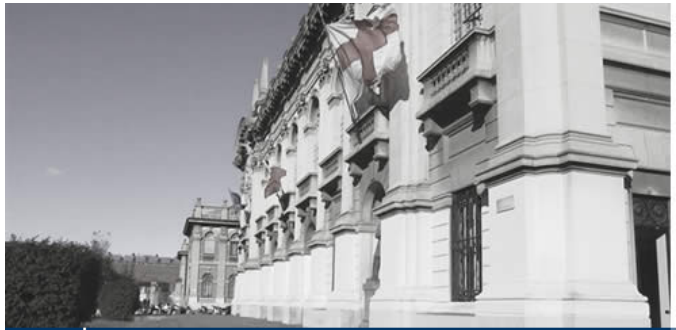


POLITECNICO DI MILANO

Computing Infrastructures





Performance Bounds

Prof. Gianluca Palermo

Credits: Hilston, Marzolla, Mirandola, Gribaudo, Zahorian, Lazowska, Ardagna

Performance bounds

- Provide valuable insight into the primary factors affecting the performance of computer systems
- Can be computed quickly and easily therefore serve as a first cut modeling technique
- Several alternatives can be treated together

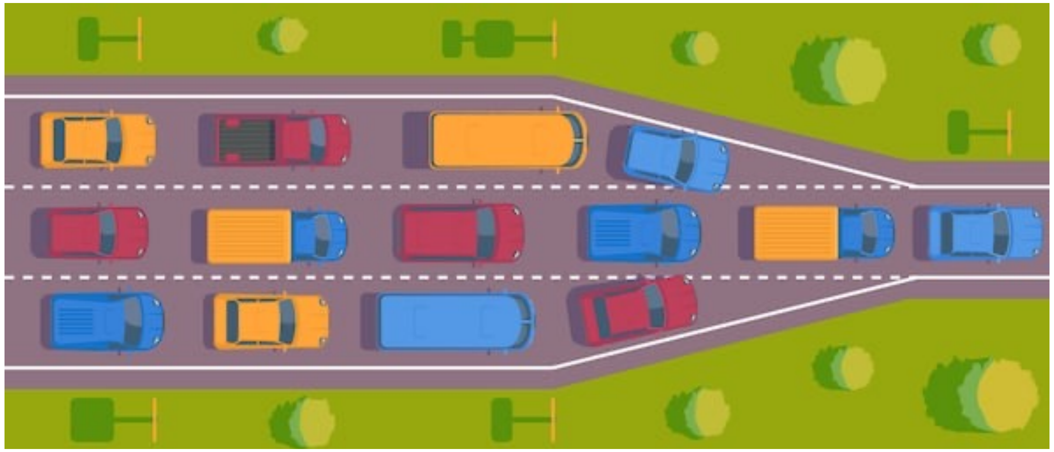
- **Bound Analysis:**
  - We will consider single-class systems only
  - Determine *asymptotic bounds*, *i.e.*, ***upper*** and ***lower bounds*** on a system’s performance indices X and R:
    - In our case, we will treat X and R bounds as functions of Number of users/Arrival Rate (i.e., λ/N)
  - Advantages 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



Advantage 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…

Notation

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

And the following performance quantities are considered:

- X, the system throughput
- R, the system response time

Bounding Analysis - *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*
  - *Heavy load*

- **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

Bounding Analysis - *Asymptotic bounds*

Open models: less information than in closed models…

***X* bound** = the $\underbrace{\textit{maximum arrival rate}}$ that the system can process

if λ > *X* bound → the system $\underbrace{\textbf{\textit{SATURATES}}}$

new jobs have to wait an indefinitely long time

Remember that $U_k = XD_k$

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

The *X* bound is calculated as:

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

Open models:

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

2 extreme situations:

- If no customers interferes with any other **(= no queue time**)

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

Open models:

- If ***n*** customers arrive together every ***n/λ*** time units (the system arrival rate is n /(n/ λ)= λ) there is no pessimistic bound on ***R***
  - ***BATCH OF ARRIVAL!***
- 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
  - ***THE BATCH CAN BE EXTREMELY LONG! N->Infinite***
- **There is no pessimistic bound on response times, regardless of how small the arrival rate λ might be**

