CSM: Operational Analysis

2016-17 Computer Science Tripos
Part II
Computer Systems Modelling:
Operational Analysis
by Ian Leslie

Richard Gibbens, Ian Leslie

Operational Analysis

Based on the idea of *observation* rather than a probabilistic description of system behaviour.

It is also concerned with quantities 'directly related' to these observed quantities.

Operational analysis makes very weak assumptions about the system being modelled...

... unlike simulation which requires detailed system knowledge, or the techniques from queuing theory which depend extensively on the probability distributions involved

We begin by examining some *fundamental quantities* and *operational laws*.

Basic Quanitites

We examine a system for some time recording the customer arrivals and departures, and define the quantities of interest:

- T, the length of time we observe the system
- A, the number of arrivals observed
- C, the number of departures (or completions of service) observed
- W, the job-time product: the sum of the durations of all customers over the observation period

If the system is a single resource, then we can also measure:

- B, the time for which the resource was busy.

Basic Quantities (2)

From these we can define the following quantities:

- *Arrival rate*, $\lambda := \frac{A}{T}$
  — the mean number of arrivals per unit time
- *Throughput*, $X := \frac{C}{T}$.
  — the mean number of departures per unit time
- *Mean number of customers*, $N := \frac{W}{T}$
  — the job-time product in terms of N and T
- *Mean residence time*, $R := \frac{W}{C}$
  — the job-time product in terms of R and C

Basic Quantities (3)

For a single resource, we can also define:

- *Utilization*, $U := \frac{B}{T}$

  — the proportion of the time that the resource is busy

- *Average service requirement*, $S := \frac{B}{C}$

  — the mean time that the resource spends for each departure

The utilization law

Our first "law" is just an algebraic identity

$$U := \frac{B}{T} = \frac{C}{T}\frac{B}{C} = XS$$

This is termed the *utilization law*.

For example, if the throughput (X) is 5 departures/sec and the service demand (S) is 0.1 sec/departure then the utilization (U) is 50%.

Little's law

Similarly, we can derive the familiar *Little's Law*

$$N := \frac{W}{T} = \frac{C}{T}\frac{W}{C} = XR$$

For example, if the throughput (X) is 5 customers/sec and the mean residence time is 1 sec then the average number in the system is 5.

Little's Law (2)

- Very weak assumptions about the system
- Applicable to a wide range of systems
- Can be applied recursively to subsystems and to individual resources — but take care that mutually-consistent values are used for X and R; in particular, whether they apply to the queue, the server or the entire system

An example

Observe system for $T = 10$ sec

4 customers spend 10 s in the system

One customer spends 5 s in the system

Then we have $W = 4 \times 10 + 1 \times 5 = 45$ s.

If also $A = C = 5$ then

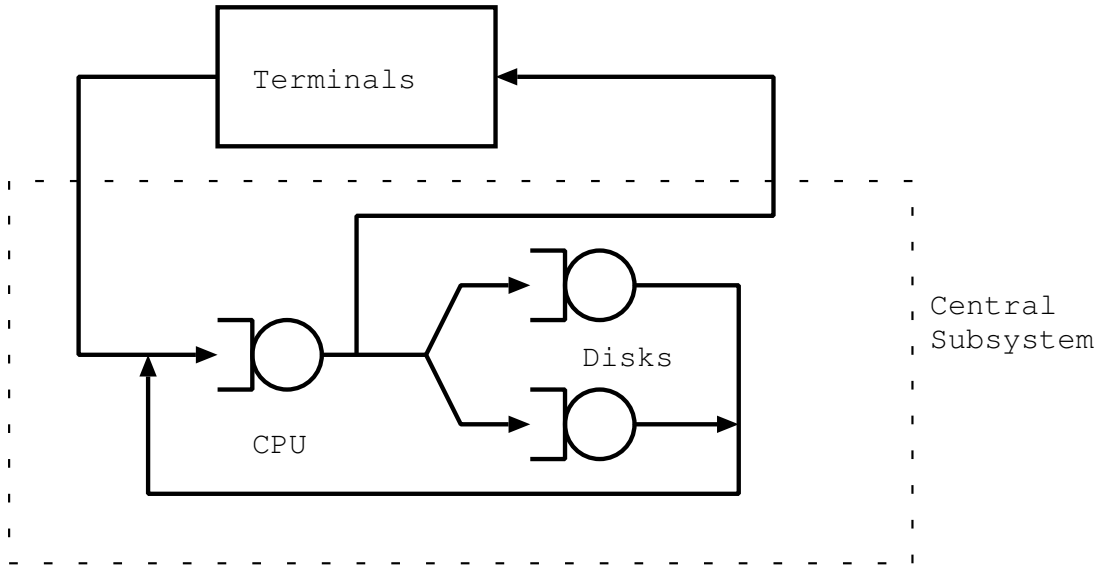
$X = C/T = 5/10 = 0.5$ customers per second

$\lambda = 5/10 = 0.5$ customers per second

$N = 45/10 = 4.5$ customers

$R = 45/5 = 9.0$ s per customer

A simple interactive system



Fixed number, M, of users logged on.

Customer is at the terminal whilst thinking.

The *think time*, Z, is the average time a user spends between receiving a prompt and responding.

A customer not thinking is inside the central subsystem.

Simple interactive system (2)

We can use Little's Law to relate some observable quantities in the central subsystem:

- N is the number of customers in the central subsystem $0 \leq N \leq M$
- X is the rate at which customers complete in the central subsystem
- R is the average time a customer spends in the central subsystem (intuitively equivalent to "response time")

Simple Interactive systeem (3)

If we observe that system throughput is $\mathbf{0.5}$ interactions per second and we find on average $\mathbf{7.5}$ users in the subsystem then

$$R = \frac{N}{X} = \frac{7.5}{0.5} = 15 \text{ s}$$

from Little's Law applied to the central subsystem.

Simple interactive system (4)

We can also apply Little's law to the entire system.

This is a closed system so the number of customers is fixed as M.

We can split the time spent during an interaction into the response time (R) and the *think time* (Z). The *residence time* is $R + Z$.

We consequently derive the *interactive system* version of Little's Law:

$$M = X(R + Z)$$

Simple interactive system (5)

With $\mathbf{10}$ users logged on, $\mathbf{5}$ s average think time and an average response time of $\mathbf{15}$ s,

$$X = \frac{M}{R+Z} = \frac{10}{15+5} = 0.5 \quad \text{interactions/sec}$$

Under heavy load (M large or Z small)

$$U \approx 1$$

Using the Utilization Law the throughput $X \approx \frac{1}{S}$ and hence

$$R = \frac{M}{X} - Z \approx MS - Z$$

Visit counts and forced flow

We now extend our notation to allow the modelling of multiple devices. Use subscripts $i = 1, 2, \ldots, K$ to identify each device, e.g. X_i is the throughput at device i. Assume that the service required by a customer is an inherent property of the *customer* not of the state of the system.

A *visit count* for a device is the number of completions at that device for every completion from the system

$$V_i := \frac{C_i}{C}$$

Where C_i is the number of completions at device i.

(Recall that in a *feed-forward* queueing network, $0 \leq V_i \leq 1$ because each customer visits a given device at most once.)

The forced flow law

Since $X = \frac{C}{T}$, we have that

$$X_i = \frac{C_i}{T} = \frac{C_i}{C}\frac{C}{T} = V_i X$$

the *Forced Flow Law*.

For example, if the throughput from the entire system is **20** customers per second and each customer visits a given device **3** times then the throughput of that device must be **60** completions per second.

If devices are load independent, then define the *service demand* a customer makes on a device i by

$$D_i := V_i S_i$$

Queue lengths at a server

Applying the utilization law at each device:

$$U_i = X_i S_i = (X V_i) S_i = X(V_i S_i) = X D_i$$

Similarly, applying Little's law at each device:

$$N_i = X_i R_i \,.$$

R_i is the residence time at device i and can be decomposed into the time spent queuing and the time spent in service, approximated by R_i^*:

$$\begin{aligned} R_i^* &= N_i S_i + S_i \\ &= R_i^* X_i S_i + S_i \\ &= R_i^* U_i + S_i \,. \end{aligned}$$

Queue lengths at a server (2)

Hence

$$R_i^* = \frac{S_i}{1 - U_i}.$$

So that

$$N_i = X_i R_i^* = \frac{X_i S_i}{1 - U_i} = \frac{U_i}{1 - U_i}.$$

Observe that

- N_i is zero when U_i is zero;
- N_i grows rapidly without bound as U_i approaches one.

Bottleneck analysis

A bottleneck in a system is a hindrance to progress.

Given the forced flow assumption, at high loads system performance is determined by the device with the highest utilization: *the bottleneck*.

