saturation; this difference is a result of the LQNS solution strategy and may be due to LQNS ignoring service time variance in all its multiserver approximations (the coefficients of variation of the service time of the Router and the FrontEnd were estimated at 2.97 and 1.78 respectively).

6. Related Work

Related work on FIFO multiservers with closed workloads begins with separable models with exact solutions by Mean Value Analysis (MVA) [10]. The high complexity of MVA, due to the need for marginal distributions of queue length, has led to AMVA approximations to multiservers, such as Linearizer [4], improved in SCAT [2] and in [1] by using an ad hoc distribution shape based on the mean. QD-MVA [3] takes a different approach and reformulates the solution as an optimization problem.

For multiclass queues de Souza e Silva and Muntz approximated the queue state by a multinomial distribution [13] based on the mean values, to get the class waiting times, and combined these with AMVA methods. The single class version of the multinomial is the binomial distribution used in AB. Conway [5] adapted their method to AMVA, as it is used in LQNS. Rolia and Sevcik [11] described a much simpler AMVA calculation (RF in this work) that approximates the probability that all servers are busy, by the product of the individual probabilities (effectively assuming the servers are independently busy). Franks made a small improvement described in [7].

There are many other works on multiserver approximations that address different models, such as open arrivals, general service times and multiple classes.

7. Conclusions

The AB approximation has been shown to be a useful addition to numerical methods for multiserver queue waiting. It is a little more accurate, and a little faster than the best alternative which is RF, and is more robust in terms of convergence. In defense of RF it should be mentioned that using a smaller relaxation coefficient a in Eq (6) improves its convergence, at the expense of longer computation times. In other experiments not reported here we have been forced to make a as small as 0.02 to obtain convergence.

The example in Section 4 shows that the faster queue solution by AB translates into faster overall solutions.

The in-depth investigation of AB, SS and RF show that SS is surprisingly good given its simplicity. It has substantial errors for many cases but its maximum errors are the least of the three, to it has a safety factor.

The results are reported against the traffic ratio T, and the relationship of T to the server utilization is of interest. Utilization for a given m is close to T or less than T; it increases with T and reaches saturation between T=1 and T=2, in most cases, depending on the balance of m, N and Z. For larger T the server is saturated but the error behaviour still depends on T, indicating that it varies according to the balance of N and Z. Larger T gives behaviour more like an open queue.

The extension of AB and SS to multiple classes will be reported elsewhere.

ACKNOWLEDGMENTS

This research was funded by the Natural Sciences and Engineering Research Council of Canada, grant 06274-2016.

REFERENCES

[1] I.F. Akyildiz, G. Bolch, "Mean Value Analysis Approximation for Multiple Server Queueing Networks", Performance Evaluation, v 8, pp 77-91, 1988.

[2] G. Bolch, S. Greiner, H. Meer, K. Trivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications, Second Edition, April 2006.

[3] G. Casale, J. Perez, W. Wang, "QD-AMVA: evaluating systems with queue-dependent service requirements", Performance Evaluation, v 91, pp 80-98, 2015.

[4] K.M. Chandy, Doug Neuse, "Linearizer: A heuristic algorithm for queueing network models of computing systems", Comm. of the ACM, v 25 n 2, pp 126-134, 1982.

[5] A.E. Conway, "Fast approximate solution of queueing networks with multi-server chain-dependent FCFS queues", in *Modeling Techniques and Tools for Computer Performance Evaluation*, pp 385-396, Plenum, New York, 1989.

[6] G. Franks et al, "Layered Queueing Network Solver and Simulator User Manual", http://www.sce.carleton.ca/rads/lqns/LQNSUserMan-jan13.pdf.

[7] G. Franks, Performance analysis of distributed server systems, PhD thesis, Carleton University, 1999.

[8] A.I. Gias, G. Casale, M. Woodside, "ATOM: Model-Driven Autoscaling for Microservices", ICDCS 2019.

[9] E. Lazowska, J. Zahorjan, G.S. Graham, K. Sevcik, Quantitative System Performance, Wiley, 1984.

[10] M. Reiser, S.S Lavenberg, "Mean-value analysis of closed multichain queueing networks", J. of ACM, v27 n 2, pp 312–322, 1980.

[11] J.A. Rolia, K.C. Sevcik, The Method of Layers, IEEE Trans Software Engineering, v 21 pp 689 – 700, 2015.

[12] P.J. Schweitzer, "Approximate analysis of multiclass closed networks of queues", Proc. Int. Conf. on Stochastic Control and Optimization, pp 25-29, Amsterdam, 1979.

[13] E. de Souza e Silva, R.R. Muntz, "Approximate solutions for a class of non-product form queueing network models", Performance Evaluation, v 7, pp 221-242, 1987.

[14] Q. Zhang, Q. Zhu, M.F. Zhani, R. Boutaba, J.L. Hellerstein, "Dynamic Service Placement in Geographically Distributed Clouds", IEEE J. on Selected Areas in Communications, v 31, n 10, Oct. 2013.

[15] Q. Zhang, Y. Xiao, F. Liu, J.C.S. Lui, J. Guo, T. Wang, "Joint Optimization of Chain Placement and Request Scheduling for Network Function Virtualization", Int. Conf. Distrib. Computing Systems (ICDCS), pp 731-741, 2017.