3 Lecture 3: Queueing Networks and their Fluid and Diffusion Approximation

- Fluid and diffusion scales
- Queueing networks
- Fluid and diffusion approximation of Generalized Jackson networks.
- Reflected Brownian Motions
- Routing in single class queueing networks.
- Multi class queueing networks
- Stability of policies for MCQN via fluid
- Optimization of MCQN — MDP and the fluid model.
- Formulation of the fluid problem for a MCQN

3.1 Fluid and diffusion scales

As we saw, queues can only be studied under some very detailed assumptions. If we are willing to assume Poisson arrivals and exponetial service we get a very complete description of the queuing process, in which we can study almost every phenomenon of interest by analytic means.

However, much of what we see in M/M/1 simply does not describe what goes on in non M/M/1 systems. If we give up the assumption of exponential service time, then it needs to be replaced by a distribution function G. Fortunately, M/G/1 can also be analyzed in moderate detail, and quite a lot can be said about it which does not depend too much on the detailed form of G. In particular, for the M/G/1 system in steady state, we can calculate moments of queue length and waiting times in terms of the same moments plus one of the distribution G, and so we can evaluate performance of M/G/1 without too many details on G, sometimes $E(G)$ and c_G^2 suffice.

Again, behavior of M/G/1 does not describe what goes on in GI/G/1 systems. Unfortunately, to study almost any properties of GI/G/1 requires the exact inter-arrival and service distributions. So one can work with cetrain classes of distributions (e.g. phase type) and get detailed results, or one needs to simulate whatever one wants to know.

Nevertheless, if the system is almost always empty (light traffic), quite a lot can be said about it without specific dependence on the distributions of inter-arrival and service. Also, extremes and rare events can be studied through large deviation theory. In the opposite situation, if there are many customers in the system (heavy traffic) then one can use statistical averaging to say quite a lot about the performance of the system, without dependence on the details of the inter-arrival and service distributions. Since our manufacturing systems have plenty of material in them we shall study them via heavy traffic approximations.

Consider an arriving customer. If service is FIFO, he will be in the system until all the customers he found there leave, and he will then get his own service and leave. In light traffic he will often find 0 or 1 ahead of him, and he will leave 0 or 1 behind him. This whole episode will take about one or two service times, and during that time the queue will change only once or twice. The sojurn time of a customer is then comparable to the time that it takes the queue to change say from average to empty.

When the system is congested, i.e. most of the time it has many customers, the picture is different. A customer which arrives to find an average queue, say with L customers in it, will remain there for $L+1$ services, so he will have a long delay. While he is there, he will see the queue of customers ahead of him move down from L to 0. However, in this time of $L+1$ departures, there will also be many arrivals. If traffic intensity is close to 1, the number of arrivals will be of order L as well. Thus, when the customer leaves he will have a queue behind him which is comparable to what he found. For the system to go from average to empty will take more like L busy periods. Thus, in terms of traffic intensity, the time of a customer in the system is of order $1/(1-\rho)$, the time for the system to change say from average to empty is $1/(1-\rho)^2$. This is all very curious. Note for instance that an average busy period is $1/(1-\rho)$ long, which is much less then the length of a busy period in which the queue is allowed to grow to average queue length, Hence, the length of busy period and the max queue length within a busy period must be extremely variable, which indeed they are. The important observation here is that the time scale at which the system changes is much longer then the time scale of a single customer. We therefore have a time scale separation.

We study a system with many customers on two scales, fluid and diffusion. Let $Z(t)$ be some process. We let $\bar{Z}^n(t) = \frac{1}{n}Z(nt)$ be its fluid rescaling by n. This means that we measure time in units of n and we measure the state (counts of customers) in units of n. As $n \to \infty$ we shall look (in analogy to the SLLN) for $\frac{1}{n}Z(nt) \stackrel{a.s.}{\to} \bar{Z}(t)$, where $\bar{Z}(t)$ is the fluid limit.

At this scale as $n \to \infty$ the arrival process and the service process obey the FSLLN and have fluid limits λt and μt which means that they are deterministic. As we said, queueing is the result of variability, and so on a fluid scale, when input and output are not variable, there will be no real queueing behavior in the system. We may see the queue length grow linearly indefinitely ($\rho > 1$), or go to zero linearly and then stay at 0 ($\rho < 1$), or we may see it constant, ($\rho = 1$). For queueing networks we may observe piecewise linear behavior of queue lengths. This will capture changes in the queue on the fluid scale: The queue changes by n in a time of order n. The stochastic fluctuations of a queue in steady state are scaled down to be identically 0 and uninteresting.

The diffusion scaling looks at the difference between the process and its fluid limit, and measures the time in units of n and the state (counts of customers) in units of $\sqrt{n}$. The diffusion rescaling of $Z(t)$ by n is $\hat{Z}^n(t) = \sqrt{n}(\bar{Z}^n(t) - \bar{Z}(t)) = \frac{Z(nt)-\bar{Z}(nt)}{\sqrt{n}}$. As $n \to \infty$ we shall look (in analogy to the CLT) for $\hat{Z}^n(t) \stackrel{w}{\to} \hat{Z}(t)$, where $\hat{Z}(t)$, the diffusion limit, is a diffusion process, such as Brownian motion or reflected Brownian motion.

The diffusion limit captures the random fluctuation of the system around its fluid limit. When the diffusion limit is positive the unbounded variation of Brownian motion is the limiting expression of the very large number of small changes in the queue (individual arrivals and

departures) while the total queue length (in a congested situation, with large number in the queue) changes very slowly. Where the diffusion limit hits zero, the phenomena of very many extremely small busy periods in a congested queue is captured by the infinite number of returns to zero which a reflected Brownian motion has.