The ratio of the completion rates of any two devices is

$$\frac{X_i}{X_j} = \frac{V_i X}{V_j X} = \frac{V_i}{V_j}.$$

Since $U_i = X_i S_i$, we have a similar property for utilizations

$$\frac{U_i}{U_j} = \frac{X_i S_i}{X_j S_j} = \frac{V_i S_i}{V_j S_j}.$$

Bottleneck analysis (2)

A system is *load independent* if

- V_i are intrinsic properties of customers,
- S_i are independent of the queue length at i.

In such cases, the throughput and utilization ratios are the same for all loads.

This can be used to determine asymptotes for X and R.

In general, $U_i \leq 1$ and $X_i \leq \frac{1}{S_i}$.

A device i becomes *saturated* as $U_i \to 1$

Thus, as $U_i \to 1$, we have that $X_i \to \frac{1}{S_i}$: device i is working as fast as it can and consequently serves one customer every S_i units of time.

Bottleneck analysis (3)

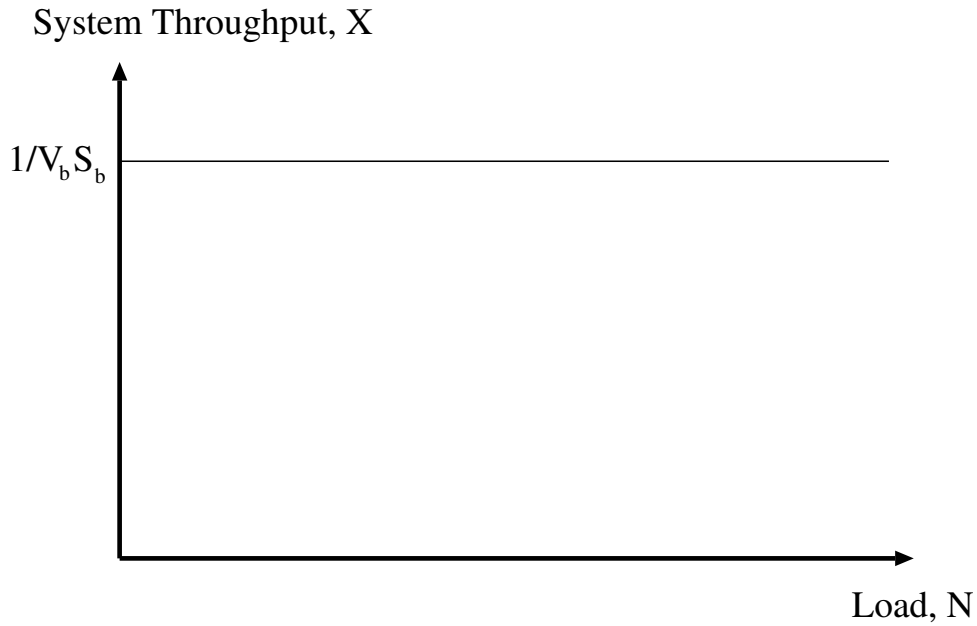
We use the subscript b to denote a device capable of saturating.

Since the utilization ratios are fixed, the device i with the largest $V_i S_i$ product will be the first to achieve 100% utilization as N increases:

$$V_b S_b = \max\{V_1 S_1, ..., V_K S_K\}$$

so the bottleneck is determined by both the device and workload (the V_i and S_i) properties.

Maximum throughput



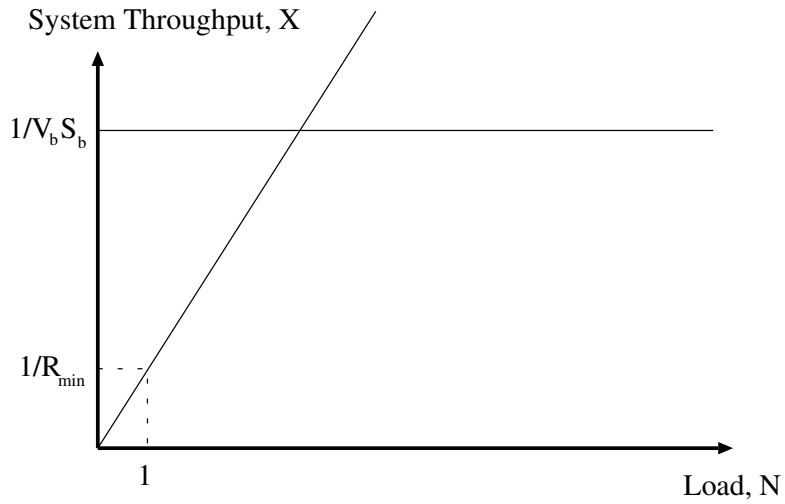
By the forced flow law

$$X = \frac{X_b}{V_b}$$

So, as $U_b \to 1$ and $X_b \to 1/S_b$

$$X_{\max} = \frac{X_b}{V_b} \to \frac{1}{V_b S_b}$$

Maximum throughput (2)

