Performance of Networked Systems

Lecture 3: Performance models for streaming and elastic traffic

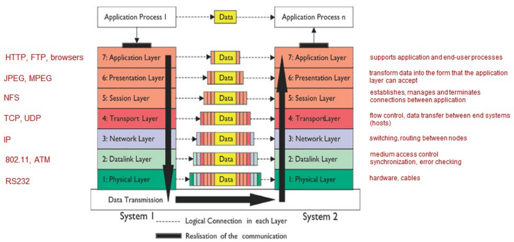
Overview of today's lecture

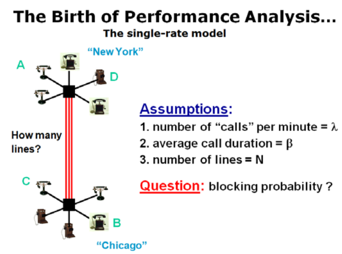
1. Streaming versus elastic traffic
2. Erlang blocking model (Erlang-B)
3. Multi-rate models and product-form solution
4. Kaufman/Roberts recursion
5. Elastic traffic: Processor Sharing models

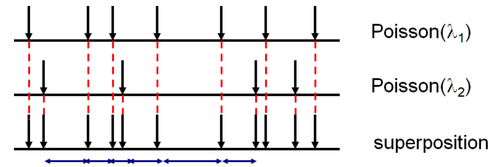
Background reading material:

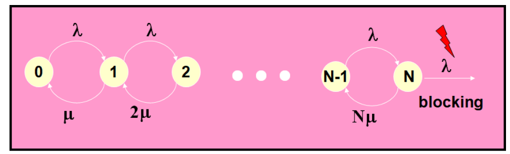
COST 242 - Multi-Rate Models for Dimensioning and Performance Evaluation of ATM Networks, Chapter 4 (only pages 17–25)

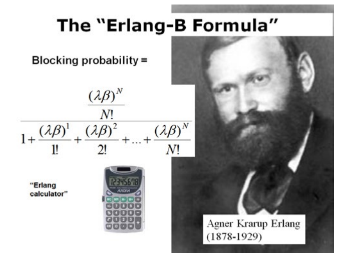
Wrap Up of Previous Lecture











What you should know:

1. Discrete and continuous probability distributions
2. Exponential distribution and memoryless property
3. Poisson processes, superposition and thinning properties
4. Poisson Arrivals See Time Averages (PASTA)
5. Poisson distribution and relation to Poisson process
6. Markov chains, equilibrium distributions
7. Erlang-B model

Background reading:

O.J. Boxma, Stochastic Performance Modelling, chapters 1, 2 and 3

Streaming versus Elastic Traffic

Streaming traffic:

- Typically requires fixed amount of bandwidth
- Examples:
  - circuit-switched voice telephone (fixed bandwidth)
  - streaming audio (bandwidth more or less constant)
  - streaming video
- **Key performance metric:** blocking probability

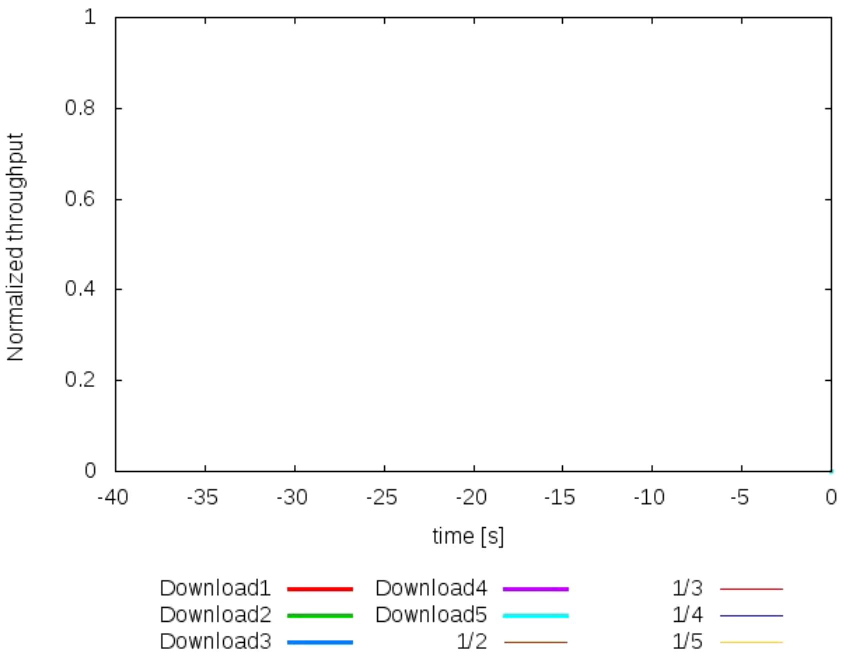
Network capacity is efficiently used by simultaneous transmission of stream traffic (speech, video) and elastic traffic (data). At time t the available bandwidth for data (ct) is divided among nt packets to be transmitted.

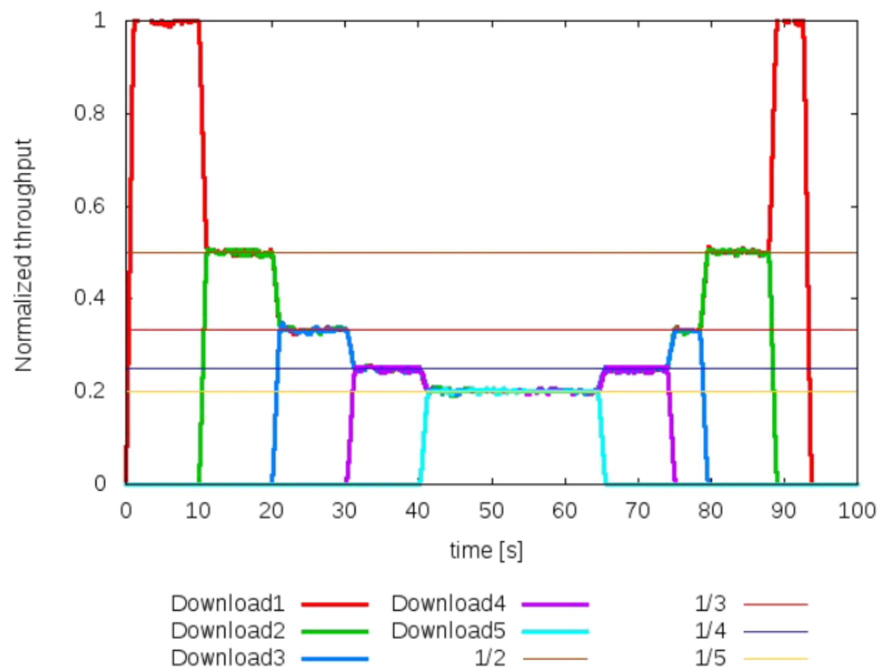
Elastic traffic:

- Mainly for data applications, not very delay-sensitive
- No fixed throughput required: **bandwidth sharing**
- Examples:
  - data transfer via TCP – usually ‘best-effort’ service
- **Key performance metrics:** throughput, transfer times

Elastic Traffic:
capacity sharing example

Experiment: **TCP file transfer over WLAN**

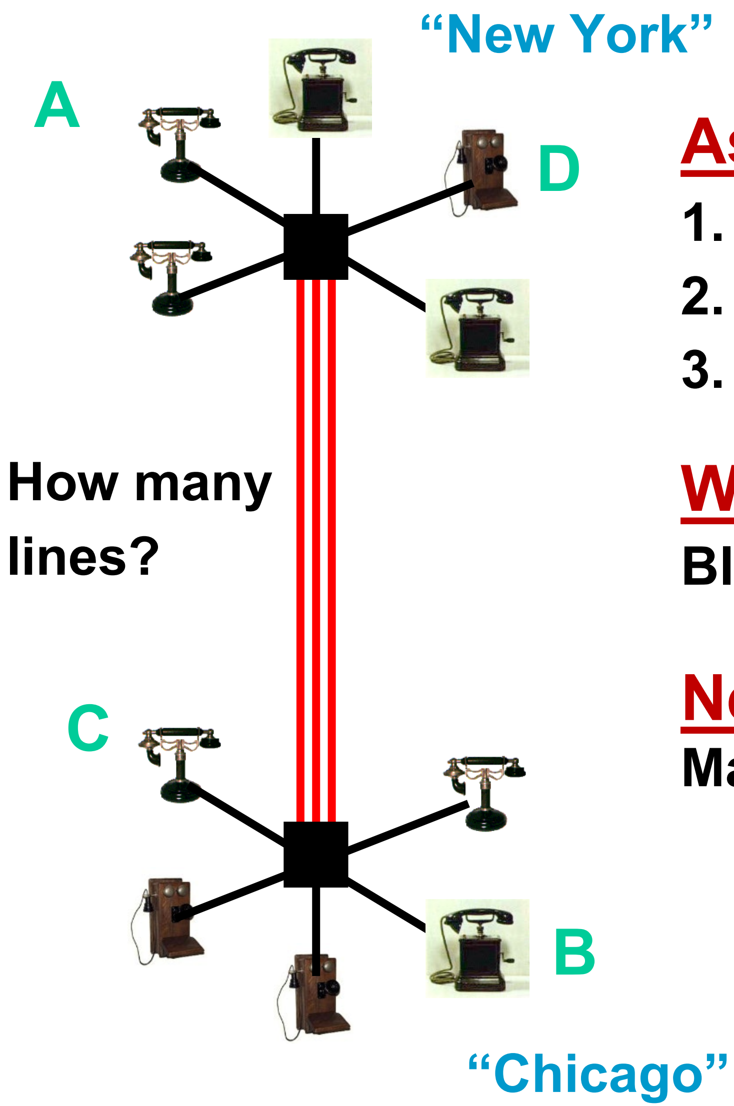




- **Flows enter network at a pre-defined time with a fixed size**
- **Called “elastic” traffic**
- **Bandwidth sharing behaves like Processor Sharing (PS) model**

The Erlang Blocking Model…

The single-rate model



Assumptions:

1. Poisson process of “calls” rate λ
2. average call duration = β
3. number of lines = N

What we want:

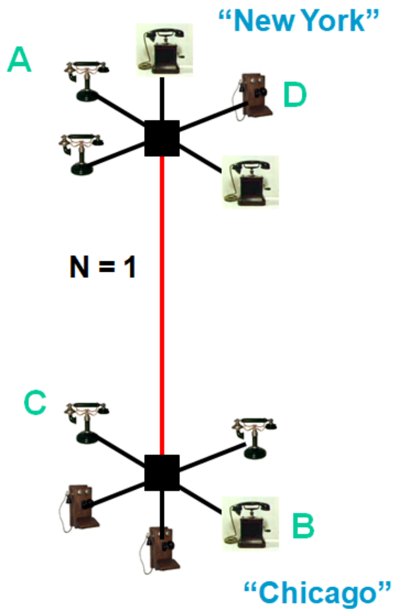
Blocking probability $p < \alpha$

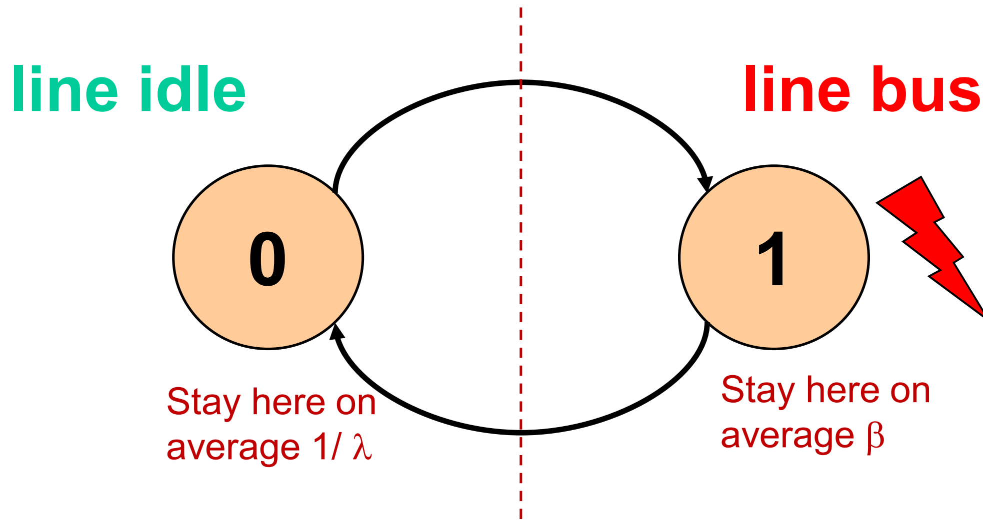
Needed:

Mathematical relation between

λ, β, N and p

Solution for N=1





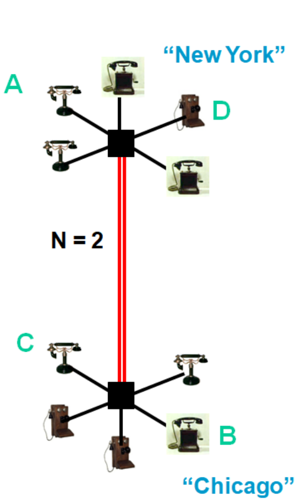
If λ = 1 and β = 2… then blocking probability = 2 / 3

If λ = 2 and β = 2… then blocking probability = 4 / 5

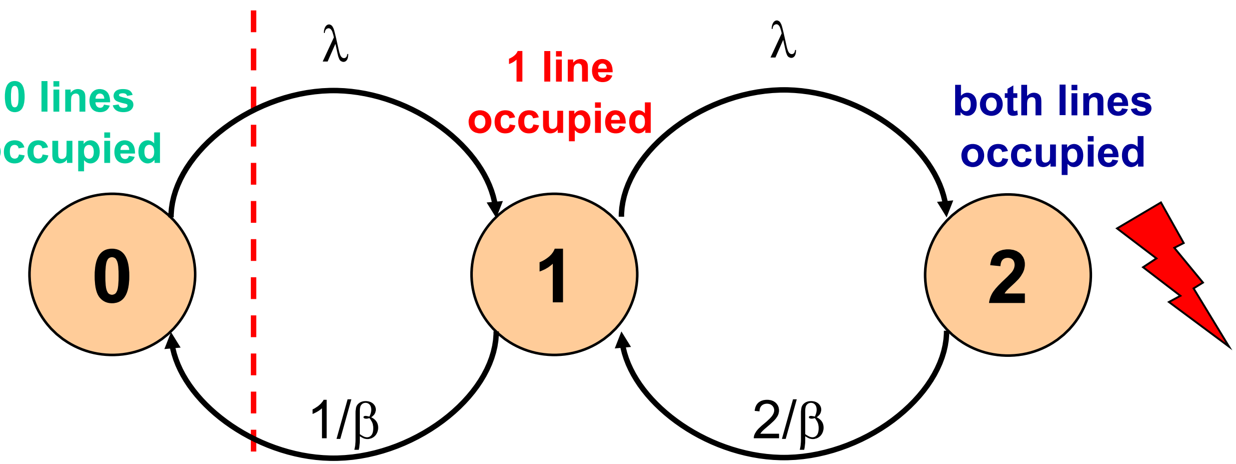
In general: $\frac{\pi_1}{\pi_0} = \frac{\beta}{1/\lambda} = \lambda\beta \Longleftrightarrow \lambda\pi_0 = (1/\beta)\,\pi_1$ and $\pi_0 + \pi_1 = 1$

PASTA!

Blocking probability = $\pi_1 = \frac{\beta}{1/\lambda + \beta} = \frac{\lambda\beta}{1+\lambda\beta}$



Solution for N=2



$\beta=1/\mu$

β is mean service time
μ is called 'service rate'

Balance equations:

$$(1/\beta)\pi_1 = \lambda\pi_0 \quad \Rightarrow \quad \pi_1 = (\lambda\beta)\times\pi_0$$

$$(2/\beta)\pi_2 = \lambda\pi_1 \quad \Rightarrow \quad \pi_2 = (\lambda\beta/2)\times\pi_1 = (\lambda^2\beta^2/2)\times\pi_0$$

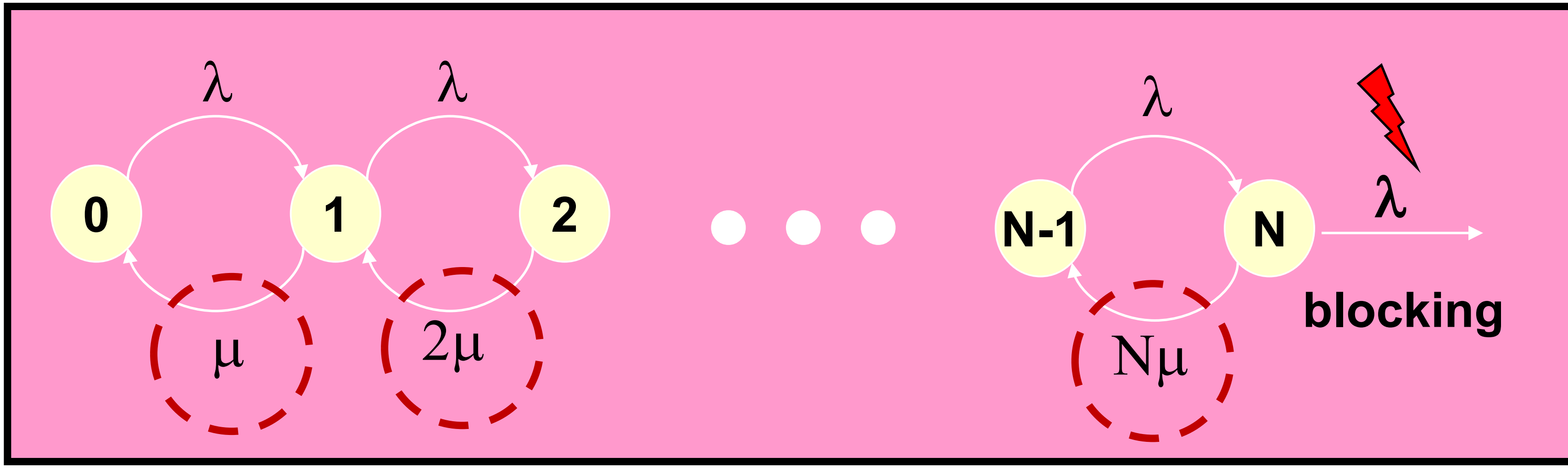
Normalisation:

$$\pi_0 + \pi_1 + \pi_2 = 1$$

PASTA!

Blocking probability = $\pi_2 = \dfrac{(\lambda\beta)^2/2}{1+\lambda\beta+(\lambda\beta)^2/2}$

Solution for General N



- N(t) := # busy lines at time t (t > 0)
- Continuous-time Markov chain { N(t), t > 0 }
- State space S := {0,1,…,N}
- Called **birth-and-death process**
- Transition rates:

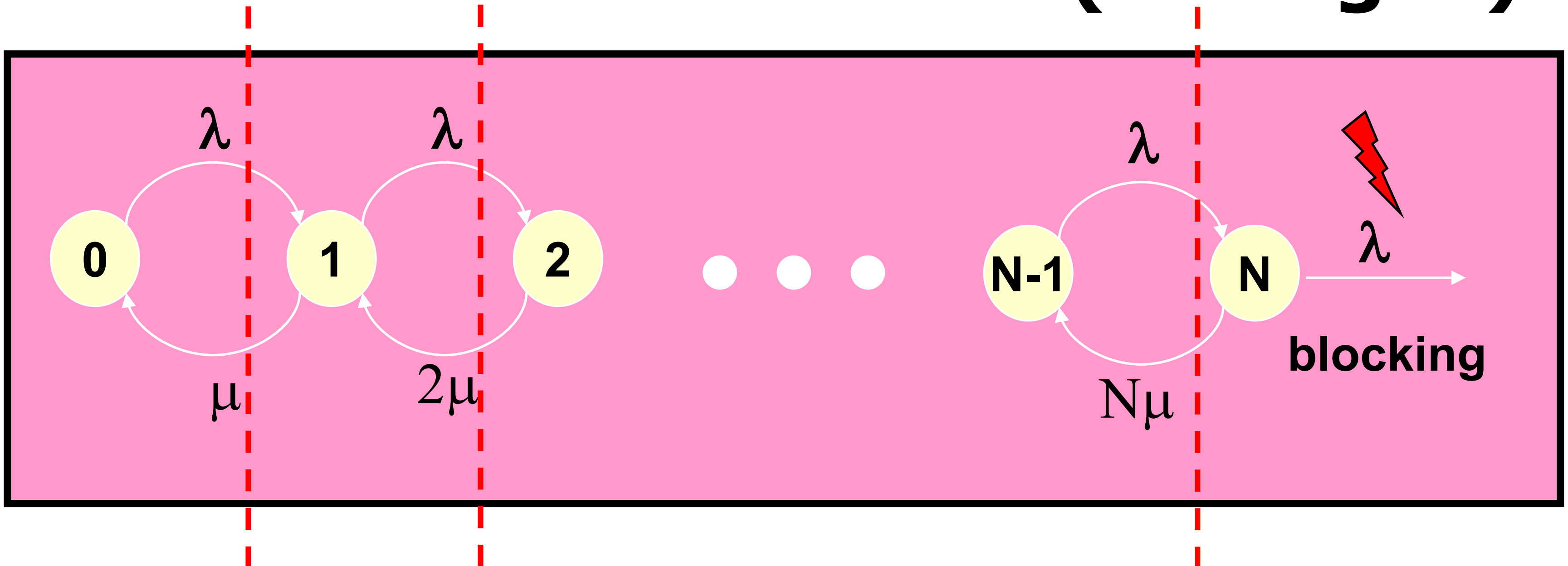
β=1/μ

β is mean service time
μ is called 'service rate'

Prob { i → i + 1 in (t; t + Δt] } = λΔt + o(Δt), for Δt ↓ 0 (arrivals) – jumps to right

Prob { i → i – 1 in (t; t + Δt] } = i□ t + o(Δt), for Δt ↓ 0 (departures) – jumps to left

Solution for General N (Erlang-B)



- **Equilibrium distribution (also called "stationary distrbution")**

  $\pi_k := \lim_{t \to \infty}$ Prob{ N(t) = k } (k = 0, 1,…, N)

- **Balance equations ("rate in = rate out")**
  1. $\lambda\pi_0 = \mu\pi_1, \dots, \lambda\pi_{N-1} = \mu\pi_N$
  2. $\pi_0 + \cdots + \pi_N = 1$ (normalisation)
  3. Recursively calculate $\pi_0 \Rightarrow \pi_1 \Rightarrow \dots \Rightarrow \pi_N$, and then normalize

Erlang Blocking Formula

Solution of Markov chain:

$$\pi_k = \frac{(\lambda\beta)^k / k!}{\sum_{i=0}^{N} (\lambda\beta)^i / i!} \quad (k = 0,1,\ldots,N)$$

Agner Krarup Erlang (1878-1929)

- **Use PASTA: Poisson Arrivals See Time Averages**
- **Blocking probability:**

  Prob { call blocked } = π_N

Insensitivity Property: formula also valid for non-exponential call holding times

The “Erlang-B Formula”

Blocking probability =

$$\frac{\dfrac{(\lambda\beta)^N}{N!}}{1+\dfrac{(\lambda\beta)^1}{1!}+\dfrac{(\lambda\beta)^2}{2!}+\ldots+\dfrac{(\lambda\beta)^N}{N!}}$$

“Erlang calculator”

Agner Krarup Erlang (1878-1929)

Example for “Erlang-B Formula”

Erlang-B formula:

Blocking probability = $$\frac{\frac{(\lambda\beta)^N}{N!}}{1+\frac{(\lambda\beta)^1}{1!}+\frac{(\lambda\beta)^2}{2!}+\ldots+\frac{(\lambda\beta)^N}{N!}}$$

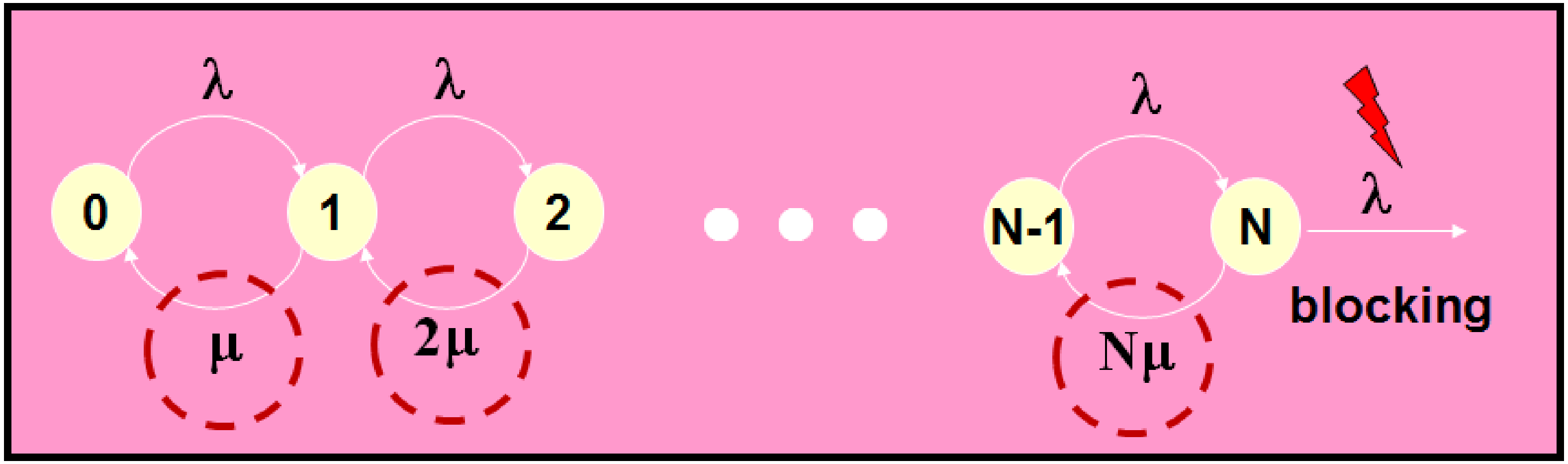
Assumptions:

- **Call arrival rate is $\lambda = 2$ calls per minute**
- **Mean call duration $\beta = 3$ minutes**
- **Number of channels *N=4***

Blocking probability:

$$\frac{\frac{(2.3)^4}{4!}}{1+\frac{(2.3)^1}{1!}+\frac{(2.3)^2}{2!}+\frac{(2.3)^3}{3!}+\frac{(2.3)^4}{4!}} = \frac{54}{1+6+18+36+54} \approx 0.46956$$