The fluid limit is a.s., the diffusion limit is in distribution. We can think of $\tilde{Z}(t) = \bar{Z}(t) + \hat{Z}(t)$ as an approximation to $Z(t)$ in the sense of strong approximations. Copies of the process can be embedded (Skorohod representation) in a joint probability space in which (under assumption of existence of moments of order r)

$$\sup_{0\le t\le T} |Z(t) - \tilde{Z}(t)| \stackrel{a.s.}{=} o(T^{1/r})$$

This expression illustrates the separation of scales in a different way: In $\tilde{Z}(t) = \alpha t + \sigma(R)BM(t)$ the fluid part is linear in t, the BM or RBM part has 'size' $\sqrt{t}$, and the error is of size $o(t^{1/r})$. At what times t these separate will then depend on the parameters.

So far we have tried to give some intuitive feel for fluid and diffusion approximations at the cost of being very vague and imprecise. The limiting behavior of queues has in fact a very complete mathematical treatment for single queue and single class queueing networks. The limit theory for multi-class queueing networks is at the research frontier. While quite a lot is known (see Bramson, Williams, Harrison and Dai) there are still many unclear issues.

3.2 Queueing networks

A (single class) queueing network consists of nodes $i = 1, \dots, I$. Customers from outside arrive at node i in an arriaval process $\mathcal{A}_i(t)$ with rate α_i. Customers at node i are served at rate μ_i according to a service process $\mathcal{S}_i(t)$ which counts the number of service completions achieved by service for a total duration t. Upon service completion a customer from node i moves to node j with probability p_{ij}, and leaves the system with probability $1 - \sum_j p_{ij}$. We let $\mathcal{R}(m_1, \dots, m_I)$ be a matrix counting process, which in its ij position counts the number of customers among the first m_i served at node i which upon service completion were switched to node j. We assume that interarrival times, service times and switching choices are independent sequences of i.i.d. random variables.

Queue balance equations

Queue balance equations for this are:

$$Q(t) = Q(0) + \mathcal{A}(t) - \mathcal{S}(B(t)) + \mathcal{R}'(\mathcal{S}'(B(t)))e \tag{3.1}$$

Notation: We denote by $\mathcal{S}(B(t))$ the vector of service completions up to time t, with components $\mathcal{S}_i(B_i(t))$, we let e be a vector of all 1's; also, e_j is the jth unit vector, and we use $x(y(t))$ to denote composition of vector functions. For vectors a, b we use $a \cdot b$ do denote elementwise product, $(a \cdot b)_i = a_i b_i$.

We assume work conserving policies, so that

$$B_i(t) = \int_0^t 1_{\{Q_i(s)>0\}} ds. \tag{3.2}$$

Clearly, the queue length process satisfies these relations. However, they do not give Q, B explicitly in terms of $\mathcal{A}, \mathcal{S}, \mathcal{R}$ so we will still need to investigate, (i) if solutions exist for any $\mathcal{A}, \mathcal{S}, \mathcal{R}$ (ii) if they are unique. Note that solutions exist for $\mathcal{A}, \mathcal{S}, \mathcal{R}$ step functions as defined for arrivals, services and switching.

Traffic equations and stability

The traffic equations determine at what rates customers can move in and out of the nodes of the network in the long run. The rates are the maximal solution of:

$$\lambda = \alpha + P'(\lambda \wedge \mu) \tag{3.3}$$

If the solution to the traffic equations are $\lambda < \mu$, then the queuing network is stable: It can be shown that the Markov process obtained by adding attained services and interarrival vectors (for non-preemptive service) is positive Harris recurrent. Hence a stationary distribution (invariant measure) exists and we can use it to calculate long term average performance measures. Other forms of stability (e.g. equality of input and output rates) under other model assumptions (e.g. $(\mathcal{A}, \mathcal{S}, \mathcal{R})$ ergodic) have also been shown.

In the special case that all the interarrival and service times are memoryless (exponenetially distributed), the queue length is a continuous time Markov chain on a countable state space, and (wonder of wonders) it has a very simple steady state distribution: The queue lengths are independent, each geometric:

$$\mathbb{P}(Q(t) = (n_1, \ldots, n_I)) = c \prod_{i=1}^{I} \rho_i^{n_i}$$

where $\rho_i = \lambda_i/\mu_i$ is the traffic intensity at node i. This result is due to Jackson, after whom these memoryless networks are called Jackson network. With general distributions they are (sometimes) called generalized Jackson networks.

If not all $\lambda < \mu$, we can partition $\{1, \ldots, I\}$ into three sets, $\mathbf{a} = \{i : \lambda_i < \mu_i\}$, $\mathbf{b} = \{i : \lambda_i = \mu_i\}$, $\mathbf{c} = \{i : \lambda_i > \mu_i\}$. The queues at nodes $\mathbf{b}, \mathbf{c}$ are unstable, they are respectively balanced and strict bottlenecks. Those in $\mathbf{c}$ grow like a random walk, and after a finite amount of time are never again 0. The queues at nodes in $\mathbf{b}$ are null recurrent, and will return to 0 but only at very irregular intervals, and as time t goes on we should expect to find $\sqrt{t}$ customers there. In particular, the output stream out of nodes in $\mathbf{b}, \mathbf{c}$ is at rate μ_i, and can be added to the exogenous input to nodes in $\mathbf{a}$. The queues at the nodes in $\mathbf{a}$ are stable, and in the memoryless case they have product form joint steady state distributions.

Centering the queue balance equation

We rewrite (3.1) (we use $w = u \cdot v$ to denote elementwise product $w_j = u_j v_j$):

$$\begin{aligned} Q(t) &= [Q(0) + (\alpha + (P' - I)\mu)t + (\mathcal{A}(t) - \alpha t) - (\mathcal{S}(B(t)) - \mu \cdot B(t)) \\ &\quad + (\mathcal{R}'(\mathcal{S}'(B(t)))e - P'\mathcal{S}(B(t))) + (P'\mathcal{S}(B(t)) - P'\mu \cdot B(t))] \\ &\quad + (I - P')\left[\mu \cdot (et - B(t))\right] \\ &= \mathcal{X}(t) + (I - P')\mathcal{Y}(t) \end{aligned} \tag{3.4}$$

We call $\mathcal{X}(t)$ the netput process, and $\mathcal{Y}(t) = \mu \cdot (et - B(t))$ is the cumulative lost (expected) production.

The fluid model of the QN

We consider the same system, but instead of discrete customers moving at random times we now assume deterministic continuous flows. The network has the same nodes, and starts with the same initial fluid levels (queue lengths) $Q(0)$. Cumulative exogenous inflow is αt, service is at maximal possible rate (work conservation) up to μ, and fluid that flows out of buffer i splits in proportions p_{ij} into other nodes, with a fraction $1 - \sum_{j \neq i} p_{ij}$ flowing out of the system.

The potential outflow from node i up to time t is $\mu_i t$. The actual flow may be less than that, and we let $y_i(t)$ denote the lost outflow. The balance equations for the fluid levels are:

$$q(t) = [Q(0) + \alpha t + (P' - I)\mu t] + (I - P')y(t) = \bar{x}(t) + (I - P')y(t) \tag{3.5}$$

With a fluid system we can have flow out of an empty buffer (at a rate equal to the inflow). There is therefore no direct analog to (3.2). However, we can certainly say that while $q_k > 0$ there will be no lost flows (under work conserving policies), hence:

$$\int_0^t q_i(t) dy_i(t) = 0, \quad i = 1, \ldots, I. \tag{3.6}$$

In fact the flow rates turn out to be piecewise constant and the queue lengths and busy times are continuous piecewise linear. We return to that later.

Oblique reflection mapping

The queue length process satisfies (3.2), which implies also that

$$\int_0^t Q_k(t) d\mathcal{Y}_k(t) = 0, \quad i = 1, \ldots, I. \tag{3.7}$$

We have for both the QN and the fluid network the following problem:
The oblique reflection problem (Skorohod problem). Let $x(t)$ be a vector function in D^n, the space of RCLL functions, with $x(0) \geq 0$, and let M be an $n \times n$ matrix. Find functions y, z in D^n such that:

(i) $z = x + My \geq 0$,

(ii) $y(0) = 0$ and y non decreasing,

(iii) $\int z' dy = 0$.

Theorem 3.1 *If $M = I - P'$ where P is a non-negative matrix with spectral radius less than 1, then (i,ii,iii) have a unique solution. Furthermore, one can replace the requirement (iii) by the equivalent requirement*

(iii') y is the minimal vector function satisfying (i,ii).

Also, the mapping $x \to y, z$ is Lipshitz continuous in the metric of D.

We call such M an $\mathcal{M}$-matrix. The metric of D is the Skorohod metric, but if one of the functions is continuous then the Skorohod metric is equivalent to the supremum metric.

It follows from this that the balance equations and work conservation conditions (3.1,3.7) determine the queue length $Q(t)$ and the expected lost processing $\mathcal{Y}(t)$ and hence also the busy time $B(t)$, uniquely. Similarly, (3.5,3.6) determine the fluid level process $q(t)$ and lost flows $y(t)$ uniquely.

Connection to the linear complementarity problem

For given vector q and matrix M the linear complementatity proplem LCP(q, M) is to find η, ζ such that:

(i) $\zeta = q + M\eta \geq 0$,

(ii) $\eta \geq 0$,

(iii) $\zeta'\eta = 0$.

A matrix is a $\mathcal{P}$-matrix if all its principal minors are positive.

Proposition 3.2 *LCP(q, M) has a unique solution for all q if and only if M is a $\mathcal{P}$-matrix.*

The linear complementarity proplem is associated with the above oblique reflection problem (Skorohod problem) as follows: If we look at the processes x, y, z only at discrete times $t = 0, 1, \ldots$ then the conditions (i,ii,iii) of oblique reflection hold if and only if we can successively solve for each $t = 1, 2, \ldots$ LCP$(z(t-1) + x(t) - x(t-1), M)$ in which case $z(t) = \zeta$, $y(t) = y(t-1) + \eta$. Mandelbaum calls this the discrete dynamic complementarity problem.

The oblique reflection problem needs however to satisfy (i,ii,iii) for all $t > 0$. Mandelbaum refers to it as the continuous dynamic complementarity problem (DCP). If x is piecewise constant or piecewise linear then we can solve DCP for all t by solving LCP for a discrete sequence of points. In general however DCP is harder. In fact, no necessary and sufficient conditions are known for uniqueness of the solution of DCP. A sufficient condition is that M is an $\mathcal{M}$-matrix.

Consider now the traffic equations. It is immediately seen that $\zeta = (\lambda - \mu)^+$, $\eta = (\mu - \lambda)^+$ solve LCP$(\alpha + (P' - I)\mu,\ I - P')$.

Fluid approximation

Consider fluid scaling: We look at the centered queueing equation, counting time in units of n and space (customers) in units of n.

$$\frac{1}{n}Q(nt) \quad = \quad [\frac{1}{n}Q(0) + (\alpha + (P' - I)\mu)t + (\frac{1}{n}\mathcal{A}(nt) - \alpha t) - (\frac{1}{n}\mathcal{S}(B(nt)) - \mu \cdot \frac{1}{n}B(nt))$$

$$
\begin{aligned}
&+(\frac{1}{n}\mathcal{R}'(\mathcal{S}'(B(nt)))e - P'\frac{1}{n}\mathcal{S}(B(nt))) + (P'\frac{1}{n}\mathcal{S}(B(nt)) - P'\mu\cdot\frac{1}{n}B(nt))] \quad (3.8)\\
&+(I-P')\left[\mu\cdot(et-\frac{1}{n}B(nt))\right]\\
= \;& \frac{1}{n}\mathcal{X}(nt) + (I-P')\frac{1}{n}\mathcal{Y}(nt)
\end{aligned}
$$

As written we would have $\frac{1}{n}Q(0) \to 0$ as $n \to \infty$, and since we want to study systems with customers in them we need to look at a sequence of systems, with different initial queue lengths, and let $\frac{1}{n}Q^n(0) \to \bar{Q}(0)$. We use bar notation to denote the fluid scaling, $\bar{x}(t) = \frac{1}{n}x(nt)$. We obtain for the sequence of systems (where we use $B^n(nt) = n\frac{1}{n}B^n(nt) = n\bar{B}^n(t)$):

$$
\begin{aligned}
\bar{Q}^n(t) &= [\bar{Q}^n(0) + (\alpha + (P'-I)\mu)t + (\bar{\mathcal{A}}^n(t) - \alpha t) - (\bar{\mathcal{S}}^n(\bar{B}^n(t)) - \mu\cdot\bar{B}^n(t)) \quad (3.9)\\
&\quad +(\bar{\mathcal{R}}^{n\prime}(\bar{\mathcal{S}}^{n\prime}(\bar{B}^n(t)))e - P'\bar{\mathcal{S}}^n(\bar{B}^n(t))) + (P'\bar{\mathcal{S}}^n(\bar{B}^n(t)) - P'\mu\cdot\bar{B}^n(t))]\\
&\quad +(I-P')\left[\mu\cdot(et-\bar{B}^n(t))\right]\\
&= \bar{\mathcal{X}}^n(t) + (I-P')\bar{\mathcal{Y}}^n(t)
\end{aligned}
$$

We now let $n \to \infty$, and note that most terms $\to 0$. This is best seen by going back to the form:

$$
\begin{aligned}
\bar{Q}^n(t) &= [\frac{1}{n}Q^n(0) + (\alpha + (P'-I)\mu)t + (\frac{1}{n}\mathcal{A}(nt) - \alpha t) - (\frac{1}{n}\mathcal{S}(B(nt)) - \mu\cdot\frac{1}{n}B(nt))\\
&\quad +(\frac{1}{n}\mathcal{R}'(\mathcal{S}'(B(nt)))e - P'\frac{1}{n}\mathcal{S}(B(nt))) + (P'\frac{1}{n}\mathcal{S}(B(nt)) - P'\mu\cdot\frac{1}{n}B(nt))] \quad (3.10)\\
&\quad +(I-P')\left[\mu\cdot(et-\frac{1}{n}B(nt))\right]\\
&= \bar{\mathcal{X}}^n(t) + (I-P')\bar{\mathcal{Y}}^n(t).
\end{aligned}
$$

Using the FSLLN we get that:

$$\bar{\mathcal{X}}^n(t) \to \bar{Q}(0) + (\alpha + (P'-I)\mu)t$$

and by the Lipschitz continuity of the oblique reflection mapping, $\bar{Q}^n(t), \bar{\mathcal{Y}}^n(t)$ will converge to the reflection mapping of $\bar{Q}(0) + (\alpha + (P'-I)\mu)t$, which is exactly the fluid model (3.4,3.7)

Diffusion approximation

We now subtract from the system processes their fluid approximations and observe the difference on a $\sqrt{n}$ scale: $\hat{z}(t) = \frac{z(nt)-\bar{z}(nt)}{\sqrt{n}}$. The diffusion approximation provides a diffusion process which approximates this.

One again considers a sequence of systems. The initial states of the systems in the sequence satisfy $\frac{1}{\sqrt{n}}Q^n(nt) \to \hat{Q}(0)$, so that the initial queue length is still large but not as large as in fluid scaling. Also, to study the behavior of queues which are close to balanced, one considers sequences of parameters, α^n, μ^n, such that $\sqrt{n}(\alpha^n - \alpha) \to d_\alpha$, $\sqrt{n}(\mu^n - \mu) \to d_\mu$, the switching matrix is fixed P for all n, and the solutions to the traffic equations then satisfy:

$\sqrt{n}(\lambda^n - \lambda) \to d_\lambda$. One then has again a partition of the nodes into sets $\mathbf{a}, \mathbf{b}, \mathbf{c}$ of non-bottleneck, balanced bottleneck and strict bottlenecks.

Even though the parameters of each system in the sequence are different now, it is possible to generate interarrivals and processing times using a common sequence of mean 1 values, which are then scaled differently in each system.

The exact fluid approximation is given in the following Figure 1 which is copied from the paper by Chen and Mandelbaum (*Annals of Probability* 19:1463–1519).

Note the following main points:

The diffusion scaled netput process converges to a Brownian motion. The parameters are obtained from the renewal processess of interarrivals and the renewal processes of services, the latter are geometrically compounded according to the switching. One uses Wald's formulas to obtain 1st and 2nd moments. As $n \to \infty$ these converge to Brownian motion with the corresponding limiting parameters.

The initial content of the nodes $\mathbf{a}$ is switched instantly to the initial netputs for the nodes $\mathbf{b}, \mathbf{c}$.

The limiting queue length and the limiting lost processing are the reflection and regulator of the limiting Brownian netput. The busy time and other processes such as virtual workload are obtained from those.

For the non-bottleneck nodes $\mathbf{a}$, the diffusion queue $\hat{Q} = 0$ while $\hat{B}$ is (almost) a Brownian motion (correction terms for lost processing switched from nodes of $\mathbf{b}$ are added).

For the strict bottleneck nodes $\mathbf{c}$ the centered scaled busy time is 0, corresponding to the fact that the busy time converges to the fluid limit of t even on diffusion scale. The diffusion queue length is (almost) a Brownian motion (correction terms for lost processing switched from nodes of $\mathbf{b}$ are added).

3.3 Reflected Brownian Motions

We saw that the queue length can be approximated (e.g. through fluid approximation, diffusion approximation, or strong approximation) by a reflected Brownian motion. A process $Z(t)$ is a reflected Brownian motion in the non-negative orthant of d-dimensional space, denoted $Z(t) \sim \text{RBM}(z(0), \theta, \Gamma, R)$, where $z(0) \geq 0$ is the initial state, θ is the drift, Γ a positive definite covariance matris, so that $Z(t)$ is the oblique reflection of a Brownian motion $X(t) = z(0) + \theta t + \Gamma^{1/2} BM(t)$ with reflection matrix R, where $Z(t) = X(t) + RY(t)$ and $Y(t)$ is the regulator.

