Approximate Analysis of a Network of Fluid Queues

Tony Field and Peter Harrison
Department of Computing, Imperial College London SW7 2AZ, UK.
{ajf,pgh}@doc.ic.ac.uk

May 24, 2007

Abstract

Fluid models have for some time been used to approximate stochastic networks with discrete state. These range from traditional 'heavy traffic' approximations to the recent advances in bio-chemical system models. Here we use an approximate compositional method to analyse a simple feedforward network of fluid queues which comprises both probabilistic branching and superposition. This extends our earlier work that showed the approximation to yield excellent results for a linear chain of fluid queues. The results are compared with those from a simulation model of the same system. The compositional approach is shown to yield good approximations, deteriorating for nodes with high load when there is correlation between their immediate inputs. This correlation arises when a common set of external sources feeds more than one queue, directly or indirectly.

1 Introduction

Stochastic fluid flow models have for long been used to describe networks of nodes that provide service to traffic of some sort that flows amongst them. Such models can exactly describe systems with continuous state, for example volumes in literal fluid flows, but more commonly are used to approximate discrete state systems of traffic flows.

The single fluid queue has been studied in some depth and results exist under quite general assumptions about the input processes—see for example [1]. Various flavours of tandem queues, fed by on/off processes have also been analysed to obtain the steady-state joint distribution of the fluid levels in each queue, from which various measures can be derived, e.g. [2, 3, 4]. Similar exact results for more general networks have also been produced using Martingale methods [5, 6].

In our previous work we presented a simple approximate approach to the analysis of steady-state fluid queue chains with a single on/off Markov modulated external arrival process. For these linear chains, the approximation was shown to be excellent, even under high load. In this short paper, we extend the analysis to a more complex, feedforward network, which contains proportional branching and superposition of streams (joins). The accuracy of the approximation is evaluated by comparison with simulation.

2 The System

We consider a network with three stages of fluid queues (or nodes) and two independent Markov modulated on-off arrival processes (MMOAPs), each with parameters 5, 5, and 20, as shown in Figure 1. One MMAP feeds the upper queue of stage 0; the other feeds the lower queue. In stage 0 of the network, the output of each node is split, with 50% of the flow being passed to the upper queue of stage 1 and 50% to the lower. The outputs of stage 1 are similarly split in the ratio 70:30. The nodes of stage 2 feed sink nodes, so the output routing is unimportant.

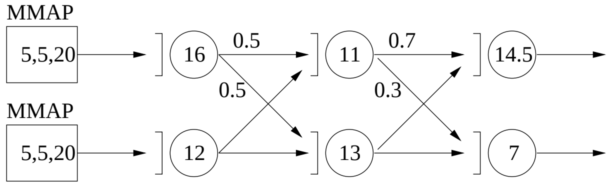

Figure 1: The fluid queueing network

The fluid service rates for stage 0 (Upper) and stage 0 (Lower) are the constants 16 and 12 respectively. Those of stage 1 are 11 and 13 respectively, and in stage 2 they are 14.5 and 7.

3 Approximate analysis

Our approximation method is motivated by the observation that the output process of each fluid queue in such a network is also an on/off process, although the busy periods will not be exponentially distributed. We approximate this on/off process by a three-state Markov chain that modulates its rate – either 0 or its specified constant service rate. One of the states of the chain reproduces the off state of the queue. The sum of the times spent in the other two states of the Markov chain seeks to approximate the busy period of the queue in question. The generators of the approximating Markov chain are estimated by matching the first three moments of the generated on-periods with the first three moments of the busy period being approximated. The latter comes from an exact analytical model of a single fluid queue.

3.1 Busy period moments

Consider a single on-off fluid arrival process with exponentially distributed off time, parameter b, and constant fluid on-rate λ. This process feeds a fluid server that has constant fluid output rate μ and infinite fluid reservoir (corresponding to a discrete queue). Let the busy period (respectively, on-period) random variable be denoted by W (respectively V), with probability distribution, density function and Laplace transform $V(t), v(t) = V'(t)$ and $V^*(\theta)$ (respectively, $W(t), w(t) = W'(t)$ and $W^*(\theta)$). Then, it is easy to show that [7]

$$W^*(\theta) = V^*\big(\theta\rho + b(\rho - 1)(1 - W^*(\theta))\big)$$

where $\rho = \lambda/\mu > 1$. When there are $n \geq 2$ MMOAP input streams, rather more effort is required and it is shown in [1] that, for $1 \leq i, j \leq n$,

