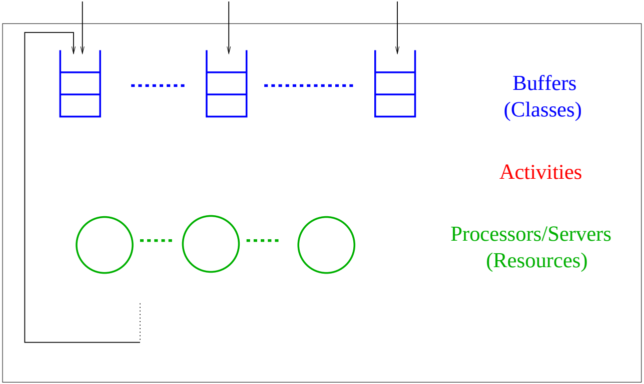
Maximum pressure policies for stochastic processing networks: throughput optimality

Jim Dai

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

Joint work with Wuqin Lin at Kellogg

Stochastic Processing Networks (Harrison 00)

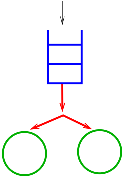


An activity

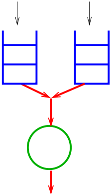
- uses certain resources to
- process certain classes and
- produce certain (possibly different) classes.

Modeling Capability

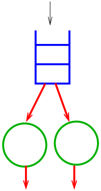
Activities are very general



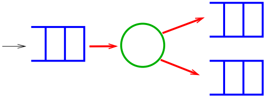
Simultaneous Resource Possession



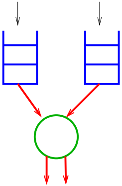
Assembly



Parallel Servers

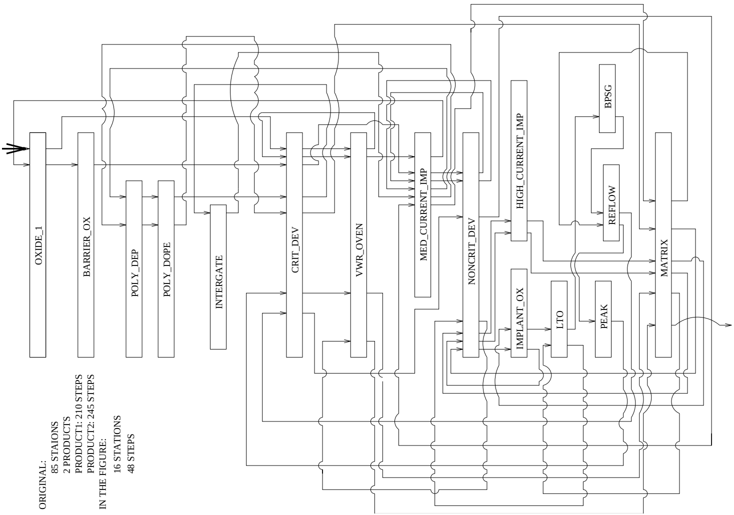


Dynamic Routing



Multiclass Station

Semiconductor Wafer Fabs (Fabrication Facilities)

