Delay Models and Queueing

The inputs to networks are unpredictable and best modeled probabilistically.

Queueing theory (customers with random service needs arrive at random times) is an appropriate model).

Little's Theorem

Under very broad conditions:

$$\begin{pmatrix}\text{Average number}\\ \text{of customers}\\ \text{in the system}\end{pmatrix} = \begin{pmatrix}\text{The rate}\\ \text{costomers}\\ \text{arrive at}\\ \text{the system}\end{pmatrix} \times \begin{pmatrix}\text{Average time}\\ \text{a customer}\\ \text{spend in}\\ \text{the system}\end{pmatrix}$$

$$N = \lambda \cdot T$$

The "system" could be a network, a queue, a queue plus a server, a server alone, a network of queues, etc..

Proof of Little's Theorem

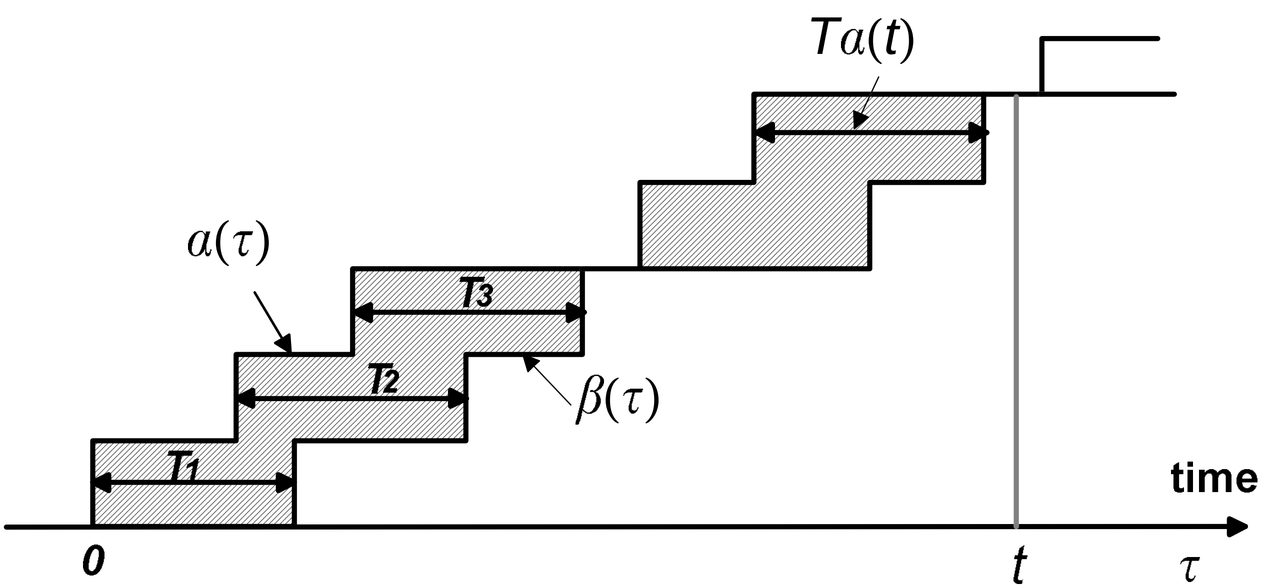
Let:

$\alpha(\tau) =$ **number of arrivals from time** 0 **to** τ

$T_i =$ **time spent in system by** ***i*** **th arrival**

$\beta(\tau) =$ **number of departures from time** 0 **to** τ

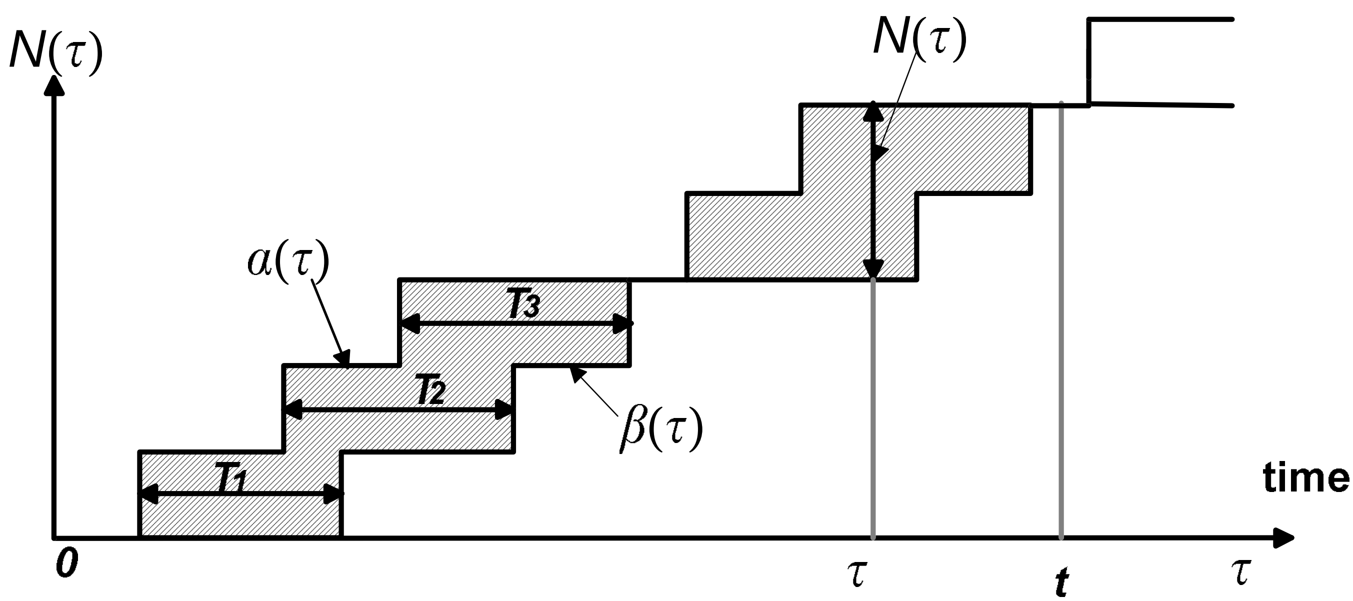
Assume that the system is empty at time 0.

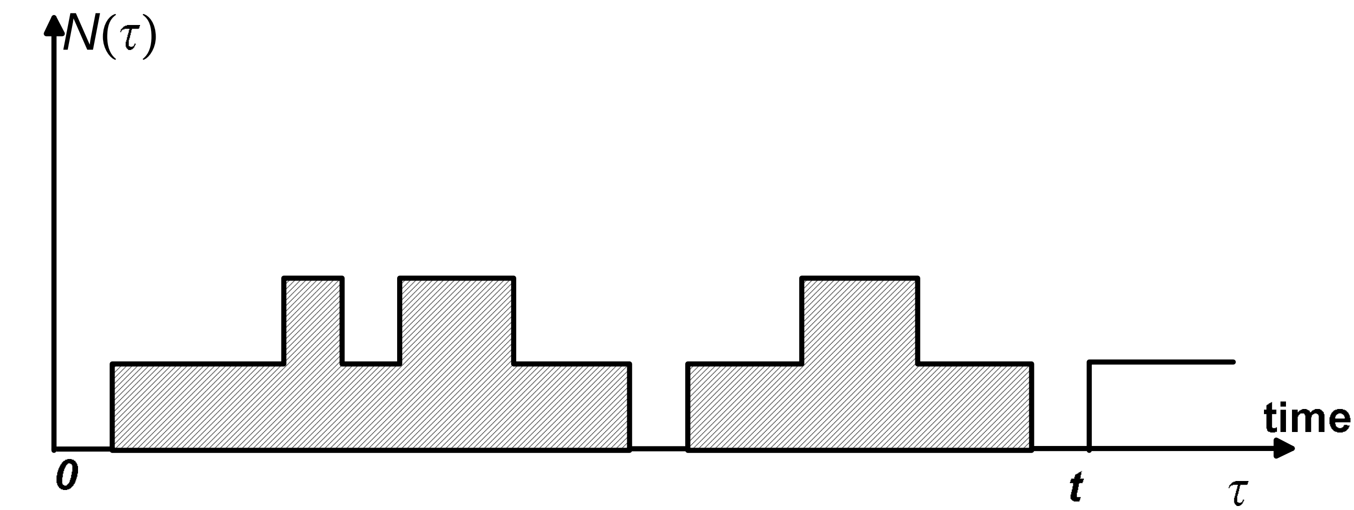


$$\textbf{Shaded area} = \sum_{i=1}^{\alpha(t)} T_i$$

(The total time customers $0, 1, \cdots\cdot \alpha(t)$ **spend in the system)**

Let:
$N(\tau) =$ **number of customers in system at time** τ
then $N(\tau) = a(\tau) - \beta(\tau)$





Shaded area = $\int_0^t N(\tau)d\tau$

Let N_t be the average number of customers from time 0 to t.

$$N_t = \frac{\int_0^t N(\tau)d\tau}{t}$$

$$N_t = \frac{1}{t}\sum_{i=1}^{\alpha(t)} T_i = \frac{\alpha(t)}{t}\cdot\frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

The average arrival rate from 0 to t as

$$\lambda_t = \frac{\alpha(t)}{t}$$

and the average time a customer is in the system as

$$T_t = \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

Thus: $$N_t = \lambda_t \cdot T_t$$

Assume that all three of these approach a limit (with probability 1) as $t \to \infty$.