$$W_i^*(\theta) = V_i^*\Big(\theta\rho_i + b_i(\rho_i - 1)(1 - W_i^*(\theta)) + \rho_i \sum_{j \neq i} b_j(1 - W_j^*(\theta))\Big)$$

where b_i, λ_i, V_i are defined in the obvious way for the ith arrival stream and W_i is the length of a busy period that starts with input stream i changing from its off- to on-state. The unconditional busy period W then has pdf with Laplace transform

$$W^*(\theta) = \sum_{i=1}^{n} b_i W_i^*(\theta)/b$$

where $b = b_1 + \ldots b_n$.

Differentiating this result three times yields the following equations for the first three moments of the busy period, M_1, M_2, M_3:

$$M_k = \sum_{i=1}^{n} b_i M_{ki}/b$$

for $k = 1, 2, 3$, where

$$\begin{aligned}
M_{i1} &= m_{i1}(\rho_i + b_i(\rho_i - 1)M_{i1} + \sum b_j\rho_i M_{j1}) \\
M_{i2} &= m_{i2}(M_{i1}/m_{i1})^2 + m_{i1}(b_i(\rho_i - 1)M_{i2} + \sum b_j\rho_i M_{j2}) \\
M_{i3} &= m_{i3}(M_{i1}/m_{i1})^3 + 3m_{i2}(M_{i1}/m_{i1})(b_i(\rho_i - 1)M_{i2} \\
&\quad + \sum b_j\rho_i M_{j2}) + m_{i1}(b_i(\rho_i - 1)M_{i3} + \sum b_j\rho_i M_{j3})
\end{aligned}$$

The summations are taken over $\{j \mid 1 \leq j \neq i \leq n\}$ and m_{ik} is the kth moment of the on-period of stream i.

3.2 Fluid queue building block

Given a set of input MMOAPs, each characterized by its first three moments, to a fluid queue with given constant rate, the first three moments of the busy period of that fluid queue can be determined by the above formulas. These are then used to approximate the queue's output as a 3-phase MMOAP, by estimating the generators of the modulating Markov chain using moment matching. Because all external streams' off-periods are exponential random variables, the same must be true at every node; any node receiving fluid must also be emitting fluid. Hence an input to a node will be idle if and only if all the upstream nodes behind it are also idle and every external source upstream is in the off-state. The transition rate from the off-state is therefore straightforward. The matching is done as described in [7] using least squares minimization. The moments of the time spent in each of the two on-states can be calculated in terms of the generators of the Markov chain, and these are used to calculate the moments of the on-period of the output process.

If the output of a node is 'split', i.e. directed proportionally to more than one other node (or out of the network), the input stream received by each downstream node is characterised by the same on-off process as the whole output process that is split. However, the fluid input rate on that stream is reduced in the given proportion.

This allows us to define a building block to approximate each queue in a feed-forward network, in which the nodes can be sequenced so that the input process of each subsequent node analysed is known, starting the sequence with a node having only external input. We examine properties of the fluid level in each node, such as the probability distribution of the fluid level at equilibrium, assuming this exists, using a standard solution for the single fluid queue with Markov modulated input:

- We assume that the input streams to each queue are independent (but note the error this approximating assumption can cause in the next section), so that the aggregate input MMOAP is simply the Kronecker sum of the constituent streams;
- The number of phases in the aggregate process is therefore the product of the numbers of phases of the constituents;
- The node can only be idle (no output) in the single state when all constituents are in the off-state;
- However, the fluid level can be zero in any aggregate phase with arrival rate less than the node's service rate – this leads to numerical complications (not present in the tandem case [7]) due to singularities in the solution and more complicated boundary conditions, especially when there are degeneracies such as occur when superposing identical MMOAPs.

We used Mathematica to perform our calculations of the mean and standard deviation of each node's fluid level at steady state.

4 Results and speculation

We applied the approximation to the network of Figure 1 and calculated the mean and standard deviation of the fluid level at each queue. The results are shown in Table 1 for both the analytical model and simulation. For the simulation results, the 90% confidence intervals are shown in parentheses, expressed as a percentage.

It can be seen that the approximate model performs reasonably at all nodes but the upper right one, where it is clearly poor. In the first stage the result is exact, of course, and the agreement serves as a check on the simulation. The accuracy in the second stage is much worse than we obtained for tandem networks which initially seemed surprising. The lower right node has low loading and so can be expected to yield better agreement than at the top right node which is heavily loaded: an average input rate of 14 and service rate of 14.5. There are a number of causes for the discrepancies, which we now consider.

