Stochastic methods for measurement-based network control
Ellens, W.

Citation for published version (APA):
Chapter 7

Numerical results for a periodically observed queue

Outline. Chapter 6 considered the behaviour of a birth-death process for which the system state is known only at particular moments in time, because of periodic measurement practices. We considered four metrics that give insight into the performance of the system between two consecutive measurements of the system state, namely the probability to exceed a predefined level $m$ during a time interval with given initial and final states and three metrics related to the ‘severity’ of such an undesirable event. In the current chapter we demonstrate the procedures to compute these metrics by applying them to a queueing system inspired by cable access networks providing Internet access.

This chapter is organised as follows. We describe the queueing system (a processor-sharing type of model) in Section 7.1. Section 7.2 gives numerical results for this model, also showing the impact of various parameters on the four considered performance metrics. Then, in Section 7.3, we give an illustration of the practical use of our models in the context of capacity management for cable access networks. Section 7.4 discusses the main simulation results as well as some of the assumptions of the model. The chapter is followed by an appendix (Section 7.5) in which we analytically derive some interesting limiting results that we observed in our numerical study.
Chapter 7 Numerical results for a periodically observed queue

7.1 The queueing model

The model for our numerical study is motivated by cable access networks providing Internet access to its users. In this model jobs (file transmissions) arrive according to a Poisson process with rate $\lambda$, file sizes are exponential with mean $1/f$ and there is a maximum per-user transmission rate $R_{\text{max}}$. Users will receive this maximum rate as long as it can be accommodated, otherwise they will equally share the available capacity (denoted by $C$). In other words, the transition rate downward in the birth-death process equals

$$\mu_i = \min\{C, R_{\text{max}}i\}/f,$$

with $i$ the number of active users. This model can be regarded as processor-sharing model with maximum access rates and is essentially a hybrid variant between the classical $M/M/\infty$ (in which the total service rate is proportional to the number of active users) and $M/M/1$ (in which the total service rate is constant) queueing models, see e.g. [60, 61].

Throughout this section we choose the following default parameter values, unless stated otherwise: the capacity is $C = 800$ Mbps, the maximum access rate is $R_{\text{max}} = 80$ Mbps, the arrival rate is $\lambda = 100$ s$^{-1}$, the average file size is $f = 6$ Mb, the utilisation is $\rho = \lambda f/C = 0.75$ (75%, that is) the sample period $t = 2$ s. Given these values, users start to share the bandwidth when there are more than $C/R_{\text{max}} = 10$ active users. In steady state, the most probable state corresponds to 7 users. We choose the threshold to be $m = 15$ users, at which level the per-user throughput is two thirds of the maximum access rate (53 Mbps).

It is emphasised that in all experiments conducted, computation times are low, even when relatively high values of $k$ are used in the approximation $S_{m,t}^* \approx (S_{m,k\tau})^k$. Just to give an impression: computing the full matrix of entries $q_{m,i,j,t}$ (i.e. determining $q_{m,i,j,t}$ for all $i, j \in \{0, \ldots, m\}$ and given $m, t$) takes in the order of tens of milliseconds for $m = 15$. Also it should be noted that the described processor-sharing type of model serves as an example; the procedure developed in the preceding chapter can be applied to any birth-death process.
7.2 Numerical evaluation of the four metrics

In this section we present the numerical results obtained from applying the procedures of Chapter 6 to the model of Section 7.1. For time intervals of length $t$ with given initial value $X_0$ and final value $X_t$, we analyse the following four performance metrics:

- the probability that the number of active clients exceeds the ‘overload’ level $m$
  \[ q_t := 1 - q_{m,i,j,t}, \]

- the expected fraction of time the process spends above $m$
  \[ u_t := \mathbb{E}(U_t \mid X_0 = i, X_t = j)/t, \]

- the expected area below the graph of the number of clients and above $m$, per time unit
  \[ a_t := \mathbb{E}(A_t \mid X_0 = i, X_t = j)/t, \]

- and the expected number of arrivals while the process is above $m$, per time unit
  \[ n_t := \mathbb{E}(N_t \mid X_0 = i, X_t = j)/t, \]

for various settings of important parameters, being the sample period $t$, the initial state $X_0$, the final state $X_t$, the arrival rate $\lambda$, the average file size $f$ and the system capacity $C$. Note that the parameters $i$, $j$ and $m$ are suppressed in the above notation. For reasons of conciseness, we only show graphs for the expected values of the metrics. However, remind that the procedures also allow the computation of the variances, which are certainly important in practical situations.

