Stability and Performance of Multi-Class Queueing Networks with Infinite Virtual Queues

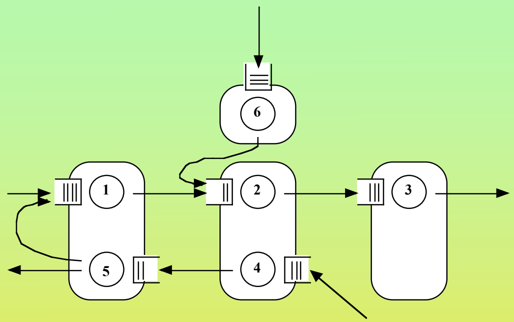
YongJiang Guo[1], Erjen Lefeber[2], Yoni Nazarathy[3],
Gideon Weiss[4], Hanqin Zhang[5]

Workshop in honor of Frank Kelly,
Eurandom, Eindhoven, 28-29/4/2011

1 Beijing University of Post and Telecommunications
2 Eindhoven University of Technology
3 Swinburne University of Technology
4 The University of Haifa
5 National University of Singapore

Multi-Class Queueing Networks with Infinite Virtual Queues

Standard MCQN



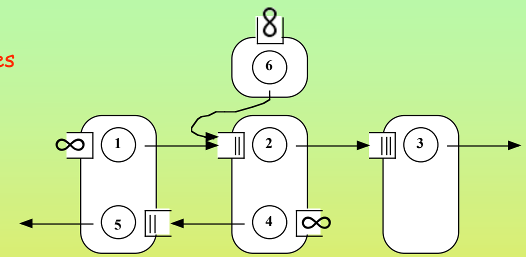
$$Q_k(t) = Q_k(0) + A_k(t) - S_k(T_k(t)) + \sum_{k' \in K} \Phi_{k'k}(S_{k'}(T_{k'}(t)))$$

Multi-Class Queueing Networks with Infinite Virtual Queues

MCQN with IVQ

Two types of classes
Standard Queues
Infinite Virtual Queues

$K_0 = \{1,2,3,5\}$
$K_\infty = \{4,6\}$



No exogenous arrivals
Nominal Input Rate

$$k \in K_0: \quad Q_k(t) = Q_k(0) - S_k(T_k(t)) + \sum_{k' \in K} \Phi_{k'k}(S_{k'}(T_{k'}(t))) \geq 0$$

$$k \in K_\infty: \quad Q_k(t) = Q_k(0) + \alpha_k\, t \quad - S_k(T_k(t))$$

A static production planning problem

Harrison static planning problem

$$\min_u \rho$$
$$Ru = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$$
$$Cu \leq \rho 1$$
$$u \geq 0$$

R input output matrix $= (I - P')\mu$

C constituency matrix $c_{i,k} = \begin{cases} 1 & k \in C(i) \\ 0 & else \end{cases}$

$\rho < 1$ sub-critical, $\rho > 1$ unstable, $\rho = 1$ critical

Static production planning problem

$$\max_{\alpha,u} \sum_{k \in K_\infty} w_k \alpha_k$$
$$Ru = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$$
$$Cu \leq 1$$
$$\alpha, u \geq 0$$

w are revenues, α are optimal production rates

u are optimal service allocations

$\rho_i = (Cu)_i = \sum_{k=1}^{K} c_{i,k} u_k$ is node i workload

$\tilde{\rho}_i = \sum_{i \notin K_\infty} c_{i,k}$ workload excluding IVQ

The question:

Static production planning problem

$$\max_{\alpha,u} \sum_{k\in K_\infty} w_k\alpha_k$$
$$Ru = \begin{bmatrix}\alpha\\0\end{bmatrix}$$
$$Cu \le 1$$
$$\alpha, u \ge 0$$

$\rho_i = (Cu)_i = \sum_{k=1}^{K} c_{i,k}u_k$ is node i workload

$\tilde{\rho}_i = \sum_{i\notin K_\infty} c_{i,k}$

$\rho = \max \rho_i \qquad \tilde{\rho} = \max \tilde{\rho}_i$

typically, $\rho = 1$

MCQN:

$\rho < 1$ network is stable under some policies, e.g. max pressure, but, becomes congested as $\rho \to 1$

$\rho = 1$ network is rate stable under some policies, e.g. max pressure

MCQN w IVQ:

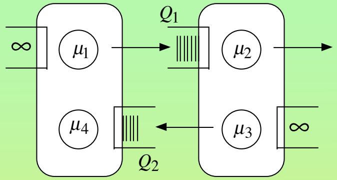
$\rho = 1$ but $\tilde{\rho} < 1$ Can it be stabilized ? will it remained uncongested ?

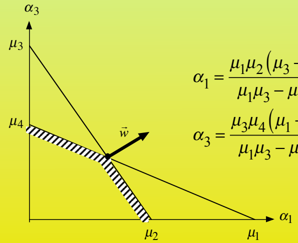
Answers very far away . . . , we discuss some examples . . .

example: Push pull system - KSRS with IVQ

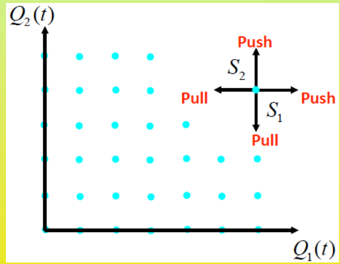
Static production planning problem

$$\max_{\alpha,u} \; w_1\alpha_1 + w_3\alpha_3$$
$$\begin{bmatrix}\mu_1 & 0 & 0 & 0\\ -\mu_1 & \mu_2 & 0 & 0\\ 0 & 0 & \mu_3 & 0\\ 0 & 0 & -\mu_3 & \mu_4\end{bmatrix}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\end{bmatrix} = \begin{bmatrix}\alpha_1\\0\\\alpha_3\\0\end{bmatrix}$$
$$\begin{bmatrix}1 & 0 & 1 & 0\\ 0 & 1 & 0 & 1\end{bmatrix}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\end{bmatrix} \le \begin{bmatrix}1\\1\\1\\1\end{bmatrix}$$
$$\alpha, u \ge 0$$





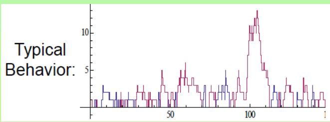
$$\alpha_1 = \frac{\mu_1\mu_2(\mu_3 - \mu_4)}{\mu_1\mu_3 - \mu_2\mu_4}$$
$$\alpha_3 = \frac{\mu_3\mu_4(\mu_1 - \mu_2)}{\mu_1\mu_3 - \mu_2\mu_4}$$

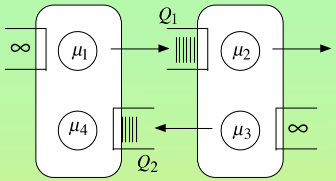


example: Push pull system - Stable policies

Case 1: $\mu_2 > \mu_1,\ \mu_4 > \mu_3$

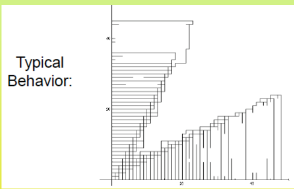
Pull priority, one queue at a time is empty

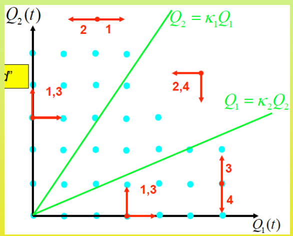




Case 2: $\mu_1 > \mu_2,\ \mu_3 < \mu_4$

Threshold policy, don't let a queue become empty





In the memoryless case, probabilities decay geometrically

Main tools for stability results

Establish that an "associated" deterministic fluid system is "stable"

The "framework" then implies the stochastic system is "stable"

Nice, since stability of deterministic system is easier to establish

This "fluid framework" was pioneered and exploited in the 90's by Dai, Meyn, Stolyar, Bramson, Williams, Chen . . .

