Computer Systems Modelling

Operational Analysis

Simon Crosby

Part II, CST, 1996

Operational Analysis

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

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

It is also concerned with quantities "directly related" to observables

Begin by examining some **fundamental quantities and operational laws**

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

- T is the length of time we observe the system
- A is the number of arrivals observed
- C is the number of departures (or completions of service) observed.

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

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

From these quantities we can define the following quantities

- Arrival rate, $\lambda \stackrel{\triangle}{=} \frac{A}{T}$
- Throughput, $X \stackrel{\triangle}{=} \frac{C}{T}$.

For a single resource, we can also define:

- Utilization, $U \stackrel{\triangle}{=} \frac{B}{T}$
- Average service requirement, $S \stackrel{\triangle}{=} \frac{B}{C}$

The Utilization Law

Our first "law" is just an algebraic identity

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

Our next law requires a further quantity to be observed

- W, the job-time product which is sum of the durations of all customers in the system over the observation period.

From this we can define

- Average number of requests in the system, $N \triangleq \frac{W}{T}$
- Average residence time per request, $R \triangleq \frac{W}{C}$

Again, simple manipulation gives the by now familiar **Little's Law**

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

- very weak assumptions about the system
- applicable to a wide range of systems
- can be applied recursively to subsystems and to individual resources

An Example

Observe system for $T = 10$ sec

4 customers spend 10 s in the system

One customer who spends 5 s

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

If also $A = C = 5$ then

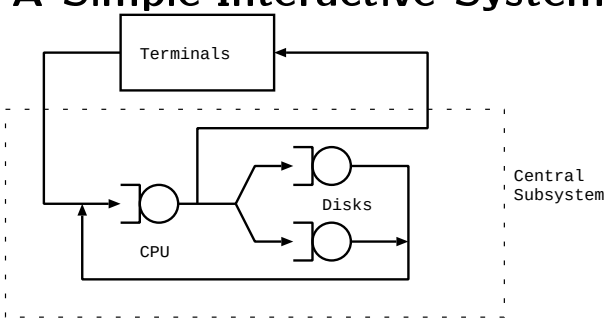
$X = C/T = 5/10 = 0.5$ customers per second

$\lambda = 5/10 = 0.5$ 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 somewhere inside the central subsystem

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")

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

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

from Little's Law

We can also "view" R as response time rather than residence time to get the **interactive system** version of Little's Law

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

where R is the response time and Z is the think time

The residence time is $R + Z$ (which we were calling R above)

With 10 users logged on, 5 s average think time and an average response time of 15 s,

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

Under heavy load (M large or Z small)

$$U \approx 1$$

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

$$R = \frac{M}{X} - Z \approx \frac{1}{S}M - Z$$

$\Longrightarrow$ The response time grows approximately linearly with the number of users M.

Visit Counts and Forced Flow

Extend our notation to allow the modelling of multiple $i = 1, 2, \ldots, K$ devices

e.g. Throughput at a device $i \stackrel{\triangle}{=} X_i$ etc

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

$$V_i \stackrel{\triangle}{=} \frac{C_i}{C}$$

Where C_i is the number of completions at device i.

The Forced Flow Law

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

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

the **Forced Flow Law**

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

$$D_i \triangleq V_i S_i$$

We can then determine the utilization at each device

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

If we also know the N_i

$$N_i = R_i X_i \Longrightarrow R_i = \frac{N_i}{X_i}$$

by applying Little's Law at each device, where R_i is the residence time at device i.

We can infer X_i from U_i but can not find R_i without further assumptions

We can approximate it by R_i^*

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

Reducing this we get

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

and

$$\begin{aligned} N_i &= R_i^* X_i \\ &= \frac{S_i X_i}{1 - U_i} \\ &= \frac{U_i}{1 - U_i} \end{aligned}$$

which looks remarkably like the M/M/1 queueing system.

Bottleneck Analysis

A bottleneck in a system is a hindrance to movement or 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

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

If:
V_i are intrinsic properties of customers and
S_i independent of the queue length at device i

then the system has **load independent** parameters

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 $X_i \leq \frac{1}{S_i}$

Device i is **saturated** if $U_i \to 1$

Thus, as $U_i \to 1$, we see $X_i \to \frac{1}{S_i}$

Use the subscript b to denote a device capable of saturating

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

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

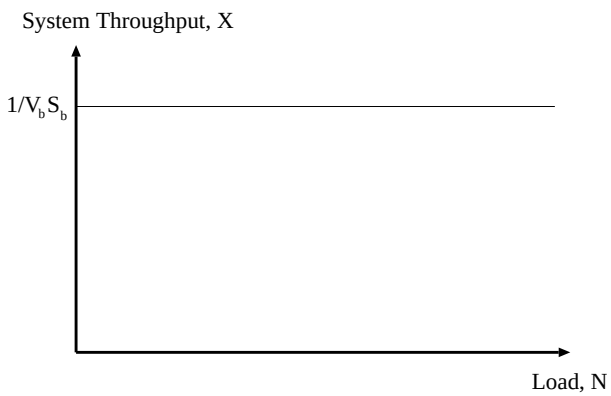
so the bottleneck is determined by device and workload (the V_i and S_i) properties