In Figure 7.1 the dependence of the four performance metrics on the measured initial and final states (the number of active users at beginning and end of the sample period) is illustrated. A first crucial observation is that all four metrics are just mildly affected by the measured initial and final state as long as both these states correspond to at most 10 active users, being the point beyond which the bandwidth is shared. This
Chapter 7 Numerical results for a periodically observed queue

legitimates the fact that the calculations are based on the number of active users although in actual measurements the per-user throughput is monitored, since the former can be uniquely inferred from the latter whenever the bandwidth is shared, while the simulations show that knowledge of the exact number of active users is not needed when it cannot be inferred from the measured per-user throughput (i.e. if the per-user throughput is equal to the maximum per-user transmission rate).

Secondly, all four metrics are sensitive to the measured initial/final states when these exceed 10 active users, showing that throughput measurements provide valuable information.

Next we consider the impact of the sample period \( t \) on the performance metrics; see Figure 7.2. Interestingly, all graphs show non-monotonous behaviour for specific combinations of initial states, final states and values of \( m \). However, in practice, one is interested in measured values (in terms of the number of users) that are smaller than the threshold \( m \), in which case the graphs are monotonous. For these measured values, all metrics increase when increasing the sample period, confirming that in general the longer the time between measurements, the less informative these measurements are.

From our simulations we observed that the limits of the metrics when \( t \) goes to zero, only depend on the initial and final states; they are independent of the other parameters and (thus) do not depend on the specific queuing system (if the queueing system is a birth-death process). This fact is proved in Section 7.5. The analytical results presented in that appendix indicate that for very small sample periods \( (t \downarrow 0) \), the direct path between the initial and final states becomes the most likely path (in the sense that the likelihood of all other paths vanishes relative to the likelihood of the direct path), while the expected time spent in each state of this path is equal. On the other hand, when \( t \) goes to infinity the dependence on the initial and final state is lost and the performance metrics can be calculated using the stationary distribution of the underlying birth-death process. Moreover, a joint central limit theorem for \( U_t, A_t \) and \( N_t \) can be derived, see [4].

Numerical computations indicate that \( U_t, A_t \) and \( N_t \) (conditionally on \( X_0 = i, X_t = j \) ) are typically highly correlated. The correlation between \( U_t \) and \( A_t \) is roughly 88% in our experiments, the correlation between \( A_t \) and \( N_t \) about 90% and the correlation between \( U_t \) and \( N_t \) even almost

104
7.2 Numerical evaluation of the four metrics

Figure 7.1. Dependence of the four performance metrics on the initial and final states

99%. These correlations do not strongly depend on the chosen initial and final states, nor on the length of the sample period (in the range of ‘practically relevant’ lengths). The high correlations indicate that it is enough to monitor just one of the metrics, since it will be a reliable predictor for the other metrics. For this reason we decided to study in the upcoming plots only the metric $u_t$. The aforementioned result gives an indication that also in practice it is sufficient to compute just one of the metrics if the process evolves according to the model described in Section 7.1.
Chapter 7 Numerical results for a periodically observed queue

Figure 7.2. Dependence of the four performance metrics on the time between measurements, for a initial state of 7 active users

(a) Probability $q_t$ for the maximum number of active users to exceed $m = 15$

(b) Expected fraction of time $u_t$ over the level of $m = 15$ active users

(c) Expected area $a_t$ above level $m = 15$ per time unit

(d) Expected number of arrivals $n_t$ above level $m = 15$ per time unit

Figure 7.3 illustrates the influence of other parameters on $u_t$. There are clear similarities between the first three figures that can be explained as follows:

- Increasing the arrival rate, increasing the average file size and decreasing the capacity all ensure either an increasing number of arrivals or a decreasing number of departures, thus explaining an increase in $u_t$.
- Decreasing $C$ and increasing $f$ both lead to a decrease in the departure rates $\mu_i$ for states $i$ larger than $C/R_{\text{max}}$ (that is, when the
7.2 Numerical evaluation of the four metrics

(a) Dependence on the arrival rate
(b) Dependence on the average file size
(c) Dependence on the capacity
(d) The arrival rate and the average file size are varied together in such a way that the utilisation stays the same.

