# 6.263/16.37: Lectures 5 \& 6 Introduction to Queueing Theory 

Eytan Modiano<br>Massachusetts Institute of Technology

## Packet Switched Networks



## Queueing Systems

- Used for analyzing network performance
- In packet networks, events are random
- Random packet arrivals
- Random packet lengths
- While at the physical layer we were concerned with bit-error-rate, at the network layer we care about delays
- How long does a packet spend waiting in buffers?
- How large are the buffers?
- In circuit switched networks want to know call blocking probability
- How many circuits do we need to limit the blocking probability?


## Random events

- Arrival process
- Packets arrive according to a random process
- Typically the arrival process is modeled as Poisson
- The Poisson process
- Arrival rate of $\lambda$ packets per second
- Over a small interval $\delta$,
$P($ exactly one arrival $)=\lambda \delta+o(\delta)$
$P(0$ arrivals $)=1-\lambda \delta+o(\delta)$
$P($ more than one arrival $)=0(\delta)$
Where $0(\delta) / \delta->0$ as $\delta \rightarrow 0$.
- It can be shown that:

$$
\mathrm{P}(\mathrm{n} \text { arrivals in interval } \mathrm{T})=\frac{(\lambda T)^{n} e^{-\lambda T}}{n!}
$$

## The Poisson Process

$\mathrm{P}(\mathrm{n}$ arrivals in interval T$)=\frac{(\lambda T)^{n} e^{-\lambda T}}{n!}$
$\mathrm{n}=$ number of arrivals in T
It can be shown that,
$\mathrm{E}[\mathrm{n}]=\lambda \mathrm{T}$
$E\left[n^{2}\right]=\lambda T+(\lambda T)^{2}$
$\sigma^{2}=\mathrm{E}\left[(\mathrm{n}-\mathrm{E}[\mathrm{n}])^{2}\right]=\mathrm{E}\left[\mathrm{n}^{2}\right]-\mathrm{E}[\mathrm{n}]^{2}=\lambda T$

## Inter-arrival times

- Time that elapses between arrivals (IA)

$$
\begin{aligned}
P(I A<=t)= & 1-P(I A>t) \\
& =1-P(0 \text { arrivals in time } t) \\
& =1-e^{-\lambda t}
\end{aligned}
$$

- This is known as the exponential distribution
- Inter-arrival CDF $=\mathrm{F}_{\text {IA }}(\mathrm{t})=\mathbf{1}-\mathrm{e}^{-\lambda \mathrm{t}}$
- Inter-arrival PDF $=\mathbf{d} / \mathbf{d t} \mathrm{F}_{\mathrm{IA}}(\mathrm{t})=\lambda \mathrm{e}^{-\lambda \mathrm{t}}$
- The exponential distribution is often used to model the service times (I.e., the packet length distribution)


## Markov property (Memoryless)

$$
P\left(T \leq t_{0}+t \mid T>t_{0}\right)=P(T \leq t)
$$

Proof :

$$
\begin{aligned}
& P\left(T \leq t_{0}+t \mid T>t_{0}\right)=\frac{P\left(t_{0}<T \leq t_{0}+t\right)}{P\left(T>t_{0}\right)} \\
& =\frac{\int_{t_{0}}^{t_{0}+t} \lambda e^{-\lambda t} d t}{\int_{t_{0}}^{\infty} \lambda e^{-\lambda t} d t}=\frac{-\left.e^{-\lambda t}\right|_{t_{0}} ^{t_{0}+t}}{-\left.e^{-\lambda t}\right|_{t_{0}} ^{\infty}}=\frac{-e^{-\lambda\left(t+t_{0}\right)}+e^{-\lambda\left(t_{0}\right)}}{e^{-\lambda\left(t_{0}\right)}} \\
& =1-e^{-\lambda t}=P(T \leq t)
\end{aligned}
$$

- Previous history does not help in predicting the future!
- Distribution of the time until the next arrival is independent of when the last arrival occurred!


## Example

- Suppose a train arrives at a station according to a Poisson process with average inter-arrival time of $\mathbf{2 0}$ minutes
- When a customer arrives at the station the average amount of time until the next arrival is 20 minutes
- Regardless of when the previous train arrived
- The average amount of time since the last departure is 20 minutes!
- Paradox: If an average of $\mathbf{2 0}$ minutes passed since the last train arrived and an average of 20 minutes until the next train, then an average of 40 minutes will elapse between trains
- But we assumed an average inter-arrival time of 20 minutes!
- What happened?


## Properties of the Poisson process

- Merging Property


Let A1, A2, ... Ak be independent Poisson Processes
of rate $\lambda 1, \lambda 2, \ldots \lambda k$

$$
\mathbf{A}=\sum \mathbf{A}_{i} \text { is also Poisson of rate }=\sum \lambda_{i}
$$

- Splitting property
- Suppose that every arrival is randomly routed with probability $\mathbf{P}$ to stream 1 and (1-P) to stream 2
- Streams 1 and 2 are Poisson of rates $P \lambda$ and (1-P) $\lambda$ respectively



## Queueing Models



- Model for
- Customers waiting in line
- Assembly line
- Packets in a network (transmission line)
- Want to know
- Average number of customers in the system
- Average delay experienced by a customer
- Quantities obtained in terms of
- Arrival rate of customers (average number of customers per unit time)
- Service rate (average number of customers that the server can serve per unit time)


## Little's theorem



- $\mathbf{N}=$ average number of packets in system
- $T=$ average amount of time a packet spends in the system
- $\lambda=$ arrival rate of packets into the system (not necessarily Poisson)
- Little's theorem: $\mathbf{N}=\lambda T$
- Can be applied to entire system or any part of it
- Crowded system -> long delays

On a rainy day people drive slowly and roads are more congested!

## Proof of Little's Theorem



- $\quad \alpha(t)=$ number of arrivals by time $t$
- $\quad \beta(t)=$ number of departures by time $t$
- $t_{i}=$ arrival time of $i^{\text {th }}$ customer
- $T_{i}=$ amount of time $i^{\text {th }}$ customer spends in the system
- $\quad \mathbf{N}(\mathbf{t})=$ number of customers in system at time $t=\alpha(t)-\beta(t)$
- Similar proof for non First-come-first-serve


## Proof of Little's Theorem

$$
\begin{aligned}
N_{t} & =\frac{1}{t} \int_{0}^{t} N(\tau) d \tau=\text { timeave.number of customers in queue } \\
N & =\text { Limit }_{t \rightarrow \infty} N_{t}=\text { steadystate time ave. } \\
\lambda_{t} & =\alpha(t) / t, \lambda=\text { Limit }_{t \rightarrow \infty} \lambda_{t}=\text { arrival rate } \\
T_{t} & =\frac{\sum_{i=0}^{\alpha(t)} T_{i}}{\alpha(t)}=\text { timeave.systemdelay, } T=\text { Limit }_{t \rightarrow \infty} T_{t}
\end{aligned}
$$

- Assume above limits exists, assume Ergodic system

$$
\begin{aligned}
& N(t)=\alpha(t)-\beta(t) \Rightarrow N_{t}=\frac{\sum_{i=1}^{\alpha(t)} T_{i}}{t} \\
& N=\lim _{t \rightarrow \infty} \frac{\sum_{i=1}^{\alpha(t)} T_{i}}{t}, \quad T=\lim _{t \rightarrow \infty} \frac{\sum_{i=1}^{\alpha(t)} T_{i}}{\alpha(t)} \Rightarrow \sum_{i=1}^{\alpha(t)} T_{i}=\alpha(t) T \\
& N=\frac{\sum_{i=1}^{\alpha(t)} T_{i}}{t}=\left(\frac{\alpha(t)}{t}\right) \frac{\sum_{i=1}^{\alpha(t)} T_{i}}{\alpha(t)}=\lambda T
\end{aligned}
$$

## Application of little's Theorem

- Little's Theorem can be applied to almost any system or part of it
- Example:


1) The transmitter: $D_{T P}=$ packet transmission time

- $\quad$ Average number of packets at transmitter $=\lambda D_{T P}=\rho=$ link utilization

2) The transmission line: $D_{p}=$ propagation delay
$-\quad$ Average number of packets in flight $=\lambda D_{p}$
3) The buffer: $D_{q}=$ average queueing delay

- $\quad$ Average number of packets in buffer $=N_{q}=\lambda D_{q}$

4) Transmitter + buffer
$-\quad$ Average number of packets $=\rho+\mathbf{N}_{q}$