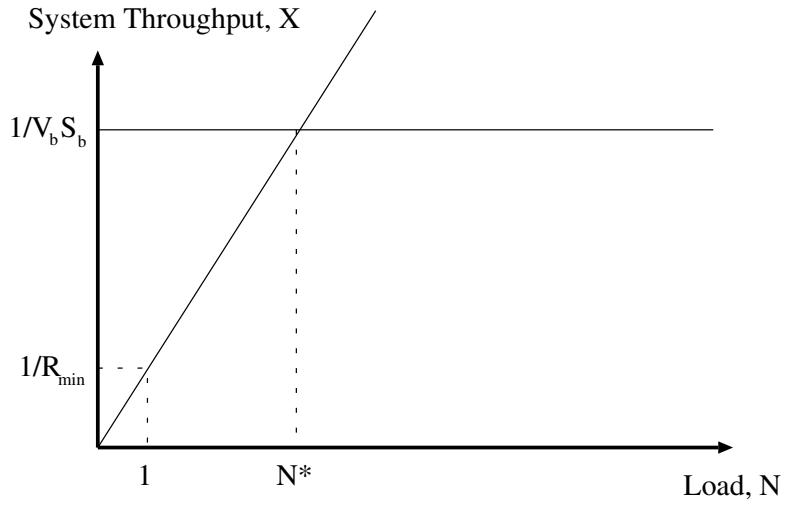


The total per-customer service required is

$$R_{\min} = \sum_{i=1}^{K} V_i S_i \qquad \Rightarrow \qquad X \leq \frac{N}{R_{\min}}$$

$R_{\min}$ denotes the smallest possible value of mean response time, occurring when $N = 1$.

Maximum throughput (3)

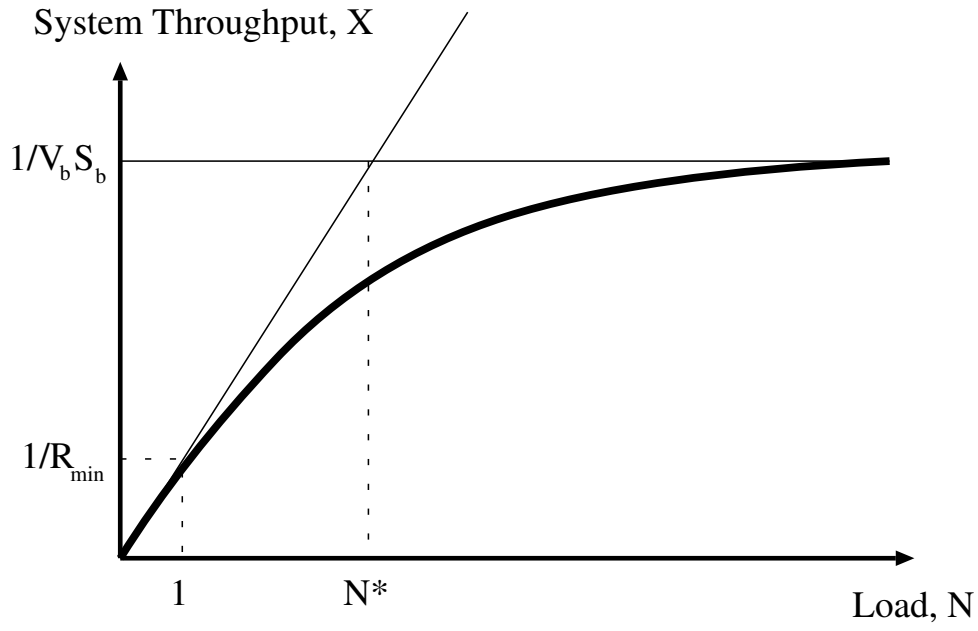


If $k \leq K$ jobs always avoid each other then

$$X = \frac{k}{R_{\min}} \leq \frac{1}{V_b S_b}$$

$$k \leq \frac{R_{\min}}{V_b S_b} = \frac{\sum_{i=1}^{K} V_i S_i}{V_b S_b} = N^*, \quad \text{say}$$

Maximum throughput (4)



- It stays below $1/(V_b S_b)$ because, at that point, a bottleneck will be operating at maximum utilization;
- It stays below the straight line $X = N/R_{min}$ because the throughput is limited by the number of customers in service.

