

Queueing theory

- Little's Law
- Average delay of $M/G/1$ queue with FCFS (FIFO) discipline
 - ❖ motivation for packet switching
- Residual life of a random variable
- Conservation Law ($M/G/1$)

Little's Law

No assumptions – applicable to *any system* whose arrivals and departures are observable

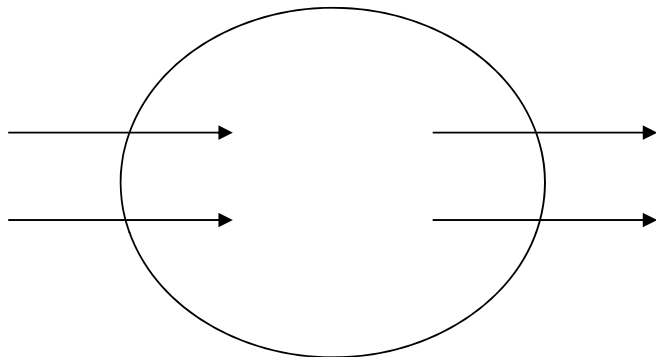
Average population
= (average delay) ×
(throughput rate)

$$\text{average delay} = \frac{1}{N} \sum_{i=1}^N \text{delay}_i$$

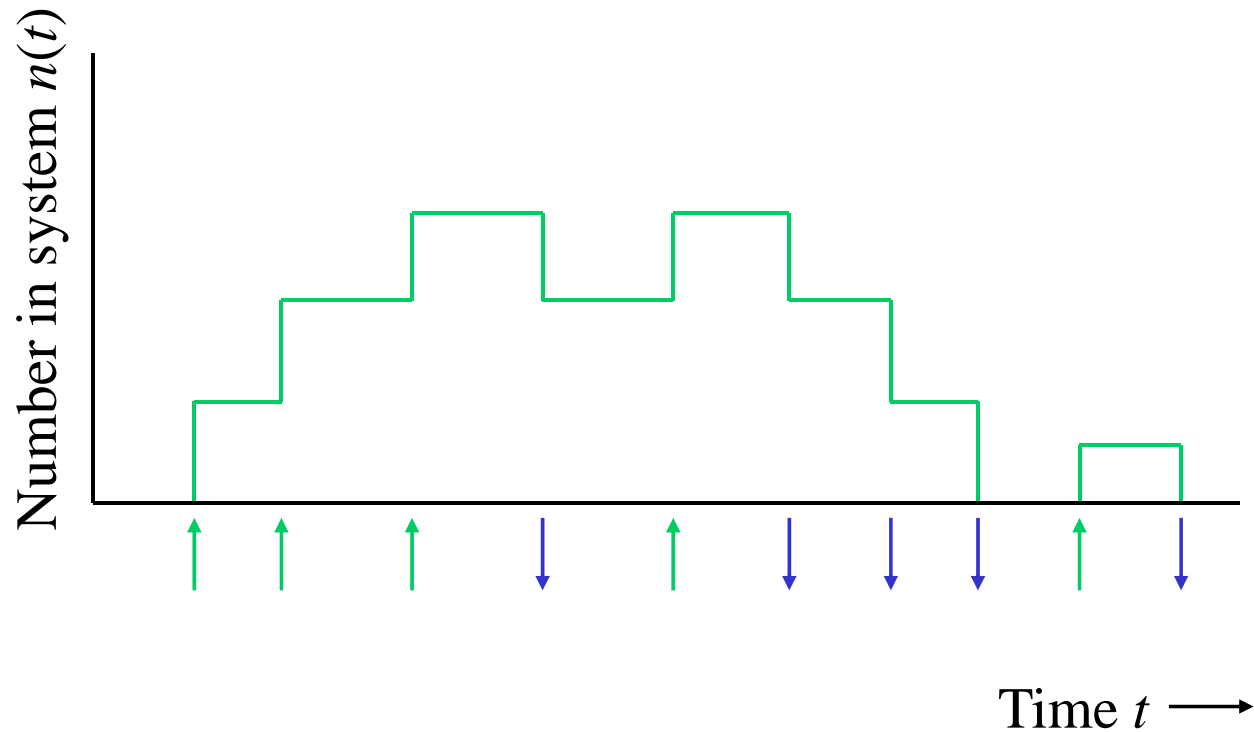
where N is number of departures

$$\text{throughput rate} = N/\tau$$

where τ is duration of experiment



average population
(to be defined)