The input MMOAPs are independent, so that the analytical results for the nodes they feed are exact. Similarly, in stage 1, each node receives as its input the output from the independent fluid queues in stage 0. So this does not explain the imprecision in this stage. However, in stage 2, when any node – say the top one – is in the idle state, with no input, all the other nodes in this stage must also have no input because they all have common input nodes, viz. the nodes in stage 2. Therefore, there is no transition in the modulating Markov chain of the top (or any other) node from the off-state to the state with a *single* input node in its on-state. Although other joint transitions may be affected similarly, there will be joint states in which one input is off and the other is on, since the asymmetric queues will not, in general, become empty simultaneously. In the case of symmetric queues in each stage, of course, there are more constraints, certain joint states are invalid, and the upper and lower tiers will work in lock-step.

To test the significance of the dependency we have in the last stage, we modified the joint state transition rate matrix by setting the rate from the (joint) idle state – i.e. with both inputs in stage 1 in off-state, to the state with both inputs on – a simultaneous transition not valid in the Kronecker sum. This gave mean and standard deviation for the equilibrium fluid level of 9.662 and 10.751 respectively. At the lower stage 2 node we similarly obtained

Level	Measure	Stage 0	Stage 1	Stage 2
Upper	Mean	0.667	3.153	6.950
	Sim Mean	0.666	2.947	14.03
		(0.08)	(0.32)	(0.94)
	S.D.	0.989	3.804	8.158
	Sim S.D.	0.989	3.939	17.226
		(0.07)	(0.28)	(1.33)
Lower	Mean	4.0	0.283	0.090
	Sim Mean	3.996	0.248	0.096
		(0.14)	(0.17)	(0.44)
	S.D.	4.733	0.447	0.132
	Sim S.D.	4.726	0.438	0.170
		(0.16)	(0.22)	(0.73)

Table 1: Mean and standard deviation of the fluid levels at each node

0.0922 and 0.132. This gives better agreement with simulation, but there is still clearly more to do, especially at the heavily loaded node.

A further problem is that our output model from each fluid node is inadequate when there are states with input rate less than the service rate of the node. This occurs when a subset of the input sources to a fluid queue are in on-state but the sum of their rates is less than the service rate – for example all of the nodes in our test case apart from the lower one in stage 2. As already described, in our MMOAP streams, there are three states, with fluid arrival rates equal to 0 in the off-state and the specified constant service rate in both of the other states. This does not allow for the lower output rates that occur when the queue is empty but there is, nevertheless, an input from one source. We propose to address this by either generalizing the node building block or perhaps by utilising the steady state probabilities that the fluid level is zero in each input state.

5 Summary and Conclusion

We have considered some of the problems that occur when we extend our highly promising fluid-flow approximation from tandem networks to more general networks. Although we have maintained excellent results for specialised cases, such as tree-like networks, where there are no correlations between different nodes, the effect of common input sources has proved problematic in highly loaded nodes. The research has highlighted specific aspects of the model that can be improved and indicated promising future directions to investigate.

References

[1] O. Boxma and V. Dumas, "The busy period in the fluid queue," in *Proceedings of the 1998 ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, (Madison), pp. 100–110, ACM, June 1998.

[2] L. Rabehasaina, "Moments of a Markov-modulated, irreducible network of fluid queues," *Journal of Applied Probability*, vol. 43, no. 2, pp. 510–522, 2006.

[3] D. Kroese and W. Scheinhardt, "Joint distributions for interacting fluid queues," *Queueing Systems*, vol. 37, no. 1–3, pp. 99–139, 2001.

[4] W. Scheinhardt and B. Zwart, "A tandem fluid queue with gradual input," *Probability in the Engineering and Informational Sciences*, vol. 16, pp. 29–45, 2002.

[5] O. Kella, "Markov-modulated Feedforward Fluid Networks," *Queueing Systems*, vol. 37, no. 1–3, pp. 141–161, 2001.

[6] O. Kella and W. Whitt, "Useful martingales for stochastic storage processes with Levy input," *Journal of Applied Probability*, vol. 29, no. 2, pp. 396–403, 1992.

[7] A. Field and P. Harrison, "An approximate compositional approach to the analysis of fluid queue networks," in *Performance '07*, (Cologne), Elsevier, 2007. Submitted.