Maximum pressure policies for stochastic processing networks

Jim Dai

Joint work with Wuqin Lin at Northwestern Univ.

The 2011 Lunteren Conference
Outline

1. Stochastic processing network models
2. A motivating example arising from wafer fabrication lines
3. Maximum pressure policies
4. Throughput optimality
5. Asymptotic optimality under complete resource pooling (single bottleneck)
6. References
7. Appendix: Supplements and multiple bottlenecks diffusion limits
Stochastic Processing Network Models
A Stochastic Processing Network Model

Basic elements:
- \( I + 1 \) buffers
- \( K \) processors
- \( J \) activities

Indexes:
- \( i \in \mathcal{I} \cup \{0\} \)  (input and service processors)
- \( k \in \mathcal{K} \)  (input and service activities)
- \( j \in \mathcal{J} \)  (input and service activities)

Material consumption:
- \( \mu_j \): service rate for activity \( j \);
- \( B_{ij} = 1 \) if activity \( j \) processes jobs in buffer \( i \) and \( B_{ij} = 0 \) otherwise;
- \( P_{ii'}^j \) is a fraction of buffer \( i \) jobs served by activity \( j \) that go next to buffer \( i' \).
Resource Allocation

- $A_{kj} = 1$ if activity $j$ requires processor $k$ and 0 otherwise; multiple processors may be needed to activate an activity.
- Allocation space $\mathcal{A}$ is the set of allocations $a \in \mathbb{R}_+^J$ satisfying

$$\sum_j A_{kj}a_j \leq 1 \text{ for each service processor,}$$
$$\sum_j A_{kj}a_j = 1 \text{ for each input processor;}$$

- $a_j$ the level at which activity $j$ is undertaken;
- more constraints on $a$ can be added.
- one input processor, one input activity; the input processor never idles.
- three service processors
Skill-Based Routing

- four input processors, each processing one input activity
- three service processors
two input processors; the left one processes two input activities and the right one processes one input activity.
In each time slot, at most one packet is sent from each input port
In each time slot, at most one packet is sent to each output port
Multiple packets can be transferred in a single time slot
A high speed switch needs to maintain thousands of flows
Operational policies

- $A = \{ a \in \mathbb{R}_+^J : Aa \leq e \}$
- $\mathcal{E} = \{ a_1, ..., a_u \}$ – set extreme points of $A$.
- $A(t)$ – set of feasible allocations at time $t$.
- $\mathcal{E}(t) = A(t) \cap \mathcal{E}$ – set of feasible, extreme allocations at time $t$.
- e.g. $a_1 = (1, 1, 1, 0, 0, 0, 0, 0), a_2 = (1, 1, 1, 1, 0, 1, 0, 0)$
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
- Long-run average cost

Operational policies can have a dramatic impact on key performance measures.
An example arising from wafer fabrication lines
Traffic intensity:

\[ \rho_1 = \lambda_1 m_1 + \lambda_2 m_4 = 0.8 \] and 

\[ \rho_2 = \lambda_1 m_2 + \lambda_2 m_3 = 0.8. \]

Pull policy – give priority to products closer to completion
WIP Levels at Two Stations
Utilization and Cycle Time

\[
\begin{align*}
\lambda_1 &= 1 \\
m_1 &= 0.1 \\
m_4 &= 0.7 \\
m_3 &= 0.1 \\
m_2 &= 0.7 \\
\lambda_2 &= 1
\end{align*}
\]

<table>
<thead>
<tr>
<th># departed</th>
<th>100</th>
<th>1,000</th>
<th>10,000</th>
<th>100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average cycle time</td>
<td>13.68</td>
<td>99.87</td>
<td>927.96</td>
<td>7277.62</td>
</tr>
<tr>
<td>Utilization A</td>
<td>0.65</td>
<td>0.48</td>
<td>0.46</td>
<td>0.71</td>
</tr>
<tr>
<td>Utilization B</td>
<td>0.49</td>
<td>0.67</td>
<td>0.73</td>
<td>0.44</td>
</tr>
<tr>
<td>Overall Utilization</td>
<td>0.57</td>
<td>0.58</td>
<td>0.60</td>
<td>0.58</td>
</tr>
</tbody>
</table>

the throughput is about 0.7.
Under the pull policy, the system is “stable” if and only if

\[ \rho_1 = \lambda_1 m_1 + \lambda_2 m_4 \leq 1, \quad \Rightarrow \quad \lambda_1^* = \lambda_2^* = \frac{1}{0.8} = 1.25, \]

\[ \rho_2 = \lambda_1 m_2 + \lambda_2 m_3 \leq 1, \]

\[ \rho_v = \lambda_1 m_2 + \lambda_2 m_4 \leq 1. \quad \Rightarrow \quad \lambda_1^* = \lambda_2^* = \frac{1}{1.4} = 0.714 \]

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)
... 

Symptoms:
- WIP is high, and
- bottleneck machines are underutilized
Maximum pressure policies
Maximum Pressure Policies

- Fix an \( \alpha = (\alpha_i) \in \mathbb{R}_+^I \) with \( \alpha_i > 0 \).
- Pressure at time \( t \) for activity \( j \),
  \[
  p_j(t) = \mu_j \left( \sum_{i \in \mathcal{I}} B_{ij} \left( \alpha_i Z_i(t) - \sum_{i'} P^j_{ii'} \alpha_{i'} Z_{i'}(t) \right) \right),
  \]
  where \( Z_i(t) \) is the number of jobs in buffer \( i \) at time \( t \).
- At any time \( t \), choose an allocation \( a \)
  \[
  a \in \arg\max_{a \in \mathcal{E}(t)} \sum_j a_j p_j(t).
  \]
Server \( k \) chooses to work on a buffer that has the highest pressure.
The pressure at buffer \( i \) is

\[
p_i(t) = \mu_i \left( Z_i(t) - Z_{i+1}(t) \right).
\]

If all \( p_i(t) \leq 0 \), idle the server.
Generalization: change \( Z_i(t) \) to \( \alpha_i Z_i(t) \)
For example, processor 1 chooses to work on buffer \( i \) that attains

\[
\max\{\mu_1 Z_1(t), \mu_2 Z_2(t), \mu_4 Z_3(t)\}.
\]

- Mandelbaum-Stolyar (04): generalized \( c_\mu \)-rule; van Mieghem (95)
- Stolyar (04): MaxWeight policies
An MPP translates into: **Join-the-shortest-queue** and server 1 idles when $Z_3(t) > Z_1(t)$.
An MPP translates into: Join-the-shortest-queue and server 1 idles when $Z_3(t) > Z_1(t)$.

MPPs can be idling policies.
Non-Idling Server 1

Number of jobs in queue 3

exponential distribution
Features of Maximum Pressure Policies

- They are simple.
- They are semi-local.
- They are throughput optimal.
- They are asymptotically optimal in workload and certain holding cost structure.
Throughput optimality
Recall that $Z_i(t)$ is the buffer level at time $t$ in buffer $i$.

**Rate stability**

With probability one,

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

for each buffer $i$, which is equivalent to that departure rate is equal to arrival rate.

**Positive Harris recurrence**
Define $\rho = \max(\rho_1, \rho_2)$, where

\[
\rho_1 = \lambda_1 m_1 + \lambda_2 m_4 \leq 1, \\
\rho_2 = \lambda_1 m_2 + \lambda_2 m_3 \leq 1.
\]
The static planning problem (Harrison 00):

\[
\begin{align*}
\text{minimize} & \quad \rho \\
\text{subject to} & \quad Rx = 0 \\
& \quad \sum_j A_{kj}x_j = 1 \text{ for each input processor } k \\
& \quad \sum_j A_{kj}x_j \leq \rho \text{ for each service processor } k \\
& \quad x \geq 0
\end{align*}
\]

- \( R_{ij} = \mu_j(B_{ij} - \sum_{i'} B_{i'j}P^j_{i',i}) \)
- \( A \): capacity consumption matrix
- \( x_j \): fraction of time for activity \( j \);
- \( \rho \): utilization of bottleneck servers.
Theorem (Dai-Lin 05)

If the stochastic processing network operating under any operational policy is rate stable, the static planning LP has a feasible solution with $\rho \leq 1$.

Theorem (Dai-Lin 05)

Conversely, suppose that Assumption 1 in the appendix is satisfied. If the static planning LP has a feasible solution with $\rho \leq 1$, the stochastic processing network operating under a maximum pressure policy is rate stable.
Theorem (Dai-Lin 05)

A stochastic processing network is rate stable if the corresponding continuous, deterministic fluid model is weakly stable.

Theorem (Dai 95)

A multiclass queueing network is positive Harris recurrent if the corresponding continuous, deterministic fluid model is stable.
Let $T_k(t)$ be the cumulative time that class $k$ jobs have received in $[0, t]$.

\[
Z_1(t) = Z_1(0) + \lambda t - \mu_1 T_1(t),
\]
\[
Z_k(t) = Z_k(0) + \mu_{k-1} T_{k-1}(t) - \mu_k T_k(t),
\]

$T_k(0) = 0$ and $T_k(\cdot)$ is nondecreasing,

\[
(T_1(t) + T_3(t) + T_5(t)) - (T_1(s) + T_3(s) + T_5(s)) \leq (t - s)
\]
\[
(T_2(t) + T_4(t)) - (T_2(s) + T_4(s)) \leq (t - s)
\]
Fluid Model under MPP

\[ \sum_{i=1}^{5} \dot{T}_i(t)p_i(t) = \max \left\{ \sum_i a_ip_i(t) : a_1 + a_3 + a_5 \leq 1, a_2 + a_4 \leq 1 \right\}, \quad (1) \]

where, the buffer \( i \) pressure \( p_i(t) = \mu_i(\bar{Z}_i(t) - \bar{Z}_{i+1}(t)) \).

- The drift of the quadratic function \( f(t) = \sum_i \bar{Z}^2_i(t)/2 \) is given by
  \[ \dot{f}(t) = \lambda Z_1(t) - \sum_i \dot{T}_i(t)p_i(t). \]
- Under a maximum pressure policy, \( \dot{f}(t) \) is minimized among all policies.
Definition (Weak Stability)

A fluid model is said to be weakly stable if for every fluid model solution with $\bar{Z}(0) = 0$, $\bar{Z}(t) = 0$ for $t \geq 0$.

- Consider the quadratic function $f(t) = \sum_i \bar{Z}_i^2(t)/2$.
- Under a maximum pressure policy, $\dot{f}(t) \leq 0$. Therefore, $\bar{Z}(t) = 0$ for all $t$ if $\bar{Z}(0) = 0$; the fluid model is weakly stable.
Fluid Limits

- Fluid model equations are justified through a fluid limit procedure.
- A function \((\bar{Z}, \bar{T})\) is said to be a fluid limit if

\[
\frac{1}{r_n}(Z(r_n t, \omega), T(r_n t, \omega)) \rightarrow (\bar{Z}(t), \bar{T}(t))
\]

as \(r_n \rightarrow \infty\) for some sample path \(\omega\)
Asymptotic optimality under complete resource pooling (CRP) or single bottleneck assumption
Each buffer $i$, the holding cost rate is $h_i(Z_i(t))^2$.

The network cost rate is

$$h(Z(t)) = \sum_i h_i(Z_i(t))^2.$$ 

Under a policy $\pi$, the expected total discounted holding cost

$$J_\pi \equiv \mathbb{E} \left( \int_0^\infty e^{-\gamma t} h(Z_\pi(t)) \, dt \right).$$
Consider a sequence of networks indexed by $r$ in heavy traffic,

$$\lim_{r \to \infty} R^r = R.$$  \hspace{1cm} (2)

Diffusion Scaling: $\hat{Z}^r(t) = Z^r(rt)/\sqrt{r}$ and

$$\hat{J}_\pi^r \equiv \mathbb{E}\left(\int_0^\infty e^{-\gamma t} h(\hat{Z}^r(t)) dt\right).$$

**Theorem (Dai-Lin 08)**

For a sequence of networks that satisfies a heavy traffic condition and a complete resource pooling condition, the maximum pressure policy with $\alpha = h$ is asymptotically optimal to minimize the quadratic holding cost, i.e.,

$$\lim_{r \to \infty} \hat{J}^r_{\text{MPP}} \leq \lim_{r \to \infty} \inf \hat{J}^r_\pi \text{ for any policy } \pi.$$
References: Stochastic Processing Networks


References on Unstable Queueing Networks

An Appendix

- Assumption 1.
- The heavy traffic assumption
- The complete resource pooling assumption
- State space collapse and semimartingale reflecting Brownian motions (SRBM) as diffusion limits
- Extension of maximum pressure policies
Assumption 1

For any vector \( z \in \mathbb{R}_+^I \), there exists an \( a \in \operatorname{arg\,max}_{a \in E} \sum_i v(a, i)z_i \) such that \( v(a, i) = 0 \) if \( z_i = 0 \), where \( v(a, i) = \sum_j a_j R_{ij} \) is the consumption rate of buffer \( i \) under allocation \( a \).

The assumption holds when each activity is associated with one buffer (in Leontief networks).
The Heavy Traffic Assumption

Consider a sequence of stochastic processing networks that satisfies assumption (2). The static planning problem (Harrison 00):

\[
\begin{align*}
\text{minimize} & \quad \rho \\
\text{subject to} & \quad Rx = 0 \\
& \quad \sum_j A_{kj}x_j = 1 \text{ for each input processor } k \\
& \quad \sum_j A_{kj}x_j \leq \rho \text{ for each service processor } k \\
& \quad x \geq 0
\end{align*}
\]

- $x_j$: fraction of time for activity $j$ is employed;
- $\rho$: utilization of bottleneck servers.

**Assumption**

The optimal solution $(\rho^*, x^*)$ is unique and $\rho^* = 1$. 
The Complete Resource Pooling (CRP) Assumption

The dual LP:

\[
\begin{align*}
\text{minimize} & \quad \sum_{k \in K_I} z_k \\
\text{subject to} & \quad \sum_{i \in I} y_i R_{ij} \leq - \sum_{k \in K_I} z_k A_{kj} \quad \text{for each input activity } j \\
& \quad \sum_{i \in I} y_i R_{ij} \leq \sum_{k \in K_S} z_k A_{kj} \quad \text{for each service activity } j \\
& \quad \sum_{k \in K_S} z_k = 1, \\
& \quad z_k \geq 0
\end{align*}
\]

**Assumption**

The dual LP has a nonnegative, unique optimal solution \((y^*, z^*)\).
An Example of CRP: Multiclass queueing networks

$\lambda = 2$  
$\mu_1 = 3$  
$\mu_2 = 4$  
$\mu_3 = 6$

$\bullet$ Heavy traffic assumption: $\rho^* = 1$

$\bullet$ Complete resource pooling condition: either $\rho_1^* = 1$ and $\rho_2^* < 1$ or $\rho_1^* < 1$ and $\rho_2^* = 1$.

$\bullet$ In the former case, $y^* = (m_1 + m_3, m_3, m_3)$; in the latter case, $y^* = (m_2, m_2, 0)$.

$\bullet$ Ata-Kumar (05) does not cover this class of networks

Unique solution $(\rho^*, x^*)$

- $x_j^* = \lambda m_j$, $j = 1, 2, 3$
- $\rho_1 = x_1^* + x_3^*$,
  $\rho_2 = x_2^*$
- $\rho^* = \max(\rho_1, \rho_2)$. 

Jim Dai (Georgia Tech)  
MPPs  
January 18, 2011  
46 / 55
An Example of CRP: Parallel server queues

- Assume $\rho^* = 1$ and $x^*$ is unique.
- Complete resource pooling: all servers communicate through basic activities.
- Harrison-Lopez (99), and Bell-Williams (05)
An Example of multiple LP Solutions

- $\rho^* = 1$, but $\lambda^*$ is not unique
Assume the complete resource pooling condition and \((y, z)\) is the unique solution to the dual LP.

Let \(W(t) = y \cdot Z(t)\) and \(\hat{W}(t) = W(rt)/\sqrt{r} = y \cdot \hat{Z}(t)\).

**Theorem (Workload Optimality (Dai-Lin 08))**

For a sequence of networks that satisfies the heavy traffic condition and the complete resource pooling condition, any the maximum pressure policy is asymptotically optimal for workload in that for each \(t \geq 0\) and \(w > 0\),

\[
P\left( \lim_{r \to \infty} \hat{W}_{MPP}(t) > w \right) \leq P\left( \liminf_{r \to \infty} \hat{W}_{\pi}(t) > w \right).
\]
Proof: A Lower Bound on Workload Process

We can write $\hat{W}^r(t)$ as

$$\hat{W}^r(t) = \hat{X}^r(t) + \hat{Y}^r(t),$$

where $\hat{Y}^r(t) \geq 0$ and nondecreasing. This implies

$$\hat{W}^r(t) \geq \hat{W}^{*,r}(t) \equiv \hat{X}^r(t) - \inf_{0 \leq s \leq t} \hat{X}^r(s).$$

Letting $\hat{W}^*(t) \equiv \hat{X}^*(t) - \inf_{0 \leq s \leq t} \hat{X}^*(s)$,

$$\mathbb{P}\left(\liminf_{r \to \infty} \hat{W}^r(t) > w\right) \geq \mathbb{P}\left(\hat{W}^*(t) > w\right).$$
Proof: A Heavy Traffic Limit Theorem

**Theorem**

For a sequence of networks that satisfies the heavy traffic condition and a complete resource pooling condition, under the maximum pressure policy with $\alpha = e$,

$$(\hat{W}^r, \hat{Z}^r) \Rightarrow (\hat{W}^*, \hat{Z}^*),$$

where $\hat{Z}^* = \frac{y\hat{W}^*}{\|y\|^2}$.

- A key to the proof of this theorem is to show a state space collapse result:

$$\sup_{0 \leq t \leq T} \left| \hat{Z}^r(t) - \frac{y\hat{W}^r(t)}{\|y\|^2} \right| \to 0 \text{ in probability as } r \to \infty.$$

- Use framework of Bramson (98)
- Unlike Chen and Mandelbaum (90), non-bottleneck stations do not disappear.
Consider the optimization problem

\[
\min \sum_{i=1}^{3} q_i^2 \\
\text{s.t.} \quad y \cdot q = w \\
q \geq 0.
\]

- The optimal solution is given by \( q^* = yw/\|y\|^2 \).
- For any given \( w \), it is optimal to distribute the workload to the buffers in proportion to \( y \).
- MPP not only minimizes the workload process \( W(t) \), but also distributes it in the optimal way.
Dai-Lin (08): for each $\epsilon > 0$, one can find an MPP policy with parameter $\alpha$ that is asymptotically $\epsilon$-optimal; choice of $\alpha$ is data heavy.

Ata and Kumar (05) uses Harrison’s BIGSTEP method; rules out multiclass networks

Bell and Williams (05) parallel-server queues; Ghamami and Ward (09)

Lin (09): $\beta$-Maximum Pressure Policies in Stochastic Processing Networks: Heavy Traffic Analysis. Fix a $\beta > 0$ and $(\alpha_i) > 0$

$$p_i(t) = \mu_i \left( \alpha_i(Z_i(t))^{\beta} - \alpha_{i+1}(Z_{i+1}(t))^{\beta} \right)$$
Let \( \{(y^\ell, z^\ell) : \ell = 1, \ldots, L\} \) denote the set of basic optimal solutions to the dual LP.

Let \( \hat{W}_r^\ell(t) = y^\ell \cdot \hat{Z}(t) \).

**Theorem (Ata-Lin 08)**

Consider a sequence of networks that satisfies the heavy traffic condition. Assume that \( y^\ell \geq 0 \) for each \( \ell \) and \( y^1, \ldots, y^L \) are linearly independent. Under a maximum pressure policy with parameter \( \alpha \),

\[
(\hat{W}^r, \hat{Z}^r) \Rightarrow (\hat{W}^*, \hat{Z}^*),
\]

where \( \hat{W} \) is an \( L \)-dimensional SRBM, and \( \hat{Z}^* = \Delta \hat{W}^* \).
Other extensions

- Rajagopalan, Shah and Shin (09): random-access algorithm to approximate a maximum pressure policy for single-hop networks