Intro to Queueing Theory

- Little's Law
- M/G/1 queue
- Conservation Law

Little's Law

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

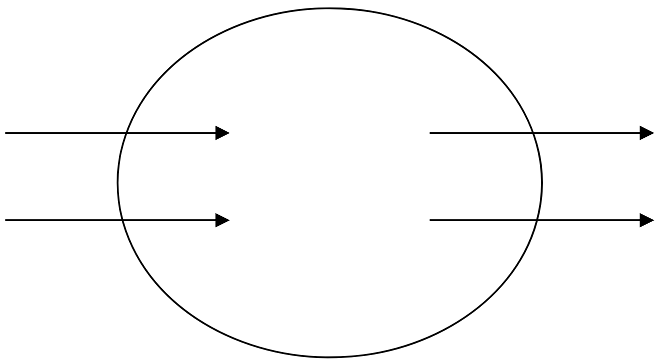
Average population = (average delay) x (throughput rate)

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

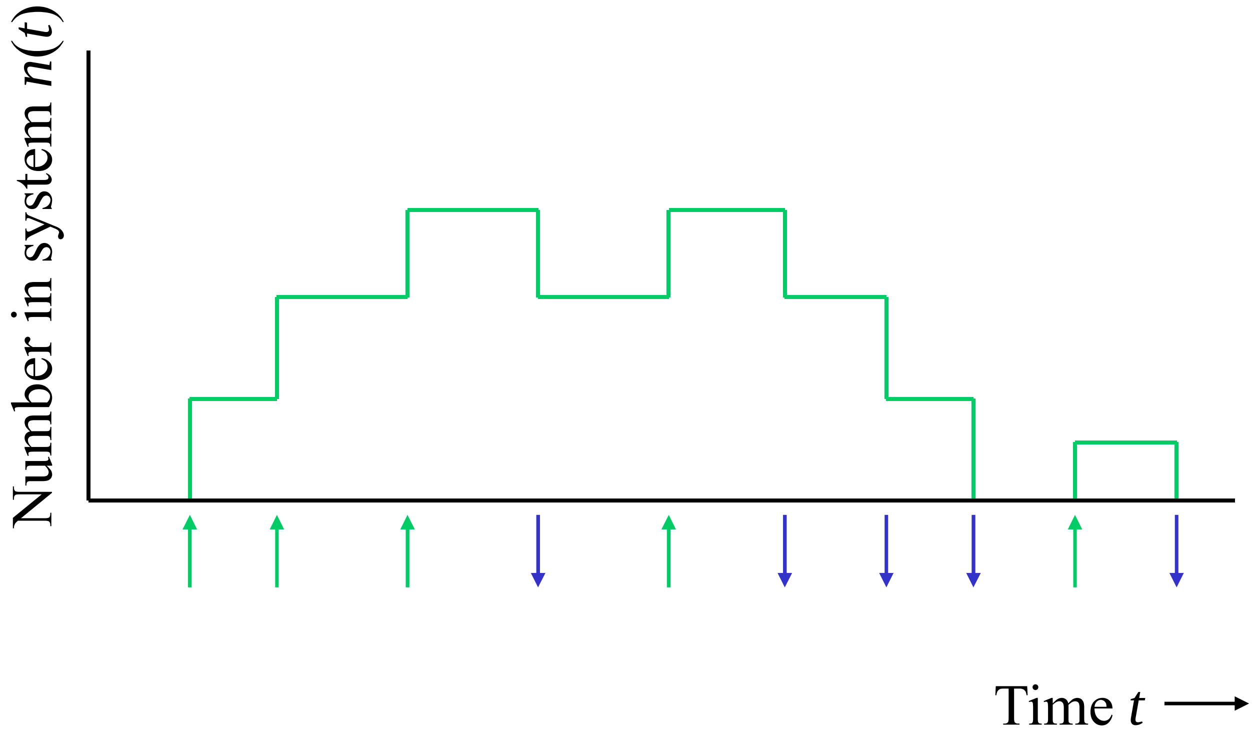
where N is number of departures

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

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]:

(a) calculate throughput rate;

(b) 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];

(c) 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 $x_1, x_2, \ldots, x_m$

Let P_i = probability $[X = x_i]$ for $i = 1, 2, \ldots, m$

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

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

Random Variable with continuous values

Random variable X with probability distribution function (PDF), $F_X(x) = \mathrm{P}[X \le x], \quad x \ge 0$

Its probability density function (pdf) is

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

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

Its second moment is

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

What if $F_X(x)$ is discontinuous?