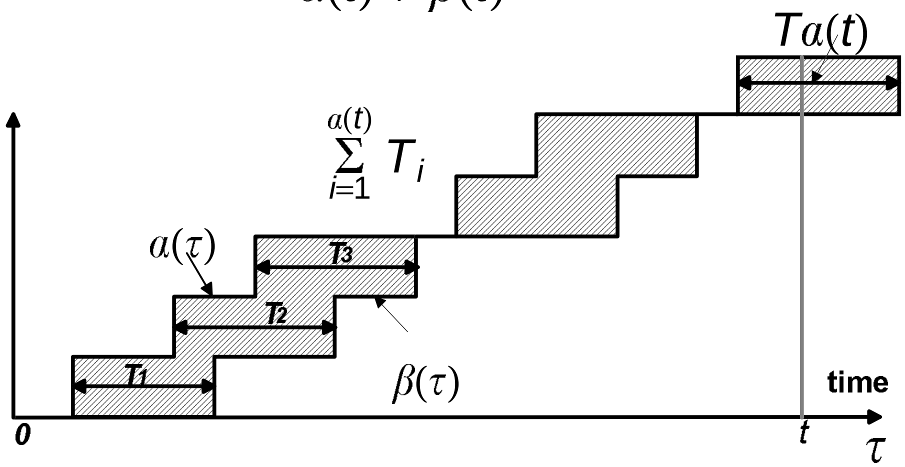
$$N = \lim_{t\to\infty} N_t$$

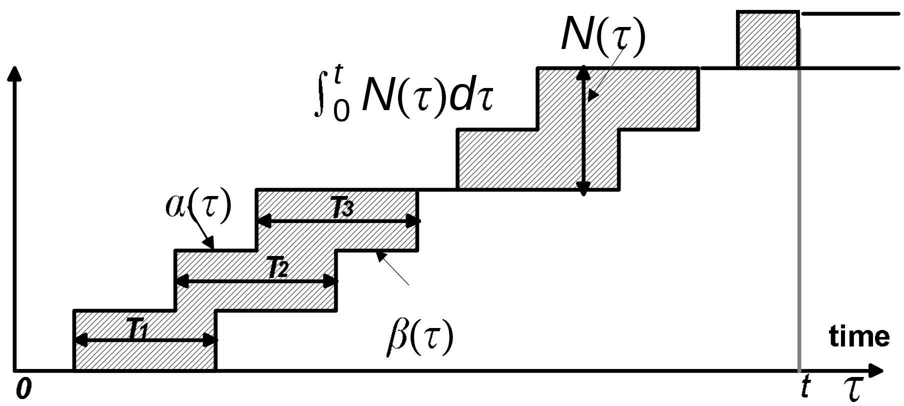
$$\lambda = \lim_{t\to\infty} \lambda_t$$