As N increases, $U_b \to 1$ and $X_b \to \frac{1}{S_b}$

Now, $XV_b = X_b$ so that

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

is the maximum possible system throughput as N increases.

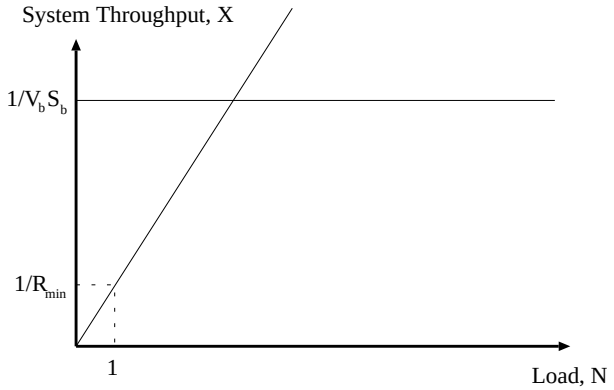


The total service required by a customer in the system is

$$R_{\min} = \sum_{i=1}^{K} V_i S_i$$

This also denotes the smallest possible value of mean response time

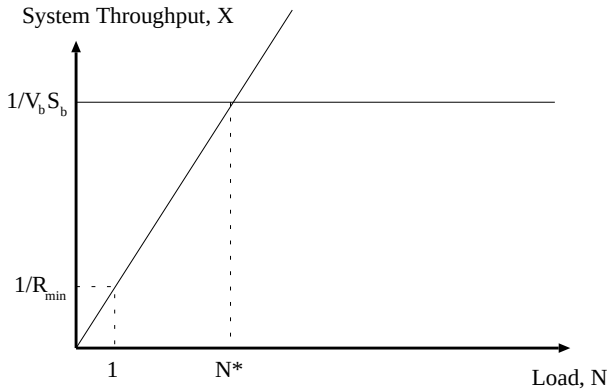
In fact, $R_{\min}$ is the response time when $N = 1$



This implies that $X = \frac{1}{R_{\min}}$ when $N = 1$ and that for $N > 1$, $X \leq \frac{N}{R_{\min}}$.

Assume that $k \leq K$ jobs always manage to avoid each other so that $X = \frac{k}{R_{\min}}$, we would require that

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

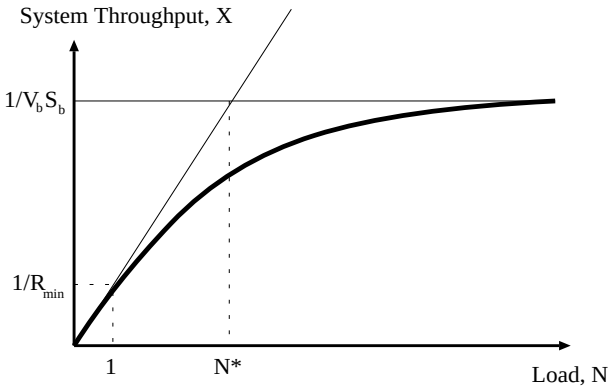


Alternatively $k \leq N^*$ where

$$N^* = \frac{R_{\min}}{V_b S_b} = \frac{\sum_{i=1}^{K} V_i S_i}{V_b S_b}$$

N^* represents a load beyond which queueing is certain to exist

System Throughput



System throughput, X, rises monotonically from $\frac{1}{R_{\min}}$ at $N = 1$ to the asymptote $\frac{1}{V_b S_b}$

