SFWR ENG 4E03

Kemal Ahmed

Fall 2015

Instructor: Vincent Maccio

Note: material covered in [Stats 3Y03 Summary](https://drive.google.com/file/d/0BxW61uJyyN8TNy1iUFE0ZlRMLTg/view) will not be covered in this summary. To find a unit CTRL-F “[<unit>]”, e.g. for Number of jobs in system, CTRL-F “[N]”

Contents

[Statistics 2](#_Toc437977067)

[Variance 2](#_Toc437977068)

[Exponential 2](#_Toc437977069)

[Uniform 3](#_Toc437977070)

[Binomial 3](#_Toc437977071)

[Operations Analysis 3](#_Toc437977072)

[Visitation Trick 5](#_Toc437977073)

[Summation Equations 5](#_Toc437977074)

[DTMC 6](#_Toc437977075)

[Balance Equations 6](#_Toc437977076)

[Matrices 6](#_Toc437977077)

[CTMC 6](#_Toc437977078)

[Poisson Process 6](#_Toc437977079)

[Kendall notation 6](#_Toc437977080)

[Variations 7](#_Toc437977081)

[Steady State 7](#_Toc437977082)

[M/M/1 7](#_Toc437977083)

[M/M/1/N 7](#_Toc437977084)

[M/M/C 8](#_Toc437977085)

[M/M/∞ 8](#_Toc437977086)

[M/G/1 8](#_Toc437977087)

[Queuing 8](#_Toc437977088)

[Square Root Staffing Rule 9](#_Toc437977089)

[e.g.) 9](#_Toc437977090)

[Jackson Networks 10](#_Toc437977091)

[Open Loop 10](#_Toc437977092)

[Traffic Equations 10](#_Toc437977093)

[Closed Loop 10](#_Toc437977094)

[Mean Value Analysis 10](#_Toc437977095)

[General Distribution 11](#_Toc437977096)

[First Come, First Serve 11](#_Toc437977097)

[Last Come, First Serve 11](#_Toc437977098)

[Processor Sharing 12](#_Toc437977099)

[Failure 12](#_Toc437977100)

# Statistics

**Poisson parameter** [λ]: rate

**Service rate** [μ]:

**Continuous Random Variable (CRV)**:



*Think chemistry, i.e. cancelling units*

**Probability Density Function (PDF) [f(x)]**:

**Cumulative Density Function (CDF) [F(x)]**:

**Second Moment** E[x2]

## Variance



* Don’t change probability, but square X for calculation only



**The higher your variance, the worse your system will perform.**

## Exponential

* **Mean** [E[X]]: 1/λ
  + a.k.a. Expected value
* **Variance**: 1/λ2
* **Probability Distribution Function (PDF)** [P(X=x)]: λe–λx/x!
* **Cumulative Distribution Function (CDF)** [f(x)]: CDF = ∫PDF, i.e. 1 – e–λx
* Memoryless
* not always for time

## Uniform

* **Variance**: (b–a)2/12
* **Mean**: (a+b)/2
* **PDF**: 1 / (b–a) , a ≤ x ≤ b
* **CDF**: x–a/b–a
* **Uniform Distribution**: no memoryless property

## Binomial

* **Mean** [E[X]]: n × probability
* **Variance**: n × p × (1 – p)
* **Probability Distribution Function (PDF)** [P(X)]: (n c x)px(1–p)n–x
* **Cumulative Distribution Function (CDF)** [f(x)]: 

# Operations Analysis

**Device** [i]: units that are in terms of *i* are specific to an individual device or node within a system

**Total devices** [k]:

**Service Time** [S]: time per specific job

1/μ

**Visitation** [V]: given or projected visits/jobs (closed system); cannot be calculated; basically a probability

[E(V)]: calculated visit/job ratio

P(visit)∙total visits in previous node

**Demand** [D]: total service time for all jobs





**Bottleneck** [Dmax]: device with largest demand, utilization

**Time in system** [T]: time the job is in the system





If E[Z] = 0, T = R

**Response Time** [R]: time the job is *being processed* in the system

If E[Z] = 0, R = T

M/M/1: E[R] = 1/(μ–λ)