In the case of a one dimensional RBM one has closed form formulae for the distributions of $Z(t)$. In particular, if $R > 0, \theta < 0, \Gamma = \sigma^2$, process $Z(t)$ possesses a stationary distribution, $Z(t \to \infty) \sim \exp(-2\theta/\sigma^2)$.

For the multivariate case a closed form solution for the stationary distribution of an RBM exists only for some very special cases. Here is the most important of them. We assume

that RBM has been scaled so that the diagonal elements of R, Γ are 1, R is inveritble, and $\max_{1\le i\le d} \left|(R^{-1}\theta)_i\right| = 1$.

Let $\gamma_i = -(R^{-1}\theta)_i$. If $\gamma_i > 0$, and $2\Gamma_{jk} = (R_{kj} + R_{jk})$ for all $0 \le i, j \ne k \le d$ then $Z(t)$ has product form stationary distribution, with $Z_i(t \to \infty)) \sim \exp(2\gamma_i)$.

For other cases there exists a recently developed numerical algorithm to obtain the distribution of $Z(t)$. This is the QNET algorithm of Jim Dai, which we describe now.

The RBM $Z(t)$ moves as a Brownian motion in the interior of the positive d-orthant $S = \{(x_1, \ldots, x_d) : x_i \ge 0\}$ and is reflected from its faces, $F_i = \{(x_1, \ldots, x_d) : x_i = 0, x_j \ge 0, j \ne i\}$. Let v_i be the ith column of R, and let $L_i(t)$ be the ith coordinate of the regulator $Y(t)$.

3.3.1 Basic adjoint relationship

Theorem 3.3 *Let π be a stationary distribution for RBM$(z(0), \theta, \Gamma, R)$. For each $i = 1, \ldots, d$ there exists a finite Borel measure ν_i on F_i so that for any Borel set in F_i:*

$$E_\pi\left(\int_0^t 1_A(Z(s))dL_i(s)\right) = \frac{1}{2}t\nu_i(A).$$

π and ν_i are absolutely continuous w.r.t. the Lebesgue measures $dx, d\sigma_i$ on their domains, with densities $p_0 d\pi/dx, p_i = dp_i/d\sigma_i$, which satisfy the following Basic adjoint relationship*:*

$$\int_S (\mathcal{G}f \cdot p_0)dx + \frac{1}{2}\sum_{i=1}^d \int_{F_i} (\mathcal{D}_i f \cdot p_i)d\sigma_i = 0, \text{ for all } f \in C_b^2(S)$$

where

$$\mathcal{G}f = \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d \Gamma_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j} + \sum_{i=1}^d \theta_i \frac{\partial f}{\partial x_i},$$

$$\mathcal{D}_i f = v_i \cdot \nabla f(x), \quad x \in F_i,$$

and $C_b^2(S)$ is the space of twice differentiable functions which together with their first and second order partials are continuous and bounded on S.

Letting $\mathcal{A}f = (\mathcal{G}f, \mathcal{D}_1 f, \ldots, \mathcal{D}_d f)$, $p = (p_0, p_1, \ldots, p_d)$ and $d\lambda = (dx, \frac{1}{2}d\sigma_1, \ldots, \frac{1}{2}d\sigma_d)$, the basic adjoint relationship reads:

$$\int_S (\mathcal{A}f \cdot p)d\lambda \text{ for all } f \in C_b^2(S) \tag{3.11}$$

To obtain p we need to solve (3.11), i.e find p ortogonal to $\mathcal{A}f$. However, $\mathcal{A}f$ may not be square integrable over S which is unbounded. What we do instead is use a *reference measure*, $q = (q_0, q_1, \ldots, q_d)$. As the reference measure we use the product form exponential distribution with parameters $2\gamma_i$, as described above. We then let: $d\eta = qd\lambda = (q_0 dx, \frac{1}{2}q_1 d\sigma_1, \ldots, \frac{1}{2}q_d d\sigma_d)$ and we are now looking for unknown r which solves:

$$\int_S (\mathcal{A}f \cdot r)d\eta \text{ for all } f \in C_b^2(S) \tag{3.12}$$

and from the solution we get $p = r \cdot q$.

We get r and p as follows:

Theorem 3.4 *Let H = the closure of $\mathcal{A}f : f \in C_b^2(S)$. Let $\phi_0 = (1, 0, \ldots 0)$, let $\bar{\phi}$ be the projection of ϕ_0 on H, and $\tilde{\phi} = \phi_0 - \bar{\phi}$. If $r \in L^2(S, d\eta)$ (and assuming a conjecture about uniquness), then $p = \frac{1}{||\tilde{\phi}||^2}\tilde{\phi} \cdot q$.*

Theorem 3.5 *Let H^n be a sequence of finite dimensional subspaces of H such that $H^n \uparrow H$. Let $\bar{\psi}^n$ be the projection of ϕ_0 on H^n, and $\tilde{\phi}^n = \phi_0 - \bar{\psi}^n$. define $p^n = \frac{1}{||\tilde{\psi}^n||^2}\tilde{\psi}^n \cdot q$. Then $p^n \to p$ in $L^2(S, d\lambda)$.*

Choose as basis for H^n the functions $\{\mathcal{A}x_1^{i_1} \cdots x_d^{i_d} : i_1 + \cdots + i_d \leq n\}$.

The steps of the calculation of p^n then involve the calculation of the projection of ϕ_0 onto H^n, which is done by solving a set of linear equations, of dimension $N = \binom{n+d}{d}$.

This algorithm has been coded by Jim Dai and is available??

3.4 Single multiple output networks

Consider the following model: Nodes are $i = 1, \ldots, I$, and there are several output choices from each node. Let $j = 1, \ldots, J$ be the list of all processing options, where: Processing option j prcesses jobs out of the queue of node $i = \kappa(j)$, so that there is a partition D_i of the processing options according to the node which they process. Processing time by processing option j takes a time distributed $X \sim G_j$, with rate μ_j. Jobs which complete their processing by processing option j switch into queue k with probability P_{jk}. In other words, we have a switching matrix P with J rows and I columns.