Figure 7.3. Dependence of the expected fraction of time $u_t$ over the level of $m = 15$ active users on different parameters, for an initial state of 7 active users. For (a) and (c) the x-axis has been cut-off at a utilisation of 90%, that is $\lambda < 120$ s$^{-1}$, $f < 7.2$ Mb, $C > 667$ Mbps.

bandwidth has to be shared). However, only increasing $f$ affects the $\mu_i$s for the other states. This explains that the differences between both changes are smaller for larger final states (as a consequence of the fact that then the number of active users is above $C/R_{\text{max}}$ more frequently).

Note that increasing the arrival rate, increasing the average file size, or
decreasing the capacity, all increase the utilisation. In Figure 7.3d, the arrival rate is increased while keeping the utilisation fixed by decreasing the file size simultaneously. Essentially, this type of change corresponds to a change in time unit. Therefore, the resulting graph is exactly the same as in Figure 7.2b, where the sample period is altered. The figure illustrates that knowledge of the utilisation alone does not give much information about the throughput perceived by the users.

7.3 Illustration of practical use

In this section we again focus on the setting of cable access networks, modeled via the processor-sharing model with maximum access rates and illustrate how operators can use our technique. For an operator, to prevent SLA violations, it is relevant to quickly detect throughput degradations experienced by their users. We assume the operator can perform periodic per-user throughput measurements, e.g. via speed tests, as well as utilisation measurements over the sample periods (the periods between the throughput measurements).

We assume that the flow arrival process between two consecutive throughput measurements (say measurement $j$ and $j + 1$) can be ‘locally’ approximated by a homogeneous Poisson process with rate $\lambda(j)$, while we impose the common assumption of relatively stable flow-size characteristics over a longer time interval (i.e. many sample periods) and that the expected flow size $f$ over this time interval is known. As a result, $\lambda(j)$ can be estimated from the utilisation during the sample period.

Using our approach, given the operator’s parameter settings (i.e. $C$ and $R_{\text{max}}$), charts like the one shown in Figure 7.4 can be produced for different measured utilisations (caused by different arrival rates). These provide the operator with useful insight into the throughput perceived by the users. Given two subsequent throughput measurements, this type of graph gives a ‘safe’ estimate of the throughput between these measurements, in the sense that the real throughput will be lower than this estimate not more than $x\%$ of the time (in the example we took $x = 10\%$). Then, for instance if this throughput drops below 40 Mbps an alarm could be raised. Note that the symmetry of Figure 7.4 is a consequence of the symmetry of $U_t$, as was pointed out at the end of
Figure 7.4. Largest throughput (in Mbps) such that the expected fraction of time that the actual throughput is lower is less than 10%. The utilisation equals 75%.

Section 6.2

The throughput estimates can also be calculated on the fly, producing plots in the spirit of Figure 7.5. This particular plot is based on a (simulation) experiment in which we set the arrival rates between subsequent measurements at specific (predefined) values, resulting in a utilisation that varies over time (as depicted in Figure 7.6). In Figure 7.5 the (simulated) throughput measurements are shown with circles. The ‘safe’ throughput estimates, i.e. those calculated in the same way as in Figure 7.4 are depicted by the line. It can be seen that the measurements and the estimates of the throughput between the measurements are often quite far apart, which illustrates the added value of calculating the estimate.

7.4 Conclusion

To illustrate the practical use of the procedures derived in Chapter 6 we have presented a set of numerical experiments for a model inspired by cable access networks. The key idea behind this model is that the user throughput in such a network depends on the number of simultaneously active users; every user gets a maximum access rate whenever this can be accommodated by the available capacity, otherwise the bandwidth is equally shared. Interestingly, in the numerical experiments the correlation between the time, area and number of arrivals above \( m \) is typically very high, which suggests that it is sufficient to keep track of only one of these
Chapter 7 Numerical results for a periodically observed queue

![Figure 7.5](image1.png)  ![Figure 7.6](image2.png)

**Figure 7.5.** Simulated periodic throughput measurements (circles) and the corresponding estimated intermediate throughputs

**Figure 7.6.** The utilisation values used as an input for the simulations of Figure 7.5

Bearing in mind that the presented approach assumes Poisson arrivals and exponential service times, an important topic for further research concerns its validation in practical situations. One could for instance compare our results to simulations with real network traffic traces, or to simulations with input processes that are good approximations of real traffic, so as to assess the robustness of the procedure. The Poisson arrivals are justified by the fact that the resource is used by a large number of potential users. While following a diurnal pattern, the Poisson arrival rate can be assumed stationary during substantial periods of time; observe that the Poisson rate changes at a timescale which is orders of magnitude longer than that corresponding to the measurements.

