Operation Laws

1. BackGround
1.1. System Model
1.2. Basic Idea
1.3. Global Assumption
2. Terminology
3. Different type of time
4. Utilization Law
5. Little's Law
6. Forced Flow law
7. Bottleneck law
8. General Response Time Law
9. Asymptotic Bounds for Closed Systems

1. BackGround

1.1. System Model

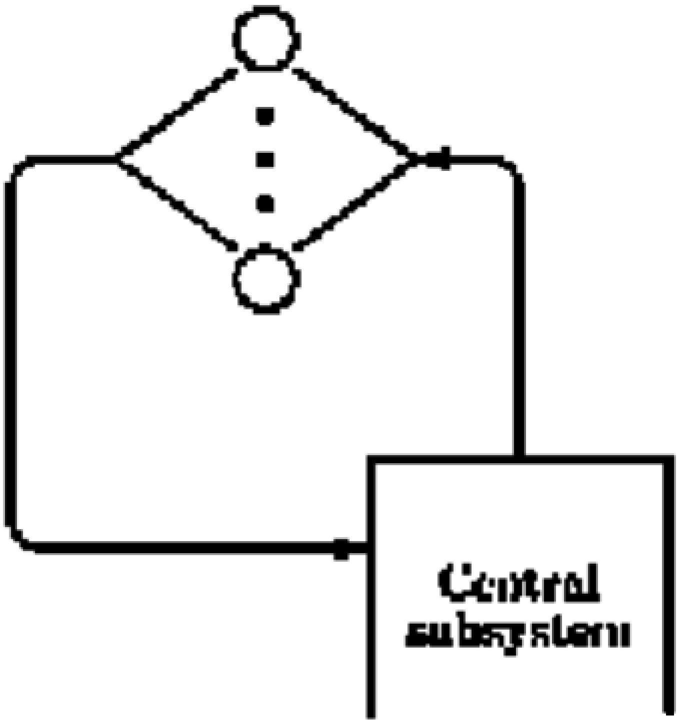
Open System

external arrivals (e.g. web server)

Closed Systems

fixed number of jobs (e.g. time-sharing system)

One classical type of closed systems:



Systems of systems

Multiples resources and connected queues

1.2. Basic Idea

In Queue Theory,

1. We regard the system we evaluate as a **black box**, that means we only know some basic work principles/ measurements/configuration of it, but we do not exactly know the structure
2. We only consider variables/measurements from a **mean** view
3. Operation Laws are relationships that **do not require any assumptions** about the distribution of service times or inter-arrival times

1.3. Global Assumption

We always consider the system has arrived at a stable state, that always means:

1. job flow balance: number of arrivals=number of completions

2. Terminology

What we can observe

A——number of arrivals

B——busy time

C——number of completions

What we can calculate

$$\begin{gathered}\text{arrival rate : } \lambda = A/t \\ \text{throughput : } X = C/t \\ \text{utilization : } U = B/t \\ \text{service time : } S = B/C \\ \text{service rate : } \mu = 1/S\end{gathered}$$

3. Different type of time

Service Time

The time spent on server for each completed task

Think Time

Time between the completion of one request and the start of the next request. That means the time spend when a request leave and re-enter the queue to start a new request.

Response Time

the response time is the sum of the service time and wait time.

4. Utilization Law

$$\begin{gathered}\text{Open System:} \quad U = \lambda S \\ \text{Closed System:} \quad U = XS\end{gathered}$$

5. Little's Law

$$\begin{gathered}\text{Open System :} \quad N = XR \\ \text{Closed System :} \quad N = X(R+Z)\end{gathered}$$

where:

N—— number of jobs in the system

R—— Response time

Z—— Think time

Intuition

Jobs in the system = enter rate $\times$ how long stay in the system/job

e.g.

