Processing Networks: Managing Cloud Computing

Keywords: Network Science, Cloud Computing, Capacity, Pricing, Networks of Queues

Extended Abstract

An important application of network analysis is the management of Cloud computing networks. Cloud computing has become a major way of delivering computing services, and the dramatic increase in the use and computational requirements of AI models and applications underscores the importance of trading off price, capacity and delay. This paper models the cloud computing system as a network of processors (for broad reviews, see [1,2,3]), builds on the modeling approaches developed in [4,5] to solve simpler problems, discusses the drivers of the solution and presents a simulation approach used to teach students how to perform the analysis and to allow practitioners to apply it.

The cloud computing provider is modeled as a profit-maximizing firm (a different version of the model, which follows a similar approach, addresses the provision of private cloud services within an enterprise). The provider faces a downward-sloping demand curve, and the cloud computing system is modeled as a network of processors. The workload is modeled as a random stream of units of work arriving into the system. The valuations and processing requirements of these units of work are independent and identically distributed, so they vary randomly across units of work. Users pay a price per unit of work and incur costly delays. The delay cost structure may be additive (delay cost independent of value) or multiplicative (delay cost proportional to value) [5].

I show that when the capacity of the cloud computing system is fixed, the optimal price for the additive delay-cost case comprises three components:
1. Direct cost per unit of work, which includes the cost of electricity, GHG emission credits, and any other direct costs of operating the system;
2. A delay externality cost, which is the marginal increase in the system's delay cost due to the addition of one unit of work to the system; and
3. An endogenously-determined profit margin charged by the cloud computing provider per unit of work.

The price is derived endogenously. I show that when the network is a Jackson Network [1] and capacity expansion is achieved by continuously increasing the capacity of each processor at a cost, the optimal price gets simplified to the sum of the marginal capacity cost, the direct cost, and the profit margin. However, most of today's cloud computing networks are more complex and capacity decisions involve both the number of processors (GPUs for AI-focused systems) and the capacity of each processor. In addition, capacities are discrete, so the solution has to be derived numerically or using simulation. Nevertheless, the analytic solution provides an initiation point for the numerical procedure and provides insights on the characteristics of the solution.

The multiplicative delay cost structure has become increasingly important as simulation modeling and AI are dominating cloud computing workflows. Consider, for example, the

development of a complex product such as an aircraft or a car. Modern development processes use digital simulations to study the effects of multiple design choices under thousands of scenarios. The simulations may use multiple GPUs to run computational fluid dynamics simulations to model and predict air flows and to repeatedly adapt the designs based on simulation results. Under these conditions, computational capacity becomes a bottleneck which drives the product development time. The higher the value of the product, the higher the loss due to bringing it to market later, which creates a multiplicative delay cost model. Similar considerations apply to the training time of Large Language Models, where time-based competition drives success or failure. A multiplicative delay cost structure is also obtained whenever we apply discounting to delayed cash flows.

Under the multiplicative delay cost structure, maximizing expected profit leads to straightforward but tedious equations that are solved numerically even for simple networks. Then, the best approach in my view is to perform a series of numerical simulations of the cloud computing system and to optimize it system by varying its key parameters.

I have built a simulation system which I use to teach these concepts to MBA students as well as to perform the optimization. In the educational setting, students configure a cloud computing system which is designed to perform computational fluid dynamics analyses using GPUs and to optimize three decision variables: the number of GPUs, the MFLOPS performance of each GPU (i.e., the number of computations it can perform per unit of time, which is a variable taking on discrete values), and the price for each unit of work. One of the key insights is that for complex systems with expensive GPUs, statistical economies of scale imply that they should be operated at a high utilization. As a result, a good set of initial values for the numerical analysis is the solution obtained under a fluid model (with no uncertainty), which can be derived analytically. Once the analytic solution is obtained, the simulation is used to refine the solution and incorporate the elaborate delay cost structure. This approach results in fast convergence unless there are multiple near-optimal solutions, in which case qualitative considerations are brought in to determine the final solution (for example, a high-quality vs. low-price business strategy). The system shows nicely how one may incorporate multiple, complementary approached to solve complex network optimization problems.

References

[1] Mor Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, Cambridge University Press, 2013.
[2] E. J. Ghomi, A. M. Rahmani and N. N. Qader, “Applying queue theory for modeling of cloud computing: “A systematic review,” *Concurrency and Computation*, Vol. 31(17), 2019.
[3] I. Bambrik, “A Survey on Cloud Computing Simulation and Modeling,” *SN Computer Science*, Vol. 1 (249), 2020.
[4] H. Mendelson, “Pricing computer services: Queuing effects,” *Communications of the ACM*, Vol. 28(3), 1985, pp. 312-321.
[5] P. Afeche and H. Mendelson, “Pricing and Priority Auctions in Queueing Systems with a Generalized Delay Cost Structure,” *Management Science*, Vol, 50 (7), pp. 855-1000, 2004.