

CAPACITY ALLOCATION IN FEEDFORWARD QUEUEING NETWORKS

Ton Dieker

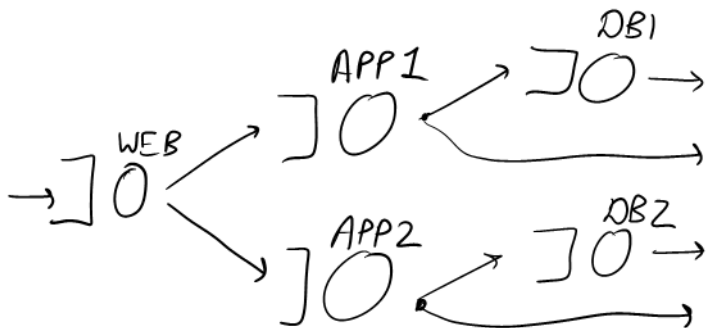
ISyE, Georgia Institute of Technology

June 10, 2010

joint work with Soumyadip Ghosh and Mark Squillante (IBM Research)

- Service capacity allocation ('staffing') in queueing networks aims to manage the trade-off between costs and delay
- Applications include workforce management, call centers, and health care delivery
- In practice, there is essentially exclusive use of (computationally expensive) black-box simulation for the analysis and optimization of business processes
- Our goal is to develop generic methods which exploit the structure of the network

A NETWORK OF QUEUES



- 1 Single server queues, service speed β_i
- 2 Service speed at the i -th queue costs c_i per unit
- 3 The total budget is K
- 4 Objective: minimize the 'expected' long-run weighted queue length

- The capacity-allocation problem has been solved completely for exponential interarrival and service times (PhD thesis of Kleinrock).
- It is unrealistic to hope for an 'explicit' solution for general service times and arrival times
- Existing approximation approaches rely on some kind of decomposition result (Whitt's QNA, ...)
 - but there is interaction between the stations in the network!
- We developed a simulation-based algorithm which *does* take this interplay into account
- During this talk, we discuss the theoretical foundation for this algorithm

Suppose:

- The external arrival rate is 0.1
- The budget is $K = 5$
- Service speed costs $c_i = 1$
- All distributions Coxian

SOME NUMERICAL RESULTS

Suppose:

- The external arrival rate is 0.1
- The budget is $K = 5$
- Service speed costs $c_i = 1$
- All distributions Coxian

C_A	C_W	C_{A1}	C_{DB1}	C_{A2}	C_{DB2}	# iter.	obj. rel. err.
0.75	2.75	1.5	1.5	3	0.75	34	0.61%
0.75	4.75	1.5	0.75	3	0.75	32	0.37%
2.25	2.75	3	1.5	1.5	0.75	22	0.13%
2.25	4.75	3	1.5	1.5	1.75	25	0.04%

Consider two queues in series, with external arrivals only at the first queue (arrival rate α).

Goal:

$$\begin{aligned} & \text{minimize } T(\boldsymbol{\beta}) := \mathbf{E}(Q_1^\beta(\infty) + Q_2^\beta(\infty)) \\ & \text{over } \{\boldsymbol{\beta} \in (\alpha, \infty)^2 : c_1\beta_1 + c_2\beta_2 \leq K\}. \end{aligned}$$

One expects that the minimum is attained on the face $\{c_1\beta_1 + c_2\beta_2 = K\}$.