Bounding analysis: Open models

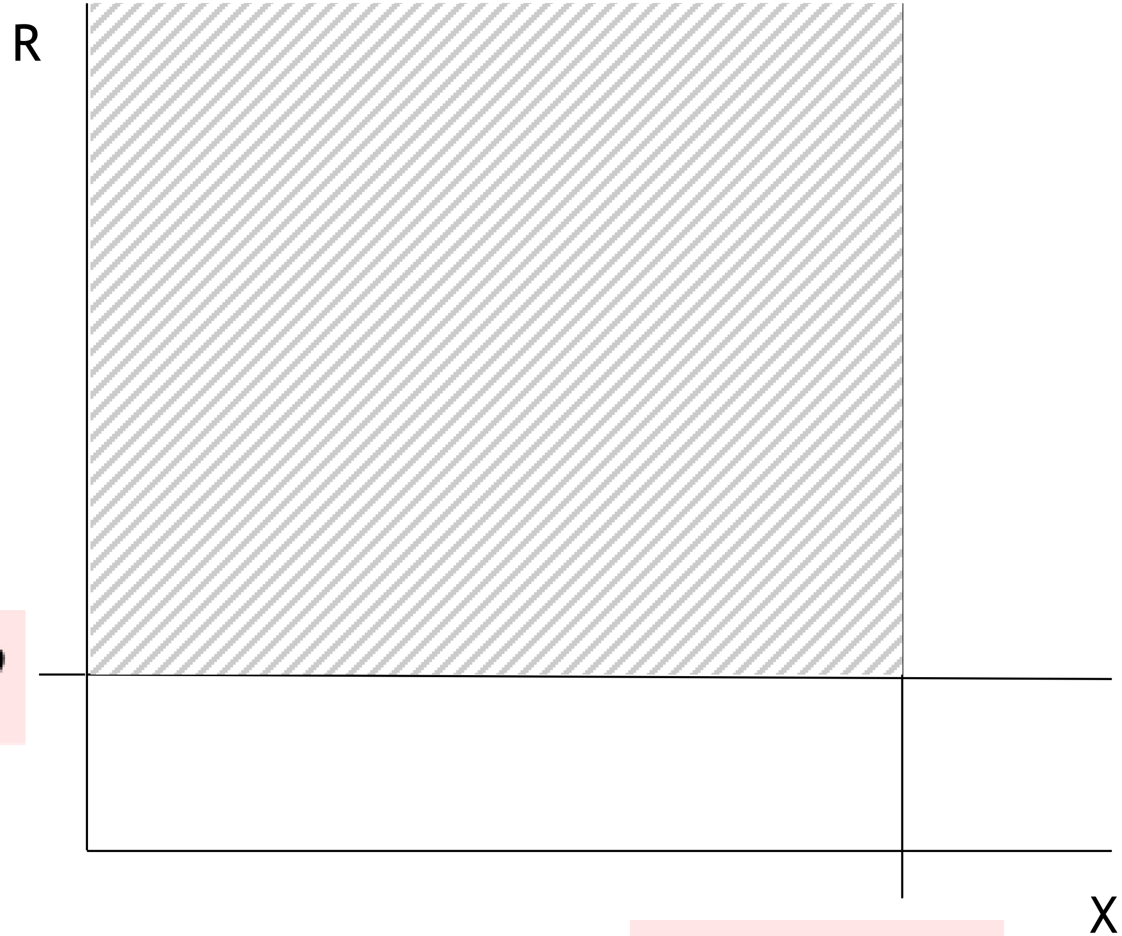
Bound for X(λ)

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

Bound for R(λ)

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

Bounding analysis: Open models



$R(\lambda) \geq D$

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

Bounding Analysis - *Asymptotic bounds*

Closed models:

***X* bounds** considered first, then converted in ***R* bounds** using Little's Law

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

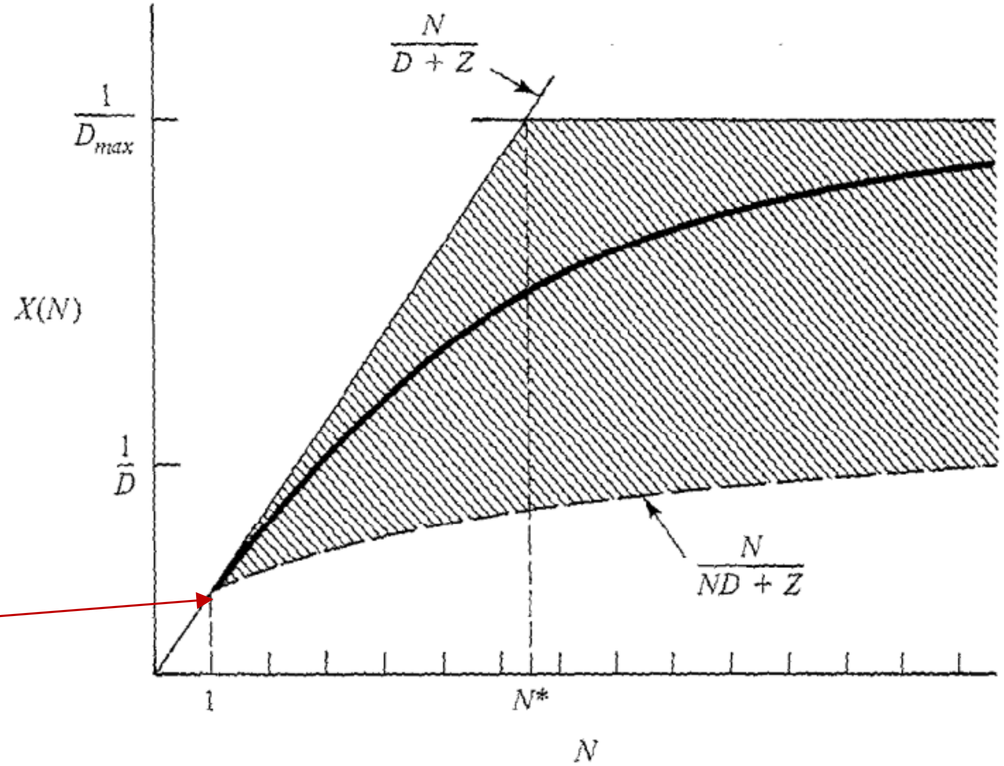
<u>1 customer case:</u>

$$N = X\,(R + Z)$$

$$1 = X\,(D + Z)$$

Then ***X*** is:

$$X = 1 / (D + Z)$$



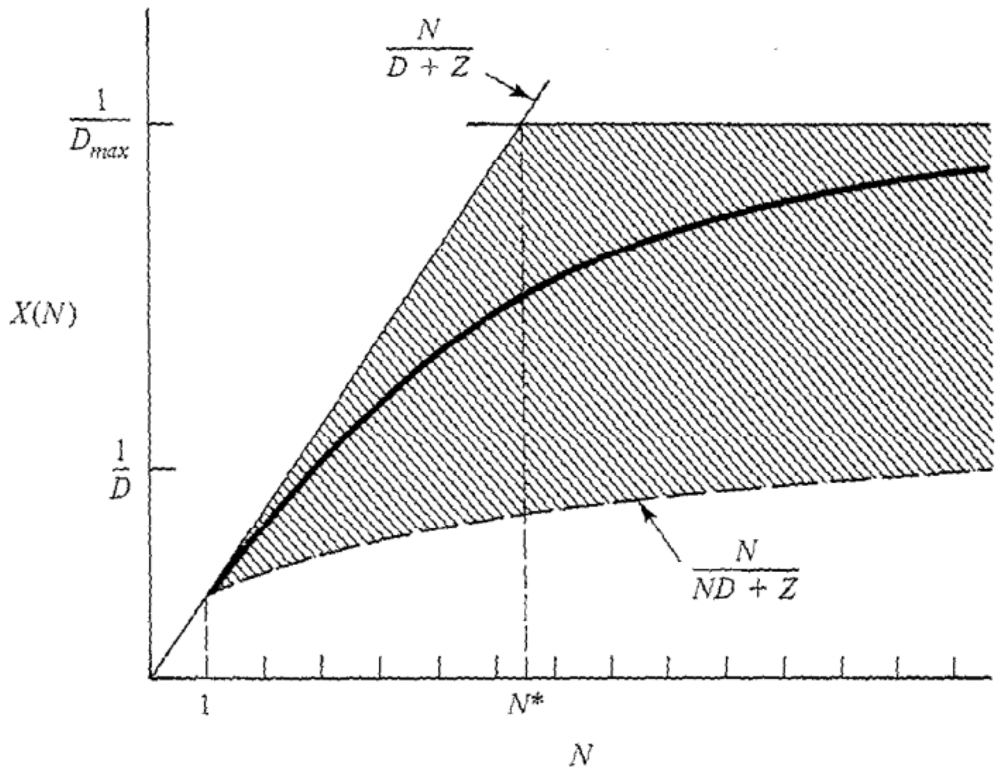
Bounding Analysis - *Asymptotic bounds*

Closed models:

Light Load situation (lower bounds):

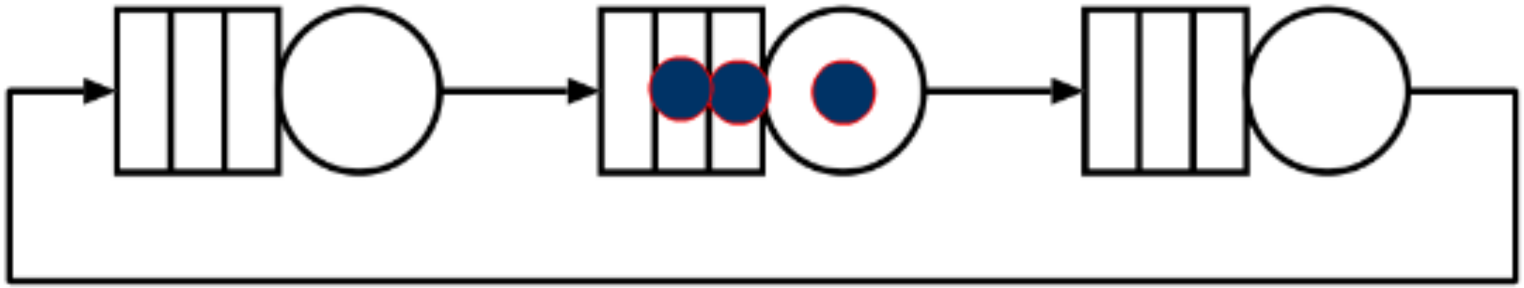
Adding customers:

Smallest *X* obtained with largest R, i.e., new jobs queue behind others already in the system



Bounding Analysis - *Asymptotic bounds*

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 -> **R=ND**



Closed models:

Light Load situation (lower bounds):

