Maximum pressure policies for stochastic processing networks

Jim Dai

Georgia Institute of Technology
The H. Milton Stewart School of Industrial and Systems Engineering

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:

- $\mathbf{I+1}$ buffers
- $\mathbf{K}$ processors
- $\mathbf{J}$ activities

Indexes:

- $i \in \mathcal{I} \cup \{0\}$
- input and service processors $k \in \mathcal{K}$
- input and service activities $j \in \mathcal{J}$

Material consumption:

- μ_j: service rate for activity j;
- $B_{ij} = 1$ if activity j processes jobs in in buffer i and $B_{ij} = 0$ otherwise;
- $P^j_{ii'}$ is a fraction of buffer i jobs served by activity j that go next to buffer i';

Resource Allocation

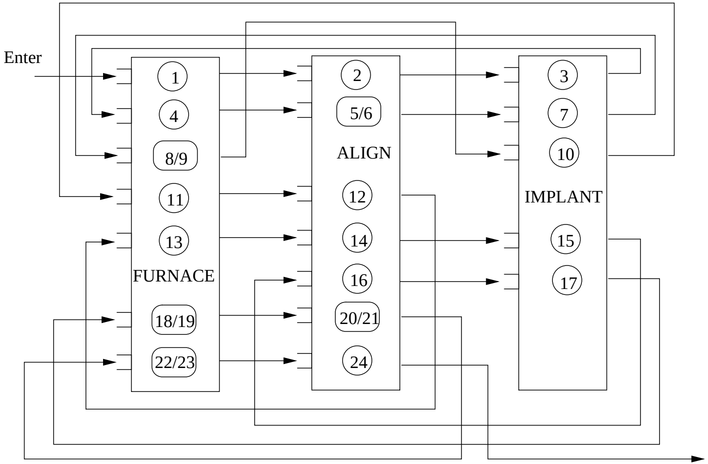
- $A_{kj} = 1$ if activity j requires processor k and 0 otherwise; multiple processors may be needed to activate an activity.
- Allocation space $\mathcal{A}$ is the set of allocations $a \in \mathbb{R}_+^{\mathbf{J}}$ satisfying

$$\sum_j A_{kj} a_j \le 1 \text{ for each service processor,}$$

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

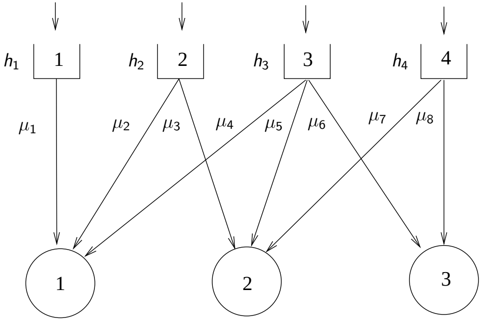
- a_j the level at which activity j is undertaken;
- more constraints on a can be added.

Multiclass Queueing Networks: A Re-Entrant Line



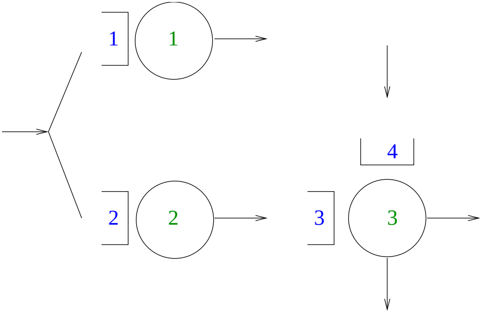
- one input processor, one input activity; the input processor never idles.
- three service processors

Skill-Based Routing



- four input processors, each processing one input activity
- three service processors

Queueing Networks with Alternate Routes



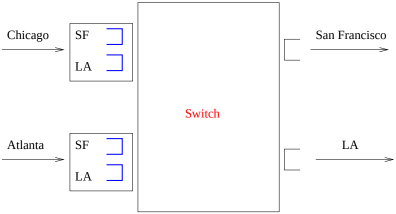
Laws and Louth (1990)

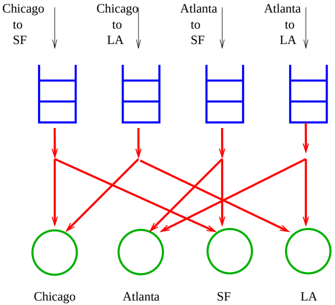
Kelly and Laws (1993)

Dai, Hasenbein and Kim (2007)

- two input processors; the left one processes two input activities and the right one processes one input activity.