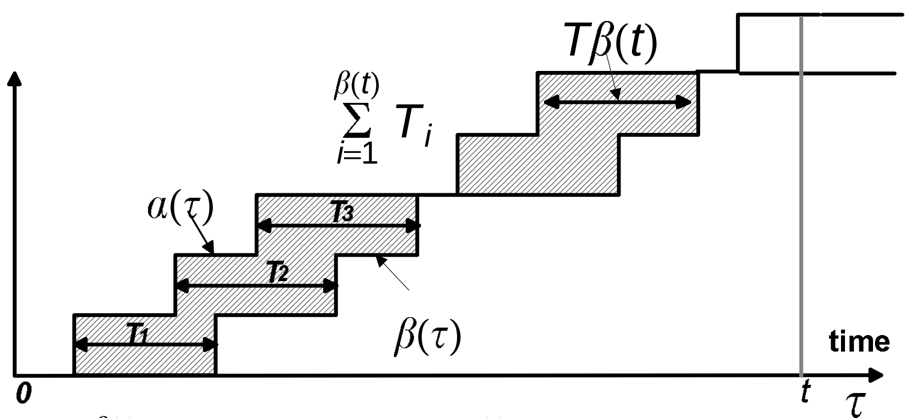
$$T = \lim_{t\to\infty} T_t$$

then $N = \lambda \cdot T$

$$\alpha(t) \neq \beta(t)$$





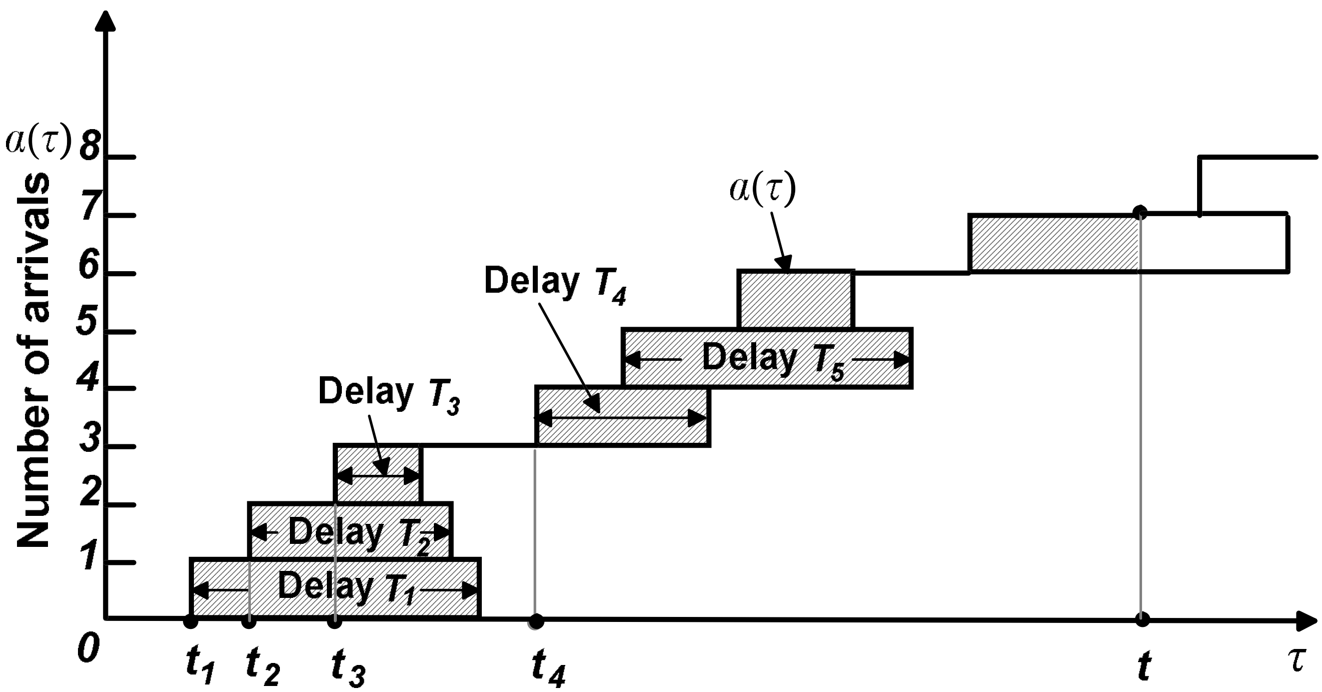


$$\sum_{i=1}^{\beta(t)} T_i \le \int_0^t N(\tau)d\tau \le \sum_{i=1}^{\alpha(t)} T_i$$

$$\frac{\sum_{i=1}^{\beta(t)} T_i}{t} = \frac{\beta(t)}{t} \cdot \frac{\sum_{i=1}^{\beta(t)} T_i}{\beta(t)} \le \frac{\int_0^t N(\tau)d\tau}{t} \le \frac{\sum_{i=1}^{\alpha(t)} T_i}{t} = \frac{\alpha(t)}{t} \cdot \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

$$t \to \infty, \lambda T \quad = \quad N \quad = \quad \lambda T$$

General Queueing Discipline



$$\sum_{i=1}^{\beta(t)} T_i \le \text{shaded area} = \int_0^t N(\tau)d\tau \le \sum_{i=1}^{\alpha(t)} T$$

Example:

$$N = \lambda \cdot T$$

Fast food restaurant (small *T*) require small dining are (small *N*) for a given λ.

