Fluid models and Stability of Multiclass Queueing Networks

Jim Dai

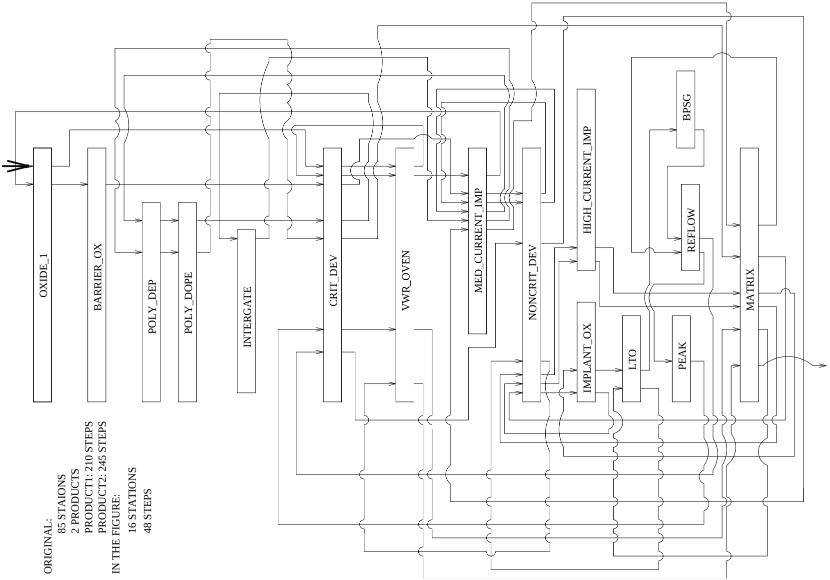
School of Industrial and Systems Engineering
Georgia Institute of Technology

Joint work with John Hasenbein and John Vande Vate
April 28, 2009

Outline

- Part I: Importance of an operational policy in a wafer fab
- Part II: Fluid models and their stability
- Part III: For a queueing network operating under a service policy, its stability region can depend on
  - its distributions, not just means;
  - the preemption mechanism.

Flow in a Wafer Fab



Performance Measures

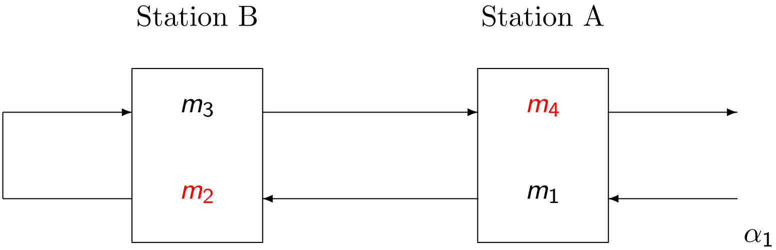
First order ones:

- Throughput: rate at which entities leave a system
- Utilization

Second order ones:

- Cycle time: processing times plus waiting time of an entity; average and variance of cycle time

An Re-Entrant Line (Lu-Kumar network)

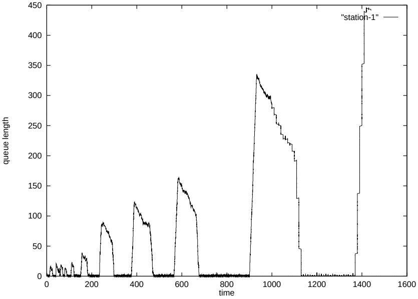


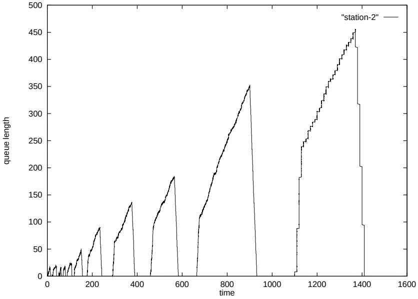
$$\alpha_1 = 1, \quad m_1 = .2, \quad m_2 = .6, \quad m_3 = .1, \quad m_4 = .6.$$

Operational policy: LBFS at Station A, FBFS at Station B.

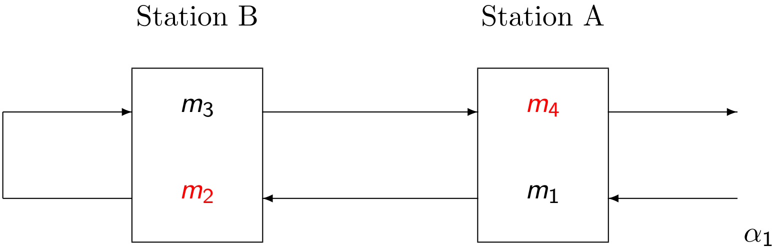
$$\rho_1 = 80\%, \quad \rho_2 = 70\%.$$

WIP Levels at Two Stations



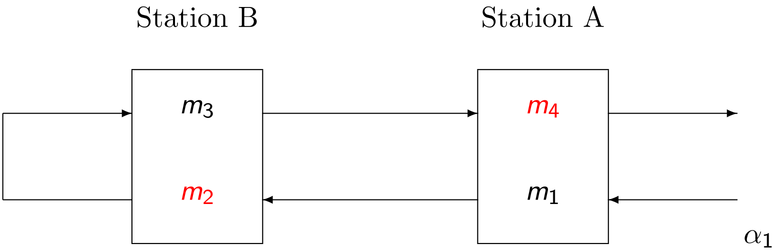


Utilization and Cycle Time



# departed	100	1,000	10,000	100,000
cycle time	15.8	183.7	1740.2	17043.7
utilization A	0.65	0.60	0.61	0.65
utilization B	0.59	0.68	0.67	0.61

Theorem

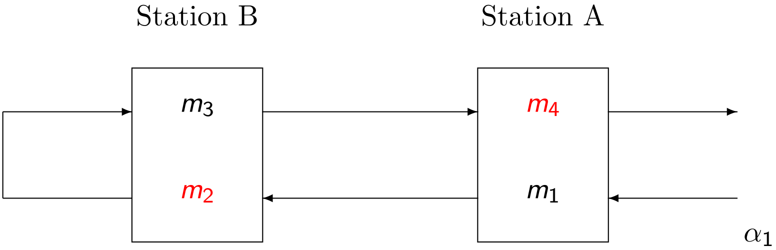


Under the operational policy, the system is "stable" if and only if

$$\rho_1 = \alpha_1(m_1 + m_4) \leq 1,$$
$$\rho_2 = \alpha_1(m_2 + m_3) \leq 1,$$
$$\rho_v = \alpha_1(m_2 + m_4) \leq 1.$$

Dai and Vande Vate, *Operations Research*, 721–744, 2000.

Maximum Throughput

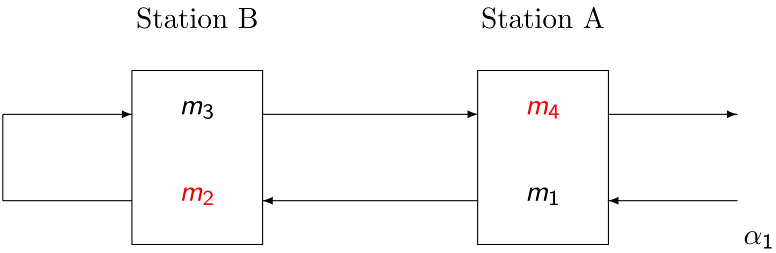


If the static-buffer-priority policy is used, the maximum throughput is

$$\min\left\{\frac{1}{m_1+m_4}, \quad \frac{1}{m_2+m_3}, \quad \frac{1}{m_2+m_4}\right\}.$$

In our example, the maximum throughput is 0.83 instead of 1.25, a 50% relative difference.

Virtual Station



Lemma (Harrison-Nguyen 95, Dumas)

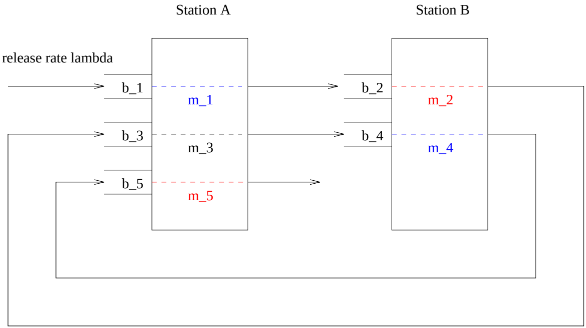
Under the operational policy,

$$Z_2(t)Z_4(t) = 0 \quad \textit{for all } t \geq 0$$

if $Z_2(0)Z_4(0) = 0$. *Thus, classes* $\{2, 4\}$ *form a virtual station.*

If $\rho_v = \alpha_1(m_2 + m_4) > 1$, with probability one, the total number of jobs goes to infinity.

Virtual Station



If the red buffers have higher priority than the blue buffers, jobs in buffer 2 and buffer 5 can *never* be processed simultaneously. Mathematically,

$$Z_2(t)Z_5(t) = 0, \quad t \geq 0 \quad \text{if} \quad Z_2(0)Z_5(0) = 0.$$

Steps 2 and 5 form a virtual station under the priority dispatch policy.

Virtual Bottleneck?

Phenomenon:

- WIP is high, and
- bottleneck machines are underutilized

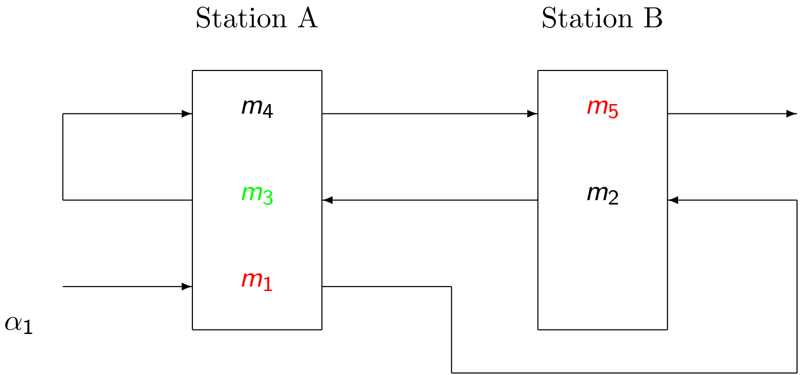
Efficient and Inefficient Policies

- Inefficient policies:
  - First-in-first-out (FIFO) (Bramson 1994, Seidman 1994)
  - Static buffer priority (Lu-Kumar 1992)
  - Shortest processing time first
  - Shortest remaining processing time first
  - Exhaustive service (Kumar-Seidman 1990)
  - ...
- Under an efficient policy, the throughput is constrained by actual machine speed.
- Many operational policies have been discovered and proved to be efficient.
- Fluid model is the main tool for proofs.

Motivation I: Power of Fluid Models

- Sufficiency: a queueing network is stable if the corresponding fluid model is stable. [Rybko-Stolyar (92), Dai (95), Stolyar (95), Dai-Meyn]
  Powerful in showing "good policies" for a stochastic network are indeed "good" . [generalized HL processor sharing, HL proportional processor sharing (Bramson), global LIFO (Rybko-Stolyar-Suhov) , global FIFO (Bramson), ...]
- Partial converses: Meyn (95), Dai (96), Rybko-Pulhaski (99)

Push Start



$$\frac{m_3}{1 - \alpha_1 m_1}$$

$$\alpha_1 \left(\frac{m_3}{1 - \alpha_1 m_1} + m_5 \right) \leq 1.$$

Motivation II: Fluid Model Stability Region Characterization

- the usual traffic conditions: $\rho_i < 1$
- virtual station conditions: $\rho_v < 1$
- push start conditions: $\rho_{\mathrm{ps}} < 1$

Dai-Vande Vate (00) for general 2-station fluid networks

Fluid Model

- deterministic, continuous analog
- defined through a set of equations
- non-unique fluid model solutions

Definition

A fluid model is said to be stable if every fluid solution model empties eventually.

Fluid Model Equations

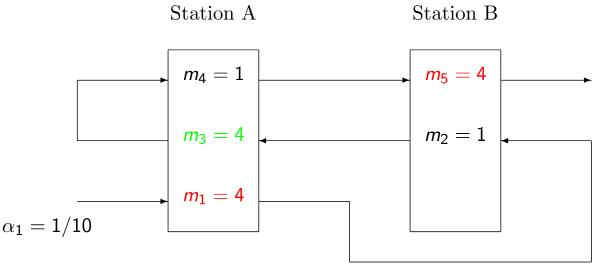
$$Z_k(t) = Z_k(0) + \mu_{k-1} T_{k-1}(t) - \mu_k T_k(t), \tag{1}$$

$$T_k(t) \text{ is nondecreasing}, \tag{2}$$

$$Z_5(t) > 0 \Rightarrow \dot{T}_5(t) = 1, \tag{3}$$

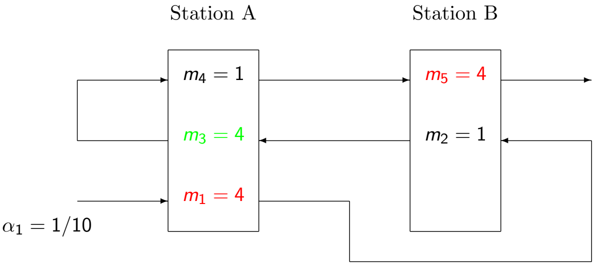
$$Z_2(t) + Z_5(t) > 0 \Rightarrow \dot{T}_2(t) + \dot{T}_5(t) = 1, \tag{4}$$

$$\ldots$$



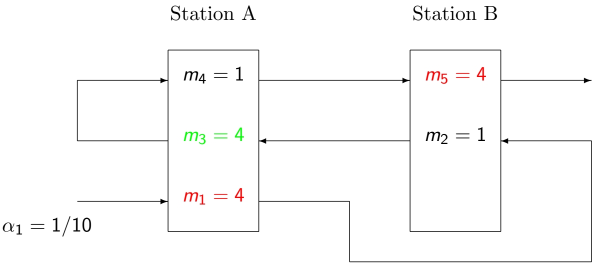
An Unstable Fluid Model Solution: Part I

Let $d_k(t) = \mu_k \dot{T}_k(t)$ be the departure rate from buffer k. Assume that $Z(0) = (0, 0, 0, 1, 0)$.
$d_5(t) = \mu_5 = 1/4$, $d_4(t) = \mu_4(1 - \alpha_1 m_1) = \mu_4(0.6) > d_5(t)$.
Buffer 2 accumulates as long as buffer 5 is non-empty. Buffer 5 empties at time $t_1 = m_5$.



Unstable Fluid Model Solution: Part II

$Z(t_1) = (0, \alpha_1 m_5, 0, 0, 0)$.
$d_2(t) = \mu_2 = 1,\ d_3(t) = \mu_3(1 - \alpha_1 m_1) = 0.28 < d_2(t)$.
Buffer 4 accumulates until buffer 3 empties.
Buffers 1-3 empty at time $t_1 + t_2$ with $t_2 = \frac{\alpha_1 m_5}{d_3(t) - \alpha_1}$.



Unstable Fluid Model Solution: Back to Part I

$Z(t_1 + t_2) = (0, 0, 0, \square, 0)$ with

$$
\begin{aligned}
\square &= \alpha_1 t_1 + \alpha_1 t_2 \\
&= \frac{\alpha_1 m_5}{1 - \frac{\alpha_1 m_3}{(1-\alpha_1 m_1)}}.
\end{aligned}
$$

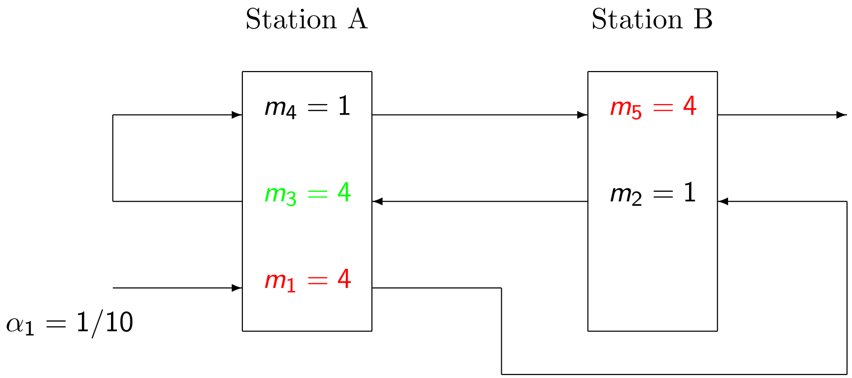
The last expression > 1 if and only if the push start condition is violated, i.e.,

