Chapter 5
Bounds on Performance

Asymtotic Bounds and Balanced System Bounds

These are very quick rule-of-thumb calculations for systems, able to be carried out by hand or with a basic calculator

They are correspondingly rough. They give a range of possible response times / throughputs If the range is informative, great. If not, you haven't lost much time.

They can be used to provide a quick graphic display of the performance of the system under a range of input values. Computationally, plug in a variety of input values. Get the corresponding ranges of response times or throughputs

To keep the virtue of computational simplicity, use a single class model.

The following are the only parameters required to calculate projections for these models: K, the number of service centers;
Dmax, the largest service demand at any single center;
D, the sum of the service demands at all the centers;
the type of workload (batch, terminal, or transaction);
Z, the average think time.

These bounds require only one assumption in order to be valid: service demand at a center may not depend on the number or arrangement of customers currently in the system. This allows any model in which service demands at each center are intrinsic to the job.

What violates this assumption?

Examples include memory queues (why?) and adaptive resource use such as directing a job to the processor with the lightest load in a multiprocessor system. Other examples?

Transaction Workloads: A couple rough basic conclusions are possible. (We'll do better with the closed workloads.)

One bound on the throughput X comes from the maximum arrival rate λ that the system can handle without necessarily generating a backlog that grows indefinitely ~ system saturation.
Can you think of a simple bound on X based on the D_k's?

Recall that in a system that is not saturated $\lambda = X$. Also, for all k

$$U_k = X D_k \qquad \text{ie}$$

$$U_k = \lambda D_k \qquad \text{but}$$

$0 \leq U_k \leq 1$, so

$$\lambda D_k \leq 1, \quad \lambda \leq \frac{1}{D_k} \quad \text{for all } k,$$

ie $\lambda \leq (D_{max})^{-1}$ where

$$D_{max} = \max_k \{ D_k \mid 1 \leq k \leq K \}$$

Can a throughput of $\lambda = (D_{max})^{-1}$ be realized? If $D_{max} < D$?

What about response time?

Best case is that every job goes through without any queuing time:

$$R \geq D = \sum_k D_k$$

Worst Case? There is no worst case average response time. It can get arbitrarily long.

For a given $\lambda < D_{max}$, consider the case of n customers arriving together every $t = n/\lambda$ time units, for an arrival rate

$$\frac{n}{t} = \frac{n}{(n/\lambda)} = \lambda.$$

Of the n customers in a cohort, the last to finish has a response time of at least nD_{max} (why?), the second to last has a response time of at least $(n-1)D_{max}$. Generally, the ith to finish has a response time of at least $(i)D_{max}$. The average response time is at least

$$\frac{\sum_{i=1}^{n} iD_{max}}{n} = \frac{n(n+1)}{2n} D_{max}$$

$$= \frac{n+1}{2} D_{max}.$$

In this way, by increasing n you can create a workload for the system with average arrival rate λ and an arbitrarily long average response time.

This is actually a useful observation: response time degrades as jobs cluster.

Conclusion: For transaction workloads

$$X \leq \frac{1}{D_{max}}$$

$$R \geq D$$

We can make stronger statements about batch and terminal workloads. Since batch corresponds to terminal with $Z=0$, we'll derive results for terminal workloads. First, derive bounds on X. From these and Little's Law, calculate bounds on R.

Note that the lower bound on throughput is a 'worst case' or pessimistic bound, while the upper bound represents a best case or optimistic bound.

Upper Bounds on Throughput $X(N)$

We still have $X(N) \leq \frac{1}{D_{max}}$

We also know that the fastest possible average response time is $D+Z$ sec/job.

One user cycling through with this average response time has a personal throughput of $\leq \frac{1}{D+Z}$ jobs/sec.

N users each with personal throughput $\leq \frac{1}{D+Z}$ give a system throughput

$$X(N) \leq \frac{N}{D+Z}$$

$$X \leq \min\left(\frac{1}{D_{max}}, \frac{N}{D+Z}\right)$$

Lower Bound on Throughput $X(N)$

How long can a job possibly spend in the system?
It can have to wait while all the other $N-1$ jobs complete, requiring D time each.

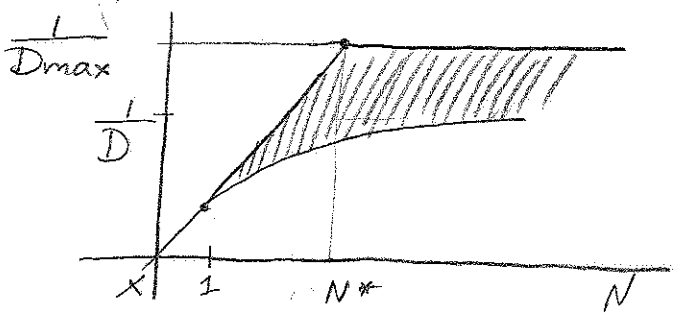
Then the time required for the job must be $\leq Z+(N-1)D+D = ND+Z$.

The throughput for an individual is thus $\geq \frac{1}{ND+Z}$

Conclude that $X(N) \geq \frac{N}{ND+Z}$

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

The upper bound is formed by the intersection of two lines, $X = \frac{N}{D+Z}$ for small N and the horizontal line (on an (N,X) coordinate plane) $X = \frac{1}{D_{max}}$. These intersect at $N^* = \frac{D+Z}{D_{max}}$



To use these bounds to obtain bounds for $R(N)$ apply

$$R(N) = \frac{N}{R(N)} - Z, \text{ so}$$

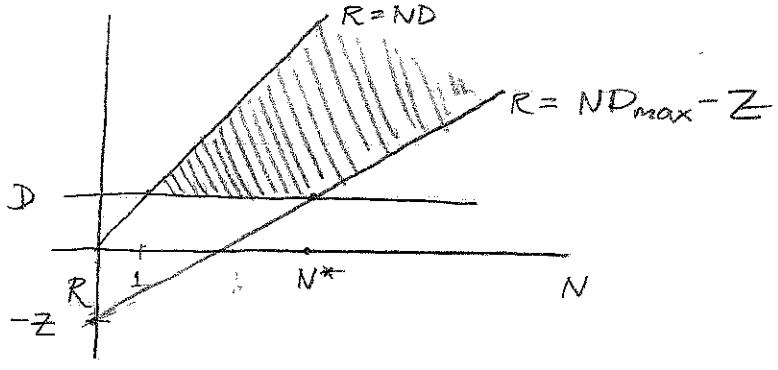
$$X = \frac{N}{R(N)+Z}, \text{ so}$$

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

Manipulate this algebraically to get a pair of inequalities with $R(N)$ in the middle

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

The lines forming the lower bound again intersect at $N^* = \frac{D+Z}{D_{max}}$



The text presents several case studies that illustrate the application of asymptotic bounds.

In the first, a company is examining options for an interim upgrade to a system that they model with

terminal type workload

$K = 2$

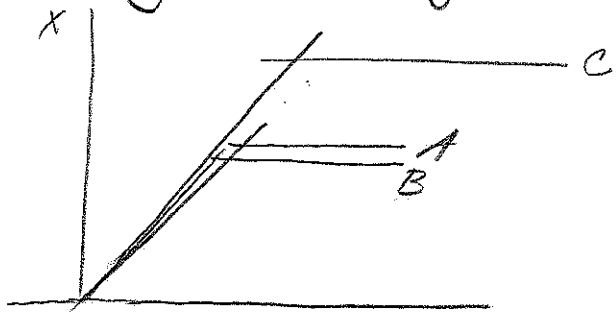
$Z = 20$.

The systems A, B, and C under consideration have the following values for D and Dmax

	A (current)	B	C
D	8.6	7.0	5.0
Dmax	4.6	5.1	3.1

The text shows optimistic bounds on throughput and response time.

Qualitatively, the light load bounds



are similar for all three. The heavy load bound for A is actually better than B

The company planning the interim upgrade investigated option B further with benchmark tests and concluded B was not useful.

Effect of Bottleneck Removal

The lesson of asymptotic bounds regarding bottleneck removal is that the heavy load optimistic bound on throughput, $\frac{1}{D_{max}}$, can be reduced by reducing the service time at the corresponding center. This will increase this bound on throughput until that service time drops below the old second-largest service time, making it the new D_{max}.

This same analysis holds for the heavy load optimistic bound on response time, $R \geq N D_{max} - Z$.

Other bounds will improve only if total service time decreases.

The next case study examines 3 system improvements to a system modeled as having a CPU, a slow disk, and a fast disk.

Based on the following measurements,

$T = 900$ sec
$B_1 = 400$ sec (CPU busy)
$B_2 = 100$ sec (slow disk busy)
$B_3 = 600$ sec (fast disk busy)

$C = 200$ jobs (completed jobs)
$C_2 = 2{,}000$ (slow disk operations)
$C_3 = 20{,}000$ (fast disk operations)

$Z = 15$ sec (think time)

calculate

$$D_1 = \frac{400 \text{ sec}}{200 \text{ jobs}} = 2, \quad D_2 = \frac{100 \text{ sec.}}{200 \text{ jobs}} = .5, \quad D_3 = \frac{600 \text{ sec}}{200 \text{ jobs}} = 3.$$

$$V_2 = \frac{2000 \text{ operations}}{200 \text{ jobs}} = 10 \qquad V_3 = \frac{20{,}000 \text{ operations}}{200 \text{ jobs}} = 100$$

$$S_2 = \frac{100 \text{ sec}}{2000 \text{ op.}} = .05 \qquad S_3 = \frac{600 \text{ sec.}}{20{,}000 \text{ op.}} = .03$$

$D = 5.5, \quad D_{max} = 3$

1. Replace CPU with one twice as fast.
   $D_1 = 1$
   $D = 4.5 \quad D_{max} = 3$

2. Shift some files from the faster disk to the slower disk to balance demands.

To calculate the primary effect of this, assume the same average number of visits will be required:

$$V_2 + V_3 = V_2' + V_3' = 110.$$

$V_i' = \frac{D_i'}{S_i}$, and we're setting $D_2' = D_3' = D'$

$$D'\left(\frac{1}{S_2} + \frac{1}{S_3}\right) = 110, \quad D'\left(\frac{1}{.05} + \frac{1}{.03}\right) = 110$$

$$D' \approx 2.06 = D_2 = D_3$$

$$D = 6.12, \quad D_{max} = 2.06$$

3. Add a second fast disk (center 4) to handle half the load of the busier, fast, disk.

$D_3 = 1.5$, $D_4 = 1.5$

$D = 5.5$ $D_{max} = 2$

4. Replace the CPU with the fast CPU and balance the load across 1 slow disk and two fast disks.

$D_1 = 1$

$D_2 = D_3 = D_4 = D'$ and

$$D'\left(\frac{1}{S_2} + \frac{1}{S_3} + \frac{1}{S_3}\right) = 110,$$

$$D'\left(\frac{1}{.05} + \frac{1}{.03} + \frac{1}{.03}\right) = 110, \quad D' \approx 1.27$$

$D \approx 4.8$, $D_{max} \approx 1.27$

The results show that option 1 does little to the optimistic bounds. Options 2 and 3 produce similar effects (while option 2 doesn't require any new equipment). Option 4 produces dramatic improvement

	D	Dmax
orig.	5.5	3
1	4.5	3
2	≈ 6.12	≈ 2.06
3	5.5	2
4	≈ 4.8	≈ 1.27

Balanced System Bounds

These are more refined bounds. Their validity rests on the validity of the separability assumptions detailed at the end of chapter 6, but basically requiring that the arrival rate (when relevant) doesn't depend on the number or position of jobs in the system,

that the completion rate at a center depends only on the number of jobs at that center (if that).

that the probability of a jobs moving from center j to center k is independent of the queue lengths at any of the centers, for all j and k.

The basic idea is to bracket the system in question with systems that are easier to analyze because they have the property of being <u>balanced</u>: the utilizations of all service centers are equal.

The techniques of Chapter 6 can be used to show that for a balanced system $U_k(N) = \frac{N}{N+K-1}$. Assume this.

We'll work through the consequences for a batch workload and use the results given in the text for transaction and terminal workloads.

It is intuitively satisfying that for separable systems with K service centers and total demand D, the one with $D_k = \frac{D}{K}$ has the highest throughput due to the absence of a bottleneck. This is eligible for theoretical verification.

The throughput of that balanced system is $X_{bal}(N) = \frac{U_k}{D_k} = \frac{N}{N+K-1} \frac{1}{D_{ave}}$

$$= \frac{N}{D+(N-1)D_{ave}}.$$

Thus $$X(N) \leq \min\left(\frac{1}{D_{max}}, \frac{N}{D+(N-1)D_{ave}}\right)$$

It is also intuitively satisfying that, of all systems with demand D and maximum demand D_{max}, the worst performance comes from the system with D/D_{max} centers with demand D_{max} and demand 0 at the remaining centers (this scenario gives us the worst possible bottlenecks).

- Yes $D/D_{max} \notin \mathbb{Z}$ makes the interpretation of K a little weird, but the math works anyway.

Using the balanced system utilization $\frac{N}{N+D/D_{max}-1}$, conclude that

$\frac{N}{N+D/D_{max}-1} * \frac{1}{D_{max}} \leq X(N)$. The full balanced bound inequalities are

$$\frac{N}{D+(N-1)D_{max}} \leq X(N) \leq \min\left(\frac{1}{D_{max}}, \frac{N}{D+(N-1)D_{ave}}\right)$$

For response time, use $N=XR$, $R=\frac{N}{X}$

$D+(N-1)D_{max} \geq \frac{N}{X(N)} \geq \max\left(ND_{max}, D+(N-1)D_{ave}\right)$ ie

$$\max\left(ND_{max}, D+(N-1)D_{ave}\right) \leq R(N) \leq D+(N-1)D_{max}$$