Solving Blocking Probabilities



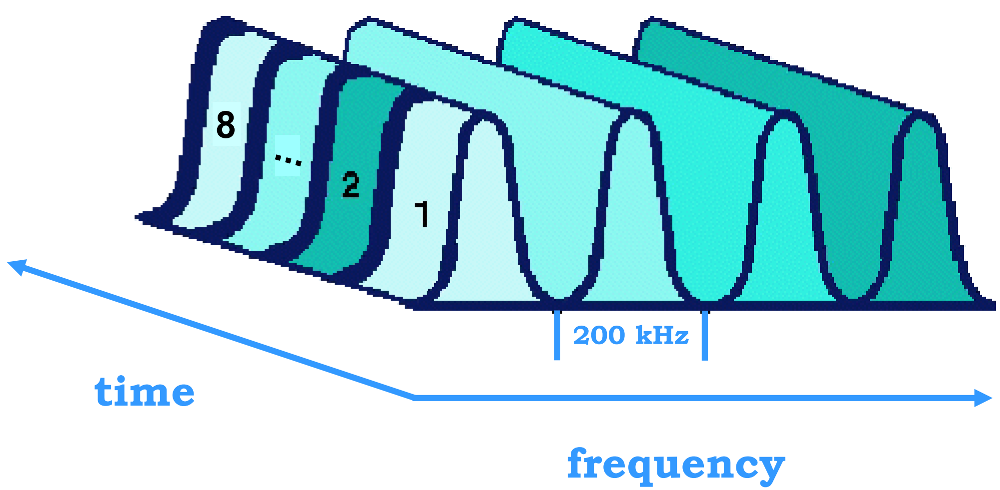
Stepwise approach

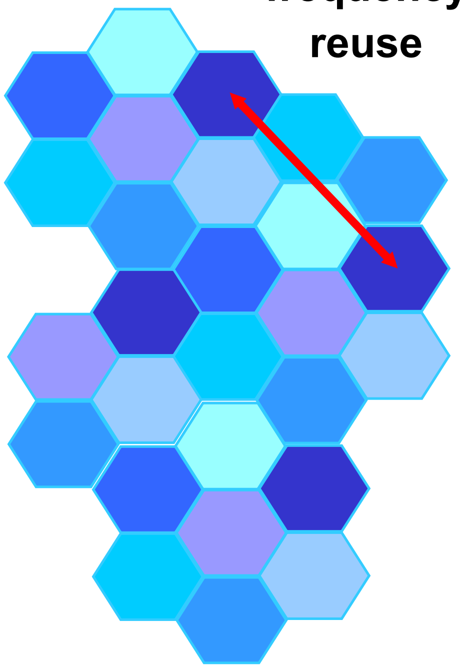
1. Define Markov Chain to describe the system
2. Write down balance equations
3. Determine steady-state distribution $\underline{\pi} = (\pi_0, \ldots, \pi_N)$
4. Use PASTA
5. Express blocking probabilities in terms of $\underline{\pi}$

Global System for Mobile communications (GSM)

GSM radio interface basics

- 2 × 25 MHz ⇒ [FDMA] ⇒ 2 × 125 frequencies
- each carrier ⇒ [TDMA] ⇒ 8 “channels”
- planning is done on the basis of frequencies



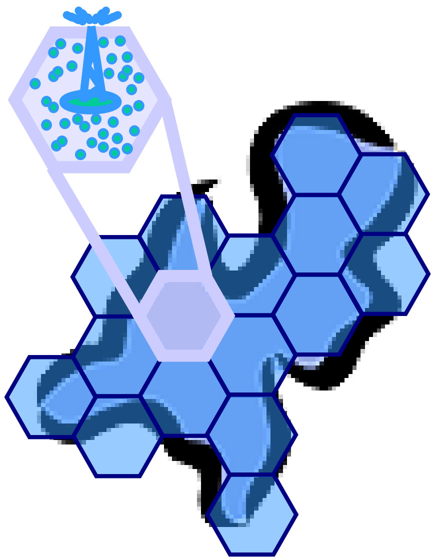


neighbouring cells may interfere

Key performance model: Erlang-B model

Capacity Planning of Cellular Voice Networks

few base stations, large cells



many base stations, small cells

Question:

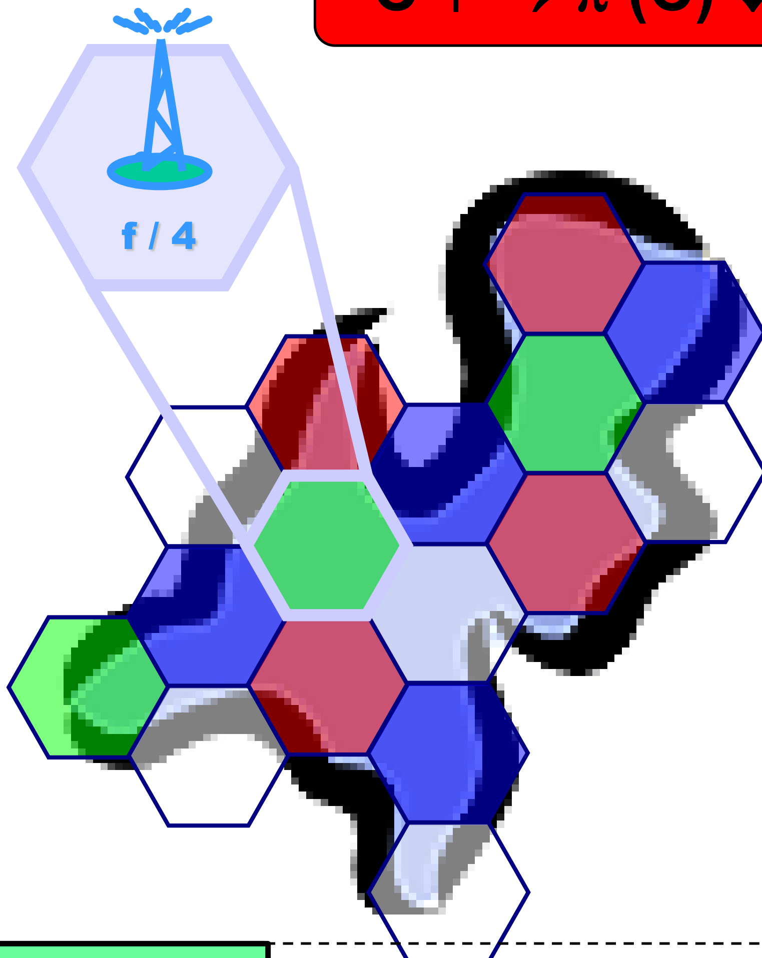
How many base stations do we need such that blocking probability less than some target value α?

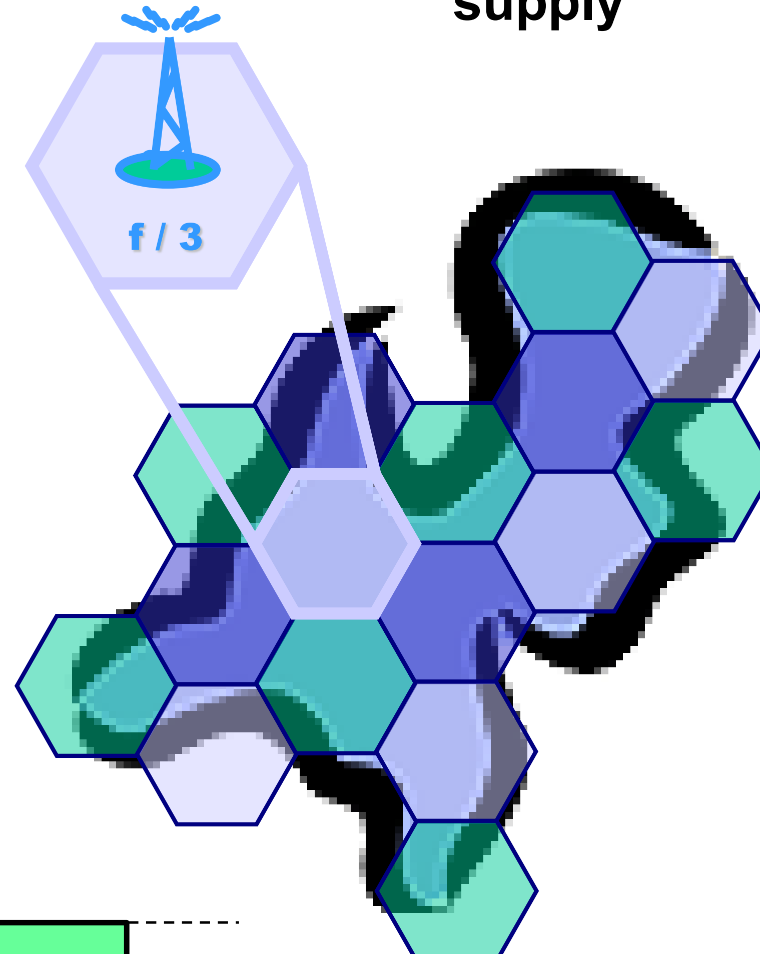
1

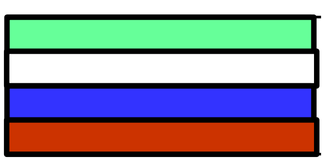
Denser Frequency Reuse

$C \uparrow \rightarrow \pi(C) \downarrow$

"more capacity supply"







frequency band divided in 4 parts

frequency band divided in 3 parts

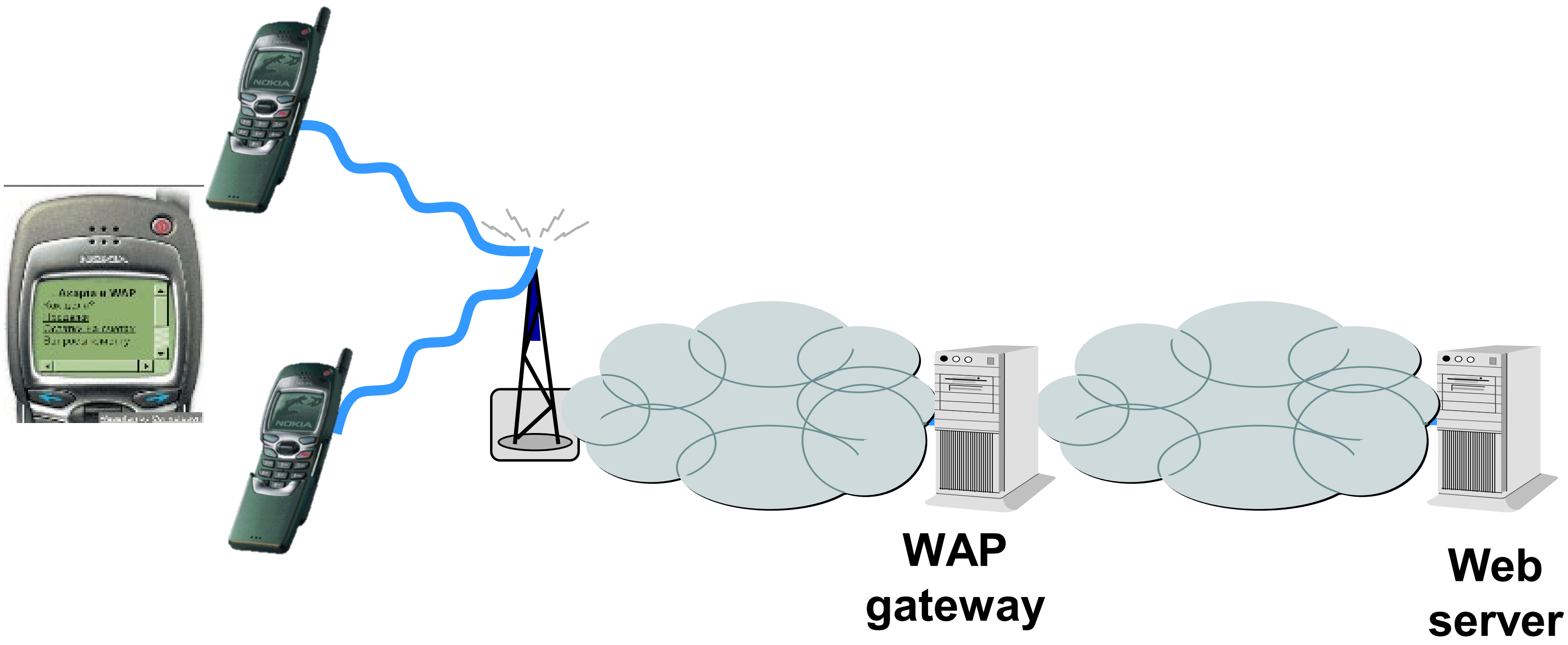
2

Cellular Densification

$\rho \downarrow \rightarrow \pi(C) \downarrow$

"demand per cell goes down"

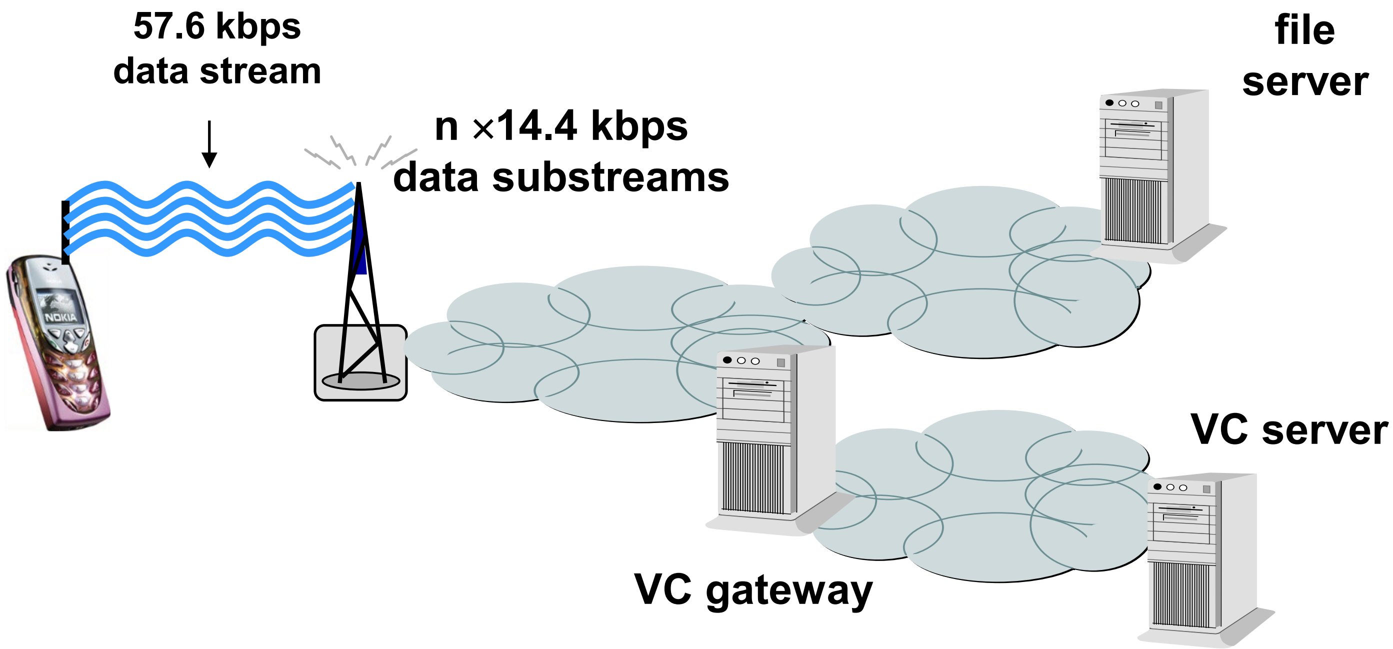
Circuit Switched Data (CSD)

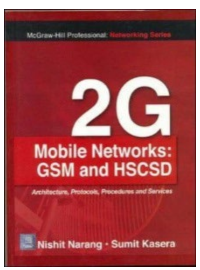


GSM data

- Net data rate varies: 2.4, 4.8, 9.6, 14.4, 22.8 kbps
- Data rate depends on choice of the codec
- Often called "First generation mobile data services"
- Wireless Access Protocol (WAP)

High Speed Circuit Switched Data

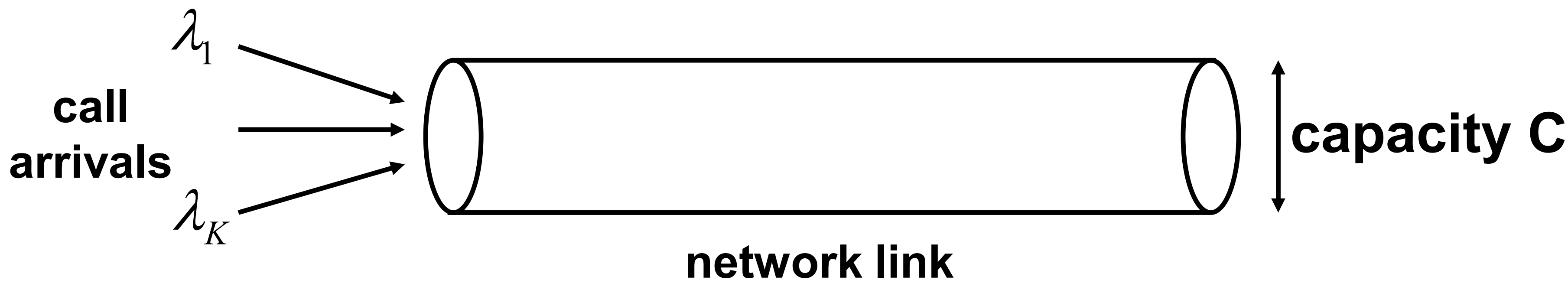




High Speed Circuit-Switched Data (for higher data dates)

- bundling of traffic channels: adjacent multi-channel transmission
- attainable data rates n x 14.4 kbps (n up to 8)
- services: “large” file transfer, video-conferencing

Multi-Rate Models



Network dimensioning problem:

How much capacity is needed, such that blocking probabilities are small enough?

Multi-rate model:

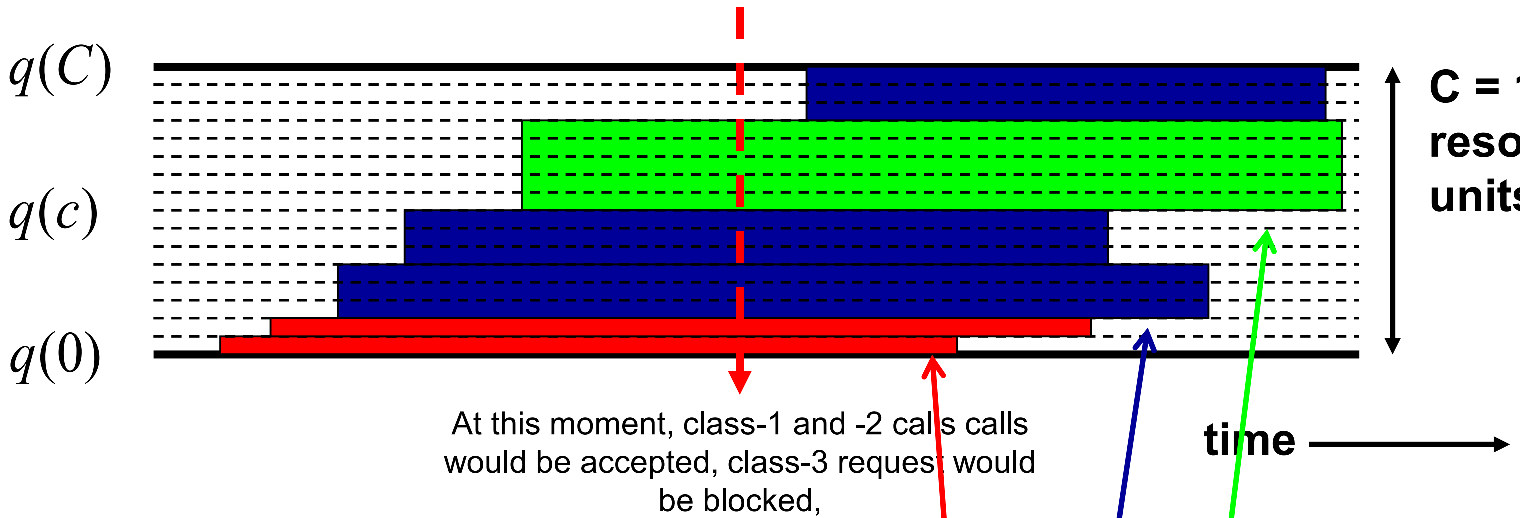
- K call classes
- Poisson call arrivals rate $\lambda_1, \lambda_2, \ldots, \lambda_K$
- required capacity ("effective bandwidth"): $b_1, b_2, \ldots, b_K$
- mean call duration $d_1, d_2, \ldots, d_K$ (or: service rates $\mu_1, \mu_2, \ldots, \mu_K$)
- number of class-k calls in progress (k = 1,…,K)
- $n = (n_1, \ldots, n_K)$ state of the system

capacity used when system is in state n

Blocking: class-k call is accepted if and only if there is room: $\sum_{i=1}^{K} b_i n_i \le C - b_k$

capacity available

Multi-Rate Model

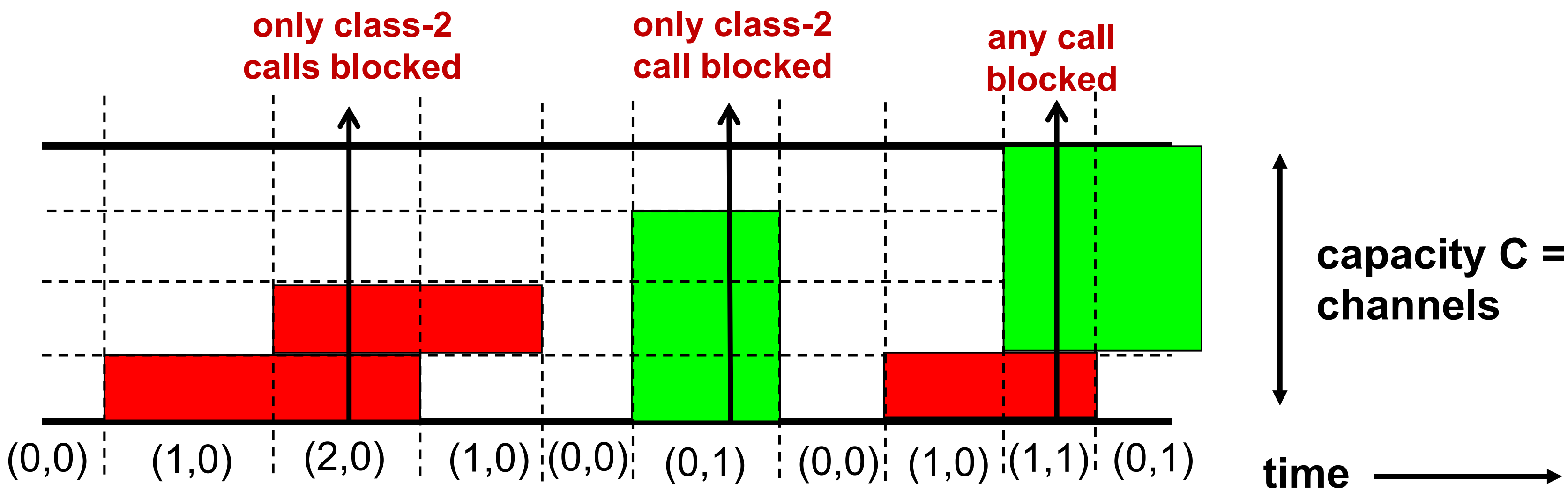


Model example:

- K = 3 call classes
- Poisson call arrivals rate λ_k (k = 1, 2, 3)
- required capacity ("effective bandwidth"): $b_1=1$, $b_2=3$, $b_3=5$
- mean call duration $d_k = 1/\mu_k$ (k = 1, 2, 3)
- load $\rho_k = \lambda_k / \mu_k$ (k = 1, 2, 3)
- n_k := number of class-k calls in progress (k = 1, 2, 3)

Blocking: class-k call is accepted if and only if $\sum_{i=1}^{K} b_i n_i \leq C - b_k$

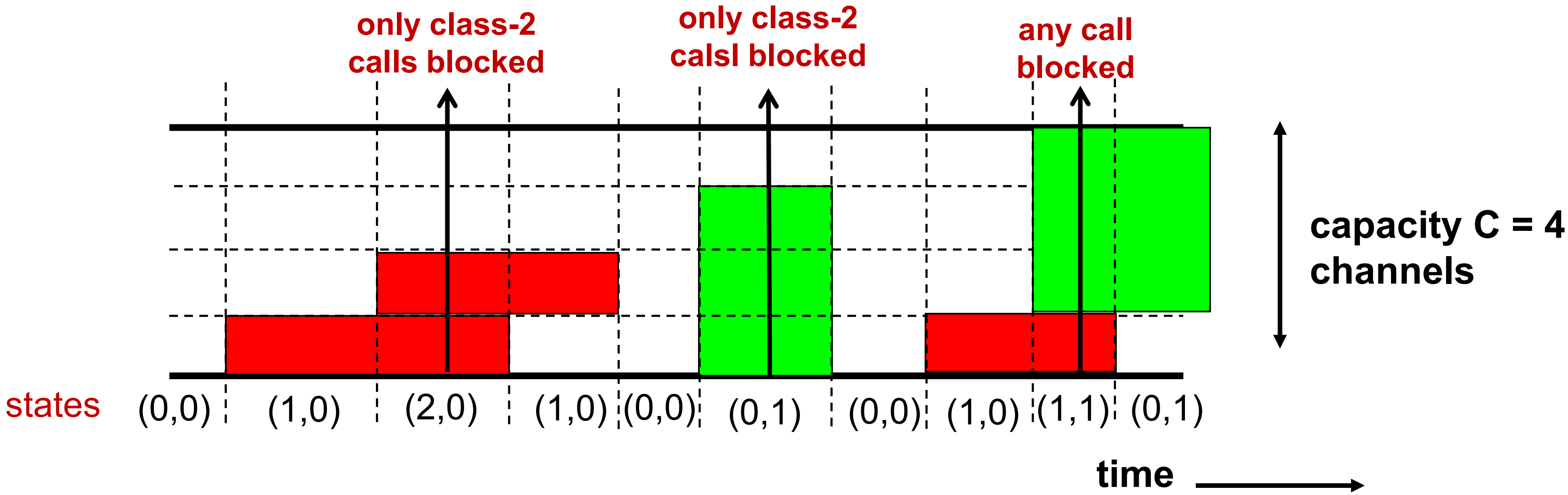
Multi-Rate Model



Model:

- K = 2 call classes, capacity C = 4
- Poisson call arrivals rate λ_1 , λ_2
- required capacity ("effective bandwidth"): $b_1 = 1$, $b_2 = 3$
- state (n_1 , n_2), n_1 := number of class-1 calls, n_2 := number of class-2 calls
- state space S = {(0,0), (1,0), (2,0), (3,0), (4,0), (0,1), (1,1)}
- class-1 call is accepted only if system in states (0,0), (1,0), (2,0), (3,0) or (0,1)
- class-2 call is accepted if and only if system in state (0,0) or (1,0)

Multi-Rate Model



Questions:

1. What is call blocking probability for class-1 calls?
2. What is call blocking probability for class-2 calls?

Solution approach:

1. Formulate two-dimensional Markov chain
2. Calculate the state probabilities
3. Determine blocking probabilities (using PASTA)