Stochastic system vs fluid model

$k \in K_0: \quad Q_k(t) = Q_k(0) - S_k(T_k(t)) + \sum_{k' \in K} \Phi_{k'k}(S_{k'}(T_{k'}(t))) \geq 0$

$k \in K_\infty: \quad Q_k(t) = Q_k(0) + \alpha_k\, t \quad - S_k(T_k(t))$

$T_k(0) = 0,\ T_k \nearrow,\ \sum_{k \in C(i)} T_k(t) - T_k(s) \leq t - s$ *+ Policy implied equations*

Markov process $\quad X(t) = (Q(t), T(t), more)$

Fluid limits:

$$\bar{Q}^n(t) = \frac{Q^n(nt)}{n}, \quad \bar{T}^n(t) = \frac{T^n(nt)}{n},$$

$$\bar{Q}^{n_r}(t,\omega) \xrightarrow[r\to\infty]{} \bar{Q}(t), \quad \bar{T}^{n_r}(t,\omega) \xrightarrow[r\to\infty]{} \bar{T}(t)$$

Fluid model:

$$\bar{Q}(t) = \bar{Q}(0) - (I - P')[\mu]\bar{T}(t)$$

$\bar{T}(0) = 0, \quad \dot{\bar{T}}(t) \geq 0, \quad C\bar{T}(t) \leq 1$ *+ Policy implied equations*

Stability of queueing networks

Stochastic system: (MCQN as well as MCQN with IVQ)
Rate stable if $Q(t)/t \dashrightarrow 0$ as $t \dashrightarrow \infty$
Stable if $X(t)$ positive Harris recurrent (has stationary distribution)

fluid model
Stable if there is t_0 so that starting at $|q(t)|=1$, $q(t)=0, t > t_0$
Weakly stable if starting at $q(t)=0$ it stays $=0$.

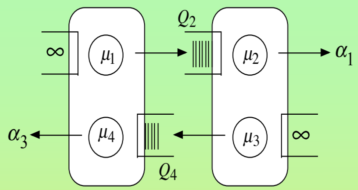
Theorem (Dai 95, holds for MCQN with IVQ) with technical condition:
"compact sets of states are petite" fluid stability implies MCQN is stable.
Fluid weak stability implies MCQN is rate stable.

Theorem (Dai &Lin 2004, Tassiulas, Stolyar) Under maximum pressure,
if $\rho \leq 1$ then fluid weakly stable, if $\rho < 1$ fluid is stable.

Observation: Extending definition of maximum pressure to
MCQN with IVQ, theorem continues to hold.

Push-pull fluid stability

Case 1: $\mu_2 > \mu_1,\ \mu_4 > \mu_3$

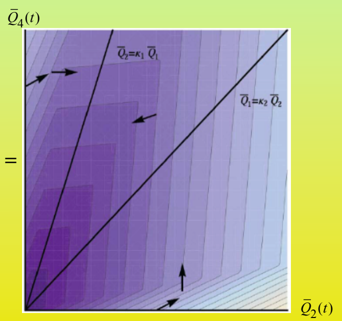


Lyapunov function: $\quad f(t) = \bar{Q}_2(t) + \bar{Q}_4(t)$

$\dot{f}(t) < \delta < 0, \quad t > 0$

Case 2: $\mu_1 > \mu_2,\ \mu_3 < \mu_4$
Piecewise linear Lyapunov function:

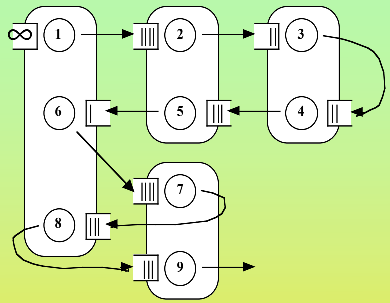
$f(t) =$



$\dot{f}(t) < \delta < 0, \quad t > 0$

Re-entrant line

For $\rho<1$, random exogenous input, stable under: LBFS, FBFS, Max-Press
Assume $\rho=1$ but $\tilde{\rho} < 1$ hence:
node 1 is bottleneck



Theorem: Under LBFS system is stable.

Cannot use $f(t) = \sum_{k=2}^{K} \bar{Q}_k(t)$

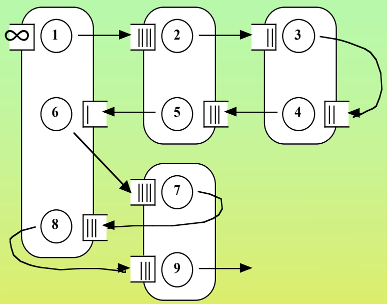
as Lyapunov function

Argument for proof:
assume total 1 unit initial fluid, processing time by node 1 is 1 per unit fluid
(1) while not empty, output at rate $\geq 1 + \delta > 1$ (by LBFS)
(2) by time 1 all original fluid cleared (service is head of the line)
(3) all output after time 1 requires 1 time unit per unit processing at node 1
(4) if not empty by T, output $\geq (1+\delta)$T, input $\leq \mu_1^{-1}$+T-1
(5) must be empty by some t_0 and stay empty

Re-entrant line

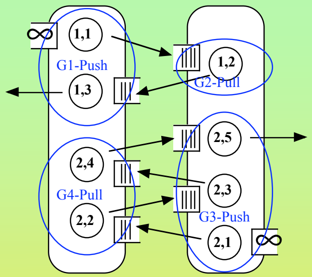
Further one can see:

- System is not stable under Max pressure, it is only rate stable (by simulation)
- Low priority to the IVQ and Max pressure for all other buffers is unstable.
- Low priority to the IVQ and FBFS is stable under some necessary and sufficient conditions on the parameters



Two Re-entrant lines

2 Re-entrant lines, starting with IVQ at 2 machines,
Buffers grouped as G1,G2,G3,G4.
Assume total work at G1,G3 is 1 per unit fluid, total work at G2,G4 is $r_2, r_4 < 1$ per unit fluid.



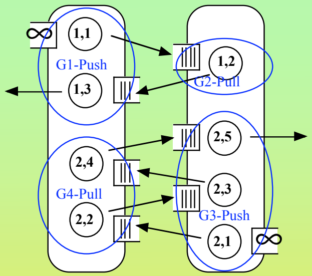
Policy: priority to G2 over G3, and G4 over G1
LBFS within each group.

Modes:
Both G2,G4 not empty, transient.
G4 empty, G2 not empty - line 2 frozen, work on line 1 only,
G2 empty, G3 not empty - line 1 frozen, work on line 2 only,
G2, G4 empty, G1+G3 not empty: work on both lines,
All empty: stays empty

Two Re-entrant lines

Theorem: Fluid stable under this policy.

Modes: $T_1(t)$ time on line 1 only
$T_2(t)$ time on line 2 only
$T_{1\&2}(t)$ sharaing both lines
Define: $\Theta_1(t), \Theta_2(t)$ average allocation of machine 1,2 to group G1,G3



Output inequalities - lower bound:

$$\bar{D}_{1,K_1}(t) \ge T_{L1}(t) + \Theta_1(t) T_{1\&2}(t)$$

$$\bar{D}_{2,K_2}(t) \ge T_{L2}(t) + \Theta_2(t) T_{1\&2}(t)$$

$$\bar{D}_{1,K_1}(t) + \bar{D}_{2,K_2}(t) \ge \left(T_{L1}(t) T_{L2}(t) + (\Theta_1(t) + \Theta_2(t)) T_{1\&2}(t)\right)(1+\delta)$$

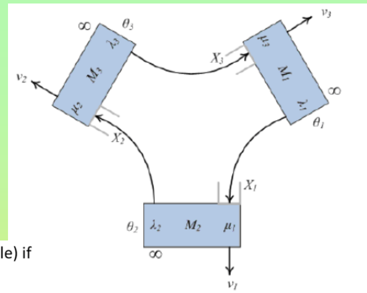
Input inequalities - upper bound:
more chalk,

Hence: system emty by t_0 and stays empty.

A ring system

M machines, M product lines each through two successive machines.
processing rate at the IVQ is 1 (w.l.g),
at the second queue it is ρ_i

using **pull priority** we have



Theorem (Guo, Lefeber, N., Weiss, Zhang): $X(t)$ is PHR (stable) if

(i) $\rho_i < 1 \;\; i = 1, \dots, M$

(ii) When M is odd and $1 < \rho_i \;\; i = 1, \dots, M$ and

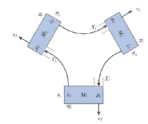
$$\Delta = \sum_{i=1}^{M} c_i \left(\frac{M-1}{2} (\rho_i - 1) - 1 \right) < 0$$

$$c_i = (\dots(\rho_{i-1} - 1)\rho_{i-2} + 1)\rho_{i-3} - 1)\rho_{i-4} + \cdots + (-1)^{1+(M-1)/2})\rho_{i+1} + (-1)^{(M-1)/2}$$

Specifically when $\rho_i = \rho$ then $\Delta<0$ iff $\rho < 1 + \frac{2}{M-1}$

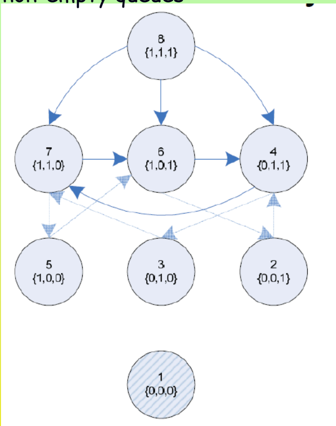
And when M=3,

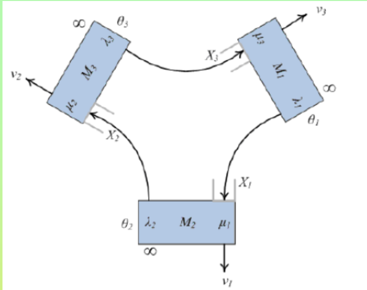
$$\Delta = \prod_{i=1}^{3} (\rho_i - 1) - 1$$



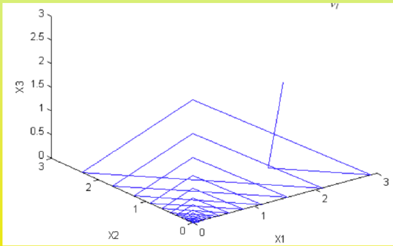
A ring system M=3, pull priority

Graph of modes according to non-empty queues





fluid trajectory cycles in iff $\Delta<0$



Stability of the stochastic systems - petiteness of compacts

Stability of the fluid implies stability of the stochastic system,

However: a technical condition is required:

In the Markov process X(t)=(Q(t),T(t),more)
Compact sets of states are petite.

This needs some sufficient conditions:
for standard MCQN: work conservation
+ input interarrivals have unbounded support and are spread out

For MCQN with IVQ:
weak pull priority - if not all empty, always work on some non-IVQ
+ input interarrivals have unbounded support and are spread out

For our threshold policy in the push-pull system we don't know conditions

Fluid and Diffusion approximation

For simplicity, consider MCQN with deterministic routing

Assume i=1, . . ., I product lines, each starting with IVQ at node i,
with steps (i,1), . . . , (i,K_i),
processing times have mean m_{ik} and squared c.o.v d_{ik}^2

Assume under some policy we have full utilization and stable queues.