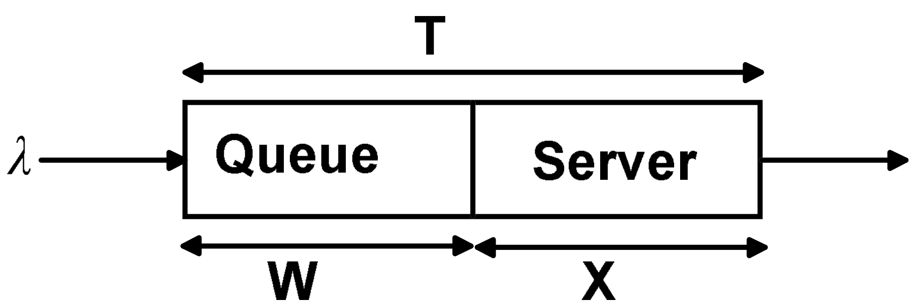
On a rainy day, people drive more slowly (*T* is large) and thus *N* is larger .

Example 3.1: Application of Little's Theorem

$$N = \lambda \cdot T$$

The "system" could be a queue, queue plus serve, network, server alone, etc..

e.g.



T = average delay in queue+server
W = average waiting time in queue
X = average service time

The average number of customers in queue or server

$$N = \lambda \cdot T$$

The average number of customers in queue alone

$$Q = \lambda \cdot W$$

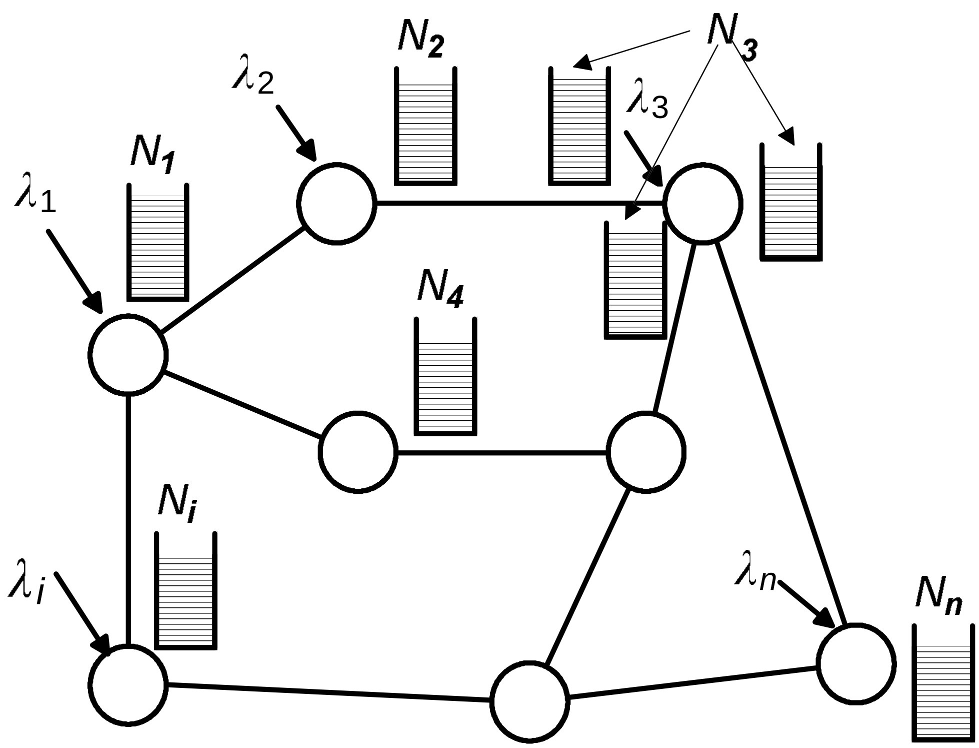
The average of customers number in server alone

$$\rho = \lambda \cdot X$$

Example 3.2:

λ_i: Arrival rate of source packets at node *i*.

N_i : Average number of packets in the queues of node *i*.