We can now pose the problem of which processing options to use, to control this system optimally.

We shall discuss this proplem later.

3.5 Multi class queueing networks

In a multiclass queueing network, each node of $i = 1, \ldots, I$ has several classes of customers. The classes are $k = 1, \ldots, K$, partitioned according to the nodes, C_i where $k \in C_i$ if k is served at node i, $i = \sigma(k)$. Average processing time of jobs of class k is m_k (processing rate $\mu_k = 1/m_k$), and on completion of service job of class k is switched to class l randomly, with switching probability P_{kl}, or it leaves the system with probability $1 - \sum_l P_{kl}$. The special feature here is that the behavior of a MCQN is highly dependent on the policy, that is on how each node handles its various classes of customers.

We write the dynamics of a MCQN similar to those of a single class network:

$$Q(t) = Q(0) + \mathcal{A}(t) - \mathcal{S}(T(t)) + \mathcal{R}'(\mathcal{S}'(T(t)))e \tag{3.13}$$

where as before, $\mathcal{A}(t)$ is the stream of external arrivals, $\mathcal{S}(t)$ the service completions, $\mathcal{R}(m_1, \ldots, m_K)$ the matrix of switch counts, and all of these are now K or $K \times K$ dimenisonal. The new feature are $T(t)$, the K vector of cumulative processing times, where $T_k(t)$ is the cumulative time

devoted to the service of class k. This replaces the busy time $B_i(t)$ of node i in the dynamics (3.1) of single class queueing networks, and to supplement the busys time we define the idle time $Y_i(t) = t - B_i(t)$. We have:

$$\begin{aligned} &T_k(0) = 0, \quad Y_i(0) = 0, \\ &T_k(t),\, Y_i(t) \text{ are non-decreasing}, \\ &\sum_{k \in C_i} T_k(t) + Y_i(t) = t, \end{aligned}$$

The condition of work conservation is

$$B_i(t) = \int_0^t 1_{\{\sum_{k \in C_i} Q_k(s) > 0\}} ds, \tag{3.14}$$

or equivqlently (for service of discrete customers)

$$\int_0^t \sum_{k \in C_i} Q_k(s) dY_i(s) = 0. \tag{3.15}$$

The big difference from single class QN is that these equations do not determine the behavior of the system. The behavior of the system, in particular the division of the busy time of node i between the classes in C_i, is determined by the policy.

Once the policy is given, the queue lengths are uniquely determined. The dependence of the system dynamics on the policy is expressed by supplementary equations of conditions. We list some of them:

FIFO

$$\mathcal{D}_k(t + W_{\sigma(k)}(t)) = Q_k(0) + \mathcal{A}_k(t)$$

where $\mathcal{D}_k$ is the departure process of service completions at buffer k and $W_i(t)$ is the total virtual workload at station i. The equation says that arrivals at t depart at t plus all the work at the station $i = \sigma(k)$.

Static Buffer Pirority Policies Let $Q_k^+(t) = \sum_j$ with priority $\geq k$ $Q_j(t)$ with a similar definition for $T_k^+(t)$. Then:

$$\int_0^\infty Q_k^+(t) d(t - T_k^+(t)) = 0$$

expresses the rule that while customers of priority $\geq k$ are in the system, all the capacity of the machine is allocated to them.

Generalized Head of the Line Processor Sharing Here all non-empty buffers are served simultaneously, the effort on buffer k is proportional to β_k and is devoted to processing the head of the line customer in the buffer.

$$T_k(t) = \int_0^t \frac{\beta_k 1_{Q_k(s)>0}}{\sum_{l \in C_i} \beta_l 1_{Q_l(s)>0}}$$

If the departure process is increasing in steps of 1, and if no sumultaneous events occur (e.g. sumultaneous departures or arrivals) then the usual equations and the supplementary equations, for work conserving head of the line policies, determine the whole queue dynamics.

Traffic equations for the network are as before:

$$\lambda = \alpha + P'(\lambda \wedge \mu) \tag{3.16}$$

and if the solution satisfies $\lambda_k < \mu_k$ for all k then λ_k is the only possible long term throughput rate for class k. $\lambda_k < \mu_k$ for all k is a necessary condition for stability of the network. In fact a more stringent condition is necessary. Define:

$$\rho_i = \sum_{k \in C_i} \lambda_k m_k$$

this is the total amount of processing required by machine i per unit time, to maintain the thorughput λ_k for all buffers $k \in C_i$. A necessary condition for stability is $\rho_i < 1$ for all nodes i.

However, unlike the case of single class queueing networks, this condition is not sufficient to guarantee stability of the queueing network under any work conserving policy. We shall examine counter examples of unstable policies as well as examples of stable policies in the next lecture. We now define the tool of examining stability, i.e. the fluid model, and quote the theorem of Jim Dai that enables us to deduce stability of the stochastic system from its fluid model.

3.5.1 Kelly networks

One kind of multiclass queueing network that can be analyzed is Kelly networks. Here it is assumed that service times of all jobs of the classes $k \in C_i$ have the same average processing time m_i, and further, all processing times and interarrival times are exponentially distributed. Under this assuption, the steady state distribution of the multivariate queue length process is of product form if the service policy is 'symmetric', e.g. if it is FIFO, LIFO. The product form solution is

$$\mathbb{P}(Q(t \to \infty) = q) = \prod_{i=1}^{I} (1 - \rho_i) \prod_{k \in C_i} (\lambda_k m_k)^{q_k}$$

3.5.2 Fluid equations for MCQN

In analogy with the single class case, similar to the queue balance equations, we get for the fluid:

$$q(t) = Q(0) + \alpha t + (P' - I)\mu \cdot T(t) \tag{3.17}$$

$$\int_0^t \sum_{k \in C_i} q_k(s) dY_i(s) = 0. \tag{3.18}$$

The supplementary equations for the various disciplines are:

FIFO

$$d_k(t + w_{\sigma(k)}(t)) = q_k(0) + \alpha_k t$$

Static Buffer Pirority Let $q_k^+(t) = \sum_j$ with priority $\geq_k q_j(t)$ with a similar definition for $T_k^+(t)$. Then:

$$\int_0^\infty q_k^+(t) d(t - T_k^+(t)) = 0$$

Generalized Head of the Line Processor Sharing

$$T_k(t) = \int_0^t \frac{\beta_k 1_{q_k(s)>0}}{\sum_{l \in C_i} \beta_l 1_{q_l(s)>0}}$$

It is no longer true that the fluid queue length process is uniquely determined by the usual plus supplementary equations. In particular, one may have a solution which is constant 0, and another which pops up from 0.

