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 Kellogg


get


dear jim I hope you
feel Better soon!.
from sashà and
Kaitlyn.

Outline

- Stochastic processing networks
- Maximum pressure policies
- Asymptotic optimality in heavy traffic

References: Stochastic Processing Networks

- Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations.
- 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.

A Stochastic Processing Network Model

Basic elements:	Indexes:
I + 1 buffers	$i \in \mathcal{I} \cup \{0\}$
K processors	input and service processors $k \in \mathcal{K}$
J activities	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 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';

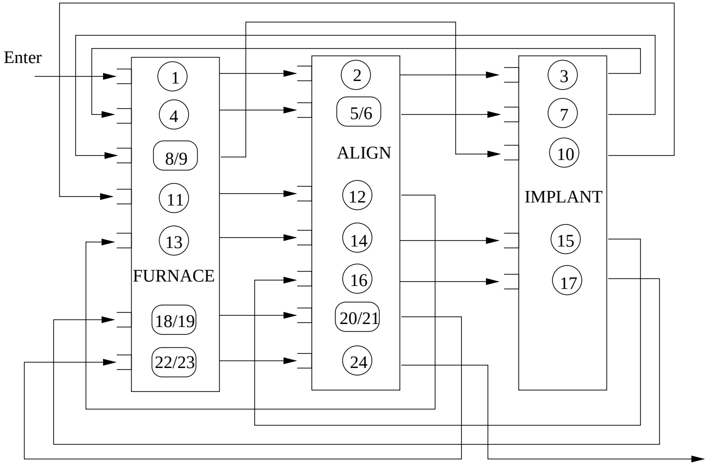
Resource Allocation

- $A_{kj} = 1$ 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

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

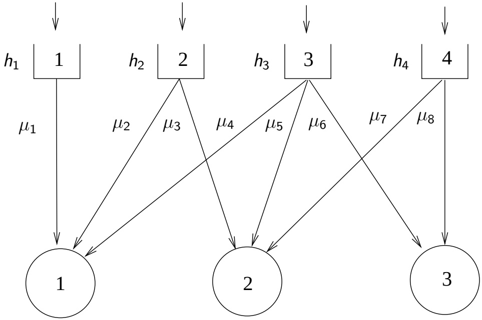
- a_j the level at which activity j is undertaken;
- more constraints on a can be added.

Multiclass queueing networks: a re-entrant line



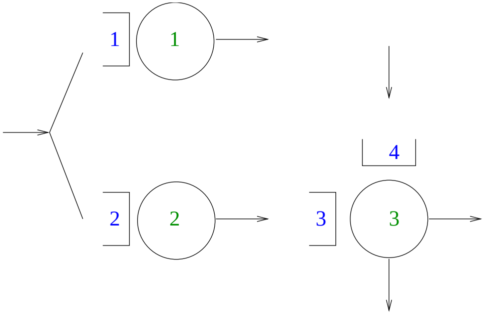
- one input processor, one input activity; the input processor never idles.
- three service processors

Parallel Server Systems



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

Queueing Networks with Alternate Routes



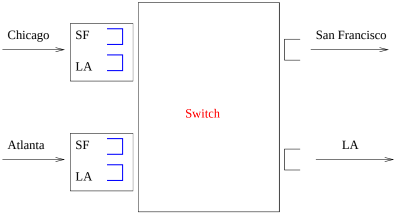
Laws and Louth (1990)

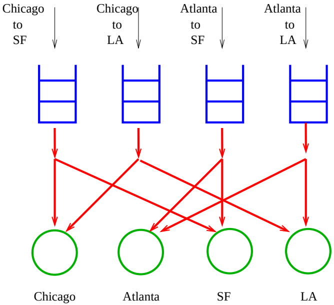
Kelly and Laws (1993)

Dai, Hasenbein and Kim (2007)

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

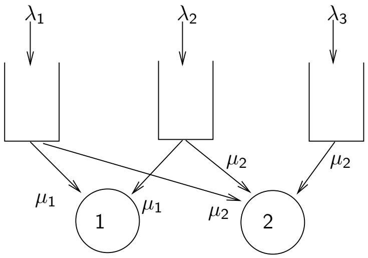
Input Queued Data Switches





- 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. $a_1 = (1, 1, 1, 0, 0, 0, 0, 0)$, $a_2 = (1, 1, 1, 1, 0, 1, 0, 0)$

Maximum Pressure Policies

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

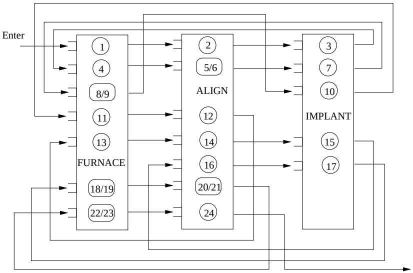
$$p_j(t) = \mu_j \left(\sum_{i \in \mathcal{I} \cup \{0\}} B_{ij} \Big(\alpha_i Z_i(t) - \sum_{i'} P^j_{ii'} \alpha_{i'} Z_{i'}(t) \Big) \right).$$

