Models of distributed system, multiclass Markov chain, burst of arrivals

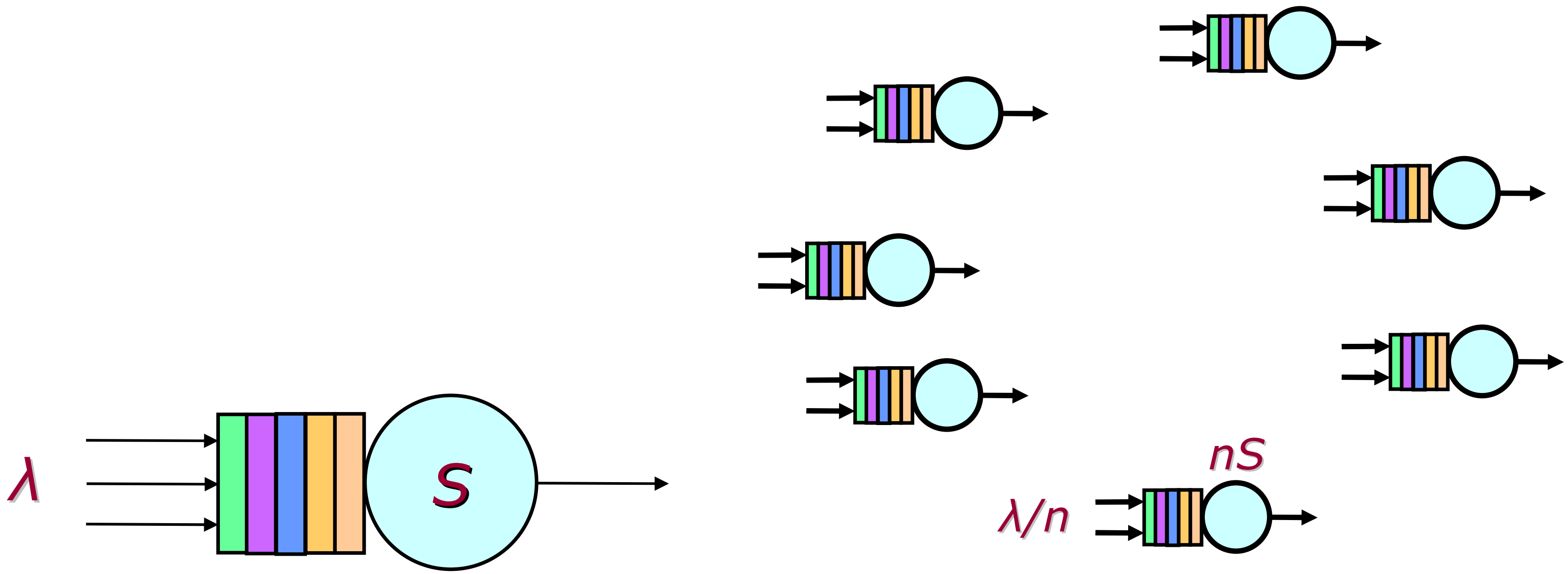
06/06/07

outline

- centralized vs distributed web architecture
- Markov Chain of a multiclass network
- burstiness model

centralized vs distributed web architecture

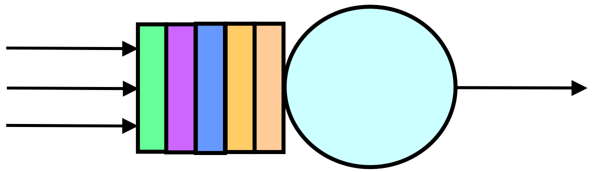
centralized vs distributed web sites

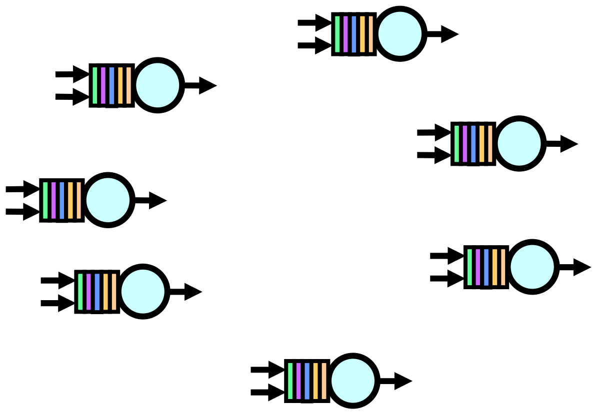


- *1* web site centralized
- capacity processing S
- workload λ

- n web sites distributed
- localized accesses
- *1/n th* capacity proc., nS
- *1/n th* workload, λ/n

