Fluid Limits of Queueing Networks with Batches

Luca Bortolussi
Department of Mathematics
and Computer Science
University of Trieste, Italy
luca@dmi.units.it

Mirco Tribastone
Institute of Informatics
Ludwig-Maximilians-Universität Munich,
Germany
tribastone@pst.ifi.lmu.de

ABSTRACT

This paper presents an analytical model for the performance prediction of queueing networks with batch services and batch arrivals, related to the fluid limit of a suitable single-parameter sequence of continuous-time Markov chains and interpreted as the deterministic approximation of the average behaviour of the stochastic process. Notably, the underlying system of ordinary differential equations exhibits discontinuities in the right-hand sides, which however are proven to yield a meaningful solution. A substantial numerical assessment is used to study the quality of the approximation and shows very good accuracy in networks with large job populations.

Categories and Subject Descriptors

I.6.5 [**Simulation and Modeling**]: Model Development—*Modeling methodologies*; D.2.8 [**Software Engineering**]: Metrics—*Performance measures*

General Terms

Performance

Keywords

queueing networks, batch services, fluid limits

1. INTRODUCTION

Batches are useful in the study of computer and communication systems for describing situations when an event gives rise to the simultaneous arrival of more than one element, or when servers accumulate a certain number of jobs before processing them so as to reduce, for instance, overheads in communication bandwidth [1].

There is a vast body of literature concerned with the analysis of batch systems, especially within queueing theory, with references which may be tracked as far back as the Twenties with Erlang's solution of the $M/E_k/1$ queue, which may also be used to yield that of the $M^k/M/1$ queue. The book by Chaudhry and Templeton provides an exhaustive account of analyses of queueing systems with batch (or *bulk*) arrivals and service, both for transient and steady-state solutions [2].

The present paper considers queueing networks with open batch arrivals and batch services which can be described in terms of a continuous-time Markov chain (CTMC), with state descriptor characterised by a population vector which gives the job population in each station of the network. Models of this kind have been studied in the past, mostly with the aim of extending classical product-form solutions of ordinary queueing networks where jobs arrive at the network, transfer between nodes, and receive service *singly* [3, 4]. The works by Henderson and Taylor [5] and Henderson *et al.* [6] have provided product forms for a class of open and closed networks, respectively. Despite the considerable value from a theoretical viewpoint, these results present the drawback that in practical applications the computational cost of the normalising constant may be prohibitive, especially when analysing networks with large job populations.

The technique herein presented is instead based on an approximation in terms of a *fluid* model. In a classical setting, for a given network under study a sequence of CTMCs indexed by a single parameter, hereafter denoted by N, is suitably constructed so as to be shown to converge asymptotically to a system of ordinary differential equations (ODEs). The parameter is usually referred to as the network's size, e.g., the larger N the larger the initial population levels in the system. The limiting fluid behaviour is shown to be undistinguishable from a sample path of the CTMC for $N \to \infty$, thus justifying the ODE solution as an analytical approximate of the average behaviour of the network for large N. The framework is that of Kurtz, who has proven this form of convergence under relatively mild assumptions on the nature of the transitions in general CTMCs with a population-based state descriptor [7]. A brief overview of related work concerning fluid models, with applications to computer and communications systems, is provided in Section 2. This is followed by Section 3, where we present the relevant notation for the fluid framework considered in this paper.

The mathematics used to describe queueing networks with batches is discussed in Section 4 by means of a queueing system with Poisson batch arrivals at a station that serves singly. Two forms of scaling will be discussed which turn out to lead to different limit behaviours. The first — and perhaps the least surprising — case concerns a sequence of

ICPE'12, April 22–25, 2012, Boston, Massachusetts, USA

CTMCs where the batch size is constant and the arrival intensity grows linearly with N (Section 4.1). This case belongs to the aforementioned standard framework of Kurtz. The other scaling considers the situation when the arrival rate is constant and the batch size is allowed to grow with N (Section 4.2). In this case, instead, the limit behaviour is a stochastic hybrid system (cf. [8]) which mixes continuous flows with Markovian jumps. However, also in this case a fluid ODE can be syntactically constructed, and its relationship with the hybrid limit will be discussed.

In Section 4.3, the running example is varied to analyse a queue with finite capacity. The (illustrative) purpose is to introduce another form of limit behaviour, namely that of an ODE with discontinuous right-hand side. To build some intuition as to how this arises from inherently discontinuities in the CTMC transitions rates, let c be the queue capacity and $b < c$ the arrival batch size. Then, the arrival rate will be some $\lambda > 0$ if the queue length is less than $c - b$ and 0 otherwise. Under these circumstances Kurtz's theorem cannot be applied, as it requires Lipschitz continuity of the ODE vector field. However, using recent developments concerned with non-smooth systems [9, 10], we show that such a fluid limit with discontinuities is meaningful. Clearly, an analogous form of discontinuity presents itself in the case of batch services. This situation is studied in detail in Section 5, which considers two forms of scaling that give rise to a deterministic limit and to a hybrid one.

Section 6 unifies these results in a general model of Markovian queueing networks with batch services and batch arrivals. The model is accompanied by a discussion in Section 6.2 concerning its applicability to practical situations, with emphasis on the impact of the forms of scaling studied in this paper. The natural question as to whether and under which conditions the deterministic trajectory may be used as an approximation to the expected behaviour of the stochastic process is investigated in Section 7 by means of a substantial numerical study. It confirms that the quality of the approximation improves with increasing population sizes, yielding accurate estimates for medium/large sized networks under a wide range of traffic conditions.

The paper ends in Section 8 with concluding remarks. In particular, we sketch a methodology to help the modeller choose between different analysis options—numerical solution of the underlying CTMC, stochastic simulation, or fluid approximations—according to the nature of the actual system under consideration.

2. RELATED WORK

Mean field and fluid approaches have a long-standing tradition in performance engineering and in queuing theory. Recently, general frameworks to apply mean-field asymptotic results, with limits defined in terms of ODEs, have been developed [11, 12, 13, 14, 10, 9]. Some of them deal with discrete-time Markov chain models, and show convergence under a suitable scaling of transition kernels and duration of a time step of the chain [11, 12]. Other deal with CTMC models [13, 14], possibly connecting the mean field approximation with high level formal languages to describe systems [13]. In all cases, Lipschitz continuity is required for the rates.

As discussed above, extensions of such frameworks to discontinuous rates and kernels, including new convergence results, have been proposed in [9, 10]. Our approach uses these works to study approximations for queueing networks with batches. However, the contribution of this paper goes beyond a mere application of these results since we also consider non-fluid forms of scaling, giving rise to hybrid systems, which are not considered in all the aforementioned papers. To the best of these authors' knowledge, there has not been any application of fluid approximation to queue models with batches.

In the literature, many of the applications of mean-field limits for specific systems are concerned with models in which the assumption on Lipschitz continuity holds. Without pretending to be exhaustive, we recall recent work on the analysis of MAC protocols [12, 15], peer-to-peer protocols [16, 17], TCP protocols (with emphasis on data centers), [18, 19]), and load balancing policies [20, 21].

Similarly to us, [17] also studies a Piecewise Deterministic Markov Process [8]. However, the simple structure of their hybrid model, which permits to decouple stochastic dynamics from deterministic behaviour, enables analytical solutions. This is harder in our setting, due to the strong bidirectional coupling of the two dynamical regimes. This is why we focussed instead on approximate techniques, see Sections 4.2 and 8.

The authors of [18], instead, consider a fluid model based on delay differential equations which contain discontinuities in the right-hand side, induced by congestion control policies. However, contrary to queues with batches, such discontinuities have no dramatic impact on the dynamics (there is no sliding motion). In [19], instead, the focus is more on the control policy, studied from the point of view of control theory. Both [17] and [18] focus on the analysis of the fluid model and provide only experimental evidence of convergence of the pure stochastic system, without discussing the quality of approximation in detail.

The paper [12] considers a mean field approach that is different from the ones used in this paper. In particular, their limit result, proved in the paper, is concerned with Lipschitz continuous rates in a rapidly varying environment, that reaches instantaneously equilibrium in the limit. In [15], instead, the authors use a more classical mean field approach, with a limit in continuous times for Lipschitz continuous rates, and apply also a central limit result (i.e. a limit in terms of Stochastic Differential Equations with Gaussian noise). Papers [20, 21] use a classic mean field approach (i.e. with limits in continuous time and Lipschitz continuous rates) to study optimisation policies for load balancing. In particular, [21] exploits mean-field properties to compute performance measures at the level of single server or job.

3. NOTATION

In order to make the paper self-contained, in this section we fix the notation that will be used throughout the remainder. Additional background on fluid limits will be given in Section 4, while discussing an example of a queueing system with batch arrivals.

We will first introduce a simple language to describe network models as *population processes*, where the variables are the number of jobs at each station.

Formally, a CTMC representation for such models is the tuple $\mathcal{X} = (X, \mathcal{S}, x_0, \mathcal{T})$, where:

- $X = (X_1, \ldots, X_n)$ is a vector of *variables*, where n is the total number of stations in the network;

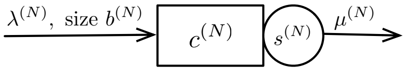


Figure 1: The queueing system with batch arrivals considered in Section 4.

- $\mathcal{S}$ is the (countable) *state space* of the CTMC;
- $x_0 \in \mathcal{S}$ is the *initial state* of the model;
- $\mathcal{T}$ is the set of *transitions*, where $\tau \in \mathcal{T}$ is in the form $\tau = (g_\tau(X), v_\tau, r_\tau(X))$; $g_\tau(X)$ is the *guard*, a conjunction of inequalities of the form $h(X) \geq 0$, for a suitably smooth function h (usually linear); $v_\tau \in \mathbb{R}^n$, is the *update vector*, i.e., a vector giving the net change on each variable caused by the transition (we require that $X + v_\tau \in \mathcal{S}$ whenever $g_\tau(X)$ is true); $r_\tau : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is a Lipschitz continuous and bounded *rate function*, which specifies the rate of the transition as a function of the current state of the system.

Given a model $\mathcal{X}$, it is straightforward to obtain the infinitesimal generator matrix Q of the CTMC, which is given by the $|\mathcal{S}| \times |\mathcal{S}|$ matrix defined by

$$q_{x,x'} = \sum\{r_\tau(x) \mid \tau \in \mathcal{T},\ g_\tau(x)\ true,\ x' = x + v_\tau\}.$$

We indicate by $X(t)$ the state of such a CTMC at time t.

We will make our network models depend upon a parameter, N, which plays the role of the *system's size*; intuitively, the larger N, the larger the system, e.g., the more clients will request service. By varying N, we obtain a sequence $(\mathcal{X}^{(N)})_{N\in\mathbb{N}}$ of models, generating a sequence of CTMCs, denoted by $X^{(N)}(t)$. We aim at finding a fluid approximation of these models, for large N. In order to compare models of different size, we carry out the usual normalisation step, which consists in dividing all the populations by N, and rescaling transitions accordingly. This is essentially a change of variables from X to $\bar{X} = X/N$.

In general, given the CTMC model for level N, denoted by $\mathcal{X}^{(N)} = (X, \mathcal{S}^{(N)}, \mathcal{T}^{(N)}, x_0^{(N)})$, we denote its normalised version by $\bar{\mathcal{X}}^{(N)} = (\bar{X}, \bar{\mathcal{S}}^{(N)}, \bar{\mathcal{T}}^{(N)}, \bar{x}_0^{(N)})$, where $\bar{\mathcal{S}}^{(N)} = \mathcal{S}^{(N)}/N$, $\bar{x}_0^{(N)} = x_0^{(N)}/N$. The transitions $\bar{\tau} \in \bar{\mathcal{T}}^{(N)}$, with $\bar{\tau}$ defined as $(\bar{g}^{(N)}(\bar{X}), \bar{v}^{(N)}, \bar{r}^{(N)}(\bar{X}))$, are obtained from the corresponding transitions $\tau = (g^{(N)}(X), v^{(N)}, r^{(N)}(X))$ by setting $\bar{g}^{(N)}(\bar{X}) = g^{(N)}(X)$, $\bar{v}^{(N)} = v^{(N)}/N$, and $\bar{r}^{(N)}(\bar{X}) = r^{(N)}(X)$.

We introduce the following *indicator function* $I\{P(X)\}$, where $P(X)$ is a logical predicate on variables X, which is useful when describing rates with discontinuities.

$$I\{P(x)\} = \begin{cases} 1 & \text{if } P(x) \text{ true,} \\ 0 & \text{otherwise.} \end{cases}$$

4. FLUID APPROXIMATION OF BATCH ARRIVALS

In this section we discuss a simple multi-server queueing system with batch arrivals. In doing so, we present all fluid limit results that are needed in the paper. The queue has an exponentially distributed service rate $\mu^{(N)}$ and server multiplicity $s^{(N)}$. The batch arrivals have exponentially distributed interarrival times with rate $\lambda^{(N)}$ and batch sizes $b^{(N)}$, where N is the scaling parameter for the CTMC sequence. The meaning of these parameters is summarised pictorially in Figure 1. The buffer is hereafter supposed to be unbounded. Then, Section 4.3 will study the case with finite capacity, indicated by $c^{(N)}$ in the figure.

The model may be formalised in the notation presented in the previous section. Its model $\mathcal{X}^{(N)}$ has a single variable, denoted by $X^{(N)}$ with domain $\mathbb{N}_0$, which counts the population of jobs in the buffer, and two transitions, τ_1 and τ_2. The arrival transition τ_1 has rate $\lambda_1^{(N)}$, no guard ($g_1 = true$), and update vector $v_1 = b^{(N)}$, while the service transition τ_2 has rate $\mu^{(N)} \min\{X^{(N)}, s^{(N)}\}$, no guard, and update vector $v_2 = -1$.

For given $\lambda, \mu > 0$ and $b, s \in \mathbb{N}$, we consider to scalings of the network parameters as follows.

S1 The batch size is constant, $b^{(N)} = b$, but the arrival rate of batches of clients increases with N, $\lambda^{(N)} = N\lambda$.

S2 Clients arrive at a constant rate $\lambda^{(N)} = \lambda$, but in batches of growing size, $b^{(N)} = Nb$.

In both cases, we need to increase the number of servers to keep up with the increased traffic, so that we always let $s^{(N)} = sN$. Notice that, in closed networks such as the one of Section 6, a natural interpretation for N is the total number of clients in the system.

4.1 Fluid Limit (S1)

The main idea behind *fluid* (or deterministic) approximations for a sequence of CTMCs $\bar{X}^{(N)}$ is that, if suitable scaling assumptions are satisfied by rates and update vectors, the sequence converges to a deterministic limit process, solution of an ordinary differential equation (ODE). Essentially, we have to require that rates increase with N and updates decrease as $1/N$ for all transitions. If this is the case, then as N gets larger and larger, the density of jumps increases while their magnitude decreases, hence jumps can be approximated as continuous derivatives in the limit.

To be more precise, the scaling conditions that we require are the following: for each transition $\tau^{(N)} \in \bar{\mathcal{T}}^{(N)}$ of the normalised model, the supremum of the rate $r_\tau^{(N)}$ must be of order N, i.e. $\sup_{x\in\bar{\mathcal{S}}^{(N)}} r_\tau^{(N)}(x) = \Theta(N)$, while the norm of the update vector $v_\tau^{(N)}$ must be of order $1/N$, i.e., $\|v_\tau^{(N)}\| = \Theta(1/N)$.

In order to construct the limit ODE, we need to define the *drift* of the model, i.e., the mean increment of variables at each step, which is

$$F^{(N)}(\bar{X}) = \sum_{\tau\in\bar{\mathcal{T}}} \bar{v}_\tau^{(N)}\, \bar{r}_\tau^{(N)}(\bar{X}).$$

Now, consider the smallest closed subset $E \subseteq \mathbb{R}^n$ containing all the state spaces $\bar{\mathcal{S}}$ of normalized models, that is $E = cl\left(\bigcup_{N\in\mathbb{N}} \bar{\mathcal{S}}^N\right)$, and assume that:

1. $F^{(N)}$ converges uniformly in E to a Lipschitz continuous function F;
2. the initial state of the CTMC sequence converges to a point in (the interior of) E, i.e. $\bar{x}_0^{(N)} \to x_0 \in E$;
3. $x(t)$ is the solution of the initial value problem $\frac{dx(t)}{dt} = F(x(t))$, $x(0) = x_0$.

Under such conditions, it is possible to prove the following theorem [7, 22, 23]:

THEOREM 1 (DETERMINISTIC APPROXIMATION). *Under the previous assumptions, for any $T < \infty$,*

$$\lim_{N\to\infty} \sup_{t\leq T} \|\bar{X}^{(N)}(t) - x(t)\| = 0 \quad \textit{in probability.}$$

This theorem states that, for any finite time horizon T, the *trajectories* of the CTMC become indistinguishable from the solution of the fluid ODE $\frac{dx(t)}{dt} = F(x(t))$. Essentially, the sequence $\bar{X}^{(N)}(t)$ behaves as a deterministic process in the limit.

Of the two forms of scaling introduced at the beginning of Section 4, we observe that only S1, i.e., $\lambda^{(N)} = N\lambda$ and $b^{(N)} = b$, is amenable to fluid approximation. In S2, instead the arrival transition does not satisfy the scaling assumptions because the suprema of both the rate and the update vector are $\Theta(1)$.

Then, considering S1 and computing the drift, we obtain

$$F^{(N)}(x) = F(x) = k\lambda - \mu \min\{x, s\}, \tag{1}$$

which is independent of N. Furthermore, assuming $\bar{x}_0^{(N)} = x_0 = 0$, the conditions of Theorem 1 are satisfied, and we can conclude that the sequence of CTMC converges uniformly to the solution of $\frac{dx(t)}{dt} = F(x(t))$ in any bounded time interval.

In this particular model, however, we can say something also about the steady-state behaviour of the sequence of CTMCs. In fact, it is easy to show that each CTMC in the sequence is irreducible and that the ODE has a unique globally attracting steady state, equal to $\frac{k\mu_1}{\mu_2}$, provided that $k\mu_1 < \mu_2 s$.[1] Under such conditions, it can be shown that

$$\lim_{N\to\infty} \lim_{t\to\infty} \|\bar{X}^N(t) - x(t)\| = 0$$

in probability [11, 24].

Finally, we point out that the equation $\frac{dx(t)}{dt} = F^{(N)}(x(t))$ has another interpretation, namely as an approximate equation for the average of the CTMC at level N [25, 26]. Using either a direct manipulation of the Chapman-Kolmogorov equations, or by Dynkin's formula in differential form [8], the exact equation for the derivative of the expected values of the CTMC reads

$$\frac{d\mathbb{E}[X]}{dt} = \mathbb{E}\left[F^{(N)}(X)\right].$$

By approximating $\mathbb{E}[\min\{x, y\}]$ with $\min\{E[x], E[y]\}$, one obtains

$$\frac{d\mathbb{E}[x]}{dt} = \mathbb{E}\left[F^{(N)}(x)\right] \approx F^{(N)}(\mathbb{E}[x]),$$

which is the fluid equation (using the level-N drift). In our example, however, as $F^N(x) = F(x)$, this is the proper fluid limit equation.

4.2 Hybrid Fluid Limit (S2)

Recalling that the scaling laws for case S2 are $\lambda^{(N)} = \lambda$ and $b^{(N)} = bN$, one observes that in this situation the temporal density of batch arrivals remains constant with respect to N, while the jump is constant in the normalised variables, meaning that its magnitude increases in the unscaled model. Intuitively, the dynamics of such a transition maintains a similar structure for all N, always showing a stochastic behaviour. On the other hand, the service rate does enjoy a suitable scaling, hence this dynamics should intuitively become deterministic asymptotically. Therefore, we expect that, overall, this sequence of CTMCs still exhibits a limit behaviour, although not a purely deterministic one in the sense of Theorem 1, since its scaling conditions are not satisfied. Indeed, it turns out that the limit process is hybrid, mixing discrete/stochastic with continuous/deterministic evolution.

In order to put this intuition into a formal framework, we introduce Piecewise Deterministic Markov Processes, a model of stochastic hybrid systems interleaving periods of continuous evolution with discrete jumps [8]. Following [27], we consider a simple version with Markovian jumps (while in general also *instantaneous* jumps are allowed, happening as soon as their guard becomes true).

DEFINITION 1. *A* simple *Piecewise Deterministic Markov Process (PDMP) is a tuple $(D, X, \mathcal{D}, \varphi, \mathcal{T}, d_0, x_0)$, where:*

- *D is a vector of discrete variables, taking values in the finite set $\mathcal{D}$. An element $d \in \mathcal{D}$ is usually called a* discrete mode. *D can be the empty vector, and in this case $\mathcal{D}$ contains a single point;*
- *X is a vector of n continuous variables, taking values in (a subset of) $\mathbb{R}^n$;*
- *$\varphi : \mathcal{D} \times \mathbb{R}^n \to \mathbb{R}^n$ is a function that defines a Lipschitz continuous vector field for each mode $d \in \mathcal{D}$;*
- *$\mathcal{T}$ is a set of Markovian transitions in the form (r_i, R_i), where $r_i : \mathcal{D} \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is the rate of the transition, and $R_i : \mathcal{D} \times \mathbb{R}^n \to \mathcal{D} \times \mathbb{R}^n$ is the reset map.*
- *$(d_0, x_0) \in \mathcal{D} \times \mathbb{R}^n$ is the initial state.*

Intuitively, the dynamics of a simple PDMP is as follows. The process starts in the initial state (d_0, x_0) and the continuous variables evolve following the solution of the ODE $\frac{dX(t)}{dt} = \varphi(d_0, X(t))$, while the discrete variables remain constant. Such a continuous evolution is followed until a time T_1, when Markovian jump happens with rate

$$r(d_0, X) = \sum_{(r_i, R_i)\in\mathcal{T}} r_i(d_0, X).$$

A discrete transition i is chosen with probability proportional to its rate, i.e. $r_i(d_0, X(T_1))/r(d_0, X(T_1))$, and the system jumps to the new state $(d_1, x_1) = R_i(d_0, X(T_1))$. Then, the system starts again to evolve continuously from (d_1, x_1), until a new jump occurs. Applying this scheme iteratively yields the piecewise continuous trajectories of the PDMP.

We briefly sketch how to define the simple PDMP associated with a sequence of models $\mathcal{X}^{(N)}$, referring to [27] for more details. The idea is to partition transitions of model $\mathcal{X}^{(N)}$ into two classes, those amenable of continuous approximation (i.e., those satisfying the scaling conditions of Theorem 1), and those to be kept discrete (having rate and update vectors both independent of N in the normalised model). Then, variables not affected by continuous transitions will constitute the discrete variables of the PDMP,

[1] For $k\mu_1 = \mu_2 s$, the ODE has an infinite number of equilibria, namely all points $x \geq s$, while for $k\mu_1 > \mu_2 s$, the ODE goes to infinity.

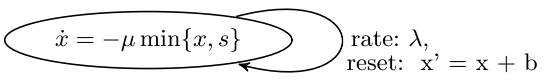


Figure 2: Hybrid-automaton representation of the PDMP associated with the queueing system in Figure 1 to which scaling S2 is applied. There is one single mode and one single variable, x, subject to continuous evolution given by $\frac{dx}{dt} = -\mu \min\{x, s\}$ and to a stochastic jump happening with rate λ and changing the system from x to $x + b$.

while all other variables will become continuous variables. The vector field φ is defined like the drift, but restricting the summation to continuous transitions. Finally, Markovian transitions of the PDMP are obtained straightforwardly from the discrete transitions of the CTMC.

Applying this construction to the batch arrival example with the scaling S2, we obtain the following PDMP: it has one continuous variable ($\bar{X}$) and one discrete mode (there is no discrete variable), the vector field is $\varphi(\bar{X}) = -\mu \min\{\bar{X}, s\}$, and its unique Markovian transition has rate λ and reset map $R(\bar{X}) = \bar{X} + b$. A visual representation of this PDMP, in the usual style of hybrid automata, is shown in Figure 2.

Applying a result of [27], it can be shown that the sequence $\bar{X}^{(N)}(t)$ of CTMC converges in distribution to the PDMP $\bar{X}(t)$ obtained by the previously sketched construction, provided that the vector field is Lipschitz continuous (and rate functions are integrable).

The average behaviour of the limit PDMP can also be described by an ODE, using a more general version of the Dynkin Formula [8]. For any suitably smooth function f, it holds that

$$\frac{d\mathbb{E}[f(d,x)]}{dt} = \mathbb{E}\Bigg[\nabla f(d,x) \cdot \varphi(d,x) + \sum_i r_i(d,x)\left(f(R_i(d,x)) - f(d,x)\right)\Bigg].$$

Let us now specialise the previous formula for the average $\mathbb{E}[d,x](t)$, and apply it to a PDMP obtained from a CTMC model described with the language of Section 3. For this, it holds that $R_i(d,x) = (d,x) + v_i$, therefore we obtain

$$\frac{d\mathbb{E}[d,x]}{dt} = \mathbb{E}\left[\varphi(d,x) + \sum_i v_i r_i(d,x)\right] = \mathbb{E}[F(d,x)],$$

where $F(d,x)$ is the (limit) drift of the CTMC model.

If we compute the equation of the average for the PDMP in Figure 2, we obtain

$$\frac{d\mathbb{E}[x]}{dt} = \lambda k - \mu \mathbb{E}[\min\{x, s\}],$$

which, given the approximation $\mathbb{E}[\min\{x, s\}] \approx \min\{\mathbb{E}[x], s\}$, is exactly the fluid differential equation (1) we obtained for the scaling S1.

4.3 Discontinuity in Rates

All the convergence results presented in the previous section require Lipschitz continuity of the drift or of the PDMP vector field. This is needed to ensure existence and uniqueness of the solutions of the fluid ODEs. Unfortunately, the presence of guards in the CTMC transitions may introduce discontinuities in these functions, thus preventing an application of classical deterministic approximation results.

As an example, consider again the batch arrival model, with scaling S1, but additionally assume that the queue of the service station has bounded capacity $c^{(N)}$. Furthermore, we assume that the batch arrival is suspended whenever an arrival will overcome the capacity $c^{(N)}$. If we let $c^{(N)} = c_0^{(N)} + b^{(N)}$, then arrivals are suspended whenever $X > c_0^{(N)}$. Such a modification is easily accounted for by adding the guard $X \leq c_0^{(N)}$ to the batch arrival transition. In computing the drift for this modified model, we have to take into account the suspension policy, multiplying the arrival rate by the indicator function $I\{X \leq c_0^{(N)}\}$, so that the drift becomes

$$F^{(N)}(x) = F(x) = k\lambda I\{x \leq c_0\} - \mu \min\{x, s\}.$$

As $F(x)$ is discontinuous, the associated fluid equation $\frac{dx(t)}{dt} = F(x(t))$ is not an ODE, but rather a Piecewise Smooth dynamical system (PWS) [28].

PWS have continuous trajectories showing in general more complex behaviour than ODE solutions, even if in many circumstances the solutions of the initial value problems associated to a PWS exist and are unique.

Intuitively, the dynamics of a PWS within a continuity region of the vector field behaves like that of the solution of the corresponding ODE. However, differences arise in the proximity of a discontinuity surface. To fix the notation, suppose that a discontinuity surface $\mathcal{H}$ is defined as the set of zeros of a (sufficiently) smooth function $h : \mathbb{R}^n \to \mathbb{R}$, i.e., $\mathcal{H} = \{x \mid h(x) = 0\}$. This surface separates $\mathbb{R}^n$ in two regions: $R_1 = \{x \in \mathbb{R}^n \mid h(x) < 0\}$ and $R_2 = \{x \in \mathbb{R}^n \mid h(x) > 0\}$. Denote the restriction of the vector field F to R_1 by F_1 and the restriction of F to R_2 by F_2.

In the example, there is a single discontinuity surface, defined by the equation $x = c_0$, which defines the two regions $R_1 = \{x < c_0\}$ and $R_2 = \{x > c_0\}$. The vector field in R_1 is $F_1(x) = k\lambda - \mu \min\{x, s\}$, while $F_2(x) = -\mu \min\{x, s\}$.

The behaviour of a trajectory of the PWS when it hits the surface $\mathcal{H}$ in a point x essentially depends on the relative orientation of F_1 and F_2 around x. If both vector fields point towards the same region, then the trajectory crosses $\mathcal{H}$, possibly with a discontinuity in its derivative (*transversal crossing*). Formally, this happens if the projections of the vector fields along the normal $\nabla h(x)$ to the surface $\mathcal{H}$ in x (assumed to be always different from zero), $F_1(x) \cdot \nabla h(x)$ and $F_2(x) \cdot \nabla h(x)$, have the same sign.

In our example, $\nabla h(x) = 1$ and $F_2(c_0) < 0$, hence transversal crossing happens whenever $F_1(c_0) < 0$. This corresponds to the condition $\lambda < \frac{\mu}{k} \min\{c_0, s\}$. Notice that $\mathcal{H}$ can be crossed only from R_2 to R_1, an unfeasible situation as the initial conditions are always in R_1.

On the other hand, if the vector fields point in opposite directions of the surface $\mathcal{H}$ (in particular, the vector field in R_1 points towards R_2 and vice versa, meaning that $F_1(x) \cdot \nabla h(x) > 0$ and $F_2(x) \cdot \nabla h(x) < 0$), then the trajectory is constrained to move along $\mathcal{H}$, a behaviour known as *sliding motion*. In fact, the PWS moves along $\mathcal{H}$ following the so called *sliding vector field* $G(x)$, which is obtained as the convex combination of $F_1(x)$ and $F_2(x)$ tangential to $\mathcal{H}$.[2]

[2] Formally, $G(x) = \alpha(x)F_1(x) + (1 - \alpha(x))F_2(x)$, where $\alpha(x)$ satisfies $G(x) \cdot \nabla h(x) = 0$.

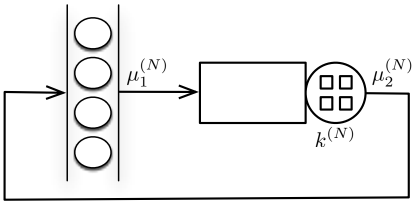


Figure 3: Closed queueing network with batch services (indicated by the small boxes within the service centre) studied in Section 5.

In our example, sliding motion happens whenever $F_1(c_0) > 0$, i.e. whenever $\lambda > \frac{\mu}{k}\min\{c_0, s\}$. In this case, the sliding vector field is $G(x) = 0$, hence once a trajectory hits the surface $\mathcal{H}$, it remains there forever.

Existence and uniqueness of solutions of a PWS is guaranteed if in each point of the surface $\mathcal{H}$ either $F_1(x)\cdot\nabla h(x) > 0$ or $F_2(x)\cdot\nabla h(x) < 0$ holds (this is known as the Filippov condition). This condition is verified by the batch arrival with bounded capacity.

Starting from a sequence of CTMC with nontrivial guards (but defined by smooth functions, for instance linear functions), and computing the drift as in the fluid approximation, we therefore obtain a PWS. In [9, 10], the authors have shown that such a sequence of CTMC converges to the solution of the associated PWS,[3] provided that such solution exists and is unique (and additionally it crosses a finite number of times discontinuity surfaces in any finite amount of time). This allows the use of fluid approximation also in these situations, once the regularity conditions on the PWS are proved.

For the batch arrival with bounded capacity, these regularity conditions hold hence the sequence of CTMC converges to the limit PWS.

5. BATCH SERVICES

We study suitable scalings for queues with batch services by considering the closed tandem network in Figure 3, which is also convenient to highlight the form of scaling to which the job population is subjected. Let the population vector be denoted by $X^{(N)} = (X_1^{(N)}, X_2^{(N)})$, where $X_1^{(N)}$ and $X_2^{(N)}$ represent the queue length at the delay station and at the batch service station, respectively. Let $X_0^{(N)} = (X_{1,0}^{(N)}, X_{2,0}^{(N)})$ be the initial condition, with $X_{1,0}^{(N)} + X_{2,0}^{(N)} = N$, i.e., the population grows with N. Let $k^{(N)}$ be the batch service size and $\mu_1^{(N)}$ and $\mu_2^{(N)}$ be the service rates at the delay station and at the queue with batches, respectively. The model is described by two transitions: τ_1 (service at the delay station) has rate $\mu_1^{(N)}$, no guard ($g_1 = true$), and update vector $v_1 = (-1, 1)$; τ_2 (batch service) has rate $\mu_2^{(N)}$, guard $X_2^{(N)} \geq k^{(N)}$, and update vector $v_2 = (k^{(N)}, -k^{(N)})$.

As in the case of batch arrivals, we consider two different scalings, for given $k \in$ ‘N, $\mu_1, \mu_2 > 0$, and $X_0 \in \mathbb{N}^2$.

S3 $k^{(N)} = k$, $\mu_1^{(N)} = \mu_1$, $\mu_2^{(N)} = N\mu_1$, $X_0^{(N)} = NX_0$, i.e., the batch size is kept constant but the service rates grow with N, to keep up with increasing population sizes.

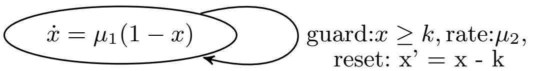


Figure 4: HA-like representation of the limit PDMP for the example of Section 5.2.

S4 $k^{(N)} = Nk$, $\mu_1^{(N)} = \mu_1$, $\mu_2^{(N)} = \mu_2$, $X_0^{(N)} = NX_0$, i.e., the batch grows but the service rates are maintained constant. Population sizes are increased as in S3.

In either case, the rate at the delay station is not varied.

5.1 Constant batch sizes, increasing rates (S3)

Classical limit theorems are not applicable under these circumstances because of the guard $X_2^{(N)} \geq k^{(N)}$, which in the N-th normalised model becomes $\bar{X}_2 \geq k^{(N)}/N$. This will give rise to a service rate which may be written in the form $\mu_2 I\{\bar{X}_2 \geq k/N\}$. The drift $F^{(N)}$ of the normalised CTMC at level N is therefore

$$F^{(N)}(x_1, x_2) = (-1, 1)\mu_1 x_1 + (k, -k)\mu_2 I\{x_2 \geq k/N\},$$

which converges to the limit drift

$$F(x_1, x_2) = (-1, 1)\mu_1 x_1 + (k, -k)\mu_2 I\{x_2 \geq 0\},$$

thus defining the limit PWS $\frac{dx(t)}{dt} = F(x(t))$. Although the presence of discontinuities prevents the use of classical limit theorems, the results discussed in Section 4.3 allow us to derive the convergence to F for this sequence of CTMCs.

5.2 Increasing batch sizes, constant rates (S4)

Under scaling S4, in the normalised model the update vector as well as the rates of the CTMC transitions are $\Theta(1)$. This independence from N is the characteristic scaling of the hybrid fluid limit introduced in Section 4.2. More specifically, the batch service will remain discrete and stochastic in the limit PDMP model, which is shown in Figure 4.

However, also in this case we can compute the drift and construct the fluid equation, which is now interpreted as a first-order approximation to the average behaviour of the limit PDMP. The drift of the CTMC at level N is

$$\begin{aligned} F^{(N)}(x_1, x_2) &= F(x_1, x_2) \\ &= (-1, 1)\mu_1 x_1 + (k, -k)\mu_2 I\{x_2 \geq k\}, \end{aligned}$$

which is independent from N and gives rise to the ODE (with discontinuous right hand side) $\frac{dx(t)}{dt} = F(x(t))$.

5.3 Properties of the fluid equation

The fluid equation $\frac{dx(t)}{dt} = F^{(N)}(x(t))$, constructed using the N-dependent drift, is an approximation of the average behaviour both for scaling S3 and S4. In general, this equation is

$$\frac{dx(t)}{dt} = (-1, 1)\mu_1 x_1 + \left(\frac{k^{(N)}}{N}, -\frac{k^{(N)}}{N}\right)\mu_2^{(N)} I\left\{x_2 \geq \frac{k^{(N)}}{N}\right\},$$

where $\mu_2^{(N)}$ and $k_2^{(N)}$ scale either as S3 or as S4. It can be proved that this PWS system has a unique solution for any

[3] Convergence is uniform in probability for any finite time horizon, as for Theorem 1.

possible initial state $x \in [0,1]^2$, $x_1 + x_2 = 1$. Furthermore, there is a unique globally attracting steady state, which is

$$\begin{cases} \left(\frac{k^{(N)}\mu_2^{(N)}}{N\mu_1}, 1 - \frac{k^{(N)}\mu_2^{(N)}}{N\mu_1}\right) & \text{if } \mu_2^{(N)} \leq \mu_1 \left(\frac{N}{k^{(N)}} - 1\right), \\ \left(1 - \frac{k^{(N)}}{N}, \frac{k^{(N)}}{N}\right) & \text{otherwise.} \end{cases}$$

The latter case corresponds to sliding motion along the discontinuity surface $x_2 = \frac{k^{(N)}}{N}$, and the equilibrium is reached in a finite amount of time.

The existence and uniqueness of solutions for any initial conditions and the presence of a global attractor bring us to conjecture that the results about limit behaviour holding for Lipschitz continuous fluid limits with a single globally attractive steady state extend also to this PWS system, allowing the use of the fluid equation to estimate the steady state behaviour. A formal proof of this result is current ongoing work.

6. NETWORKS WITH BATCHES

We are now ready to define fluid limits for a general class of queueing networks with batch services and arrivals. The general model is provided in Section 6.1, which also introduces two forms of scaling which combine those already discussed in Sections 4 and 5.

6.1 General model

Using standard notation and terminology, we consider an open network of n stations with exponential services and arrivals. In the following, let J_b and J_s be a partition of the set $\{1, 2, \ldots, n\}$, where J_b denotes the batch service stations and J_s denotes the single-job multi-server stations. The model is characterised by the following parameters.

- $\lambda = (\lambda_1, \ldots, \lambda_n)$ is the vector of the (Poisson) intensities of the exogenous arrivals at each station;
- $b = (b_1, \ldots, b_n)$ is the vector of the sizes of the batch arrivals;
- $P = (p_{ij})_{1 \leq i,j \leq n}$ is the routing probability matrix of size $n \times n$. Upon service at station i, jobs leave the network with probability $1 - \sum_{j=1}^n p_{ij}$;
- $\{k_i \mid i \in J_b\}$ is the set of batch service sizes;
- $\{s_i \mid i \in J_s\}$ is the set of server multiplicities at single-job stations. Let $s_i = \infty$ define an infinite-server (i.e., a delay) station;
- $\mu = (\mu_1, \ldots, \mu_n)$ is the vector of service rates. For single-job stations, it is the rate for each server in that station;
- $X = (X_1, \ldots, X_n)$ is a reachable state of the CTMC that defines the network, with X_i being the queue length at station i, including the jobs in service or currently accumulated in the batch;
- $X_0 = (X_{1,0}, \ldots, X_{n,0})$ is the initial state of the CTMC.

In order to define the family $\mathcal{X}^{(N)}$ of CTMCs, let $\lambda^{(N)}$, $b^{(N)}$, ... be the network configuration of the N-th CTMC of the sequence. We assume that the routing probabilities do not scale with N, i.e., $P^{(N)} = P$ for all N. Now, let e_i be a vector of length n of all zeros except the i-th element which is set to 1. For all $1 \leq i,j \leq n$, the transitions of the N-th CTMC $\mathcal{X}^{(N)}$ are as follows.

batch arrival: $(\cdot, b_i^{(N)} e_i, \lambda_i^{(N)})$;

batch service: if $i \in J_b$,
$(X_i^{(N)} \geq k_i^{(N)}, -k_i^{(N)} e_i + k_i^{(N)} e_j, p_{ij}\mu_i^{(N)})$;

batch service (leaving network): if $i \in J_b$,
$(X_i^{(N)} \geq k_i^{(N)}, -k_i^{(N)} e_i, (1 - \sum_{j=1}^n p_{ij})\mu_i^{(N)})$;

single job service: if $i \in J_s$,
$(\cdot, -e_i + e_j, p_{ij}\mu_i^{(N)} \min\{X_i^{(N)}, s_i^{(N)}\})$

single job service (leaving network): if $i \in J_s$,
$(\cdot, -e_i, 1 - \sum_{j=1}^n p_{ij})\mu_i^{(N)} \min\{X_i^{(N)}, s_i^{(N)}\})$

In the remainder of this section, we study two distinct forms of scaling:

S5 $\lambda_i^{(N)} = N\lambda_i$, $b_i^{(N)} = b_i$, and $X_0^{(N)} = NX_0$. Furthermore, $\mu_i^{(N)} = N\mu_i$, $k_i^{(N)} = k_i$ for $i \in J_b$, i.e., for all batch service stations, while $\mu_j^{(N)} = \mu_j$ and $s_j^{(N)} = s_j N$ for $j \in J_s$, i.e., for stations that serve singly. This essentially corresponds to combining scaling S1 (for arrivals) and S3 (for batch services). The initial job populations scale as in the case of the closed network analysed in Section 5, whereas multiplicity levels at ordinary stations have the scaling as in Section 4.

S6 $\lambda_i^{(N)} = \lambda_i$, $b_i^{(N)} = b_i N$, and $X_0^{(N)} = NX_0$. Moreover, $\mu_i^{(N)} = \mu_i$, $k_i^{(N)} = k_i N$, for $i \in J_b$, while $\mu_j^{(N)} = \mu_j$ and $s_i^{(N)} = s_j N$ for $j \in J_s$. Giving the same dependence upon N to initial job populations and to server multiplicities, this scaling considers S2 for arrivals and S4 for batch services.

Similarly to the limit results presented in Sections 4 and 5, also in this general case the scaling will determine the kind of limit process. Scaling S5 results in a sequence of CTMCs having a fluid limit in terms of a PWS, due to the presence of discontinuities in the rate functions induced by batches. Also in the general case, we can invoke the results of [9, 10] to conclude convergence of the sequence of CTMCs to this limit. However, care has to be taken to ensure that the PWS has the regularity properties requested by the limit theorems (essentially, existence and uniqueness of the solutions everywhere). At the moment we still do not have a general result for the class of PWS models considered, hence we need to check all generated PWS for satisfaction of regularity properties. However, all the examples we studied enjoyed the requested properties, and we are currently working on a proof for the general case, or for reasonably large subsets.

On the other hand, if we consider scaling S6, we are in a situation leading to a stochastic hybrid limit, where all batch arrivals and services remain stochastic, and all other transitions are approximated by deterministic flows. In any case, we can always derive a fluid equation also for this scaling, using the drift of the CTMC at level N, to be interpreted as an approximation of the average of the stochastic process, or of the limit stochastic hybrid system.

Specifically, we can derive the following set of differential equations, that are a PWS: For all $i = 1, \ldots, n$, let

$$\begin{aligned}\frac{dx_i}{dt} &= \lambda_i b_i + \sum_{j \in J_b} p_{ji} k_j \mu_j I\left\{x_j \geq \kappa_j^{(N)}\right\} \\ &\quad + \sum_{j \in J_s} p_{ji}\mu_j \min(x_j, s_j) - \mu_i \min(x_i, s_i), \quad \text{if } i \in J_s, \\ \frac{dx_i}{dt} &= \lambda_i b_i + \sum_{j \in J_b} p_{ji} k_j \mu_j I\left\{x_j \geq \kappa_i^{(N)}\right\} \\ &\quad + \sum_{j \in J_s} p_{ji}\mu_j \min(x_j, s_j) - k_i\mu_i I\left\{x_i \geq \kappa_i^{(N)}\right\}, \text{if } i \in J_b,\end{aligned}$$

where

$$\kappa_z^{(N)} = \begin{cases} k_z/N & \text{for scaling S5,} \\ k_z & \text{for scaling S6.}\end{cases}$$

6.2 Discussion

Mixing scalings.

In assuming S5 or S6, we are requesting that all batch transitions scale in the same way, i.e. with constant batch size and increasing rate (S5) or with increasing batch size and constant rate (S6). However, it is possible to consider a mixed scaling, in which some batch arrivals or services scale as S5 and some scale as S6. These models give rise to a limit PDMP, in which only transitions with increasing batch sizes are kept discrete. Therefore, the PDMP may exhibit discontinuous rates, unlike the previous cases. However, we conjecture that the limit results of [9, 27] can be combined so that convergence still holds, provided the PWS system has a sufficiently regular structure (existence and uniqueness of solutions for any initial condition).

Practical considerations.

In real applications, we generally do not have a sequence of CTMCs, but rather a specific model, with a given set of parameter values. Furthermore, it may not be known how the parameters are to scale with respect to N. This suggests to adopt the following policy. Construct the fluid limit equation, using the drift at level N, and approximate the average behaviour of the system by the solution of such an equation. Under the proper scaling conditions, as discussed above, then this equation also gives the limit behaviour of the model. However, for a fixed set of parameters, we wish to assess the accuracy error, i.e., how close the solution of the fluid PWS system is to the real average. This is problematic from a theoretical viewpoint as currently known error bounds are shown to grow doubly exponentially with the time horizon [23].

As a rule of thumb, we expect that if the population/scaling factor is large and the batch sizes are small (compared to the population/scaling factor), the behaviour is close to S5, hence the fluid equation will perform better. On the other hand, for relatively large batch sizes, the behaviour is close to the hybrid limit and the fluid equation may perform worse.

Indeed, let us consider again the example of Section 5. Applying the Dynkin formula to the generic drift $F^{(N)}$, we can see that the real average of the system follows the equation

$$\begin{aligned}\frac{d\mathbb{E}[x]}{dt} &= (-1, 1)\mu_1 \mathbb{E}[x_1] \\ &\quad + (k/N, -k/N)\, \mu_2^{(N)} \mathbb{E}\left[I\{x_2 \geq k^{(N)}/N\}\right],\end{aligned}$$

from which we obtain the limit fluid equation by the approximation

$$\begin{aligned}\mathbb{P}\left\{\bar{X}_2 \geq k^{(N)}/N\right\} &= \mathbb{E}\left[I\{x_2 \geq k^{(N)}/N\}\right] \\ &\approx I\{\mathbb{E}[x_2] \geq k^{(N)}/N\}.\end{aligned}$$

This can be quite crude, especially for values of the probability $\mathbb{P}\left\{\bar{X}_2 \geq k^{(N)}/N\right\}$ far from 0 or 1. In fact, with the considered approximation, we are just checking if the (approximate) average of the stochastic process is above or below the threshold $k^{(N)}/N\}$. Now, if the average is far away from such a threshold, we can expect that the probability $p = \mathbb{P}\left\{\bar{X}_2 \geq k^{(N)}/N\right\}$ to be close to 0 or 1, hence the approximation is good. However, when the average is close to the threshold, then p will have an intermediate value between 0 and 1, hence we expect the approximation to be worse. This phenomenon is less severe for large N (and small batch sizes), as we can assume S5 scaling, for which there is convergence to the solution of the fluid equation. On the other hand, for small N or large batch sizes (relatively to N), we expect large errors, because we are "closer" to scaling S6, for which the fluid equation is only an approximation of the average of the limit PDMP.

In the following section we provide numerical evidence showing that this fluid approach works quite well in many cases, but may introduce large errors, expecially for large batch sizes and when the limit PWS system shows sliding motion, i.e., when the process remains close to the switching threshold of the indicator function.

7. NUMERICAL EVALUATION

The quality of the accuracy provided by the approximate deterministic models was assessed by a numerical evaluation over a large parameter space. In studies of this kind two major routes may be taken. One is to carry out tests on a randomly generated validation dataset; the other approach is to perform an exhaustive exploration of a relatively small parameter space to subject the network under consideration to a wide variety of operating conditions. The numerical evaluation herein presented is based on the latter method and is inspired by early literature which deals with accuracy estimation in queueing networks [29, 30, 31].

7.1 Set-up and metrics

The simple tandem network presented in Figure 3 was used in this study. A summary of the parameter ranges is provided in Table 1. The job population sizes were kept purposely small across all tests. Hardly do these networks need to be subjected to approximate solution techniques since the state spaces of their underlying Markov chains is only of a few hundred states, which can be even easily dealt with by ordinary numerical CTMC solvers. These conditions are particularly problematic for deterministic approximations, thus the intent of this section is to stress this technique under its most unfavourable circumstances.

Case	k	*Range of* μ_1	μ_2	*Range of job population*
A1	5	$[0.005, 0.500]$	1.0	$15, \ldots, 80$
A2	5	$[0.005, 0.500]$	5.0	$75, \ldots, 400$
B1	10	$[0.010, 1.000]$	1.0	$15, \ldots, 80$
B2	10	$[0.010, 1.000]$	5.0	$75, \ldots, 400$
C1	15	$[0.020, 2.000]$	1.0	$15, \ldots, 80$
C2	15	$[0.020, 2.000]$	5.0	$75, \ldots, 400$

Table 1: Network parameters used for the assessment of the approximation accuracy. The labels in the first column are referred to in Section 7.2. For each dataset 700 equally spaced points in the parameter space were considered.

In each validation dataset, denoted by A1, A2, and so forth, the value of μ_2 was kept fixed whereas μ_1 and the job populations were changed so as to obtain different utilisation levels for the batch queue. This utilisation is here measured as the fraction of the network's steady-state throughput divided by the maximum attainable throughput at the batch queue, which is given by $k\mu_2$. To study the speed of convergence to the deterministic approximation, each configuration in the validation datasets labelled with 1 was scaled up according to the scaling laws S5. For instance, the model with $\mu_1 = 0.005$, $\mu_2 = 1.0$, $N = 15$ in A1 has the same fluid limit as that in A2 with $\mu_1 = 0.005$, $\mu_2 = 5 \times 1.0 = 5.0$, and $N = 5 \times 15 = 75$. Different batch sizes were also tested. In order to roughly maintain the same spectrum of network utilisations across all batch sizes, the ranges of μ_1 where adjusted in each validation dataset.

The approximation accuracy was measured as the percentage relative error of the throughput with respect to its statistical expectation, as computed by stochastic simulation with 95% confidence intervals below 1% radius. The fluid estimate was computed by standard numerical integration of the ODE, enhanced with an *event-detection* mechanism to check for sliding motion and to alter the vector field accordingly. This information was also used to test the hypothesis whether the ODE solutions that undergo sliding motion are generally less accurate than those that do not cross discontinuity regions.

7.2 Results

The results for the case A1 are shown as a contour plot in Figure 5a, where the levels are labelled with the relative error of the throughput for each network configuration. The axes are organised in such a way that points in the bottom-left part represent situations with light loads, as they are characterised by relatively low rates at the delay station and/or small population levels. Conversely, the points in the top-right part are related to high loads. The graph shows that the fluid model is particularly accurate in the latter situation (cf. absence of contours) whereas it suffers large errors in the former. This is perhaps not surprising since fluid models are generally not usable for networks with small population levels. Notably, a region where the approximation does not behave well is found in the middle of the chart; this corresponds to mid- to high-utilisation conditions for the batch queue of around 70–80% (cf. Figure 5d, which shows the utilisations for case A1). The error plots for cases B1 and C1, shown in Figures 5b and 5c, respectively, show similar trends. For the sake of conciseness, the figures regarding the remaining cases are not provided.

The dotted curve in the error plots divides the N-μ_1 plane into two regions according to the behaviour of the ODE solutions. Parameters lying below the curve give rise to solutions that undergo sliding motion, whereas those above the curve do not cross discontinuity surfaces. In order to quantify the differences in accuracy between these two cases, let us consider the aggregated error statistics for all cases, collectively reported in Table 2. Aggregating the statistics according to these two regions of the parameter space sustains the hypothesis that such discontinuities do have a negative impact on the quality of the approximation. These results also indicate that the accuracy tends to degrade with increasing batch sizes (compare, e.g., A1, B1, and C1). However, the errors in the cases without sliding motion tend to be comparable across all cases, whereas significant differences may be noted in the cases exhibiting sliding motion.

Finally, the table clearly shows how the quality of the approximation improves with increasing population sizes — compare, for instance, the error statistics of A1 and A2. Let us remark that the accuracy is already satisfactory for most practical purposes for all configurations in A2, B2, and C2, although, as discussed above, those cases are not intended to be served by deterministic approximations given their excellent computational tractability. It is therefore not unreasonable to except very good accuracy in the case of large-scale models, where explicit enumeration is unfeasible and stochastic simulation costly.

8. CONCLUSION

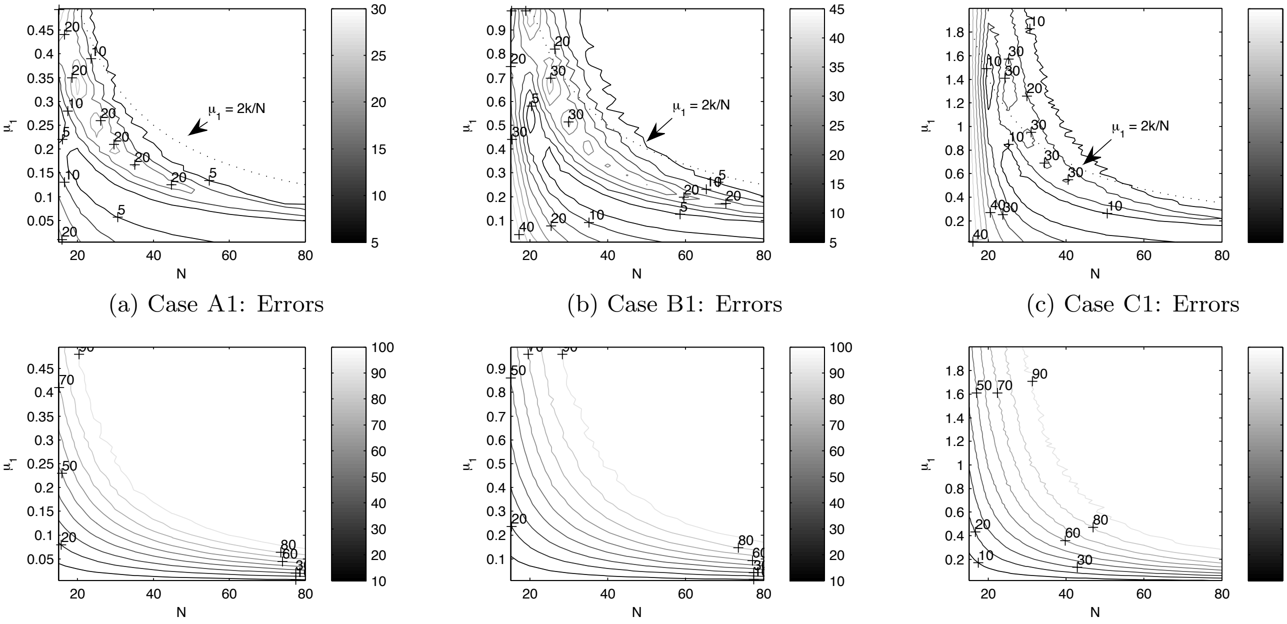
Summary of contributions.

This paper has discussed deterministic approximations for queueing networks with batch arrivals and batch services. In some cases it is possible to straightforwardly apply classical limit theorems and interpret the solution to the resulting ODE system as the sample-path trajectory of a suitable sequence of CTMCs, thus justifying, for instance, its practical application as an estimate of the stochastic behaviour for large-scale models. However, for other network configurations, it turns out that the limit behaviour is an ODE with discontinuous right-hand side, or that it is not a deterministic process but rather a stochastic hybrid one.

In the former case, by appealing to recently established results we have been able to show that the discontinuous ODE model is still a meaningful limit trajectory in the sense of an extension of Kurtz's theory (which is originally valid for smooth ODEs). In the latter situation, instead, we have derived an approximating ODE system which may be interpreted as a first-order approximation to the limit hybrid automaton. A numerical study has highlighted that the quality of the approximation increases quite rapidly with larger populations of the network under consideration, with errors less than 3% on average even for networks of moderate size (i.e., a few hundred jobs).

Practical implications.

There are two main assumptions in our models. The first, which is common to all analyses based on CTMCs, is that every activity in the system under study can be reasonably considered as being distributed exponentially. The second,



(a) Case A1: Errors

(b) Case B1: Errors

(c) Case C1: Errors

(d) Case A1: Utilisations

(e) Case B1: Utilisations

(f) Case C1: Utilisations

Figure 5: Contour plots of the percentage relative errors and the utilisations of the batch queue for the validation datasets in Table 1. The dotted line in the error plots divides the plane into two regions characterised by network parameters which lead to ODE solutions with sliding motion (below) and without discontinuities (above).

which is specific to our approach, is that of deterministic batch sizes. Under those assumptions, the techniques herein discussed are readily usable for the analysis of batch networks with general topologies. Taken together, the theoretical results and the numerical assessment suggest the following strategy for the performance evaluation of systems exhibiting batch behaviour.

- If the system is sufficiently small, the traditional numerical routes to transient or steady-analysis of the underlying CTMC may be taken [32]. In this case, the computational cost tends to be dominated by the population sizes of the jobs more than by the number of queues in the network.
- Larger models are typically more difficult to solve numerically due to the size of the generator matrix. In this case, stochastic simulation appears to be a viable option; stringent enough confidence interval give performance estimates which are usable for all practical purposes.
- The deterministic approximations presented in this paper could be used for massive systems: the numerical investigation suggests excellent accuracy for networks with thousands of jobs and more under all situations of traffic conditions and all parameter configurations (cf. last four columns for cases C1 and C2 in Table 2).
- Differential equations may still be preferred for systems of medium size in virtue of the low computational cost of the solution. This is particularly appealing for parameter sweeps over large configuration spaces during early-stage capacity planning, when the modeller may be willing to trade accuracy for speed. Practically useful error bounds are not available, however the numerical results suggest some heuristic approaches to assessing the quality of the approximation. For instance, situations of sliding motion, which can be easily checked with a simple routine integrated with the numerical ODE solver, may flag potential inaccuracies.

Future work.

Though this paper was concerned with batch processing, there are other forms of service in queueing networks which may enjoy similar results in terms of nonstandard fluid limits. Ongoing work is being devoted to studying the case of multiclass networks with priorities.

Another interesting research line could be to study the hybrid limits further. Here, they have been approximated with deterministic equations. However, a hybrid automaton may be considered *per se* as an approximate representation. This raises the question whether it is possible to provide exact solutions; in any case, its simulation may be seen as an intermediate approach which is expected to be faster than the simulation of the overall CTMC, and more accurate than the deterministic approximation. The authors have been recently involved in (numerical) studies of this kind which confirm this behaviour, although in a different context [33]. Along this direction, a promising approach which we wish to investigate is that of using suitable moment-closure techniques [34] in order to improve the approximation of the expected behaviour of the hybrid automaton.

	With Sliding Motion					*Without Sliding Motion*					*Overall*			
Case	*%*	*5th*	*Avg.*	*50th*	*95th*	*%*	*5th*	*Avg.*	*50th*	*95th*	*5th*	*Avg.*	*50th*	*95th*
A1	33	0.74	8.63	6.80	21.41	67	0.11	3.14	1.13	15.04	0.15	4.96	1.87	18.92
A2	73	0.14	2.59	1.44	9.06	27	0.03	0.94	0.41	4.04	0.04	1.40	0.52	6.81
B1	38	1.21	15.23	12.56	38.08	62	0.15	5.22	1.71	21.54	0.23	9.16	4.57	30.12
B2	29	0.25	4.12	2.77	12.83	71	0.04	1.45	0.44	8.84	0.05	2.21	0.63	10.63
C1	34	1.86	28.28	18.29	79.80	66	0.15	5.22	1.84	26.51	0.19	13.80	5.27	73.79
C2	23	0.31	5.52	3.80	15.86	77	0.05	1.65	0.43	10.11	0.06	2.53	0.59	12.30

Table 2: Error statistics (5th quantile, average, median, and 95th quantile) for the validation sets in Table 1. The overall results (cf. last four columns) are disaggregated into two groups according to the nature of the ODE solution (with/without sliding motion). The first column for each group gives the fraction of models considered.

Acknowledgement

The authors thank the anonymous reviewers for constructive comments which helped improve the quality of this paper.

This work has been partially sponsored by the EU project ASCENS, 257414.

9. REFERENCES

[1] V.O.K. Li, Wanjiun Liao, Xiaoxin Qiu, and E.W.M. Wong. Performance model of interactive video-on-demand systems. *Selected Areas in Communications, IEEE Journal on*, 14(6):1099–1109, aug 1996.

[2] M.L. Chaudhry and J.G.C. Templeton. *A First Course in Bulk Queues*. John Wiley and Sons, 1983.

[3] F.P. Kelly. *Reversibility and Stochastic Networks*. Cambridge University Press, 2011.

[4] Forest Baskett, K. Mani Chandy, Richard R. Muntz, and Fernando G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM*, 22(2):248–260, 1975.

[5] W. Henderson and P. Taylor. Product form in networks of queues with batch arrivals and batch services. *Queueing Systems*, 6:71–87, 1990. 10.1007/BF02411466.

[6] W. Henderson, C. Pearce, P. Taylor, and N. van Dijk. Closed queueing networks with batch services. *Queueing Systems*, 6:59–70, 1990. 10.1007/BF02411465.

[7] T. G. Kurtz. Solutions of ordinary differential equations as limits of pure Markov processes. *J. Appl. Prob.*, 7(1):49–58, April 1970.

[8] M.H.A. Davis. *Markov Models and Optimization*. Chapman & Hall, 1993.

[9] Luca Bortolussi. Hybrid limits of continuous time Markov chains. In *Proceedings of Eighth International Conference on the Quantitative Evaluation of Systems, QEST 2011*, pages 3–12. IEEE Computer Society, 2011.

[10] N. Gast and B. Gaujal. Mean field limit of non-smooth systems and differential inclusions. *SIGMETRICS Perform. Eval. Rev.*, 38:30–32, October 2010.

[11] M. Benaïm and J.Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Perform. Eval.*, 65(11-12):823–838, 2008.

[12] C. Bordenave, D. McDonald, and A. Proutiére. A particle system in interaction with a rapidly varying environment: Mean field limits and applications. *NHM*, 5(1):31–62, 2010.

[13] M. Tribastone, S. Gilmore, and J. Hillston. Scalable Differential Analysis of Process Algebra Models. *IEEE Transactions on Software Engineering*, 2010.

[14] A. Bobbio, M. Gribaudo, and M. Telek. Analysis of large scale interacting systems by mean field method. In *Proceedings of Fifth International Conference on the Quantitative Evaluaiton of Systems (QEST 2008)*, pages 215–224, 2008.

[15] G. Sharma, A.J. Ganesh, and P.B. Key. Performance analysis of contention based medium access control protocols. *IEEE Transactions on Information Theory*, 55(4):1665–1682, 2009.

[16] D. Qiu and R. Srikant. Modeling and performance analysis of bittorrent-like peer-to-peer networks. In *Proceedings of ACM SIGCOMM 2004*, pages 367–378, 2004.

[17] F. Clévenot and P. Nain. A simple model for the analysis of squirrel. In *Proceedings of INFOCOM 2004*, 2004.

[18] Mohammad Alizadeh, Adel Javanmard, and Balaji Prabhakar. Analysis of dctcp: stability, convergence, and fairness. In *Proceedings of ACM SIGMETRICS 2011*, pages 73–84, 2011.

[19] M. Alizadeh, A. Kabbani, B. Atikoglu, and B. Prabhakar. Stability analysis of qcn: the averaging principle. In *Proceedings of ACM SIGMETRICS 2011*, pages 49–60, 2011.

[20] A. Ganesh, S. Lilienthal, D. Manjunath, A. Proutiere, and F. Simatos. Load balancing via random local search in closed and open systems. In *Proceedings of ACM SIGMETRICS 2010*, pages 287–298, 2010.

[21] N. Gast and B. Gaujal. A mean field model of work stealing in large-scale systems. In *Proceedings of ACM SIGMETRICS 2010*, pages 13–24, 2010.

[22] R.W.R. Darling. Fluid limits of pure jump Markov processes: A practical guide. `arXiv. org`, 2002.

[23] R.W.R. Darling and J.R. Norris. Differential equation approximations for Markov chains. *Probability Surveys*, 5, 2008.

[24] M. Benaïm and J. Weibull. Deterministic

approximation of stochastic evolution in games. *Econometrica*, 2003.

[25] R. Hayden and J. T. Bradley. A fluid analysis framework for a Markovian process algebra. *Theoretical Computer Science*, 2010.

[26] L. Bortolussi. A master equation approach to differential approximations of stochastic concurrent constraint programming. In *Proceedings of the Sixth Workshop on Quantitative Aspects of Programming Languages (QAPL 2008)*, volume 220 of *ENTCS*, pages 163–180, 2008.

[27] L. Bortolussi. Limit behavior of the hybrid approximation of stochastic process algebras. In *Proceedings of 17th International Conference on Analytical and Stochastic Modeling Techniques and Applications, ASMTA 2010*, volume 6148 of *Lecture Notes in Computer Science*, pages 367–381. Springer, 2010.

[28] J. Cortes. Discontinuous dynamical systems: A tutorial on solutions, nonsmooth analysis, and stability. *IEEE Control Systems Magazine*, pages 36–73, 2008.

[29] Raymond M. Bryant, Anthony E. Krzesinski, and Peter Teunissen. The MVA pre-empt resume priority approximation. In *Proceedings of the 1983 ACM SIGMETRICS conference on Measurement and modeling of computer systems*, SIGMETRICS '83, pages 12–27, New York, NY, USA, 1983. ACM.

[30] Raymond M. Bryant, Anthony E. Krzesinski, M. Seetha Lakshmi, and K. Mani Chandy. The MVA priority approximation. *ACM Trans. Comput. Syst.*, 2:335–359, November 1984.

[31] Derek L. Eager and John N. Lipscomb. The AMVA priority approximation. *Performance Evaluation*, 8(3):173–193, 1988.

[32] William Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994.

[33] Luca Bortolussi, Vashti Galpin, Jane Hillston, and Mirco Tribastone. Hybrid semantics for pepa. In *QEST 2010, Seventh International Conference on the Quantitative Evaluation of Systems*, pages 181–190, Williamsburg, Virginia, USA, September 2010. IEEE Computer Society.

[34] A. Singh and J.P. Hespanha. Lognormal moment closures for biochemical reactions. In *Proceedings of 45th IEEE Conference on Decision and Control*, 2006.