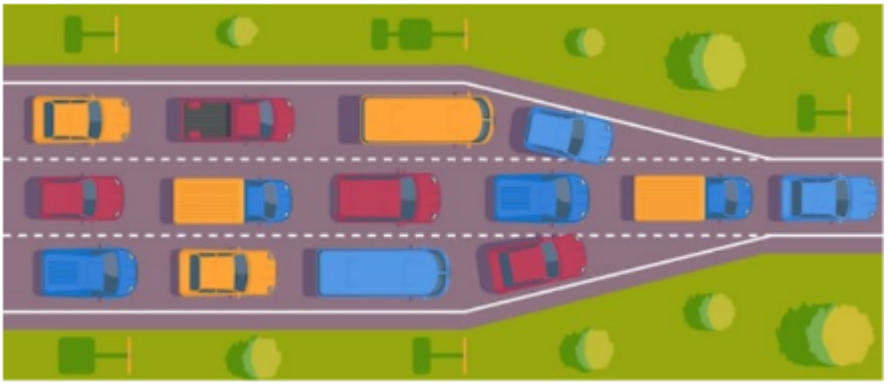
# Performance Bounds

Provide valuable insight into the primary factors affecting the performance of computer system
Can be computed quickly and easily therefore serve as a first cut modeling technique
Several alternatives can be treated together
What happens in bottleneck analysis. Important as the formulas are very simple and moreover since we will see what matters in a system and what is a bottleneck and see the impact on the performance of each single component.
We will consider single class systems only
Determine asymptotic bounds, i.e., upper and lower bounds on a system's performance indices X and R: In our case, we will treat X and R bounds as functions of number
of users or arrival rate (i.e., l or N)
Advantages of bounding analysis:
Highlight and quantify the critical influence of the system bottleneck

## Bottleneck

The resource within a system which has the greatest service demand is known as the bottleneck resource or bottleneck device, and its service demand is $max_k\{D_k\}$, denoted $D_{max}$
The bottleneck resource is important because it limits the possible performance of the system
This will be the resource which has the highest utilisation in the system



## Bounding Analysis

Advantages of bounding analysis:

- Highlight and quantify the critical influence of the system bottleneck
- Can be computed quickly, even by hand
- Useful in System Sizing:
- Based on preliminary estimates (quickness)
- This kind of studies involve typically a large number of candidate configurations with a single critical resource (e.g., CPU) dominant and the other configured accordingly: treated as one alternative

- Useful for System Upgrades…
  The considered models and the bounding analysis make use of the following parameters:
- K, the number of service centers
- D, the sum of the service demands at the centers, so $D = \sum_k D_k$
- $D_{max}$, the largest service demand at any single center
- Z, the average think time, for interactive systems
- X, the system throughput
- R, the system response time
  And the following performance quantities are considered:
- X, the system throughput
- R, the system response time
  **Asymptotic Bounds**
  Are derived by considering the (asymptotically) extreme conditions of light and heavy loads:
- Optimistic: X upper bound and R lower bound
- Pessimistic: X lower bound and R upper bound
  Under the extreme conditions of:
- Light load(Optimistic bound)
- Heavy load(Pessimistic bound)
  Under the 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
  Not totally true that demands not depend on requests at the current time(example is cashier in a discount market).

## Opens models

## *Open models:* less information than in closed models...

***X* bound** = the *maximum arrival rate* that the system can process

if $\lambda$ > ***X*** bound → the system ***SATURATES***

new jobs have to wait an indefinitely long time

Remembering that $U_k = XD_k$ $$U_{max}(\lambda) = \lambda D_{max} \leqslant 1$$

The ***X*** bound is calculated as: $$\lambda_{sat} = \frac{1}{D_{max}}$$

***R* bounds** = the largest and smallest possible **R** experienced at a given $\lambda$ investigated only when $\lambda < \lambda_{sat}$ (otherwise the system is unstable!)