Regarding the service times, it is first noted that in our birth-death model the stationary distribution of the number of customers in the system is *insensitive*, in that it depends on the service-time distribution through its mean only. The transient distribution, however, *does* depend on the full service-time distribution. To make our approach more widely applicable, it may seem a logical step to try to lift the exponentiality assumption by approximating the service-time distribution by its phase-type counterpart. At the conceptual level, however, such extensions are non-trivial: one can typically measure the number of jobs in progress (as assumed in our numerical example), but not the phase each of them is in.
7.5 Appendix: limits for small time-scales

In Section 7.2 we made some observations regarding the behaviour of \( q_t, u_t, a_t \) and \( n_t \) for small \( t \). In this appendix we analyse this behaviour analytically. The intuition is that in this limiting regime, the process jumps, with overwhelming probability, exactly \( |i - j| \) times between the initial state \( X_0 \) and the final state \( X_t \). Perhaps surprisingly, we prove that in the limit the portion of time in each of the \( |i - j| + 1 \) states of the path is equal (and hence equals \( 1/(|i - j| + 1) \)) — this fraction does not depend on the specific values of the transition rates. Based on these findings, we obtain the asymptotics of \( u_t, a_t \) and \( n_t \).

We start by a useful lemma. Write \( S_m := \sum_{i=1}^{m} V_i \), with the \( V_i \) independent and exponentially distributed, with rates \( \tau_1 \) up to \( \tau_m \).

**Lemma 7.1.** As \( t \downarrow 0 \),

\[
\mathbb{P}(S_m \leq t) = \frac{1}{m!} \left( \prod_{i=1}^{m} \tau_i \right) t^m + o(t^m).
\]

**Proof.** We have that, with \( f(\cdot) \) denoting the density of \( S_m \) and \( f^{(n)}(\cdot) \) its \( n \)-th derivative,

\[
\mathbb{P}(S_m \leq t) = \sum_{k=0}^{m-1} \frac{f^{(k)}(0)}{(k+1)!} t^{k+1} + o(t^m). \tag{7.1}
\]

By induction to \( k \) it is verified that, relying on repeated integration by parts in the induction step,

\[
f^{(k)}(0) = \lim_{s \to \infty} \int_{0}^{\infty} s e^{-sx} f^{(k)}(x) dx = \lim_{s \to \infty} \int_{0}^{\infty} s^2 e^{-sx} f^{(k-1)}(x) dx
\]

\[
= \cdots = \lim_{s \to \infty} s^{k+1} \mathcal{R}(s) = \begin{cases} 
\tau_1 \tau_2 \cdots \tau_m & \text{for } k = m - 1 \\
0 & \text{for } k < m - 1,
\end{cases}
\]

with \( \mathcal{R}(s) \) the Laplace transform of \( S_m \), i.e.

\[
\mathcal{R}(s) := \mathbb{E}(e^{-sS_m}) = \prod_{i=1}^{m} \frac{\tau_i}{\tau_i + s}.
\]
Filling in the expressions for $f^{(k)}(0)$ (for $k \leq m - 1$) in (7.1), we find the stated.

In the following result we return to our birth-death process and establish the property that in the regime $t \downarrow 0$ the number of transitions of $X_t$, conditioning on $\{X_0 = i, X_t = j\}$, is, with overwhelming probability, $|i - j|$.

**Proposition 7.2.** Let $\bar{N}_t$ be the number of transitions during $[0, t]$. Then,

$$
\lim_{t \downarrow 0} P(\bar{N}_t = |i - j| \mid X_0 = i, X_t = j) = 1.
$$

**Proof.** We first observe the relation

$$
P(\bar{N}_t = |i - j| \mid X_0 = i, X_t = j) = \frac{\varrho_t(|i - j|)}{\sum_{k=0}^{\infty} \varrho_t(|i - j| + 2k)},
$$

with $\varrho_t(k) := P(\bar{N}_t = k, X_t = j \mid X_0 = i)$, which holds since the number of transitions from $i$ to $j$ is at least $|i - j|$ and the extra number of transitions must be even. Let us consider the case $j \geq i$; the complementary case can be dealt with analogously. In this case we have