M/M/1/N: E[R] = E[N]/λ'

M/M/C: E[R] = E[RQ] + E[S]

**Users** [M]:

**Optimal users** [M\*]:



**Total Jobs** [N]: N=M in a closed system

* *Little’s Law*:
* 
* M/M/1:
  + E[N] = λ/(μ–λ) = *ρ*/(1–*ρ*), if you have overall system λ
  + E[N] = ← probability × #jobs, if your λ or μ is different for each state
* M/M/1/N: E[N] is expected # jobs, N is max # jobs
* M/M/C: go through Little’s law
  + E[N] = E[NQ] + ρ
* M/M/∞:
* Jackson Network: E[N] = ΣE[Ni] = ΣPλ/(μi–Pλ) = Σ(λi/(μi–λi))

**Think time** [Z]: time it takes the user to put a request in and start, it’s kinda like the frequency that users put in requests (seconds / request)



**Throughput** [X]: out-rate, max jobs / hour of full system



Note: andconverge at their lowest point, so equate them



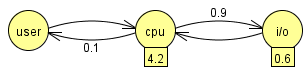
**Utilization** [ρ]: ratio that the time is busy



ρ = λ/ciμ

## Visitation Trick

If determining visitation at a node, establish a reference node from one of the incoming nodes, usually the user node, that has a returning percentage



Vuser = 1 = 0.1 ∙ VCPU

## Summation Equations

**Geometric Series**: , where 0 ≤ r ≤ 1 (because otherwise it would be unstable)



**Geometric Sequence**: 



Removing the annoying factors:



Take out a value so the integral takes out the i and i+1



# DTMC

**Discrete Time Markov Chains (DTMC)**:

[n]: number of tasks in queue / system

Steady state: n->∞

For discrete: use the sum of the X’s, so E[X] = Σ(P(X=i)∙Xi) and E[X2] = Σ(P(X=i)∙Xi2)

## Balance Equations



^think of it like series / parallel, where you add multiple connections out in different directions (parallel) and multiply connections stacked onto each other (series)

OR jobsin = jobsout

## Matrices

Rows: equations for nodes going out (add up to 1)

Columns: equations for nodes coming in

# CTMC

## Poisson Process

**Counting Process**: a way of determining the time between consecutive occurrences of an event

**Poisson Process**: a *counting process*, whose time between arrivals uses Exponential Distribution

* λtotal = Σλi
  + you can also split up λ into multiple λs
* Not only do you see each second as time independent, each stream of probabilities is independent
* P(x;λ) = e–λλx/x!
  + [x]: things will happen
  + [λ]: rate; λ = αt
* [α]: expected number of events during unit interval
* [t]: time interval length
* 

## Kendall notation

**Job Processing time** [μ]: rate of jobs leaving system (jobs/sec)

μ = 1/ processing\_time\_per\_job

M/M/1 Queue

[M]: time between arrivals is Markovian (Memoryless) ~ exp(λ)

[M]: job processing times are Markovian (Memoryless) ~ exp(μ)

[1]: single server

(Σpout)×πi = Σpjπj, j=0..n, j≠i

π0: percent of time that the queue is empty

Attributes:

* FIFO
* Infinite buffer

### Variations

* M/M/2 Queue: same, except 2 servers
* M/M/C Queue: *C* servers
* M/Ek/C: Erlang k, i.e. series of exponential
* H()/M/C: hyperexpontial distribution
* PH/M/C: phase type, i.e. any combination of any number of exponentials with any rate
* M/G/C: Memoryless, general distribution of service time
* G/G/1: has not been solved yet
* M/M/1/1: 1 server, 1 job

[c]: number of servers

Think: one queue goes to multiple servers



## Steady State

### M/M/1

π0 = 1 – λ/μ

πi = ρi(1 – ρ)



### M/M/1/N

When you can only have up to N jobs in system queue.

[λ']: rate jobs enter the system, until the queue is full

λ' = λ(1 – πN)





**Waiting**: jobs put into the queue

**Blocked**: jobs not allowed in the queue

### M/M/C

Useful if multiple jobs are sharing the same queue

Does the μ you use for equations double in M/M/2? No, but you’ll see jobs coming out of a system at a rate of c∙μ.



