Queuing Theory Quick Reference

0.1 Important Variables

Name	**Variable(s)**	**Description**
Arrival rate	λ or A	The rate at which jobs arrive to the system
Mean interarrival time	$1/\lambda$	The mean time between jobs arriving in the system
Mean service time	$1/\mu$ or S	The mean time that a job waits upon reaching the head of the queue
Service rate	μ.	The rate at jobs are served
Traffic intensity	ρ	A measure of load offered to the system
Utilization	U	The proportion of time the server is busy
Throughput	X	The rate at which the whole system processes jobs
Response time	R	Time from a job's arrival to its service completion, aka "sojurn time"
Mean queue length	$\bar{n}$	Average length of a given queue
Visit ratio	V_i	Relative number of visits to the entire system for queue i.
Think time	Z	The average time a user starts thinking before re-launching a task in closed systems

0.2 Kendall Notation

$A/B/m/K/n/D$, where:

A	Distribution function of interarrival times
B	Distribution function of service times
m	Number of servers
K	Capacity, or maximum number of jobs in the system including the one being serviced
	Population size
D	Service discipline (FIFO, FCFS etc.)

M means *Markovian*, or expoentially distributed. G means *Generally* distributed (usually with C^2 as coefficient of variation)

If not specified, $K = \infty$, $n = \infty$ D is FIFO

0.3 Properties and Laws

$f(x) = \lambda e^{-\lambda x}$	Exponential PDF
$F(x) = \int_{-\infty}^{x} \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$	Exponential CDF
$\rho = \lambda/\mu = \lambda S$	Definition of Traffic Load
$U = \rho$	single server system, when the system does not drop jobs
$U \leq \rho$	when the system may drop jobs
$U = XS$	Utilization Law of a single server
$\bar{n} = XR$	Little's Law
$\bar{n} = \frac{\rho}{1-\rho}$	for $M/M/1$ queues, for $0 \leq \rho < 1$ (eq 3.8)
$X_i = V_i X_{global}$	Force Flow Law on a given server i
$D_i = V_i S_i$	The demand on a given server i
$U_i = X_{global} D_i$	Utilization and Forced Flow Law
$X_{global} < \frac{1}{D_{max}}$	Bottleneck analysis
$V_{cpu} = 1 + V_{io_1} + ... + V_{io_k}$	Visit ratio in a central server system
$R_{global} = \sum_{i=i}^{K} V_i S_i$	Overall system response time
$U_k = \lambda V_k S_k$	Jackson's theorem. Open models when $U_k < 1, \forall k$
$R(N) = \frac{N}{X_{global(N)}} - Z$	Interactive Response time law, for total circulating tasks N