A fluid solution is any solution to the usual plus supplementary fluid equations.

3.5.3 Fluid limits

To investigate a MCQN under some policy, we consider not just Q, but the whole sixtuplet: $\mathbf{X}(t) = (\mathcal{A}(t), \mathcal{D}(t), T(t), \mathcal{W}(t), Y(t), Q(t)), t \geq 0$.

We let the fluid scaling of this process be:

$$\bar{\mathbf{X}}^r(t) = r^{-1}\mathbf{X}(rt)$$

and if we consider a sequence of systems the fluid scaling is:

$$\bar{\mathbf{X}}^r(t) = r^{-1}\mathbf{X}^r(rt)$$

We shall usually take a sequence of syetems which differ only in their initial conditions ($Q(0)$ and for FIFO also the departures up to time $t < W(0)$). We let $|Q(0)|$ be the total number in system at time 0. Dai shows that if $|Q^r(0)|/r$ is bounded, then $\bar{\mathbf{X}}^r(\cdot, \omega)$ is pre-compact, and so each sequence of r values has a subsequence for which $\bar{\mathbf{X}}^{r'}(\cdot, \omega)$ converges in the Skorohod topology. We call these fluid limits.

Proposition 3.6 *Every fluid limit is a fluid solution.*

3.5.4 Stability and instaiblity of the fluid model

Definition 3.7 *The fluid model is stable if there exists $\delta > 0$ such that for each fluid solution $x(t)$ with $|Q(0)| < 1$, $Q(t) = 0$ for $t > \delta$.*

Definition 3.8 *The fluid model is weakly stable if for each fluid solution $x(t)$ with $Q(0) = 0$, $Q(t) = 0$ for $t > 0$.*

Definition 3.9 *fluid solution $x(t)$ is unstable if there exists a sequence $t_n \to \infty$ such that $Q(t_n) > 0$ for all n. The fluid model is unstable if there exists an unstable fluid solution.*

Definition 3.10 *The fluid model is weakly unstable if there exists $\delta > 0$ such that for each fluid solution $x(t)$ with $Q(0) = 0$, $Q(\delta) \neq 0$.*

Theorem 3.11 *If the fluid model is weakly unstable then the corresponding fluid network is unstable in the sense that with probability 1: $|Q(t)| \to \infty$ as $t \to \infty$.*

Definition 3.12 *A MCQN is rate stable if for each fixed $\mathbf{X}(0)$, with probability 1, $\lim_{t\to\infty} \mathcal{D}_k(t)/t = \lambda_k$.*

Theorem 3.13 *The queueing network is rate stable if and only if for each fixed initial data, with probability 1, the fluid limit $\bar{\mathbf{X}}$ is uniquely given by*

$$\begin{array}{ll} \bar{\mathcal{A}}(t) = \lambda t, & \bar{\mathcal{D}}(t) = \lambda t, \\ \bar{T}(t) = \mu \cdot \lambda t, & \bar{\mathcal{W}}(t) = 0, \\ \bar{Y}(t) = (e - \rho)t, & \bar{Q}(t) = 0, \end{array}$$

The previous results hold whenever we have $\mathcal{A}, \mathcal{S}$ with stable rates. A stronger definition of stability is positive Harris recurrence. To obtain Harris recurrence, we must have a more structured model: Assume interarrivals and services are i.i.d. One then needs to identify an appropriate Markovian process $\mathcal{X}(t)$ which contains in its state the queue length at time t as well as additional information which is necessary to implement the poilicy, as well as attained service times, so that given $\mathcal{X}(t)$ determines the future of $\mathbf{X}$, and makes past and future independent. This can be done for most policies, in such a way that the resulting state space is a finite dimensional vector space.

Theorem 3.14 *Conisder a head of the line work conserving MCQN with independent i.i.d. interarrival and service time sequences. Assume the distribution of interarrivals is unbounded and spread out. If the corresponding fluid limit model is stable, then $\mathcal{X}(t)$ is positive Harris recurrent.*

6.1. For the nth network, the exogenous arrival process is given by $A^n = \{A^n(t) = A^0(\lambda^{0,n}t),\ t \geq 0\}$, the service process by $S^n = \{S^n(t) = S^0(\mu^n t),\ t \geq 0\}$ and the routing sequence R by 2.2.C–2.2.D. The queue lengths Q^n and busy-times B^n are constructed via (2.2)–(2.3). We assume that for some J-dimensional vector c^{λ^0}, c^{μ} and a random vector $\hat{Q}(0) \geq 0$, the following limits exist as $n \to \infty$:

6.1.A $\qquad \sqrt{n}\,(\lambda^{0,n} - \lambda^0) \to c^{\lambda^0},$

6.1.B $\qquad \sqrt{n}\,(\mu^n - \mu) \to c^{\mu},$

6.1.C $\qquad \frac{1}{\sqrt{n}} Q^n(0) \to_d \hat{Q}(0).$

In the formulation of the theorem we use a J-dimensional driftless Brownian motion

6.1.D $\qquad \hat{\xi} = \mathrm{BM}(0, \hat{\Lambda})$

which starts at $\hat{\xi}(0) = 0$. The covariance matrix $\hat{\Lambda} = [\hat{\Lambda}_{jk}]$ is given by