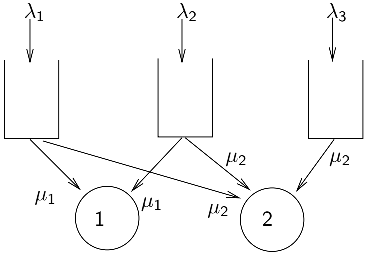
Input Queued Data Switches





- 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



- $\mathcal{A} = \{a \in \mathbb{R}_+^{\mathbf{J}} : Aa \le e\}$
- $\mathcal{E} = \{a_1, ..., a_u\}$ – set extreme points of $\mathcal{A}$.
- $\mathcal{A}(t)$ – set of feasible allocations at time t.
- $\mathcal{E}(t) = \mathcal{A}(t) \cap \mathcal{E}$ – 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:

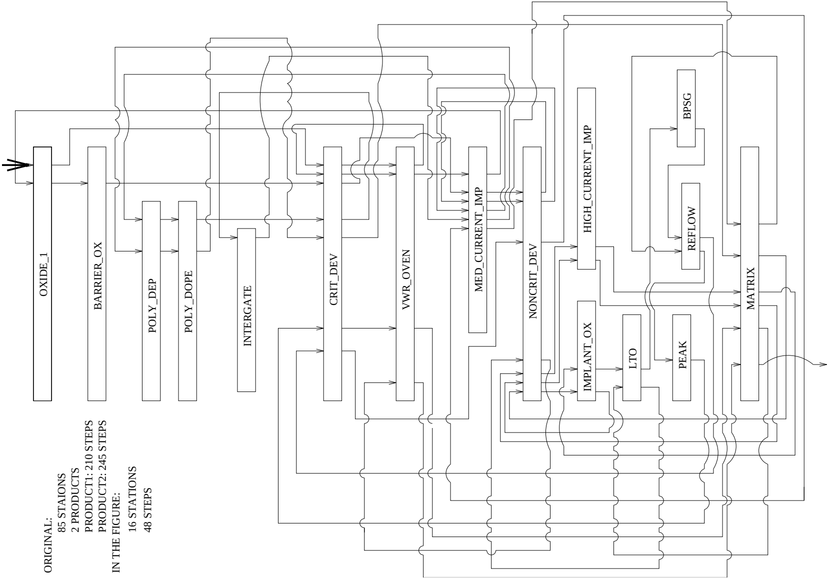
- Cycle time: processing times plus waiting time of an entity; average and variance of cycle time
- Long-run average cost

Operational policies can have a dramatic impact on key performance measures.

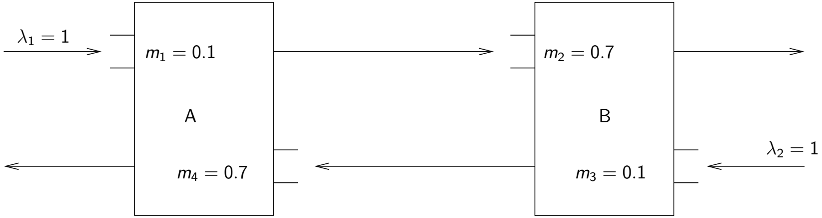
Part II

An example arising from wafer fabrication lines

Flow in a Wafer Fab



The Kumar-Seidman, Rybko-Stolyar Network

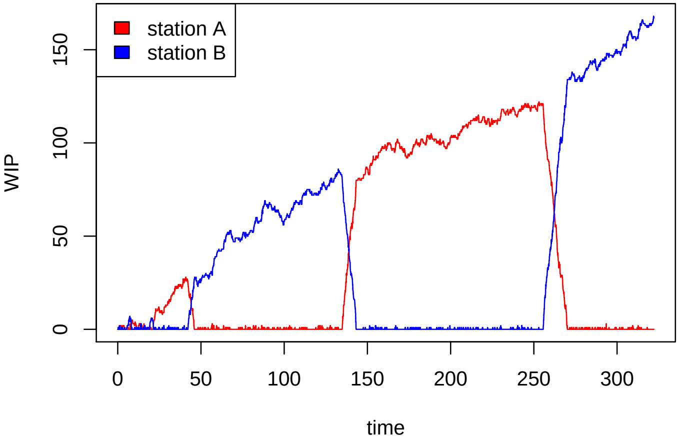


- Traffic intensity:

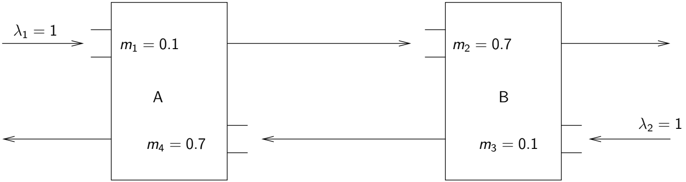
$$\rho_1 = \lambda_1 m_1 + \lambda_2 m_4 = 0.8 \text{ and } \rho_2 = \lambda_1 m_2 + \lambda_2 m_3 = 0.8.$$

- Pull policy – give priority to products closer to completion

WIP Levels at Two Stations



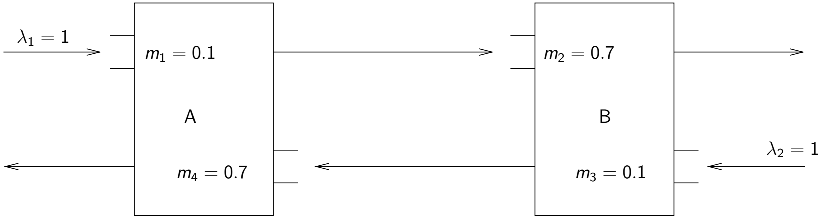
Utilization and Cycle Time



# departed	100	1,000	10,000	100,000
Average cycle time	13.68	99.87	927.96	7277.62
Utilization A	0.65	0.48	0.46	0.71
Utilization B	0.49	0.67	0.73	0.44
Overall Utilization	0.57	0.58	0.60	0.58

the throughput is about 0.7.

Maximum throughput



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

$$\rho_1 = \lambda_1 m_1 + \lambda_2 m_4) \le 1, \quad \Rightarrow \quad \lambda_1^* = \lambda_2^* = \frac{1}{.8} = 1.25,$$

$$\rho_2 = \lambda_1 m_2 + \lambda_2 m_3 \le 1,$$

$$\rho_v = \lambda_1 m_2 + \lambda_2 m_4 \le 1. \quad \Rightarrow \quad \lambda_1^* = \lambda_2^* = \frac{1}{1.4} = 0.714$$

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

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

Part III

Maximum pressure policies

Maximum Pressure Policies

- Fix an $\alpha = (\alpha_i) \in \mathbb{R}^{\mathbf{I}}_{+}$ with $\alpha_i > 0$.
- Pressure at time t for activity j,

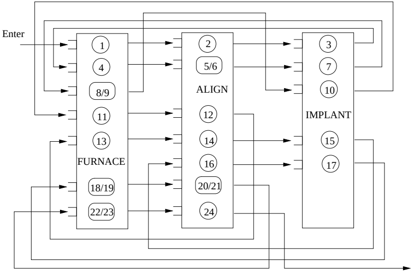
$$p_j(t) = \mu_j \left(\sum_{i \in \mathcal{I} \cup \{0\}} B_{ij} \Big(\alpha_i Z_i(t) - \sum_{i'} P^j_{ii'} \alpha_{i'} Z_{i'}(t) \Big) \right),$$

  where $Z_i(t)$ is the number of jobs in buffer i at time t.
- At any time t, choose an allocation a

$$a \in \operatorname*{argmax}_{a \in \mathcal{E}(t)} \sum_j a_j p_j(t).$$

- Tassiulas (1995): Adaptive back-pressure congestion control based on local information.

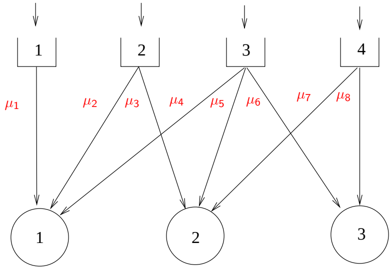
Maximum Pressure Policies for Multiclass Queueing Networks



- Server k chooses to work on a buffer that has the highest pressure.
- The pressure at buffer i is

$$p_i(t) = \mu_i \Big(Z_i(t) - Z_{i+1}(t) \Big).$$

- If all $p_i(t) \le 0$, idle the server.
- Generalization: change $Z_i(t)$ to $\alpha_i Z_i(t)$

