The M/M/n+G Queue:
Summary of Performance Measures

Avishai Mandelbaum and Sergey Zeltyn*

Faculty of Industrial Engineering & Management
Technion
Haifa 32000, ISRAEL

emails: avim@tx.technion.ac.il, zeltyn@ie.technion.ac.il

January 16, 2009

Contents

* **Acknowledgements.** The research of both authors was supported by BSF (Binational Science Foundation) grant 2001685/2005175, ISF (Israeli Science Foundation) grants 388/99, 126/02 and 1046/04, the Niderzaksen Fund and by the Technion funds for the promotion of research and sponsored research.

1 M/M/n+G: primitives and building blocks

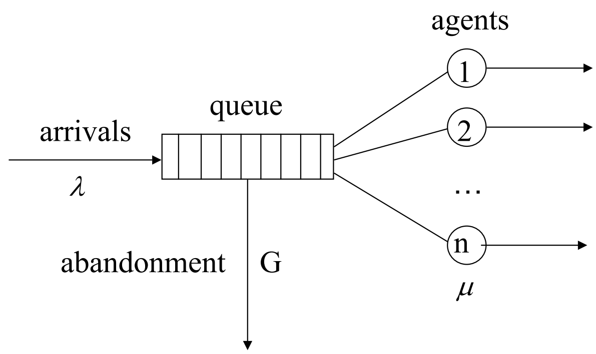
Primitives:

λ – arrival rate,

μ – service rate (= reciprocal of average service time),

n – number of servers,

G – patience distribution ($\bar{G} = 1 - G$: survival function).



Building blocks.

Define

$$H(x) \triangleq \int_0^x \bar{G}(u)du\,.$$

Let

$$J \triangleq \int_0^\infty \exp\{\lambda H(x) - n\mu x\}\,dx\,,$$
$$J_1 \triangleq \int_0^\infty x \cdot \exp\{\lambda H(x) - n\mu x\}\,dx\,,$$
$$J_H \triangleq \int_0^\infty H(x) \cdot \exp\{\lambda H(x) - n\mu x\}\,dx\,.$$

In addition, let

$$J(t) \triangleq \int_t^\infty \exp\{\lambda H(x) - n\mu x\}\,dx\,,$$

and

$$J_H(t) \triangleq \int_t^\infty H(x) \cdot \exp\{\lambda H(x) - n\mu x\}\,dx\,.$$

Finally, introduce

$$\mathcal{E} \triangleq \frac{\sum_{j=0}^{n-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}}\,.$$

1.1 Special case. Deterministic patience (M/M/n+D).

Patience times equal to a constant D. Then

$$H(x) \;=\; \begin{cases} x, & 0 \le x \le D \\ D, & x > D \end{cases} \;.$$

If $\lambda - n\mu \neq 0$,

$$J \;=\; \frac{1}{n\mu - \lambda} - \frac{\lambda}{n\mu(n\mu - \lambda)} \cdot e^{-(n\mu-\lambda)D}\,,$$

$$J(t) \;=\; \begin{cases} \dfrac{1}{n\mu - \lambda} \cdot e^{-(n\mu-\lambda)t} - \dfrac{\lambda}{n\mu(n\mu - \lambda)} \cdot e^{-(n\mu-\lambda)D}\,, & t < D \\[2ex] \dfrac{1}{n\mu} \cdot e^{\lambda D - n\mu t}\,, & t \ge D \end{cases}$$

$$J_1 \;=\; \frac{1}{(n\mu - \lambda)^2} - \left[\frac{1}{(n\mu - \lambda)^2} - \frac{1}{(n\mu)^2} + \frac{\lambda D}{n\mu(n\mu - \lambda)}\right] \cdot e^{-(n\mu-\lambda)D}\,,$$

$$J_H \;=\; \frac{1}{(n\mu - \lambda)^2} \cdot [1 - e^{-(n\mu-\lambda)D}] - \frac{\lambda D}{n\mu(n\mu - \lambda)} \cdot e^{-(n\mu-\lambda)D}\,,$$

$$J_H(t) \;=\; \begin{cases} \dfrac{1}{(n\mu - \lambda)^2} \cdot [e^{-(n\mu-\lambda)t} - e^{-(n\mu-\lambda)D}] + \dfrac{t}{n\mu - \lambda} \cdot e^{-(n\mu-\lambda)t} - \dfrac{\lambda D}{n\mu(n\mu - \lambda)} \cdot e^{-(n\mu-\lambda)D}\,, & t < D \\[2ex] \dfrac{D}{n\mu} \cdot e^{\lambda D - n\mu t}\,, & t \ge D \end{cases}$$

If $\lambda - n\mu = 0$,

$$J \;=\; D + \frac{1}{n\mu}\,,$$

$$J(t) \;=\; \begin{cases} D - t + \dfrac{1}{n\mu}\,, & t < D \\[2ex] \dfrac{1}{n\mu} \cdot e^{\lambda D - n\mu t}\,, & t \ge D \end{cases}$$

$$J_1 \;=\; \frac{D^2}{2} + \frac{D}{n\mu} + \frac{1}{(n\mu)^2}\,,$$

$$J_H \;=\; \frac{D^2}{2} + \frac{D}{n\mu}\,,$$

$$J_H(t) \;=\; \begin{cases} \dfrac{D^2 - t^2}{2} + \dfrac{D}{n\mu}\,, & t < D \\[2ex] \dfrac{D}{n\mu} \cdot e^{\lambda D - n\mu t}\,, & t \ge D \end{cases}$$

1.2 Special case. Exponential patience (M/M/n+M, Erlang-A).

Patience times are iid $\exp(\theta)$. Then

$$H(x) \;=\; \frac{1}{\theta}\cdot\left(1-e^{-\theta x}\right).$$

Define the *incomplete Gamma function*

$$\gamma(x,y) \;\triangleq\; \int_0^y t^{x-1}e^{-t}dt, \qquad x>0, \quad y\geq 0.$$

($\gamma(x,y)$ can be calculated in Matlab.) Then

$$\begin{aligned}
J &= \frac{\exp\left\{\frac{\lambda}{\theta}\right\}}{\theta}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}}\cdot\gamma\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)\\
J(t) &= \frac{\exp\left\{\frac{\lambda}{\theta}\right\}}{\theta}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}}\cdot\gamma\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}e^{-\theta t}\right)\\
J_H &= \frac{J}{\theta}-\frac{\exp\left\{\frac{\lambda}{\theta}\right\}}{\theta^2}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}+1}\cdot\gamma\left(\frac{n\mu}{\theta}+1,\frac{\lambda}{\theta}\right)\\
J_H(t) &= \frac{J(t)}{\theta}-\frac{\exp\left\{\frac{\lambda}{\theta}\right\}}{\theta^2}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}+1}\cdot\gamma\left(\frac{n\mu}{\theta}+1,\frac{\lambda}{\theta}e^{-\theta t}\right)
\end{aligned}$$

Remark. J_1 cannot be expressed via the incomplete Gamma function. Consequently, formulae that involve J_1 (see the next page), must be calculated either numerically, or by approximations, as discussed in the sequel.

2 Performance measures, exact formulae

Many important performance measures of the M/M/n+G queue can be conveniently expressed via the building blocks above. Define

P{Ab} – probability to abandon,
P{Sr} – probability to be served,
Q – queue length,
W – waiting time,
V – offered wait (time that a customer with infinite patience would wait).
Then

$$
\begin{aligned}
\mathrm{P}\{V>0\} &= \frac{\lambda J}{\mathcal{E}+\lambda J}\,, \\
\mathrm{P}\{W>0\} &= \frac{\lambda J}{\mathcal{E}+\lambda J}\cdot \bar{G}(0)\,, \\
\mathrm{P}\{\mathrm{Ab}\} &= \frac{1+(\lambda-n\mu)J}{\mathcal{E}+\lambda J}\,, \\
\mathrm{P}\{\mathrm{Sr}\} &= \frac{\mathcal{E}+n\mu J-1}{\mathcal{E}+\lambda J}\,, \\
\mathrm{E}[V] &= \frac{\lambda J_1}{\mathcal{E}+\lambda J}\,, \\
\mathrm{E}[W] &= \frac{\lambda J_H}{\mathcal{E}+\lambda J}\,, \\
\mathrm{E}[Q] &= \frac{\lambda^2 J_H}{\mathcal{E}+\lambda J}\,, \\
\mathrm{E}[W \mid \mathrm{Ab}] &= \frac{J+\lambda J_H-n\mu J_1}{(\lambda-n\mu)J+1}\,, \\
\mathrm{E}[W \mid \mathrm{Sr}] &= \frac{n\mu J_1-J}{\mathcal{E}+n\mu J-1}\,, \\
\mathrm{P}\{W>t\} &= \frac{\lambda\bar{G}(t)J(t)}{\mathcal{E}+\lambda J}\,, \\
\mathrm{E}[W \mid W>t] &= \frac{J_H(t)-(H(t)-t\bar{G}(t))\cdot J(t)}{\bar{G}(t)J(t)}\,, \\
\mathrm{P}\{\mathrm{Ab} \mid W>t\} &= 1-\frac{n\mu}{\lambda\bar{G}(t)}+\frac{\exp\{\lambda H(t)-n\mu t\}}{\lambda\bar{G}(t)J(t)}\,.
\end{aligned}
$$

3 Performance measures, QED approximations

Definitions and assumptions:

Patience-time density at the origin is positive: $G'(0) \triangleq g_0 > 0$.

Arrival rate $\lambda \to \infty$, and the number of agents is given by

$$n \;=\; \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda})\,, \qquad -\infty < \beta < \infty.$$

Let

$\Phi(x)$ – cumulative distribution function of the standard normal distribution (mean=0, std=1),

$\bar{\Phi}(x)$ – survival function ($\bar{\Phi} = 1 - \Phi$),

$\phi(x) \triangleq \Phi'(x)$ – density,

$h(x) \triangleq \phi(x)/\bar{\Phi}(x)$ – hazard rate.

Approximation formulae: Use $\beta = \left(n - \frac{\lambda}{\mu}\right) \Big/ \sqrt{\frac{\lambda}{\mu}}$, $\quad \hat{\beta} \triangleq \beta\sqrt{\frac{\mu}{g_0}}$.

$$\mathrm{P}\{\mathrm{V} > 0\} \;\approx\; \mathrm{P}\{\mathrm{W} > 0\} \;\approx\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1},$$

$$\mathrm{P}\{\mathrm{Ab}\} \;\approx\; \frac{1}{\sqrt{n}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{g_0}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1},$$

$$\mathrm{E}[V] \;\approx\; \mathrm{E}[W] \;\approx\; \mathrm{E}[W \mid \mathrm{Sr}] \;\approx\; \frac{1}{\sqrt{n}} \cdot \frac{1}{g_0} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{g_0}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1},$$

$$\mathrm{E}[W \mid \mathrm{Ab}] \;\approx\; \frac{1}{\sqrt{n}} \cdot \frac{1}{2\sqrt{g_0\mu}} \left[\frac{1}{h(\hat{\beta}) - \hat{\beta}} - \hat{\beta}\right],$$

$$\mathrm{E}[Q] \;\approx\; \sqrt{n} \cdot \frac{\mu}{g_0} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{g_0}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1},$$

$$\mathrm{P}\left\{W > \frac{t}{\sqrt{n}}\right\} \;\approx\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \cdot \frac{\bar{\Phi}\left(\hat{\beta} + \sqrt{g_0\mu} \cdot t\right)}{\bar{\Phi}(\hat{\beta})},$$

$$\mathrm{E}\left[W \,\middle|\, W > \frac{t}{\sqrt{n}}\right] \;\approx\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1}{g_0\mu}} \cdot \left[h\left(\hat{\beta} + \sqrt{g_0\mu} \cdot t\right) - \hat{\beta}\right],$$

$$\mathrm{P}\left\{\mathrm{Ab} \,\middle|\, W > \frac{t}{\sqrt{n}}\right\} \;\approx\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h\left(\hat{\beta} + \sqrt{g_0\mu} \cdot t\right) - \hat{\beta}\right].$$

Remark. For $\exp(\theta)$ patience (Erlang-A) simply replace g_0 in the formulae above by θ. (Equivalently, let g_0 be the reciprocal of the average service time.)

4 Performance measures, efficiency-driven approximations

Definitions and assumptions:

Arrival rate $\lambda \to \infty$, and the number of agents is given by

$$n \;=\; \frac{\lambda}{\mu} \cdot (1-\gamma) + o(\lambda)\,, \qquad \gamma > 0.$$

Assume that the equation $G(x) = \gamma$ has a unique solution x^*, namely $G(x^*) \;=\; \gamma$, or

$$x^* \;=\; G^{-1}(\gamma).$$

Approximation formulae.

$$\begin{aligned}
\mathrm{P}\{\mathrm{V} > 0\} \;\approx\; \mathrm{P}\{\mathrm{W} > 0\} \;&\approx\; 1\,,\\
\mathrm{P}\{\mathrm{Ab}\} \;&\approx\; \gamma\,,\\
\mathrm{P}\{\mathrm{Sr}\} \;&\approx\; 1-\gamma\,,\\
\mathrm{E}[V] \;&\approx\; x^*\,,\\
\mathrm{E}[W] \;&\approx\; H(x^*)\,,\\
\mathrm{E}[W \mid \mathrm{Sr}] \;&\approx\; x^*\,,\\
\mathrm{E}[W \mid \mathrm{Ab}] \;&\approx\; \frac{H(x^*) - x^*(1-\gamma)}{\gamma}\,,\\
\mathrm{E}[Q] \;&\approx\; \frac{n\mu}{1-\gamma} \cdot H(x^*)\,,\\
\mathrm{P}\{W > t\} \;&\approx\; \begin{cases} \bar{G}(t), & t < x^* \\ 0, & t > x^* \end{cases}\,,\\
\mathrm{E}[W \mid W > t] \;&\approx\; t + \frac{H(x^*) - H(t)}{\bar{G}(t)}\,, \qquad 0 \le t < x^*\,,\\
\mathrm{P}\{\mathrm{Ab} \mid W > t\} \;&\approx\; \frac{\gamma - G(t)}{\bar{G}(t)}\,, \qquad 0 \le t < x^*\,.
\end{aligned}$$

5 Guidelines for applications

5.1 Exact formulae: numerical calculations

The central issue is the computation of the building blocks (see the first page). Calculations for J, J_1 and J_H require two-stage integration which, numerically, could be time-consuming. However, $H(x)$ (the inner integral) often has a closed analytical form (as in the Exponential and Deterministic cases), and one is left with the external integrals that, as a rule, are to be evaluated numerically.

External integration: requires non-trivial programming for $n = 100$'s. (We did not go above $n = 1000$.) Special attention should be given to the approximation of integrals with ∞-upper-limit by finite-upper limits.

For calculating $\mathcal{E}$, define

$$\mathcal{E}_k \triangleq \frac{\sum_{j=0}^{k} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k}, \qquad k \geq 0,$$

and use recursion

$$\mathcal{E}_0 = 1; \qquad \mathcal{E}_k = 1 + \frac{k\mu}{\lambda} \cdot \mathcal{E}_{k-1}, \quad 1 \leq k \leq n-1; \qquad \mathcal{E} = \mathcal{E}_{n-1}.$$

5.2 QED approximation

Can be performed using any software that provides the standard normal distribution (e.g. Excel). This approximation works well for

- Number of servers n from 10's to 1000's;
- Agents highly utilized but not overloaded (~90-95%);
- Probability of delay 10-90%;
- Probability to abandon: 3-7% for small n, 1-4% for large n.

5.3 Efficiency-driven approximation

Requires solving the equation $G(x) = \gamma$, as well as integration (calculating $H(x^*)$). Both can be performed either numerically or analytically, depending on the patience distribution. This works well for

- Number of servers $n \geq 100$. (One can cautiously use n=10's, if the probability to abandon is large (>10%).)
- Agents very highly utilized (>95%);
- Probability of delay: more than 85%;
- Probability to abandon: more than 5%.