- 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

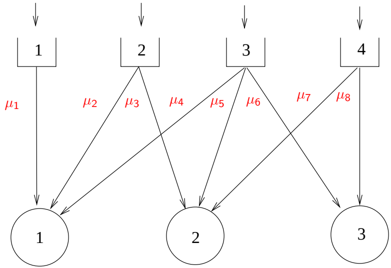


- Server k chooses to work on a buffer that has the highest pressure.
- The pressure at buffer i is

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

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

Maximum Pressure Policies: Parallel Server Systems

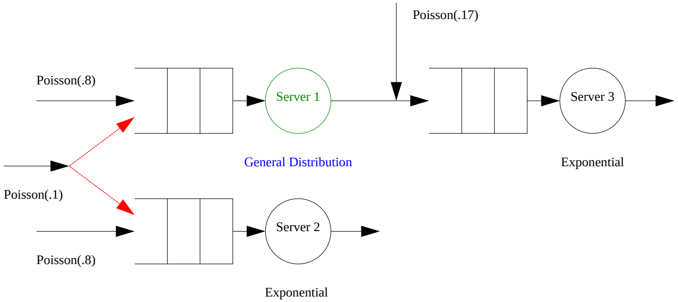


For example, processor 1 chooses to work on buffer i that attains

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

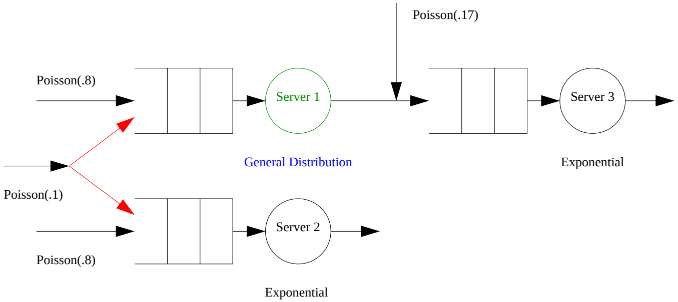
- Mandelbaum-Stolyar (04): generalized $c\mu$-rule; van Mieghem (95)
- Stolyar (04): MaxWeight policies

Maximum Pressure Policies: Alternate Routing



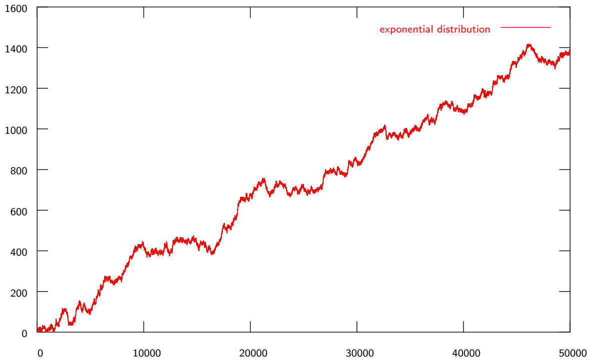
- An MPP translates into: Join-the-shortest-queue and server 1 idles when $Z_3(t) > Z_1(t)$.

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

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.
- 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.,

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

The Heavy Traffic Assumption

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 Assumption

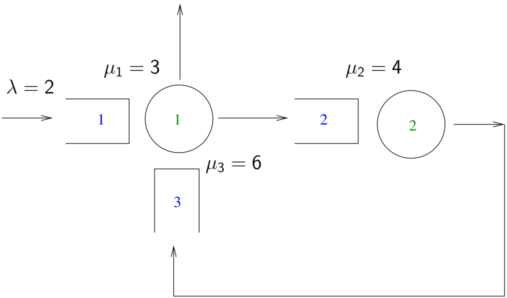
The dual LP:

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

Assumption

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

Multiclass queueing networks

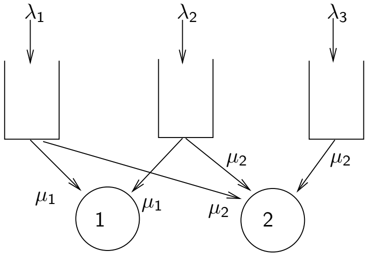


Unique solution (ρ^*, x^*)

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

- 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

Parallel server queues: complete resource pooling



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

Asymptotic Optimality on Workload Process

- 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

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

- 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 $y^1, ..., y^L$ are linearly independent. Under a maximum pressure policy with parameter α,

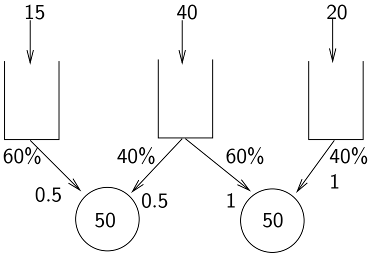
$$(\widehat{W}^r, \widehat{Z}^r) \Rightarrow (\widehat{W}^*, \widehat{Z}^*),$$

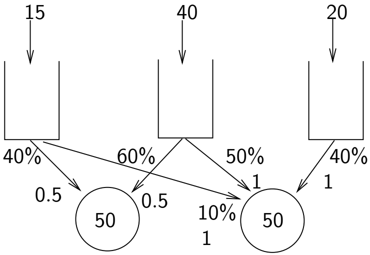
where $\widehat{W}$ is an L-dimensional SRBM, and $\widehat{Z}^* = \Delta \widehat{W}^*$.

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.

Parallel server queues: multiple LP Solutions





- $\rho^* = 1$, but x^* is not unique

Assumption 3

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).