6.1.E
$$\hat{\Lambda}_{jk} = \left[\lambda_j^0\left(a_j^2 - 1\right) + \lambda_j + (\lambda_j \wedge \mu_j)b_j^2\right]\delta_{jk} - (\lambda_j \wedge \mu_j)b_j^2 p_{jk} - (\lambda_k \wedge \mu_k)b_k^2 p_{kj} - \sum_{l=1}^{J} (\lambda_l \wedge \mu_l)p_{lj}p_{lk}\left[1 - b_l^2\right],$$

where λ is the inflow capacity vector of the open network (λ^0, P, μ), as determined by (2.1). We maintain the "bar" convention that was introduced for closed networks in Subsection 4.4. The "hat" convention changes, however, because we rescale open networks differently. Indeed, our diffusion limits for open networks arise as time is accelerated by a factor of n while space is aggregated by a factor of $\sqrt{n}$. In accordance with this rescaling, let

$$\hat{Q}^n(t) = \sqrt{n}\left[\bar{Q}^n(t) - (\lambda - \mu)^+ t\right], \qquad \hat{W}^n(t) = \sqrt{n}\left[\bar{W}^n(t) - (\rho - e)^+ t\right],$$

$$\hat{D}^n_{j,h}(t) = \sqrt{n}\,\bar{D}^n_{j,h}(t), \qquad \hat{B}^n(t) = \sqrt{n}\left[\bar{B}^n(t) - (\rho^n \wedge e)t\right],$$

$$\hat{S}^n(t) = \sqrt{n}\left[\bar{S}^n(t) - \mu^n t\right], \qquad \hat{V}_j^n(t) = \sqrt{n}\left[\bar{V}_j^n(t) - t\right]/\mu_j^n,$$

$$\hat{A}^n(t) = \sqrt{n}\left[\bar{A}^n(t) - \lambda^{0,n}t\right], \qquad \hat{R}^n(t) = \sqrt{n}\left[\bar{R}^n(t) - P^t t\right].$$

Here ρ is the traffic intensity vector of the network (λ^0, P, μ). Recalling that the sets α, β and γ stand for nonbottleneck, balanced and strict bottleneck stations, respectively, we now have:

THEOREM 6.1. *Consider the above sequence of open networks. Assume that* 6.1.A–6.1.C *hold and let* $\hat{\xi}$ *be the Brownian motion* 6.1.D. *Then the weak convergence*

(6.1) $$\left(\hat{Q}^n, \hat{W}^n, \hat{B}^n, \hat{D}^n\right) \to_d \left(\hat{Q}, \hat{W}, \hat{B}, \hat{D}\right) \quad \text{in } t > 0,$$

holds as $n \to \infty$. *The limit is described for* $t > 0$ *by*

(6.2) $$\hat{Q}_\alpha = 0,$$

(6.3) $$\hat{Q}_\beta = X + \left[I - \bar{P}'_\beta\right]Y,$$

(6.4) $$X(t) = X(0) + \hat{\xi}_\beta(t) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{\xi}_\alpha(t) + \left\{\bar{c}_\beta^{\lambda^0} + \left[\bar{P}'_\beta - I\right]c_\beta^\mu + \bar{P}'_{\gamma\beta}c_\gamma^\mu\right\}t,$$

(6.5) $$X(0) = \hat{Q}_\beta(0) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{Q}_\alpha(0),$$

(6.6) $$\bar{c}_\beta^{\lambda^0} = c_\beta^{\lambda^0} + P'_{\alpha\beta}[I - P'_\alpha]^{-1}c_\alpha^{\lambda^0},$$

(6.7) $$\bar{P}_{\gamma\beta} = P_{\gamma\beta} + P_{\gamma\alpha}[I - P_\alpha]^{-1}P_{\alpha\beta},$$

(6.8) $$\bar{P}_\beta = P_\beta + P_{\beta\alpha}[I - P_\alpha]^{-1}P_{\alpha\beta},$$

(6.9) $$Y = \Psi_{\bar{P}_\beta}(X),$$

(6.10) $$\hat{Q}_\gamma = \left[\hat{Q}_\gamma(0) + \hat{\xi}_\gamma\right] + P'_{\alpha\gamma}[I - P'_\alpha]^{-1}\left[\hat{Q}_\alpha(0) + \hat{\xi}_\alpha\right] - \bar{P}'_{\beta\gamma}Y,$$

(6.11) $$\bar{P}_{\beta\gamma} = P_{\beta\gamma} + P_{\beta\alpha}[I - P_\alpha]^{-1}P_{\alpha\gamma},$$

(6.12) $$\hat{W} = \mathrm{diag}(\mu^{-1})\left[\hat{Q} - \mathrm{diag}(\rho - e)^+ c^\mu\right],$$

(6.13) $$\hat{D}_{j,h} = \sum_{k \in \beta} \frac{h_k}{\mu_k}\hat{Q}_k = h\hat{W}, \qquad h_\gamma = 0,$$

(6.14) $$\hat{B}_\alpha = \mathrm{diag}(\mu_\alpha^{-1})[I - P'_\alpha]^{-1}\left[\hat{Q}_\alpha(0) + \hat{\xi}_\alpha - P'_{\beta\alpha}Y + P'_{\beta\alpha}\left[c_\beta^\mu - c_\beta^\lambda\right]^+ t\right],$$

(6.15) $$c_\beta^\lambda = \left\{c_\beta^{\lambda^0} + P'_{\alpha\beta}[I - P'_\alpha]^{-1}c_\alpha^{\lambda^0} + \bar{P}'_{\gamma\beta}c_\gamma^\mu\right\} + \bar{P}'_\beta\left(c_\beta^\lambda \wedge c_\beta^\mu\right),$$

(6.16) $$\hat{B}_\beta = -\mathrm{diag}\left(\mu_\beta^{-1}\right)Y,$$

(6.17) $$\hat{B}_\gamma = 0.$$

Figure 1: Diffusion Approximation for a QN