Poisson arrival process at rate λ

- It is a counting process with *independent* increments. Let $X(s, s+t)$ be the number of arrivals in the time interval $(s, s+t)$. For any time s,

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

- The above can be derived from the binomial distribution by dividing t into n small intervals and let n go to infinity:

$$P[X(s,s+t)=k]=\lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k} \quad k\geq 0,\ t\geq 0$$

Time between consecutive arrivals in a Poisson process has the exponential distribution

- Consider the random variable *T* which is the time between consecutive arrivals
- Probability distribution function of *T* is

$$A(t) = P[T \le t] = 1 - P[T > t]$$
$$= 1 - P[X(s, s+t) = 0] = 1 - e^{-\lambda t}, \quad t \ge 0$$

- Probability density function of *T* is

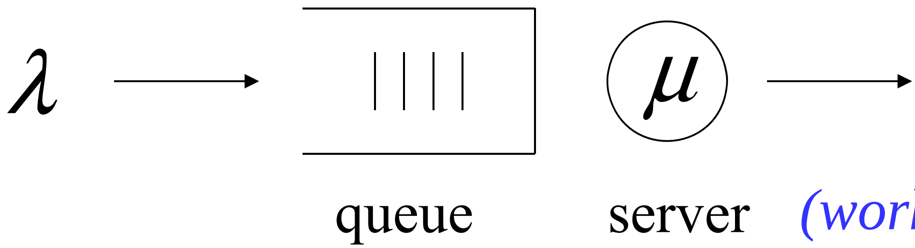
$$a(t) = \frac{dA(t)}{dt} = \lambda e^{-\lambda t} \quad t \ge 0$$

memoryless property

Topics

- **Average delay of M/G/1 queue with FCFS (FIFO) scheduling**
  - Pollaczek-Khinchin formula
  - ***motivation for packet switching***
- **Residual life of a random variable**
- **Conservation Law (M/G/1)**

Single-Server Queue



(work conserving)

(Note: For packets, service time is transmission time)

$\bar{x}$ average service time, in seconds

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

λ arrival rate, in packets/second

ρ utilization of server

Conservation of flow ($\lambda < \mu$, unbounded buffer)

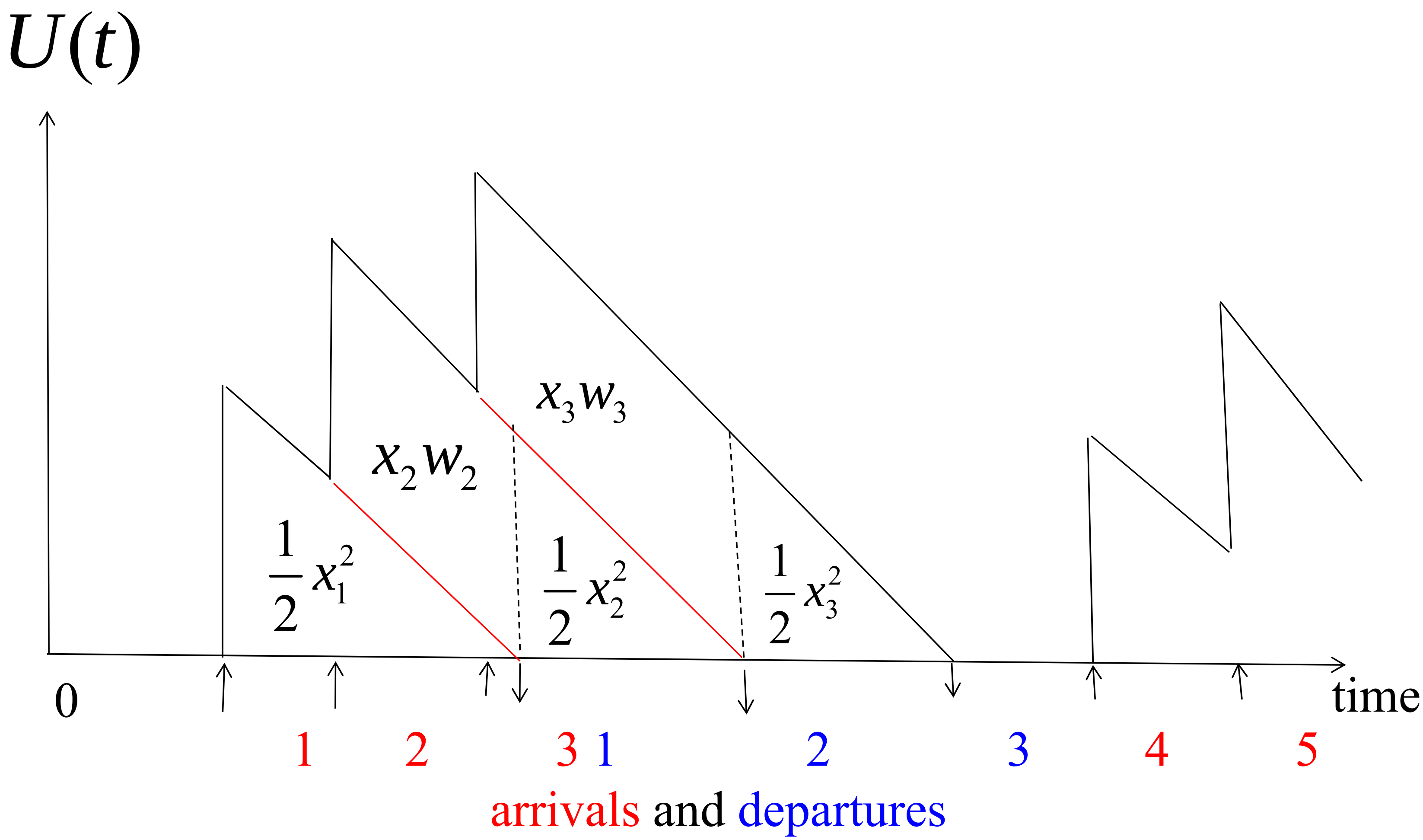
$$\boxed{\lambda = \rho\mu}$$

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

M/G/1 queue

- Single server
  - work-conserving – it does not idle when there is work, also 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 $x_1, x_2, \ldots, x_i \ldots$ 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 W

Time average of unfinished work is

$$\overline{U} = \frac{1}{\tau}\int_0^{\tau} U(t)dt$$

$$= \frac{1}{\tau}\left(\frac{1}{2}\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i w_i\right)$$

x_i and w_i are independent

$$= \frac{n}{\tau}\left(\frac{1}{2}\overline{x_i^{\,2}} + \overline{x_i} \times \overline{w_i}\right) \quad \text{where } \overline{x_i w_i} = \overline{x_i} \times \overline{w_i}$$

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

Derivation of W (cont.)

- The average wait is

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

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

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

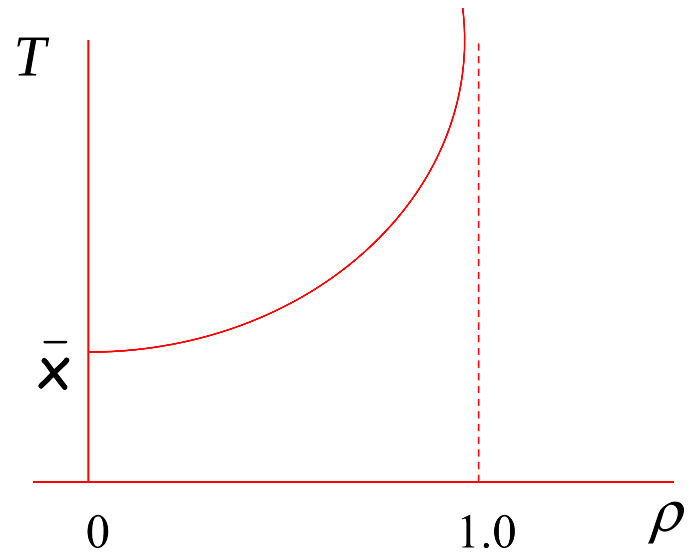
Pollaczek-Khinchin (P-K) mean value formula

M/G/1 queue

Markovian

Poisson

General



Average delay is

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

Also called Pollaczek-Khinchin (P-K) mean value formula

Special Cases

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

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

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

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

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

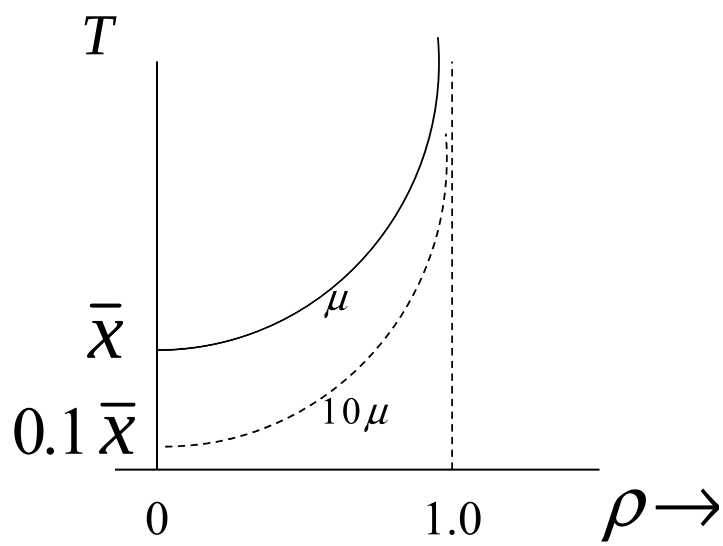
$$= \frac{\overline{x}}{1-\rho} = \boxed{\frac{\rho}{1-\rho}\frac{1}{\lambda}}$$

T decreases as λ increases

$$\lambda \rightarrow 10\lambda$$

$$\mu \rightarrow 10\mu$$

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



2. Service times are constant (deterministic)

$\downarrow$

M/D/1

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

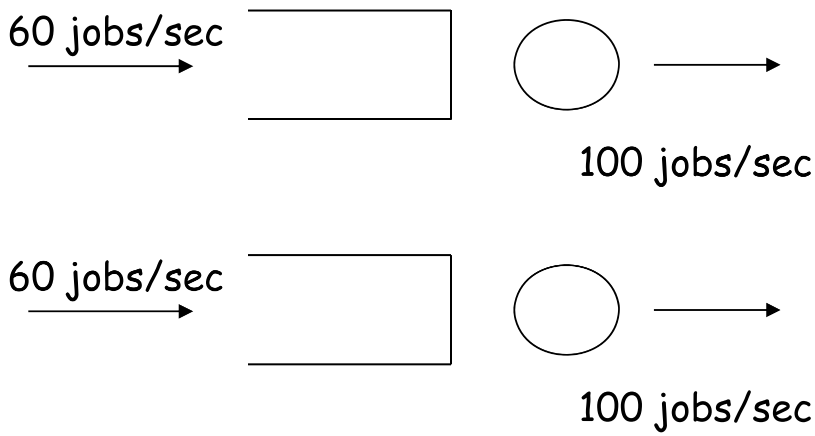
$$W = \frac{\lambda(\bar{x})^2}{2(1-\rho)} = \frac{\rho\bar{x}}{2(1-\rho)}$$

$$\boxed{T = \frac{\rho\bar{x}}{2(1-\rho)} + \bar{x}} = \frac{(\rho+2-2\rho)\bar{x}}{2(1-\rho)}$$

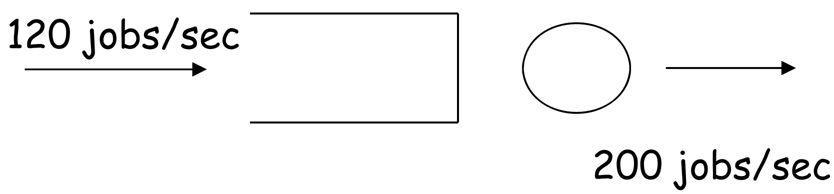
$$\boxed{T = \frac{\rho(2-\rho)}{2(1-\rho)}\frac{1}{\lambda}}$$

T decreases as λ increases

Two Servers and Two Queues:



Single Higher Speed Server:



Delay performance of packet switching over circuit switching

Consider how to share a 10 Gbps channel

1. Circuit switching : Divide 10 Gbps of bandwidth into 10,000 channels of 1 Mbps each and allocate them to 10,000 sources
2. Packet switching: Packets from 10,000 sources queue to share the 10 Gbps channel

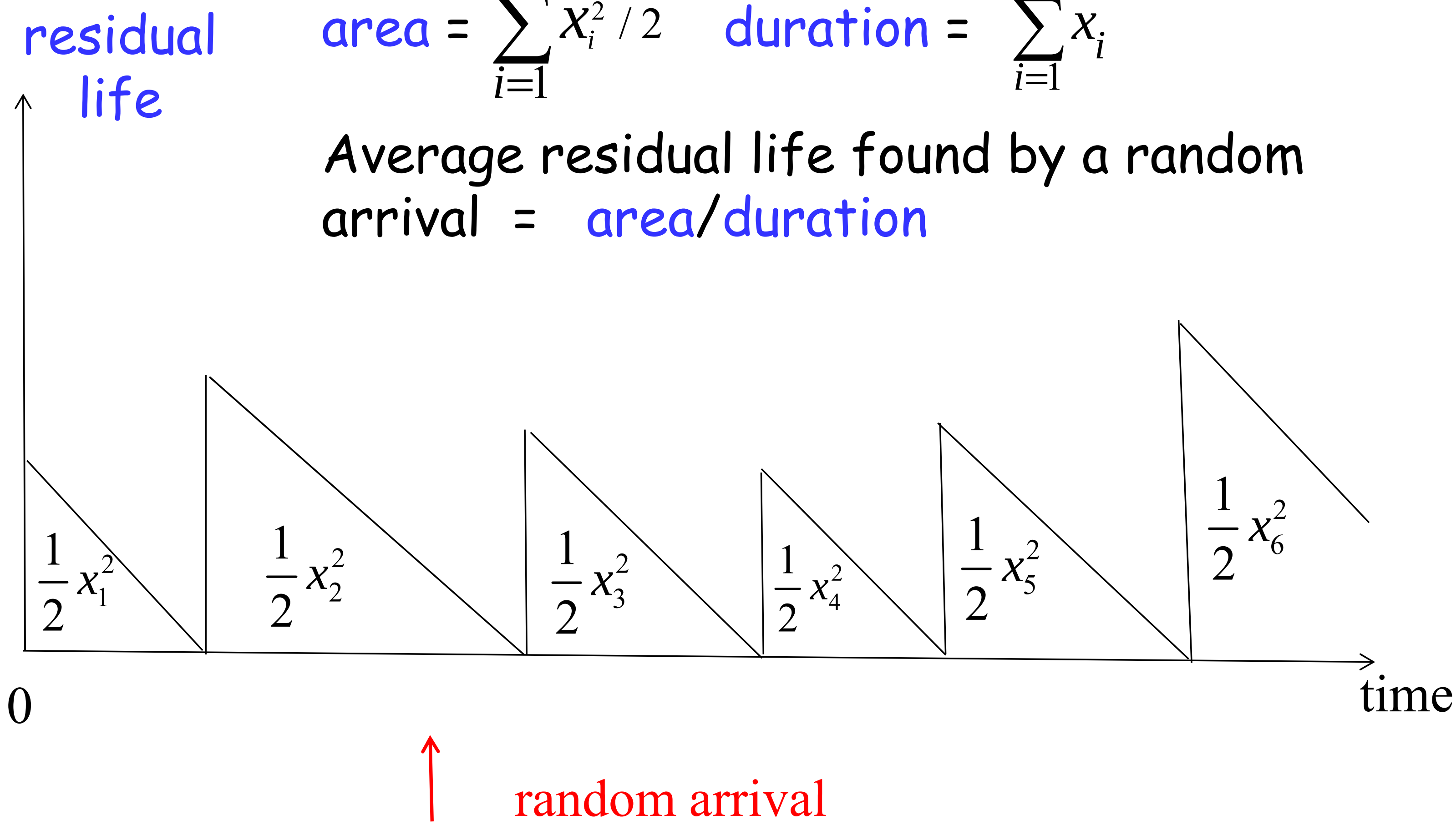
Packet switching delay is 10^{-4} of circuit switching delay

Contribution of queueing theory!

