Course Summary

- Introduction to queuing systems
- Basics of probabilistic theory and Markov chains.
- Modeling and dimensioning of communication systems in terms of delay, packet loss probability, system utilization etc.
- Markovian queuing systems
  - one server and multiple servers
  - unlimited/limited queue
  - unlimited/limited population of customers
  - Poisson/non Poisson arrivals
  - arrivals in batches or one at a time
- Semi-Markovian queuing systems
- Queuing networks

Aim

After the course students shall be able to:

- define the basic queuing models for different communication systems
- dimension the systems in terms of system ~~router~~ capacity, delay, utilization, throughput and packet loss probability, blocking probability

k = integers

Positive random variable $X \geq 0$

sequences

DISCRETE RANDOM VARIABLE

Probability function: $p_k = P(X = k)$

Properties: $p_k \geq 0$; $\sum_{k=0}^{\infty} p_k = 1$

Probability distribution of X:

$$F_X(x_i) = P(X \leq x_i) = \sum_{k=0}^{i} p_k$$

Properties: $F_X(x_i) \geq 0$; $F_X(0) = p_0$; $F_X(\infty) = 1$;

$F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$

Expected (mean) value (first moment) of X:

$$E[X] = m = \sum_{i=0}^{\infty} x_i p_i$$

Second moment of X: $E[X^2] = \sum_{i=0}^{\infty} x_i^2 p_i$

Variance of X: $Var[X] = E[(X-m)^2] = E[X^2] - m^2$

Squered coefficient of variance: $C^2 = Var[X]/ m^2$

CONTINUOUS RANDOM VARIABLE

Probability density function: $f_X(x)$

Properties: $f_X(x) \geq 0$; $\int_0^{\infty} f_X(u)du = 1$

Probability distribution of X:

$$F_X(x_i) = P(X \leq x) = \int_0^{x} f_X(u)du$$

Properties: $F_X(x_i) \geq 0$; $F_X(0) = 0$; $F_X(\infty) = 1$;

$F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$

Expected (mean) value (first moment) of X:

$$E[X] = m = \int_0^{\infty} x f_X(x)dx$$

Second moment of X: $E[X^2] = \int_0^{\infty} x^2 f_X(x)dx$

Variance of X: $Var[X] = E[(X-m)^2] = E[X^2] - m^2$

Squered coefficient of variance: $C^2 = Var[X]/ m^2$

System dimensioning problems

- Given arrival intensity and traffic characteristic
  - Design a system that meets requirements on
    - Delay (waiting time and service time)
    - Loss probability
    - Number of customers in the system
    - Blocking probability

  etc

- Given system and requirements
  - Define the arrival process that fits the system and requirements
    - Arrival rate can't be too high
    - Arrival pattern should be appropriate

  etc

System dimensioning problems

- Given arrival intensity and traffic characteristic
  - Design a system that meets requirements on
    - Delay (waiting time and service time)
    - Loss probability
    - Number of customers in the system
    - Blocking probability

  etc

- Given system and requirements
  - Define the arrival process that fits the system and requirements
    - Arrival rate can't be too high
    - Arrival pattern should be appropriate

  etc

Classification of stochastic processes

- Stochastic process SP: $X(t,\omega)$

Random variable X

states (state space)

Discrete X **Continuous X**

  - Discrete-time SP (discrete ***t***): $X(t,\omega)$
  - Continuous-time SP (continuous ***t***): $X(n,\omega)$
- Markov process (MP): a *memoryless* SP
- Markov chain (MC): MP with discrete $\mathbf{X}$
  - Discrete time MC (**DTMC**)
  - Continuous time MC (**CTMC**)
    - Birth-death process (**B-D** process), a special case of CTMC
    - **Poisson process**