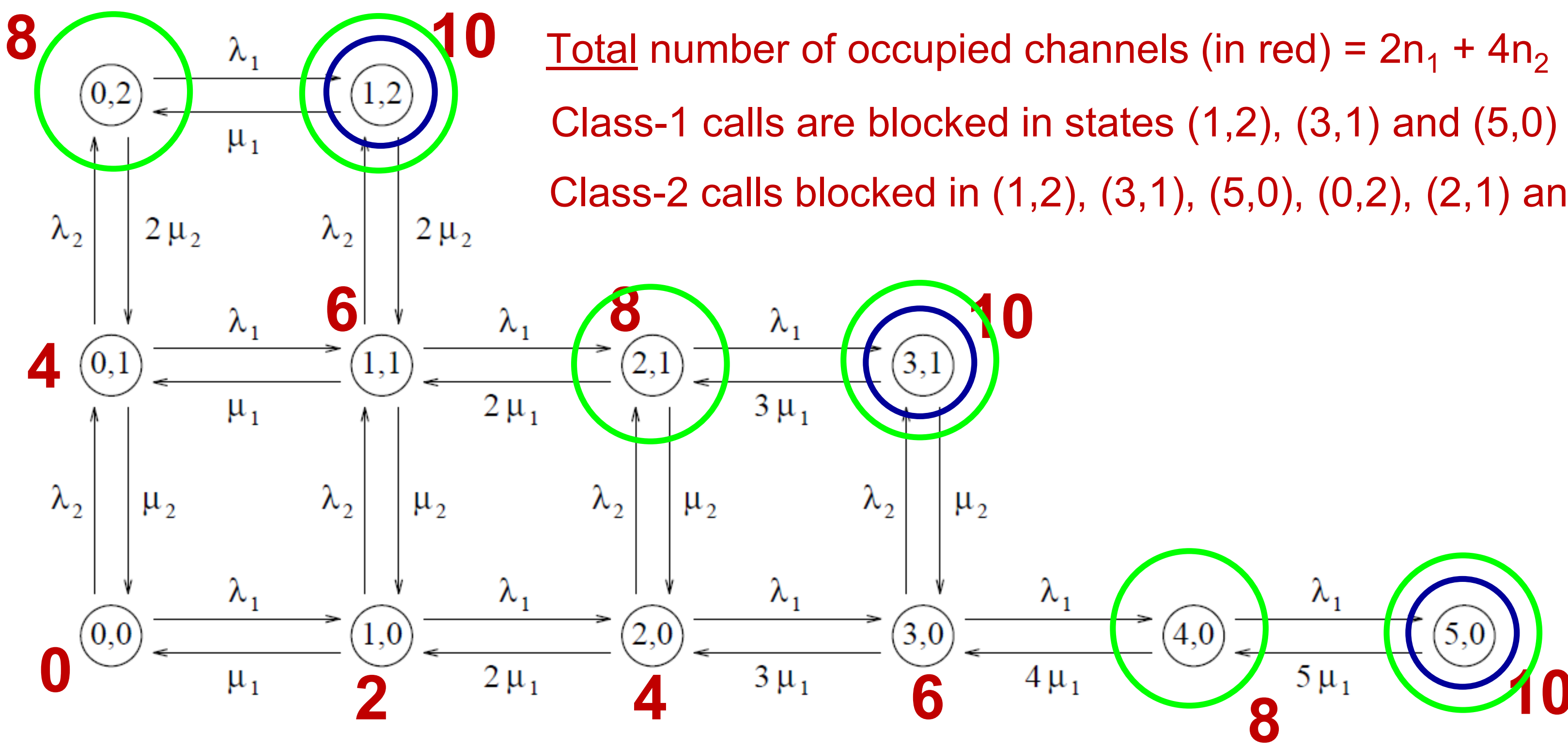
Another Example

Assume: K = 2 classes, C = 10 channels, b_1 = 2, b_2 = 4

Total number of occupied channels (in red) = $2n_1 + 4n_2$

Class-1 calls are blocked in states (1,2), (3,1) and (5,0)

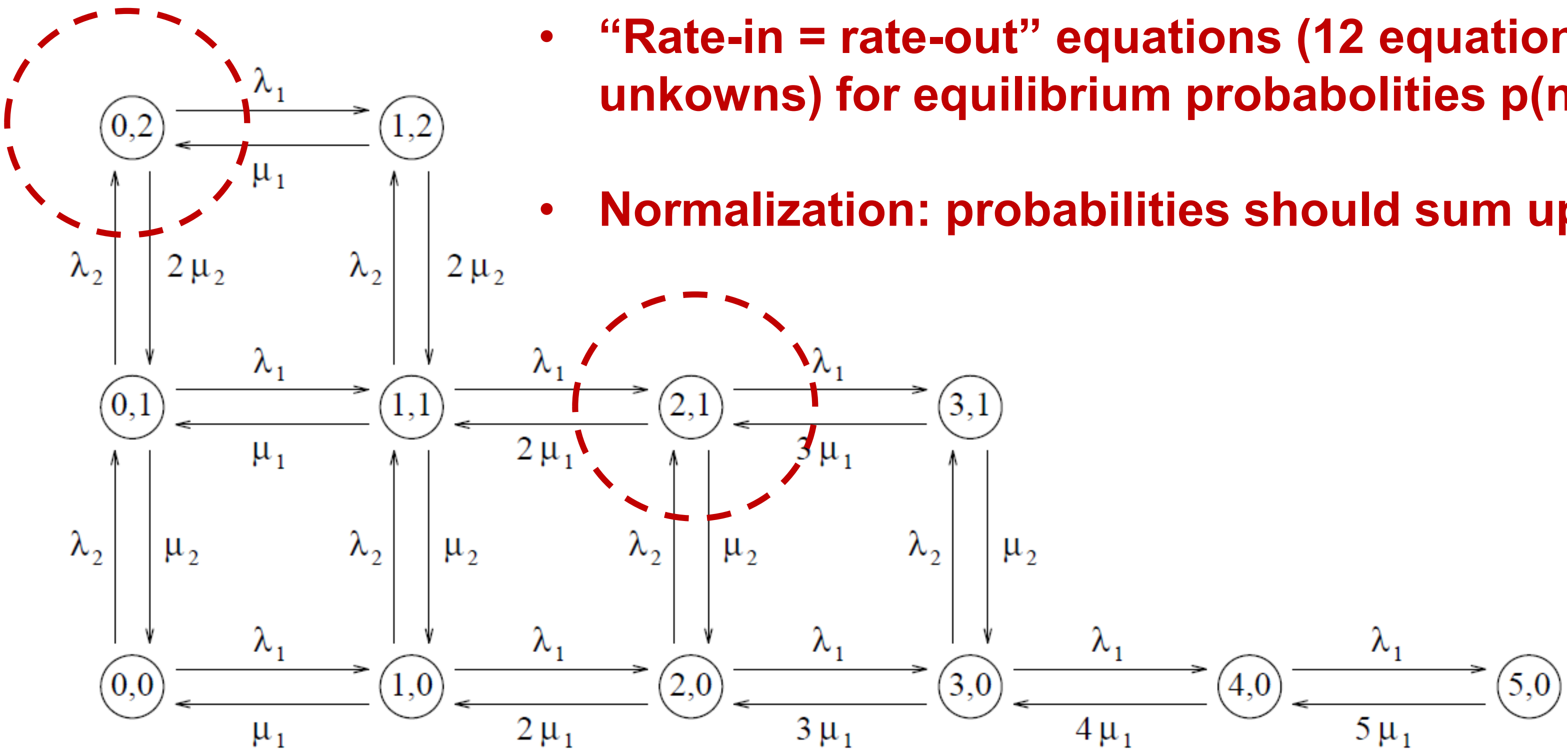
Class-2 calls blocked in (1,2), (3,1), (5,0), (0,2), (2,1) and (4,0)



- State space S = {(0,0), (0,1),…, (5,0)}   (12 states)
- Transition rates
- $\pi(n_1,n_2)$ = probability that system is in state (n_1,n_2)

Another Example

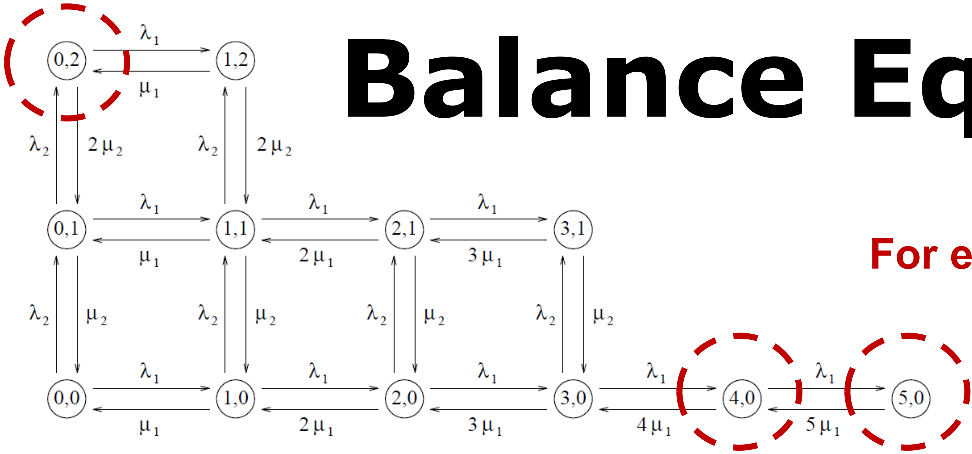
Case: K = 2, C = 10, b_1 = 2, b_2 = 4



- **"Rate-in = rate-out" equations (12 equations, 12 unkowns) for equilibrium probabolities p(n_1,n_2)**
- **Normalization: probabilities should sum up to 1**

- state space S = {(0,0), (0,1),…, (5,0)}   (12 states)
- transition rates
- π(n_1,n_2) = probability that system is in state (n_1,n_2), for all states

Balance Equations



For each state: "rate in = rate out"

For state (0,2): $(\lambda_1 + 2\mu_2)\,\pi(0,2) = \mu_1\pi(1,2) + \lambda_2\pi(0,1)$

Do this for each of the 12 states

For state (4,0): $(\lambda_1 + 4\mu_1)\,\pi(4,0) = \lambda_1\ (3,0) + 5\mu_1\pi(5,0)$

For state (5,0): $5\mu_1\pi(5,0) = \lambda_1\pi(4,0)$

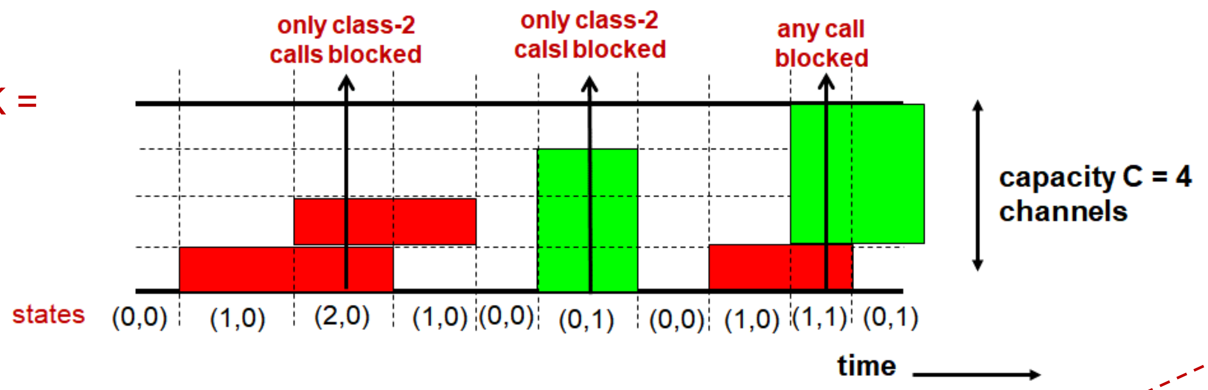
12 equations

Normalization: $\pi(0,0) + \pi(0,1) + \cdots + \pi(5,0) = 1$

13 equations, 12 unknowns, but <u>unique</u> solution (throw away 1 equation)

Analysis of Multi-Rate Model

Example with K = 2 and C = 4



$I^K = \mathbb{N}_0^K$

K-dimensional Markov chain with states $n = (n_1, \ldots, n_K)$

State space $S := \{n = (n_1, \ldots, n_K) \in I^K : \sum_{k=1}^{K} b_k n_k \le C\}$

S= set of states n for which total number of occupied channels ≤ C

Theorem: the equilibrium distribution is given by

$$\pi(n) = \pi(n_1, \ldots, n_K) = \frac{1}{G} \prod_{j=1}^{K} \frac{\rho_j^{n_j}}{n_j!}, \text{ for } n \in S$$

"product-form solution"

where

$$G := \sum_{n \in S} \prod_{j=1}^{K} \frac{\rho_j^{n_j}}{n_j!}, \text{ with } \rho_j := \lambda_j d_j$$

"normalizing constant"

Analysis of Multi-Rate Model

Example with K = 2 and C = 4

$$S_j := \{n \in S : \sum_{k=1}^{K} b_k n_k \le C - b_j\}$$

S_j = subset of states in which class-j calls are accepted

<u>Blocking probability</u> for each class-j is given by:

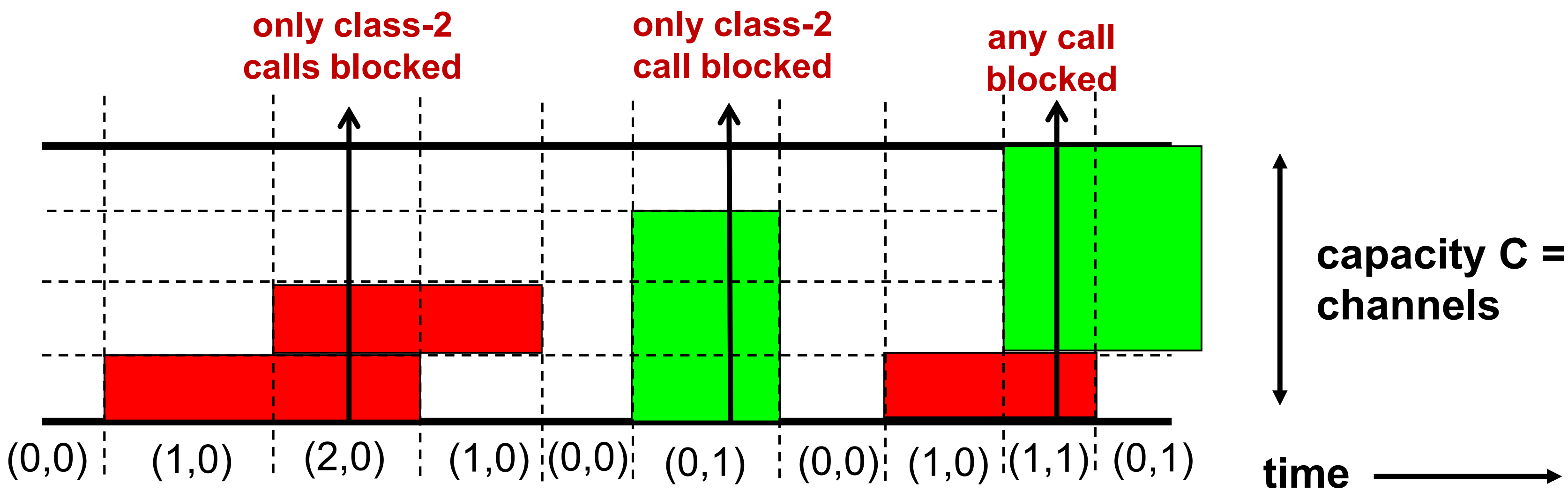
sum of probabilities of states in which class-j calls are <u>not</u> accepted

$$B_j = 1 - \sum_{n \in S_j} \pi(n) = 1 - \frac{\sum_{n \in S_j} \prod_{i=1}^{K} \rho_i^{n_i} / n_i!}{\sum_{n \in S} \prod_{i=1}^{K} \rho_i^{n_i} / n_i!}$$

$\pi(n)$ are the known probabilities from product form expression

<u>BUT</u>: numerical evaluation may be a problem due to large state spaces S_j and S! Efficient algorithms needed for calculating blocking probabilities!

Multi-Rate Model



Model:

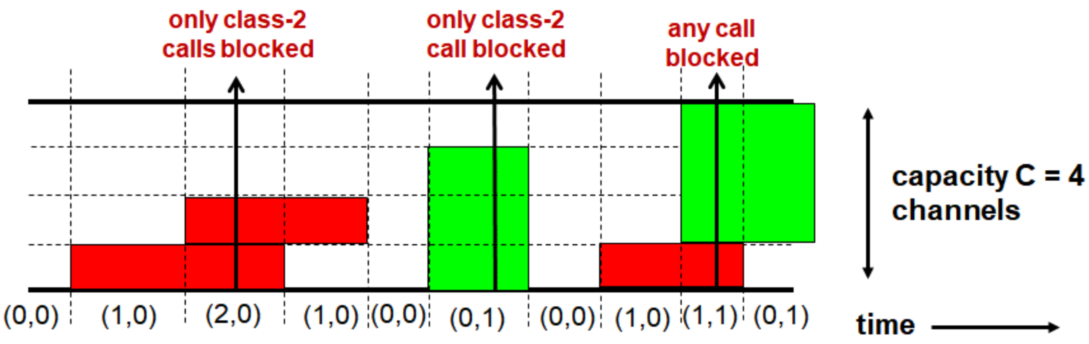
- K = 2 call classes, capacity C = 4
- Poisson call arrivals rate λ_1 , λ_2
- required capacity (“effective bandwidth”): b_1 = 1, b_2 = 3
- state (n_1 , n_2), n_1 := number of class-1 calls, n_2 := number of class-2 calls
- state space S = {(0,0), (1,0), (2,0), (3,0), (4,0), (0,1), (1,1)}
- class-1 call is accepted only if system in states (0,0), (1,0), (2,0), (3,0) or (0,1)
- class-2 call is accepted if and only if system in state (0,0) or (1,0)

Product-form Solution

Example:

$\mathbf{K = 2}$ **classes**

$\mathbf{b_1 = 1, b_2 = 3}$



State space: $S = \{(0,0), (1,0), (2,0), (3,0), (4,0), (0,1), (1,1)\}$

Product-form solution:

$$\pi(0,0) = \frac{1}{G} \cdot \frac{\rho_1^0}{0!} \frac{\rho_2^0}{0!} = \frac{1}{G}$$

$$\pi(1,0) = \frac{1}{G} \cdot \frac{\rho_1^1}{1!} \frac{\rho_2^0}{0!} = \frac{\rho_1}{G}$$

$$\vdots$$

$$\pi(4,0) = \frac{1}{G} \cdot \frac{\rho_1^4}{4!} \frac{\rho_2^0}{0!} = \frac{\rho_1^4}{24G}$$

$$\pi(0,1) = \frac{1}{G} \cdot \frac{\rho_1^0}{0!} \frac{\rho_2^1}{1!} = \frac{\rho_2}{G}$$

$$\pi(1,1) = \frac{1}{G} \cdot \frac{\rho_1^1}{1!} \frac{\rho_2^1}{1!} = \frac{\rho_1 \rho_2}{G}$$

G = normalizing constant, so

$$G = \frac{1}{1 + \rho_1 + \cdots + \frac{\rho_1^4}{24} + \rho_2 + \rho_1 \rho_2}$$

Class-1 accepted in states in $S_1 := \{(0,0), (1,0), (2,0), (3,0), (0,1)\}$

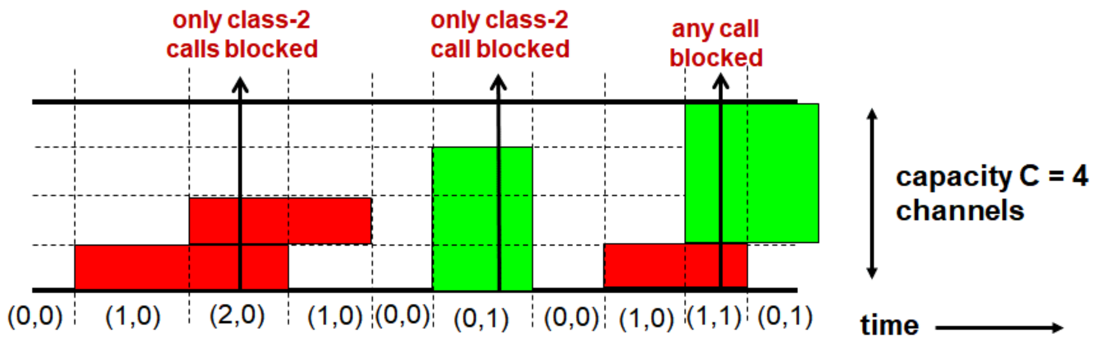
Class-2 accepted in states in $S_2 := \{(0,0), (1,0)\}$

Product-form Solution

Example:

K = 2 classes

$\mathbf{b_1 = 1, b_2 = 3}$



State space:

$$S = \{(0,0), (1,0), (2,0), (3,0), (4,0), (0,1), (1,1)\}$$

Blocking probabolities:

Class-1 accepted in states in $S_1 := \{ (0,0), (1,0), (2,0), (3,0), (0,1)\}$

Class-1 blocked in states in $S/S_1 := \{(4,0), (1,1)\}$

Blocking probability for class-1 = $\pi(4,0) + \pi(1,1)$

All these probabilities were calculated from the product-form solution on previous slide

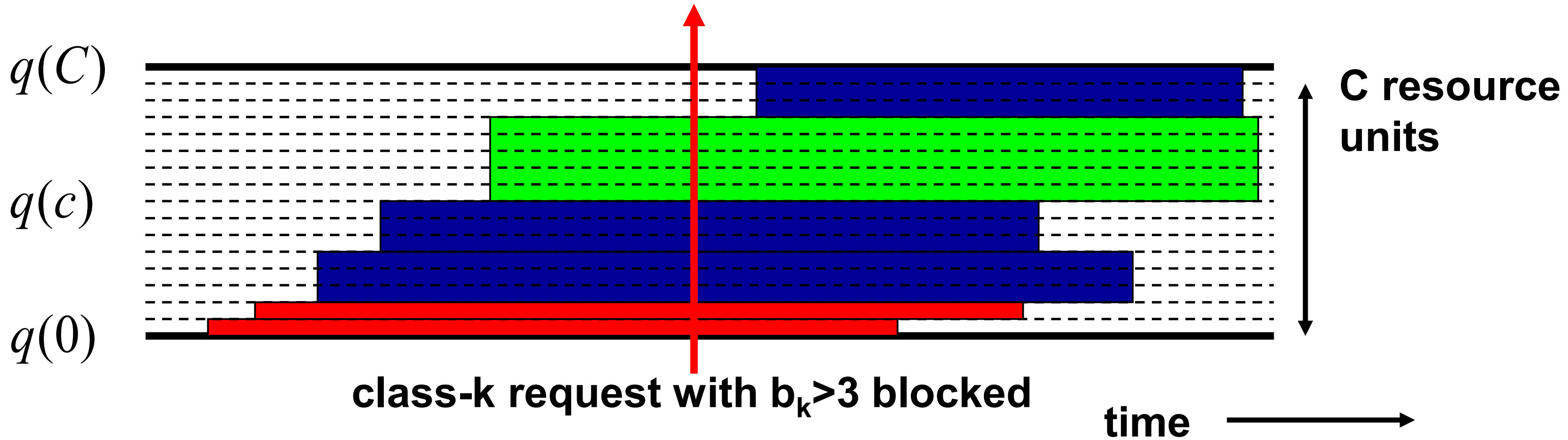
Class-2 accepted in states in $S_2 := \{(0,0), (1,0)\}$

Class-2 blocked in states in $S/S_2 := \{(2,0), (3,0), (4,0), (0,1), (1,1)\}$

Blocking probability for class-2 = $\pi(2,0) + \pi(3,0) + \pi(4,0) + \pi(0,1) + \pi(1,1)$

All these probabilities were calculated from the product-form solution on previous slide

Multi-Rate Models: Kaufman-Roberts Recursion



Define: $S(c) := \left\{ n \in S : \sum_{k=1}^{K} b_k n_k = c \right\}$ for $0 \le c \le C$

S(c) is the set of states where in total c channels are occupied

$$q(c) := \sum_{n \in S(c)} \pi(n)$$

probability that in total exactly c channels are occupied

Theorem: the occupancy probabilities q(c) satisfy

$$q(c) = \frac{1}{c} \sum_{k=1}^{K} \rho_k b_k q(c - b_k) \qquad (c = 1, \ldots, C)$$

Example for Kaufmann-Roberts

Assumptions:

- K = 10 classes, link capacity C = 100
- Call arrival rate for each classes is 1: $\lambda_1 = \lambda_2 = \ldots \lambda_{10} = 1$
- Mean call holding times $d_1=10$, $d_2=10/2=5$, $d_3=10/3$, $d_4=10/4=2.5$, … , $d_{10}=10/10=1$
- Per-class capacity requirements $b_1=1$, $b_2=2$, … , $b_{10}=10$

Intermediate steps:

- Service rates: $\mu_1 = 1/d_1 = 1/10$, $\mu_2 = 1/d_2 = 2/10$, $\mu_3 = 3/10$, … , $\mu_{10} = 10/10 = 1$
- Per-class load factors: $\rho_1 = \lambda_1/\mu_1 = 10$, $\rho_2 = \lambda_2/\mu_2 = 5$, …, $\rho_{10} = \lambda_{10}/\mu_{10} = 1$

Probabilities of total number of busy channels (not yet normalized):

- Basis for recursion: $g(-10) = g(-9) = \cdots = g(-1) := 0$ and $g(0) := 1$
- $g(1) = 10\big(g(0) + g(-1) + \cdots + q(-9)\big) = 10g(0) = 10$
- $g(2) = \frac{10}{2}(g(1) + g(0) + g(-1) \ldots + g(-8)) = \frac{10*11}{2} = 55$

repeat this step 100 times…

- $g(100) = \frac{10}{100}\big(g(99) + g(98) + \cdots + g(90)\big) = \ \ldots . = 88802067002$
- Normalizing constant $G = (g(0) + g(1) + \ldots + g(100)) = 277485714278$

Example for Kaufmann-Roberts

Probabilities for number of busy channels (after normalization)

- $q(0) = \frac{g(0)}{G} = 3.60379 * 10^{-13}$
- $q(1) = \frac{g(1)}{G} = 3.60379 * 10^{-12}$