Residual life of a random variable x

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

Average residual life found by a random arrival = area/duration



Mean residual life for 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 $f_X(x) = \mu e^{-\mu x}$

Recall that the exponential distribution is ***memoryless***

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

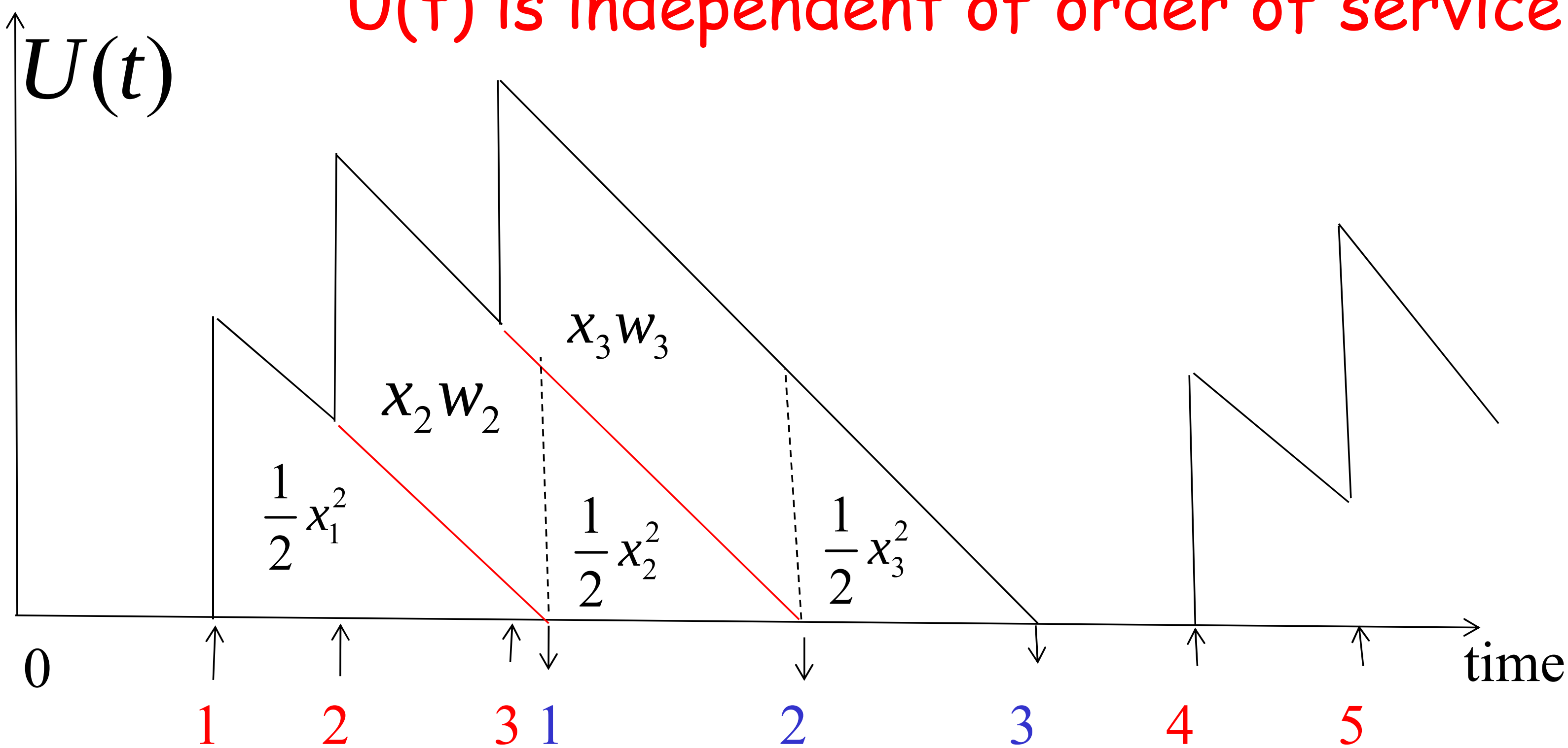
$$\text{mean residual life} = \overline{X} = 1/\mu$$

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



arrivals and departures (FCFS)

R classes of packets

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

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

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

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

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

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

where U_S is mean residual service, ρ_r is the fraction of time a class r packet is in service, and $\frac{\overline{x_r^2}}{2\overline{x_r}}$ is ave. residual service of the class r packet found by arrival

M/G/1 Conservation Law

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

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

$$\overline{U} = \boxed{U_S + \sum_{r=1}^{R} N_{q,r}\overline{x_r}} = \boxed{U_S + \sum_{r=1}^{R} \lambda_r W_r \overline{x_r}} = \boxed{U_S + \sum_{r=1}^{R} \rho_r W_r}$$

We already have from P-K formula (for $\rho<1$)

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

M/G/1 Conservation Law (cont.)

Therefore, $U_S + \sum_{r=1}^{R} \rho_r W_r = \dfrac{U_S}{1-\rho}$ for $\rho < 1$

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

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

The end