Performance Modeling
of Computer Systems and Networks

Prof. Vittoria de Nitto Personè

Operational analysis:

Queueing Networks

Università degli studi di Roma Tor Vergata
Department of Civil Engineering and Computer Science Engineering

1

Analytical models

Queueing network (QN) modelling is a particular approach to computer system (CS) modeling in which the CS is represented as a *network of queues* which is evaluated *analitically*

- Many analists experienced doubts on its accuracy
- A series of assumptions:

  ✓ the system is modeled by a *stationary stochastic process*;

  ✓ jobs are *stochastically independent*;

  ✓ job steps from device to device follow a *Markov chain*;

  ✓ the system is in stochastic equilibrium;

  ✓ the service time requirements at each device conform to an *exponential distribution*;

  ✓ the system is *ergodic*, i.e. long-term time averages converge to the values computed for stochastic equilibrium

2

Analytical models

Some of these concepts are difficult and cannot be proved to hold by observing the system in a <u>finite time period</u>
Most can be disproved empirically

- ✓ parameters change over time
- ✓ jobs are dependent
- ✓ systems are observable only for short periods
- ✓ …

Prof. Vittoria de Nitto Personè   3

3

Analytical models

In applying or validating the results of Markovian QN theory, analists substituted *operational* values for stochastic parameters

directly measured

The Markovian QN equations are also valid among operational variables

They hold under different assumptions and apply to
a large class of real systems[1]

[1]

BUZEN, J.P. "Operational analysis: the key to the new generation of performance prediction tools," in Proc. IEEE COMPCON, 1976, IEEE, New York.
DENNING, P. J.; AND BUZEN, J. P. "Operational analysis of queueing networks," in Proc. Third Int. Symp. Computer Performance Modeling, Measurement, and Evaluatlon, 1977, North-Holland Publ. Co., Amsterdam, The Netherlands.

Prof. Vittoria de Nitto Personè   4

4

Operational Analysis

Three *operational principles*:

1. All quantities should be *precisely measurable* and all assumptions should be *directly testable*

2. The system must be *flow balanced*

3. The devices must be *homogeneous*,

    i.e., the routing must be independent of queue lengths (ql)
    the mean service time at a given device must not depend on ql of other devices

the same mathematical equations

but the operational assumptions can be tested

much more confidence and understanding of the QN technology

5

Operational Analysis

Def.

Hypotheses whose veracity can be established beyond doubt by measurement will be called *operationally testable*

Operational analysis provides a rigorous mathematical discipline for studying CS performance based solely on operationally testable hypotheses

two basic components:

    a system (real or hypothetical)
    a (finite) time period

        the observation period

6

Operational Analysis

According to the *operational* approach, let us consider

Basic quantities

T	the length of the observation period (op)
A	the number of arrivals during op
B	the total amount of time during which the system is busy $B \le T$ during op
C	the number of completions during op

Prof. Vittoria de Nitto Personè   7
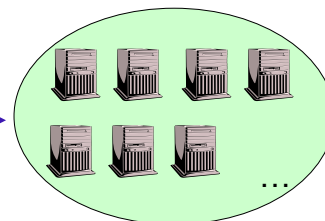
7

Multi Resources Systems   **1**

- a collection of servers, working together for incoming requests
- homogeneous servers

requests

Data centers
Server farm

What is the minimum number of servers needed to guarantee that only a small fraction of jobs are delayed?

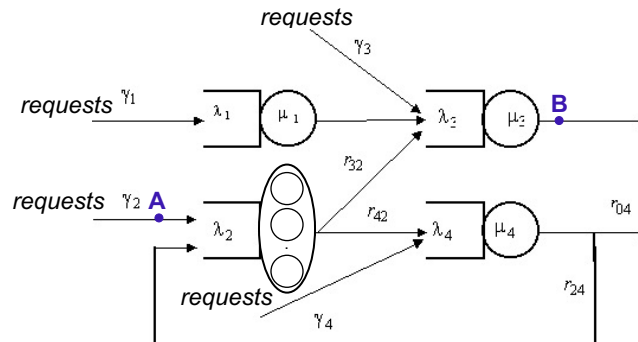*Is a single central queue superior to having a queue at each server?*

Prof. Vittoria de Nitto Personè   8

8

Multi Resources Systems **2**

- a collection of different resources, connected, where incoming requests can require services more than once to each resources

requests γ_1

requests γ_3

requests γ_2 **A**

requests γ_4

λ_1 μ_1

λ_2 μ_2 **B**

λ_2

λ_4 μ_4

r_{32}

r_{42}

r_{04}

r_{24}

Which is the "load" of each resource?