**2 extreme situations:**

1. If no customers interferes with any other (= no queue time)

   Then ***R = D, with D*** = $\sum_k D_k$

There is no pessimistic bound on R:

- if n customers arrives together every $n/\lambda$ time units (the system arrival rate is $n/(n/\lambda) = \lambda$)
- 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 $\lambda$, it is possible to pick a batch size n sufficiently large that the bound is exceeded
  There is no pessimistic bound on response times, regardless of how small the arrival rate $\lambda$ might be

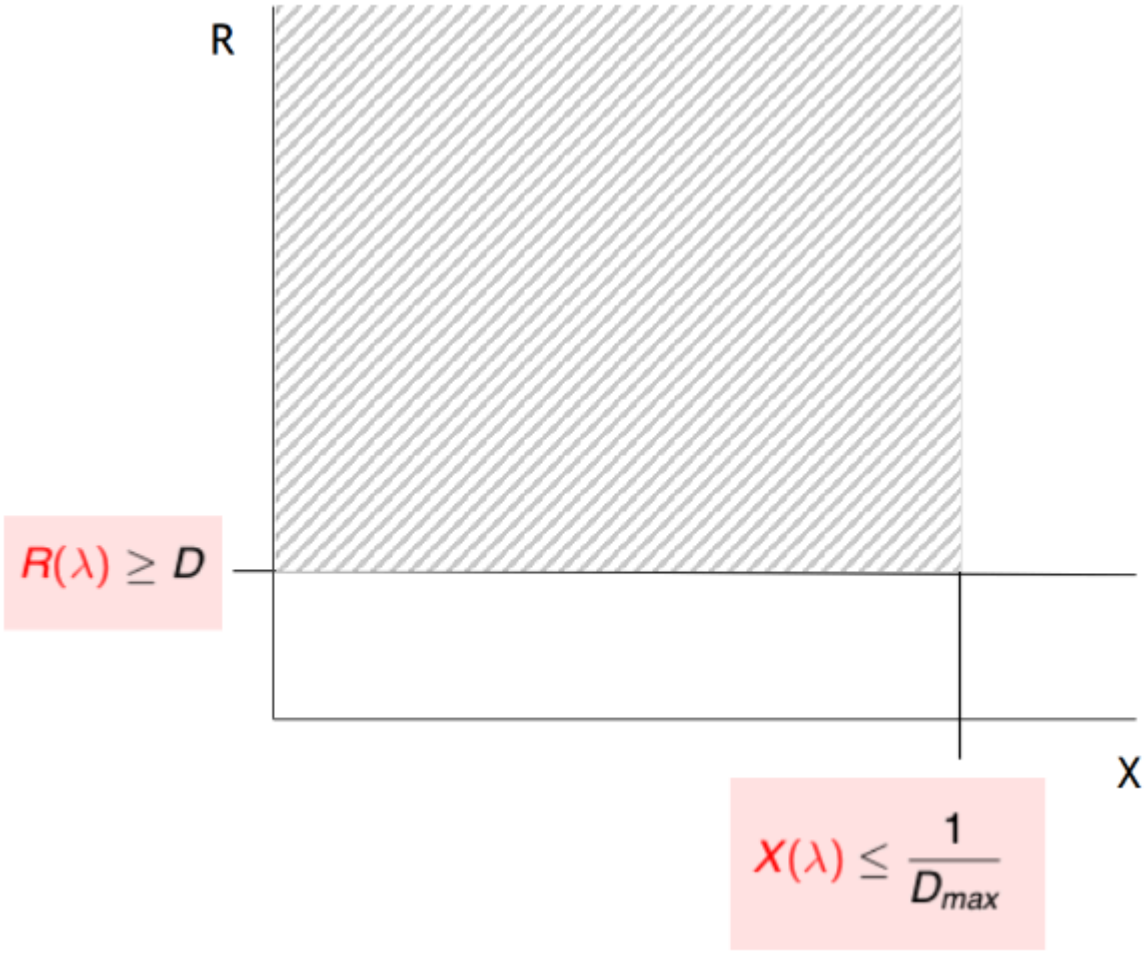
In the worst case($\lambda$ is low and N is large) we can adjust N in a way that we can identify every response time bound