centralized vs distributed web sites



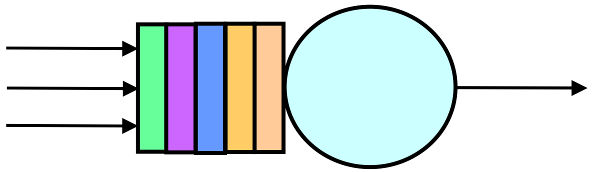


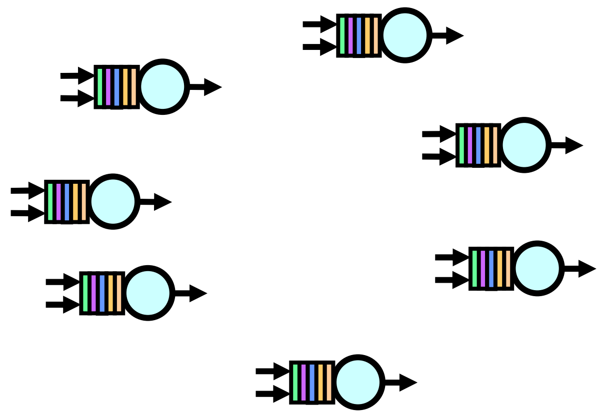
$$U_{centr} = \lambda S$$

$$U_{dec} = \frac{\lambda}{n} \; x \; nS = \lambda S = U_{centr}$$

the utilizations are the same, the troughput is the same,
thus the response times are the same!

centralized vs distributed web sites





$$R_{centr} = \frac{S}{1-U_{centr}} \qquad R_{dec} = \frac{nS}{1-U_{dec}} = n\,\frac{S}{1-U_{centr}} = n\;R_{centr}$$

the response times if the distributed architecture are *n times* the ones of the centralized architecture!!!

10 distributed web sites

$$U_{centr} = 0.7 \qquad\qquad S = 2\sec$$

$$R_{centr} = \frac{S}{1 - U_{centr}} = \frac{2}{1 - 0.7} = 6.66\sec$$

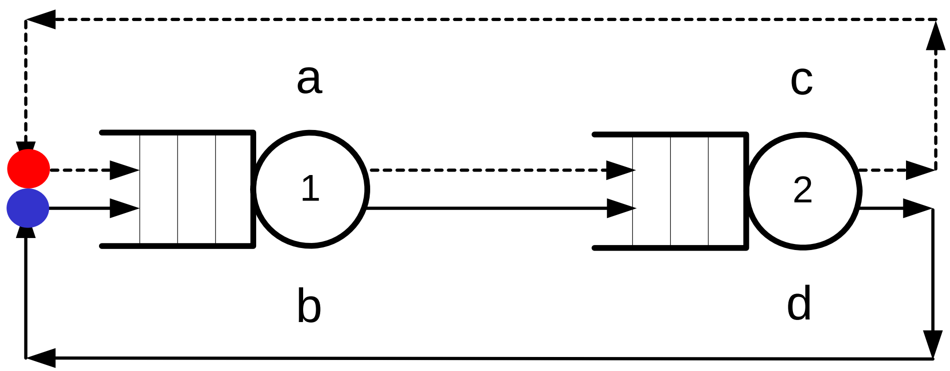
$$U_{dec} = 0.7 \qquad R_{dec} = \frac{nS}{1 - U_{dec}} = \frac{10\, x\, 2}{1 - 0.7} = 66.6\sec$$