Which is the time spent on each resource?

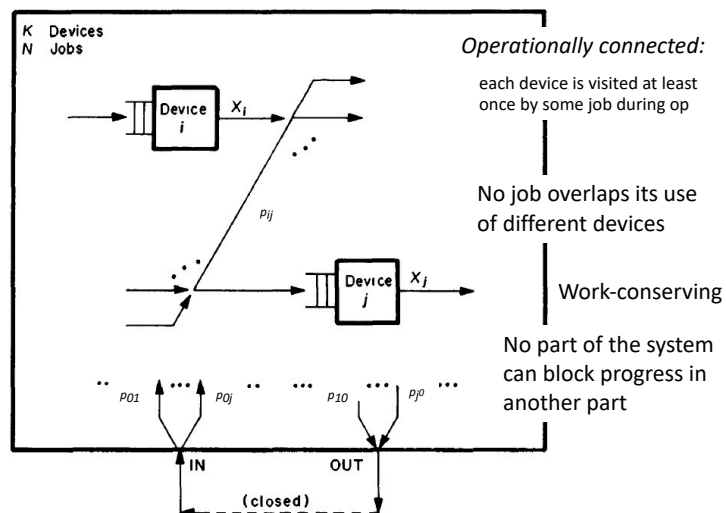*Which is the time to go from point A to point B?*
(end-to-end response time)

Prof. Vittoria de Nitto Personè   9

9

Queueing Networks
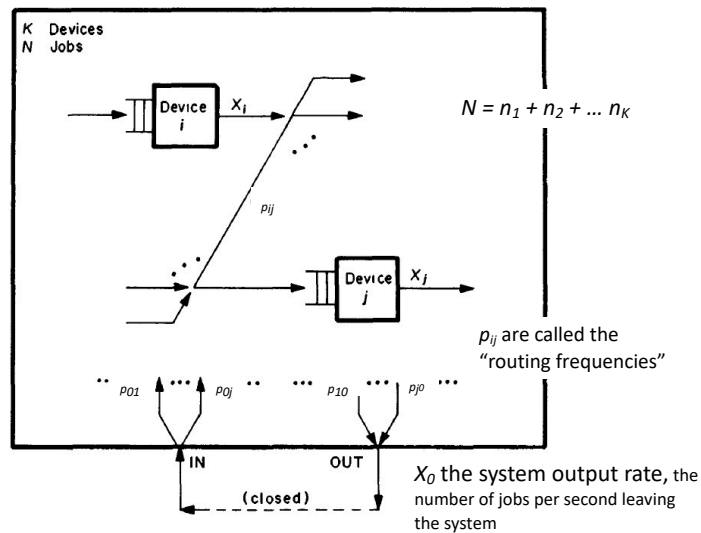

K Devices
N Jobs

Device i x_i

p_{ij}

Device j x_j

p_{01} p_{0j} p_{10} p_{j0}

IN   OUT

(closed)

Operationally connected:

each device is visited at least once by some job during op

No job overlaps its use of different devices

Work-conserving

No part of the system can block progress in another part

Prof. Vittoria de Nitto Personè   10

10

Queueing Networks

$N = n_1 + n_2 + \dots n_K$

p_{ij} are called the "routing frequencies"

X_0 the system output rate, the number of jobs per second leaving the system
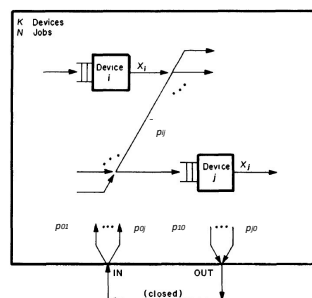
Prof. Vittoria de Nitto Personè

11

11

Queueing Networks

If the system is *open*, X_0 is known and N varies as jobs enter or leave the system

An analysis of an open system assumes that X_0 is known and seeks to characterize the distribution of N

If the system is *closed* the number of jobs N is fixed

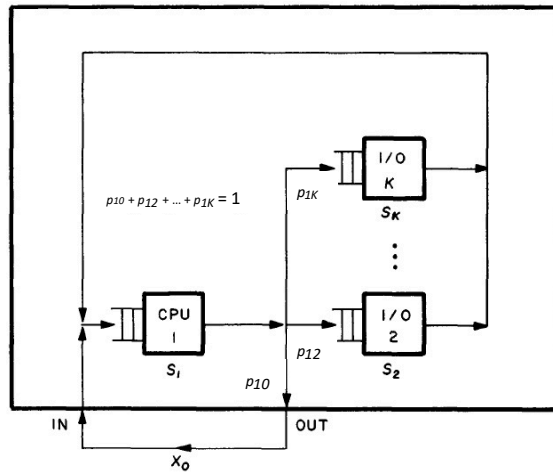An analysis of a closed system begins with N given and seeks to determine the resulting X_0 along the OUT/IN path