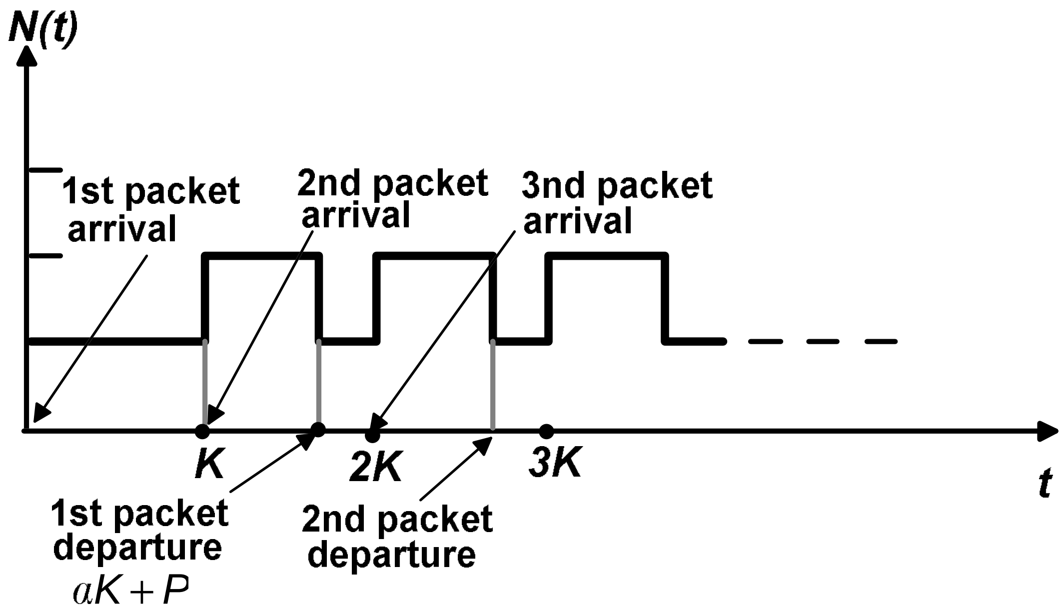
Average delay per packet

$$T = \frac{\sum_{i=1}^{n} N_i}{\sum_{i=1}^{n} \lambda_i}$$

Example 3.3:

A packet arrives every *K* seconds.