4000 students enter school, each student 4 years, how many in the school in total: $4000 \times 4 = 16000$

Short Prove

J = hatched area = total time spent in the system by all jobs

$$N = J/t$$
$$R = J/C$$
$$X = A/t = C/t$$

6. Forced Flow law

Calculate the throughput of components from the whole system

$$X_k = V_k X$$

where

$$V_k \text{ : number of visits to device k per job divided}$$
$$V_k = C_k/C_0$$

Short Prove

$$X_k = C_k/t = C_k/C_0 \cdot C_0/t = V_k X$$

7. Bottleneck law

Calculate the utility of components from the whole system

$$U_k = D_k X$$

where:

$$D_k: \text{ total service demand (time) on device k for all visits of a job}$$
$$D_k = V_k S_k = C_k/C_0 \cdot B_k/C_k = B_k/C_0$$

- The device with the **highest utilization** (demand) is the **bottleneck** in the system
- Delay centers can have utilizations more than one without any stability problems. Therefore, **delay centers cannot be a bottleneck device.**

8. General Response Time Law

Assuming there is one terminal per user and the rest of the system is shared by all users

$$R = \sum_{i=1}^{M} R_i V_i$$

Proof:

- For central subsystem, using **Little's Law**: $Q = XR$
  Q: total number of jobs in the system
  R: system response time
  X: system throughput
- $Q = Q_1 + Q_2 + \cdots + Q_M$

  $XR = X_1 R_1 + X_2 R_2 + \cdots + X_M R_M$
- Dividing both sides by X and using **forced flow law**:
  - $R = V_1 R_1 + V_2 R_2 + \cdots + V_M R_M$
  - or write as: $R = \sum_{i=1}^{M} R_i V_i$

9. Asymptotic Bounds for Closed Systems

$$X \leq \min(\frac{1}{D_{max}}, \frac{N}{D+Z})$$
$$R \geq \max(D, D \cdot D_{max} - Z)$$

where

$$D = \sum D_k$$

Simple Prove

We only prove X, R can be induce from Little's Law

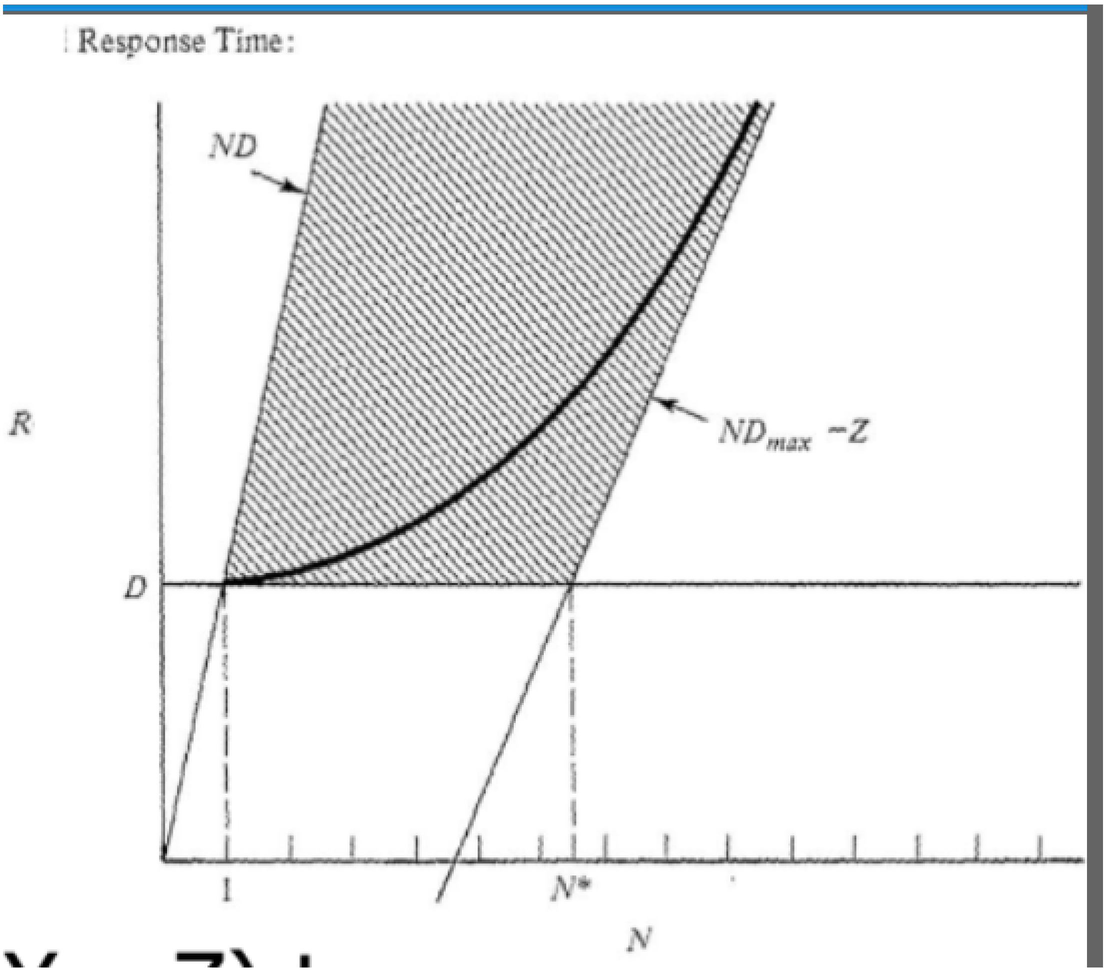
- when loading the system, the slowest device becomes the bottleneck