Interactive response time

- X throughput;
- M terminals;
- Average think time Z;
- Recall the interactive system version of Little's law:

$$R = \frac{M}{X} - Z\,.$$

Intuitively the minimal response time, $R_{\min}$ is achieved when $M = 1$.

Similarly, the throughput is bound by the bottleneck device.

Interactive response time (2)

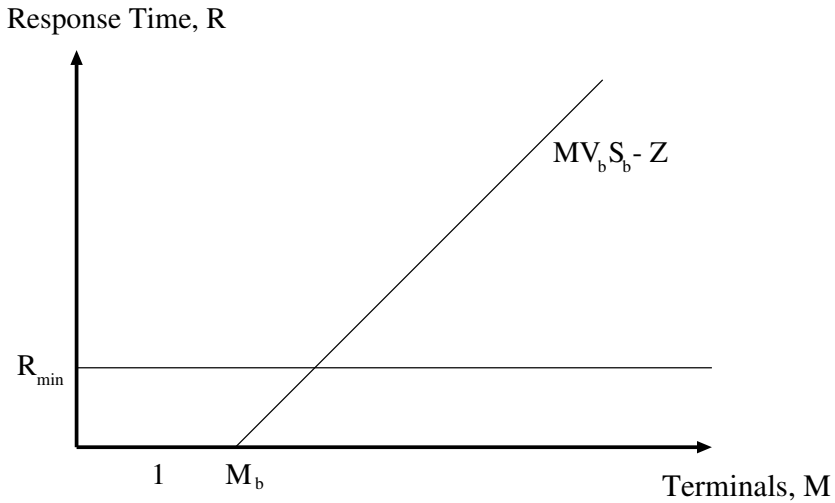
By considering a bottleneck device b:

$$X \leq \frac{1}{V_b S_b}$$

$$\Rightarrow \qquad R \geq M V_b S_b - Z$$

$$\Rightarrow \qquad R \geq M V_i S_i - Z \qquad \forall i \in \{1 \ldots K\}$$

Interactive response time (3)



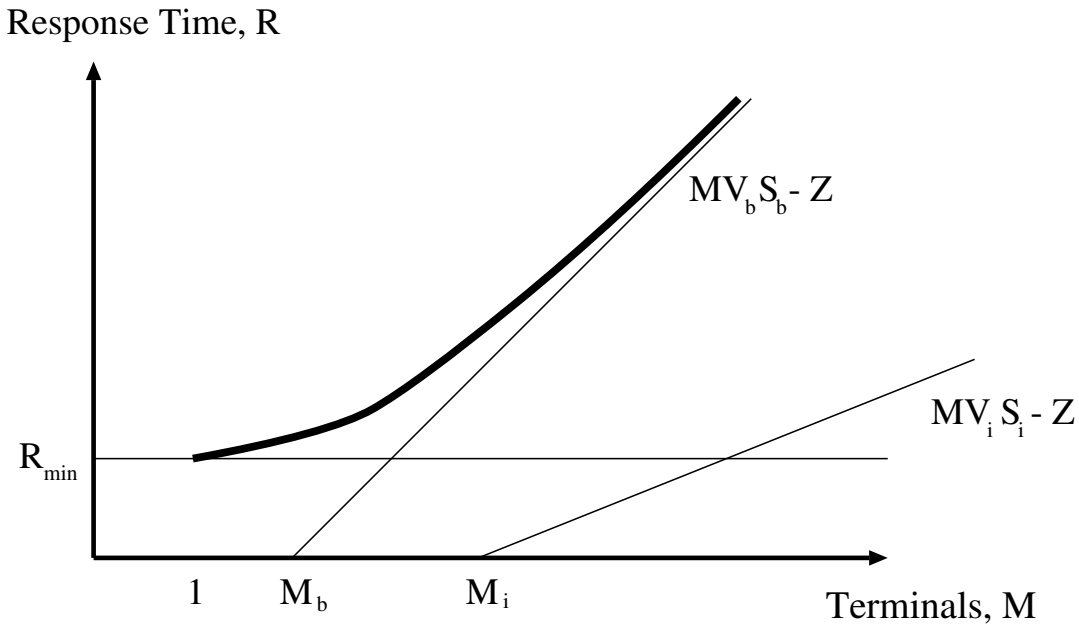
The response time asymptote meets the horizontal axis at

$$M_b = \frac{Z}{V_b S_b}$$

It intersects the minimum response time $R_{\min}$ at M_b^* (say) where

$$M_b^* V_b S_b - Z = R_{\min}\,.$$

Interactive response time (4)



Thus

$$M_b^* = \frac{R_{min} + Z}{V_b S_b} = N^* + M_b$$

When there are more than M_b^* terminals, queueing is *certain* to exist.