***** the response times of the decentralized architecture are *n times* the ones of the centralized architecture!!!

Markov Chain of a network with two class workload

network with 2 class of requests

2 customers, 2 classes, 2 resources

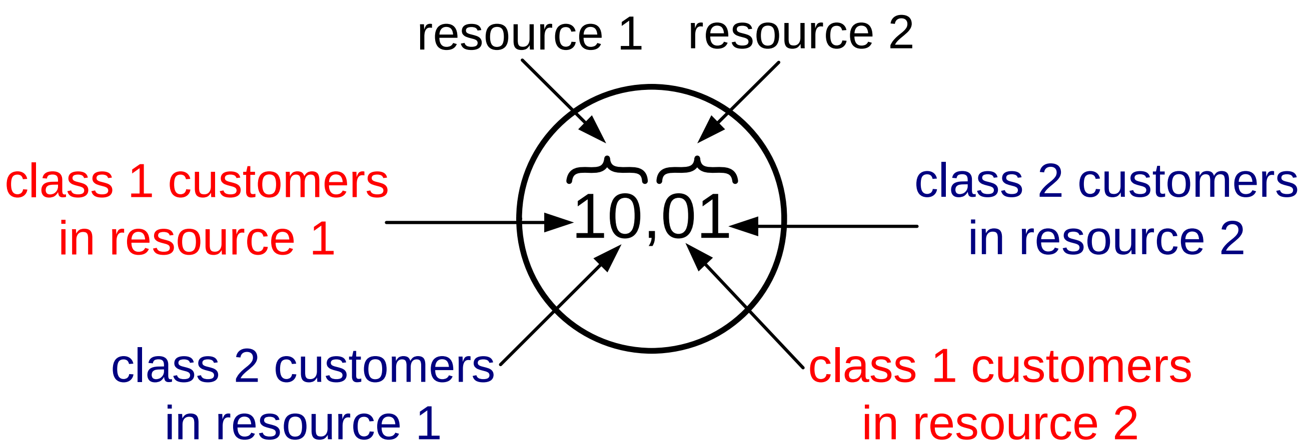


a, b, c, d
rates of execution
(μ=1/S)

$$\text{resources}\quad \begin{array}{c|cc|} & \multicolumn{2}{c}{\text{classes}} \\ & 1 & 2 \\ 1 & 1/a & 1/b \\ 2 & 1/c & 1/d \end{array}$$

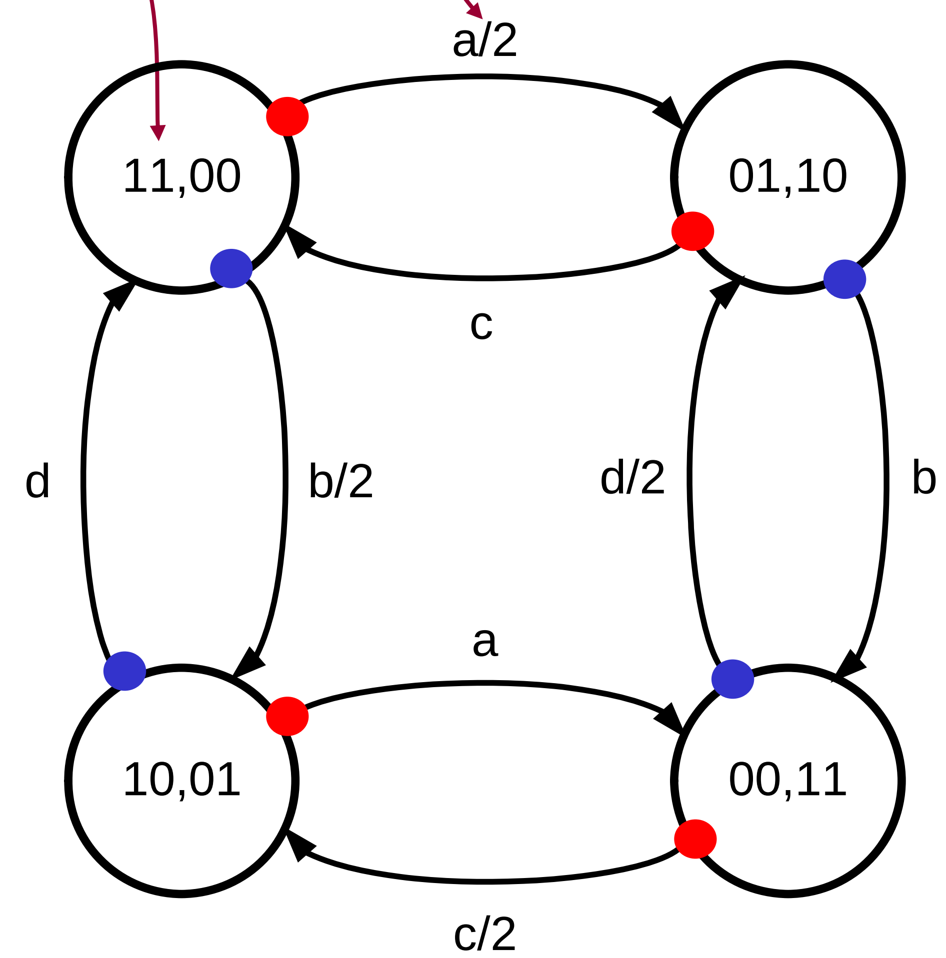
matrix of the D's

state of the Markov chain



Markov Chain

2 customers
share
the capacity of
the same
resource



state probabilities

$$P_{01,10} = \frac{a}{2c} P_{11,00}$$

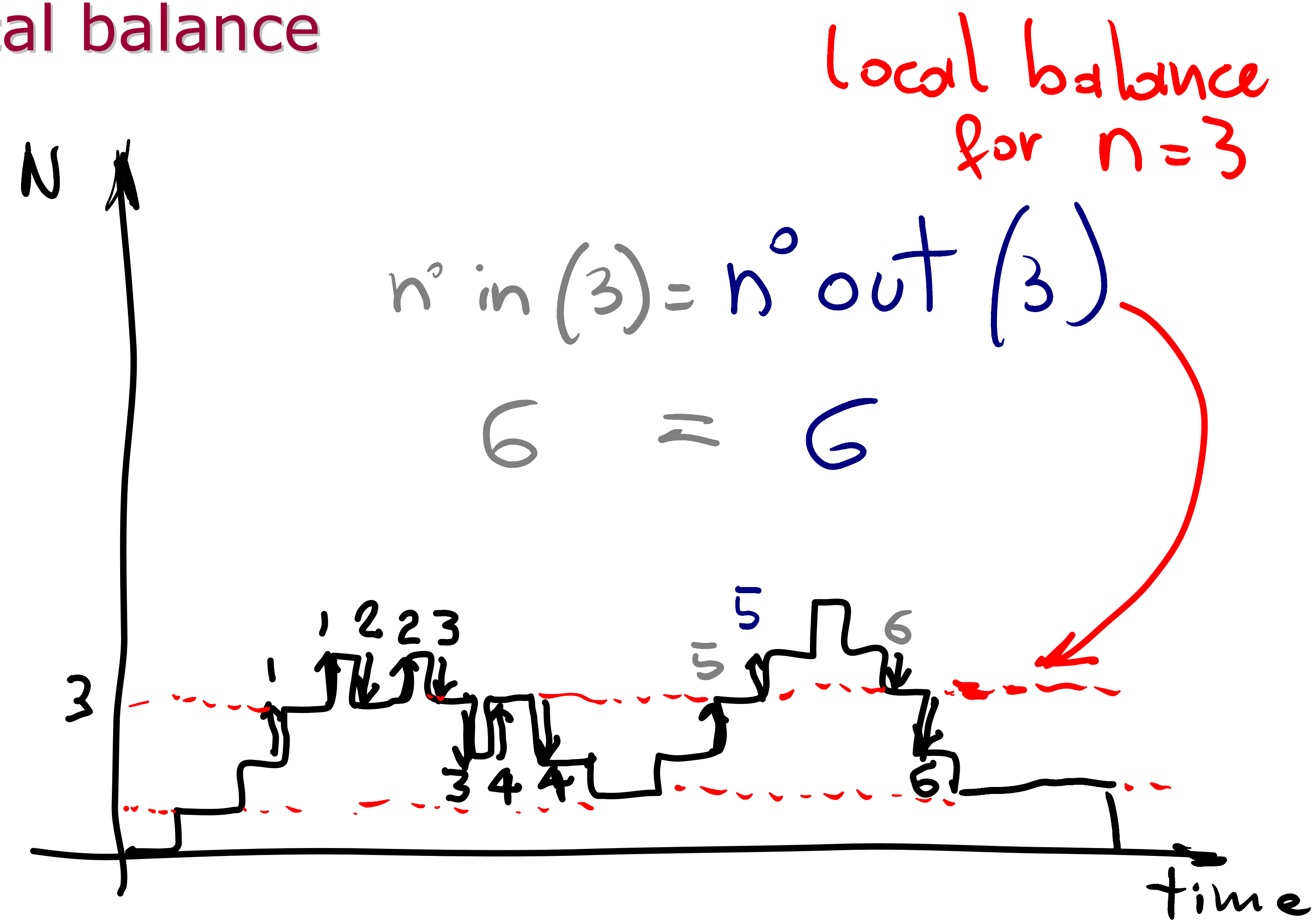
$$P_{10,01} = \frac{b}{2d} P_{11,00}$$

$$P_{00,11} = \frac{2b}{d} P_{01,10} = \frac{2b}{d} \frac{a}{2c} P_{11,00} = \frac{ab}{cd} P_{11,00}$$

$$P_{11,00} + P_{01,10} + P_{10,01} + P_{00,11} = 1$$

$$P_{11,00} + \frac{a}{2c} P_{11,00} + \frac{b}{2d} P_{11,00} + \frac{ab}{cd} P_{11,00} = 1$$

local balance



state probabilities

$$P_{11,00} = \frac{1}{1 + \frac{a}{2c} + \frac{b}{2d} + \frac{ab}{cd}} = \frac{2cd}{2cd + ad + bc + 2ab}$$

$$P_{01,10} = \frac{a/2c}{1 + \frac{a}{2c} + \frac{b}{2d} + \frac{ab}{cd}} = \frac{ad}{2cd + ad + bc + 2ab}$$

$$P_{10,01} = \frac{b/2d}{1 + \frac{a}{2c} + \frac{b}{2d} + \frac{ab}{cd}} = \frac{bc}{2cd + ad + bc + 2ab}$$

$$P_{00,11} = \frac{ab/cd}{1 + \frac{a}{2c} + \frac{b}{2d} + \frac{ab}{cd}} = \frac{2ab}{2cd + ad + bc + 2ab}$$

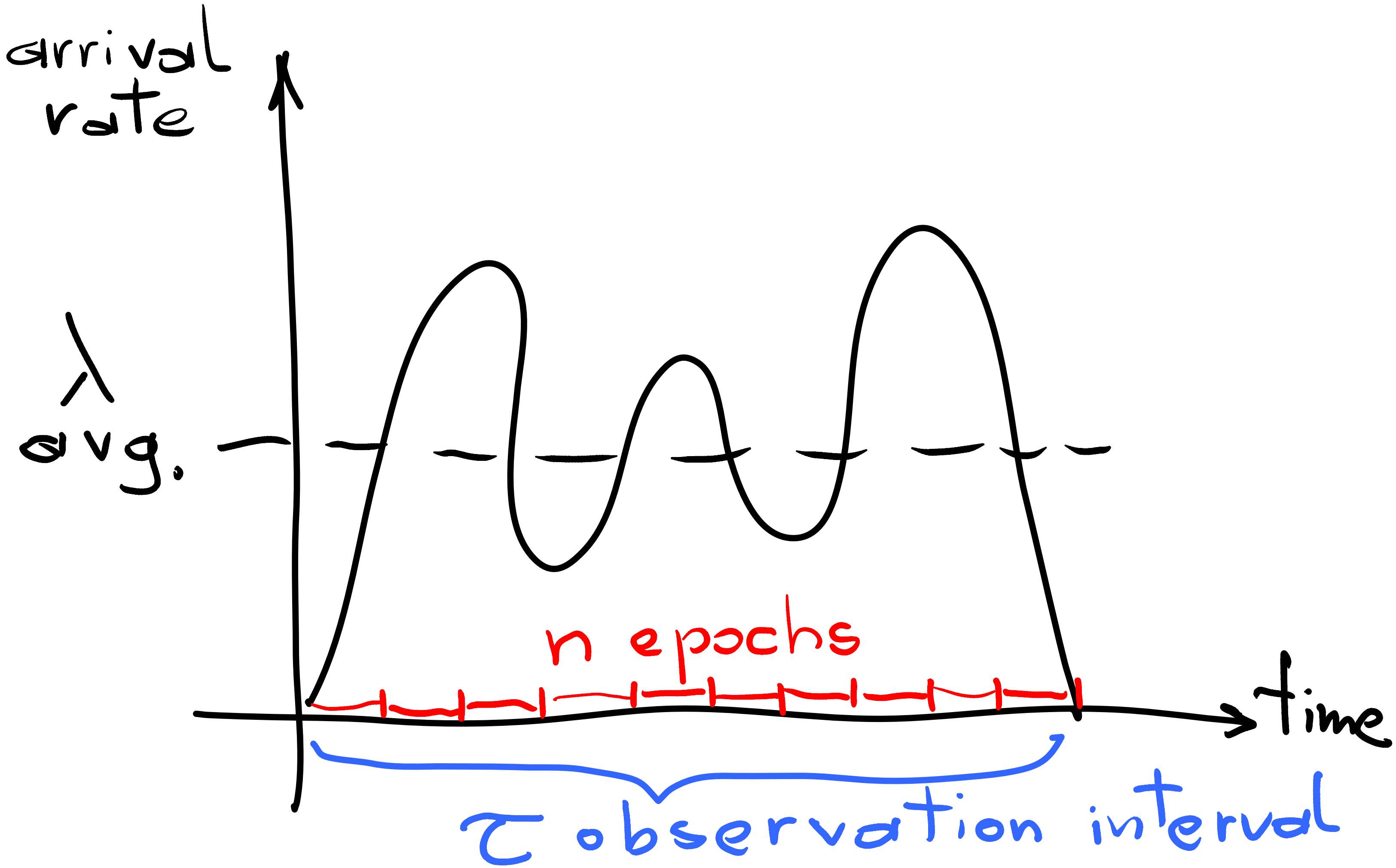
throughput, utilization

$$
\begin{aligned}
throughput \;\; res.1 &= \left(\frac{a}{2}+\frac{b}{2}\right) P_{11,00} + bP_{01,10} + aP_{10,01} \\
&= \frac{acd+bcd+abd+abc}{2cd+ad+bc+2ab} \\
utilization \;\; res.1 &= P_{01,10} + P_{10,01} + P_{11,00} \\
utilization \;\; res.2 &= P_{01,10} + P_{10,01} + P_{00,11}
\end{aligned}
$$

Burstiness modeling

(from: D.Menasce, V.Almeida, Web Performance, Prentice Hall, 1998, Chapter 10)

arrival behavior of HTTP requests



variables definition

- *L* number of HTTP requests of the log file
- *T* duration of the observation interval
- $\lambda = L/T$ average arrival rate of requests in *T*
- subdivide *T* into *n* subintervals (equal) of duration *T/n* referred to as epochs
- *Arr(n)*: number of requests that arrive in epoch *n*
- λ_n arrival rate of requests during epoch *n*

variables definition

$$\lambda = \frac{L}{T} \qquad \lambda_j = \frac{Arr(j)}{T/n} = \frac{n \, * \, Arr(j)}{T}$$

- Arr^+ (Arr^-) number of requests that arrive in epochs whose arrival rate is greater (lower) than the average arrival rate λ in T

$$Arr^+ = \sum_{\forall j \;\; \lambda_j > \lambda} Arr(j)$$

$$Arr^- = \sum_{\forall j \;\; \lambda_j \leq \lambda} Arr(j)$$

burstiness parameter *b*

$$L = Arr^{+} + Arr^{-}$$

$$b = \frac{n.\ of\ epochs\ for\ which\ \lambda_j > \lambda}{n}$$

- if traffic is not bursty (uniformly distributed over *T*) it is *b=0*

$$Arr(j) = \frac{L}{n}$$

$$\lambda_j = \frac{L/n}{T/n} = \frac{L}{T} = \lambda \quad \rightarrow \quad b = 0$$

above average arrival rate

$$\lambda^+ = \frac{Arr^+}{n.epochs\ \lambda_j > \lambda\ *\ epoch\ duration}$$

$$= \frac{Arr^+/n}{\frac{n.epochs\ \lambda_j > \lambda}{n}\ *\ \frac{T}{n}}$$

$$= \frac{Arr^+}{b\ *\ T}$$

parameter a

- ratio between the *above average* arrival rate and the average arrival rate λ in *T*

$$a = \frac{\lambda^+}{\lambda} = \frac{Arr^+}{b*T} \ \frac{1}{L/T} = \frac{Arr^+}{b*L}$$

- a and b are related, both can be used as indicator of burstiness (we will use b)

inflation of service demands to burstiness

- we want to capture the effects of bustiness on the performance
- it is kown that the maximum throughput is given by

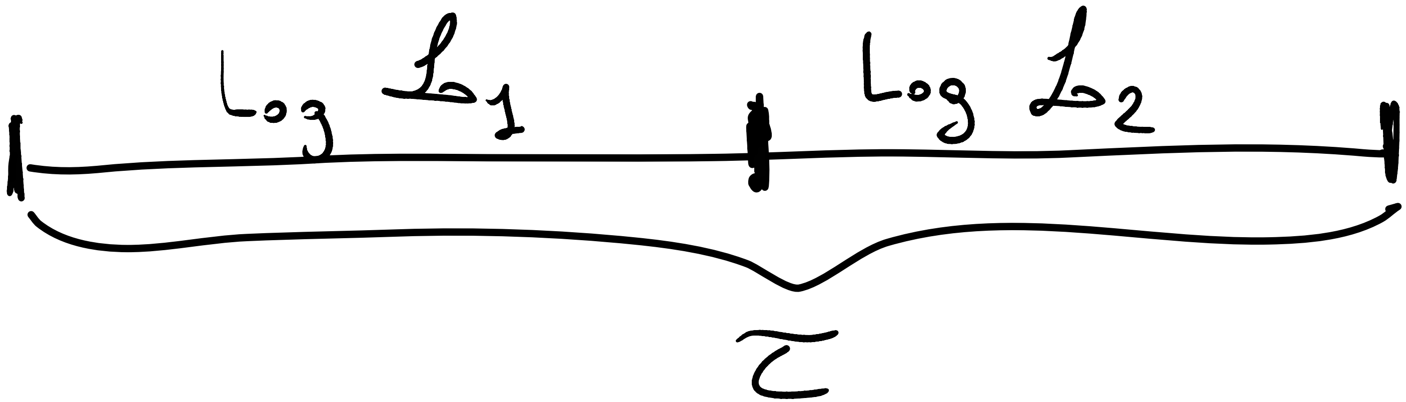
$$X_{max} = \frac{1}{D_{bott}}$$

- to account for the burstiness effect

$$D_{bott} = D_f + \alpha * b \qquad \alpha > 0$$

- D_f: portion of service demand that does not depend on burstiness, $\alpha * b$ is a term used to inflate the service demand of the bottleneck

computation of α



- two subintervals of T, T^1 and T^2
- U^1 , U^2 utilization of the bottleneck in the two subintervals T^1 , T^2
- X^1 , X^2 throughputs [$L^1/(T/2)$, $L^2/(T/2)$]

computation of α

$$U_i = X * D_i$$

$$\frac{U^1_{bott}}{X^1} = D_f + \alpha * b^1 \qquad \frac{U^2_{bott}}{X^2} = D_f + \alpha * b^2$$

- D_f :is assumed to be fairly homogeneous during the entire T
- b^1, b^2 : burstiness factors computed for L^1 , L^2

$$\alpha = \frac{\frac{U^1_{bott}}{X^1} - \frac{U^2_{bott}}{X^2}}{b^1 - b^2}$$