$$\text{average population} = \frac{1}{\tau} \int_0^{\tau} n(t) dt$$

where τ is duration of the experiment

Exercise - check Little's Law for the following example

Consider 6 jobs that have gone through a system during the time interval $[0, 15]$, where time is in seconds, as shown in the table:

| Job | Arrival Time | Departure Time |
|-----|--------------|----------------|
| 1 | 0.5 | 4.5 |
| 2 | 1.5 | 3.0 |
| 3 | 6.0 | 11.0 |
| 4 | 7.0 | 14.0 |
| 5 | 8.5 | 10.0 |
| 6 | 12.0 | 13.0 |

For the time duration $[0, 15]$:

- calculate throughput rate;
- plot number of jobs in the system as a function of time from 0 to 15 and calculate the average number over the duration $[0, 15]$;
- calculate the average delay of the 6 jobs.

Verify that Little's Law is satisfied by the results in (a), (b), and (c).

Random variable with discrete values

Random variable X with discrete values x_1, x_2, \dots, x_m

Let $P_i = \text{probability } [X = x_i]$ for $i = 1, 2, \dots, m$

Its expected value (mean) is $\bar{X} = \sum_{i=1}^m x_i P_i$

Its second moment is $\overline{X^2} = \sum_{i=1}^m x_i^2 P_i \geq (\bar{X})^2$

Random Variable with continuous values

Random variable X with probability distribution function (PDF), $F_X(x) = P[X \leq x]$, $x \geq 0$

Its probability density function (pdf) is

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Its expected value is $\bar{X} = \int_0^{\infty} x f_X(x) dx$

Its second moment is

$$\overline{X^2} = \int_0^{\infty} x^2 f_X(x) dx \geq (\bar{X})^2$$

Poisson arrival process at rate λ

- It is a random process with *independent* increments, that is:

Let $X(s, s+t)$ be the number of arrivals in the time interval $(s, s+t)$. For any time s ,

$$P[X(s, s+t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad k \geq 0, t \geq 0$$

Inter-arrival time of a Poisson process has the exponential distribution

□ Consider the random variable \tilde{t} which is the time between consecutive arrivals

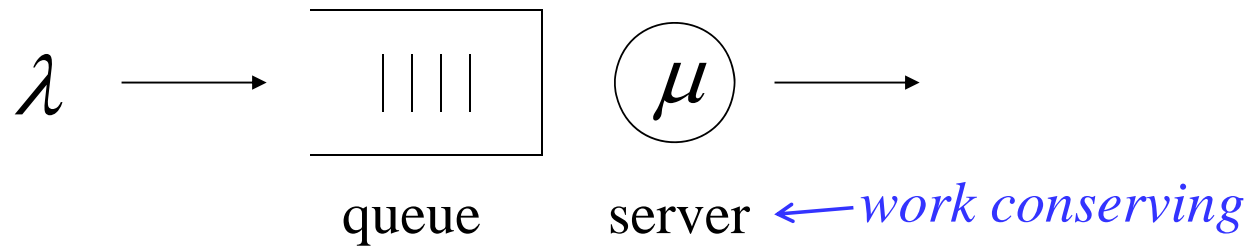
□ Probability distribution function of \tilde{t} is

$$\begin{aligned} A(t) &= P[\tilde{t} \leq t] = 1 - P[\tilde{t} > t] \\ &= 1 - P[X(s, s+t) = 0] = 1 - e^{-\lambda t}, \quad t \geq 0 \end{aligned}$$

□ Probability density function of \tilde{t} is

$$a(t) = \frac{dA(t)}{dt} = \lambda e^{-\lambda t} \quad t \geq 0$$

Single-Server Queue



\bar{x} average service time, in seconds

μ service rate, in jobs/second ($\mu = 1/\bar{x}$)

λ arrival rate, in jobs/second

ρ utilization of server

Conservation of flow

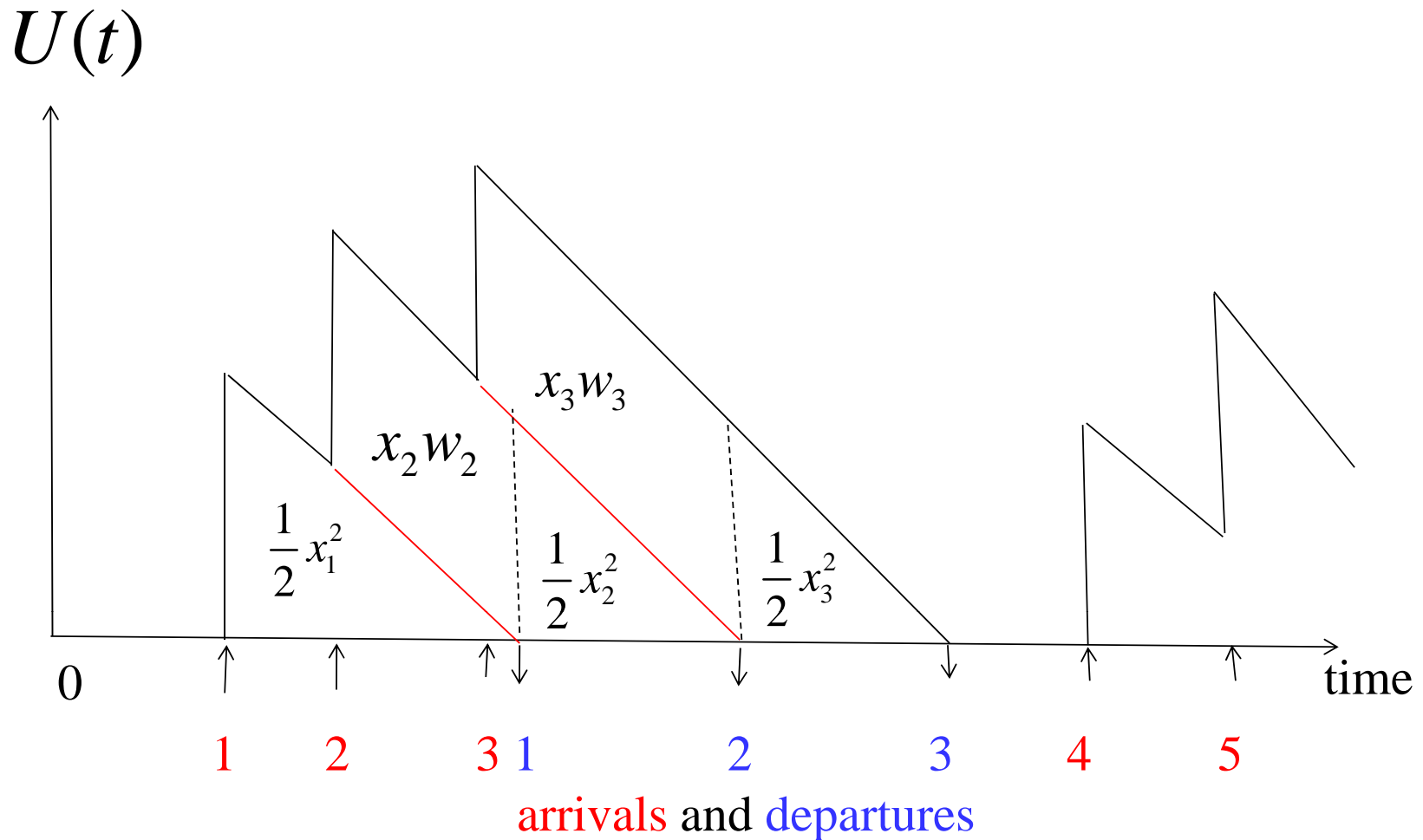
$$\lambda = \rho\mu$$

$$\rho = \frac{\lambda}{\mu} = \lambda\bar{x}$$

M/G/1 queue

- Single server
 - **work-conserving** - it does not idle when there is work, no overhead, i.e., it performs 1 second of work per second
 - FCFS service
- Arrivals according to a Poisson process at rate λ jobs/second
- Service times of arrivals are $x_1, x_2, \dots, x_i \dots$ which are *independent, identically distributed* (with a **general** distribution)
- Average service time is \bar{x} , average wait is W , average delay is $T = W + \bar{x}$

Let $U(t)$ be the unfinished work at time t



Derivation of \bar{W}