$$
\varrho_t(|i - j|) = \left(\prod_{k=1}^{j-i} \frac{\lambda_{i+k-1}}{\nu_{i+k-1}}\right) \left(\mathsf{P} \left(\sum_{k=1}^{j-i} T_k < t\right) - \mathsf{P} \left(\sum_{k=1}^{j-i+1} T_k < t\right)\right),
$$

where $\nu_i := \lambda_i + \mu_i$ and the $T_k$ are independent, with $T_k$ exponentially distributed with rate $\nu_{i+k-1}$. Relying on Lemma 7.1, we thus find that, for a constant $K$,

$$
\varrho_t(|i - j|) = K t^{j-i} + o(t^{j-i}).
$$

The next step is to realise that

$$
\sum_{k=0}^{\infty} \varrho_t(|i - j| + 2k) \leq \varrho_t(|i - j|) + \mathsf{P} \left(\sum_{k=1}^{j-i+2} \bar{T}_k < t\right),
$$

where the $\bar{T}_k$ are independent, with each of the $\bar{T}_k$ exponentially distributed with rate $\bar{\nu} := \max_{i \leq l \leq j} \nu_l$. Using standard results for the
Erlang distribution (or Lemma 7.1 again), we have that
\[
P\left(\sum_{k=1}^{j-i+2} \bar{T}_k < t\right) = \frac{\nu^{j-i+2}}{(j-i+2)!} t^{j-i+2} + o(t^{j-i+2}). \tag{7.5}\]

Combining the expressions \((7.2)-(7.5)\) implies the stated. \(\square\)

The next result concerns the portion of time that is spend in each of the states when the total time \(t\) goes to zero. Consider, as before, the situation of \(S_m := \sum_{i=1}^{m} V_i\), where the \(V_i\) are independent and exponentially distributed, with rates \(\tau_1\) up to \(\tau_m\).

**Proposition 7.3.** For \(m \geq 1\) it holds that
\[
\lim_{t \downarrow 0} \frac{\mathbb{E}(V_m \mid S_m = t)}{t} = \frac{1}{m}.
\]

**Proof.** In self-evident notation,
\[
\mathbb{E}(V_m \mid S_m = t) = \frac{\int_{0}^{t} s f_{V_m}(s) f_{S_{m-1}}(t-s) ds}{tf_{S_m}(t)}.
\]

Using the same technique as above in Lemma 7.1,
\[
f_{S_m}(t) = \frac{1}{(m-1)!} \left(\prod_{i=1}^{m} \tau_i\right) t^{m-1} + o(t^{m-1})
\]
and observing that
\[
\int_{0}^{t} s \tau_m \frac{1}{(m-2)!} \left(\prod_{i=1}^{m-1} \tau_i\right) (t-s)^{m-2} ds = \frac{1}{m},
\]
we arrive at the claimed property. \(\square\)

We now combine the findings of the above propositions. Based on Proposition 7.2 we know that as \(t\) becomes very small, with high likelihood
Chapter 7 Numerical results for a periodically observed queue

The birth-death chain has $|i - j|$ (i.e. the minimal number) transitions. Proposition 7.3 entails that the sojourn times in states $i$ up to $j$ all have an expected duration in the order of $t/(|i - j| + 1)$.

Regarding the four overload metrics, we thus observe that

- for the probability $q_t$ to exceed $m$,

$$\lim_{t \downarrow 0} q_t = \begin{cases} 
0 & \text{if } i, j \leq m \\
1 & \text{otherwise};
\end{cases}$$

- for the expected fraction of time $u_t$ above $m$,

$$\lim_{t \downarrow 0} u_t = \begin{cases} 
0 & \text{if } i, j \leq m \\
(j - m)/(j - i + 1) & \text{if } i \leq m < j \\
(i - m)/(i - j + 1) & \text{if } j \leq m < i \\
1 & \text{if } i, j > m;
\end{cases}$$

- for the expected area $a_t$ above $m$ per time unit,

$$\lim_{t \downarrow 0} a_t = \begin{cases} 
0 & \text{if } i, j \leq m \\
(j - m + 1)(j - m)/(2(j - i + 1)) & \text{if } i \leq m < j \\
(i - m + 1)(i - m)/(2(i - j + 1)) & \text{if } j \leq m < i \\
(i + j - 2m)/2 & \text{if } i, j > m;
\end{cases}$$

- for the expected number of arrivals $n_t$ above $m$ per time unit,

$$\lim_{t \downarrow 0} n_t = \begin{cases} 
0 & \text{if } i, j \leq m \\
\infty & \text{otherwise}.
\end{cases}$$