## Application to complex system



- We have complex network with several traffic streams moving through it and interacting arbitrarily
- For each stream i individually, Little says $\mathbf{N}_{i}=\lambda_{i} \mathbf{T}_{i}$
- $\quad$ For the streams collectively, Little says $\mathbf{N}=\lambda T$ where
- 

$$
\mathbf{N}=\sum_{i} \mathbf{N}_{\mathbf{i}} \quad \boldsymbol{\&} \boldsymbol{\lambda}=\sum_{\mathbf{i}} \lambda_{\mathbf{i}}
$$

From Little's Theorem:

$$
T=\frac{\sum_{i=1}^{i=k} \lambda_{i} T_{i}}{\sum_{i=1}^{i=1} \lambda_{i}}
$$

## Single server queues


$\mu$ packet per second
Service time $=1 / \mu$

- M/M/1
- Poisson arrivals, exponential service times
- M/G/1
- Poisson arrivals, general service times
- M/D/1
- Poisson arrivals, deterministic service times (fixed)


## Markov Chain for M/M/1 system



- $\quad$ State $k \Rightarrow k$ customers in the system
- $P(I, j)=$ probability of transition from state I to state $\mathbf{j}$
- As $\boldsymbol{\delta} \Rightarrow \mathbf{0}$, we get:

$$
\begin{array}{ll}
P(0,0)=1-\lambda \delta, & P(\mathbf{j}, \mathbf{j}+1)=\lambda \delta \\
P(\mathbf{j}, \mathbf{j})=1-\lambda \delta-\mu \delta & P(\mathbf{j}, \mathbf{j}-1)=\mu \delta \\
\mathbf{P}(\mathbf{I}, \mathbf{j})=0 \text { for all other values of } \mathbf{I}, \mathbf{j} . &
\end{array}
$$

- Birth-death chain: Transitions exist only between adjacent states
- $\quad \lambda \delta, \mu \delta$ are flow rates between states


## Equilibrium analysis

- We want to obtain $P(n)=$ the probability of being in state $n$
- At equilibrium $\lambda \mathbf{P}(n)=\mu P(n+1)$ for all $n$
- Local balance equations between two states ( $n, n+1$ )
$-\quad \mathbf{P}(\mathbf{n}+1)=(\lambda / \mu) \mathbf{P}(\mathbf{n})=\rho \mathbf{P}(\mathbf{n}), \rho=\lambda / \mu$
- It follows: $P(n)=\rho^{n} P(0)$
- Now by axiom of probability: $\quad \sum_{i=0}^{\infty} P(n)=1$
$\Rightarrow \sum_{i=0}^{\infty} \rho^{n} P(0)=\frac{P(0)}{1-\rho}=1$
$\Rightarrow P(0)=1-\rho$
$P(n)=\rho^{n}(1-\rho)$


## Average queue size

$$
\begin{aligned}
& N=\sum_{n=0}^{\infty} n P(n)=\sum_{n=0}^{\infty} n \rho^{n}(1-\rho)=\frac{\rho}{1-\rho} \\
& N=\frac{\rho}{1-\rho}=\frac{\lambda / \mu}{1-\lambda / \mu}=\frac{\lambda}{\mu-\lambda}
\end{aligned}
$$

- $\quad \mathbf{N}=$ Average number of customers in the system
- The average amount of time that a customer spends in the system can be obtained from Little's formula ( $\mathbf{N}=\lambda T \Rightarrow T=N / \lambda$ )

$$
T=\frac{1}{\mu-\lambda}
$$

- $\quad$ T includes the queueing delay plus the service time (Service time $=D_{T P}=1 / \mu$ )
- $W=$ amount of time spent in queue $=T-1 / \mu \Rightarrow$

$$
W=\frac{1}{\mu-\lambda}-\frac{1}{\mu}
$$

- Finally, the average number of customers in the buffer can be obtained from little's formula

$$
N_{Q}=\lambda W=\frac{\lambda}{\mu-\lambda}-\frac{\lambda}{\mu}=N-\rho
$$

## Example (fast food restaurant)

- Customers arrive at a fast food restaurant at a rate of $\mathbf{1 0 0}$ per hour and take 30 seconds to be served.
- How much time do they spend in the restaurant?
- $\quad$ Service rate $=\mu=\mathbf{6 0} / \mathbf{0 . 5}=120$ customers per hour
$-\quad T=1 / \mu-\lambda=1 /(120-100)=1 / 20 \mathrm{hrs}=3$ minutes
- How much time waiting in line?
$-\quad W=T-1 / \mu=2.5$ minutes
- How many customers in the restaurant?
$-\quad N=\lambda T=5$
- What is the server utilization?
$-\rho=\lambda / \mu=5 / 6$