Time average of unfinished work is

$$\begin{aligned}\bar{U} &= \frac{1}{\tau} \int_0^{\tau} U(t) dt \\ &= \frac{1}{\tau} \left(\frac{1}{2} \sum_{i=1}^n x_i^2 + \sum_{i=1}^n x_i w_i \right) \quad x_i \text{ and } w_i \text{ are independent} \\ &= \frac{n}{\tau} \left(\frac{1}{2} \overline{x_i^2} + \overline{x_i} \times \overline{w_i} \right) \quad \text{where } \overline{x_i w_i} = \overline{x_i} \times \overline{w_i}\end{aligned}$$

For Poisson arrivals, the average wait is equal to \bar{U} from the *Poisson arrivals see time average (PASTA)* Theorem

Derivation of W (cont.)

We then have

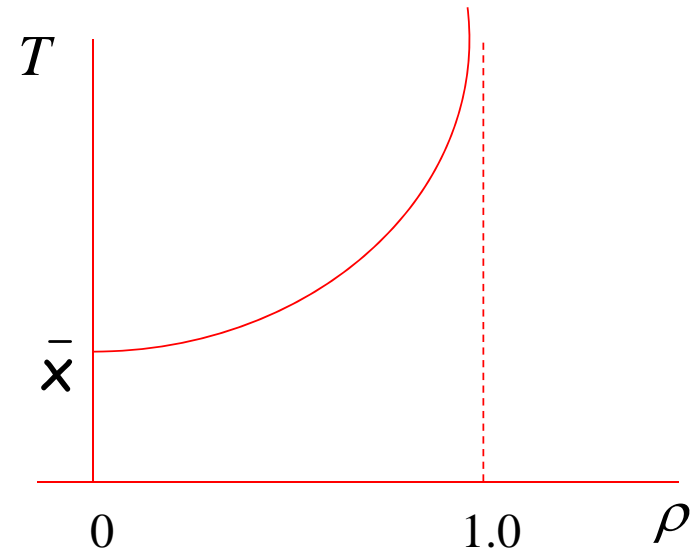
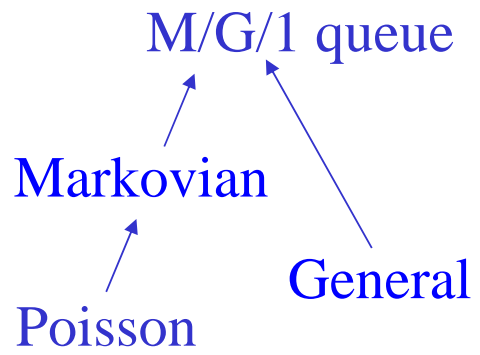
$$W = \lambda \left(\frac{1}{2} \overline{x^2} + \overline{x}W \right)$$

$$W(1 - \rho) = \frac{\lambda \overline{x^2}}{2}$$

$$W = \frac{\lambda \overline{x^2}}{2(1 - \rho)}$$

$$T = \overline{x} + W = \overline{x} + \frac{\lambda \overline{x^2}}{2(1 - \rho)}$$

The bottom two equations are known as Pollaczek-Khinchin (P-K) mean value formulas



Special Cases

1. Service times have an exponential distribution (M/M/1). We then have

$$\overline{x^2} = \frac{2}{\mu^2} = 2(\overline{x})^2$$

$$W = \frac{\lambda(2)(\overline{x})^2}{2(1-\rho)} = \frac{\lambda(\overline{x})^2}{1-\rho} = \frac{\rho(\overline{x})}{1-\rho}$$

$$T = W + \overline{x}$$

$$= \frac{\rho\overline{x}}{1-\rho} + \overline{x} = \frac{\rho\overline{x} + \overline{x} - \rho\overline{x}}{1-\rho}$$

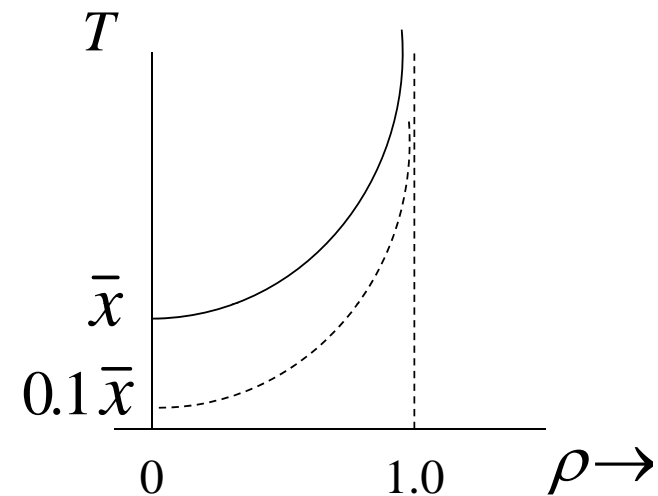
$$= \frac{\overline{x}}{1-\rho} = \frac{\rho}{1-\rho} \frac{1}{\lambda}$$

T decreases as λ
increases

$$\lambda \rightarrow 10\lambda$$

$$\mu \rightarrow 10\mu$$

$$\rho = \frac{10\lambda}{10\mu} = \frac{\lambda}{\mu}$$



M/G/1 queue (Simon S. Lam) 15

2. Service times are constant (deterministic)

↓
M/D/1

$$\overline{x^2} = (\bar{x})^2$$

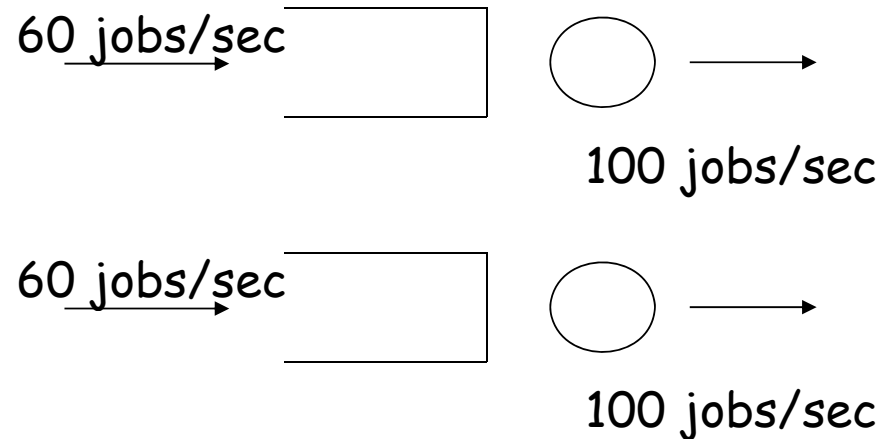
$$W = \frac{\lambda(\bar{x})^2}{2(1-\rho)} = \frac{\rho\bar{x}}{2(1-\rho)}$$

$$T = W + \bar{x}$$

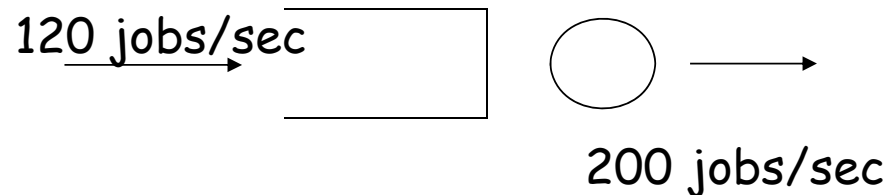
$$T = \frac{\rho(2-\rho)}{2(1-\rho)} \frac{1}{\lambda}$$