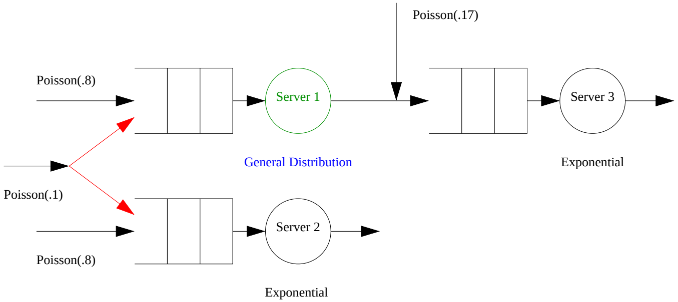


For example, processor 1 chooses to work on buffer i that attains

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

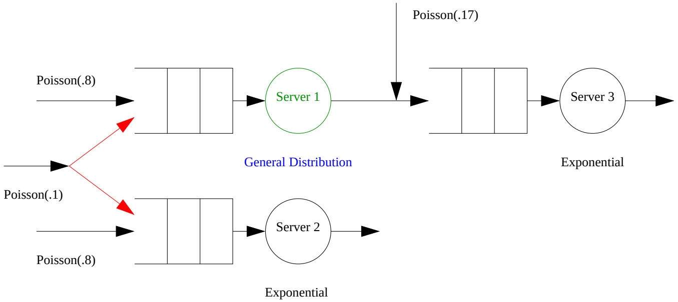
- Mandelbaum-Stolyar (04): generalized $c\mu$-rule; van Mieghem (95)
- Stolyar (04): MaxWeight policies

Maximum Pressure Policies: Alternate Routing



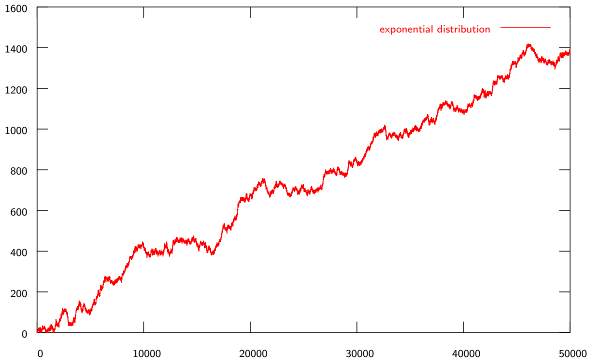
- An MPP translates into: Join-the-shortest-queue and server 1 idles when $Z_3(t) > Z_1(t)$.

Maximum Pressure Policies: Alternate Routing



- 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

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.

Part IV

Throughput optimality

Stability

Recall that $Z_i(t)$ is the buffer level at time t in buffer i.

Rate stability

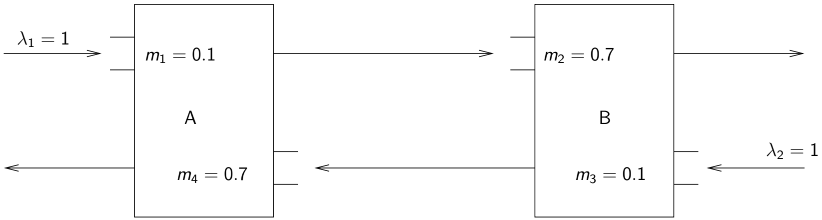
With probability one,

$$\lim_{t\to\infty} Z_i(t)/t = 0, \text{ for each buffer } i$$

which is equivalent to that departure rate is equal to arrival rate.

Positive Harris recurrence

Traffic Intensity



Define $\rho = \max(\rho_1, \rho_2)$, where

$$\rho_1 = \lambda_1 m_1 + \lambda_2 m_4) \leq 1,$$
$$\rho_2 = \lambda_1 m_2 + \lambda_2 m_3 \leq 1.$$

Static Planing Problem

The static planning problem (Harrison 00):

$$
\begin{aligned}
\text{minimize} \quad & \rho \\
\text{subject to} \quad & Rx = 0 \\
& \sum_j A_{kj} x_j = 1 \text{ for each input processor } k \\
& \sum_j A_{kj} x_j \le \rho \text{ for each service processor } k \\
& x \ge 0
\end{aligned}
$$

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

Stability Result

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.

Proof: Fluid Model Approach

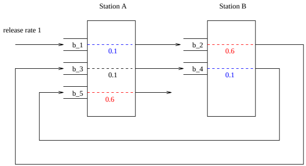
Theorem (Dai-Lin 05)

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.

Fluid Model Equations



Let $T_k(t)$ be the cumulative time that class k jobs have received in $[0, t]$.

$$Z_1(t) = Z_1(0) + \lambda t - \mu_1 T_1(t),$$
$$Z_k(t) = Z_k(0) + \mu_{k-1} T_{k-1}(t) - \mu_k T_k(t),$$
$$T_k(0) = 0 \text{ and } T_k(\cdot) \text{ is nondecreasing},$$
$$(T_1(t) + T_3(t) + T_5(t)) - (T_1(s) + T_3(s) + T_5(s)) \le (t - s)$$
$$(T_2(t) + T_4(t)) - (T_2(s) + T_4(s)) \le (t - s)$$

Fluid Model under MPP

$$\sum_{i=1}^{5} \dot{\bar{T}}_i(t) p_i(t) = \max\Big\{ \sum_i a_i p_i(t) : a_1 + a_3 + a_5 \le 1, a_2 + a_4 \le 1. \Big\}, \quad (1)$$

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 $f(t) = \sum_i \bar{Z}_i^2(t)/2$ is given by $\dot{f}(t) = \lambda Z_1(t) - \sum_i \dot{\bar{T}}_i(t) p_i(t)$.
- Under a maximum pressure policy, $\dot{f}(t)$ is minimized among all policies.

Weak Stability of Fluid Model

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

$$\frac{1}{r_n}(Z(r_n t, \omega), T(r_n t, \omega)) \to (\bar{Z}(t), \bar{T}(t))$$

as $r_n \to \infty$ for some sample path ω

Part V

Asymptotic optimality under complete resource pooling (CRP) or single bottleneck assumption

Quadratic Holding Cost

- Each buffer i, the holding cost rate is $h_i\big(Z_i(t)\big)^2$.
- The network cost rate is

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

- Under a policy π, the expected total discounted holding cost

$$J_\pi \equiv \mathbb{E}\left(\int_0^\infty e^{-\gamma t} h\big(Z^\pi(t)\big)dt\right).$$

Asymptotic Optimality on Quadratic Holding Cost

- Consider a sequence of networks indexed by r in heavy traffic,

$$\lim_{r\to\infty} R^r = R. \tag{2}$$

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

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

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.,

$$\lim_{r\to\infty} \widehat{J}^r_{\mathrm{MPP}} \le \liminf_{r\to\infty} \widehat{J}^r_\pi \quad \textit{for any policy } \pi.$$

References: Stochastic Processing Networks

- Harrison, J. M. (2000). Brownian models of open processing networks: canonical representation of workload. *Annals of Applied Probability*, **10** 75–103. Correction: **13**, 390–393 (2003).
- Harrison, J. M. (2002). Stochastic networks and activity analysis. In *Analytic Methods in Applied Probability* (Y. Suhov, ed.). In Memory of Fridrik Karpelevich, American Mathematical Society, Providence, RI.
- Harrison, J. M. (2003). A broader view of Brownian networks. *Annals of Applied Probability*, **13** 1119–1150.

References: Maximum Pressure Policies

- Dai, J. G. and Lin, W. (2005). Maximum pressure policies in stochastic processing networks. *Operations Research*, **53** 197–218.
- Dai, J. G. and Lin, W. (2008). Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Annals of Applied Probability*, **18** 2239–2299.
- Ata, B. and Lin, W. (2008). Heavy traffic analysis of maximum pressure policies for stochastic processing networks with multiple bottlenecks. *Queueing Systems*, **59** 191–235.

References on Unstable Queueing Networks

- 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.

Part V

An Appendix

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

Assumption 1

Assumption

For any vector $z \in \mathbb{R}^{\mathbf{I}}_{+}$, there exists an $a \in \arg\max_{a \in \mathcal{E}} \sum_i v(a,i) z_i$ 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):

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

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

Assumption

The optimal solution (ρ^*, x^*) is unique and $\rho^* = 1$.

The Complete Resource Pooling (CRP) Assumption

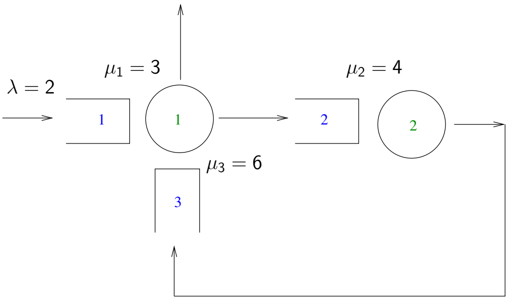
The dual LP:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{k\in\mathcal{K}_I} z_k \\
\text{subject to} \quad & \sum_{i\in\mathcal{I}} y_i R_{ij} \le -\sum_{k\in\mathcal{K}_I} z_k A_{kj} \text{ for each input activity } j \\
& \sum_{i\in\mathcal{I}} y_i R_{ij} \le \sum_{k\in\mathcal{K}_S} z_k A_{kj} \text{ for each service activity } j \\
& \sum_{k\in\mathcal{K}_S} z_k = 1, \\
& z_k \ge 0
\end{aligned}
$$

Assumption

The dual LP has a nonnegative, unique optimal solution (y^*, z^*).

