Control Policies for HGI Performance in Resource Sharing Networks

Amarjit Budhiraja
Department of Statistics & Operations Research
University of North Carolina at Chapel Hill

Joint work with Dane Johnson

Institute for Mathematics and its Applications
Background


- Goal: Allocate capacities of various resources to the various job types in a ‘suitable manner’.


- HGI strives for two things:
  - Workload minimization
  - Minimization of holding cost subject to a given workload.

- Advantages of HGI motivated policies illustrated in HMSY (2014) through simulation and formal analysis.
Background

- **Open Problem** [HMSY(2014)]: Formulate control policies for resource sharing network that achieve HGI performance in the heavy traffic limit.

- **Main Result:**
  - We formulate a set of sufficient conditions under which such policies exist.
  - These conditions cover all the examples studied in HMSY(2014), including the ‘negative example’ [Due to Srikant].
  - One of the main ingredient in implementation is identifying a certain viable ranking map.
  - The policy is given in terms of suitable ‘thresholds’ and ‘safety stocks’.
  - Proof of convergence to HGI performance uses large deviation estimates of the form first introduced in Bell and Williams [(2001), (2005)].
Resource Sharing Networks

- $J$ types of jobs and $I$ resources for processing jobs.

- An $I \times J$ dim. matrix $K$ describes relations between jobs and resources.

- $K_{ij} = 1$ if job $j$ requires processing from resource $i$. 0 otherwise.

- Capacity of resource $i$ is $C_i$.

- Jobs of type $j$ arrive with iid exponential inter arrival times $\{\eta^r_j(k)\}_{k \in \mathbb{N}}$ with rates $\lambda^r_j$. Each such job brings an exponentially distributed work $\{\Delta^r_j(k)\}$ with mean size $1/\mu^r_j$.

- A job of type $j$ is simultaneously processed by all associated resources at a rate $x_j$ decided by the controller.

- Rate allocations must satisfy capacity constraints: $Kx \leq C$.

- A job departs when the integrated flow rate assigned to it equals the size of the job.
Example

- $I = 3, J = 6$
  - Resource 1: $K_{11} = 1, K_{14} = 1, K_{16} = 1, K_{1j} = 0$ otherwise.
  - Resource 2: $K_{22} = 1, K_{24} = 1, K_{25} = 1, K_{26} = 1, K_{2j} = 0$ otherwise.
  - Resource 3 $K_{33} = 1, K_{35} = 1, K_{36} = 1, K_{3j} = 0$ otherwise.
Assumptions

- **Stability and Heavy Traffic:**
  - $C > K \varrho_j^r$, where $\varrho_j^r = \lambda_j^r / \mu_j^r$.
  - $\lim_{r \to \infty} \lambda_j^r = \lambda_j > 0$, $\lim_{r \to \infty} \mu_j^r = \mu_j > 0$.
  - $\lim_{r \to \infty} r(C - K \varrho_j^r) = \nu^* > 0$.

- **Local Traffic:** [Kang, Kelly, Lee, Williams (2009)]

  For each resource $i$ there is a unique job type $\tilde{j}(i)$ that only uses resource $i$. 


State Equations and Control Policies

- **Basic Poisson Processes:**
  \[ A_j^r(t) = \max \left\{ k : \sum_{i=1}^{k} \eta_j^r(i) \leq t \right\}, \quad S_j^r(t) = \max \left\{ k : \sum_{i=1}^{k} \Delta_j^r(i) \leq t \right\}. \]

- **Rate Allocation Policy:** A \( J \)-dimensional, absolutely continuous, nonnegative, non-decreasing, ‘non-anticipative’, stochastic process \( \{B^r(t)\} \)
  \( B_j^r(t) \) is the amount of type \( j \) work processed by time \( t \).

- **Queue Length Process:** \( J \) dimensional.
  \[ Q^r(t) = q^r + A^r(t) - S^r(B^r(t)) \geq 0. \]

- **Capacity Utilization Process:** \( I \) dimensional. \( T^r(t) = KB^r(t) \).
  Require: \( T^r(t) \leq Ct \).

- **Unused Capacity Process:** \( I \) dimensional. \( I^r(t) = Ct - T^r(t) \).
Let $\hat{Q}^r(t) = Q^r(r^2t)/r$ be scaled Queue-length associated with a policy $B^r$.

Consider two types of cost: Fix $h \in \mathbb{R}^J$, $h > 0$.

- Infinite horizon discounted cost:
  \[ J_D^r(B^r, q^r) = \int_0^{\infty} e^{-\theta t} E(h \cdot \hat{Q}^r(t)) \, dt. \]

- Long-term cost per unit time:
  \[ J_E^r(B^r, q^r) = \limsup_{T \to \infty} \frac{1}{T} \int_0^T E(h \cdot \hat{Q}^r(t)) \, dt. \]

**A Difficult Goal:** Construct simple form asymptotically optimal policies.

- $B^{r,*}$ is asymptotically optimal if for any other policy $\{B^r\}$,
  \[ \limsup_{r \to \infty} J^r(B^{r,*}, q^r) \leq \limsup_{r \to \infty} J^r(B^r, q^r). \]

**A Less Ambitious Goal:** [HMSY(2014)] Construct simple form policies that achieve HGI performance asymptotically.
Equivalent Workload Formulation


Let $G^r = K \text{diag}(1/\mu^r)$ and consider the $l$-dimensional workload process

$$\hat{W}^r(t) = G^r \hat{Q}^r(t).$$

Two step procedure for constructing asymptotically optimal control

- construct an asymptotically optimal workload process $\hat{W}^{r,*}(t)$.
- construct an optimal way to distribute workload among various job-types.

