On Statistical Stability Analysis of Discrete-Time Markov Chains

Sergey Astafiev, Alexander Rumyantsev

Institute of Applied Mathematical Research Karelian Research Centre, RAS,
Petrozavodsk State University, Petrozavodsk, Russia.

The research is partly supported by Russian Foundation for Basic Research,
projects 19-07-00303, 19-57-45022.

Petrozavodsk, Russia, 2021

Introduction

Multicore systems are widely used both on large and small scale computing systems

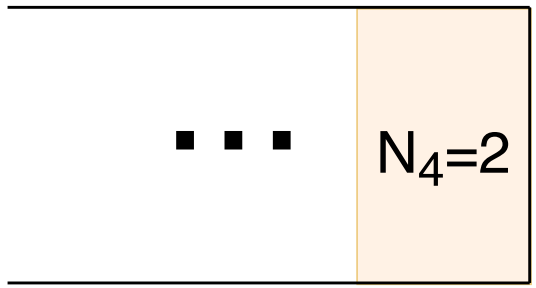
- Sensors, IoT, wireless connectivity, edge computing
- Handheld computing, smart health
- High-Performance Computing

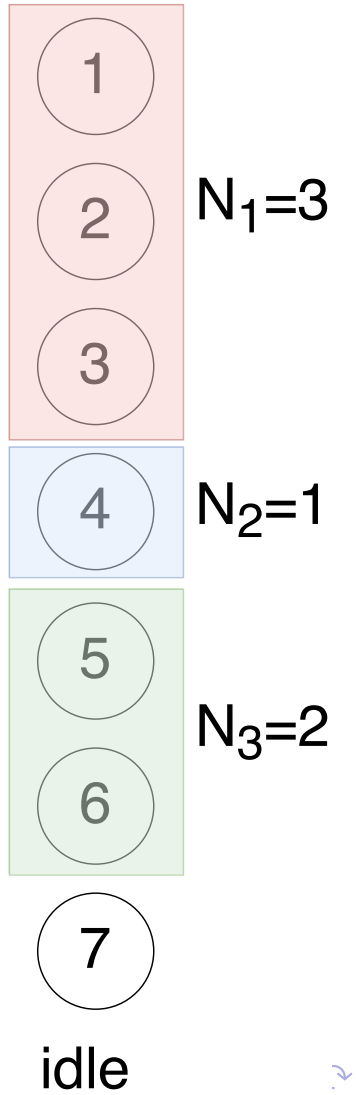
In these models, the problem of checking their stability for a given set of input parameters arises.

Analytical verification of the stability conditions for such Markov models is rather difficult. However, statistical methods of checking for instability can be used.

Case: G/G/c-type Supercomputer Model

- c homogeneus servers, FCFS;
- interarrival times $\tau_i, i \geqslant 1$ (input rate λ);
- service times S_i;
- client i requires N_i servers *simultaneously* ($\{p_j = \mathrm{P}(N = j), 1 \leqslant j \leqslant c\}$);
- *idle servers & non-empty queue*.





Earlier Works

Despite simplicity, explicit stability condition is known only for some special cases, e.g. two-server case with Poisson[1] or MAP arrivals[2], exponential[3] distribution of S_i, which does not hold for general distribution[4], and it remains an open problem[5].

---

[1][D. Filippopoulos, H. Karatza. An M/M/2 parallel system model with pure space sharing among rigid jobs. MCM, 2007, DOI: 10.1016/j.mcm.2006.06.007]

[2][S.R. Chakravarthy, H.D. Karatza. Two-server parallel system with pure space sharing and Markovian arrivals. COR, 2013, DOI: 10.1016/j.cor.2012.08.002]

[3][Rumyantsev, A., Morozov, E.: Stability criterion of a multiserver model with simultaneous service. AOR, 2017, DOI: 10.1007/s10479-015-1917-2]

[4][L. Afanaseva et al., Stability Analysis of a Multi-server Model with Simultaneous Service and a Regenerative Input Flow, MCAP, 2019, DOI: 10/ggnsgj]

[5][M. Harchol-Balter, Open problems in queueing theory inspired by datacenter computing, QS, 2021, DOI: 10.1007/s11134-020-09684-6]

Stability: M/M/c-type Supercomputer

Continuous-time QBD $\{X(t), Y(t)\}_{t\geqslant 0}$,

level $X(t)$ — number of customers in the system at t,

state $Y(t) = (m_1(t), \ldots, m_c(t)) \in \{1, \ldots, c\}^c$,
$m_i(t)$ – number of servers required by i-th oldest customer.

Stability criterion of M/M/c-type model [3]

QBD process is positive recurrent iff

$$C\lambda \mathrm{E}S < 1, \tag{1}$$

where

$$C = \sum_{i=1}^{c} \frac{1}{i} \sum_{j=i}^{c} p_j^{*i} \sum_{t=c-j+1}^{c} p_t, \tag{2}$$

and p^{*i} is a discrete convolution of the vector p with itself.

Stability: M/PH/2-type Supercomputer

It was shown in [4] that (1) fails for generalized Erlang distribution

$$S =_d \xi_1 + \xi_2,$$

where $\xi_i, i = 1, 2$ is independent exponential distribution with rate μ_i. However, explicit analysis of stability conditions of M/G/c-type model is not available yet.

Question to address

What can be done for a general (Markovian) model to statistically prove (in)stability of such a model?

Formulation of the problem

Consider a discrete-time Markov chain X^λ depending on the λ parameter:

- Let X_τ^λ be it state at the moment of time $\tau \in \mathbb{Z}_+$.
- $\lambda \in \mathcal{L}$, where $\mathcal{L}$ is the set of possible λ parameters.
- $\overline{\mathcal{L}} \in \mathcal{L}$ is a set of parameters at which the Markov chain becomes unstable.
- $\mathcal{B}(\lambda)$ is some neighborhood of λ.
- $f(X_\tau^\lambda)$ is a Lyapunov function that maps the Markov chain states to the set of real numbers.

We need to find the $\lambda^* \in \overline{\mathcal{L}}$ at which the Markov chain X^{λ^*} becomes unstable.

If λ is the *input rate* of a queueing system, we may obtain stability boundary of the model.

Statistical Instability Analysis Algorithm

[M. Mandjes, B. Patch, N. S. Walton. Detecting Markov Chain Instability: A Monte Carlo Approach. Stochastic Systems 7 (2), 2017, 289-314 DOI: 10.1287/stsy.2017.0003]

Initialization: Let $k = 1, T_0 = 0, \lambda_0 \sim U(\mathcal{L}), Y_0 = X_0^{\lambda_0}$

1. $\tau = aY_{k-1} + b, T_k = T_{k-1} + \tau, \gamma \sim U(\mathcal{B}(\lambda_{k-1})), \lambda = \lambda_{k-1}$
2. $y = X_\tau^\gamma | X_0^\gamma = Y_{k-1}$
3. $x = X_\tau^\lambda | X_0^\lambda = Y_{k-1}$
4. $(Y_k, \lambda_k) = \begin{cases} (y, \gamma) & \text{with probability } e^{\eta \min(f(y)-f(x),0)} \\ (x, \lambda) & \text{otherwise} \end{cases}$

Stop condition: $f(Y_k) > q_k^{(\alpha)}$ **OR** $k > K$.

Estimation of $q_k^{(\alpha)}$

$q_k^{(\alpha)}$ is the upper quantile of the alpha level for the process:

$$W_k = W_{k-1} + Z(W_{k-1}), \quad W_0 = 0.$$

where $Z(w)$ stochastically majorities the increment $f(Y_k)$ at the kth step. The following distribution is used to build the stochastic recursion for $\{W_k\}_{k\geq 0}$:

$$P(Z(w) \geqslant z) = \begin{cases} \min\left[1, e^{-\frac{(z-\alpha_1(w))^2}{2\alpha_2(w)}} + n(w)e^{-\frac{(z-\alpha_3(w))^2}{2\alpha_4(w)}}\right], & \text{if } z > 0, \\ 1, & \text{otherwise,} \end{cases}$$

where:

$$\alpha_1(w) = \sigma\phi - \sigma n(w)\delta, \quad \alpha_2(w) = (\phi+\delta)^2\sigma^2 n(w),$$
$$\alpha_3(w) = \sigma\phi - w + \kappa, \quad \alpha_4(w) = \phi^2\sigma^2 n(w).$$

Estimation of $q_k^{(\alpha)}$

In these formulas:

- ϕ is an upper bound for the increment of the Lyapunov function;
- $\tau(w)$ is a simulation time;
- $n(w)$ is the smallest integer such that $\sigma n(w) \geqslant \tau(w)$;
- κ is a lower bound for the modulus of the initial state of the chain;
- σ is a lower bound for the number of steps;
- δ is a negative drift that a process must demonstrate to be stationary.

In practice, we used binary search and summation to find $q_k^{(\alpha)}$.

Setting up a numerical experiment

Consider M/PH/2-type and M/M/2-type systems with the following set of parameters:

- Distribution for the number of servers requested by a customer: $p_1 = p_2 = 0.5$.
- Service rate of a customer in M/M/2-type system: $\mu = \frac{1}{3}$.
- Service time parameters for generalized Erlang in M/PH/2-type system: $\mu_1 = 1$, $\mu_2 = 0.5$.

Consider the stability boundary for the arrival rate given in [4] for these systems:

- M/PH/2-type: $\lambda_{max} = \frac{12}{31} \approx 0.3870967742$
- M/M/2-type: $\lambda_{max} = \frac{8}{21} \approx 0.3809523810$

We will search for λ^* in the set $\mathcal{L} = (0, \frac{12}{31}) \approx (0, 0.3870967742)$.

Setting Up a Numerical Experiment: Required Parameters

For the models under consideration, the required parameters take the following values:

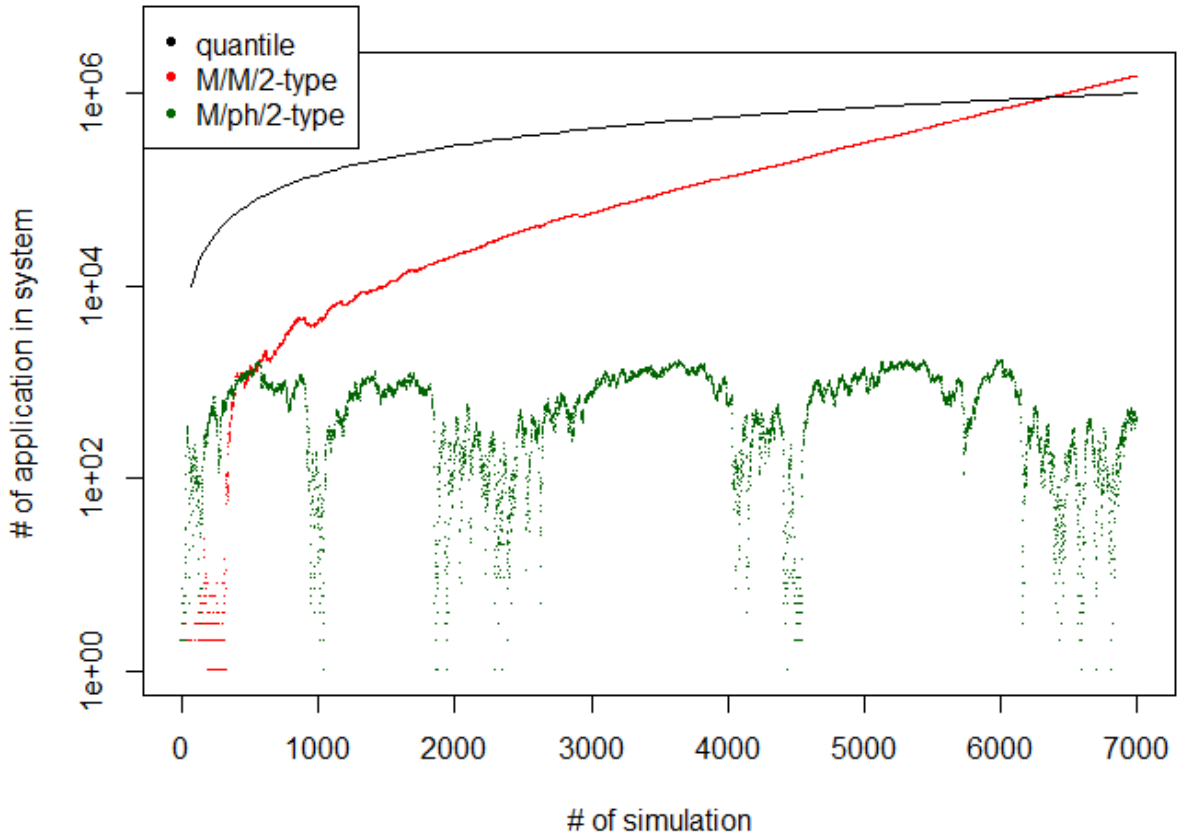
Lyapunov function $f(x)$ – number of customers in the system;

Parameters $\phi = 1, \kappa = 0, \sigma = 1, \delta = 0.01, \alpha = 0.05, K = 7000,$

Neighborhood $\mathcal{B}(\lambda) = (max(0, \lambda - 0.05), min(\frac{250}{651}, \lambda + 0.05))$

To demonstrate the operation of the algorithm, we removed the comparison with quantile from the stop condition.

Preliminary Numerical Results



Observe existence of $\lambda^* \in \mathcal{L}$ such that M/M/2-type model is unstable.

Future work

- Improve this algorithm to increase the rate of convergence to the unstable region $\overline{\mathcal{L}}$.
- Modification of this algorithm to check for instability for a given λ.
- Simplify the use of the algorithm by automatically determining required additional parameters of the Markov chain.

Thank you for attention!

Astafiev Sergey
Institute of Applied Mathematical Research
Karelian Research Centre RAS
seryymail@mail.ru

ORCID: orcid.org/0000-0002-4486-798X

ResearchGate: https://www.researchgate.net/profile/Sergey-Astafiev