We define $\tau_1(\beta_1)$ and $\tau_2(\beta_1, \beta_2)$ so that

$$\begin{aligned} \mathbf{E}(Q_1^\beta(\infty)) &= \frac{\tau_1(\beta_1)}{\beta_1 - \alpha} \\ \mathbf{E}(Q_2^\beta(\infty)) &= \frac{\tau_2(\beta_1, \beta_2)}{\beta_2 - \alpha}. \end{aligned}$$

AN APPROXIMATION (2)

For given $\tau_1, \tau_2 > 0$, the problem

$$\begin{aligned} & \text{minimize } \frac{\tau_1}{\beta_1 - \alpha} + \frac{\tau_2}{\beta_2 - \alpha} \\ & \text{over } \{\boldsymbol{\beta} \in (\alpha, \infty)^2 : c_1\beta_1 + c_2\beta_2 \leq K\}. \end{aligned}$$

is solved by $\boldsymbol{\beta} = \rho(\tau_1, \tau_2)$, where

$$\rho_i(\tau_1, \tau_2) = \alpha + \frac{K - (c_1 + c_2)\alpha}{c_i} \frac{\sqrt{c_i\tau_i}}{\sqrt{c_1\tau_1} + \sqrt{c_2\tau_2}}.$$

The optimal capacity allocation rule is approximated by an iteration of the ρ map:

$$\boldsymbol{\beta}^{(n+1)} = \phi(\boldsymbol{\beta}^{(n)}),$$

where $\phi(\boldsymbol{\beta}) = \rho(\tau_1(\beta_1), \tau_2(\beta_1, \beta_2))$.

The optimal capacity allocation rule is approximated by an iteration of the ρ map:

$$\boldsymbol{\beta}^{(n+1)} = \phi(\boldsymbol{\beta}^{(n)}),$$

where $\phi(\boldsymbol{\beta}) = \rho(\tau_1(\beta_1), \tau_2(\beta_1, \beta_2))$.

Questions:

- Does this fixed point exist? (Not equal to the 'true' optimum!) If so, is it unique?
- Does the algorithm converge?

The optimal capacity allocation rule is approximated by an iteration of the ρ map:

$$\boldsymbol{\beta}^{(n+1)} = \phi(\boldsymbol{\beta}^{(n)}),$$

where $\phi(\boldsymbol{\beta}) = \rho(\tau_1(\beta_1), \tau_2(\beta_1, \beta_2))$.

Questions:

- Does this fixed point exist? (Not equal to the 'true' optimum!) If so, is it unique?
- Does the algorithm converge?

We study these questions in the heavy traffic setting for feedforward networks. We exploit properties of (reflected) Brownian motion.

LEMMA

There exists a fixed point of the above equation.

LEMMA

There exists a fixed point of the above equation.

The function ϕ depends on the objective function T .

- In the 'relevant' area we always seems to have $|\phi'| < 1...$
- but we were unable to prove this.
- We were not even able to prove monotonicity of ϕ .

LEMMA

There exists a fixed point of the above equation.

The function ϕ depends on the objective function T .

- In the 'relevant' area we always seems to have $|\phi'| < 1$...
- but we were unable to prove this.
- We were not even able to prove monotonicity of ϕ .

After deriving some properties of Brownian feedforward networks, we were able to prove that:

LEMMA

The fixed point is unique.

In the Brownian model, we have $\tau_1(\beta_1) = \tau_1$ for some $\tau_1 > 0$.

Define

$$x_2(\beta) = \frac{\beta_2 - \alpha}{\beta_1 - \alpha}.$$

Rewriting $\beta = \phi(\beta)$ yields

$$(\beta_1 - \alpha)T(\beta) = \tau_1 + \tau_1 \frac{c_2}{c_1} x_2(\beta).$$

Observations:

- the left-hand side is a nondecreasing function of $x_2(\beta)$ by the theory of Brownian networks
- the right-hand side is an increasing function of $x_2(\beta)$

The story is a little messier here, but still:

- We make a small modification of the algorithm, and then the fixed-point algorithm always converges
 - Oscillations can be ruled out with bisection
 - The proof relies on a version of the implicit function theorem

- We have introduced a method for (approximate) capacity allocation in queueing networks
- Our methodology compares favorably with black-box stochastic approximation algorithms
 - We have implemented state-of-the-art stochastic approximation algorithms
- We don't (always) get the optimal value, but the fixed point could serve as a starting point for a gradient-based algorithm

AN ADVERTISEMENT