**Appendix**

This appendix presents the generalization of the state-to-state transitions of the hypercube model (Beojone et al., 2020). Equation A.1 shows the major equation of this model:

|  |  |  |
| --- | --- | --- |
|  | $$P\_{r}\left[L\left(r\right)\sum\_{j}^{}\sum\_{l}^{}I\left(R\left(r,jl\right)\ne ∅\right)λ\_{jl}+\left(1-L\left(r\right)\right)λ+\sum\_{t}^{}S\left(r,k\right)μ\_{t}\right]=\sum\_{q,t:D\left(q,r,t\right)=1}^{}P\_{q}S\left(q,t\right)μ\_{t}+\sum\_{q,l:Q\left(q,r,l\right)=1}^{}P\_{q}\sum\_{t}^{}S\left(q,t\right)μ\_{t}+\sum\_{q,t:D\left(r,q,t\right)=1}^{}P\_{q}\sum\_{jl\in A(q,t)}^{}λ\_{jl}+P\_{q}\sum\_{q,l:Q\left(r,q,l\right)=1}^{}λ\_{jl}$$ | (A.1) |

In Equation A.1, q and r are the states belonging to the set of states M. Pr and Pq are the probabilities of the steady states r and q, respectively, while t and jl are defined as the groups of servers and subatoms of the system, respectively. I($\*$) is the indicator of a function (1, if $I(\*)$ is true, otherwise 0). L(r) indicates whether state r is likely to miss a call. R (r, jl) represents the set of servers that are still available to serve subatom jl in the state r. $S\left(r,t\right)$ shows the number of servers in the group t occupied in the state r. The arrival rate from subatom jl is λjl, while λ is the arrival rate for all subatoms. The service rate of each group t is µt. The service rate is obtained considering the server's journey time and on-site service time. In addition, as server locations vary from the original scenario, there is a need to calibrate the service times *μ*-1 (Larson & Odoni, 2007). The process is similar to the calculation of the initial response times, using the average service time and the journey time of the servers in the original scenario. Here, the service time is the same, but the journey time considered is that of the last solution of the hypercube model for the given configuration. The process is repeated until the changes in the service times of the servers are below a tolerance of 0.1. $A\left(q,t\right)$ is the set of subatoms in which the server group t is preferable in the state q (based on the dispatch system preference list). $D(q,r,t)$ is another function with definition based on Equation A.2:

|  |  |  |
| --- | --- | --- |
|  | $$D\left(q,r,t\right)=\left\{\begin{array}{c}1, \& if H\left(q,r\right)=1, S\left(q,t\right)=S\left(r,t\right)+1\\0, \&otherwise\end{array}\right.$$ | (A.2) |

Where *H (q, r)* is the minimum number of transitions between states *q* and *r*. $Q\left(q,r,l\right)$ is another function defined in Equation A.3 and is related to the transitions between the queue states:

|  |  |  |
| --- | --- | --- |
|  | $$D\left(q,r,t\right)==\left\{\begin{array}{c}1, \& if H\left(q,r\right)=1, S\left(q,k\right)=S\left(r,t\right) ∀k, W\left(q,r,l\right)=1\\0, \&otherwise\end{array}\right.$$ | (A.3) |

In this equation, $W\left(q,r,l\right)$ represents the function that indicates the number of classes $l$ under service, in which state q has more states than in r.