Mean value analysis of single server retrial queues

Jesus Artalejo and Jacques Resing

January 5, 2009, Eindhoven
Ordinary queue:

Retrial queue:
Starting Point:

Although retrial queues are in general much harder to analyze than ordinary queues (e.g., due to the varying arrival rates), mean performance measures are sometimes simple.

Example: $M/G/1$ retrial queue with exponential retrial times:

$$E(W) = \frac{\rho}{1 - \rho} \left( E(R) + \frac{1}{\mu} \right)$$

Can we explain these simple formulas?
Mean Value Analysis:

Approach to determine mean queue lengths and waiting times based on

- The PASTA property: Poisson Arrivals See Time Averages
- Little’s law: $E(L) = \lambda E(W)$

Advantage:
Probabilistic intuitive approach to determine mean values, resulting in linear equations only involving first moments
Overview:

1. Mean value analysis of:
   - the standard $M/G/1$ queue;
   - the main $M/G/1$ retrial queue;
   - the variant with batch arrivals;
   - the variant with only retrial of head of the orbit;
2. Other retrial models for which mean value analysis works
3. Retrial models for which mean value analysis does not work.
Mean Value Analysis:

Derive (i) arrival relation for the mean waiting time and use (ii) Little’s law

The standard $M/G/1$ queue

Arrival relation:

$$E(W) = E(L)E(B) + \rho E(R)$$

Little’s law:

$$E(L) = \lambda E(W)$$

Combination of the two equations yields

$$E(W) = \frac{\rho}{1 - \rho} E(R) = \frac{\rho}{1 - \rho} \frac{E(B^2)}{2E(B)}$$
Many variants of the standard $M/G/1$ queue can also be analyzed via mean value analysis, e.g.

- batch arrivals;
- priorities;
- unreliable servers;
- setup times;
- vacations;


Recently, the technique has also been applied to polling systems: Winands, Adan and van Houtum (Questa, 2006).

In this talk, we will concentrate on models with retrials.
The main $M/G/1$ retrial queue

Assumption: customers in orbit retry after an exponentially distributed time with parameter $\mu$.

Arrival relation:

$$E(W) = E(L)E(B) + \rho \left( E(R) + \frac{1}{\mu} \right)$$  \hspace{1cm} (1)

Little’s law:

$$E(L) = \lambda E(W)$$

Combination of the two equations yields

$$E(W) = \frac{\rho}{1 - \rho} \left( E(R) + \frac{1}{\mu} \right).$$
The main ideas to obtain (1) are the following:

- The waiting time $W$ can be written as the sum of the idle time of the server during the waiting time, $W_0$, and the busy time of the server during the waiting time, $W_1$,

\[ W = W_0 + W_1. \] (2)

- Given that a customer goes in orbit, the idle time of the server during the customer’s waiting time is exponentially distributed with parameter $\mu$ (follows from memoryless property of exponential retrial times).

Because, due to the PASTA property, the probability that an arbitrary customer goes into orbit equals the fraction of time, $\rho$, that the server is busy, we obtain

\[ E(W_0) = \frac{\rho}{\mu}. \] (3)
Mean waiting times are the same for systems with random order of service (ROS) and systems with first come first served (FCFS) service discipline. The same holds for mean idle times and mean busy times of the server during these waiting times.

Here, we mean the following with FCFS service discipline:

There is a FCFS discipline for the customers in orbit. Given that the number of customers in orbit is $j$, the retrial rate is $j\mu$. If a retrial occurs, the customer at the head of the orbit occupies the server.

Fresh arrivals finding the server free go to the orbit and, at that moment, the customer at the head of the orbit occupies the server.
• In the system with FCFS service discipline, the total expected busy time of the server during the waiting time of a customer is given by $E(L)E(B) + \rho E(R)$. Hence, we obtain

$$E(W_1) = E(L)E(B) + \rho E(R).$$

Now, the arrival relation (1) follows from a combination of the relations (2), (3) and (4):

$$E(W) = E(W_0) + E(W_1)$$

$$= \frac{\rho}{\mu} + E(L)E(B) + \rho E(R)$$

$$= E(L)E(B) + \rho \left( E(R) + \frac{1}{\mu} \right).$$
The $M/G/1$ queue retrial queue with batch arrivals

Batch size distribution: $x_k = P(X = k), \quad k \geq 1$.

Mean value relations

$$
E(W) = \left( E(L) + \sum_{k=1}^{\infty} r_k (k - 1) \right) E(B) + \rho E(R) + \frac{1 - r_1 (1 - \rho)}{\mu},
$$

$$
E(L) = \lambda E(X) E(W),
$$

where

$$
\rho = \text{ server utilization } = \lambda E(X) E(B),
$$

and

$$
r_k = \text{ probability that a customer is the } k\text{-th customer in his batch }.
$$

The probability an arbitrary customer is going into orbit equals $1 - r_1 (1 - \rho)$. 
The probabilities \( r_k \) are given by

\[
r_k = \frac{1}{E(X)} \sum_{n=k}^{\infty} x_n.
\]

Hence, it immediately follows that \( r_1 = 1/E(X) \) and

\[
\sum_{k=1}^{\infty} r_k(k - 1) = \frac{E(X^2) - E(X)}{2E(X)}.
\]

Combination of the results gives

\[
E(W) = \frac{1}{1 - \rho} \left( \frac{E(X^2) - E(X)}{2E(X)} E(B) + \rho E(R) \right) + \frac{E(X) - 1 + \rho}{(1 - \rho)\mu E(X)}.
\]
The model with general retrial times and only retrial of head of the orbit (Gómez-Corral (1999))

Only the customer at the head of the orbit is allowed to conduct retrials.

A new retrial time begins at a service completion epoch instead of at the moment of a service attempt failure.

Denote with the random variable $A$ a retrial time and with $\alpha(s)$ its LST. The minimum between $A$ and an exp($\lambda$) random variable is denoted as $\xi_A^\lambda$.

Mean value relations

\[
E(W_0) = (E(L) + \rho)E(\xi_A^\lambda),
\]
\[
E(W_1) = E(L)E(B) + \rho E(R),
\]
\[
E(L) = \lambda E(W).
\]

Remark: We now also use the FCFS service discipline to calculate the total expected idle time of the server during the waiting time.
Combination of these mean value relations yields

\[ E(W) = \frac{\rho(E(R) + E(\xi_A^\lambda))}{1 - \rho - \lambda E(\xi_A^\lambda)}. \]

Now, using the fact that

\[ E(\xi_A^\lambda) = \frac{1 - \alpha(\lambda)}{\lambda}, \]

we obtain

\[ E(W) = \frac{\rho}{\alpha(\lambda) - \rho} \left( E(R) + \frac{1 - \alpha(\lambda)}{\lambda} \right). \]
Other retrial models for which mean value analysis works:

- model with priority subscribers;
- model with impatient subscribers;
- model with network blocking;
- model with two-way communication.

Retrial models for which mean value analysis does not seem to work:

- discrete-time models;
- models with general retrial times;
- multi-server models;
- models with finite buffers and retrials.