An Example of CRP: Multiclass queueing networks

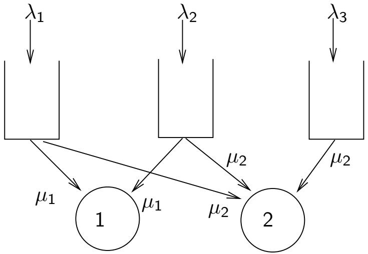


Unique solution (ρ^*, x^*)

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

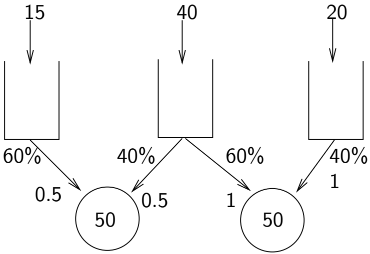
- Heavy traffic assumption: $\rho^* = 1$
- Complete resource pooling condition: either $\rho_1^* = 1$ and $\rho_2^* < 1$ or $\rho_1^* < 1$ and $\rho_2^* = 1$.
- In the former case, $y^* = (m_1 + m_3, m_3, m_3)$; in the latter case, $y^* = (m_2, m_2, 0)$.
- Ata-Kumar (05) does not cover this class of networks

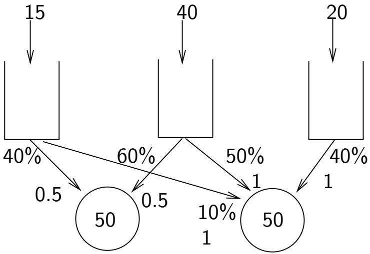
An Example of CRP: Parallel server queues



- 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 x^* 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 $\widehat{W}^r(t) = W(rt)/\sqrt{r} = y \cdot \widehat{Z}^r(t)$.

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 $w > 0$,

$$\mathbb{P}\Big(\lim_{r\to\infty} \widehat{W}^r_{\mathrm{MPP}}(t) > w\Big) \leq \mathbb{P}\Big(\liminf_{r\to\infty} \widehat{W}^r_{\pi}(t) > w\Big).$$

Proof: A Lower Bound on Workload Process

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

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

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

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

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

$$\mathbb{P}\Big(\liminf_{r \to \infty} \widehat{W}^r(t) > w\Big) \geq \mathbb{P}\Big(\widehat{W}^*(t) > w\Big).$$

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 $\alpha = e$,

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

where $\widehat{Z}^* = y\widehat{W}^*/\|y\|^2$.

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

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

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

Asymptotic Optimality Proof (for $h = e$)

Consider the optimization problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{3} q_i^2 \\ s.t. \quad & y \cdot q = w \\ & q \geq 0. \end{aligned}$$

- 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.

Extension: Linear holding cost

- Dai-Lin (08): for each $\epsilon > 0$, one can find an MPP policy with parameter α that is asymptotically ϵ-optimal; choice of α 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): β-Maximum Pressure Policies in Stochastic Processing Networks: Heavy Traffic Analysis. Fix a $\beta > 0$ and $(\alpha_i) > 0$

$$p_i(t) = \mu_i\Big(\alpha_i(Z_i(t))^\beta - \alpha_{i+1}(Z_{i+1}(t))^\beta\Big)$$

Extension: More than one bottleneck

- Let $\{(y^\ell, z^\ell) : \ell = 1, \ldots, L\}$ denote the set of basic optimal solutions to the dual LP.
- Let $\hat{W}^r_\ell(t) = y^\ell \cdot \hat{Z}^r(t)$.

Theorem (Ata-Lin 08)

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

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

where $\widehat{W}$ is an L-dimensional SRBM, and $\widehat{Z}^* = \Delta \widehat{W}^*$.

Other extensions

- Rajagopalan, Shah and Shin (09): random-access algorithm to approximate a maximum pressure policy for single-hop networks
- Shah and Wischik (09): optimal scheduling algorithms for switched networks under light load, critical load, and overload; performance of MaxWeight policies in overloaded fluid networks.