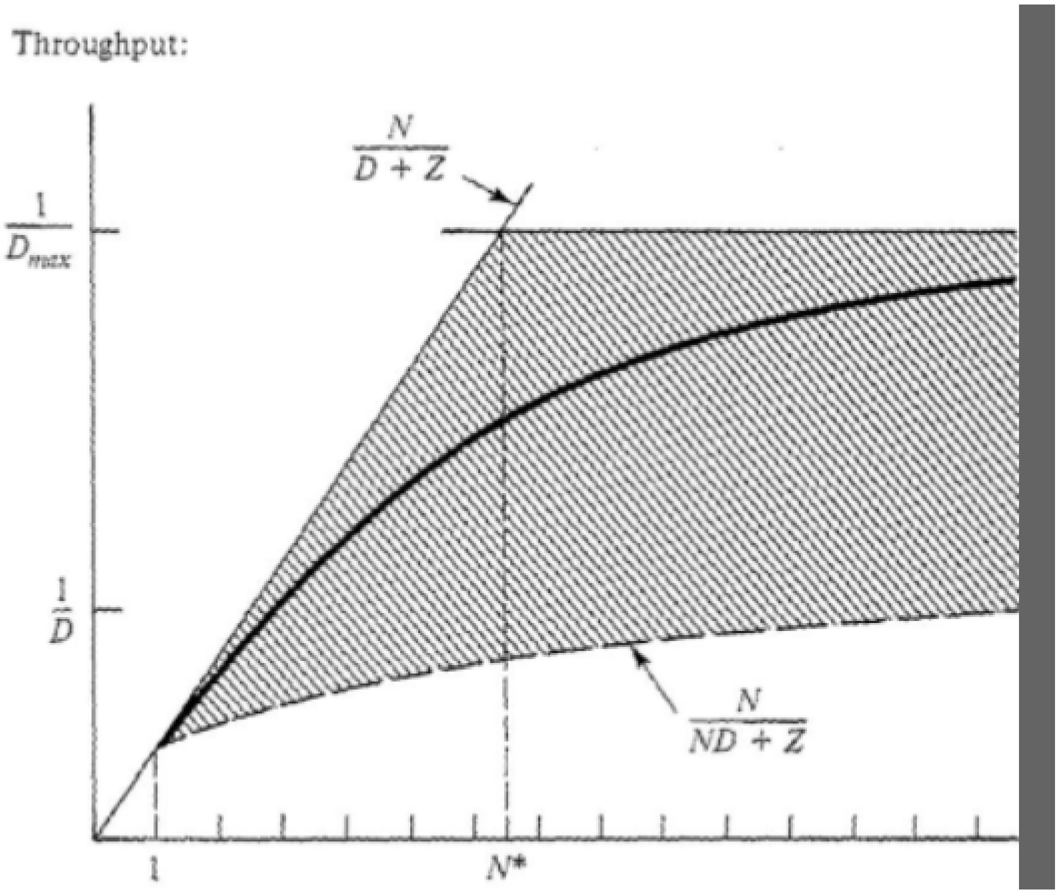
$$X = U_k/D_k \leq 1/D_l \leq 1/D_{max}$$

- max throughput when no queueing occurs (R ⩾ D)

$$X = N/(R+Z) \leq N/(D+Z)$$

Intuition


Response Time:
ND
R
$ND_{max} - Z$
D
1
N*
N



<u>Usage of Asympototic Bounds</u>

We always call the crossing **Knee**, at knee, the number in the system is annoted by N^*

If the number of jobs is more than N^*, then we can say with certainty that there is queueing somewhere in the system