****

### M/M/∞

Same as M/M/C, except:



and just find the unit

### M/G/1

General Distribution of service time

Heaviside function: 1 if not zero

[E[A]]: arrivals

E[A] = ρ

## Queuing

**Blocking Probability** [PQ]: probability that a process will be blocked when entering the system and be placed in the queue

Erlang-C Equation: 

Given λ and μ, what should c be so PQ < ρ

**Waiting time in queue** [RQ]: response time of queue



M/M/1: 

M/M/C: 

M/M/∞: E[RQ] = 0

**Number of jobs in queue** [NQ]:

M/M/1: ρ2/(1 – ρ)

M/M/C: 

M/M/∞: E[NQ] = 0

You need to know what is in the progression of each step

#### e.g.

When you have varying





## Square Root Staffing Rule

Given an M/M/c queue with arrival rate, λ, server speed, μ, and ρ is *large* (assume this means over 100, but we don’t actually know what it means), α is a bound on PQ, let denote the least # of servers needed to ensure that PQ < α. Then

, where k = is the solution to

, where Φ(∙) is the CDF of the standard normal and φ() is its pdf

[K]: minimum # servers to stay stable λ/μ or ρ

[k]: a constant…just assume 1 for now

Essentially, the perfect number of servers is ρ + √ρ

### e.g.)

|  |  |  |
| --- | --- | --- |
| α | k | ρ + k√ρ |
| 0.8 | 0.178 | 10, 018 |
| 0.5 | 0.506 | 10, 051 |
| 0.2 | 1.06 | 10, 106 |
| 0.1 | 1.42 | 10, 142 |

[Q]: transition matrix



Replace i <--> j to get qjj and qji.

# Jackson Networks

## Open Loop

P(N1 = n1) =



π**n~** = P(state of system n~)P(n jobs at node i)



**Poisson Arrivals See Time Averages property (PASTA)**: the probability of a state (i.e. πi) as seen by an outside random observer is the same as the probability of the state seen by an arriving customer. It is the open loop counterpart to arrival theorem

λtotal = Σλin,i

## Traffic Equations

For each node, what is the number of jobs entering?

λx = R + ΣPi,entering ∙ λi,entering

response rate + probability of each job entering

## Closed Loop

Since your values will become linearly independent, you cannot simply use your regular traffic equations. You need to estimate a fake value for one of your λ’s and evaluate your probabilities using them.

# Mean Value Analysis

**(MVA)**: Finds E[R] of each node of a **closed Jackson network**.

I think it is n2, whereas other methods are nn

**Visit Ratio** [v]: based on a reference node, usually set vref = c

, e.g. 

1. Base case: λ(1) = 1/μ
2. For k = 1, ..., K, compute: 
3. Little’s Law: 
4. Plug it in: 

* Performs better than balance equations or Jackson Network, but can’t find steady state distribution or PDF
* Recursive algorithm
* Only finds E[N], i.e. mean queue length

The higher your variance, the worse your system will perform.

**Arrival Theorem**: when a job arrives at a node within a closed Jackson network, there will be a number of jobs at the node, M – 1, where M is the expected number of jobs in the given node.

**Pareto Power [α]**: 0 < α < 2

**Pareto distribution**: an exponential which doesn’t start at 0 (a.k.a. **zipfian**)

CDF:

PDF: 



Just think: 99% controls 50% and 1% controls the rest

integral of the density function between k and p come out to 1

**Inspection Paradox**:

# General Distribution

**Baskett, Chandy, Muntz and Palacios (BCMP) theorem**: named after the authors of the paper

## First Come, First Serve

**First Come First Serve (FCFS)**:

## Last Come, First Serve

**Last Come First Serve (LCFS)**:

## Processor Sharing

**Processor Sharing (PS)**: fair, overhead, switching costs, thrashing

# Failure

**Failure / Hazard Rate** [h]:



**Increasing Failure Rate (IFR)**:

**Decreasing Failure Rate (DFR)**:

Reaming Processing Time: