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](#_Toc438074808)

[Variance 2](#_Toc438074809)

[Exponential 3](#_Toc438074810)

[Uniform 3](#_Toc438074811)

[Binomial 3](#_Toc438074812)

[Operations Analysis 3](#_Toc438074813)

[Visitation Trick 5](#_Toc438074814)

[Summation Equations 5](#_Toc438074815)

[DTMC 6](#_Toc438074816)

[Balance Equations 6](#_Toc438074817)

[Matrices 6](#_Toc438074818)

[CTMC 6](#_Toc438074819)

[Poisson Process 7](#_Toc438074820)

[Kendall notation 7](#_Toc438074821)

[Variations 7](#_Toc438074822)

[Steady State 8](#_Toc438074823)

[M/M/1 8](#_Toc438074824)

[M/M/1/N 8](#_Toc438074825)

[M/M/C 8](#_Toc438074826)

[M/M/C/N 8](#_Toc438074827)

[M/M/∞ 9](#_Toc438074828)

[M/G/1 9](#_Toc438074829)

[Response Time 9](#_Toc438074830)

[Job Queue Size 9](#_Toc438074831)

[Cost 10](#_Toc438074832)

[Square Root Staffing Rule 10](#_Toc438074833)

[e.g.) 10](#_Toc438074834)

[Jackson Networks 11](#_Toc438074835)

[Open Loop 11](#_Toc438074836)

[Traffic Equations 11](#_Toc438074837)

[Closed Loop 11](#_Toc438074838)

[Mean Value Analysis 11](#_Toc438074839)

[Excess 12](#_Toc438074840)

[Cycles 12](#_Toc438074841)

[General Distribution 13](#_Toc438074842)

[First Come, First Serve 13](#_Toc438074843)

[Last Come, First Serve 13](#_Toc438074844)

[Shortest Remaining Processing Time 13](#_Toc438074845)

[Processor Sharing 13](#_Toc438074846)

[Longest Remaining Processing Time 14](#_Toc438074847)

[Random 14](#_Toc438074848)

[Failure 14](#_Toc438074849)

# 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]:

**Standard Deviation**: √var

## 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

* n = trials, x = successes
* **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/C­1/C2: 1/μ, C1 ≥ C2

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



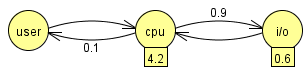
**Utilization** [ρ]: probability that the processor is busy



ρ = λ/cμ

## Visitation Trick

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



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)**: probability

[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

OR ratein × probin = rateout × probout

## Matrices

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

Columns: equations for nodes coming in

# CTMC

**Continuous Time Markov Chain (CTMC)**: rate

## 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]: distribution of time between arrivals is Markovian (Memoryless) ~ exp(λ)

[M]: distribution of 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, maximum 1 job in queue
* arrivals/processing/servers/queue size

[c]: number of servers

Think: one queue goes to multiple servers



## Steady State

**Steady State Probability**: probability x many jobs will be in system (not just in each server!)

**Blocking Probability** [PQ]: probability that a process will be blocked when entering the system and be placed in the queue. This is the same as *Steady State Probability*, since the number of jobs in a system dictates if a job will be blocked.

### 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. After it is full, jobs will be booted from system

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

λ' = λ(1 – πN)





Then remember, 1 = Σπi to find

**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

Erlang-C Equation: 

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

### M/M/C/N

The following two equations are for the probability of entering the 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

## Response Time

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



M/M/1: 

M/M/C: 

M/M/∞: E[RQ] = 0

## Job Queue Size

**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





## Cost

Jobs incur a cost only if they’re waiting

**Hourly Cost of job** [h]:

**Waiting cost** [k]:

C1 = h ∙ E[N]

C2 = k ∙ λ P(jobs have to wait)

## 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) =

balance pick



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.

**Correction Variable** [C]:

**Jobs in system** [M]:

1 = C(Σstates) = C(Σρ , such that sum of powers for each state = M)

# 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

1. Do traffic equations relative to reference node
2. Calculate λs, i.e. v’s relative to this node

, e.g. 

1. Base case:
   1. R(1) = 1/μ
   2. λ(1) = M/(pi ∙ R(1))
2. For k = 1..M (jobs), compute:
   1. We need to find:
   2. Instantiate λdenominator = 0
   3. for i = 1..N (servers):
      1. 
      2. λden += 
3. Plug it in: 

* Performs better than balance equations or Jackson Network, but can’t find steady state distribution or PDF
* Recursive algorithm, but I found it faster to implement it without recursion
* 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.

# Excess

**Inspection Paradox**: earlier you come, longer you have to wait. Think if you just missed the bus vs people who come right before the bus arrives

**Current Excess Time** [Te]:

**Age** [Ta]: how long job has been processed



# Cycles

**Personal Reward Theorem**: the expected excess is equal to the total excess accumulated over a single “cycle”, distributed by said cycle length

# General Distribution

For M/M/1, use the same formula for E[N]. For M/G/1, you may have to do something different when you have non-memoryless functions, e.g. FIFO

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

w.p. Width Probability

o.w. OtherWise

## First Come, First Serve

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

optimal if IFR

If exponential, then same as M/M/1

**Variance** [σs]:



## Last Come, First Serve

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

* Problems:
  + Context switch/ overhead
  + Isn’t fair!
* Assume stable
* Inspection paradox: could be good when you have few larger jobs

E[N] = M/M/1



**LCFS-PRe-emptive (LCFS-PR)**:

## Shortest Remaining Processing Time

**Shortest Remaining Processing Time (SRPT)**:

* # of jobs low
* response time low (optimal)
* need job size info
* Overhead
  + Starvation (fairness)

## Processor Sharing

**Processor Sharing (PS)**:

* Everyone is equal
* Constantly switch between all the jobs
* a.k.a. **thrashing**
* e.g. X = 5s, μ = 1/5
* Problems: overhead / switching costs
* a.k.a. Round Robin
* E[N] = M/M/1



## Longest Remaining Processing Time

**Longest Remaining Processing Time (LRPT)**:

* only useful if highest priority jobs
* would eventually become PS because the length of time remaining will reach the next longest processing time

E[RPS] = E[RLCFS] 

## Random

* Can be unfair
* Problems:
  + Large jobs starved

# Failure

**Pareto Power [α]**: 0 < α < 2

[K]:

**Pareto distribution**: a continuous exponential which doesn’t start at 0

**Zipfian distribution**: discrete equivalent of Pareto

x range = k..p

Kmin is the lowest value of x

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

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



* **Uniform**: 1/(b–t)
* **Exponential**: λ

**Increasing Failure Rate (IFR)**:

**Decreasing Failure Rate (DFR)**:

**Both**: constant, since it’s memoryless

**Neither**: when there are parts that increase and parts that decrease

Reaming Processing Time: