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

YongJiang Guo¹, Erjen Lefeber², Yoni Nazarathy³, Gideon Weiss⁴, Hanqin Zhang⁵


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

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

A static production planning problem

Harrison static planning problem

\[
\begin{align*}
\min_u \rho \\
Ru = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \\
Cu \leq \rho 1 \\
u \geq 0
\end{align*}
\]

\[ R_{u} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \]

\[ C = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \]

\[ c_{i,k} = \begin{cases} 1 & k \in C(i) \\ 0 & \text{else} \end{cases} \]

\[ \rho < 1 \text{ sub-critical, } \rho > 1 \text{ unstable, } \rho = 1 \text{ critical} \]

Static production planning problem

\[
\begin{align*}
\max_{\alpha, u} \sum_{k=1}^{K} w_k \alpha_k \\
Ru = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \\
Cu \leq 1 \\
\alpha, u \geq 0
\end{align*}
\]

\[ w \text{ are revenues, } \alpha \text{ are optimal production rates} \]

\[ u \text{ are optimal service allocations} \]

\[ \rho_i = (Cu_i) = \sum_{k=1}^{K_{\infty}} c_{i,k} \tilde{u}_k \text{ is node } i \text{ workload} \]

\[ \tilde{\rho}_i = \sum_{i \in K_{\infty}} c_{i,k} \text{ workload excluding IVQ} \]
The question:
Static production planning problem

\[
\max_{\alpha,\mu} \sum_{k \in K_m} w_k \alpha_k \\
\rho_i = (C_{\alpha}) = \sum_{k=1}^{K} c_{i,k} \mu_k \quad \text{is node } i \text{ workload} \\
R_i = \alpha_k \\
C_{\alpha} \leq 1 \\
\alpha, \mu \geq 0
\]

\[
\rho = \max \rho_i \quad \beta = \max \tilde{\rho}_i \\
\text{typically, } \rho = 1
\]

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 \[\rho < 1\]. Can it be stabilized? Will it remained uncongested?
Answers very far away . . ., we discuss some examples . . .

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: \(\mu_2 > \mu_3, \mu_4 < \mu_1\)
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

$k \in K_0$: \[ Q_k(t) = Q_k(0) - S_k(T_k(t)) + \sum_{k \in K} \Phi_k S_k(T_k(t)) \geq 0 \]

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

\[ T_k(0) = 0, \quad T_k \not\rightarrow \sum_{k \in C(i)} T_k(s) \leq t - s \quad + \text{Policy implied equations} \]

Markov process \[ X(t) = (Q(t), T(t), \text{more}) \]

Fluid limits:

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

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

Fluid model:

\[ \bar{Q}(t) = \bar{Q}(0) - (I - P^n)[\mu]T(t) \]

\[ \bar{T}(0) = 0, \quad \bar{T}(t) \geq 0, \quad \bar{C}T(t) \leq 1 \quad + \text{Policy implied equations} \]

Push-pull fluid stability

Case 1: \( \mu_3 > \mu_1, \mu_4 > \mu_3 \)

Lyapunov function:

\[ f(t) = \bar{Q}_2(t) + \bar{Q}_4(t) \]

\[ f(t) = \bar{Q}_2(t) + \bar{Q}_4(t) \]

Case 2: \( \mu_1 > \mu_2, \mu_4 < \mu_4 \)

Piecewise linear Lyapunov function:

\[ f(t) = \left\{ \begin{array}{ll}
Q(t) & \text{if } t < \delta \\
0 & \text{if } t \geq \delta \end{array} \right. \]

Stability of queueing networks

Stochastic system: (MCQN as well as MCQN with IVQ)

Rate stable if \( Q(t)/T \to 0 \) as \( t \to \infty \)

Stable if \( \lambda(t) \) positive Harris recurrent (has stationary distribution)

Fluid model

Stable if there is \( \mu \) so that starting at \( \{q(t) = 1, \quad q(t) = 0, \quad t > \delta \} \)

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 '04, 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.

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 \( \leq \mu_1\delta + T-1 \)

(5) must be empty by some \( t_0 \) and stay empty
Re-entrant line

Further one can see:

• System is not stable under \(\text{Max pressure}\), it is only rate stable (by simulation)

• Low priority to the IVQ and \(\text{Max pressure}\) 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

A ring system

\(M\) machines, \(M\) product lines each through two successive machines, processing rate at the IVQ is 1 (w.l.o.g), at the second queue it is \(\rho_i\), using pull priority we have

Theorem (Guo, Lefever, N., Weiss, Zhang): \(X(t)\) is PHR (stable) if
(i) \(\rho_i < 1\) \(i = 1, ..., M\)
(ii) When \(M\) is odd and 1 < \(\rho_i\) \(i = 1, ..., M\) and
\[\Delta = \sum_{i=1}^{M} \left( \frac{M-1}{M} (\rho_i - 1) - 1 \right) < 0\]

\(\rho = (-\rho_{m-1} - \rho_{m-2} - \rho_{m-3} - \cdots - \rho_{1})^{2} + (-1)^{m-2} \rho_{m-2} + (-1)^{m-3} \rho_{m-3}\)

Specifically when \(\rho_i = \rho\) then \(\Delta = \) if \(\rho < 1 + \frac{2}{M-1}\)
And when \(M=3\),
\[\Delta = \frac{3}{2} (\rho_1 - 1) - 1\]
**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 $X(t) = (Q(t), T(t), \text{more})$

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 $i=1, \ldots, I$ product lines, each starting with IVQ at node $i$, with steps $(i,1), \ldots, (i,K_i)$, 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 cumulative allocated processing times and the departure processes

**Fluid approximation**

Fluid scaling: $T_{ik}^{(n)}(t) = \frac{T_{ik}(nt)}{n}$, $Q_{ik}^{(n)}(t) = \frac{Q_{ik}(nt)}{n}$, $D_{ik}^{(n)}(t) = \frac{D_{ik}(nt)}{n}$.

Static production planning

$max_{\alpha} \sum_{i=1}^{I} w_i \alpha_i$

Production on line $i$ at rate $\alpha_i$

all steps are at rate $\alpha_i$

departure processes at rate $\alpha_i$

$\sum_{(i',k) \in C(i)} m_{i',k} \alpha_i \leq 1$, $i = 1, \ldots, I$

$\alpha \geq 0$

Queues are stable

Fluid approximation $T_{ik}^{(n)}(t) = m_{ik} \alpha_i t$, $Q_{ik}^{(n)}(t) = 0$, $D_{ik}^{(n)}(t) = \alpha_i t$

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

Diffusion scaling:
$$\hat{T}_{i,k}(t) = \frac{T_{i,k}(m) - nT_{i,k}(t)}{\sqrt{n}}, \quad \hat{Q}_{i,k}(m) = \frac{Q_{i,k}(m)}{\sqrt{n}}.$$  
$$\hat{D}_{i,k}(n)(t) = \frac{D_{i,k}(m) - nD_{i,k}(t)}{\sqrt{n}}, \quad \hat{A}_i(t) = \frac{A_i(m) - nA_i(t)}{\sqrt{n}}.$$  

Standard MCQN:
if $\rho<1$ and queues are stable, $\hat{Q}_{i,k}(m)(t) \to 0$ and $\hat{A}_i(t) \to 0$ independent, $\hat{A}_i(t) = \hat{D}_i(t) = \cdots = \hat{D}_{i,K}$.  

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

If $\rho = 1$ queues are unstable, and output depends on input and service:  
$$\hat{A}_i(t) \to \hat{A}_i(t) \sim BM, \quad \hat{Q}_{i,k}(m)(t) \to RBM,$$  
$$\hat{D}_{i,k} \to combination (Iglehart Whitt, 71)$$

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 $\hat{\rho} < 1$
- Question: Can this be stable, and remain uncongested
- Old example - KSR 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

Diffusion approximation

MCQN with IVQ:
If $\rho = 1$ but queues are stable, time allocation to IVQ absorbs the variability
$$D_{i,k}(t) = S_{i,k}(T_{i,k}(t)), \quad \hat{Q}_{i,k}(m) = D_{i,k-1}(t) - D_{i,k}(t), \quad \sum_{(i',K)\in C(i)} T_{i',k}(t) = t$$  
$$\hat{D}_{i,k}(n)(t) = S_{i,k}(\hat{T}_{i,k}(t)) + \mu_{i,k}\hat{\tau}_{i,k}(t),$$  
$$\hat{Q}_{i,k}(m)(t) = \hat{D}_{i,k}(n)(t) - \hat{D}_{i,k}(n)(t),$$  
$$\hat{\tau}_{i,k}(t) = -\sum_{(i',K)\neq C(i)} \hat{\tau}_{i,k}(t)$$  
$$\hat{Q}_{i,k}(m)(t) \to 0, \quad \hat{\tau}_{i,k}(t) \to BM, \quad \hat{D}_{i,k}(n)(t) \to BM$$

we get exact expressions for the covariances - typically negative correlations