Queue lengths and response times at the devices may be sought in both cases

Prof. Vittoria de Nitto Personè

12

12

Central Server Network

$p_{10} + p_{12} + \dots + p_{1K} = 1$

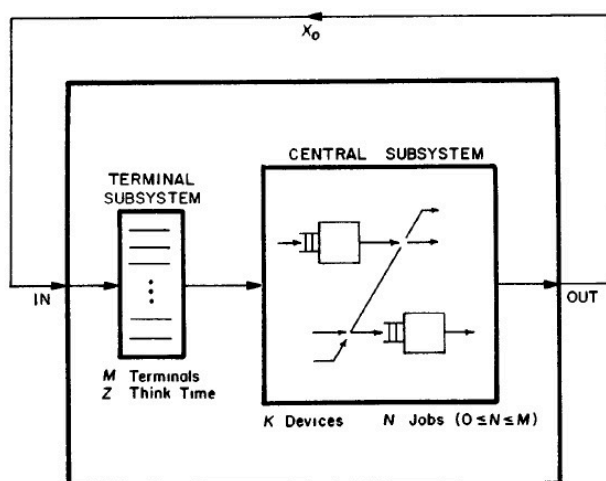**A batch processing system under a backlog:**

- a job begins with a CPU service interval and continues with zero or more I/O intervals which alternate with further CPU bursts
- a new job enters the system as soon as an active job terminates

Prof. Vittoria de Nitto Personè   13

13

Terminal Driven System

Time sharing systems, driven by interactive terminals:

a user alternates between *thinking* and *waiting*

the mean time a user spends in a thinking interval: *Z think time*
The mean time a user spends in a waiting interval: *R response time*

- *Z* is independent of *M*
- *R* is a function of *M*:

    jobs delay each other while contending for resources

Prof. Vittoria de Nitto Personè   14

14

7

Queueing Networks

The system is measured for an op of T seconds, the following data are collected for each device $i=1, 2, ...K$

A_i the number of arrivals;

B_i total busy time, during which $n_i > 0$

C_{ij} number of times a job requests service at device j immediately after completing a service request at device i;
note that it is possible $C_{ii} > 0$.

If we treat the "outside world" as device "0", we can define also

A_{0j} number of jobs whose first service request is for device j;

C_{i0} number of jobs whose last service request is for device i.

Prof. Vittoria de Nitto Personè   15

15

Queueing Networks

Prof. Vittoria de Nitto Personè   16
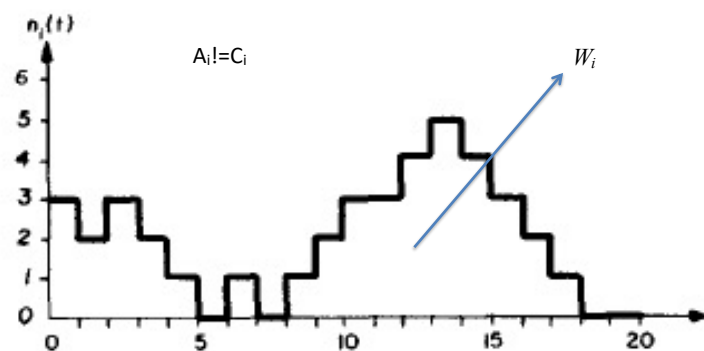
16

Operational Equations

$$X_j = \sum_{i=0}^{K} X_i p_{ij}$$ Job Flow Balance equations

$$\begin{cases} V_0 = 1 \\ V_j = p_{0j} + \sum_{i=1}^{K} V_i p_{ij} \end{cases}$$ Visit Ratio equations

Prof. Vittoria de Nitto Personè   17

17

Operational Equations

$$U_i = X_i S_i$$ *Utilization Law*

$$\bar{n}_i = X_i R_i$$ *Little's Law*

$$X_0 = \sum_{i=0}^{K} X_i p_{i0}$$ *Output Flow Law*

$$R = \sum_{i=1}^{K} V_i R_i$$ *General Response Time Law*

$$R = M / X_0 - Z$$ *Interactive Response Time Formula*
(Assumes flow balance)

Prof. Vittoria de Nitto Personè   18

18

9

IN

OUT