Jump Transitions in Second Order FSPNs

Katinka Wolter*

Technische Universität Berlin, Institut für Technische Informatik,
Franklinstr. 28/29, 10587 Berlin, Germany.
E-mail: katinka@cs.tu-berlin.de

Abstract

In this paper jump transitions for the fluid model part in second order fluid stochastic Petri nets (FSPNs) are introduced. The extended formalism is defined, and the underlying partial differential equations are derived. With a jump a certain amount of fluid is added to a fluid place or taken out at once. This amount, the jump height, is sampled from a probability distribution. The dynamics of a model are described by second order partial differential equations that include integrals. As examples the virtual waiting time (or unfinished work) in a queueing system and a performability model of a multi-processor system are modelled.

1. Introduction

Stochastic Petri nets are a modelling formalism that is well suited for performance and reliability evaluation of complex computer and communication systems. To make them applicable for systems that have continuous components, like liquid in manufacturing systems or time, fluid places have been added, that are emptied and filled at a continuous rate [5, 16, 9]. This concept is known from queueing theory, where two kinds of continuous approximations to heavily loaded queues have been used, first and second order (fluid flow and diffusion) approximation [13]. The original fluid stochastic Petri nets (FSPNs) use the first order approximation, here the second order approximation is applied. In first order fluid models the flow rate at which fluid places are emptied and filled is deterministic whereas in second order models it is a normally distributed random variable. So, more complex systems can be modelled with second order models and the approximation quality is better. With zero variance again a first order model is obtained.

Whereas the formalisms in [16, 9] are based on stochastic Petri nets, the continuous Petri nets (CPN) in [5] extend the timed and non-timed Petri nets. The latter are evaluated with simulation. For FSPNs numerical solution [9, 8, 17] as well as simulation [3] have been applied.

In both first and second order FSPNs the content of a fluid place increases and decreases continuously in infinitesimal small amounts in a short time interval. So the random variable measuring the content of a fluid place glides in infinitesimal steps up and down the continuous axis. Opposed to this, jump transitions cause an immediate change of the level in the fluid place. There are several application fields, where an immediate increase or decrease of the fluid level is needed. When modelling industrial systems one might want to model the resetting of a clock, or the almost instantaneous emptying of a container. In a communication system buffers can be represented by fluid places and a reinitialization of the buffer needs a jump transition to the initial level. Furthermore, in performability models fluid places can be used to hold the accumulated reward. The incorporation of impulse rewards needs jump transitions in the fluid level. A special kind of jump arcs are the *flush out* arcs, that empty the fluid place. They have been introduced in [2, 7] to model non-exponentially distributed firing times with an FSPN and for modelling software reliability.

The idea to define fluid jump transitions has been evoked by the Takács process [15]. This is a special stochastic process with a continuous state space in which jumps and a deterministic decrease take place. A jump is defined by the rate at which it occurs and by its height distribution. In the FSPN context the rate of occurrence of a jump is given by the firing of an exponentially distributed transition. The underlying stochastic process is a Markov process with mixed continuous and discrete state space. The equations describing its dynamics are second order partial differential equations that include integrals. The integrals are over the convolution of the state probability density with the jump height distribution. The numerical solution of the equations involves numerical evaluation of the integrals as well as the numerical integration of the differential terms.

A simple queueing model is solved to study the behaviour of the formalism. The extended second order FSPN is used to model an M/M/1 and a M/D/1 queueing system.

*The author's work is supported by the German Research Council (DFG) under grant Ho 1257/11-1

The fluid place represents the amount of unfinished work in the system (or the virtual waiting time) that increases upon each arrival of a customer with a jump by it's service time and decreases linearly with time. From the full probability density for that measure the mean will be computed and compared with the known analytical stationary values.

In the subsequent example the fluid jumps are used to instantaneously empty the fluid place that represents the buffer in a multiprocessor system. The processors may fail independently, so the system has several configurations (discrete states). Upon failure of the buffer its complete content is lost. Mean buffer content and throughput are the measures that are computed for this model.

In Section 2 second order FSPNs with jumps are described, formally defined and the computation of parameters as they are needed for the transient analysis is demonstrated. In Section 3 the differential equations for the special case of only one fluid place, as it is used in the examples, are derived and initial and boundary conditions are introduced. Section 4 holds the examples and with Section 5 this paper concludes.

2. The Formalism

For the definition of second order FSPNs the common notations for GSPNs [1] and for FSPNs [9] are taken over and extended as follows.

A second order FSPN is formally defined as a 9-tuple FSPN = $(\mathcal{P}, \mathcal{T}, \mathcal{A}, m_0, g, \lambda, r, b, w)$ where the set of places ($\mathcal{P} = \mathcal{P}_c \cup \mathcal{P}_d$) is divided into the *fluid (continuous)* and the *discrete* places. Discrete places (the elements of $\mathcal{P}_d$) are drawn as single lined circles and hold an integer number of *tokens*, whereas the continuous, or fluid, places (the elements of $\mathcal{P}_d$) are drawn as two concentrical circles and they hold a real-valued amount of *fluid*. The set of transitions $\mathcal{T} = \mathcal{T}_E \cup \mathcal{T}_I$ is composed of the exponentially distributed, and the immediate transitions, respectively. Exponentially distributed transitions are drawn as empty rectangles, the generally distributed ones as hatched rectangles and the immediate transitions as thin bars.

The set of arcs $\mathcal{A} = \mathcal{A}_d \cup \mathcal{A}_c \cup \mathcal{A}_j$ is divided into three subsets, the discrete arcs (the elements of $\mathcal{A}_d$), the continuous, or fluid, arcs ($\mathcal{A}_c$) and the jump arcs ($\mathcal{A}_j$). The discrete and the jumps arcs are drawn as single lined arrows, whereas the fluid double-lined arcs are drawn like pipes. The jump arcs only connect transitions and fluid places, and vice versa. Immediate transitions can be connected to a fluid place by a jump arc only if the immediate transition is also connected to a discrete place. In other words, the fluid jumps caused by the firing of an immediate transition can not occur without a discrete model part connected to that same immediate transition. Opposed to this, jumps caused by the firing of a timed transition may stand alone.

The *marking* $\mathbf{m} = (s, \mathbf{z}) \in \mathcal{M}$ consists of a discrete part $s = (\#p_i, i \in \mathcal{P}_d)$, where $\#p_i$ denotes the number of discrete tokens in the i-th discrete place p_i, and a continuous part, a vector representing the fluid level in each fluid place, $\mathbf{z} = (z_k, k \in \mathcal{P}_c) \in \mathcal{Z}$. The initial marking is $\mathbf{m}_0 = (s_0, \mathbf{z}_0)$. The total number of discrete states is S, the set of all discrete states is denoted $\mathcal{S}$.

$\mathrm{r}_{t,p}$, $\mathrm{r}_{p,t}$: $\mathcal{A}_c \times \mathcal{M} \to \mathbb{R}^2$ is the flow rate function along the fluid arc connecting the timed transition t and the fluid place p or vice versa. The flow rate $r = (\mu, \sigma^2)$ is a normally distributed random variable specified by expectation μ and variance σ^2. Let Z_k be the fluid level in place $k \in \mathcal{P}_c$. The change rate of the fluid level is again a normally distributed random variable with

$$\frac{dZ_k}{dt} = \sum_{t \text{ enabled in } \mathbf{m}} r_{t,p}(\mathbf{m}) - \sum_{t \text{ enabled in } \mathbf{m}} r_{p,t}(\mathbf{m}).$$

Reflecting barriers assure, that the fluid level in each place stays within its range $[z^{\min}, z^{\max}]$.

The jump height function $\mathrm{b}_{t,p}$, $\mathrm{b}_{p,t}$: $\mathcal{A}_j \times \mathcal{M} \to \mathcal{F}$ is a probability density function associated to a jump arc. The size of a fluid jump is a random variable drawn from that distribution.

$g_t : \mathcal{M} \to \mathbb{B}$ is the guard of transition t that can be a function of the discrete and the continuous state. It evaluates to *true* or *false*. $\lambda_t : \mathcal{M} \to \mathbb{R} \cup \{\infty\}$ is a function of both the continuous and the discrete marking. Immediate transitions have firing rate $\lambda_t = \infty$.

The weight function $w_t : \mathcal{S} \to \mathbb{R}$ is defined for immediate transitions. The firing probability of each of the enabled immediate transitions in a vanishing state is $w_t(s) / \sum_{t_i \text{enabled in } \mathbf{m}} w_{t_i}(s)$.

The enabling rules for the transitions are the same as in discrete Petri net models. The fluid and the jump arcs act only on the fluid places and do not influence the enabling conditions of the transitions. Transitions are enabled only by discrete arcs and a guard that can be a function of the discrete and the continuous state. Also the firing rates can be functions of the discrete and the continuous state.

2.1. The model parameters

The dynamics of the CTMC underlying the discrete model part are defined through the generator matrix $\mathbf{Q}$, whereas the continuous dynamics are defined by the fluid change rate $\frac{dZ_k}{dt}$ that describes the change in the fluid level over time. The change in the fluid level is determined by the difference between the inflow and the outflow, each of which is normally distributed. The parameters of the normal distribution are either specified by the user or calculated as the diffusion approximation to the discrete transition at the origin of the inflow arcs and at the destination of the outflow arcs. The diffusion approximation has been carried out in detail in [17], here only the results will be repeated.

Let μ_a and σ_a^2 be the expectation and variance of an arbitrary distribution of the discrete input transition to the fluid place P_k. Then the infinitesimal mean and variance of the fluid inflow rate are

$$\mu_\alpha = \frac{1}{\mu_a}, \qquad \sigma_\alpha^2 = \frac{\sigma_a^2}{(\mu_a)^3}. \tag{1}$$

and for the fluid outflow rate μ_δ and σ_δ^2 are analogous. (See [11, 13] for a derivation of this formula).

All the incoming and outgoing streams in one marking add and for the flow rate $R_k(s, \mathbf{z})$ of the fluid place k in every marking $s \in \mathcal{S}$ the following parameters are valid

$$\mu_{k,s} = \sum_{inflows} \mu_\alpha - \sum_{outflows} \mu_\delta \quad \forall k \in \mathcal{P}_c, s \in \mathcal{S}, \tag{2}$$

$$\sigma_{k,s}^2 = \sum_{inflows} \sigma_\alpha^2 + \sum_{outflows} \sigma_\delta^2 \quad \forall k \in \mathcal{P}_c, s \in \mathcal{S}. \tag{3}$$

because variances must add. Note, that this simple linear combination only holds, because all the single fluid streams are normally distributed. The expectations and variances for each fluid place are collected into diagonal matrices (if there is none in a discrete state, the entry is equal to zero):

$$\boldsymbol{M}_k = \mathbf{diag}(\mu_{k,1}, \dots, \mu_{k,S}), \tag{4}$$

$$\mathbf{\Sigma}_k^2 = \mathbf{diag}(\sigma_{k,1}^2, \dots, \sigma_{k,S}^2). \tag{5}$$

If a hybrid system is modelled $\mu_\alpha, \sigma_\alpha^2, \mu_\delta$ and σ_δ^2 are given by the user and (1) is not needed.

The jump height function $b(.)$ can be an arbitrary probability density. If for instance the exponential distribution is chosen jumps may be very large. Depending on the rate of the exponential distribution the probability of having a large jump may not be neglectable. What is large also depends on the range of the fluid place. Since the fluid places may only have positive content there always is a boundary at zero. For the numerical solution of the equations there also must be an upper bound, which may be quite large. To make sure, that the fluid jumps will not go across the boundaries the jump height distribution must be transformed. Two different transformations have been considered. In the one a jump, that would go beyond the boundary is cancelled (this jump strategy is called *preserve*) and in the other the height of the jump is reduced, so that the destination just lies inside the range of the fluid place (this jump strategy is called *force_jump*).

If z is the observed fluid level at time t and β is the jump height drawn from the distribution function $b(.)$ then the transformed function $b(.)$ is

$$\text{force\_jump} \;\cong\; b(.) = \begin{cases} z^{\max} - z & \text{if } z + \beta > z^{\max} \\ \beta & \text{else} \end{cases} \tag{6}$$

$$\text{preserve} \;\cong\; b(.) = \begin{cases} 0 & \text{if } z + \beta > z^{\max} \\ \beta & \text{else} \end{cases} \tag{7}$$

The definitions (6) and (7) are for jumps up, those for jumps down are formulated analogously.

3. Analysis

The derivation of the differential equations will be carried out for the special case of only one fluid place, since the examples presented later use only one fluid place. Extension to more continuous dimensions is straight forward and is carried out along the same lines. The difficulty with fluid models lies more in the solution of the differential equations than in their formulation.

The stochastic process under consideration is a Markov process in continuous time with mixed discrete and continuous state space (see [4]),

$$\{(S(t), Z(t)), t \in \mathbb{R}_0^+\}$$

where $S(t)$ is the discrete marking at time t, and $Z(t)$ is a random variable, representing the fluid level in the fluid place at time t.

The jumps up are enumerated by $\hat{l} = 1, \dots, \hat{L}$, the jumps down by $\check{l} = 1, \dots, \check{L}$ the union of both, i.e. all jumps in total are enumerated by $l = 1, \dots, L$. Associated to a jump is the rate λ at which it occurs and its height distribution $b(.)$ (In vector notation over all discrete states they are $\mathbf{\Lambda}$ and $\mathbf{B}(.)$). We denote by $\mathbf{\Lambda}(z) = \sum_{l=1}^{L} \mathbf{\Lambda}_l(z)$ the sum over all jump rates.

At time t the transient probability of being in discrete state i with fluid level in an infinitesimal environment around z, is called the *fluid density* and is denoted by $\pi_i(t, z) = d/dz \; P(S(t) = i, Z(t) \leq z)$. The probabilistic behaviour of the model is described only by this density. The initial impulse is approximated numerically by a triangular density. No mass component is taken into account. This can be considered an approximation. The incorporation of a probability mass is left for future work.

In the following the differential equations describing the dynamics of an FSPN are derived. The Chapman-Kolmogorov equation of the process is constituted by looking ahead an infinitesimal step in time

$$\boldsymbol{\pi}(t+\Delta, z) = \underbrace{\int \boldsymbol{\pi}(t+\Delta, z|t, z')\, \mathbf{diag}\left(\boldsymbol{\pi}(t, z')\right) dz'}_{\text{fluid flow}}$$
$$+ \boldsymbol{\pi}(t,z)\Big(\underbrace{\mathbf{Q}(z)}_{\text{discrete part}} - \underbrace{\boldsymbol{\Lambda}(z)}_{\text{jumps}} \Big)\Delta +$$
$$\underbrace{\sum_{\hat{l}=1}^{\hat{L}} \int_{z^{\min}}^{z} \boldsymbol{\pi}(t,y)\mathbf{B}_{\hat{l}}(z-y)\, dy\, \boldsymbol{\Lambda}_{\hat{l}}(z)\,\Delta +}_{\text{jumps up}}$$
$$\sum_{\check{l}=1}^{\check{L}} \underbrace{\int_{z}^{z^{\max}} \boldsymbol{\pi}(t,y)\mathbf{B}_{\check{l}}(y-z)\, dy\, \boldsymbol{\Lambda}_{\check{l}}(z)}_{\text{jumps down}}\,\Delta + o(\Delta) \quad (8)$$

This equation is formulated by looking at what can happen in a system in a small fraction of time, as it is done for deriving the differential equations for a continuous time Markov chain (CTMC). In the case of the CTMC balance equations are formulated by summing for each state over all the ingoing rates minus all the outgoing rates. The same is done here for the discrete state transitions. Another possibility for leaving the mixed discrete and continuous state is to move on to a different continuous state. The first term on the right hand side of (8) collects the probability of observing the continuous state z at time $t+\Delta$, conditional on all possible previous continuous states z' which the system could have been in at time t.

Both the last terms in (8) collect the probability of observing a jump from level y to the level z with height $z-y$ for a jump up or with height $y-z$ for a jump down. In both cases integration is over all possible previous levels y smaller than z for the jumps up and greater z for the jumps down.

Now the first integral of (8) is regarded in more detail. Every component $\pi_s(t,z')$ of the vector $\boldsymbol{\pi}(t,z')$ is expanded in a Taylor series around z

$$\pi_s(t,z') = \pi_s(t,z) + (z'-z)\frac{\partial \pi_s(t,z)}{\partial z} + \frac{(z'-z)^2}{2}\frac{\partial^2 \pi_s(t,z)}{\partial z^2} + o((z'-z)^2) \quad (9)$$

The second part of the integrand is substituted by the Taylor series, so the first integral in (8) becomes

$$\begin{aligned}
&\int \pi_s(t+\Delta, z|t,z')\Big(\pi_s(t,z) + (z'-z)\frac{\partial \pi_s(t,z)}{\partial z} \\
&\quad + \frac{(z'-z)^2}{2}\frac{\partial^2 \pi_s(t,z)}{\partial z^2} + o(z'-z)\Big)dz' \\
&= \int \pi_s(t+\Delta, z|t,z')\pi_s(t,z)dz' \\
&\quad + \int \pi_s(t+\Delta, z|t,z')(z'-z)\frac{\partial \pi_s(t,z)}{\partial z}dz' + \\
&\int \pi_s(t+\Delta, z|t,z')\frac{(z'-z)^2}{2}\frac{\partial^2 \pi_s(t,z)}{\partial z^2}dz' + o((z'-z)^2)
\end{aligned} \quad (10)$$

Since the integration is over z' all terms not containing z' can be taken outside the integrals. These are the partial derivatives. The conditional density $\pi_s(t+\Delta, z|t,z')$ represents the change in the continuous variable in a time interval of length Δ. It is the distribution of $Z(t+\Delta)-Z(t)$, which is normally distributed with mean $\mu\Delta$ and variance $\sigma^2\Delta$. Hence,

$$\pi_s(t+\Delta, z|t,z') = \frac{1}{\sqrt{2\pi\sigma^2\Delta}}\exp\left(\frac{-\left((z-z')-\mu\Delta\right)^2}{2\sigma^2\Delta}\right) \quad (11)$$

is the normal distribution $\Phi(z-z')$. It should be noted, that (11) is a normal distribution also with respect to $-z'$ (with mean $(-z+\mu\Delta)$). The normal distribution is substituted in (10)

$$\begin{aligned}
&\int \pi_s(t+\Delta, z|t,z')\pi_s(t,z')\, dz' = \pi_s(t,z)\underbrace{\int_\Omega \Phi(z-z')\, dz'}_{=1} \\
&\quad + \frac{\partial \pi_s(t,z)}{\partial z}\underbrace{\int_\Omega (z'-z)\Phi(z-z')\, dz'}_{-\mu(z)\Delta \text{ (negative first moment)}} \\
&+ \frac{1}{2}\frac{\partial^2 \pi_s(t,z)}{\partial z^2}\underbrace{\int_\Omega (z-z')^2\Phi(z-z')\, dz'}_{\sigma^2(z)\Delta+\mu^2(z)\Delta^2 \text{ (second moment)}} + o((z-z')^2) \\
&= \pi_s(t,z) - \frac{\partial \pi_s(t,z)}{\partial z}\mu(z)\Delta + \frac{1}{2}\frac{\partial^2 \pi_s(t,z)}{\partial z^2}(\sigma^2(z)\Delta + \mu^2(z)\Delta^2)
\end{aligned} \quad (12)$$

Expectation and variance are defined as the limits (see [6], p. 344)

$$\lim_{\Delta\to 0}\frac{1}{\Delta}\int (z'-z)\pi_s(t+\Delta, z|t,z')dz' = \mu(z) \qquad \text{and} \quad (13)$$

$$\lim_{\Delta\to 0}\frac{1}{\Delta}\int (z'-z)^2\pi_s(t+\Delta, z|t,z')dz' = \sigma^2(z). \quad (14)$$

In vector notation (12) is

$$\boldsymbol{\pi}(t,z) - \frac{\partial \boldsymbol{\pi}(t,z)}{\partial z}\mathbf{M}(z)\Delta + \frac{1}{2}\frac{\partial^2 \boldsymbol{\pi}(t,z)}{\partial z^2}(\boldsymbol{\Sigma}^2(z)\Delta + \mathbf{M}^2(z)\Delta^2) \forall t, \forall z \in \overset{\circ}{\Omega} = (z^{\min}, z^{\max}). \quad (15)$$

The first integral in (8) is replaced by (15), $\boldsymbol{\pi}(t,z)$ is subtracted from both sides and the equation is divided by Δ:

$$\frac{\boldsymbol{\pi}(t+\Delta,z) - \boldsymbol{\pi}(t,z)}{\Delta} = -\frac{\partial}{\partial z}\Big(\boldsymbol{\pi}(t,z)\mathbf{M}(z)\Big) + \frac{1}{2}\frac{\partial^2}{2\partial z^2}\Big(\boldsymbol{\pi}(t,z)(\Delta\mathbf{M}^2(z) + \boldsymbol{\Sigma}^2(z))\Big) + \boldsymbol{\pi}(t,z)\Big(\mathbf{Q}(z) - \boldsymbol{\Lambda}(z)\Big) + \sum_{\hat{l}=1}^{\hat{L}} \int_{z^{\min}}^{z} \boldsymbol{\pi}(t,y)\mathbf{B}_{\hat{l}}(z-y)\, dy\, \boldsymbol{\Lambda}_{\hat{l}}(z) + \sum_{\check{l}=1}^{\check{L}} \int_{z}^{z^{\max}} \boldsymbol{\pi}(t,y)\mathbf{B}_{\check{l}}(y-z)\, dy\, \boldsymbol{\Lambda}_{\check{l}}(z) + o(\Delta) \quad (16)$$

As Δ approaches zero the second Kolmogorov forward equation is obtained

$$\frac{\partial \boldsymbol{\pi}(t,z)}{\partial t} + \frac{\partial}{\partial z}\Big(\boldsymbol{\pi}(t,z)\mathbf{M}(z)\Big) - \frac{1}{2}\frac{\partial^2}{\partial z^2}\Big(\boldsymbol{\pi}(t,z)\boldsymbol{\Sigma}^2(z)\Big) - \boldsymbol{\pi}(t,z)\Big(\mathbf{Q}(z) - \boldsymbol{\Lambda}(z)\Big) - \sum_{\hat{l}=1}^{\hat{L}} \int_{z^{\min}}^{z} \boldsymbol{\pi}(t,y)\mathbf{B}_{\hat{l}}(z-y)\, dy\, \boldsymbol{\Lambda}_{\hat{l}}(z) + \sum_{\check{l}=1}^{\check{L}} \int_{z}^{z^{\max}} \boldsymbol{\pi}(t,y)\mathbf{B}_{\check{l}}(y-z)\, dy\, \boldsymbol{\Lambda}_{\check{l}}(z) = 0 \qquad \forall t, \forall z \in \overset{\circ}{\Omega}. \quad (17)$$

The derivatives form the well known multi-dimensional convection-diffusion equation, also known as Fokker-Planck equation and the generator matrix introduces a coupling term of the independent equations for each discrete state. The integrals are equivalent to those in the Takács integro-differential equation.

An initial as well as a boundary condition on either end of the regarded domain must be imposed. The initial condition is defined by the models' configuration at time zero, whereas the boundary conditions are determined by the fact that the integral over a density always evaluates to one.

Let $\boldsymbol{\pi}_0$ be the vector of initial discrete state probabilities and let $\mathbf{z}_0$ be the vector of initial fluid levels, then the inital conditions are

$$\boldsymbol{\pi}(0,z) = \boldsymbol{\pi}_0 \cdot \delta(z - z_0), \quad (18)$$

where $\delta(z - z_0)$ is the Dirac impulse at $z = z_0$.

A reflecting barrier at the origin and at the upper bound define the boundary condition for the fluid flow [4].

$$\left[\frac{1}{2}\frac{\partial}{\partial z}\boldsymbol{\pi}(t,z)\boldsymbol{\Sigma}^2(z) - \boldsymbol{\pi}(t,z)\mathbf{M}_k(z)\right]_{z=\{z^{\min},z^{\max}\}} = 0. \quad (19)$$

Eqn. (17) is a parabolic second order partial differential equation. It has been solved with a theta method, that is a flexible implicit-explicit discretization scheme [18]. For the first order derivative in space the upwind method has been used and the second order derivative is discretized in a central difference. The integration and the term with the generator matrix are discretized explicitly in each time step. The integration has been approximated by a simple rectangular sum. If the variance of the fluid change is zero in a state the second derivative term cancels out and the equation is of hyperbolic instead of parabolic type. It appeared, however, that the numerical algorithm designed for the parabolic equations works correctly also if the variance equals zero.

In a model with one fluid place a fraction of the real axis as long as the range of the fluid place has to be discretized. Solving a model with two fluid places in an analogous way requires the discretization of the cross product of the ranges of the two fluid places. For three fluid places it would be the cross product of all three ranges. Hence, the number of discretization points increases exponentially with the number of dimensions (fluid places). In consequence the algorithm becomes slow. For one and two fluid places the method is implemented and works well. To increase its speed in the more-dimensional case alternating-directions implicit (ADI) schemes can be used, where the dimensions in space are decoupled and only the sum of the grid points in each direction is needed, not their product. ADI schemes, however, are less accurate. Hence, models with three fluid places will still be solvable but there is a limit due to numerical complexity.

The numerical methods will be discussed in detail in a separate paper.

4. Examples

The behaviour of the second order FSPN model with jumps is first studied in a very simple queueing example. The unfinished work (or virtual waiting time) of a queueing system, as shown in figure 1 on the left is regarded. In the fluid model, shown in figure 1 on the right not the queue is modelled but the unfinished work itself. An M/M/1 and a M/D/1 queue will be regarded. With the fluid place not the customers but their waiting time is modelled. Transition t_3 represents the arrival process to the queue (analogous to transition t_1 in the discrete model). With each arrival of a customer its service time is added to the unfinished work

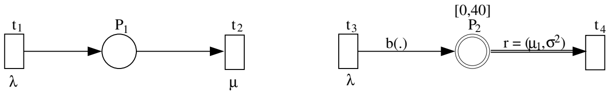

Figure 1. Discrete and fluid Petri net model

that is held by the fluid place. The addition of the service time of a customer causes an immediate increase of the fluid level. The height of this jump up is determined by the service time distribution. So, the distribution of transition t_2 in the discrete model equals the jump height distribution $b(.)$.

The capacity of the fluid place has been chosen sufficiently large, so that no probability mass reaches the upper bound and the boundary condition does not effect the solution. Hence, the solution of the fluid model is identical to the solution of an unrestricted fluid model. The z-axis has length 40, that is $z^{\min} = 0, z^{\max} = 40$.

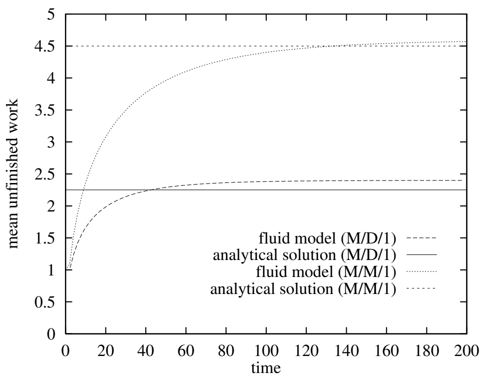

Figure 2. Mean unfinished work

The arrival rate to the queue, that is the firing rate of the transitions t_1 and t_3, is set to $\lambda = \frac{1}{2}$. The parameters of the service time distribution are chosen in such a way that the utilization ρ of the system is the same for the exponentially distributed and the deterministic service times. In the discrete model the exponentially distributed service rate, which is the firing rate of t_2 is set to $\mu = \frac{2}{3}$ (and consequently $\rho = \frac{3}{4}$). The corresponding fluid model has an exponential jump height distribution with rate μ as well. So, $b(.) \cong$ Exponential(μ). The number of tokens in P_1 is the number of customers in the system, whereas the amount of fluid in P_2 corresponds to the amount of unfinished work in the system. The fluid outflow represents the work performed per time unit by the server. In a single server system the work that is performed by the server is deterministic at rate one, as long as the queue is not empty. Hence, the flow rate $r \sim N(1, 0)$.

The mean unfinished work in an M/M/1 queue (see [10]) is

$$\overline{U} = \frac{\rho/\mu}{1-\rho} = 4.5, \tag{20}$$

for the values of the parameters chosen here. The fluid model is solved transiently. The solution is a full probability density from which the mean is computed. It is expected to approach the analytical solution as the model reaches equilibrium.

In the M/D/1 queue a service time of $3/2$ time units has been used. The analytical result for the stationary unfinished work is given by the Pollaczek-Khinchin formula (see e.g. [10])

$$\overline{U} = \frac{\lambda\overline{x^2}}{2(1-\rho)} = 2.25. \tag{21}$$

for the parameters chosen here. The curves for both service time distributions are shown in figure 2.

Another model shall be regarded in order to study the usefulness of the fluid jumps. It is a performability model of a degradable buffered multiprocessor system taken from [12]. Each of the processors fails at rate λ_p. The repair rate is r_p. The buffer has L stages, each of which may fail independently at rate λ_b causing a complete system crash. The failure rate of the whole buffer is then $L\lambda_b$ and the repair rate of the whole system after a buffer failure is r_b. Jobs

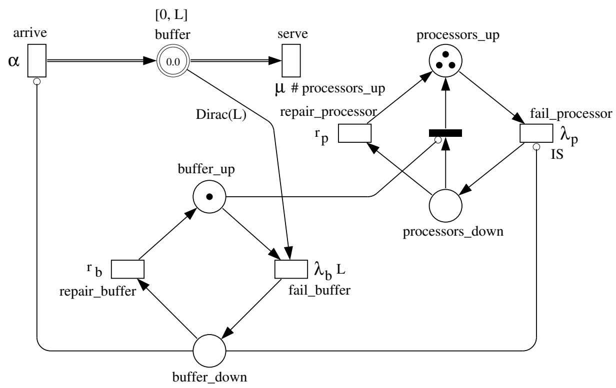

Figure 3. FSPN model of the performability model

arrive to the system at rate α and they are served at rate μ_p per processor. Upon failure of the buffer all data gets lost. When repair takes place not only the buffer but also the failed processors are repaired and the whole system is reinitialized.

Figure 3 shows an FSPN model of the system. The buffer is represented by the fluid place. Its capacity is $L = 4$. A jump arc connects the buffer to transition *fail_buffer*. The jump height distribution is deterministic with height 4.0 and the strategy is force_jump. So, when the buffer fails it is emptied immediately. The jump height is chosen in such a way, that the maximum buffer content is removed. If the buffer contains less than its maximum content then the jump strategy makes sure that less fluid is removed, so the place is emptied. When the buffer is down all the failed processors are put back into the set of operational processors. As long as repair of the buffer takes all the processors are in maintenance and repair and the transition representing the buffer failure is disabled. The arrival of jobs then is disabled too. No service takes place when the buffer is empty.

The reachability graph is shown in figure 4. Only the discrete states are labelled, the component for the fluid level is omitted. The numbers in the small circles denote the numbering of the discrete states as it is used for the parameter vectors and the results. The failure rate of the buffer has been set to $\lambda_b = 10^{-4}$ per hour and the processor failure rate is $\lambda_p = 10^{-3}$ per hour. The repair time for a failed processor $1/r_p$ is on the average one hour, whereas the repair time $1/r_b$ for the whole system is 6 hours. The mission time is one week (i.e. 168 hours).

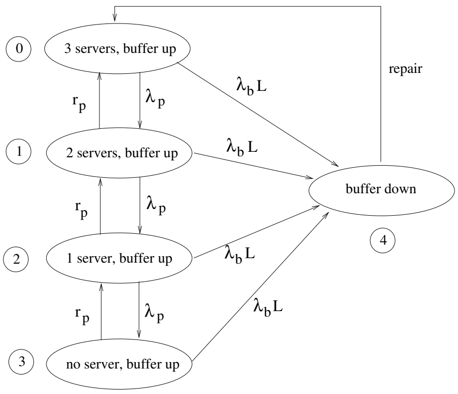

Figure 4. Reachability graph of the FSPN model

For the computation of the fluid change rates (1), (2) and (3) have been used. Two different service time parameters have been used. In all states s where the buffer is operational the expectation E_r of the fluid change rate r is computed as

$$E_r = E_\alpha - \#\text{servers\_up} \cdot E_\mu$$

and the variance V_r equals

$$V_r = V_\alpha + \#\text{servers\_up} \cdot V_\mu.$$

The arrival stream is assumed to be the approximation of a distribution with mean $\mu_a = \frac{4}{3}$ and variance $\sigma_a^2 = \frac{16}{9}$. This corresponds to an exponential distribution with rate $\alpha = 0.75$, since the exponential distribution has mean $\frac{1}{\alpha}$ and variance $\frac{1}{\alpha^2}$. Hence (1) evaluates to $E_\alpha = \frac{3}{4}$ and

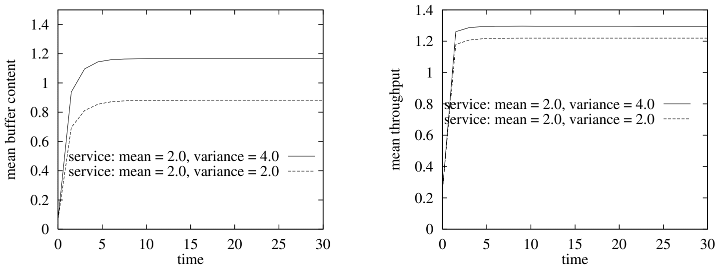

Figure 5. Mean buffer content and throughput

$V_\alpha = \frac{3}{4}$. The first service time distribution has mean $\mu_{s1} = 2.0$ and variance $\sigma_{s1}^2 = 4.0$. It can be interpreted as an approximation of exponentially distributed service with rate $\mu = 0.5$. Here (1) evaluates to $E_{\mu_{s1}} = \frac{1}{2}$ and $V_{\mu_{s1}} = \frac{1}{2}$. For this model the fluid flow parameters evaluate to

$$\boldsymbol{M} = \textbf{diag}(-0.75, -0.25, 0.25, 0.75, 0.0) \tag{22}$$

and

$$\boldsymbol{\Sigma}^2 = \textbf{diag}(2.25, 1.75, 1.25, 0.75, 0.1). \tag{23}$$

The second set of parameters for the service time distribution is $\mu_{s2} = 2.0$ and $\sigma_{s2}^2 = 4.0$. This can not be an approximated exponential distribution, because the relation of mean and variance is fixed by the parameter of the distribution. It could instead be an approximated normal distribution ($N(\mu_{s2}, \sigma_{s2}^2)$), which allows more flexibility with its two parameters. The expectation equals still $E_{\mu_{s2}} = \frac{1}{2}$ and the variance is $V_{\mu_{s2}} = \frac{1}{4}$, so that the variance of the fluid change now is

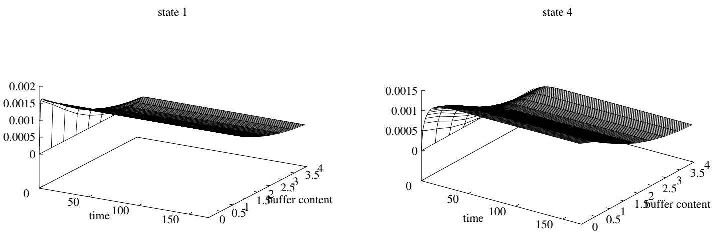

Figure 6. Probability densities in states one and four

$$\boldsymbol{\Sigma}^2 = \textbf{diag}(1.5, 1.25, 1.0, 0.75, 0.0). \tag{24}$$

Note, that using only the first two moments of a distribution is a simplification of the model and the results are not the exact ones. The quality of the diffusion approximation has been studied for queueing systems in [14].

The mean queue length is computed as the mean content of the fluid place and the throughput as the service rate times the probability of the queue being non-empty multiplied with the number of operational processors. Due to the high repair rate of the processors (compared to the failure rates) the system very quickly reaches steady-state. Therefore in figure 5 only the first 30 time units are shown.

The variance of the second service time distribution is half of the variance of the first service time distribution while the means are equal. It can be seen in figure 5 that the mean buffer content is larger for the model with the larger variance and so is the throughput. It is known that higher variability in the service time distribution increases the waiting time and the queue length. A larger mean queue length reduces the probability of an empty queue and hence increases the throughput.

The densities in two of the five discrete states are shown in figure 6.

5. Conclusions

Second order FSPNs have been extended by introducing jump transitions in the continuous model part. The jumps have a height following a probability distribution and occur together with an exponentially distributed state transition. Three different probability densities have been implemented, the exponential distribution, the uniform distribution and for deterministic jump height the Dirac impulse.

The dynamics of an FSPN with jumps are described by a system of differential equations that include convolution integrals for the jumps. The differential equations have been discretized with a flexible mixed explicit-implicit scheme and the integrals have been evaluated numerically.

The first example is equivalent to a simple discrete model that can be solved analytically. The fluid model gives the full probability density for the unfinished work in the system, a measure that usually is very hard to obtain.

In the second example the jump arc is used to empty the fluid place when the buffer fails. A height distribution must be specified for each jump arc. But it is also possible to completely empty or fill the fluid place by choosing a large deterministic jump height and the strategy force_jump.

A number of other applications should be investigated. On the one hand there is the modelling of hybrid systems where the immediate change of a continuous variable is often needed and on the other hand there are performability models, where impulse rewards can be modelled with fluid jump arcs. The accumulated reward then is modelled as the content of the fluid place and upon the firing of transition a certain amount is added.

References

[1] M. Ajmone Marsan, G. Balbo, G. Conte, and S. Donatelli. *Modelling with Generalized Stochastic Petri Nets*. Series in Parallel Computing. John Wiley & Sons, 1995.

[2] A. Bobbio, S. Garg, M. Gribaudo, A. Horváth, M. Sereno, and M. Telek. Modeling Software Systems with Rejuvenation, Restoration and Checkpointing through Fluid Stochastic Petri Nets. In *Proc. Eighth International Workshop on Petri Nets and Performance Models - PNPM'99*, Zaragoza, Spain, 1999. to appear.

[3] G. Ciardo, D. Nicol, and K. Trivedi. Discrete-event Simulation of Fluid Stochastic Petri Nets. In *Proc. Seventh International Workshop on Petri Nets and Performance Models - PNPM'97*, pages 217–225, Saint Malo, France, June 3–6 1997. IEEE-CS Press.

[4] D. Cox and H. Miller. *The Theory of Stochastic Processes*. Chapman and Hall, 1965.

[5] R. David and H. Alla. Autonomous and Timed Continuous Petri nets. In G. Rozenberg, editor, *Advances in Petri Nets 1993*, volume 674 of *Lecture Notes in Computer Science*, pages 71 – 90. Springer-Verlag, 1993.

[6] B. V. Gnedenko. *The Theory of Probability*. Chelsea Publishing Company, New York, 1962.

[7] M. Gribaudo, M. Sereno, and A. Bobbio. Fluid Stochastic Petri Nets: An Extended Formalism to Include non-Markovian Models. In *Proc. Eighth International Workshop on Petri Nets and Performance Models - PNPM'99*, Zaragoza, Spain, 1999. to appear.

[8] G. Horton. Computation of the Distribution of Accumulated Reward with Fluid Stochastic Petri nets. In *Proc. 2nd IEEE Int. Computer Performance and Dependability Symposium (IPDS)*, pages 90–95, Urbana-Champaign, Illinois, USA, 4-6 September 1996. IEEE-CS Press.

[9] G. Horton, V. G. Kulkarni, D. M. Nicol, and K. S. Trivedi. Fluid Stochastic Petri Nets: Theory, Applications and Solution. *European Journal of Operations Research*, 105(1):184–201, February 1998. Also published as ICASE report no. 96-5.

[10] L. Kleinrock. *Queueing Systems Vol I: Theory*. John Wiley, 1975.

[11] L. Kleinrock. *Queueing Systems Vol II: Computer Applications*. John Wiley, 1976.

[12] J. F. Meyer. Closed-Form Solutions of Performability. *IEEE Transactions on Computers*, C-31(7):648–657, July 1982.

[13] G. F. Newell. *Applications of Queueing Theory*. Chapman and Hall, Ltd. (London), 1971.

[14] M. Reiser and H. Kobayashi. Accuracy of the Diffusion Approximation for Some Queuing Systems. *IBM Journal of Research and Development*, 18(2):110–124, Mar. 1974.

[15] L. Takács. Investigation of waiting time problems by reduction to Markov processes. *Acta Math. Acad. Sci. Hungar.*, 6:101–129, 1955.

[16] K. S. Trivedi and V. G. Kulkarni. FSPNs: Fluid Stochastic Petri Nets. In *Proc. 14th Int. Conf. on the Application and Theory of Petri Nets*, pages 24–31, Chicago, 1993.

[17] K. Wolter. Second Order Fluid Stochastic Petri Nets: an Extension of GSPNs for Approximate and Continuous modelling. In *Proc. 1st World Congress on Systems Simulation (WCSS'97)*, pages 328–332, Singapore, Sept. 1–3 1997.

[18] A. Zisowsky. Entwurf und Implementierung eines Verfahrens für die transiente Analyse fluider stochastischer Petri-Netze. Master's thesis, TU Berlin, August 1998.