We obtain fluid and diffusion approximation to the cucmulative allocated processing times and the departure processes

Fluid approximation

Fluid scaling: $\bar{T}_{i,k}{}^{(n)}(t)=\frac{T_{i,k}(nt)}{n},\quad \bar{Q}_{i,k}{}^{(n)}(t)=\frac{Q_{i,k}(nt)}{n},\quad \bar{D}_{i,k}{}^{(n)}(t)=\frac{D_{i,k}(nt)}{n},$

static production planning

$$\max_\alpha \sum_{i=1}^{I} w_i\alpha_i$$
$$\sum_{(i',k)\in C(i)}^{I} m_{i',k}\alpha_{i'}\le 1,\quad i=1,\ldots,I$$
$$\alpha\ge 0$$

Production on line i at rate α_i
all steps are at rate α_i
departure processes at rate α_I

Queues are stable

Fluid approximation $\bar{T}_{i,k}(t)=m_{i,k}\alpha_i t,\quad \bar{Q}_{i,k}(t)=0,\quad \bar{D}_{i,K_i}(t)=\alpha_i t$

The actual system is approximated by this deterministic fluid
however stochastic deviation accumulate at a rate of $\sqrt{n}$
we approximate the deviations by diffusion scaling

Diffusion approximation

Diffusion scaling: $\hat{T}_{i,k}{}^{(n)}(t) = \frac{T_{i,k}(nt) - n\bar{T}_{i,k}(t)}{\sqrt{n}}, \quad \frac{n!}{r!(n-r)!} = \frac{Q_{i,k}(nt)}{\sqrt{n}},$

$$\hat{D}_{i,k}{}^{(n)}(t) = \frac{D_{i,k}(nt) - n\bar{D}_{i,k}(t)}{\sqrt{n}}, \quad \hat{A}_i(t) = \frac{A_i(nt) - nA_i(t)}{\sqrt{n}}$$

Standard MCQN:

if $\rho<1$ and queues are stable, $\hat{Q}_{i,k}{}^{(n)}(t) \to 0$

and $\hat{A}_i{}^{(n)}(t) \Rightarrow \hat{A}_i(t) \sim BM\ independent, \quad \hat{A}_i(t) = \hat{D}_{i,1}(t) = \cdots = \hat{D}_{i,K_i}$

the processing times have no effect on output - just copies input

If $\rho \approx 1$ queues are unstable, and output depends on input and service

$$\hat{A}_i{}^{(n)}(t) \Rightarrow \hat{A}_i(t) \sim BM, \quad \hat{Q}_{i,k}{}^{(n)}(t) \Rightarrow RBM,$$

$$\hat{D}_{i,k} \Rightarrow combination\ (Iglehart\ Whitt, 71)$$

Diffusion approximation

MCQN with IVQ:

If $\rho=1$ but queues are stable, time allocation to IVQ absorbs the variability

$$D_{i,k}(t) = S_{i,k}(T_{i,k}(t)), \quad Q_{i,k}(t) = D_{i,k-1}(t) - D_{i,k}(t), \quad \sum_{(i',k)\in C(i)} T_{i',k}(t) = t$$

$$\hat{D}_{i,k}^{(n)}(t) = \hat{S}_{i,k}^{(n)}(\bar{T}_{i,k}(t)) + \mu_{i,k}\hat{T}_{i,k}^{(n)}(t),$$

$$\hat{Q}_{i,k}^{(n)}(t) = \hat{D}_{i,k-1}^{(n)}(t) - \hat{D}_{i,k}^{(n)}(t),$$

$$\hat{T}_{i,1}^{(n)}(t) = -\sum_{\substack{(i',k)\in C(i)\\(i',k)\neq(i,1)}} \hat{T}_{i',k}^{(n)}(t)$$

$$\hat{Q}_{i,k}^{(n)}(t) \Rightarrow 0, \quad \hat{T}_{i,1}^{(n)}(t) \Rightarrow BM, \quad \hat{D}_{i,K_i}^{(n)}(t) \Rightarrow BM$$

we get exact expressions for the covariances - typically negative correlations

In summary:

- We formulated a new Static Production Planning problem
- Solution will typically imply $\rho=1$, hence MCQN congested
- For MCQN with IVQ, we may have $\rho=1$ but $\tilde{\rho}<1$
- Question: Can this be stable, and remain uncongested
- Old example - KSRS network with IVQ
- Extensions: we show stability under pull priority for:
  - re-entrant line, 2 re-entrant lines, ring network
- We derive diffusion approximation,
- New difficulties in ensuring that compact sets are petite