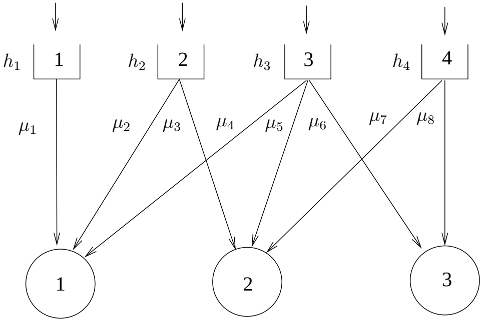


- Multiclass queueing networks

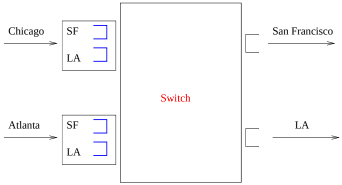
Call Centers

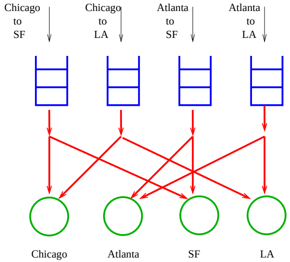
- picture from Larréché et al. 1997

Parallel Server Systems



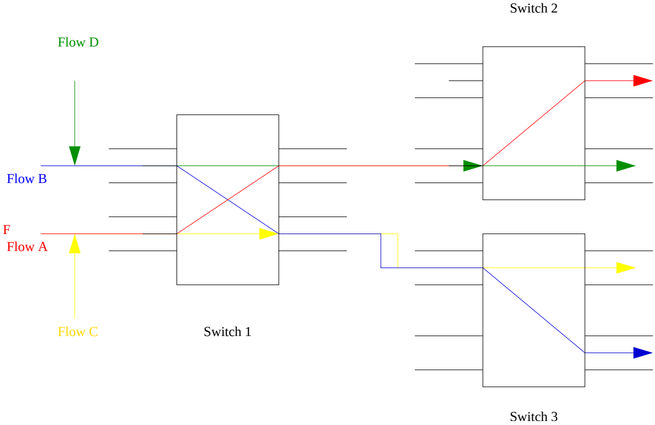
Input Queued Data Switches



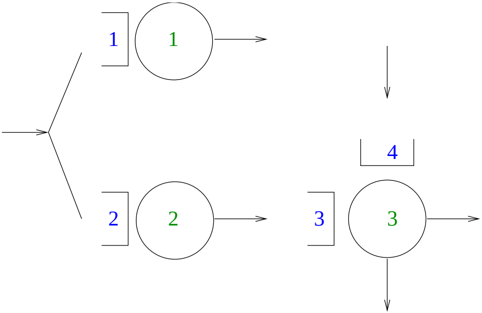


- In each time slot, at most one packet is sent from each input port
- In each time slot, at most one packet is sent to each output port
- Multiple packets can be transferred in a single time slot
- A high speed switch needs to maintain thousands of flows

Networks of Switches



Networks with Alternate Routes



Laws and Louth (1990)

Kelly and Laws (1993)

Dai and Kim (2004)

- Allow dynamic routing decision.
- Model applications in communication networks, supply chains, and road traffic.

Performance Measures

First order ones:

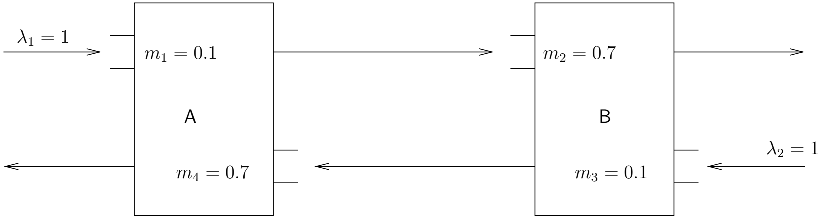
- Throughput: rate at which entities leave a system
- Utilization

Second order ones:

- Cycle time: processing time plus waiting time of an entity; average and variance of cycle time
- Holding cost.

Control decisions can have dramatic impact on key performance measures.

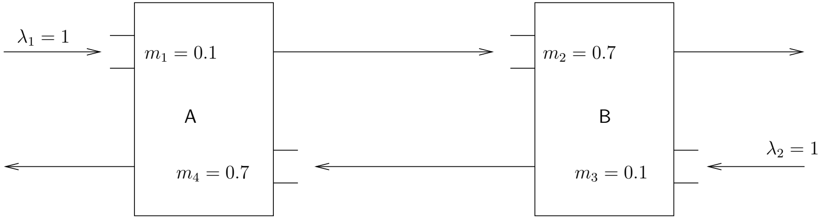
Kumar-Seidman Network



- Traffic intensity:

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

Kumar-Seidman Network

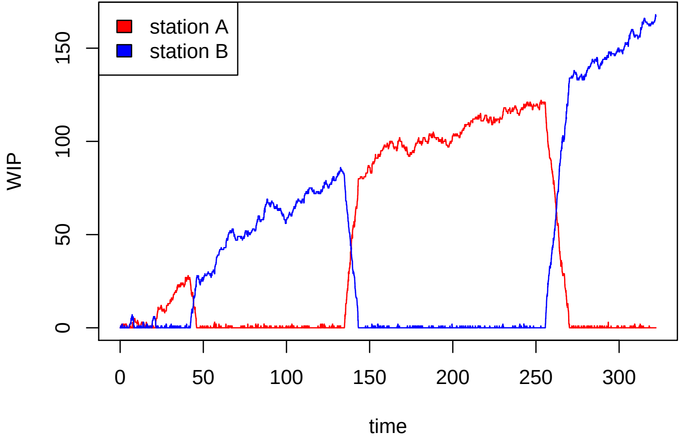


- Traffic intensity:

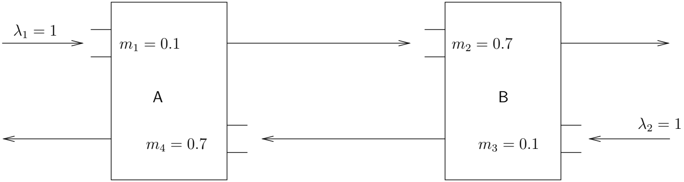
$$\rho_1 = \lambda_1 m_1 + \lambda_2 m_4 = 0.8 \text{ and } \rho_2 = \lambda_1 m_2 + \lambda_2 m_3 = 0.8.$$

- 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.

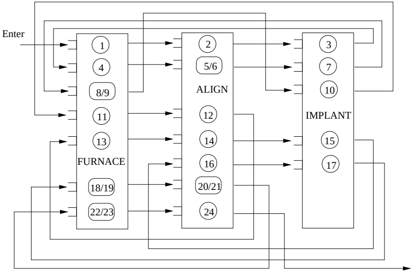
Inefficient Sequencing Policies

- First-in-first-out (FIFO) (Bramson 1994, Seidman 1994)
- $c\mu$ rule (Harrison 99)
- 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: Semiconductor Wafer Fabs

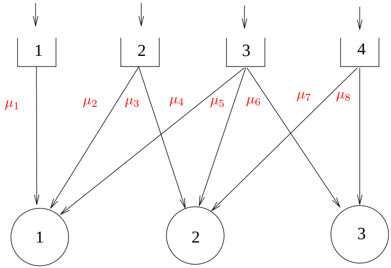


Server k chooses to work on a buffer that has the highest pressure. The pressure at buffer i is

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

Generalization: $\alpha_i Z_i^{\beta}(t)$

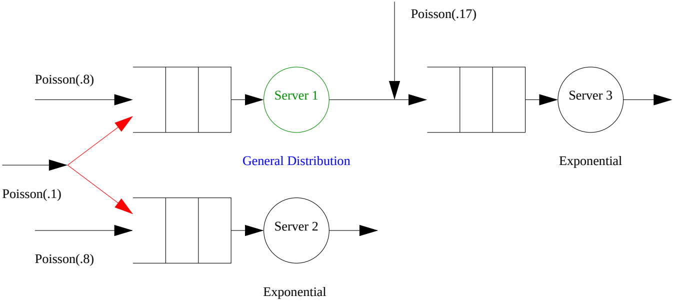
Maximum Pressure Policies: Parallel Server Systems



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

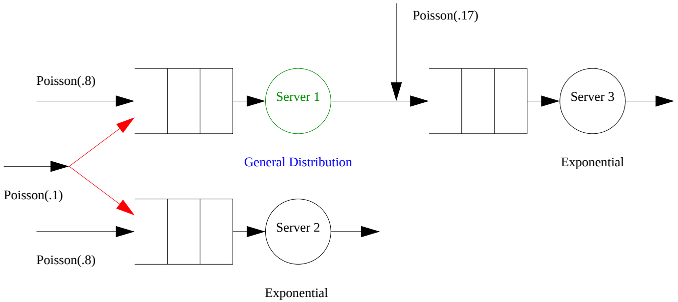
$$\max\{\mu_1 Z_1(t), \mu_2 Z_2(t), \mu_4 Z_3(t)\}.$$

Maximum Pressure Policies: Alternate Routing



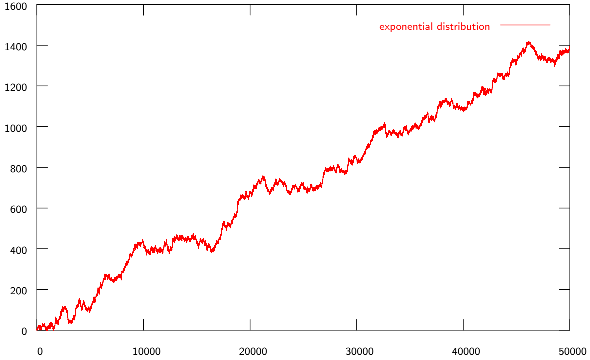
- An MPP translates into: Join-the-shortest-queue and server 1 idles when $Z_3(t) > Z_1(t)$.

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.