It stays below the straight line $X = \frac{N}{R_{min}}$

Interactive Response Time

- M terminals
- Average think time Z
- Mean response time

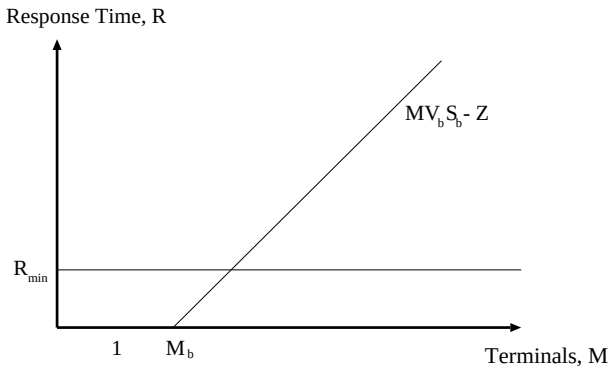
$$R = \frac{M}{X} - Z$$

When $M = 1$, we must have $R = R_{\min}$

The throughput, X, cannot exceed $\frac{1}{V_b S_b}$, so

$$R \geq M V_b S_b - Z \geq M V_i S_i - Z, \qquad i = 1, \ldots, K$$

by the choice of device b



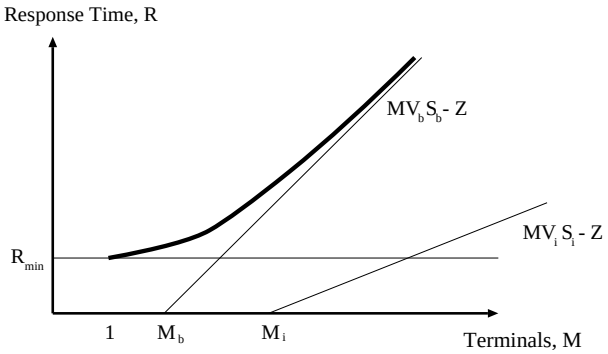
The response time asymptote intersects the x axis at

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

It intersects the minimum response time R_{min} at M_b^* where

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

System Response Time



Thus

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

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

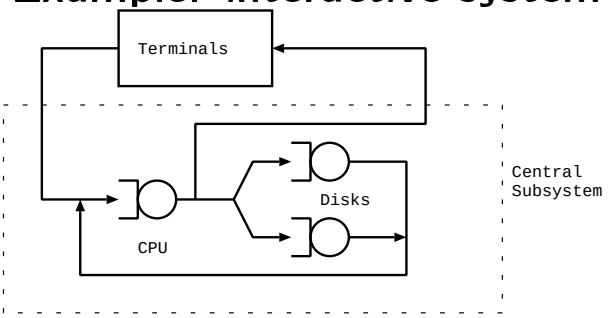
Summary

- the largest of the $V_i S_i$ determines the bottleneck b.
- the sum of these products determines the smallest possible response time R_{min}.
- queueing cannot be avoided when N exceeds $N^* = \frac{R_{\text{min}}}{V_b S_b}$.
- queueing cannot be avoided in an interactive system when $N^* + \frac{Z}{V_b S_b}$ terminals are logged on

Summary

- Reducing $V_i S_i$ for a non-bottleneck device will not move the $\frac{1}{V_b S_b}$ asymptote
- It may improve R_{min}, but it will not improve response time at high loads
- Reducing $V_i S_i$ for the bottleneck will
  - raise the $\frac{1}{V_b S_b}$ asymptote on the throughput graph
  - lower the $MV_b S_b - Z$ asymptote on the response time graph
  - reduce R_{min}
- Asymptotes will remain relevant only as long as $V_b S_b$ remains the largest of the $V_i S_i$

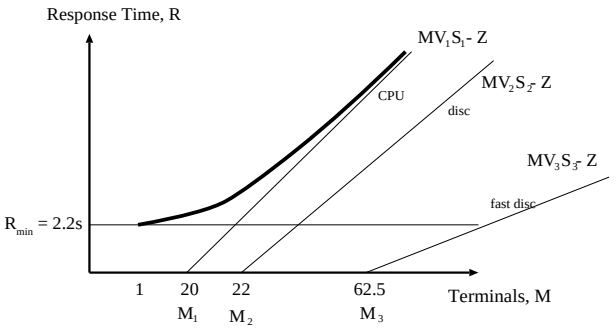
Example: interactive system



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 an 8 second response time feasible with 30 users logged on? If not, what changes are required?



The $V_i S_i$ products are $V_1 S_1 = 1$ s, $V_2 S_2 = 0.88$ s and $V_3 S_3 = 0.32$ s

The minimum response time is 2.2 s (since $1 + 0.88 + 0.32 = 2.20$) and the CPU is the bottleneck

For $M = 30$, the response time asymptote requires that

$$R \geq 30 \times 1 - 20 = 10 \quad \text{s}$$

So the answer is **no**

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

$$MV_1 S_1' - Z \leq 8$$

or

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

which is a 7% speed up in the CPU.

Then $V_1 S_1' = 0.93$ is still the largest product so the CPU is still the bottleneck.

