Performance Modeling of Computer Systems and Networks

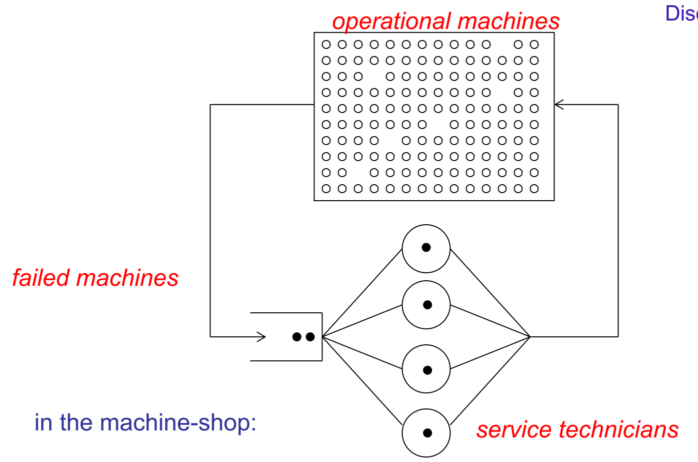
Prof. Vittoria de Nitto Personè

Trace-driven simulation
Case study 1

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

1

Discrete-event simulation
conceptual model

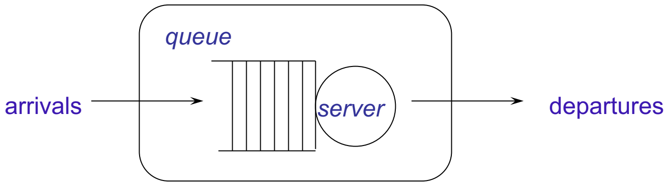


in the machine-shop:

- The "jobs" are the failed machines and the "servers" are the technies
- If there is only one service technician, the conceptual model includes a single-server queue

2

Single server queue



Service node

• *def. 1* a *single server service node* consists of a *server* plus its *queue*

terminology

synonymous

- queue/center/node
- job/user/request

usual		in the book
• waiting time	→	delay
• response/sojourn time	→	wait

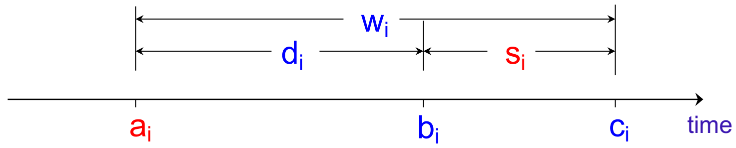
For a job *i*:

- The *arrival time* is a_i
- The *service time* is s_i

input variables

- The *delay* in the queue is d_i (delay, usually this is known as "waiting time")
- The time that service begins is $b_i = a_i + d_i$ (begin)
- The *wait* in the node is $w_i = d_i + s_i$ (wait, aka response time)
- The departure time is $c_i = a_i + w_i$ (completion)

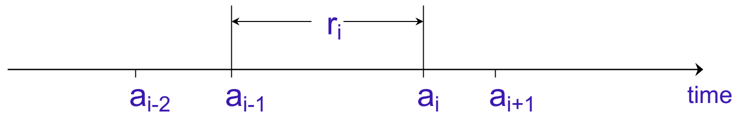
output variables



The *interarrival time* between jobs *i − 1* and *i* is

$$r_i = a_i - a_{i-1}$$

where, by definition, $a_0 = 0$



Assume only one arrival per time instant

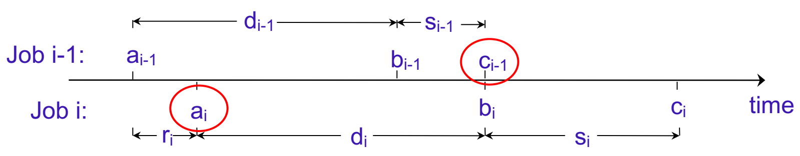
$r_i > 0, \forall i$

NO *bulk*

trace-driven simulation

- The model is driven by external data:
  Given the arrival times a_i and service times s_i, can the delay times d_i be computed?
- For some queue disciplines, this question is difficult to answer
- If the queue discipline is FIFO, d_i is determined by when a_i (the arrival) occurs relative to c_{i-1} (the previous departure)

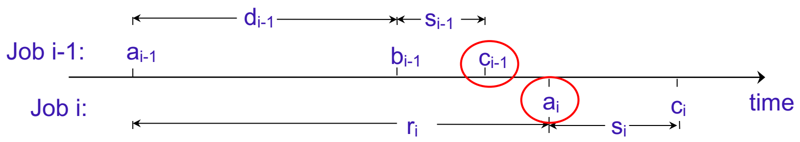
Case 1. The job arrives *before* the previous job completes
$a_i < c_{i-1}$



$\longrightarrow d_i=c_{i-1}-a_i$

Case 2. The job arrives *after* the completion of the previous job

$$a_i \geq c_{i-1}$$



$\longrightarrow$ $d_i = 0$

Output statistics

- The purpose of simulation is insight — gained by looking at statistics
- The importance of various statistics varies on perspective:
  - User perspective (job): wait time is most important
  - Manager perspective: utilization is critical
- Statistics are broken down into two categories
  - Job-averaged
  - Time-averaged

Job-averaged statistics

- *Average interarrival time* $\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i = \frac{a_n}{n}$

  Arrival rate $\frac{1}{\bar{r}}$

- *Average service time* $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$

  Service rate $\frac{1}{\bar{s}}$

Job-averaged statistics

- The *average delay* and *average wait* are defined as

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad \bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$$

Recall $w_i = d_i + s_i \ \forall$ i, hence

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n} (d_i + s_i) = \frac{1}{n}\sum_{i=1}^{n} d_i + \frac{1}{n}\sum_{i=1}^{n} s_i = \bar{d} + \bar{s}$$

Sufficient to compute any two of $\bar{w}, \bar{d}, \bar{s}$

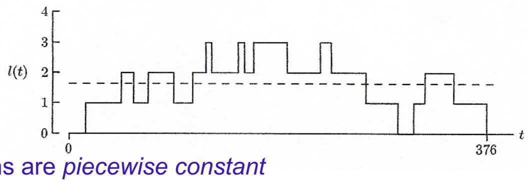
time-averaged statistics

For SSQ, need three additional functions

- l(t): number of jobs in the service node at time t
- q(t): number of jobs in the queue at time t
- x(t): number of jobs in service at time t

By definition l(t)=q(t)+x(t) $\forall$ t

l(t) = 0, 1, 2, …
q(t) = 0, 1, 2, …
x(t) = 0, 1



The three functions are *piecewise constant*

Discrete-event simulation
Output statistics

Over the time interval (0, τ):

time-averaged number in the node: $\bar{l} = \frac{1}{\tau}\int_0^\tau l(t)dt$

time-averaged number in the queue: $\bar{q} = \frac{1}{\tau}\int_0^\tau q(t)dt$

time-averaged number in service: $\bar{x} = \frac{1}{\tau}\int_0^\tau x(t)dt$

Def. *Utilization*
The proportion of time that the server is busy

Since l(t)=q(t)+x(t) $\forall$ t $\quad \bar{l} = \bar{q} + \bar{x}$

How are job-averaged and time-average statistics related?

Little's Law (1961)

If (a) queue discipline is FIFO,

(b) service node capacity is infinite, and

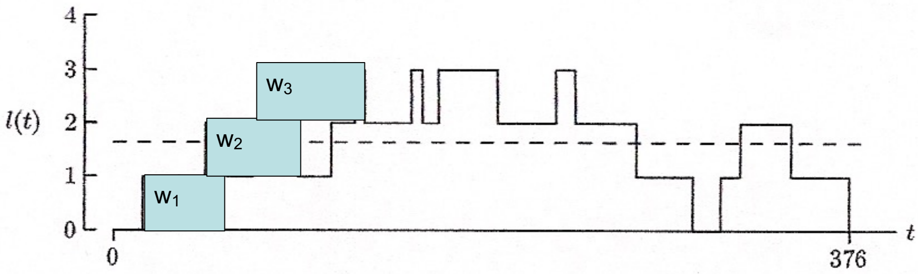
(c) server is idle both at the beginning and end of the observation interval ($t = 0$, $t = c_n$)

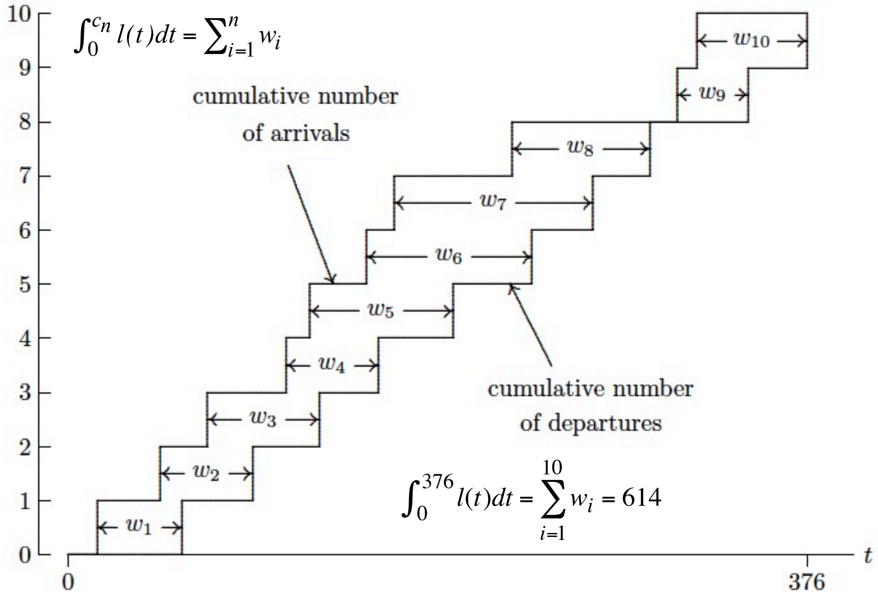
then

$$\int_0^{c_n} l(t)dt = \sum_{i=1}^{n} w_i$$

$$\int_0^{c_n} q(t)dt = \sum_{i=1}^{n} d_i$$

$$\int_0^{c_n} x(t)dt = \sum_{i=1}^{n} s_i$$





Using $\tau = c_n$ in $\bar{l} = \frac{1}{\tau}\int_0^{\tau} l(t)dt$

along with Little's Theorem, we have:

$$c_n \bar{l} = \int_0^{c_n} l(t)dt = \sum_{i=1}^{n} w_i = n\bar{w}$$

As a consequence: $\bar{l} = \frac{n}{c_n}\bar{w}$

Same holds for: $\bar{q} = \frac{n}{c_n}\bar{d}$ $\quad \bar{x} = \frac{n}{c_n}\bar{s}$

$\frac{n}{c_n}$ represents the *average system throughput* in c_n
Note that, for infinite queue, this corresponds to the average arrival rate

Def. Traffic intensity

The ratio of the arrival rate to the service rate

$$\frac{1/\bar{r}}{1/\bar{s}} = \frac{\bar{s}}{\bar{r}} = \frac{\bar{s}}{a_n/n} = \left(\frac{c_n}{a_n}\right)\bar{x}$$

$$\bar{x} = \frac{n}{c_n}\bar{s}$$

When c_n/a_n is close to 1.0, the traffic intensity and utilization will be nearly equal