First Step:

$$\mathcal{C}(w) \triangleq \inf \{ h \cdot q : q \geq 0, Gq = w \}, \ w \in \mathbb{R}^I_+.$$ 

Let $\hat{W}^{r,*}$ minimize the cost

$$\limsup_{T \to \infty} \frac{1}{T} \int_0^T E \left( \mathcal{C}(\hat{W}^r(t)) \right) \, dt.$$ 

Second Step: Let $q^*: \mathbb{R}^I_+ \to \mathbb{R}^J_+$ be a continuous map s.t. $h \cdot q^*(w) = \mathcal{C}(w)$. Then

$$\hat{Q}^{r,*} \approx q^*(\hat{W}^{r,*}).$$
EWF: Asymptotic Formulation

- **State Equation:**
  \[ W(t) = w + \Lambda X(t) - v^*t + Y(t) \geq 0 \]
  where \( \Lambda \) is \( I \times J \) with rank \( I \), \( X \) is a \( J \)-dimensional BM.

- **Here** \( Y \) is the control process – nondecreasing, nonnegative, nonanticipative, RCLL.

- **Cost Function:**
  \[ \tilde{J}(Y, w) = \limsup_{T \to \infty} \frac{1}{T} \int_0^T E(C(W(t))) \]
EWF: Asymptotic Formulation

- In general optimal $Y^*$ in the asymptotic formulation is hard to get.
- Instead consider the $\tilde{Y}$ that minimizes $W$ coordinate wise.
- The corresponding state process $\tilde{W}$ is a $I$-dimensional RBM:
  \[ \tilde{W}(t) = \Gamma(w + \Lambda X - \nu^* \iota)(t). \]
- Let $\pi$ be the unique stationary distribution of the RBM. Then the cost associated with $\tilde{Y}$ is
  \[ \int_{\mathbb{R}^I_+} C(w) \pi(dw) = HGI. \]
- If $C$ is monotonic, $\tilde{Y}$ is in fact optimal and $HGI$ is the optimal cost in the EWF.
Main Result

- **Restatement of HMSY Open Problem**: Construct a simple form policy $B^r$ such that

$$\lim_{r \to \infty} J^r(B^r, q^r) = HGI = \int_{\mathbb{R}_+^\text{c}} C(w)\pi(dw).$$

- **Main Result**: Under the assumption that there exists a ranking map for the job-types, there is an explicit policy $B^r$ that achieves HGI asymptotically.

- The policy is given in terms of certain thresholds/safety stocks of order $r^\alpha$ (with $\alpha \in (0, 1/2)$).

- Proofs use large deviations estimates and Lyapunov function constructions.
Main Result

- In general finding a ranking map may be computationally hard but in many cases it takes a simple explicit form.

- In particular all examples in HMSY\([2\text{LLN}, 3\text{LLN}, \text{C3LN}]\) are covered with a simple and explicit ranking map.

- The ‘negative’ example in HMSY also admits a simple ranking map.

- All ‘linear networks’ [Massoulié and Roberts] are covered.

- If \(C\) is monotonic then the result gives an asymptotically optimal policy.

- There are non-monotonic \(C\) for which a (explicit) ranking exists and there are monotonic \(C\) for which the ranking does not exist.
Three Types of Jobs

- A type $j$ arriving job produces a holding cost $h \cdot e_j = h_j$.

- The resulting increase in workload vector is
  
  $$
  \mu_j^{-1}[K_{1,j}, K_{2,j}, \ldots K_{I,j}]' = g_j.
  $$

- The best way to distribute this amount of workload among queues produces the cost $C(g_j)$. Note $C(g_j) \leq h_j$.

- If $C(g_j) < h_j$ we should get rid of this job.

- Such job types are referred to as primary. Collection of primary jobs denoted as $S^p$. 
Three Types of Jobs

- All remaining jobs are called secondary, the collection of such jobs is $S^s$.

- Jobs that require service from only one resource are always secondary.

- Let $S^m$ be collection of all secondary jobs that need more than one resource and $S^1$ the jobs that require only one resource.

$$S^s = S^m \cup S^1.$$  

- Then jobs can be partitioned as $S^p \cup S^1 \cup S^m$. 
A Final Assumption

- **Definition** Let $|S^m| = M$. A ranking is a map $\mathcal{R} : \{1, \ldots, M\} \rightarrow S^m$ with certain properties.

- **Assumption** There exists a ranking map $\mathcal{R}$.