Question: Is 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:

$$R \geq MV_2 S_2 - Z$$

For $M = 50$, this is

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

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 that

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

So the system throughput is given by

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

For example,

$N = 1$	$K = 2$	$U_1 = U_2 = \frac{1}{2}$
$N = 1$	$K = 100$	$U_1 = \ldots = U_{100} = \frac{1}{100}$
$N = 100$	$K = 2$	$U_1 = U_2 = \frac{100}{101}$
$N = 2$	$K = 2$	$U_1 = U_2 = \frac{2}{3}$

Look at the system and find

- $D_{\max}$ – 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 $D_{\text{av}} = D/K$

We can construct 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

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

So for the system being modelled we have

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

And since $N = XR$,

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

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

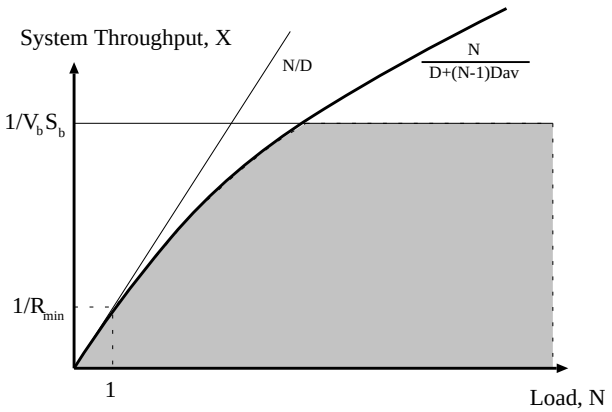
We can do even better using the following line of argument

What is the **best** performance that the system can achieve?

Answer: spread the load out evenly among all the devices.

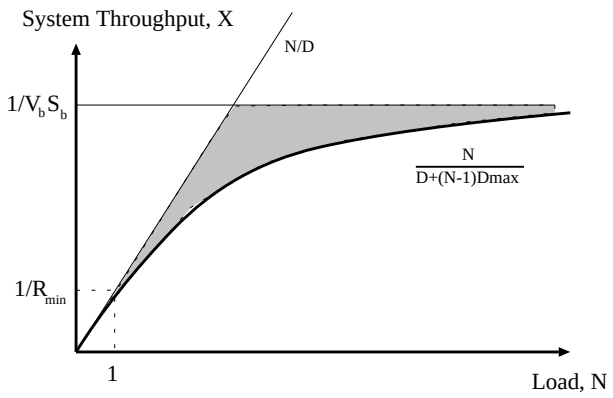
opt2: D_{av} at each of the K devices with throughput

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



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 $\frac{D}{D_{\text{max}}}$ devices with throughput

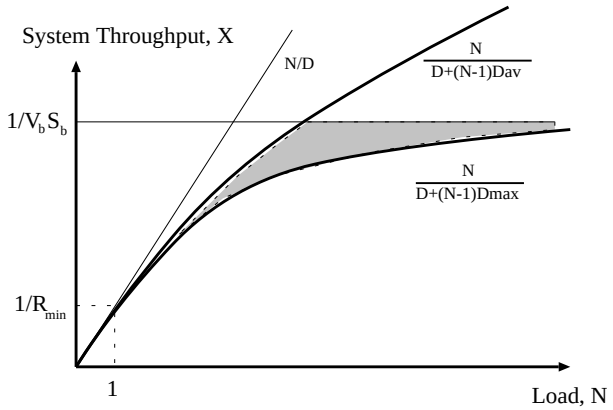
$$\frac{N}{N+K-1} \times \frac{1}{D_{\text{max}}} = \frac{N}{D+(N-1)D_{\text{max}}}$$

So that for the system being modelled

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

And

$$D+(N-1)D_{\text{max}} \geq R(N) \geq D+(N-1)D_{\text{av}}$$



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

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

So that

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

System Measurement

We call devices which measure system performance **monitors**

Can be in hardware, software or hybrid software/hardware

Whatever form a monitor takes, we can describe it in terms of monitor **artifact** and monitor **power**.

Monitor artifact describes the side effects that operation of the monitor has on the system being observed

Software monitors will have artifact depending on processing requirements

Monitor power has at least 5 dimensions:

- domain
  - the class of activities observable
- input rate
  - the maximum frequency at which events can be recognized and recorded
- input width
  - the amount of information the monitor can extract and process when a monitored event occurs
- recording capacity

  - the amount of memory available for storing extracted information
  - for hardware monitors (in particular):
    * event activity can be measured by a single counter
    * average duration can be derived from two values: accumulated duration of all occurrences of the activity and the total number of occurrences
    * distributional information requires much more memory

- monitor resolution
  - the resolution of the time clock from which the monitor derives timing information