$$
\rho_{\mathrm{ps}} \equiv \frac{\alpha_1 m_3}{1 - \alpha_1 m_1} + \alpha_1 m_5 = \frac{16}{15} > 1.
$$

Summary of Part II

- For a queueing network operating under a HL service policy, its stability region can depend on
  - its distributions, not just means;
  - the preemption mechanism.
- stability
  - total number of jobs being stochastically bounded
  - rate stability
  - positive recurrence

The 2-Station, 5-Class Queueing Network



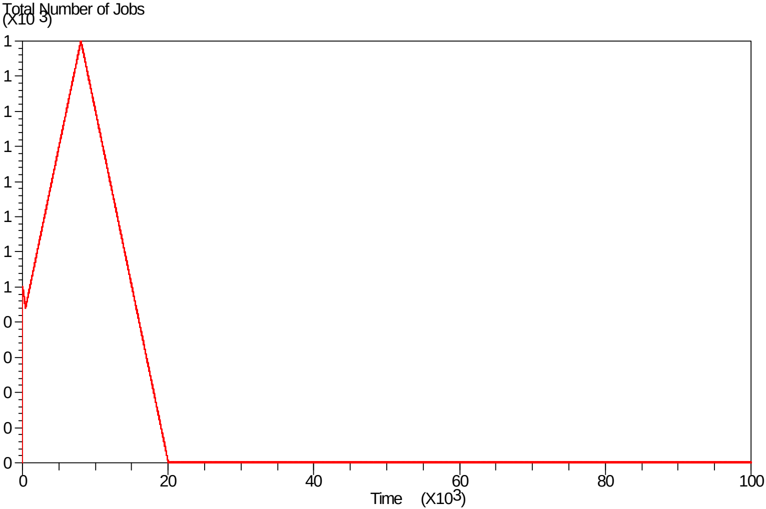
Static buffer priority (SBP) policy: $\{(1, 3, 4), (5, 2)\}$. Red buffers have the highest priority. Black buffers have the lowest priority.

$$\rho_1 = \alpha_1(m_1 + m_3 + m_4) = 0.9,$$
$$\rho_2 = \alpha_1(m_2 + m_5) = 0.5.$$

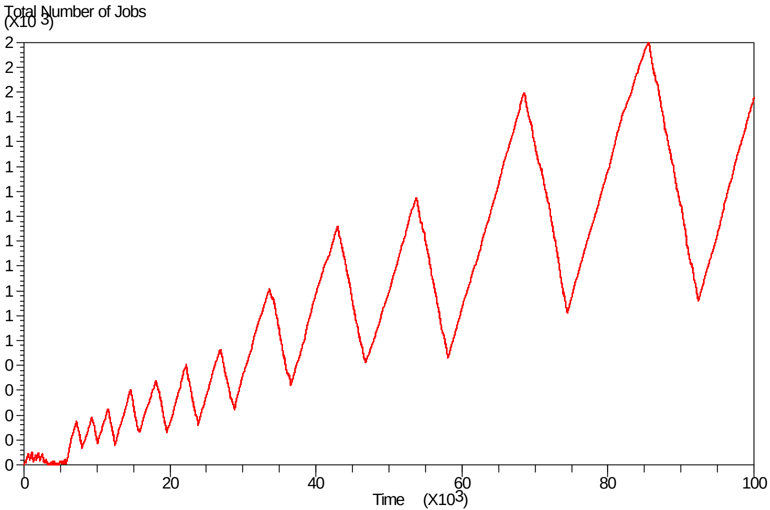
Distributions Matter

- deterministic
- exponential
- uniform with small width
- uniform with large width

$Z(0) = (100, 100, 100, 100, 100)$.

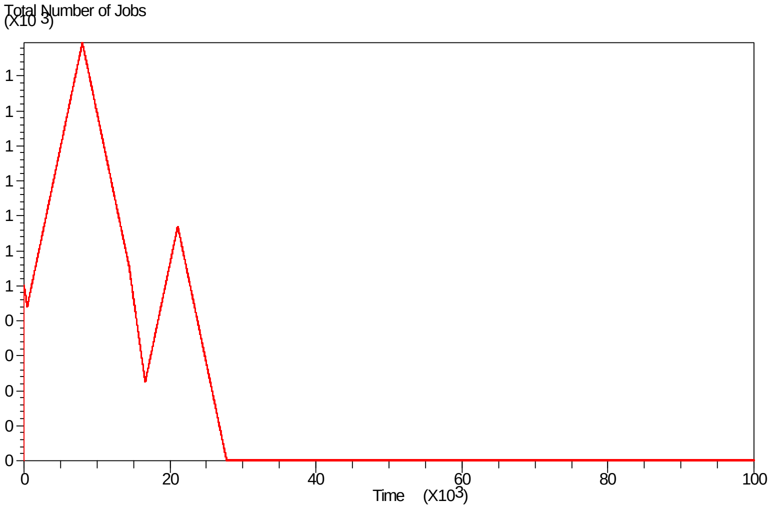


$Z(0) = 0.$


Total Number of Jobs
(X10 3)
2
2
2
1
1
1
1
1
1
1
1
1
1
0
0
0
0
0
0
20
40
60
80
100
Time (X10 3)

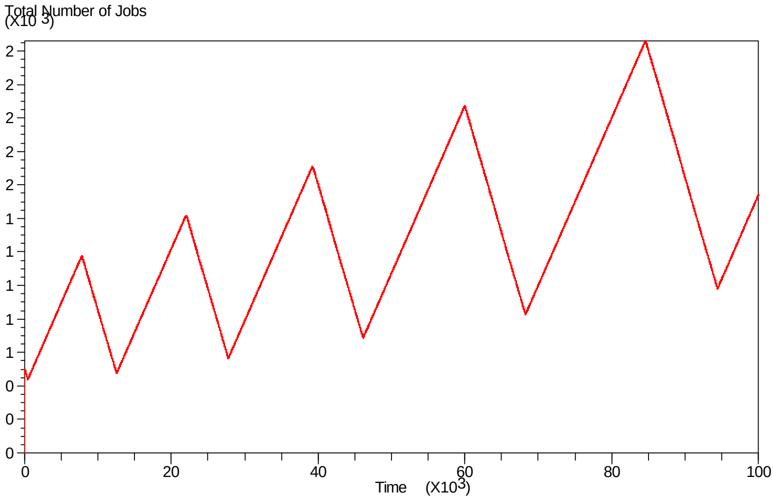
Uniform Distribution: $\epsilon = 0.01$

$Z(0) = (100, 100, 100, 100, 100)$.



Uniform Distribution: $\epsilon = 1.0$

$Z(0) = (100, 100, 100, 100, 100)$.

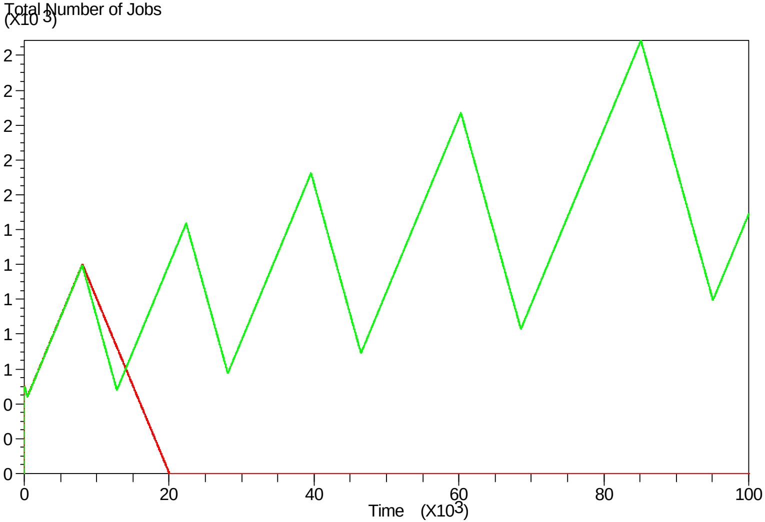


Preemption Matters

- non-preemptive
- preemptive

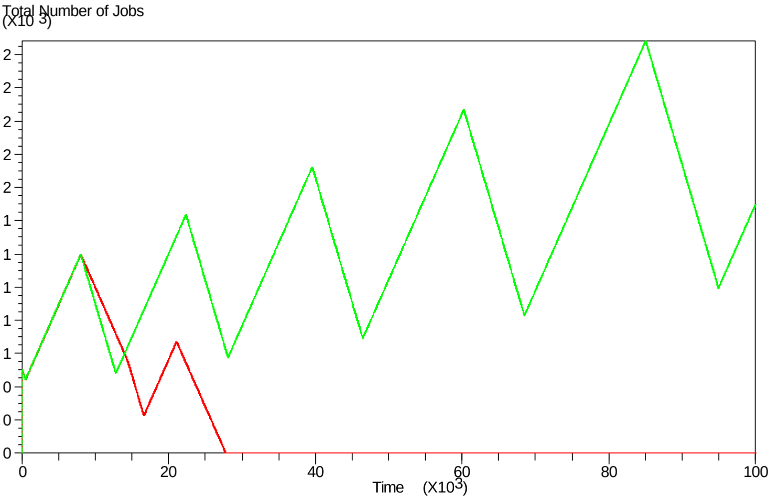
Deterministic Case

$Z(0) = (100, 100, 100, 100, 100)$.



Uniform Distribution with $\epsilon = 0.01$

$Z(0) = (100, 100, 100, 100, 100).$



Theorem 1: Instability for Exponential Case

Theorem

Assume that all distributions are exponential and the non-preemptive SBP policy is used.
Starting from any state, with probability one, the total number of jobs goes to infinity.

Theorem 2: Stability for Deterministic Case

Theorem

Assume that all distributions are deterministic, and the non-preemptive SBP policy is used.
Starting from any state, $Z(t)$ reaches a limit cycle in finite time.
Furthermore, the limit cycle is unique with at most two jobs in the system.

Theorem 3: Instability for Preemptive Case

Theorem

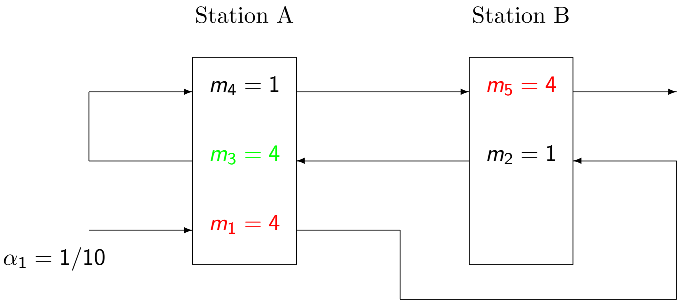
Assume that all distributions are deterministic (or random with small enough supports), and the preemptive SBP policy is used.
Starting from state $Z(0) = (0, n, 0, 0, 0)$ with large enough n, $Z(t)$ cycles to infinity as $t \to \infty$.

Proofs

Follow fluid model solutions!
But which solution?

A Stable Fluid Model Solution

- In Part II, when $Z(t_1) = (0, \alpha_1 m_5, 0, 0, 0)$, do we have to have $d_2(t) = \mu_2 = 1$, $d_3(t) = \mu_3(1 - \alpha_1 m_1) = 0.28$?
- No. One can verify that $d_2(t) = 1/(m_2 + m_5)$, and $d_5(t) = d_4(t) = d_3(t) = d_2(t)$ is another solution.



Which Solution?

- Exponential network follows the unstable fluid model solution.
- Deterministic network follows the stable fluid model solution.
- Deterministic network with preemption follows the unstable fluid model solution.

Let $X^x(t)$ be the state of a queueing network at time t with initial state x.

$$\bar{X}^{x,r}(t,\omega) = \frac{1}{r} X^x(rt,\omega)$$

If there exist sequences $r_n \to \infty$ and x_n with $\limsup_n |x_n|/r_n \leq 1$ such that as $n \to \infty$

$$\bar{X}^{x_n,r_n} \to \bar{X},$$

$\bar{X}$ is then said to be a fluid limit.

Fluid limits can be defined pathwise or distributionally.

Fluid Model v.s. Fluid Limits

- Each fluid limit is a fluid model solution.
- Which fluid model equation should one add?
- Practical fluid models should depend on means only, not on distributions.

The Fluid Model Fails

- Bramson (99): there is a stable exponential queueing network whose fluid model is unstable.
- No matter how many fluid model equations one adds, the fluid model cannot determine the stability of our queueing network.

Proof of Theorem 1: Exponential Case

Proposition. Suppose that $Z(0) = (0, z_2, 0, n, 0)$. There exist $\theta > 1$ and $\delta > 0$ such that for all large n and any z_2,

$$\mathbb{P}\{Z_4(T) \geq \theta n\} \geq 1 - \exp(-\delta\sqrt{n}),$$

where T is some random time with $Z(T) = (0, Z_2(T), 0, Z_4(T), 0)$. Furthermore,

$$\mathbb{P}\{|Z(t)| \geq \kappa n \text{ for all } t \in [0, T]\} \geq 1 - \exp(-\delta\sqrt{n}).$$

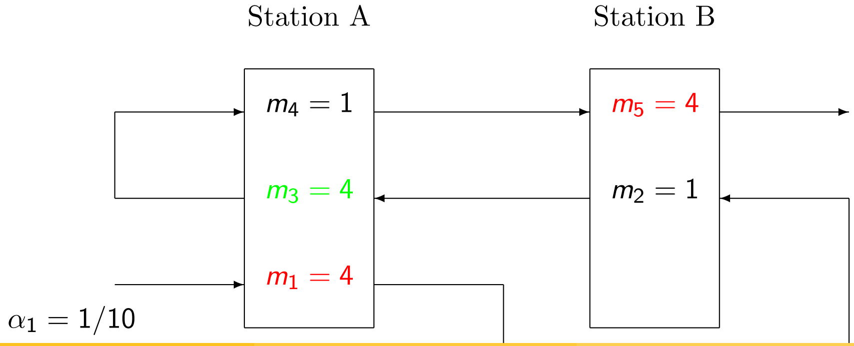
Follows the unstable fluid model solution with high probability!

Proof of Theorem 2: Deterministic without Preemption

State $(0, 0, 0, 1, 0; 1)$ starts a period.

$$Z(0) = (0, 0, 0, n, 0; 1), \qquad Z(1) = (1, 0, 0, n-1, 1; 10),$$
$$Z(5) = (0, 1, 0, n-1, 0; 6), \qquad Z(6) = (0, 0, 1, n-2, 1; 5),$$
$$Z(10) = (0, 0, 0, n-1, 0; 1).$$

Follows the stable fluid model solution.



Proof of Theorem 3: Almost Deterministic with Preemption

With probability one, follow the unstable fluid model solution.

Summary

- For a queueing network operating under some service policy, its stability region can depend on
  - its distributions,
  - the preemption mechanism,
  - the way that simultaneous events are handled.
- Practical fluid models cannot capture these fine factors, and hence cannot be used to sharply determine stability of the corresponding queueing network.
- Fluid model is useful in designing and verifying good policies.
- Fluid model may still be possible to determine sharply the global stability of a queueing network.

References

- Dai, Hasenbein and VandeVate, Stability and instability of a two-station queueing network, *Annals of Applied Probability*, 2004.
- Dai, Hasenbein and VandeVate, Stability of a Three-Station Fluid Network, *Queueing Systems*, 1999
- Bramson, A Stable Queueing network with unstable fluid network, *Annals of Applied Probability*, 1999.