**Bound for X($\lambda$)**

$$X(\lambda) \leq \frac{1}{D_{max}}$$

**Bound for R($\lambda$)**

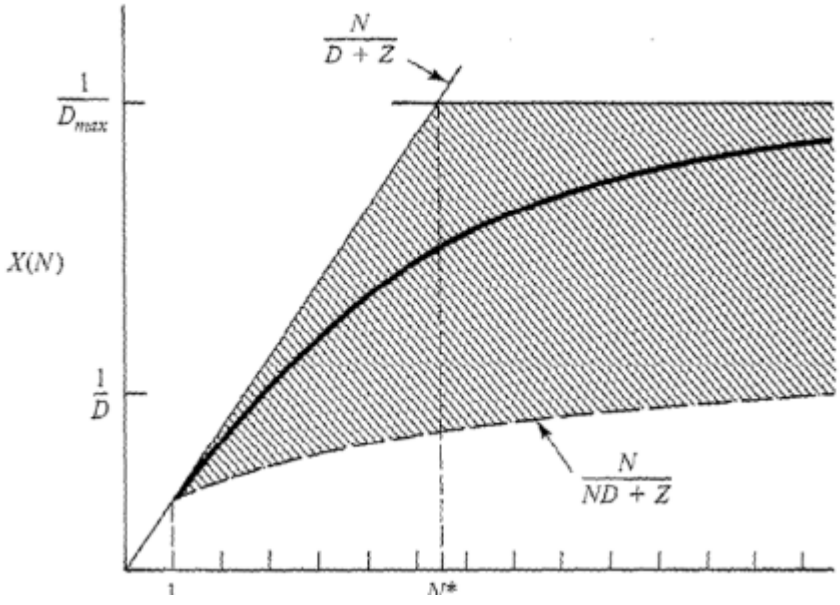
$$R(\lambda) \geq D$$




## Closed models

System where there is a determined number of users in the system at each time.
X bounds considered first, then converted in R bounds using Little's Law Light Load situation

(lower bounds):




1 customer case:
N = X (R + Z)
1 = X (D + Z)
Then X is:
X = 1 / (D + Z)

**Light Load situation (lower bounds):**

Adding customers:
Smallest X obtained with largest R,i.e., new jobs queue behind others already in the system
In closed models, the highest possible system response time occurs when each job, at each station, founds all the other N-1 costumers in front of it
In this case the X is:
X = N / (ND + Z)
$\lim_{N\to\infty} N/(ND+Z) = 1/D$
With a small X I'll have to consider a greater R and is a case where I am moving requests from one resources to another.

**Light Load situation (upper bounds):**

Adding customers:
Largest X obtained with the lowest response time R
The lowest response time can be obtained if a job always finds the queue empty and always starts being served immediately
In this case the X is:
X = N / (D + Z)

**Heavy Load situation (upper bound):**

$$U_k(N) = X(N)D_k \leq 1$$

Since the first to saturate is the Bottleneck (max):

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

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

N*: Particular population size determining if the light or the heavy load optimistic bound is to be applied*

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

$**R(N) bounds: **Letussimplyrewritethepreviousequation, consideringthat: X(N) = N/(R(N$

$$\frac{N}{ND+Z}\le\frac{N}{R(N)+Z}\le min(\frac{1}{D{max}},\frac{N}{D+Z})$$

$And to have R as numerator we invert the members and we have $max(D_{max}, \frac{D+Z}{N}) \leq \frac{R(N)+Z}{N} \leq \frac{ND+Z}{N}$

From which we have Bound for R(N): $max(D, ND_{max} - Z) \leq R(N) \leq ND$