T decreases as λ
increases

Two Servers and Two Queues:



Single Higher Speed Server:

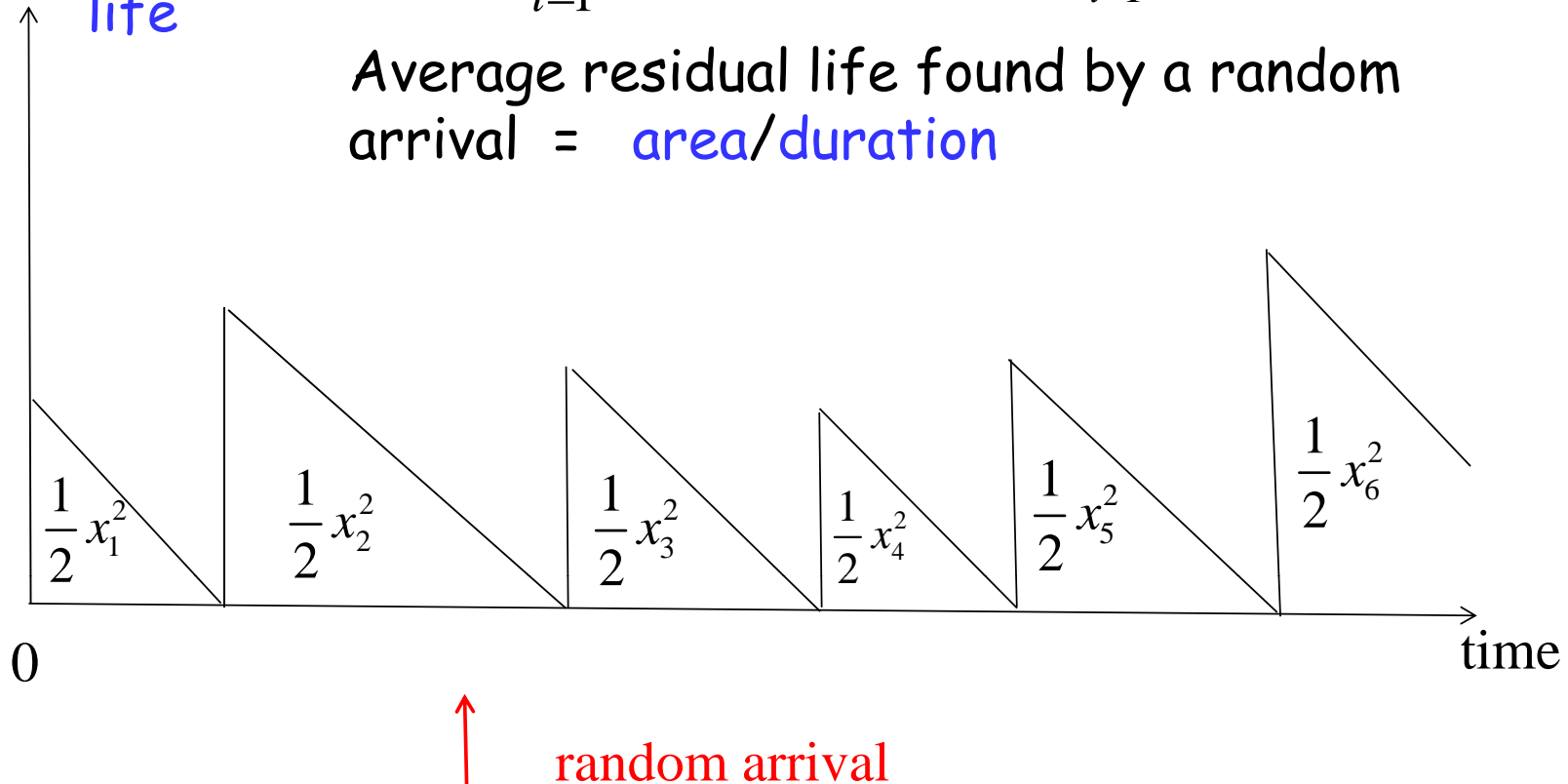


Residual life of a random variable x

residual
life

$$\text{area} = \sum_{i=1}^n x_i^2 / 2 \quad \text{duration} = \sum_{i=1}^n x_i$$

Average residual life found by a random arrival = $\text{area}/\text{duration}$



Mean residual life for 2 r.v. examples

$$\text{mean residual life} = \frac{\overline{X^2}}{2\overline{X}} \geq \frac{\overline{X}}{2}$$

