Some Dynamical Routing Algorithms in Large Systems

Nikita D. Vvedenskaya
Institute of Information Transmission Problems
Russian Academy of sciences
Moscow 101447, GSP-4, 19 Bolshoi Karetnyi
Russia

November 3, 1997

[summary by Philippe Robert]

Abstract

We consider a system with N servers and messages arriving according to a Poisson process. The service time of a message is exponentially distributed. Two strategies to process the messages are compared. In the first strategy, an arriving message is sent randomly to one of the servers. In the second strategy, for each message two servers are selected randomly, and the message is directed to the least busy one. The queue length distribution is investigated as N tends to infinity.

1. Introduction

We assume that we have a set of N servers with N queues, these servers process a stream of jobs arriving to that system. Once an arriving job has been allocated to the queue of one of the servers, it cannot switch to another queue. We assume that the arrivals are Poisson with rate λN and the processing times are exponentially distributed with rate 1. Inside the queues, the service discipline is First In First Out. We shall assume that $\lambda < 1$, with this condition the system will not explode with the service disciplines which we consider.

The simplest strategy, S_{ind} say, to allocate the jobs is to distribute them at random to the servers. In this case the system is equivalent to a set of N independent queues with arrival rate λ and service rate 1. It is well known that, at equilibrium, in each queue the number of jobs has geometric distribution with parameter λ, in particular the probability that there is at least k jobs in a queue is λ^k.

A more efficient strategy S_{sh}, consists in choosing the shortest queue at the arrival of the job. Unfortunately, this kind of discipline is very difficult to analyze, in particular it would be desirable to compare it quantitatively with our first discipline. The exact analysis has been carried out by Flatto and Mac Kean [2] in the case $N = 2$, using uniformisation techniques. This approach does not seem to extend to a higher dimension.

The object of this talk is to analyze an intermediate discipline, S_{int}, for which it is possible to derive some quantitative results. An arriving job takes two servers at random and chooses the one with the shortest queue. Notice that for S_{sh}, there is a tight correlation between the queues, all of them are considered to allocate an arriving job. Here, only two of them determine the destination of the job. For a fixed N, a quantitative analysis remains difficult; however, as N goes to infinity, a given queue depends weakly of any other queue. As we shall see, asymptotically the queues behave

independently. This phenomenon allows to write down one-dimensional equations. This approach is called the mean field method in statistical physics.

2. The Differential Equations

For this model, it is convenient to describe the queues in the following way: $u_{k,N}(t)$ will denote the fraction of the N queues which have at least k jobs at time t. Clearly

$$1 = u_{0,N}(t) \geq u_{1,N}(t) \geq \cdots u_{k,N}(t) \geq u_{k+1,N}(t) \geq \ldots,$$

and $0 \leq u_{k,N}(t) \leq 1$, the vector $U_N(t) = (u_{k,N}(t))$ belongs to the state space $\mathcal{S} = [0,1]^{\mathbb{N}}$ which is compact for the point-wise convergence. It is easily seen that $(U_N(t))$ is a Markov process since the vector of the number of jobs in each queue is a Markov process and the order of the queues does not matter. If F is a measurable functional on $\mathcal{S}$, then $F(U_N(t))$ satisfies the stochastic differential equation,

$$(1)\quad dF(U_N)(t) = \sum_{k=1}^{+\infty} N(u_{k,N}(t) - u_{k+1,N}(t)) \left(F(U_N(t) - \frac{e_k}{N}) - F(U_N(t))\right) dt$$
$$+ \sum_{k=1}^{+\infty} \lambda N(u_{k-1,N}^2(t) - u_{k,N}^2(t)) \left(F(U_N(t) + \frac{e_k}{N}) - F(U_N(t))\right) dt + dM_N(t),$$

where e_k is the vector $(1_{\{i=k\}})$. The first term in the right hand side is the contribution of departures, $N(u_k(t) - u_{k+1}(t))$ is the number of queues with k jobs hence the rate at which $U_N(t) \to U_N(t) - \frac{e_k}{N}$. The second term concerns the arrivals, $(u_{k-1}^2(t) - u_k^2(t))$ is the probability that two queues chosen at random have a size $\geq k$. The last term $M_N(t)$ is a martingale (which depends on F), i.e. roughly speaking, a stochastic perturbation; in particular $E(M_N(t)) = E(M_N(0))$ for all $t \geq 0$. It is easily seen using standard results concerning Poisson processes that