Summary

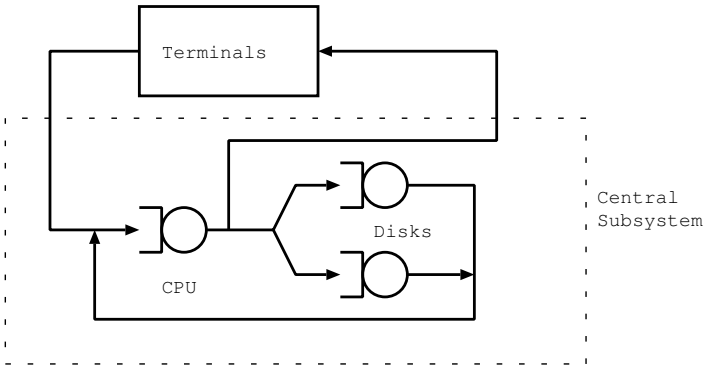
- The largest of the products $V_i S_i$ determines the bottleneck b.
- The sum of these products determines the smallest possible response time $R_{\min}$.
- Queueing cannot be avoided when N exceeds

$$N^* = \frac{R_{\min}}{V_b S_b}$$

- Queueing cannot be avoided in an interactive system when the number of logged-on terminals exceeds

$$M_b^* = N^* + \frac{Z}{V_b S_b}.$$

Example: interactive system

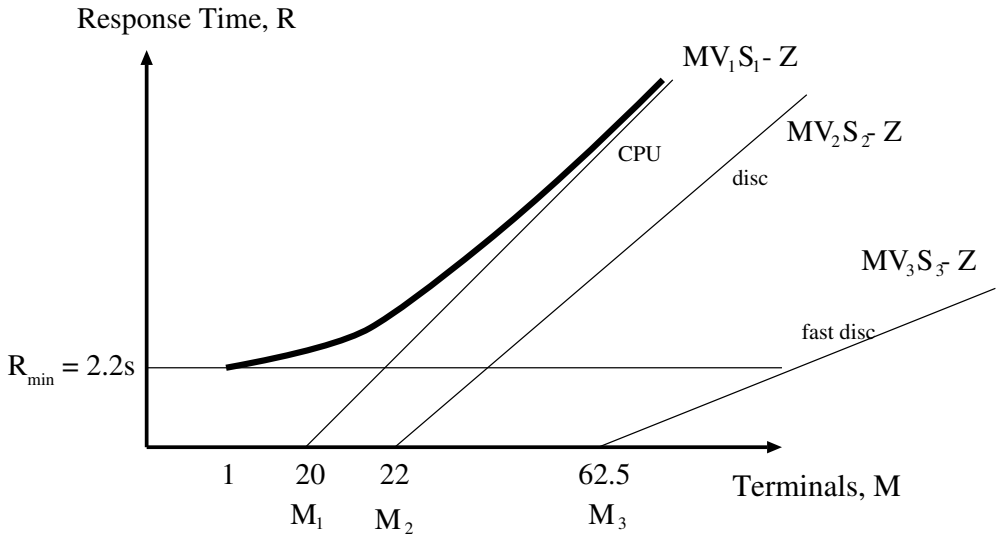


Suppose that $Z = 20$ s.

No.	device	S_i (s)	V_i	$D_i = V_i S_i$
1	CPU	0.05	20	1.00
2	disk	0.08	11	0.88
3	fast disk	0.04	8	0.32
			$R_{\min}$	2.20

Question: Is a 8 second response time feasible with 30 users logged on? If not, what changes are required?

Example: interactive system (2)



$$V_1S_1 = 1\text{s} \quad \textit{(bottleneck)}$$

$$V_2S_2 = 0.88\text{s}$$

$$V_3S_3 = 0.32\text{s}$$

$$\Rightarrow R_{\min} = \sum_{i=1}^{3} V_iS_i = 2.2\text{s}$$

Example: interactive system (3)

For $M = 30$, the response time asymptote requires $R \geq 30 \times 1 - 20 = 10 \ \ \mathbf{s}$.

So the answer is *no*, 8 second response time is not feasible with 30 users logged on.

We need to speed up the CPU. How much?

Example: interactive system (4)

To make a 8 second response time feasible, we need to speed up the CPU, so that the new service time obeys the condition

$$MV_1S_1' - Z \leq 8$$

or

$$S_1' \leq \frac{20+8}{30 \times 20} = .047 \ \ \mathrm{s}$$

which is a 7% speed up in the CPU.