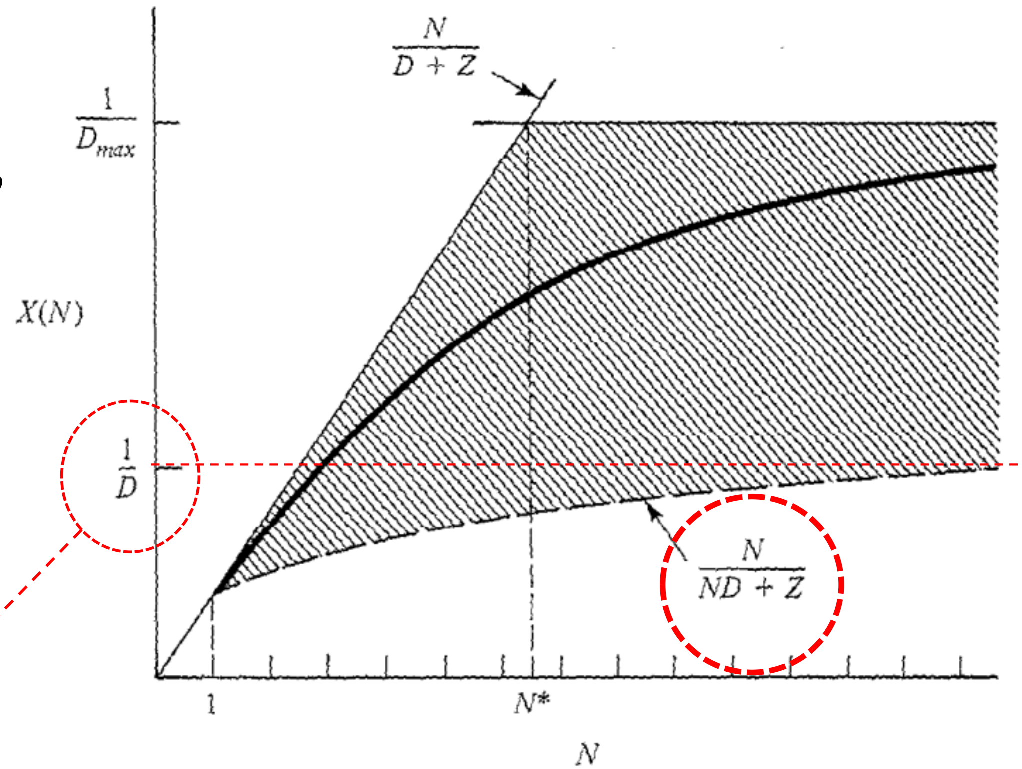
Adding customers:

Smallest *X* obtained with largest R, i.e., new jobs queue behind others already in the system

In this case **R=ND** and **X** is:

$$X = N / (ND + Z)$$

$$\lim_{N \to \infty} N / (ND + Z) = 1 / D$$



Bounding Analysis - *Asymptotic bounds*

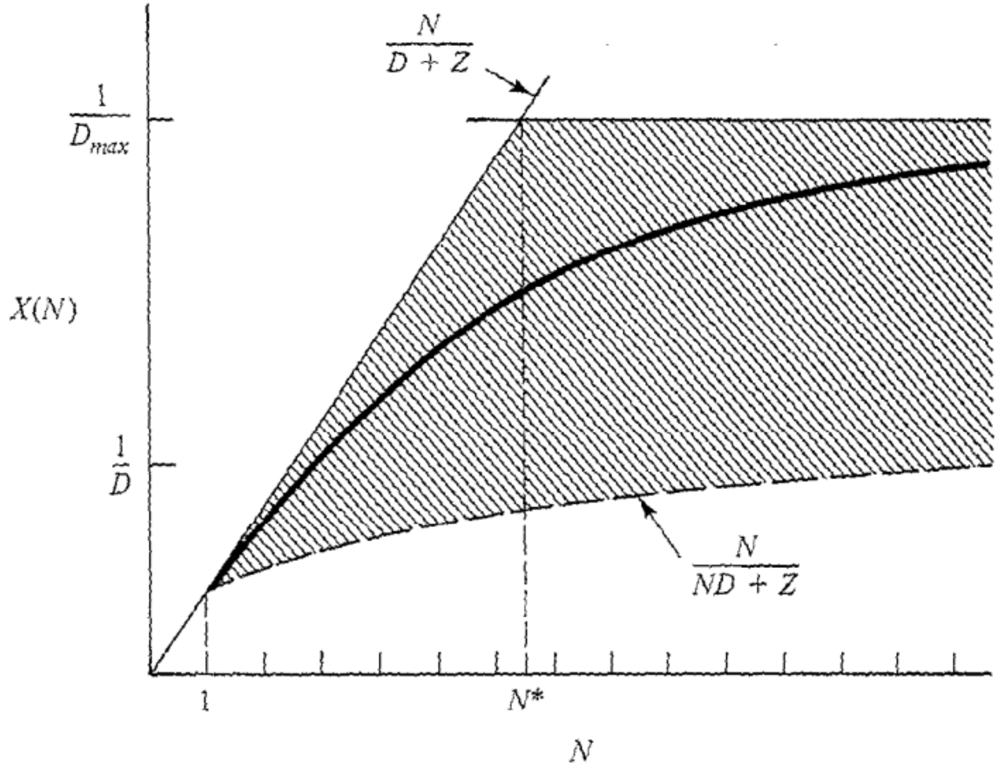
Closed models:

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

Adding customers:

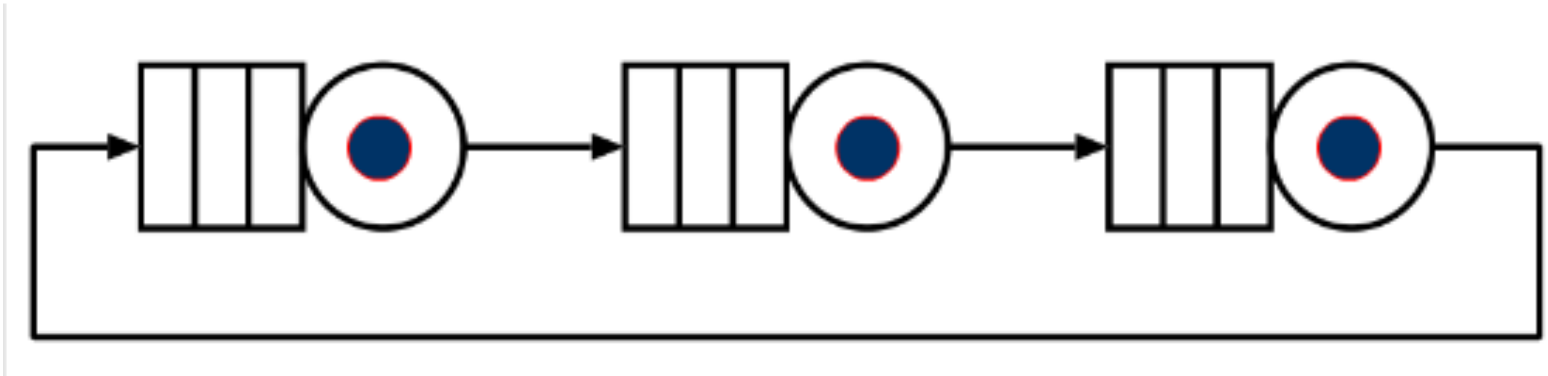
Largest ***X*** obtained with the lowest response time R

i.e No Conflicts



Asymptotic Bounds - Closed Models

The lowest response time can be obtained if a job always finds the queue empty and always starts being served immediately



Bounding Analysis - *Asymptotic bounds*

Closed models:

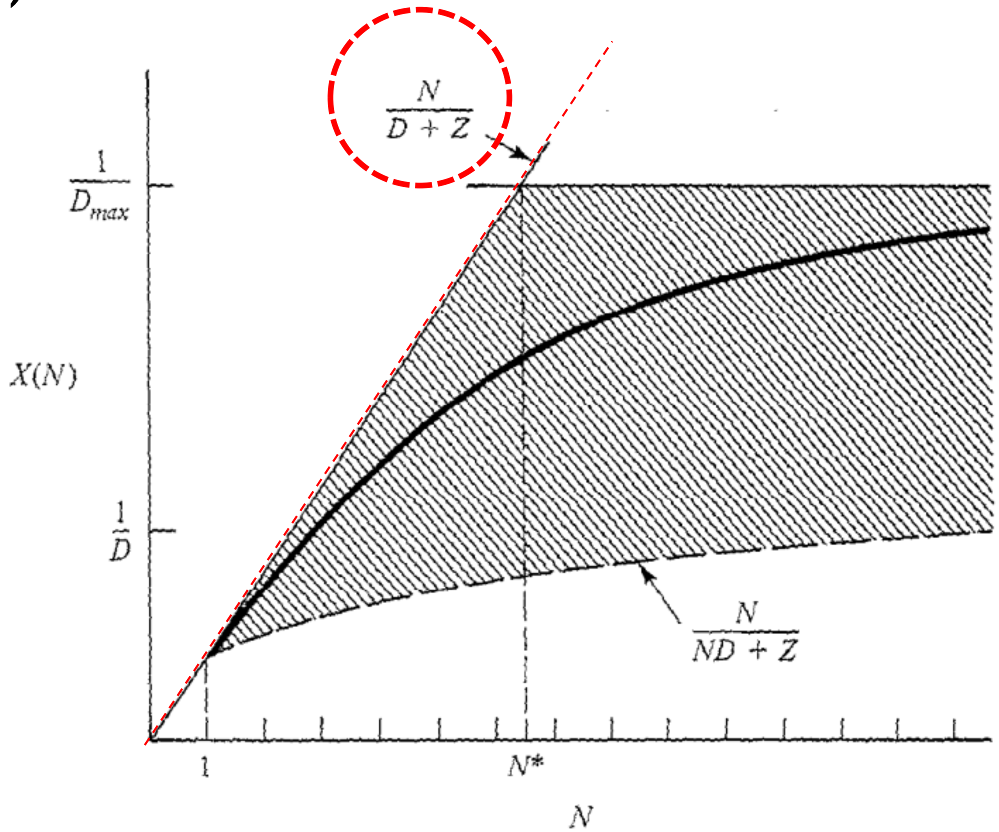
Light Load situation (upper bounds)**:**

Adding customers:

Largest ***X*** if new jobs never queue behind other already in the system:

In this case **R=D** and **X** is:

$$X = N / (D + Z)$$



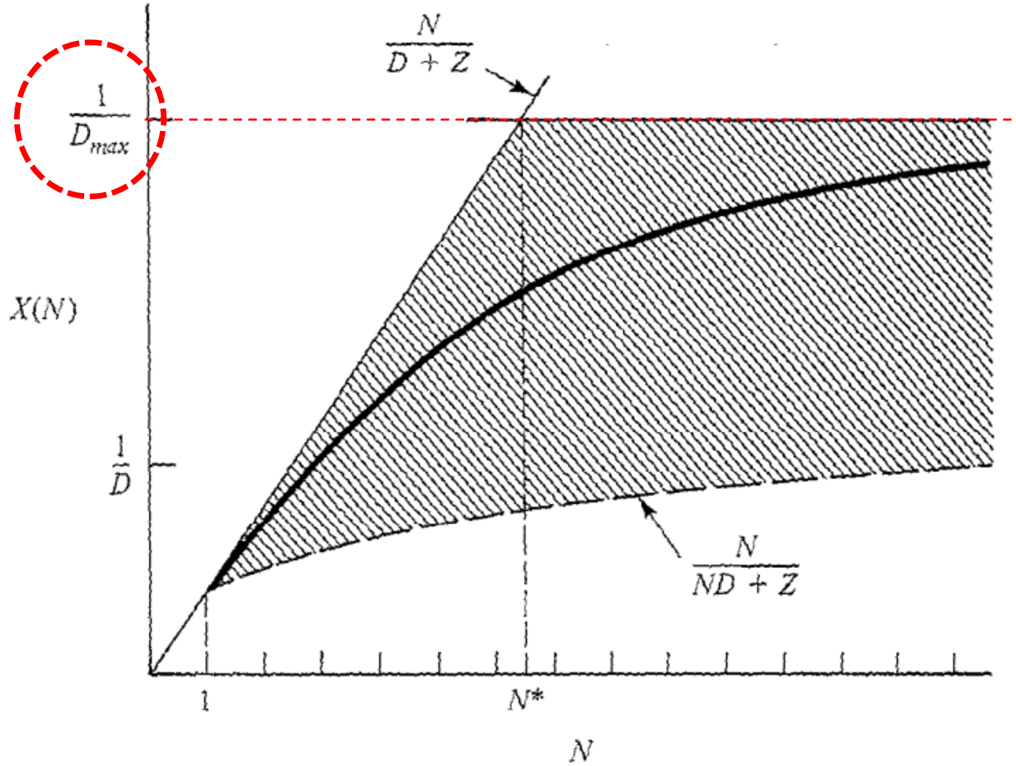
Closed models:

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

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

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

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



Bounding Analysis - *Asymptotic bounds.*

Closed models:

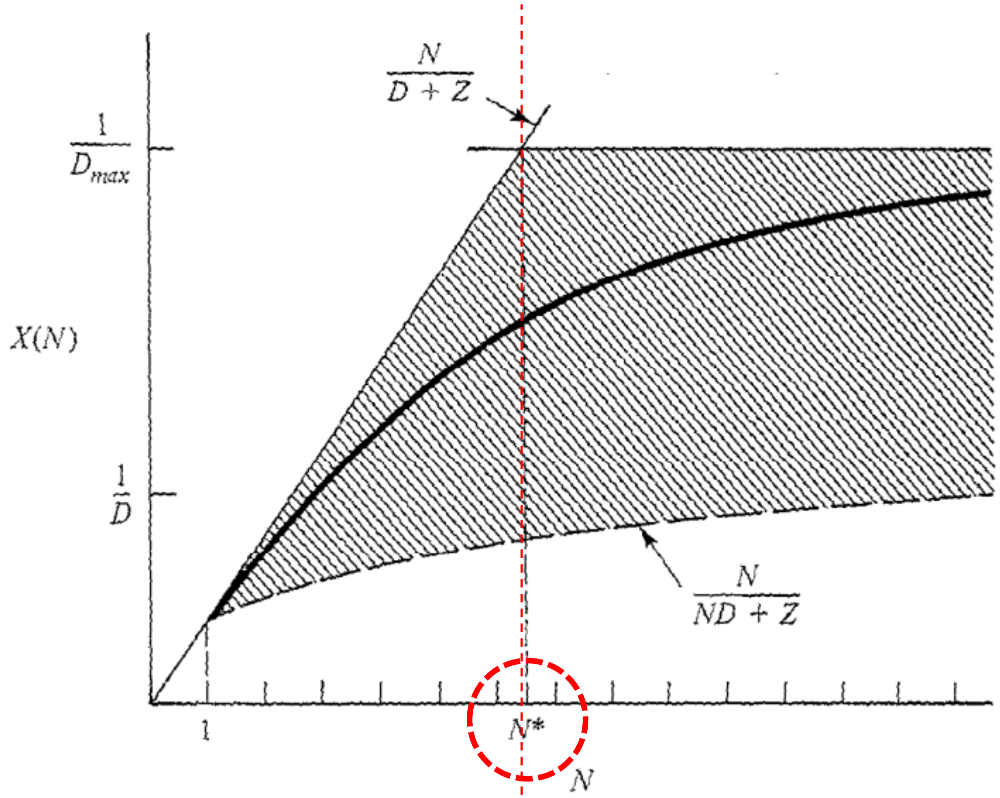
X(N) bounds:

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

N*:

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

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



Bounding Analysis - *Asymptotic bounds.*

R(N) bounds:

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

Let us simply rewrite the previous equation, considering that: X(N)=N/(R(N)+Z), we have:

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

And to have R as numerator we invert the members and we have

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

From which we have

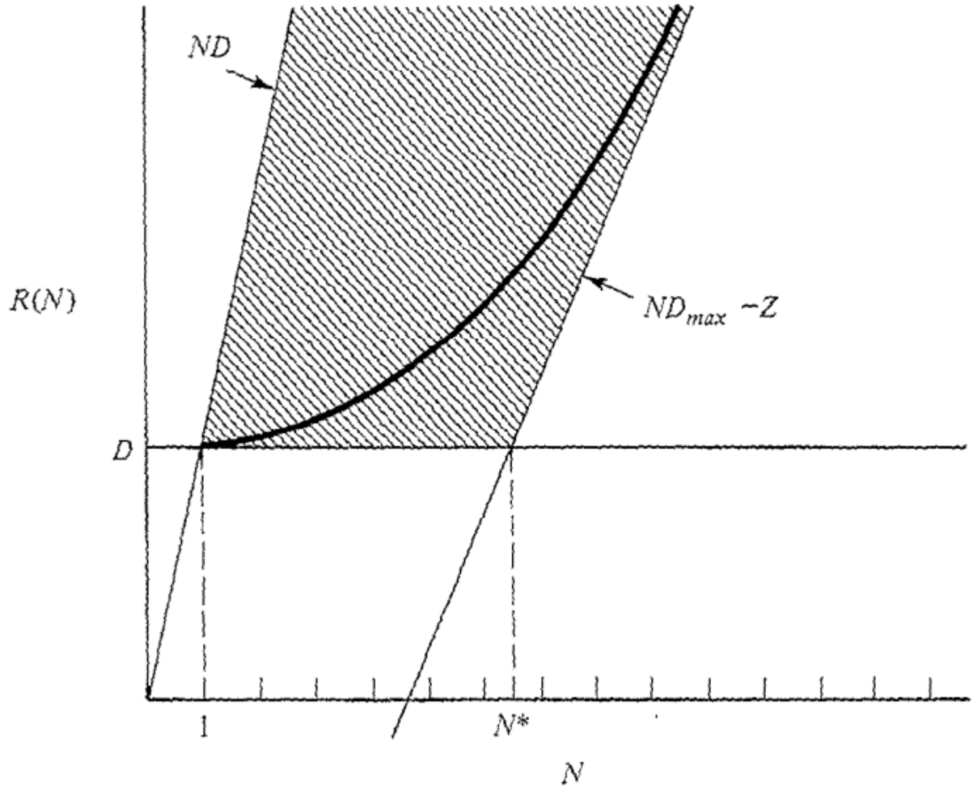
Bound for R(N)

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

Closed models:

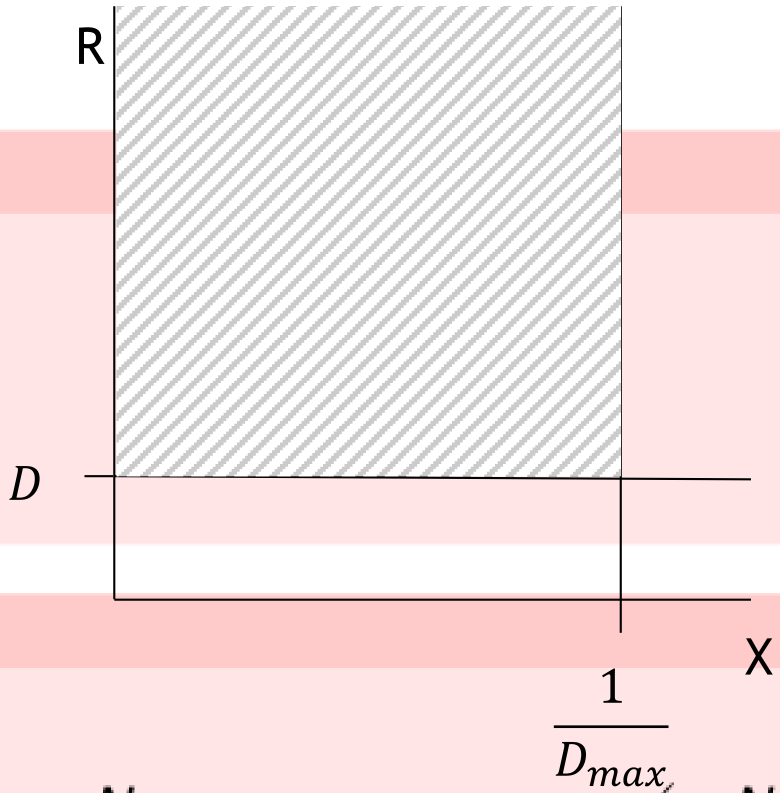
R(N) bounds:

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



Asymptotic bounds summary

Open Models

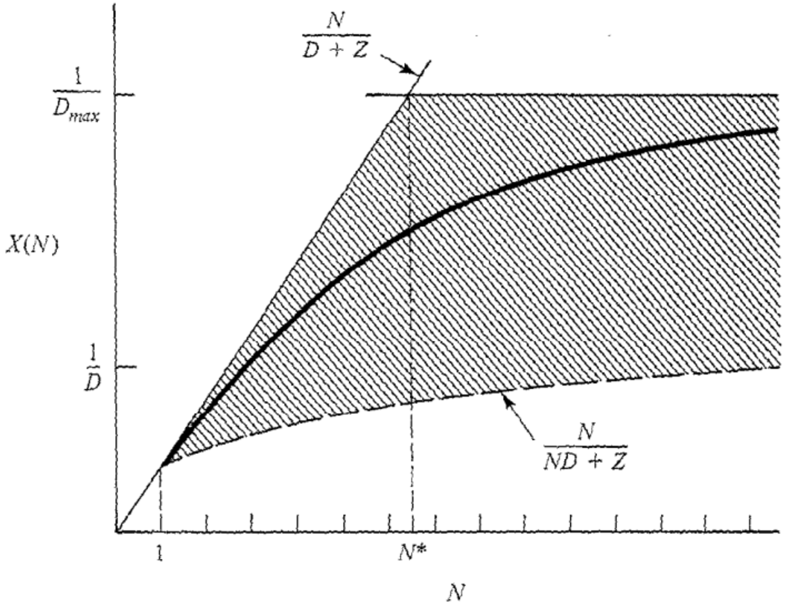


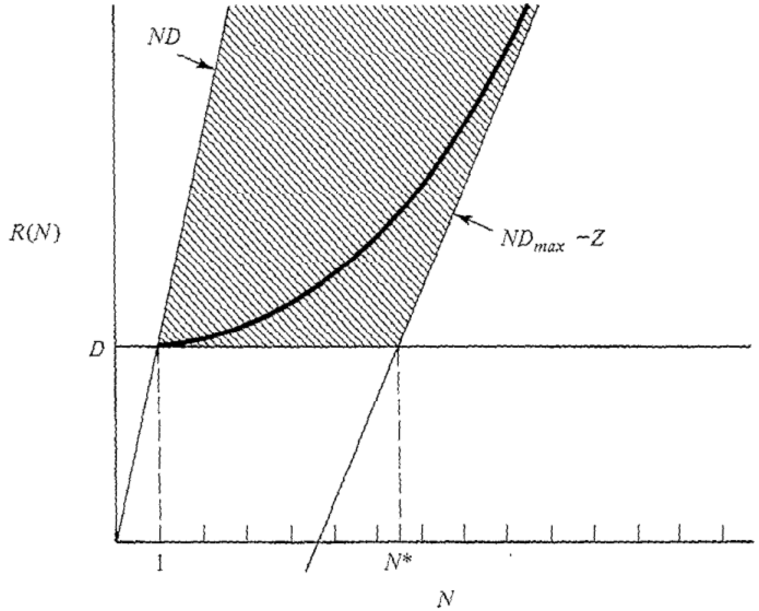
Closed Models

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

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

Asymptotic bounds summary



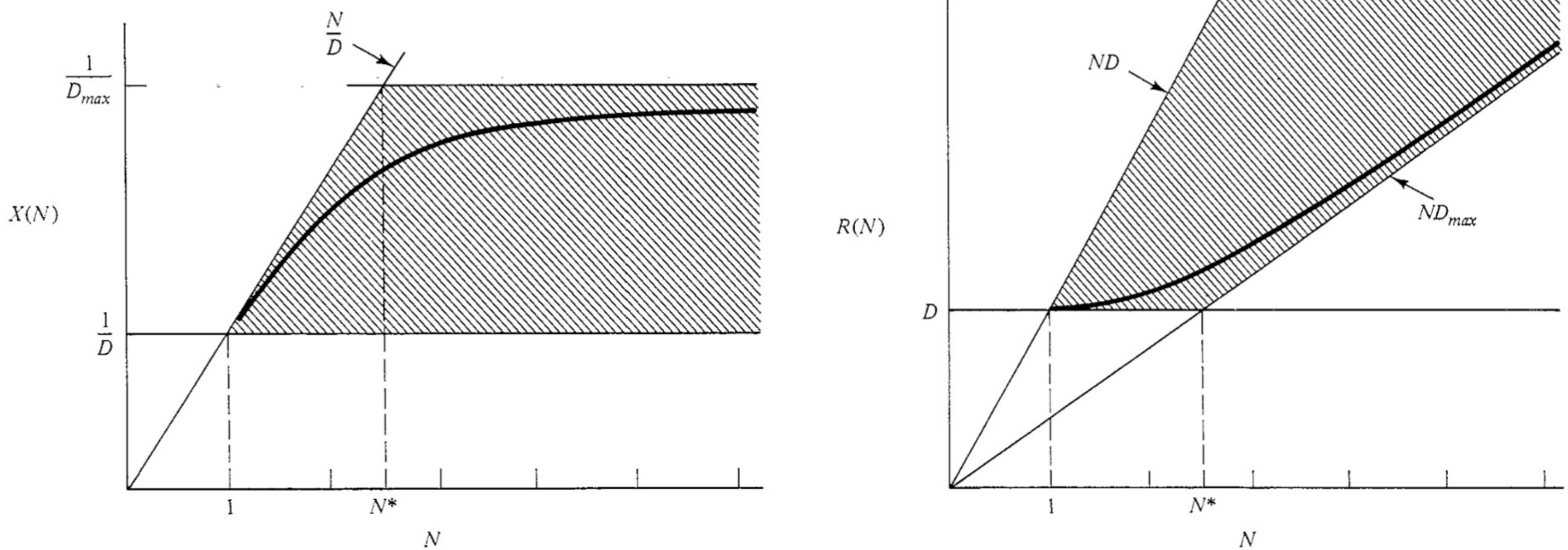


Closed Models

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

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

Asymptotic bounds summary



Closed Models

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

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

NO Think Time