, and zero demand at the remaining centers. (The fact that D/D_{max} may not be an integer hampers intuition, but not the validity of the bounds.) Therefore, pessimistic bounds are:

$$\frac{N}{N + \frac{D}{D_{max}} - 1} \times \frac{1}{D_{max}} = \frac{N}{D + (N-1)\, D_{max}} \leqslant X(N)$$

and:

$$R(N) \leqslant D + (N-1)\, D_{max}$$

Table 5.2 summarizes the balanced system bounds for batch, terminal, and transaction workloads. Algorithm 5.2 indicates how these bounds can be calculated for batch and terminal workloads. (The calculations for transaction workloads are trivial.) For batch workloads, the bounds on average response time are straight lines. Also, the optimistic bound on average response time for terminal workloads is a straight line. However, balanced system bounds on throughput and the pessimistic balanced system bound on response time for terminal workloads are not linear in N, and thus must be computed separately for each value of N of interest.

	workload type	bounds
X	batch	$\frac{N}{D + (N-1)D_{max}} \leqslant X(N)$ $\leqslant \min\left(\frac{1}{D_{max}}, \frac{N}{D + (N-1)D_{ave}}\right)$
	terminal	$\frac{N}{D + Z + \frac{(N-1)D_{max}}{1 + Z/(ND)}} \leqslant X(N)$ $\leqslant \min\left(\frac{1}{D_{max}}, \frac{N}{D + Z + \frac{(N-1)D_{ave}}{1 + Z/D}}\right)$
	transaction	$X(\lambda) \leqslant 1 / D_{max}$
R	batch	$\max(ND_{max}, D + (N-1)D_{ave}) \leqslant R(N)$ $\leqslant D + (N-1)D_{max}$
	terminal	$\max\left(ND_{max} - Z, D + \frac{(N-1)D_{ave}}{1 + Z/D}\right) \leqslant R(N)$ $\leqslant D + \frac{(N-1)D_{max}}{1 + Z/(ND)}$
	transaction	$\frac{D}{1 - \lambda D_{ave}} \leqslant R(\lambda) \leqslant \frac{D}{1 - \lambda D_{max}}$

Table 5.2 — Summary of Balanced System Bounds

5.5. Summary

In this chapter we have introduced techniques for obtaining bounds on the performance measures of systems. The bounds are summarized in Tables 5.1 and 5.2, and procedures for calculating them are given in Algorithms 5.1 and 5.2. Asymptotic bounds and balanced system bounds are important for a number of reasons:

1. Calculate the asymptotic bounds using Algorithm 5.1.
2. Determine the point at which the optimistic balanced system bound intersects the optimistic asymptotic bound. For a batch workload:

   $$N^+ = \frac{D - D_{ave}}{D_{max} - D_{ave}}$$

   For a terminal workload:

   $$N^+ = \frac{(D+Z)^2 - D\ D_{ave}}{(D+Z)D_{max} - D\ D_{ave}}$$

   The optimistic balanced system bound need be calculated only from 1 to N^+ since it is defined to coincide with the asymptotic bound beyond N^+.
3. Calculate balanced system bounds on average response time. For a batch workload, the bounds are lines through the points:

   optimistic bound :

   $(0\ ,\ D - D_{ave})$ and $(1\ ,\ D)$

   pessimistic bound :

   $(0\ ,\ D - D_{max})$ and $(1\ ,\ D)$

   For a terminal workload, the bounds are lines through the points:

   optimistic bound :

   $$(0\ ,\ D - \frac{D_{ave}}{1 + Z/D}) \text{ and } (1\ ,\ D)$$

   pessimistic bound :

   The pessimistic bound for terminal workloads is not linear in N, so must be calculated for each population of interest using the equation in Table 5.2.
4. Calculate balanced system bounds on throughput for the range of N of interest using the equations in Table 5.2. (Again, these are not linear in N.)

Algorithm 5.2 — Closed Model Balanced System Bounds

- Because they are so simple to calculate, even by hand (they require only a few arithmetic operations once D and D_{max} are known), they are a quick way to obtain a rough feel for the behavior of a system.
- They reveal the critical influence of the bottleneck service center. Changes to the system that do not affect the bottleneck center do not alter the heavy load bounds on performance. Hence, throughput curves for *all* systems with bottleneck demand D_{max} are constrained to lie below the line $1/D_{max}$. To improve performance beyond this limit, it is necessary to reduce the demand at the bottleneck center in some way.
- Diagrams that show secondary bottlenecks as well as the primary one provide insight into the extent of improvements realizable by various modifications to the system that reduce the demand on the primary bottleneck.
- In the early phases of system design and system sizing, bounding studies offer the advantage that a group of configurations may be able to be treated as a single alternative. This is the case because of the critical influence of the bottleneck center, noted above.

Using fundamental laws, bounds on center utilizations and throughputs can be calculated from the asymptotic and balanced system bounds on system throughput. The system throughput bounds of Tables 5.1 and 5.2 are transformed into bounds on center k utilization simply by multiplying through by D_k (since the utilization law states that $U_k(N) = X(N)\ D_k$). Similarly, bounds on center k throughput are obtained by multiplying through by V_k (due to the forced flow law: $X_k(N) = X(N)\ V_k$).

In the chapters that follow, we present methods for calculating specific values of performance measures rather than bounds. These values will form smooth curves that are asymptotic to the light and heavy load optimistic asymptotic bounds and to the pessimistic balanced system bounds.