Then $V_1S_1' = 0.93$ is still the largest product so the CPU is still the bottleneck.

Example: interactive system (5)

Question: Is a 10s response time feasible when 50 users are logged on? If not, how much CPU speedup is required?

If $S_1 \to 0$, the disk will become bottleneck

$$R \geq MV_2S_2 - Z$$

For $M = 50$, this is

$$R \geq 50 \times 0.88 - 20 = 24 \ \ \mathrm{s}$$

so a 10s response time is not feasible with $M = 50$ and no amount of CPU speedup is capable of achieving it.

Balanced system bounds

Balanced system bounds provide *tighter* bounds at mid-range loads than bottleneck analysis

A system is *balanced* if for any load the utilizations of all devices are equal.

Balanced systems exhibit the following important property

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

So the system throughput is given by

$$X(N) = \frac{U_i}{D_i} = \frac{N}{N + K - 1} \times \frac{1}{D_i}.$$

Balanced system bounds (2)

For example,

$$\begin{array}{lll} N = 1 & K = 2 & U_1 = U_2 = \frac{1}{2} \\ N = 1 & K = 100 & U_1 = \ldots = U_{100} = \frac{1}{100} \\ N = 100 & K = 2 & U_1 = U_2 = \frac{100}{101} \\ N = 2 & K = 2 & U_1 = U_2 = \frac{2}{3} \end{array}$$

We observe the system to determine

- $D_{\max}$ — maximum demand at any device;
- $D_{\min}$ — minimum device demand;
- D_{av} — average demand at each device;
- $D = D_{\text{tot}}$ — total demand across all devices.

So, we have $D_{\text{av}} = D/K$.

Balanced system bounds (3)

Consider imaginary balanced systems related to our system

pess1: balanced system with K devices each with a demand of $D_{\max}$;

opt1: balanced system with K devices each with a demand of $D_{\min}$.

The throughput of system **pess1** is

$$\frac{N}{N+K-1} \times \frac{1}{D_{\max}}.$$

The throughput of system **opt1** is

$$\frac{N}{N+K-1} \times \frac{1}{D_{\min}}.$$

Balanced system bounds (4)

So for the system being modelled we have

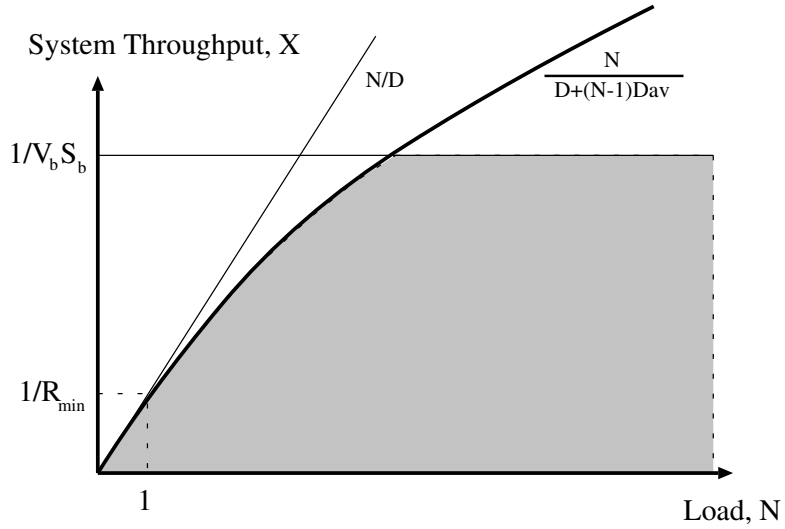
$$\frac{N}{N+K-1} \times \frac{1}{D_{\max}} \leq X(N) \leq \frac{N}{N+K-1} \times \frac{1}{D_{\min}}.$$

Since $N = XR$ we have,

$$(N+K-1)D_{\max} \geq R(N) \geq (N+K-1)D_{\min}.$$

This will give tighter bounds on the mid-range performance than the bottleneck bounds.

Balanced system bounds (5)

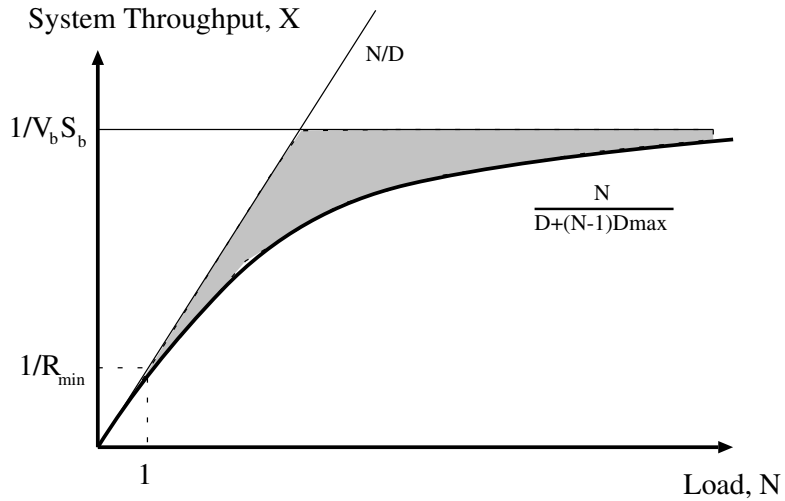


We can do even *better* by considering the *best* performance that the system can achieve which occurs when the load is spread out evenly among all the devices.

opt2: $D_{\mathbf{av}}$ at each of the K devices

$$X(N) = \frac{N}{N+K-1} \times \frac{1}{D_{\mathbf{av}}} = \frac{N}{D+(N-1)D_{\mathbf{av}}}.$$

Balanced system bounds (6)



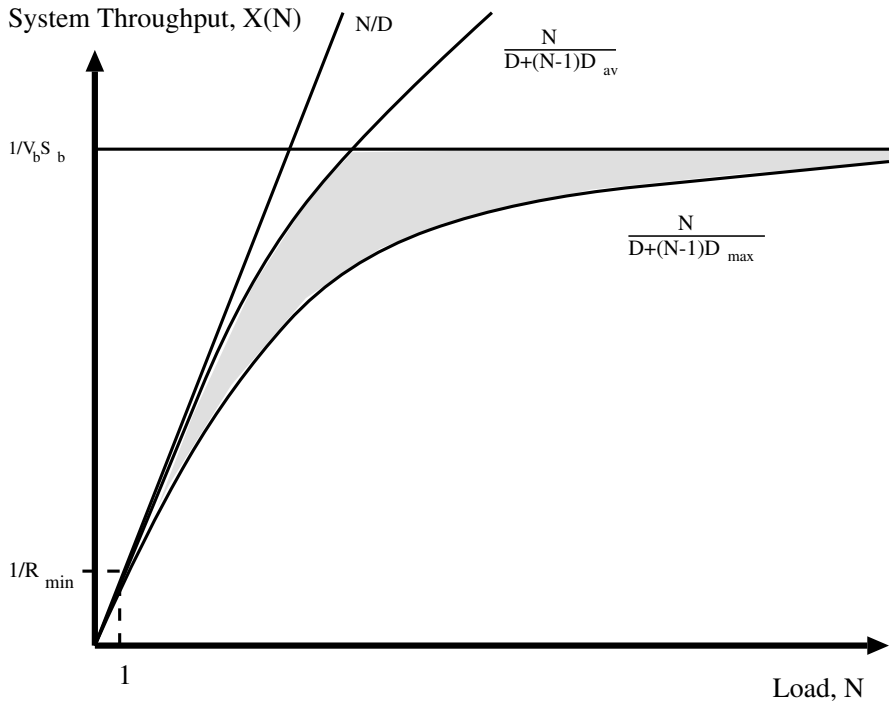
Now, what is the *worst* system subject to the constraints that D and $D_{\max}$ remain fixed?

Answer: place $D_{\max}$ at as many devices as possible and 0 at the rest

pess2: $D_{\max}$ at each of the $\frac{D}{D_{\max}}$ devices

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

Balanced system bounds (7)



$$\frac{N}{D + (N-1)D_{\max}} \leq X(N) \leq \frac{N}{D + (N-1)D_{\text{av}}}$$

Balanced system bounds (8)

As $N \to \infty$,

$$\frac{N}{D + (N-1)D_{\max}} \to \frac{1}{D_{\max}}$$

Note that at high loads the bottleneck bounds are the limiting high bound.

The asymptotic bottleneck bound $\frac{1}{D_{\max}}$ and the optimistic balanced bound intersect at $N^{\dagger}$ where

$$\frac{1}{D_{\max}} = \frac{N^{\dagger}}{D + (N^{\dagger} - 1)D_{\text{av}}}.$$

So that

$$N^{\dagger} = \frac{D - D_{\text{av}}}{D_{\max} - D_{\text{av}}}.$$