Transforms; moment generating functions

- **DISCRETE X**

$\mathcal{Z}$ – **transform of p_i**

$$P(z) = E[z^i] = \sum_{i=0}^{\infty} z^i p_i$$

$$\frac{dP(z)}{dz} = \sum_{i=0}^{\infty} i \cdot z^{i-1} p_i$$

$$\frac{d^2P(z)}{dz^2} = \sum_{i=0}^{\infty} i(1-i)z^{i-2} p_i$$

$E[X] = P\,'(z)$ for $z = 1$

$E[X^2] = P\,''(z) + E[X]$ for $z = 1$

- **CONTINUOUS X**

$\mathcal{L}$– **transform of $f_X(x)$**

$$F^*(s) = E[e^{-sX}] = \int_0^{\infty} e^{-sX} f_X(x)dx$$

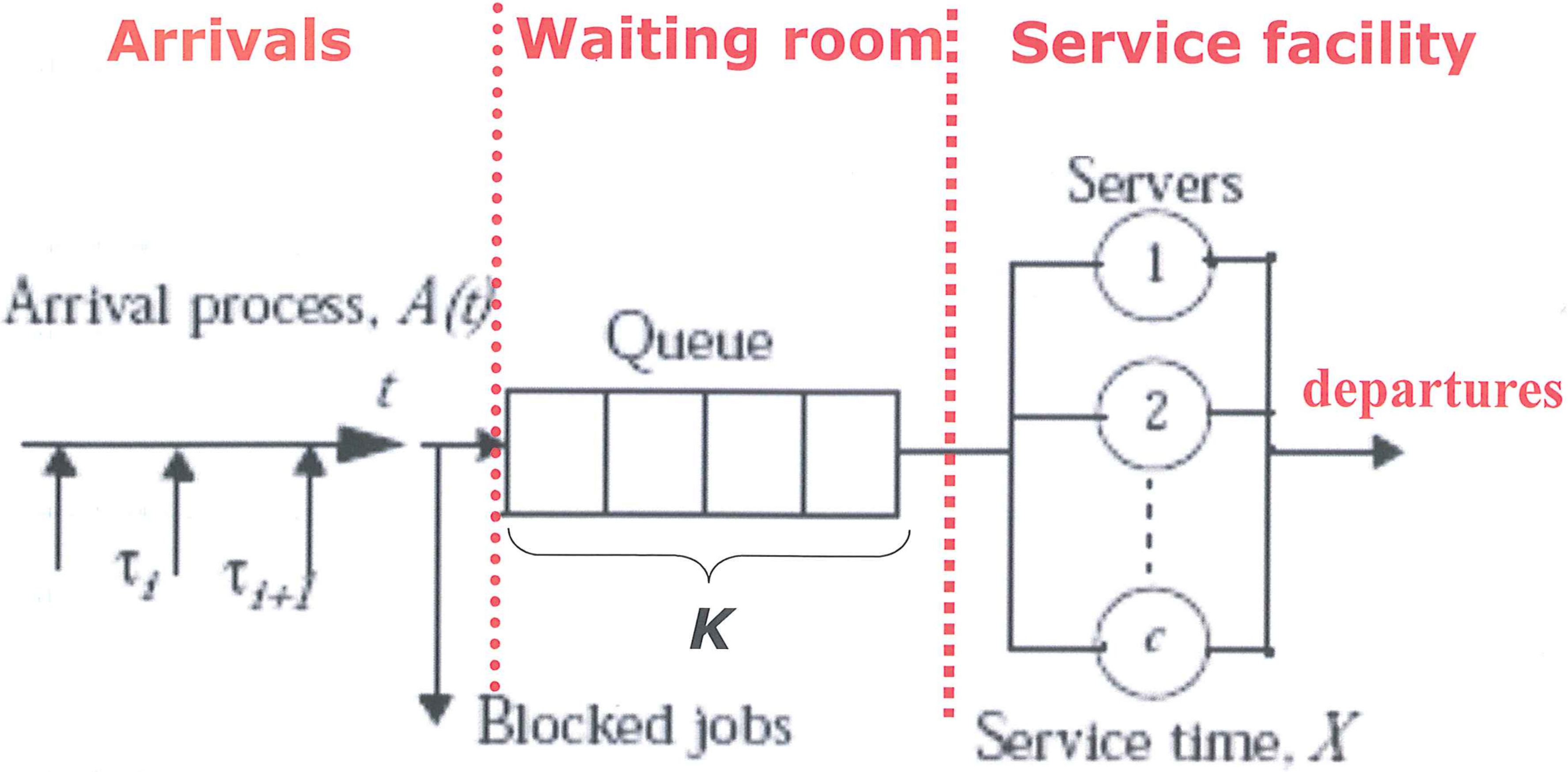
$$\frac{d\,F^*(s)}{ds} = \int_0^{\infty} -x e^{-sx} f(x)dx$$

$$\frac{d^2 F^*(s)}{ds^2} = \int_0^{\infty} x^2 e^{-sx} f(x)dx$$

$E[X] = -\,F^*\,'(s)$ for $s = 0$

$E[X^2] = F^*\,''(s)$ for $s = 0$

Queuing system



Some distributions of X

- <u>DISCRETE X</u>

Geometric distributed X:

$p_k = P(X = k) = a^{(k-1)}\ (1-a);\ 0<a<1$

$E[X] = 1/a$

Poisson distributed X:

$p_k = P(X = k) = \dfrac{a^k}{k!} e^{-a}$

$E[X] = a$

- <u>CONTINUOUS X</u>

Exponential distributed X:

$f_X(x) = a\ e^{-ax};\ 0<a<1$

$E[X] = 1/a$

Erlang$_r$ distributed X:

$f_X(x) = \dfrac{a^n}{(n-1)!} x^{n-1} e^{-ax}$

$E[X] = n/a$

Kendall's notation

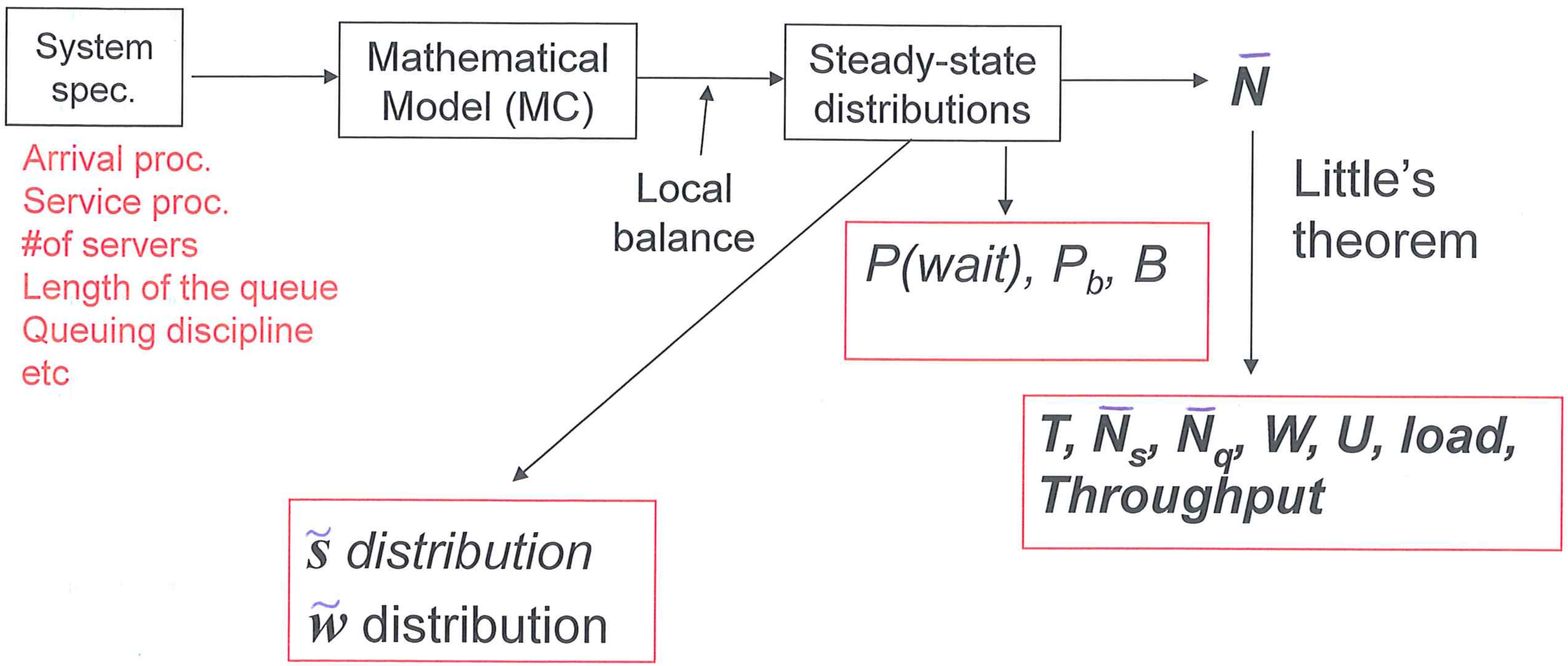
A/B/X/Y/Z

- **A**: arrival process
- **B**: service time
- X: number of servers
- Y: maximum occupancy (not indicated if unlimited buffer)
- Z: service order (not indicated if FCFS)
- <u>**A** and **B**</u> (arrival process and service time) can be:
  - ***M***: Markov (memoryless): exponential distributed time
  - ***D***: deterministic
  - E_r: ***r*** exponential distributed steps
  - H_k: hyper exponential with ***k*** branches
  - ***G***: general (but known)

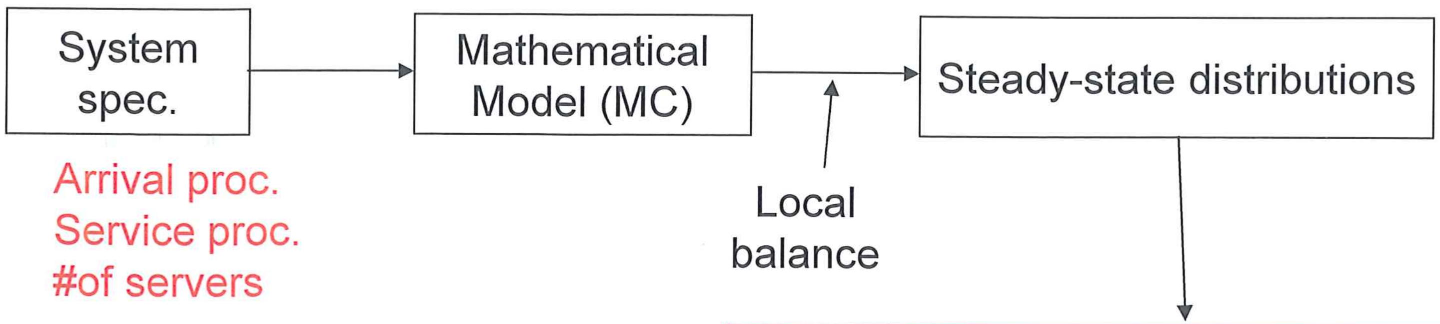
How to solve dimensioning problems?

- Analytical solution (queuing theory) for tractable systems
  - Develop a mathematical model of the system
  - The model should describe the system as accurate as possible
  - Based on your model you can be able to obtain the performance measures and dimension the system according to the requirements.
- Computer simulations for very complex systems

Dimensioning of queuing systems



Dimensioning of loss systems



Syst. performance parameters

(Time) blocking probability:

$P_b = p_c = P(c$ jobs in the system at a random time$)$

Call blocking probability:

$B = r_c = P(c$ jobs in the system at an arrival$)$

U, load, Throughput

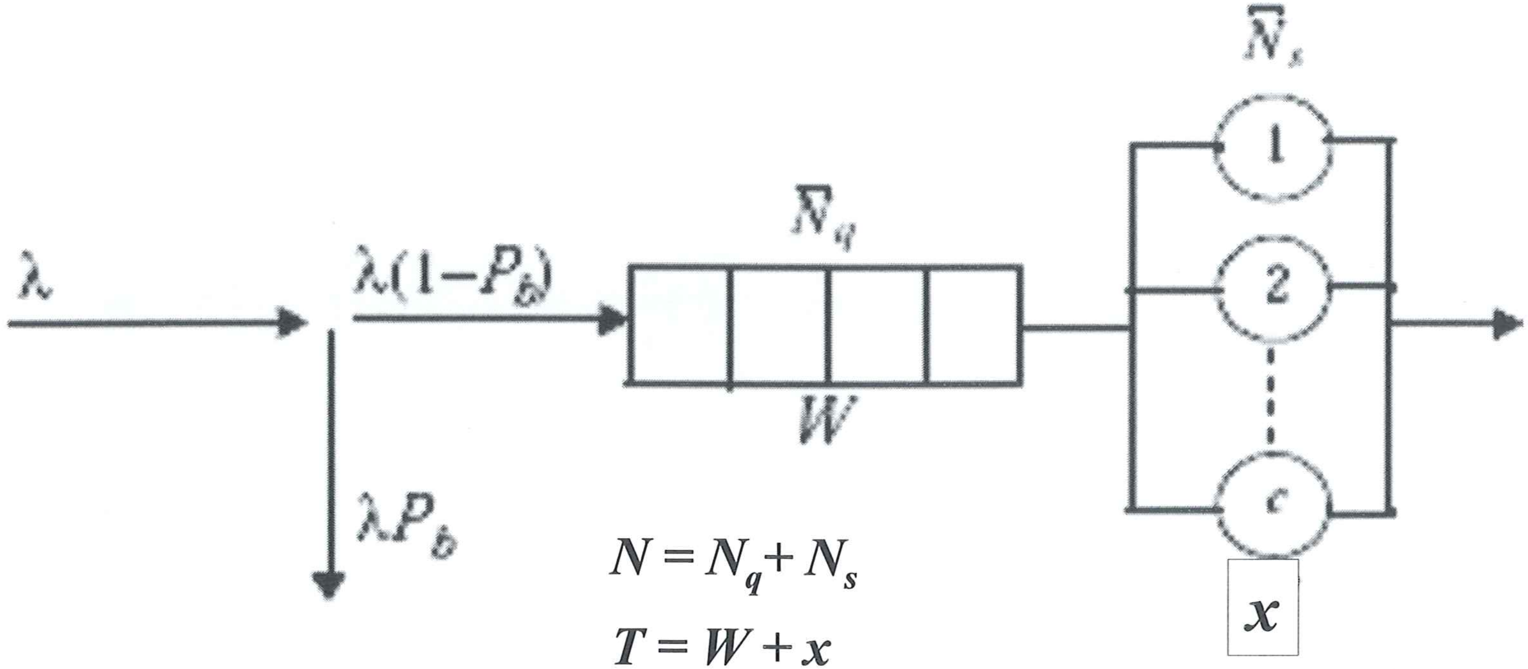
Mean number of blocked calls/time unit

Mean number of served calls/time unit

System performance parameters

- Average number of jobs
  - In the system: N
  - In the servers: N_s
  - In the queue: N_q
- Average waiting time: W
- Average service time: x
- Arrival intensity: λ
- Utilization: $U = N_s / c$
  - fraction of time the server is occupied (if one server, i.e. c=1)
  - fraction of servers that are occupied in average
- Load, expressed in Erlang [no unit]
  - Offered: $\rho = \lambda x$
  - Carried: $\rho_{\text{eff}} = \lambda_{\text{eff}}\, x$
- Throughput: λ_{eff}

System performance parameters



$$N = N_q + N_s$$

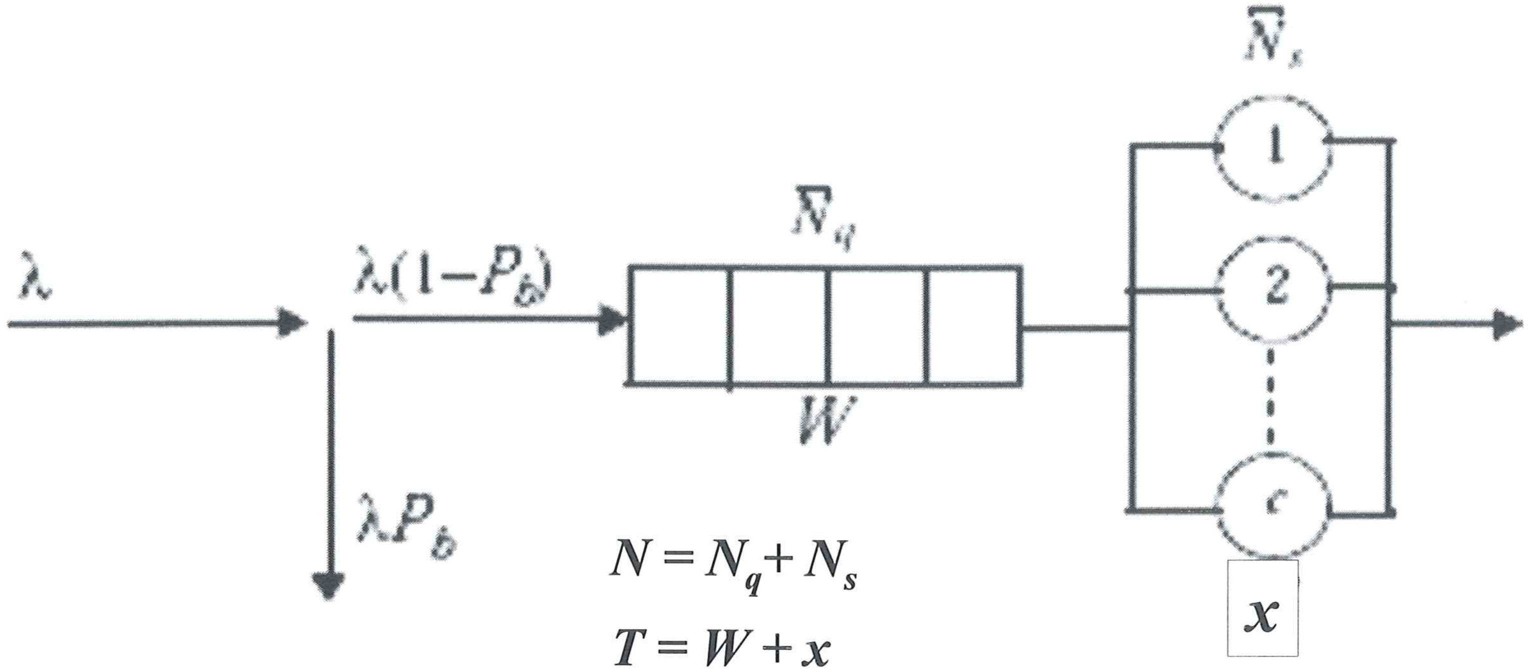
$$T = W + x$$

$$\lambda_{eff} = \lambda\,(1-P_b)$$

$$\rho = \lambda\, x$$

$$U = N_s / c$$

System performance parameters



$$N = N_q + N_s$$

$$T = W + x$$

$$\lambda_{eff} = \lambda\,(1-P_b)$$

$$\rho = \lambda x$$

$$U = N_s / c$$

Little's theorem

- The average number of customers in the system is equal to the average arrival rate times the average time spent in the system:

$$N = \lambda T$$

- The average number of customers in the queue is equal to the average arrival rate times the average waiting time:

$$N_q = \lambda W$$

- The average number of customers in the server(s) is equal to the average arrival rate times the average service time:

$$N_s = \lambda x$$

Stability conditions

- Arrival intensity is lower than departure intensity, i.e. mean time between arrivals is longer than mean time between departures
- For system with one server:
  - $\lambda < 1/x$
  - if $1/x = \mu \rightarrow \lambda < \mu$
  - Offered load $\rho < 1$
- For system with c parallel servers:
  - $\lambda < c/x$ or $\lambda < c\mu$
  - Offered load $\rho < c$