Example 1: X is a constant

$$\overline{X^2} = (\overline{X})^2$$

$$\text{mean residual life} = \overline{X} / 2$$

Example 2: X is exponentially distributed

with density function $f_X(x) = \mu e^{-\mu x}$

In general, the exponential distribution is memoryless

$$\overline{X} = 1 / \mu \quad \text{and} \quad \overline{X^2} = 2 / \mu^2$$

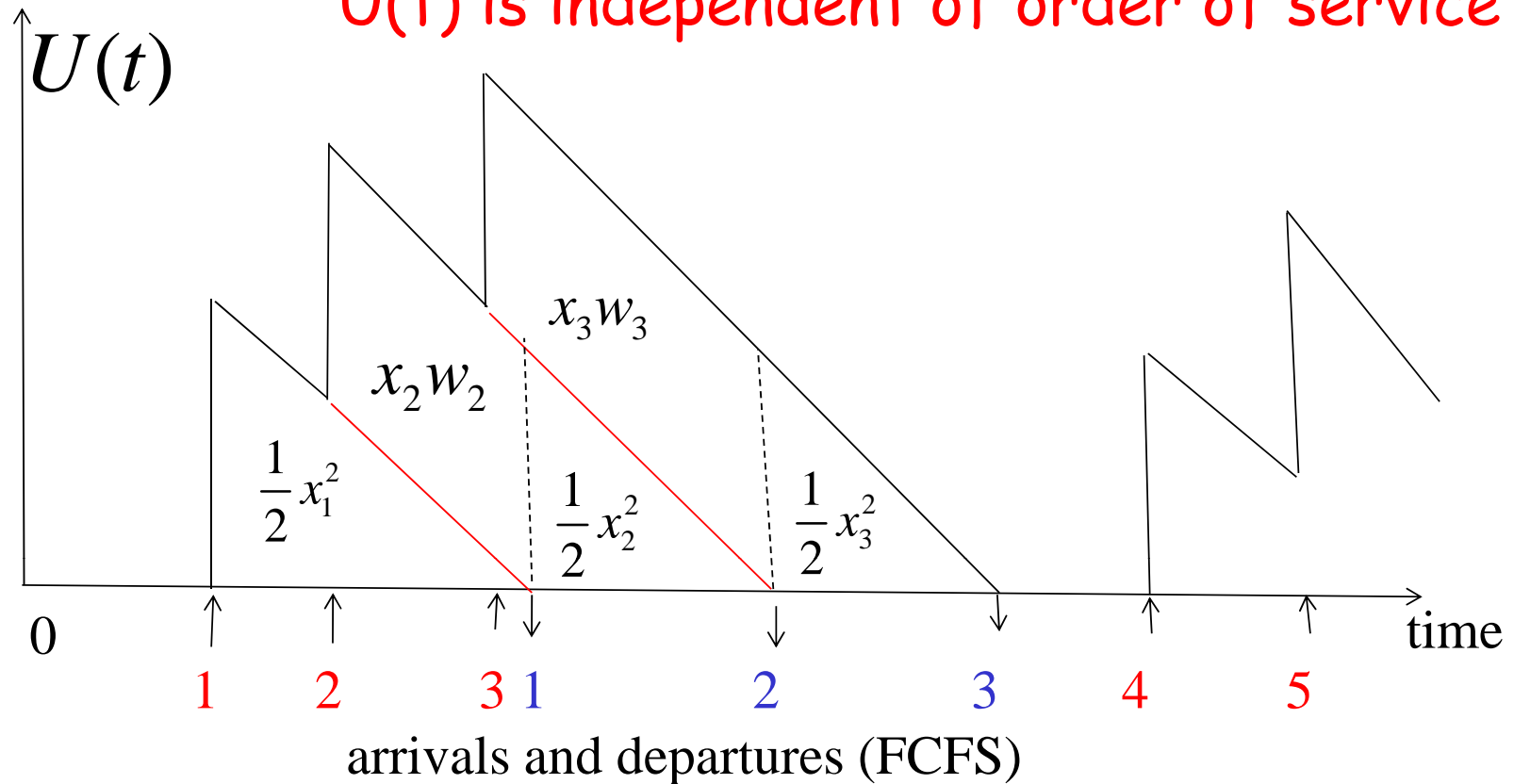
$$\text{mean residual life} = \overline{X} = 1 / \mu$$

