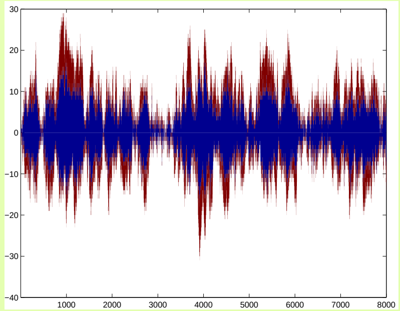
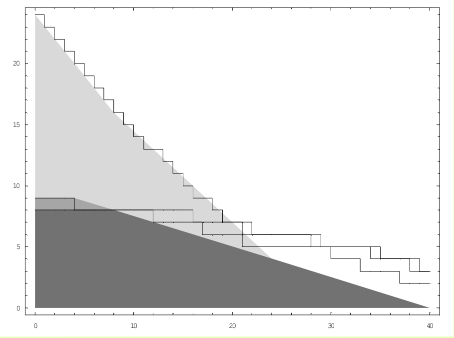
Queueing Networks with Infinite Virtual Queues

An Example, An Application and a Fundamental Question



Yoni Nazarathy
(Supervisor: Prof. Gideon Weiss)
University of Haifa



STUDENTS PROBABILITY DAY
Weizmann Institute of Science
March 28, 2007

Multi-Class Queueing Networks (Harrison 1988, Dai 1995,…)

Queues

$$\mathrm{K} = \{1,\ldots,K\}$$

$$\{Q_k(t), t > 0\}$$

Initial Queue Levels

$$Q_k(0) \quad k \in \mathrm{K}$$

Routes

$$\Phi_{kk'}(n) \nearrow \quad k,k' \in \mathrm{K}$$

$$\lim_{n\to\infty} \frac{\Phi_{kk'}(n)}{n} = P_{kk'}$$

Servers

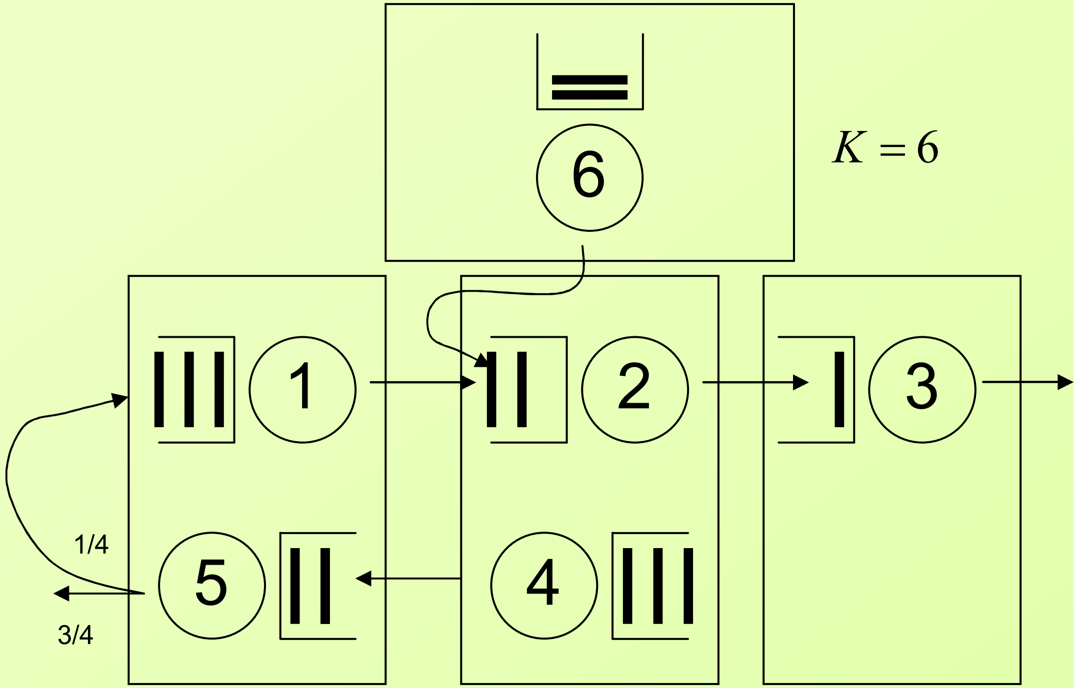
$$\mathrm{I} = \{1,\ldots,I\}$$

$$A_{I\times K} = \{A_{ik}\} \quad A_{ik} \in \{0,1\}$$

Processing Durations

$$S_k(t) \nearrow \quad k \in \mathrm{K}$$

$$\lim_{t'\to\infty} \frac{S_k(t)}{t} = \mu_k = \frac{1}{m_k}$$



Network Dynamics

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

Resource Allocation (Scheduling)

$$T_k(t) \nearrow$$

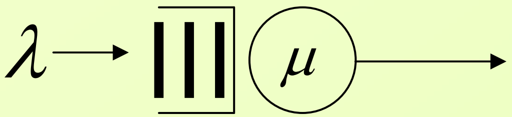
$$T_k(0) = 0 \qquad \sum_{k\in \mathrm{K}} A_{ik}\left(T_k(t) - T_k(s)\right) \le t - s \quad s < t$$

$$T_k(t) \nearrow \quad \textit{only when} \quad Q_k(t) > 0$$

INTRODUCING: Infinite Virtual Queues

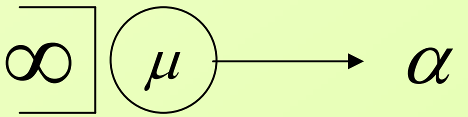
Regular Queue

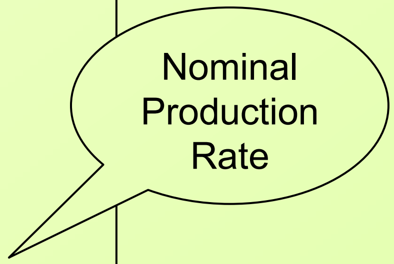
$$Q_k(t) : \mathbb{R} \to \{0,1,2,\dots\}$$



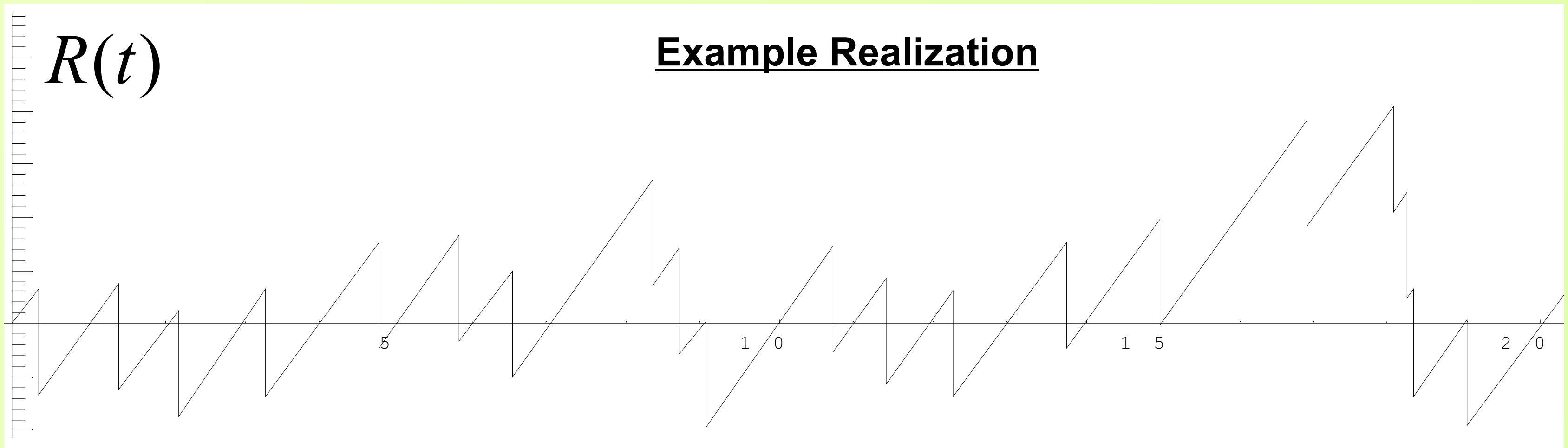
Infinite Virtual Queue

$$Q_k(t) = \infty \quad \forall\, t$$





Relative Queue Length: $$R(t) = R(0) - S(T(t)) + \alpha\, t$$



MCQN+IVQ

Queues

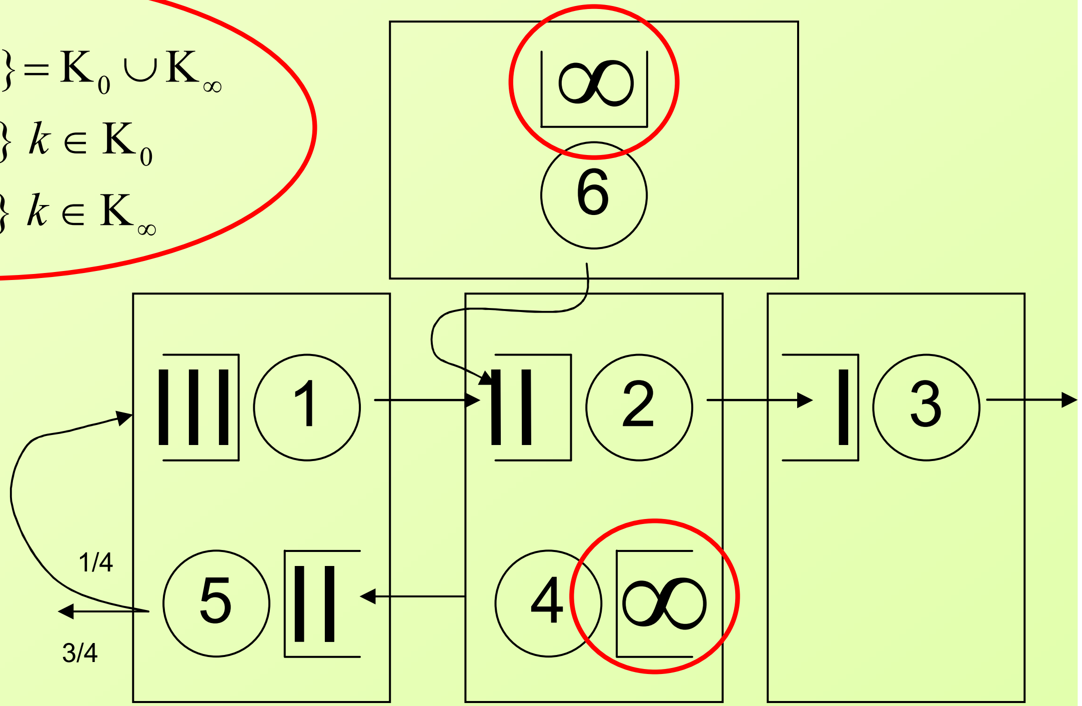
~~$\mathrm{K} = \{1,\ldots,K\}$~~

~~$\{Q_k(t), t > 0\}$~~

$$\mathrm{K} = \{1,\ldots,K\} = \mathrm{K}_0 \cup \mathrm{K}_\infty$$

$$\{Q_k(t), t > 0\} \;\; k \in \mathrm{K}_0$$

$$\{R_k(t), t > 0\} \;\; k \in \mathrm{K}_\infty$$



Initial Queue Levels

~~$Q_k(0) \;\; k \in \mathrm{K}$~~

$$Q_k(0) \;\; k \in \mathrm{K}_0$$

Routes

$\Phi_{kk'}(n) \nearrow$ $\quad k \in \mathrm{K} \quad k' \in \mathrm{K}_0$

$$\lim_{n\to\infty} \frac{\Phi_{kk'}(n)}{n} = P_{kk'}$$

Servers

$$\mathrm{I} = \{1,\ldots,I\}$$

$$A_{I\times K} = \{A_{ik}\} \quad A_{ik} \in \{0,1\}$$

Processing Durations

$$S_k(t) \nearrow \quad k \in \mathrm{K}$$

$$\lim_{t'\to\infty} \frac{S_k(t)}{t} = \mu_k = \frac{1}{m_k}$$

Network Dynamics

$$Z_k(t) = \begin{cases} Q_k(t) & = & Q_k(0) - S_k(T_k(t)) + \sum_{k' \in \mathrm{K}} \Phi_{k'k}(S_{k'}(T_{k'}(t))) \geq 0 & k \in K_0 \\ R_k(t) & = & R_k(0) - S_k(T_k(t)) + \alpha_k t & k \in K_\infty \end{cases}$$

Nominal Production Rates

Resource Allocation (Scheduling)

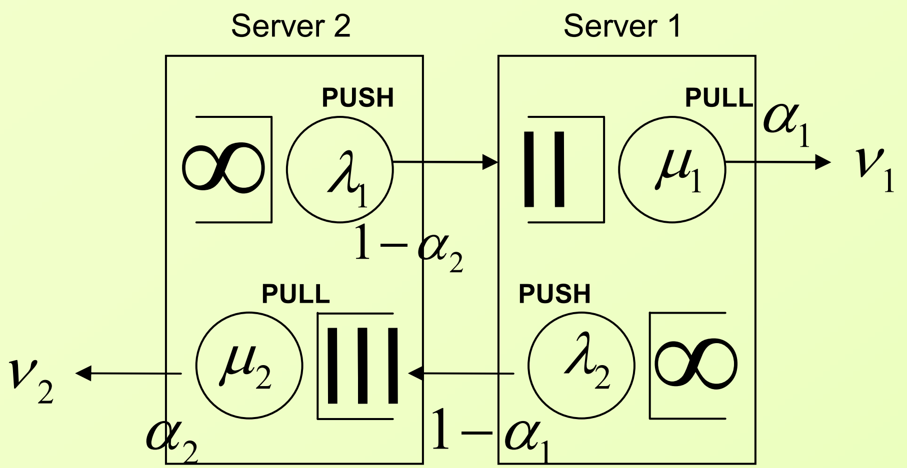
$$T_k(t) \nearrow$$

$$T_k(0) = 0 \quad \sum_{k \in \mathrm{K}} A_{ik} \left(T_k(t) - T_k(s) \right) \leq t - s \quad s < t$$

$$T_k(t) \nearrow \quad only\ when \quad Q_k(t) > 0 \quad for\ k \in \mathrm{K}_0$$

An Example

A Push-Pull Queueing System (Weiss, Kopzon 2002,2006)



$\mu_1 > \lambda_1, \mu_2 > \lambda_2$ "Inherently Stable"

or

$\lambda_1 > \mu_1, \lambda_2 > \mu_2$ "Inherently Unstable"

α_i – Proportion of time server i allocates to "Pulling"

Require Full Utilization

$$\lim_{t\to\infty} \frac{Z(t)}{t} = 0$$

Require Rate Stability

Fluid Solution:

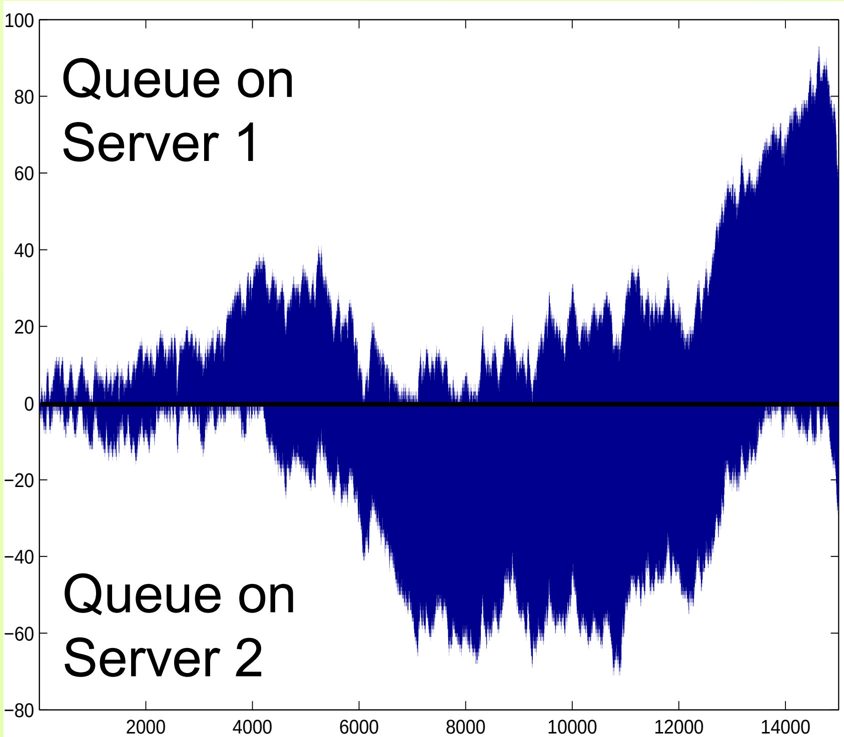
$$\alpha_1\mu_1 = (1-\alpha_2)\lambda_1$$
$$\alpha_2\mu_2 = (1-\alpha_1)\lambda_1 \Rightarrow$$

$$\alpha_1 = \frac{\lambda_1(\mu_2-\lambda_2)}{\mu_1\mu_2-\lambda_1\lambda_2}$$
$$\alpha_2 = \frac{\lambda_2(\mu_1-\lambda_1)}{\mu_1\mu_2-\lambda_1\lambda_2} \Rightarrow$$

$$\nu_1 = \mu_1\alpha_1$$
$$\nu_2 = \mu_2\alpha_2$$

Maximum Pressure (Dai, Lin 2005)

- Max-Pressure is a rate stable policy (even when ρ=1).
- Push-Pull acts like a ρ=1 System.
- As Proven by Dai and Lin, Max-Pressure is rate stable.
- But for the Push-Pull system Max-Pressure is not Positive Recurrent:

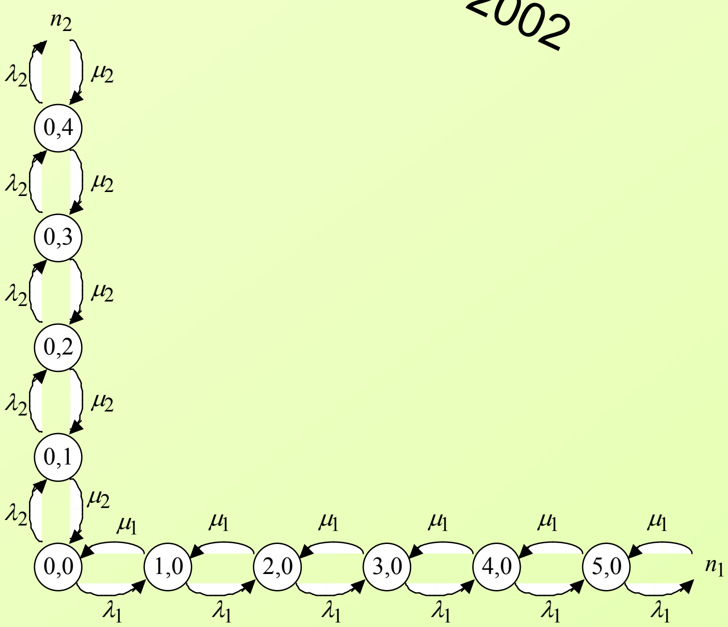


Positive Recurrent Policies Exist!!!

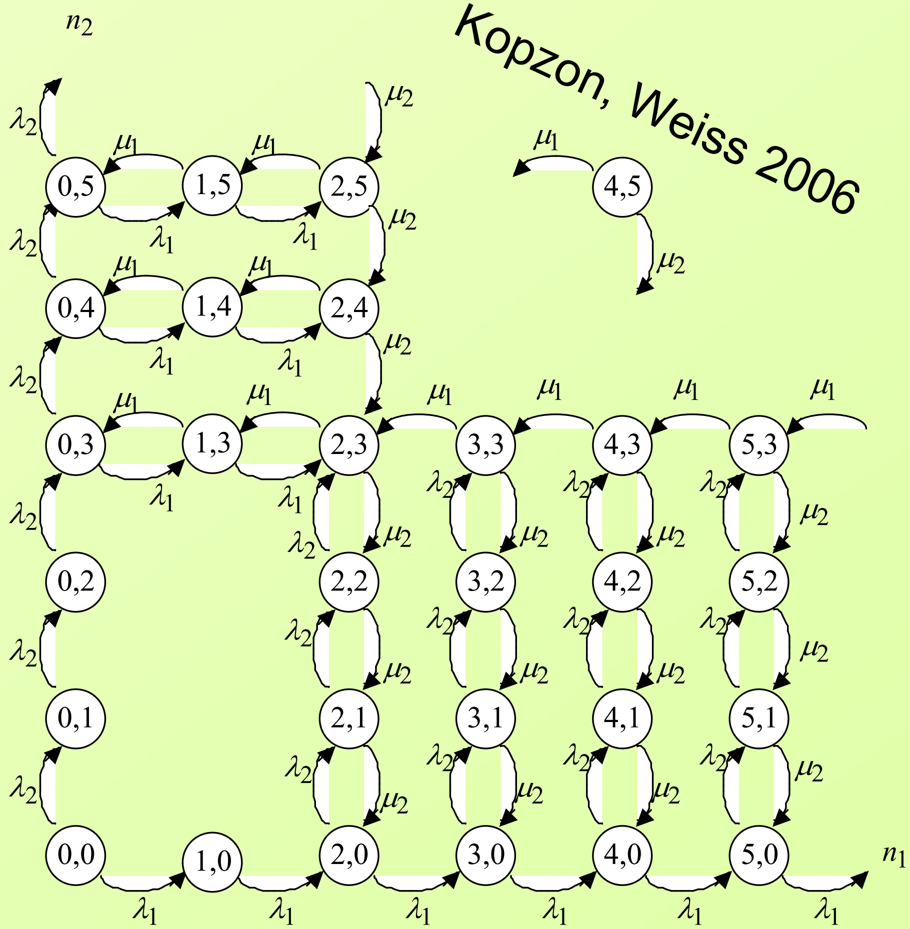
$$\mu_1 > \lambda_1, \mu_2 > \lambda_2$$

$$\lambda_1 > \mu_1, \lambda_2 > \mu_2$$

Kopzon, Weiss 2002



Kopzon, Weiss 2006



An Application

Finite Horizon Control of MCQN → Solution is intractable

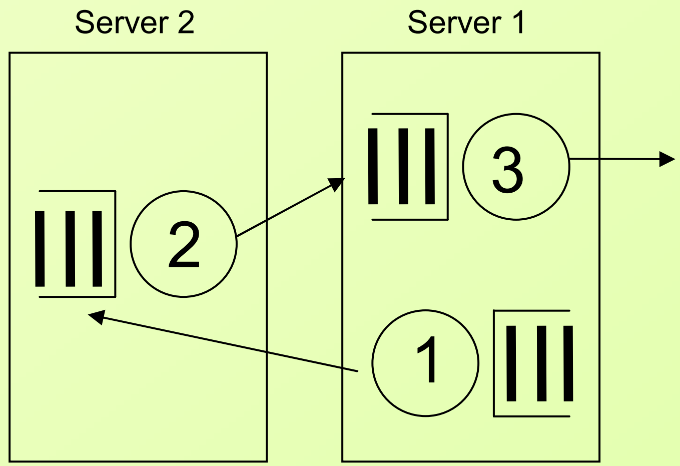
$$Min \int_0^T \sum_{k=1}^{3} Q_k(t)dt$$

Weiss, Nazarathy 2007

<u>Approximation Approach:</u>

1) Approximate the problem using a fluid system.
2) Solve the fluid system (SCLP).
3) Track the fluid solution on-line (Using MCQN+IVQs).
4) Under proper scaling, the approach is asymptotically optimal.

Fluid formulation



$$\min \int_0^T q_1(t) + q_2(t) + q_3(t)\,dt$$

s.t.

$$q_1(t) = q_1(0) - \int_0^t u_1(s)\,ds$$

$$q_2(t) = q_2(0) + \int_0^t u_1(s)\,ds - \int_0^t u_2(s)\,ds$$

$$q_3(t) = q_3(0) + \int_0^t u_2(s)\,ds - \int_0^t u_3(s)\,ds$$

$$\frac{1}{\mu_1} u_1(t) + \frac{1}{\mu_3} u_3(t) \le 1 \qquad t \in (0,T)$$

$$\frac{1}{\mu_2} u_2(t) \le 1$$

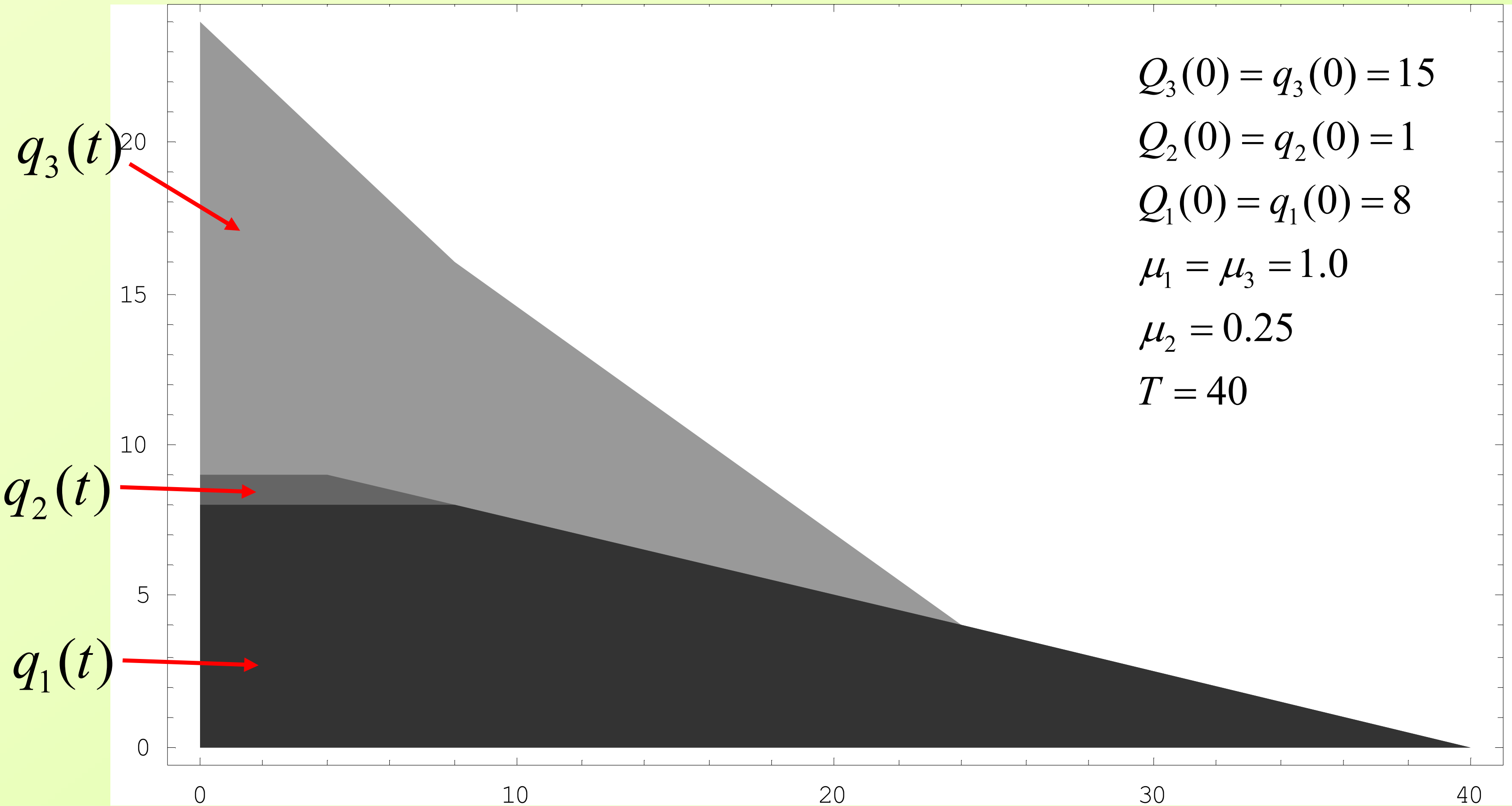
$$u, q \ge 0$$

This is a Separated Continuous Linear Program (SCLP)

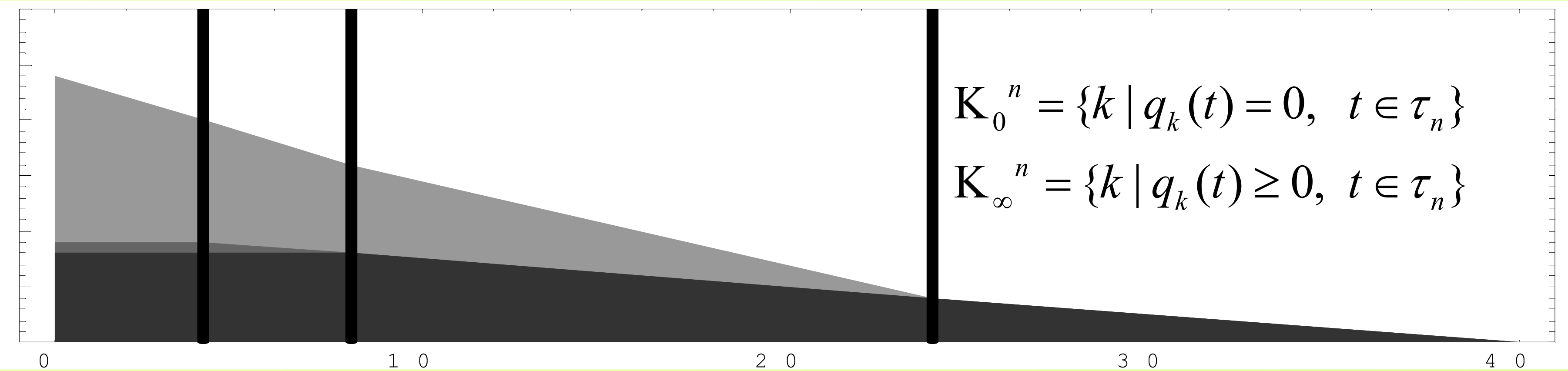
Fluid solution

- SCLP – Bellman, Anderson, Pullan, Weiss.
- Piecewise linear solution.
- Simplex based algorithm, finds the optimal solution in a finite number of steps (Weiss).

The Optimal Solution:



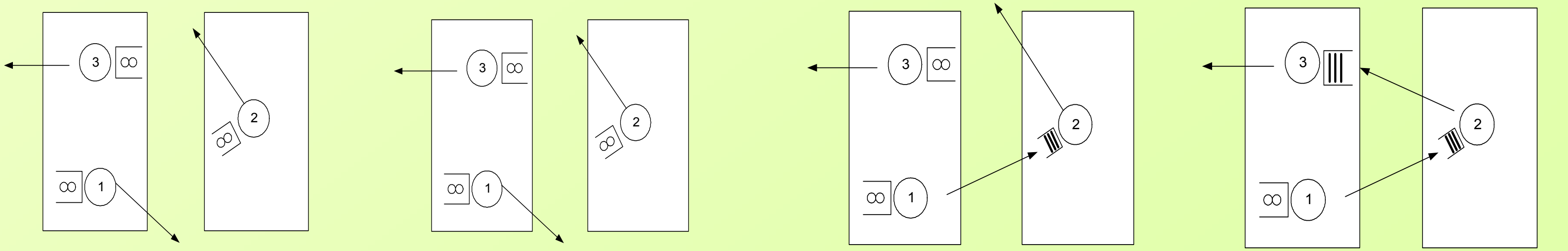
4 Time Intervals



$$K_0^n = \quad \{\} \qquad \{\} \qquad \{2\} \qquad \{2,3\}$$

$$K_\infty^n = \{1,2,3\}\ \{1,2,3\} \qquad \{1,3\} \qquad \{1\}$$

For each time interval, set a MCQN with Infinite Virtual Queues.



$$\alpha = \begin{pmatrix}0 & 0 & 1\end{pmatrix} \qquad \begin{pmatrix}0 & 1/4 & 1\end{pmatrix} \qquad \begin{pmatrix}1/4 & - & 3/4\end{pmatrix} \qquad \begin{pmatrix}1/4 & - & -\end{pmatrix}$$

Now Control the MCQN+IVQ Using a Rate Stable Policy

Maximum Pressure (Dai, Lin) is such a policy, even when $\rho=1$

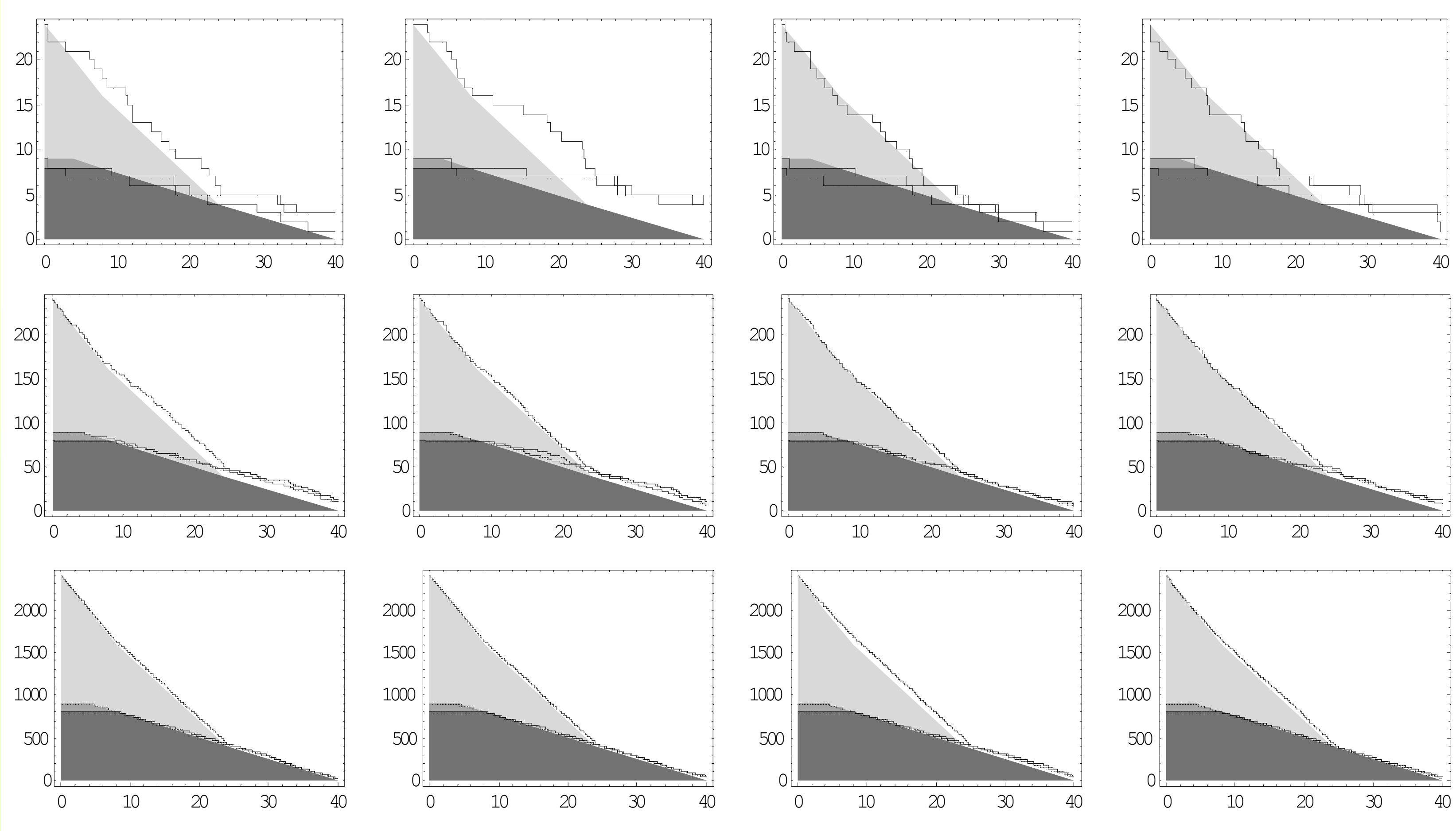
Example realizations, N={1,10,100}

seed 1 | seed 2 | seed 3 | seed 4

$N = 1$

$N = 10$

$N = 100$



A Fundamental Question

Is there a characterization of MCQN+IVQs that allows:

- Full Utilization of all the servers that have an IVQ.
- Stability of all finite queues.
- Proportional equality among production streams.

?

Thank You