- **Transmission time:** αK **seconds.**
- **Processing time:** P **seconds.**



The average time spent in the system

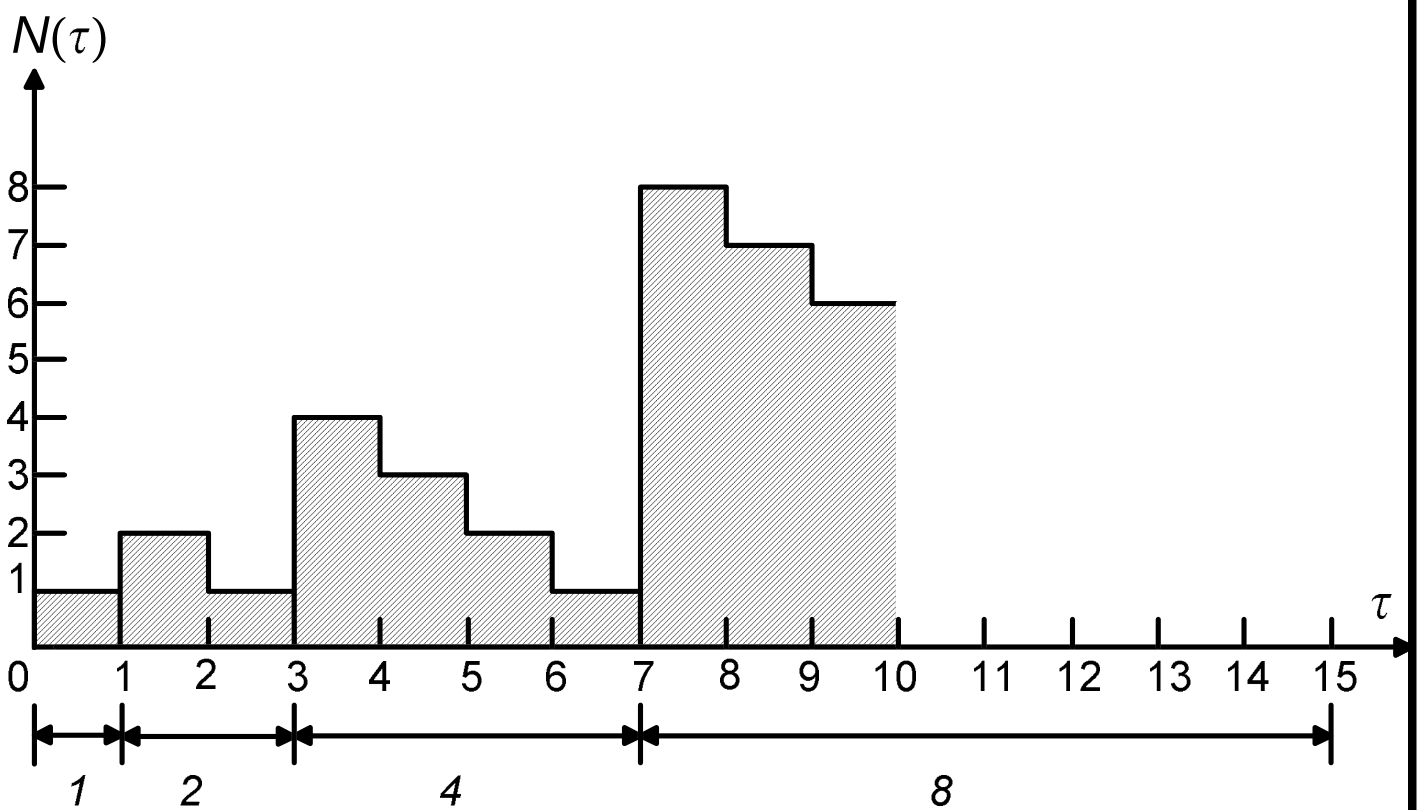
$$T = \alpha K + P$$

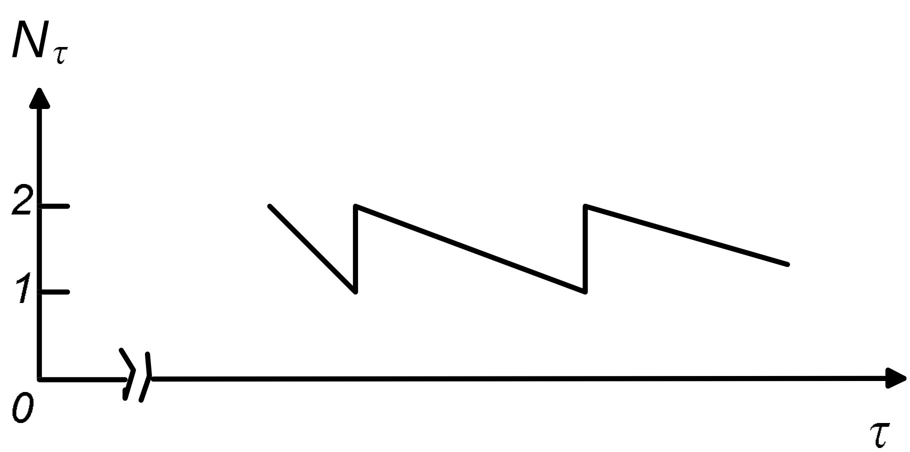
The average number of packets in the system

$$N = \lambda T = \alpha + \frac{P}{K}$$

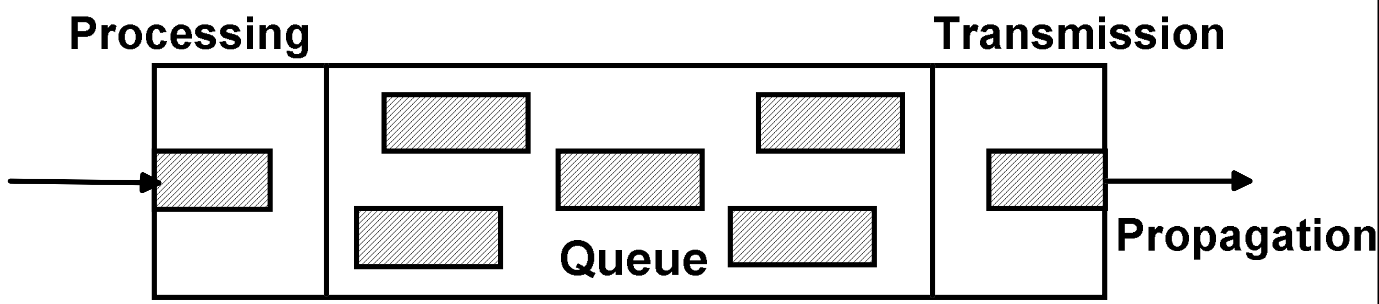
N(t) **does not converge to any value, but** *N* **does**

An example where N_τ does not converge to any value.





Components of node delay



Processing: time from end of packet reception to assignment to queue.

Queueing: time in queue until beginning of transmission (i.e. until "service" in queue jargon).

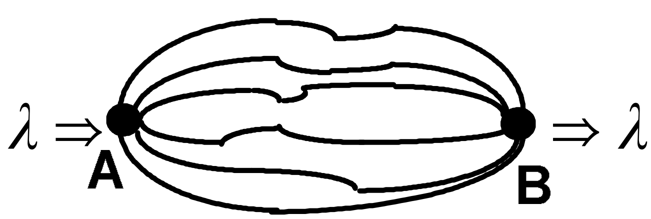
Transmission: message length/link rate.

Propagation: "flight time" of a bit.

Example 3.4: Window flow control

$$\lambda \cdot T = N \leq n$$

***n*: with go back *n* are in the network at most *n* packets**



$$\lambda \leq \lambda_{max}$$

If acknowledgements are received rightaway

$$\lambda \cdot T = N = n \approx \lambda_{max} \cdot T$$ **(when heavily loaded)**

window size

$$n \uparrow \Rightarrow \text{delay } T \uparrow$$

If delays for packets and acknowledgements are similar

$$N \approx \frac{n}{2} \approx \lambda_{max} \cdot T$$ **(heavy traffic)**

$$n \uparrow \Rightarrow T \uparrow$$

Example 3.5:

A system with *N* customers and *K* servers

- **Average service time =** $\overline{X}$
- $N \geq K$ **,** N**,** K **are constant**

The system is closed: there is a new customer arrives whenever a customer departs.

The arrival rate λ **satisfies**

$$K = \lambda \overline{X}$$

The average time a customer stay in the system

$$T = \frac{N}{\lambda} = \frac{N\overline{X}}{K}$$

Example 3.6:

A transmission line serves *m* packet streams (users) in round robin cycles.

- **Arrival rate λ_i for user i**
- **Transmission time $\overline{X_i}$**
- **Overhead A_i (Precedes the transmission)**

Average cycle length $L = ?$

Average number of packets on the transmission line

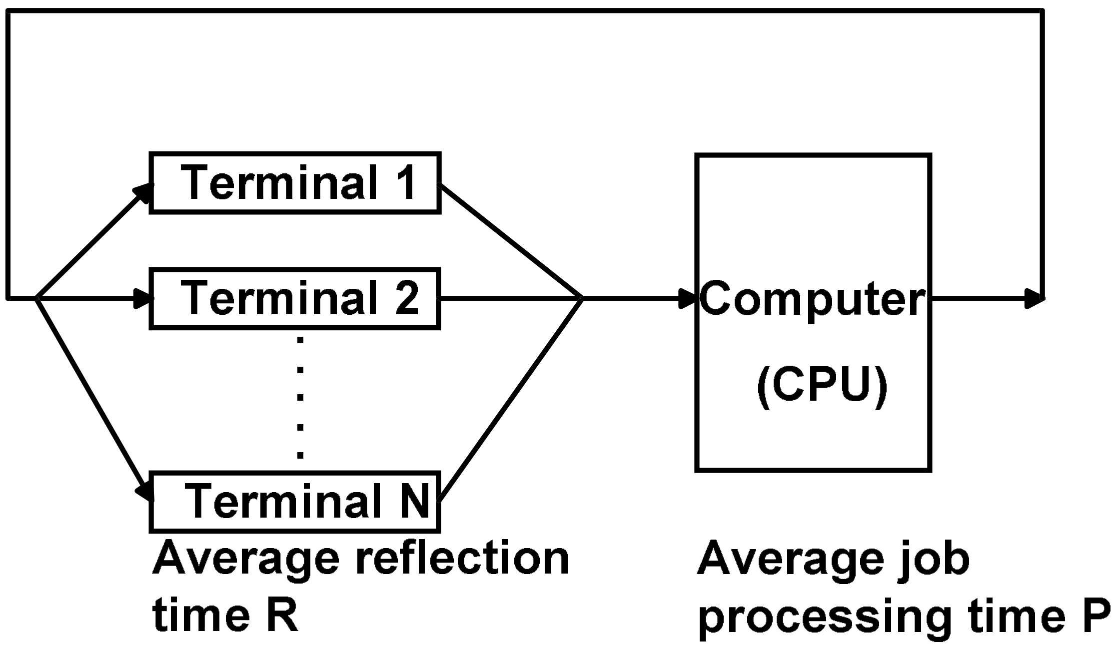
$$N = \sum_{i=1}^{m} \lambda_i \overline{X_i} \leq 1$$

The fraction of time the line is idle

$$\frac{\sum_{i=1}^{m} A_i}{L} = 1 - N = 1 - \sum_{i=1}^{m} \lambda_i \overline{X_i}$$

$$L = \frac{\sum_{i=1}^{m} A_i}{1 - \sum_{i=1}^{m} \lambda_i \overline{X_i}}$$

Example 3.7: (time sharing computer system)



$$\lambda = \frac{N}{T}$$

T**: average time a user spends in the system**

$$T = R + D$$

D**: average delay between time at which job is submitted at the CPU, and the time its execution is completed**

$$R + P \le T \le R + N \cdot P$$

$$\Rightarrow \frac{N}{R+N\cdot P} \leq \lambda \leq \frac{N}{R+P}$$

also $$\lambda \leq \frac{1}{P}$$

$$\Rightarrow \frac{N}{R+N\cdot P} \leq \lambda \leq \min\{\frac{1}{P}, \frac{N}{R+P}\}$$