5.6. References

Muntz and Wong [1974] carried out the first study devoted specifically to the asymptotic properties of closed queueing networks. Denning and Kahn [1975] obtained related results independently. Denning and Buzen [1978] describe asymptotic analysis in discussing bottlenecks as part of operational analysis. Beizer [1978] also includes bounding analysis as part of his performance evaluation methodology. Balanced system bound analysis was developed by Zahorjan et al. [1982]. The case study of Section 5.3.1 was carried out by Sevcik et al. [1980].

[Beizer 1978]
Boris Beizer. *Micro-Analysis of Computer System Performance.* Van Nostrand Reinhold, 1978.

[Denning & Buzen 1978]
Peter J. Denning and Jeffrey P. Buzen. The Operational Analysis of Queueing Network Models. *Computing Surveys 10*,3 (September 1978), 225-261.

[Denning & Kahn 1975]
Peter J. Denning and Kevin C. Kahn. Some Distribution Free Properties of Throughput and Response Time. Report CSD-TR-159, Computer Science Department, Purdue University, May 1975.

[Muntz & Wong 1974]
R.R. Muntz and J.W. Wong. Asymptotic Properties of Closed Queueing Network Models. *Proc. 8th Princeton Conference on Information Sciences and Systems* (1974).

[Sevcik et al. 1980]
K.C. Sevcik, G.S. Graham, and J. Zahorjan. Configuration and Capacity Planning in a Distributed Processing System. *Proc. 16th CPEUG Meeting* (1980), 165-171.

[Zahorjan et al. 1982]
J. Zahorjan, K.C. Sevcik, D.L. Eager, and B.I. Galler. Balanced Job Bound Analysis of Queueing Networks. *CACM 25*,2 (February 1982), 134-141.

5.7. Exercises

1. In a system serving both batch jobs and terminal users, the following observations were made during a 30 minute interval:

active terminals	40
think time	20 seconds
interactive response time	5 seconds
disk service time per access	20 milliseconds
disk accesses per batch job	100
disk accesses per terminal interaction	5
disk utilization	60%

   a. What is batch throughput?

   b. Using only the information given above, calculate the maximum batch throughput possible if interactive response times of 15 seconds are to be achievable. What assumption must you make in answering this question?

2. Consider an interactive system with a CPU and two disks. The following measurement data was obtained by observing the system:

observation interval	30 minutes
active terminals	30
think time	12 seconds
completed transactions	1,600
fast disk accesses	32,000
slow disk accesses	12,000
CPU busy	1,080 seconds
fast disk busy	400 seconds
slow disk busy	600 seconds

a. Determine the visit counts (V_k), service times per visit (S_k), and service demands (D_k) at each center.

b. Give optimistic and pessimistic asymptotic bounds on throughput and response time for 5, 10, 20, and 40 active terminals.

Consider the following modifications to the system:

1: Move all files to the fast disk.
2: Replace the slow disk by a second fast disk.
3: Increase the CPU speed by 50% (with the original disks).
4: Increase the CPU speed by 50% and balance the disk load across two fast disks.

c. For the original system and for modifications 1 through 4, graph optimistic and pessimistic asymptotic bounds on throughput and response time as functions of the number of active terminals.

d. For the original system and for modification 3, specify the maximum number of terminals that can be active such that the asymptotic bounds do not preclude the possibility of an 8 second average response time.

e. If 40 terminals were active on the original system, how much would the CPU have to be speeded up so that the bounds would not rule out the possibility of achieving 10 second average response times?

f. If 80 terminals were active on the original system, what minimum modifications to the system would be required so that the bounds would not rule out the possibility of achieving 15 second average response times?

3. An installation with a CPU intensive workload is considering moving from a centralized system with a single large CPU to a decentralized system with several smaller CPUs.

a. Suppose that 10 processors each 1/10-th the speed of the large processor can be operated at the same cost as the large processor. Use asymptotic throughput and response time bounds to investigate the conditions under which such a change clearly would be beneficial or detrimental (considering performance issues only).

b. Suppose that 15 processors each 1/10-th the speed of the large processor can be operated at the same cost. How does this affect your answer to (a)?

4. Consider a model with three service centers and service demands $D_1 = 5$ seconds, $D_2 = 4$ seconds, and $D_3 = 3$ seconds.

   a. Graph the optimistic and pessimistic asymptotic throughput and response time bounds for this model with a batch workload.

   b. On the same graphs, include balanced system bounds for the model.

   c. What is the relationship between the two sets of bounds in terms of the range of possible values to which they restrict performance measures? What is their relationship in terms of computational effort?

   d. Repeat your calculations for a terminal class with 15 second think times.

5. The assumptions introduced in deriving balanced system bounds for transaction workloads do not result in an improvement over the asymptotic bound for system throughput; we still have $X(\lambda) \leqslant 1/D_{max}$. However, they do yield an improved response time bound. The key to this improvement is the equation:

$$R_k(\lambda) = \frac{D_k}{1 - U_k(\lambda)}$$

   a. Using this equation, derive optimistic and pessimistic response time bounds based on balanced systems in which the service demands at all centers are set to D_{min} (optimistic) and D_{max} (pessimistic).

   b. Derive improved bounds by using the fact that the sum of the service demands in the original system is D. (Check your results against Table 5.2.)

   c. Compute the value of λ_{sat} for a system with three service centers with service demands of 8, 4, and 2 seconds. Sketch the two sets of response time bounds you just derived for arrival rates λ between 0 and λ_{sat}.