Outline of Rest of Talk

3. Main Results – Illustrated by Examples
   - Throughput Optimality
   - Asymptotic Optimality in Heavy Traffic
4. Main Results for General Stochastic Processing Networks
5. Conclusions

Rate Stability

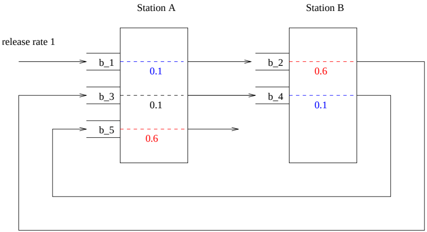
Rate stability

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.

Traffic Intensity



- $\rho_1 = \lambda(1/\mu_1 + 1/\mu_3 + 1/\mu_5), \rho_2 = \lambda(1/\mu_2 + 1/\mu_4)$
- $\rho = \max\{\rho_1, \rho_2\}$: traffic intensity of the network

Theorem

The network is rate stabilizable only if $\rho \leq 1$.

Stability Result

Theorem (Dai-Lin 05)

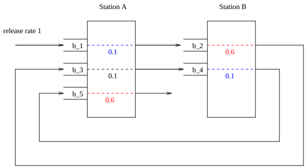
The network is rate stable under maximum pressure policies if it is stabilizable (i.e. $\rho \leq 1$).

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.

Fluid Model Equations



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

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

Fluid Model under MPP

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

- The pressure $p_i = \mu_i(\bar{Z}_i(t) - \bar{Z}_{i+1}(t))$.
- The drift of the quadratic function $f(t) = \sum_i \bar{Z}_i^2(t)/2$ is given by $\dot{f}(t) = \lambda Z_1(t) - \sum_i \dot{\bar{T}}_i(t)p_i$.
- 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.
- Weak stability of the fluid model implies the rate stability of the stochastic network.

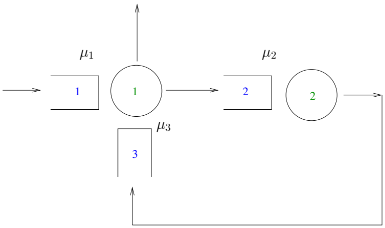
Fluid Limits

- 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

Holding Cost



Assume i.i.d. interarrival times and service times. (variance: σ_a^2 and $\sigma_j^2, j = 1, 2, 3$)

- Objective function: the expected cumulative discounted holding cost:

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

- For example, linear holding cost:

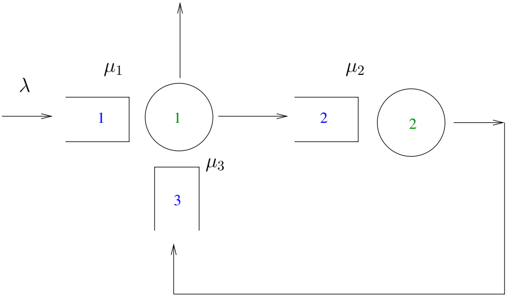
$$h(Z(t)) = h_1 Z_1(t) + h_2 Z_2(t) + h_3 Z_3(t).$$

Heavy Traffic Regime and Diffusion Approximation

- Consider networks in heavy traffic.
- Diffusion Scaling: $\widehat{Z}^r(t) = Z(rt)/\sqrt{r}$.

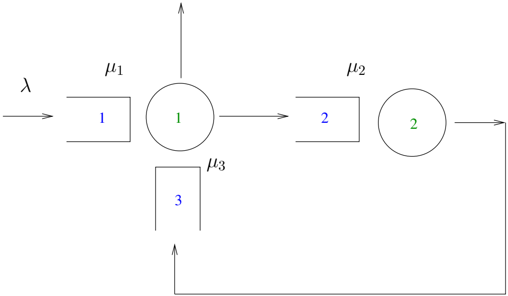
$$\widehat{J}^r_\pi \equiv \mathbb{E}\left(\int_0^\infty e^{-\gamma t} h(\widehat{Z}^r(t))dt\right).$$

Heavy Traffic Condition and Bottlenecks



$$\rho_1 = \lambda/\mu_1 + \lambda/\mu_3$$
$$\rho_2 = \lambda/\mu_2$$

Heavy Traffic Condition and Bottlenecks

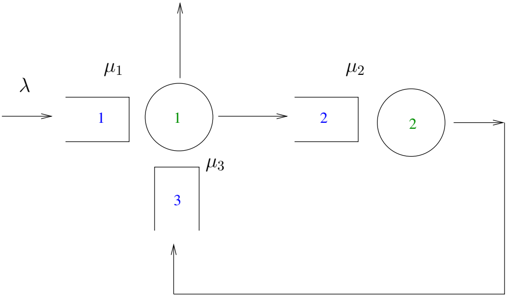


$$\rho_1 = \lambda/\mu_1 + \lambda/\mu_3$$
$$\rho_2 = \lambda/\mu_2$$

Heavy traffic condition

At least one server is critically loaded; allow some servers to be under-utilized (can be unbalanced).

Heavy Traffic Condition and Bottlenecks

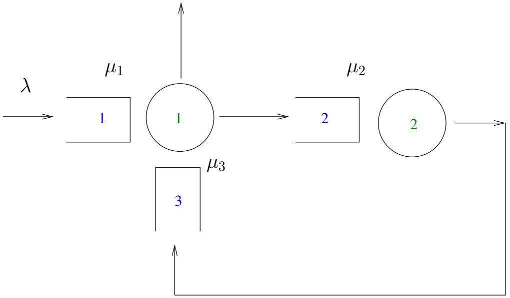


$$\rho_1 = \lambda/\mu_1 + \lambda/\mu_3$$
$$\rho_2 = \lambda/\mu_2$$
$$\lambda = 2$$
$$\mu_1 = 3, \mu_3 = 6, \mu_2 = 4$$

Heavy traffic condition

At least one server is critically loaded; allow some servers to be under-utilized (can be unbalanced).

Heavy Traffic Condition and Bottlenecks

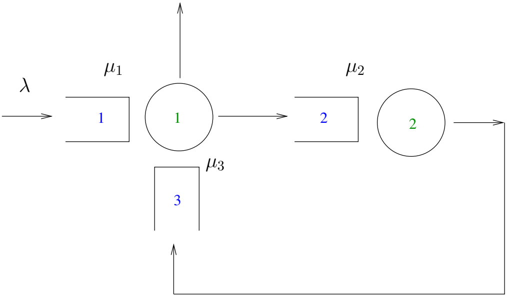


$$\rho_1 = \lambda/\mu_1 + \lambda/\mu_3$$
$$\rho_2 = \lambda/\mu_2$$
$$\lambda = 2$$
$$\mu_1 = 3, \mu_3 = 6, \mu_2 = 4$$
$$\rho_1 = 1, \rho_2 = 0.5$$

Heavy traffic condition

At least one server is critically loaded; allow some servers to be under-utilized (can be unbalanced).

Heavy Traffic Condition and Bottlenecks



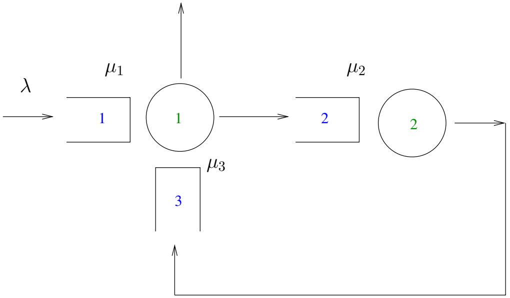
$$\rho_1 = \lambda/\mu_1 + \lambda/\mu_3$$
$$\rho_2 = \lambda/\mu_2$$
$$\lambda = 2$$
$$\mu_1 = 3, \mu_3 = 6, \mu_2 = 4$$
$$\rho_1 = 1, \rho_2 = 0.5$$

Heavy traffic condition

At least one server is critically loaded; allow some servers to be under-utilized (can be unbalanced).

- Bottlenecks: servers that are critically loaded.

Heavy Traffic Condition and Bottlenecks



$\rho_1 = \lambda/\mu_1 + \lambda/\mu_3$

$\rho_2 = \lambda/\mu_2$

$\lambda = 2$

$\mu_1 = 3, \mu_3 = 6, \mu_2 = 2$

$\rho_1 = 1, \rho_2 = 1$

Heavy traffic condition

At least one server is critically loaded; allow some servers to be under-utilized (can be unbalanced).

- Bottlenecks: servers that are critically loaded.

Asymptotic Optimality on Quadratic Holding Cost

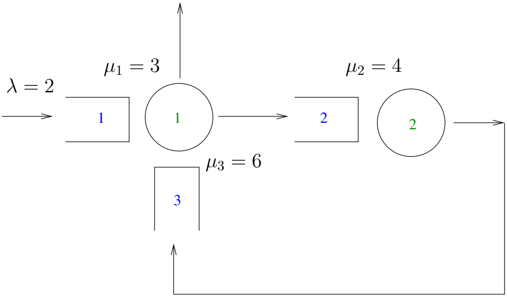
Theorem (Asymptotic Optimality (Dai-Lin 07))

