Controllable Queueing Networks for Business Process Performance Analysis

Leyla Abuzyarova, Irina Kochetkova

RUDN University, Moscow, Russian Federation, kochetkova-ia@rudn.ru

Organizational business processes are sequences of activities – successive processing stages carried out by employees of different departments – and this structure is naturally represented as a queueing network: each processing stage or department corresponds to a node of the network, and the request's path through the organization is described by a routing matrix between nodes. Queueing networks of the BCMP type (Baskett–Chandy–Muntz–Palacios) generalize this representation by allowing multiple request classes and different service disciplines within a single model, which makes it possible to describe heterogeneous request types and parallel activities executed concurrently within one coherent analytical framework, rather than building a separate model for every process variant.

A key modeling decision is the level of detail chosen for request classes, since this choice governs both the fidelity and the computational complexity of the resulting model. At the most detailed level, a class corresponds to a specific state of a request within the process – the outcome of the most recently completed stage – which yields a highly accurate description but produces a very large number of classes and hence a high-dimensional state vector. At a coarser level, classes are merged according to similarity of routing behavior or independence of service times from the request's prior history, which substantially reduces the dimensionality of the state space at some cost in descriptive precision. An intermediate level of aggregation, tied to the number of distinct routes that occur within the process, offers a practical compromise between accuracy and tractability, and the choice among these levels ultimately determines both the accuracy of the model and the computational cost of its analysis.

To move from a purely descriptive queueing network to a tool for actively managing business processes, we consider four control instruments that act on the parameters of the queueing network. First, the routing matrix $P=\left\{P_{ij}^{(k)}\right\}$, where $P_{ij}^{(k)}$ is the probability that a class-k request moves from node i to node j upon completing service; controlling routing allows workload to be redistributed among departments and the order of process stages to be changed dynamically. Second, admission control, which accepts or rejects a new request depending on the current state of the network, n, allowing, allowing the system to prevent overload by limiting the number of requests admitted when departments are already congested. The state of the controlled queueing network is described by a vector $n=(n_1,\ldots,n_M)$, where each component $n_m=(n_{m1},\ldots,n_{mK})$ gives the number of class-k requests present at node m, for M nodes (departments) and K request classes. Under Poisson arrival processes and exponentially distributed service times, this controlled queueing network reduces exactly to a Markov Decision Process, for which efficient algorithms exist to compute the optimal control policy.