These are the probabilities of the total number of occupied channels

- $q(99) = \frac{g(99)}{G} = 0.03217$
- $q(100) = \frac{g(100)}{G} = 0.03200$

Blocking probabilities per class

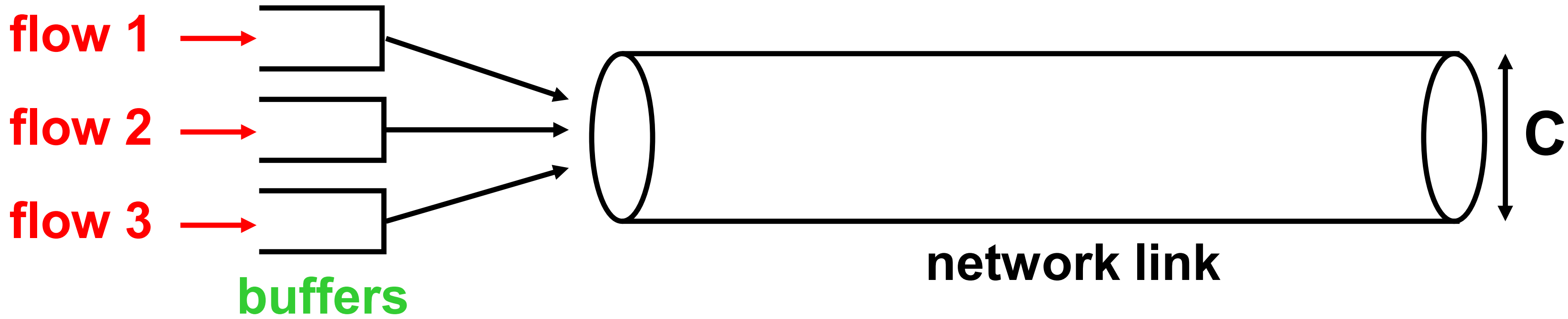
- $B_1 = q(100) = 0.03200$
- $B_2 = q(100) + q(99) = 0.06418$

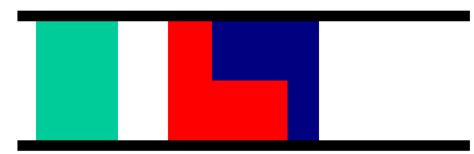
repeat this step 10 times…

- $B_{10} = q(100) + q(99) + \cdots + q(91) = 0.32091$

Elastic Traffic

Multiplexing elastic traffic flows





Properties:

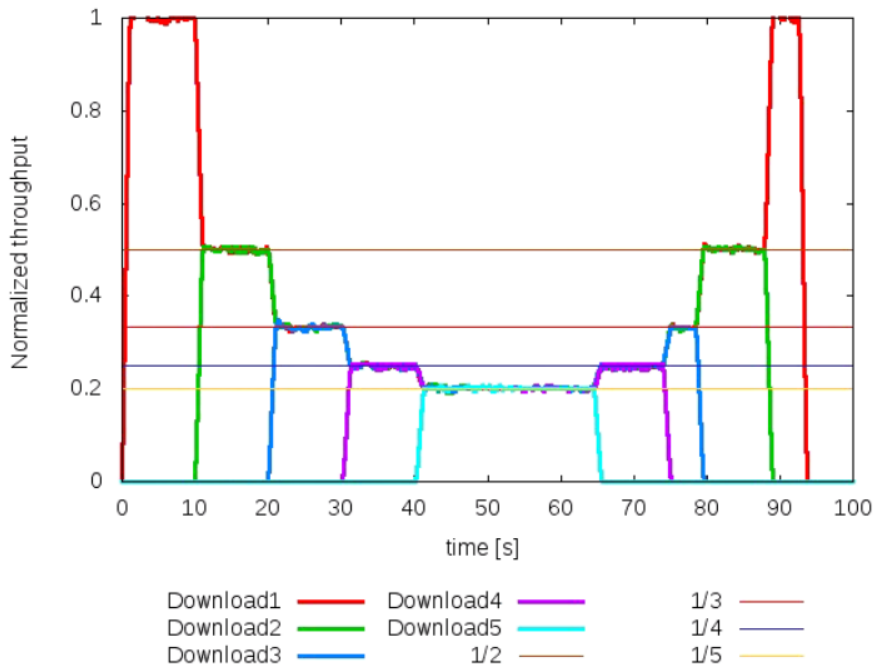
- mainly for data applications
- not very delay-sensitive
- no fixed throughput required: bandwidth sharing

- **link shared by n flows**
- **round-robin (Δt service quantum)**
- **flow throughput: C/n**
- **number of flows may vary => variable throughput**

Processor Sharing model

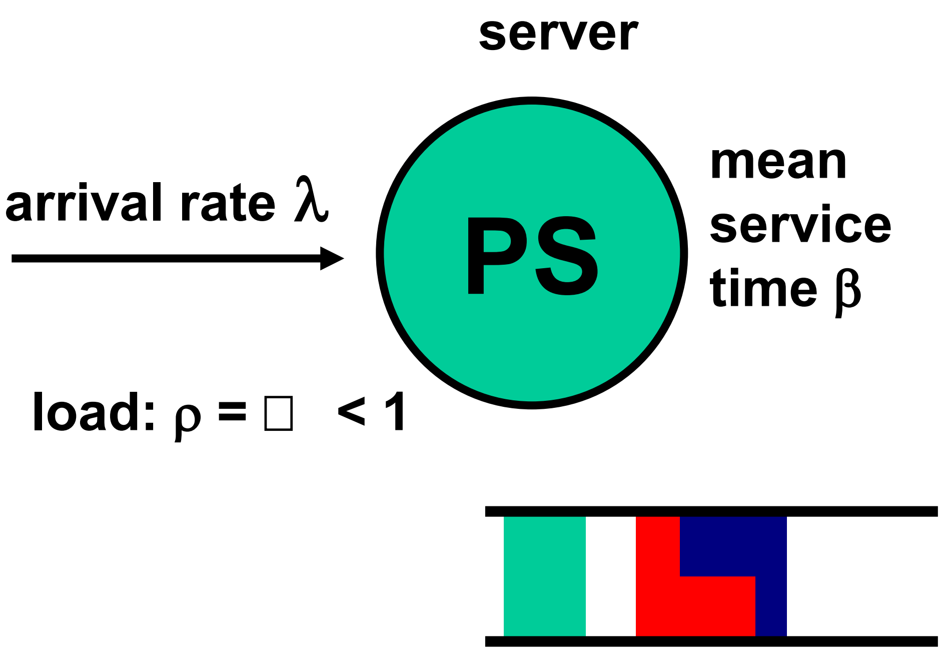
Elastic Traffic: capacity sharing example

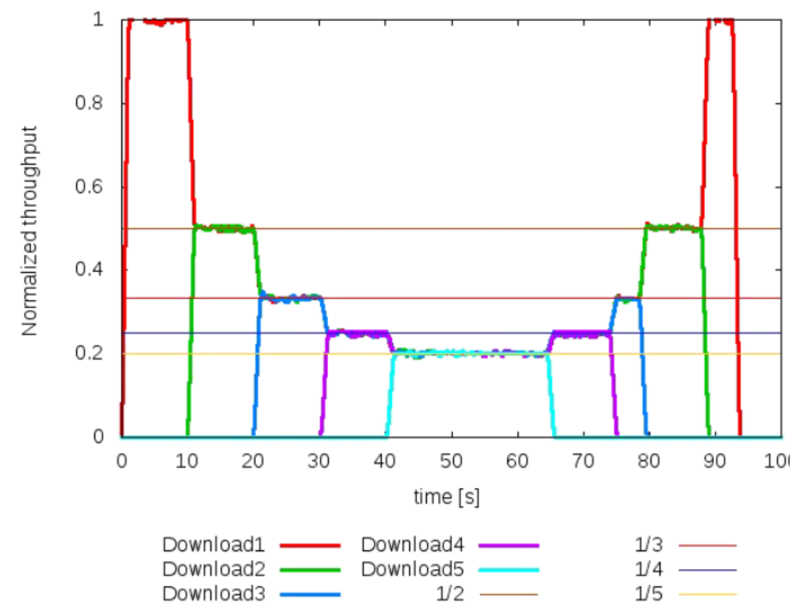
Experiment: TCP file transfer over WLAN



- **Flows enter network at a pre-defined time with a fixed size**
- **Called "elastic" traffic**
- **Bandwidth sharing behaves like Processor Sharing (PS) model**

Processor Sharing Models



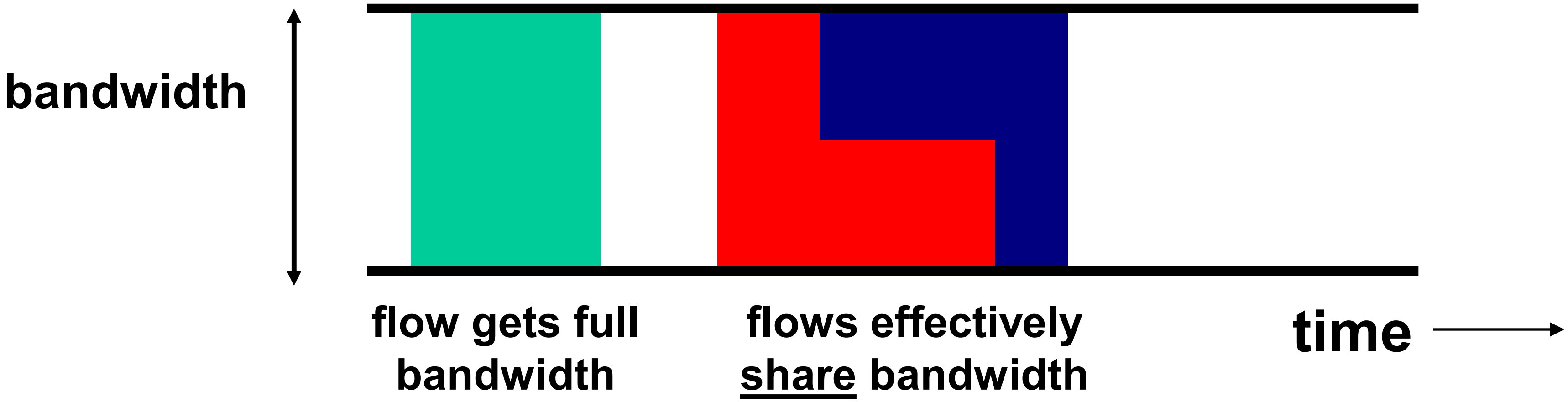


Processor Sharing (PS): If k customers in the system, then each of them gets processing speed 1/k

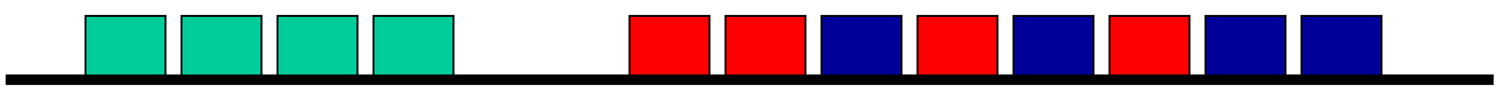
Flow-level behavior of TCP data transfers can be modeled by PS model

Key performance metric: mean sojourn time E[S] (models the transfer time of a flow)

Applications of PS models

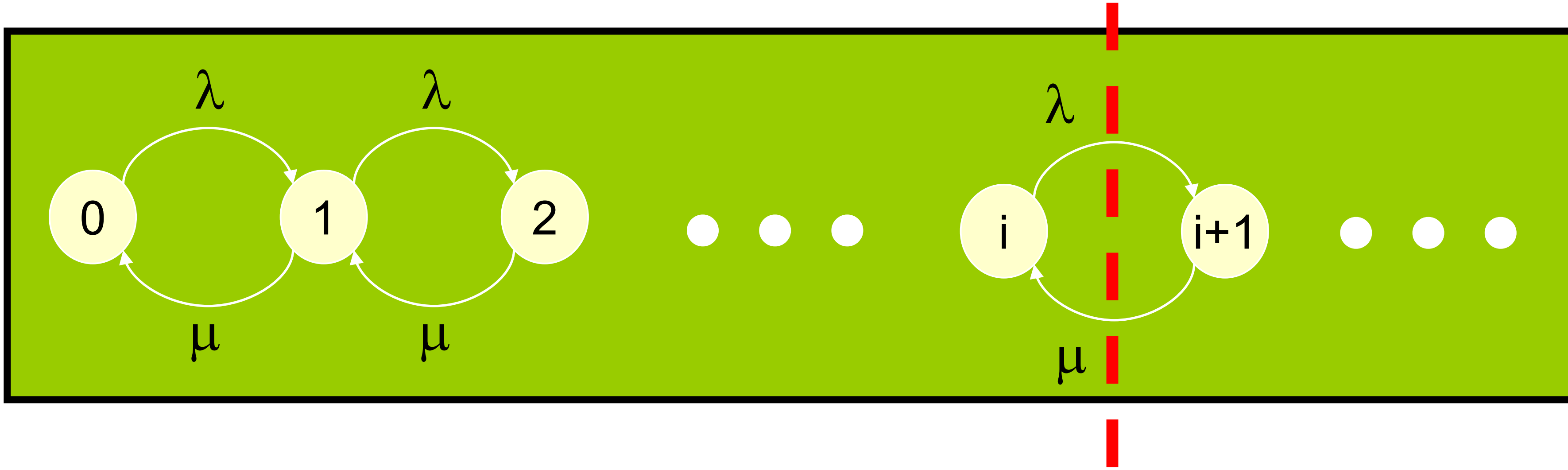


1. Multitasking: jobs fairly share CPU power
2. Bandwidth sharing among different users



"Train of packets" is seen as a <u>single</u> "flow"

Analysis of the M/M/1 PS Model



N := number of customers in system

$$\pi_j := \Pr\{N = j\} (j = 0,1,\ldots)$$

Balancing arguments:

$$\lambda \pi_i = \mu \pi_{i+1} (i = 0,1,2,\ldots)$$

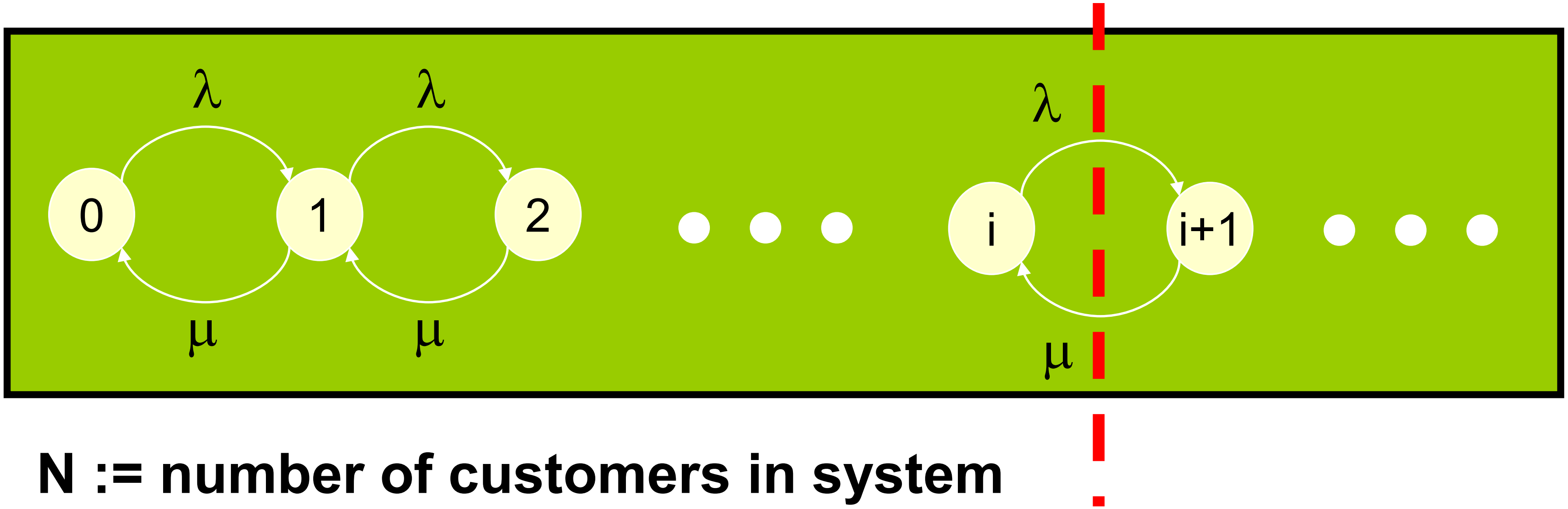
Normalization:

$$\pi_0 + \pi_1 + \cdots = 1$$

Note that the state space is unlimited (as opposed to Erlang-B model), because ther is no maximum to the number of jobs

assuming exponential service times with mean β=1/μ

Analysis of the M/M/1 PS Model



N := number of customers in system

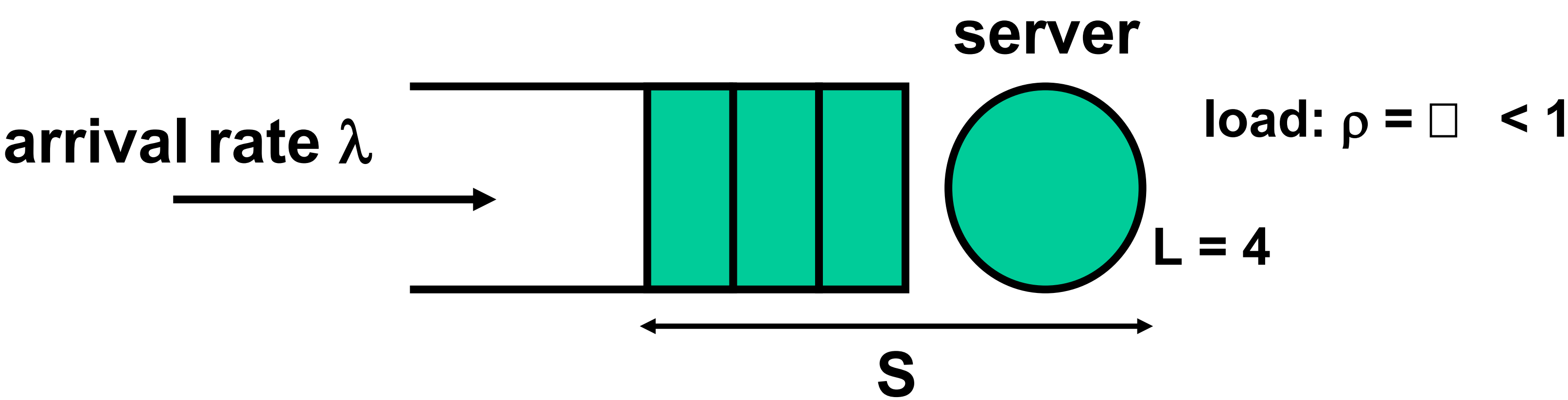
$$\pi_j = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^j = (1-\rho)\rho^j \; (j=0,1,\ldots) \qquad \rho := \lambda\beta = \frac{\lambda}{\mu}$$

Expected number of customers in system

$$E[N] = \sum_{j=0}^{\infty} j \Pr\{N=j\} = (1-\rho)\sum_{j=0}^{\infty} j\rho^j = \frac{\rho}{1-\rho}$$

Expected sojourn time $E[S] = \frac{E[N]}{\lambda} = \frac{\beta}{1-\rho}$ **(using "Little")**

Little’s Formula



S := sojourn time: total time job is in the system is (including service)

L := total number of jobs in the system (including service)

Little’s Formula: $E[L] = \lambda E[S]$

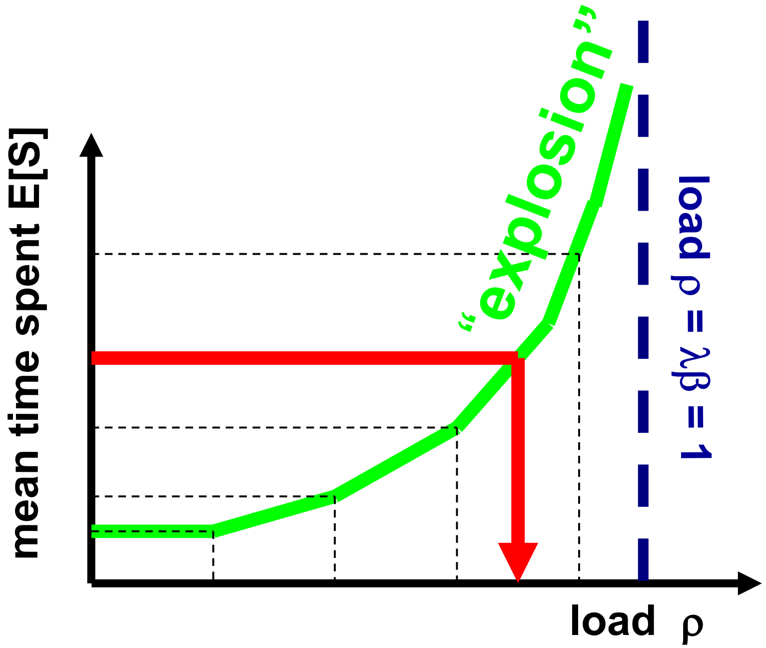
Interpretation via cost structure:

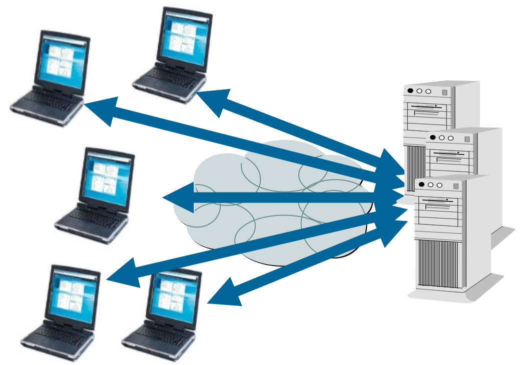
Each customer pays 1 euro per minute in system, earned in 2 ways

1: customers pay continuously in time, then total E[L] per minute

2: pay upon departure, E[S] per customer, in total λE[S] per minute

Analysis of the M/G/1 PS Model



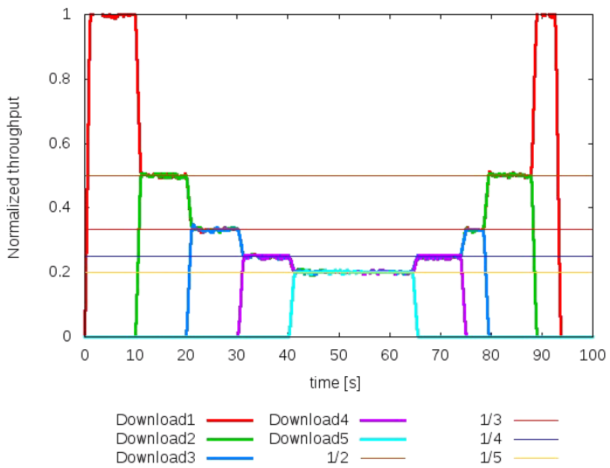


Expected sojourn time: $E[S] = \dfrac{\beta}{1-\rho}$

Insensitivity: result also holds for non-exponential service-time distributions

Interpretation: sharing the capacity with other customers leads to a slow-down factor 1/(1 – ρ)

Example of Processing Sharing Model



Assumptions:

- Network bandwidth = 100 Mbit per second
- Average file size = 5 MByte = 40 Mbit
- File transfer request rate = 2 files per second

Intermediate calculations:

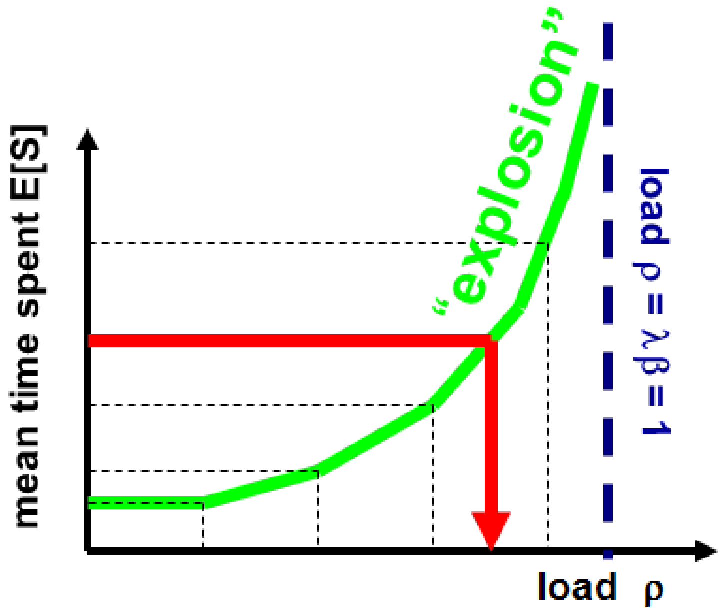
- Arrival rate $\lambda = 2$ (files per second)
- Mean processing time per file β = (40 Mbit) / (100 Mbit/s) = 0.4 seconds
- Utilization $\rho = 0.8 = 80\%$

Mean file transfer time:

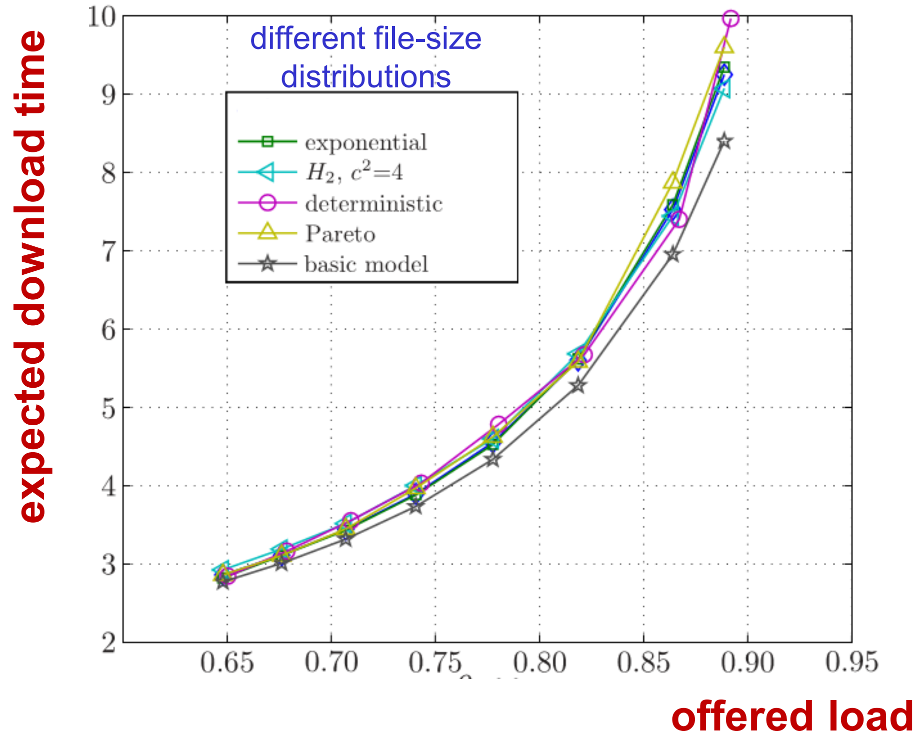
$$E[S] = \frac{\beta}{1-\rho} = \frac{0.4}{1-0.8} = 2\ seconds$$

Model Validation

basic model (theory)



practice (lab setup)



- **Lab results match very well with theoretical PS-model**
- **Transfer time indeed (fairly) insensitive to the file size distribution**

Wrap Up

Done today

1. Streaming versus elastic traffic
2. Erlang blocking model (Erlang-B)
3. Multi-rate models and product-form solution
4. Kaufman/Roberts recursion
5. Elastic traffic: Processor Sharing models

Background reading material

COST 242 - Multi-Rate Models for Dimensioning and Performance Evaluation of ATM Networks, Chapter 4 (only pages 17–25)