## Packet switching vs. Circuit switching

TDM, Time Division Multiplexing
Each user can send $\mu / \mathbf{N}$ packets/sec and has packet arriving at rate $\lambda / \mathrm{N}$ packets/sec

$$
D=M / \mu+\frac{M(\lambda / \mu)}{(\mu-\lambda)} \quad \begin{aligned}
& \text { M/M/1 } \\
& \text { formula }
\end{aligned}
$$

Packets generated at random times


## Circuit (tdm/fdm) vs. Packet switching



## M server systems: M/M/m



- Departure rate is proportional to the number of servers in use
- Similar Markov chain:



## M/M/m queue

- Balance equations:

$$
\begin{aligned}
& \lambda P(n-1)=n \mu P(n) \quad n \leq m \\
& \lambda P(n-1)=m \mu P(n) \quad n>m \\
& P(n)=\left\{\begin{array}{ll}
P(0)(m \rho)^{n} / n! & n \leq m \\
P(0)\left(m^{m} \rho^{n}\right) / m! & n>m
\end{array}, \quad \rho=\frac{\lambda}{m \mu} \leq 1\right.
\end{aligned}
$$

- Again, solve for $\mathbf{P ( 0 )}$ :

$$
\begin{aligned}
& P(0)=\left[\sum_{n=0}^{m-1} \frac{(m \rho)^{n}}{n!}+\frac{(m \rho)^{m}}{m!(1-\rho)}\right]^{-1} \\
& P_{Q}=\sum_{n=m}^{n=\infty} P(n)=\frac{P(0)(m \rho)^{m}}{m!(1-\rho)} \\
& N_{Q}=\sum_{n=0}^{n=\infty} n P(n+m)=\sum_{n=0}^{n=\infty} n P(0)\left(\frac{m^{m} \rho^{m+n}}{m!}\right)=P_{Q}\left(\frac{\rho}{1-\rho}\right) \\
& W=\frac{N_{Q}}{\lambda}, T=W+1 / \mu, N=\lambda T=\lambda / \mu+N_{Q}
\end{aligned}
$$

## Applications of M/M/m

- Bank with m tellers
- Network with parallel transmission lines


VS
One line of rate $m \mu$


- When the system is lightly loaded, PQ~0, and Single server is $\mathbf{m}$ times faster
- When system is heavily loaded, queueing delay dominates and systems are roughly the same


## M/M/Infinity

- Unlimited servers => customers experience no queueing delay
- The number of customers in the system represents the number of customers presently being served


$$
\begin{aligned}
& \lambda P(n-1)=n \mu P(n), \forall n>1, \Rightarrow P(n)=\frac{P(0)(\lambda / \mu)^{n}}{n!} \\
& P(0)=\left[1+\sum_{n=1}^{\infty}(\lambda / \mu)^{n} / n!\right]^{-1}=e^{-\lambda / \mu} \\
& P(n)=(\lambda / \mu)^{n} e^{-\lambda / \mu} / n!\Rightarrow \text { Poissondistribution }! \\
& N=\text { Averagenumber in system }=\lambda / \mu, T=N / \lambda=1 / \mu=\text { servicetime }
\end{aligned}
$$

## Blocking Probability

- A circuit switched network can be viewed as a Multi-server queueing system
- Calls are blocked when no servers available - "busy signal"
- For circuit switched network we are interested in the call blocking probability
- $\mathbf{M} / \mathbf{M} / \mathbf{m} / \mathbf{m}$ system
- $\quad \mathbf{m}$ servers $\Rightarrow \mathbf{m}$ circuits
- Last $m$ indicated that the system can hold no more than $m$ users
- Erlang B formula
- Gives the probability that a caller finds all circuits busy
- Holds for general call arrival distribution (although we prove Markov case only)

$$
P_{B}=\frac{(\lambda / \mu)^{m} / m!}{\sum_{n=0}^{m}(\lambda / \mu)^{n} / n!}
$$