For networks that satisfy the heavy traffic condition and have a single bottleneck, the maximum pressure policy is asymptotically optimal for the quadratic holding cost in that

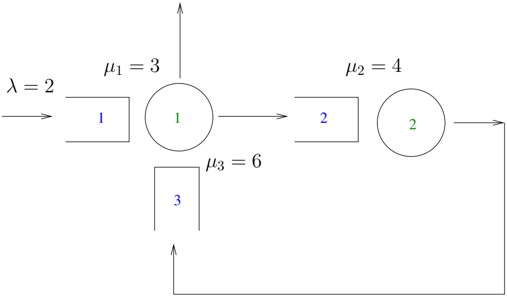
$$\lim_{r\to\infty} \widehat{J}^r_{\mathrm{MPP}} \leq \liminf_{r\to\infty} \widehat{J}^r_{\pi},$$

where $\widehat{J}^r_{\pi} = \mathbb{E}\left(\int_0^\infty e^{-\gamma t}\left(\widehat{Z}^r_1(t)^2 + \widehat{Z}^r_2(t)^2 + \widehat{Z}^r_3(t)^2\right) dt\right)$.

Workload Process



Workload Process



$$y = \left(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}\right)'$$

$$W(t) = y \cdot Z(t)$$

- $W(t) = \frac{1}{2}Z_1(t) + \frac{1}{6}Z_2(t) + \frac{1}{6}Z_3(t)$;
- $\widehat{W}^r(t) = W(rt)/\sqrt{r} = y \cdot \widehat{Z}^r(t)$.

Asymptotic Optimality on Workload Process

Theorem (Workload Optimality (Dai-Lin 07))

For networks that satisfy the heavy traffic condition and have a single bottleneck, the maximum pressure policy is asymptotically optimal for workload in that for each $t \geq 0$ *and* $w > 0$*,*

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

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

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

A Heavy Traffic Limit Theorem

Theorem

For networks that satisfy the heavy traffic condition and have a single bottleneck, under the maximum pressure policy,

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

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

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

$$\|\widehat{Z}^r(\cdot) - \frac{y\widehat{W}^r(\cdot)}{\|y\|^2}\|_T \to 0 \text{ in probability as } r \to \infty.$$

Asymptotic Optimality Proof

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.

A Stochastic Processing Network Model

Basic elements:	Indexes:
$\mathbf{I+1}$ buffers	$i \in \mathcal{I} \cup \{0\}$
$\mathbf{K}$ processors	input and service processors $k \in \mathcal{K}$
$\mathbf{J}$ 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 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 $a \in \mathbb{R}_+^{\mathbf{J}}$ 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 to suit modeling need.

Maximum Pressure Policies: SPNs

$\mathcal{E} = \{a_1, ..., a_u\}$ – the extreme points of $\mathcal{A}$.
$\mathcal{A}(t)$ - the set of feasible allocations at time t.
$\mathcal{E}(t) = \mathcal{A}(t) \cap \mathcal{E}$ - the set of feasible extreme allocations at time t.
At any time t, choose an allocation

$$a \in \arg\max_{a \in \mathcal{E}(t)} \sum_j a_j p_j,$$

where

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

is the pressure under activity j.

Static Plannning 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

If the stochastic processing network operating under any operational policy is rate stable or pathwise stable, the static planning LP has a feasible solution with $\rho \le 1$. Conversely, suppose that Assumption 1 is satisfied. If the static planning LP has a feasible solution with $\rho \le 1$, the stochastic processing network operating under a maximum pressure policy is rate stable.

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).

Asymptotic Optimality

Theorem

For networks that satisfy Assumption 1 and the heavy traffic condition, and have a single bottleneck, the maximum pressure policy is asymptotically optimal for both workload and quadratic holding cost.

- HT condition: The static planning problem has a unique optimal solution (x^*, ρ^*) with $\rho^* = 1$.
- CRP condition (single bottleneck): The dual of the static planning problem has a unique optimal solution (y^*, z^*).

Extensions

- When the networks have more than one bottlenecks, the asymptotic optimality do not hold in general. Ata-Lin (07) proves a heavy traffic limit theorem for maximum pressure policies.
- Lin is generalizing the results to a richer family of maximum pressure policies called β-maximum pressure policies.

Conclusions

- Stochastic processing networks are general.
- Maximum pressure policies are semi-local and do not use arrival rate information.
- Maximum pressure policies are throughput optimal.
- Maximum pressure policies are asymptotic optimal for workload and certain quadratic holding cost.