Conservation Law

For a work conserving server, work gets done at one second per second.

$U(t)$ depends on the arrivals only.

$U(t)$ is independent of order of service



R classes of packets

with arrival rates, $\lambda_1, \lambda_2, \dots, \lambda_R$

mean service times, $\overline{x_1}, \overline{x_2}, \dots, \overline{x_R}$

and second moments, $\overline{x_1^2}, \overline{x_2^2}, \dots, \overline{x_R^2}$

Define $\rho_r = \lambda_r \times \overline{x_r}$ for $r = 1, 2, \dots, R$

$$\rho = \rho_1 + \rho_2 + \dots + \rho_R$$

$$U_S = \sum_{r=1}^R \rho_r \frac{\overline{x_r^2}}{2\overline{x_r}} = \sum_{r=1}^R \frac{\lambda_r \overline{x_r^2}}{2} = \frac{\lambda}{2} \sum_{r=1}^R \frac{\lambda_r \overline{x_r^2}}{\lambda} = \frac{\lambda \overline{x^2}}{2} = \rho \frac{\overline{x^2}}{2\overline{x}}$$

where $\frac{\overline{x^2}}{2\overline{x}}$ is ave. residual life of the packet found in service

Conservation Law

- Non-preemptive, work-conserving server
- Let W_r be the average wait of class r packets, $r = 1, 2, \dots, R$
- Let $N_{q,r}$ be the average number of class r packets in queue

The time average of unfinished work, $U(t)$, is

$$\bar{U} = U_S + \sum_{r=1}^R N_{q,r} \bar{x}_r = U_S + \sum_{r=1}^R \lambda_r W_r \bar{x}_r = U_S + \sum_{r=1}^R \rho_r W_r$$

We already have from P-K formula

$$\bar{U} = W_{FCFS} = \frac{\lambda \bar{x}^2}{2(1-\rho)} = \frac{U_S}{1-\rho}$$

Queueing Disciplines (Simon S. Lam)

Conservation Law (cont.)

$$\text{Therefore, } U_S + \sum_{r=1}^R \rho_r W_r = \frac{U_S}{1-\rho}$$

$$\sum_{r=1}^R \rho_r W_r = \frac{U_S}{1-\rho} - U_S = \frac{\rho U_S}{1-\rho} = \rho W_{FCFS}$$

which is the Conservation Law. It holds for any non-preemptive work-conserving queueing discipline

- Any preferential treatment for one class/customer is afforded at the expense of other classes/customers