## M/M/m/m system: Erlang B formula



$$
\begin{aligned}
& \lambda P(n-1)=n \mu P(n), 1 \leq n \leq m, \Rightarrow P(n)=\frac{P(0)(\lambda / \mu)^{n}}{n!} \\
& P(0)=\left[\sum_{n=0}^{m}(\lambda / \mu)^{n} / n!\right]^{-1} \\
& P_{B}=P(\text { Blocking })=P(m)=\frac{(\lambda / \mu)^{m} / m!}{\sum_{n=0}^{m}(\lambda / \mu)^{n} / n!}
\end{aligned}
$$

## Erlang B formula

- System load usually expressed in Erlangs
- $\quad A=\lambda / \mu=$ (arrival rate)*(ave call duration) $=$ average load
- Formula insensitive to $\lambda$ and $\mu$ but only to their ratio

$$
P_{B}=\frac{(A)^{m} / m!}{\sum_{n=0}^{m}(A)^{n} / n!}
$$

- Used for sizing transmission line
- How many circuits does the satellite need to support?
- The number of circuits is a function of the blocking probability that we can tolerate

Systems are designed for a given load predictions and blocking probabilities (typically small)

- Example
- Arrival rate $=\mathbf{4}$ calls per minute, average $\mathbf{3}$ minutes per call $=>$ A = $\mathbf{1 2}$
- How many circuits do we need to provision?

Depends on the blocking probability that we can tolerate


## Multi-dimensional Markov Chains

- K classes of customers
- Class $j$ : arrival rate $\lambda_{j}$; service rate $\mu_{j}$
- State of system: $n=(n 1, n 2, \ldots, n k) ; n j=$ number of class $j$ customers in the system
- If detailed balance equations hold for adjacent states, then a product form solution exists, where:
$-\quad P(n, . n 2, \ldots, n k)=P_{1}(n 1) * P_{2}(n 2) * \ldots * P_{k}(n k)$
- Example: K independent M/M/1 systems

$$
P_{i}\left(n_{i}\right)=\rho_{i}^{n_{i}}\left(1-\rho_{i}\right), \quad \rho_{i}=\lambda_{i} / \mu_{i}
$$

- Same holds for other independent birth-death chains
- E.g., M.M/m, M/M/Inf, M/M/m/m


## Truncation

- Eliminate some of the states
- E.g., for the K M/M/1 queues, eliminate all states where n1+n2+...+nk > K1 (some constant)
- Resulting chain must remain irreducible
- All states must communicate


## Product form for stationary distribution of the truncated system

- E.g., $K$ independent $M / M / 1$ queues

$$
P\left(n_{1}, n_{2}, \ldots n_{k}\right)=\frac{\rho_{1}^{n 1} \rho_{2}^{n 2} \ldots \rho_{K}^{n K}}{G}, \quad G=\sum_{n \in S} \rho_{1}^{n 1} \rho_{2}^{n 2} \ldots \rho_{K}^{n K}
$$

- E.g., $K$ independent $M / M / i n f$ queues

$$
P\left(n_{1}, n_{2}, \ldots n_{k}\right)=\frac{\left(\rho_{1}^{n 1} / n_{1}!\right)\left(\rho_{2}^{n 2} / n_{2}!\right) \ldots\left(\rho_{K}^{n K} / n_{k}!\right)}{G}, G=\sum_{n \in S}\left(\rho_{1}^{n 1} / n_{1}!\right)\left(\rho_{2}^{n 2} / n_{2}!\right) \ldots\left(\rho_{K}^{n K} / n_{k}!\right)
$$

- $G$ is a normalization constant that makes $\mathbf{P}(\mathbf{n})$ a distribution
- $\quad S$ is the set of states in the truncated system


## Example

- Two session classes in a circuit switched system
- M channels of equal capacity
- Two session types:

Type 1: arrival rate $\lambda 1$ and service rate $\mu 1$
Type 2: arrival rate $\lambda 2$ and service rate $\mu 2$