$$E(M_N^2(t)) \leq \frac{Kt}{N},$$

this simply means that the stochastic perturbation is vanishing as $N \to +\infty$. Taking $F(U) = u_k$, this suggests that the equation (1) becomes a deterministic differential equation,

$$(2)\qquad \frac{du_k(t)}{dt} = -(u_k(t) - u_{k+1}(t)) + \lambda(u_{k-1}^2(t) - u_k^2(t)),\ \ k \geq 1.$$

If $(u_k(0)) \in L_1$, that is $\sum_{k=0}^{+\infty} u_k(0) < +\infty$, using a truncation procedure, it can be proved that (2) has a unique solution. So, as $N \to +\infty$, the $(u_{k,N}(t))$ should converge to a solution of this equation.

Rigorously, Doob's inequality [1] tells us that

$$P\left(\sup_{0 \leq s \leq t} |M_N(s)| > a\right) \leq \frac{Kt}{a^2 N},$$

hence

$$(3)\quad P\left(\sup_{0 \leq s \leq t} \left| u_{k,N}(s) - u_{k,N}(0) + \right.\right.$$
$$\left.\left. \int_0^s (u_{k,N}(x) - u_{k+1,N}(x)) - \lambda(u_{k-1,N}^2(x) - u_{k,N}^2(x))dx \right| > a\right) \leq \frac{Kt}{a^2 N}.$$

It is easy to check that if $(u_{k,N}(0))$ converges to $(u_k(0))$ as $N \to +\infty$, then the sequence of processes

$$(u_{k,N}(s))_{0\leq s\leq t},\ N \in \mathbb{N}$$

is relatively compact. The identity (3) gives that any limit $(u_k(s))_{0\leq s\leq t}$ (in distribution) satisfies the differential equation (2) with probability one. We deduce that if $(u_k(0)) \in L_1$ then $(u_{k,N}(s))$ converges in distribution to the unique solution of (2).

3. The Convergence of the Invariant Measures

Up to now, we have only looked at the transient behavior of the queues. That is, for a fixed t, we proved that the state at time t converges in distribution. For $N \in \mathbb{N}$, the model of size N has an equilibrium distribution π_N; the (delicate) question is: as $N \to +\infty$ does the sequence (π_N) converge to a stable point of (2) ?

If $U(0) \in L_1$, then it is easily seen that (λ^{2^k}) is the unique stable point of (2). Notice that the π_N are probability measures on a compact space, thus the sequence is relatively compact. If one can show that every limit point is a stable point of (2) which belongs to L_1, then necessarily the sequence converges to (λ^{2^k}). This is done using the following estimation: for any continuous function $f : L_1 \to \mathbb{R}$,

$$\lim_{N\to+\infty} \sup_{v\in L_1} \|E_v(f(U_N(t))) - f(u_v(t))\| = 0,$$

where E_v denotes the expectation with the initial condition $U_N(0) = v$ and u_v is the solution of (2) with $u(0) = v$. This gives a kind of uniform convergence of the processes U_N.

4. Conclusion

For the strategy S_{int}, the queue length has a super exponential tail. We have seen that for S_{ind}, the tail was only exponential. It is remarkable that with a little improvement of S_{ind}, the tail distribution drops significantly. For the optimal discipline S_{sh}, asymptotically, it is easy to see that there will be an unbounded number of empty queues.

Bibliography

[1] Ethier (S.N.) and Kurtz (T.G.). – *Markov Processes, characterization and convergence.* – John Wiley & Sons Ltd, 1986, *Probability and mathematical statistics.*

[2] Flatto (L.) and McKean (H.P.). – Two queues in parallel. *Communications in Pure and Applied Mathematics*, vol. XXX, 1977, pp. 255–263.

[3] Vvedenskaya (N. D.), Dobrushin (R. L.), and Karpelevich (F. I.). – A queueing system with a choice of the shorter of two queues –an asymptotic approach. *Problemy Peredachi Informatsii*, vol. 32, n° 1, 1996, pp. 20–34.