- System can support up to $M$ sessions of either class
- If $\mu 1=\mu 2$, treat system as an $M / M / m / m$ queue with arrival rate $\lambda 1+\lambda 2$
- When $\mu 1=!\mu 2$ need to know the number of calls in progress of each session type
- Two dimensional markov chain state $=(\mathrm{n} 1, \mathrm{n} 2)$
- Want P(n1, n2): n1+n2 <=m
- Can be viewed as truncated $M / M /$ Inf queues
- Notice that the transition rates in the M/M/Inf queue are the same as those in a truncated $\mathbf{M} / \mathbf{M} / \mathbf{m} / \mathbf{m}$ queue

$$
P\left(n_{1}, n_{2}\right)=\frac{\left(\rho_{1}^{n 1} / n_{1}!\right)\left(\rho_{2}^{n 2} / n_{2}!\right)}{G}, \quad G=\sum_{i=0}^{i=m} \sum_{j=0}^{j=m-i}\left(\rho_{1}^{i} / i!\right)\left(\rho_{2}^{j} / j!\right), \quad n 1+n 2 \leq m
$$

- Notice that the double sum counts only states for which $\mathbf{j}+\mathbf{i}<=\mathbf{m}$


## PASTA: Poisson Arrivals See Time Averages

- The state of an $M / M / 1$ queue is the number of customers in the system
- More general queueing systems have a more general state that may include how much service each customer has already received
- For Poisson arrivals, the arrivals in any future increment of time is independent of those in past increments and for many systems of interest, independent of the present state $S(t)($ true for $M / M / 1, M / M / m$, and M/G/1).
- For such systems, $P\{S(t)=s \mid A(t+\delta)-A(t)=1\}=P\{S(t)=s\}$
- (where $A(t)=\#$ arrivals since $t=0)$
- In steady state, arrivals see steady state probabilities


## Occupancy distribution upon arrival

- Arrivals may not always see the steady-state averages
- Example:
- Deterministic arrivals 1 per second
- Deterministic service time of 3/4 seconds
$\lambda=1$ packets/second $T=3 / 4$ seconds (no queueing)

$$
\mathbf{N}=\lambda T=\text { Average occupancy }=3 / 4
$$

- However, notice that an arrival always finds the system empty!


## Occupancy upon arrival for a $M / M / 1$ queue

$$
\begin{aligned}
& \mathbf{a}_{\mathbf{n}}=\operatorname{Lim}_{t \rightarrow \inf }(\mathbf{P}(\mathbf{N}(t)=\mathbf{n} \mid \text { an arrival occurred just after time } t)) \\
& \mathbf{P}_{\mathbf{n}}=\operatorname{Lim}_{t->\inf }(\mathbf{P}(\mathbf{N}(\mathbf{t})=\mathbf{n}))
\end{aligned}
$$

For $M / M / 1$ systems $\mathbf{a}_{\mathbf{n}}=\mathbf{P}_{\mathbf{n}}$
Proof: Let $A(t, t+\delta)$ be the event that and arrival occurred between $t$ and $t+\delta$

$$
\begin{aligned}
& \mathbf{a}_{\mathrm{n}}(\mathrm{t})=\operatorname{Lim}_{\mathrm{t}->\mathrm{inf}}(\mathbf{P}(\mathbf{N}(\mathrm{t})=\mathbf{n} \mid \mathrm{A}(\mathrm{t}, \mathrm{t}+\boldsymbol{\delta})) \\
& =\operatorname{Lim}_{t \rightarrow \inf (P(N(t)=n, A(t, t+\delta)) / P(A(t, t+\delta))} \\
& =\operatorname{Lim}_{t->\inf } \mathbf{P}(\mathbf{A}(\mathbf{t}, \mathbf{t}+\boldsymbol{\delta}) \mid \mathbf{N}(\mathrm{t})=\mathbf{n}) \mathbf{P}(\mathbf{N}(\mathbf{t})=\mathbf{n}) / \mathbf{P}(\mathbf{A}(\mathbf{t}, \mathrm{t}+\boldsymbol{\delta}))
\end{aligned}
$$

- Since future arrivals are independent of the state of the system,

$$
\mathbf{P}(\mathbf{A}(\mathbf{t}, \mathrm{t}+\delta) \mid \mathbf{N}(\mathrm{t})=\mathbf{n})=\mathbf{P}(\mathbf{A}(\mathbf{t}, \mathrm{t}+\delta))
$$

- Hence, $\mathbf{a}_{\mathbf{n}}(\mathbf{t})=\mathbf{P}(\mathbf{N}(\mathrm{t})=\mathbf{n})=\mathbf{P}_{\mathrm{n}}(\mathrm{t})$
- Taking limits as t-> infinity, we obtain $a_{n}=P_{n}$
- Result holds for M/G/1 systems as well