- Given a ranking map, under our policy, the job-types have a hierarchy of the form:

$$S^p \succ S^1 \succ \mathcal{R}(M) \succ \mathcal{R}(M-1) \cdots \succ \mathcal{R}(1).$$

- Recall that $K\varrho = C$. Can interpret $\varrho_j$ as the nominal capacity allocation to type $j$ jobs.

- The jobs higher in the hierarchy are ‘more expensive’ and, under the policy, will see more than nominal allocation in a suitable sense.

- A key consequence of the existence of a ranking is that it gives a simple form explicit minimizer $q^*(w)$ for the LP problem.
Resource Allocation Policy

- Let $0 < \alpha < 1/2$, $0 < c_1 < c_2$.

$$\tau^j_{2l} = \inf\{t \geq \tau^j_{2l-1}: Q^r_j(t) \geq c_2 r^\alpha\},$$
$$\tau^j_{2l+1} = \inf\{t \geq \tau^j_{2l}: Q^r_j(t) < c_1 r^\alpha\},$$

- $\tau_{\text{odd}}$ is when queue is depleted and $\tau_{\text{even}}$ is the next time it is stocked.

- No processing between $\tau_{\text{odd}}$ and $\tau_{\text{even}}$: Let

$$\mathcal{E}^r_j(t) = 1\{t \in [\tau^j_{2l-1}, \tau^j_{2l}) \text{ for some } l > 0\}^c$$

Rate allocation policy $B^r(t) = \int_{[0,t]} x^r(s)ds$ satisfies

$$x^r_j(t) = y^r_j(t)\mathcal{E}^r_j(t)$$
Resource Allocation Policy

- $\delta \equiv \frac{\min_j \varrho_j}{2J}$.

- **Stocked Job types:**
  \[ \sigma^r(t) \equiv \{ j : Q_j^r(t) \geq c_2 r^\alpha \} \]

- Resources associated with stocked job types:
  \[ \hat{\omega}^r(t) \equiv \{ i : \text{for some } j \in \sigma^r(t), K_{ij} = 1 \}. \]

- Job types with rank higher than $k$ associated with resource $i$.
  \[ \zeta_i^k \equiv \{ \text{job types associated with resource } i \text{ not in } \{R(1), \ldots R(k)\} \}. \]
Resource Allocation Policy

**Primary jobs.** For $j \in S^p$

$$y_j(t) \doteq [\varrho_j + \delta]1_{\{j \in \sigma^r(t)\}} + [\varrho_j - \frac{\delta}{J2M+3}]1_{\{j \notin \sigma^r(t)\}}$$

**Jobs in $S^m$.** For $k \in \{1, \ldots, M\}$

$$y_{\mathcal{R}(k)}(t) \doteq \begin{cases} 
\varrho_{\mathcal{R}(k)} - 2^k - M - 2\delta, & \text{if } \zeta_i^k \cap \sigma^r(t) \neq \emptyset \text{ for all } i \in \mathcal{N}_{\mathcal{R}(k)} \\
\varrho_{\mathcal{R}(k)} + 2^k - M - 2\delta, & \text{if } \zeta_i^k \cap \sigma^r(t) = \emptyset \text{ for some } i \in \mathcal{N}_{\mathcal{R}(k)} \text{ and } \mathcal{R}(k) \in \sigma^r(t) \\
\varrho_{\mathcal{R}(k)} - 2^{-k} - M - 2\delta, & \text{if } \zeta_i^k \cap \sigma^r(t) = \emptyset \text{ for some } i \in \mathcal{N}_{\mathcal{R}(k)} \text{ and } \mathcal{R}(k) \notin \sigma^r(t). 
\end{cases}$$

**Jobs in $S^1$.** For $j \in S^1$

$$y_j(t) \doteq [C_{\hat{i}(j)} - \sum_{l \neq j: K_{\hat{i}(j)}, l = 1} y_l(t)]1_{\{\hat{i}(j) \in \omega^r(t)\}} + [\varrho_j - \delta]1_{\{\hat{i}(j) \notin \omega^r(t)\}}$$
Primary jobs. For $j \in S^p$

$$y_j(t) = [\varrho_j + \delta]1_{\{j \in \sigma^r(t)\}} + \left[\varrho_j - \frac{\delta}{J2M+3}\right]1_{\{j \notin \sigma^r(t)\}}$$

- If the associated queue is stocked then it gets higher than nominal rate allocation and otherwise a lower than nominal allocation.
Resource Allocation Policy: Discussion

**Jobs in** $S^m$. For $k \in \{1, \ldots, M\}$

\[
y_R(k)(t) = \begin{cases} 
\varrho_R(k) - 2^{k-M-2}\delta, & \text{if } \zeta_i^k \cap \sigma^r(t) \neq \emptyset \text{ for all } i \in N_R(k) \\
\varrho_R(k) + 2^{k-M-2}\delta, & \text{if } \zeta_i^k \cap \sigma^r(t) = \emptyset \text{ for some } i \in N_R(k) \text{ and } R(k) \in \sigma^r(t) \\
\varrho_R(k) - 2^{-k-M-2}\delta, & \text{if } \zeta_i^k \cap \sigma^r(t) = \emptyset \text{ for some } i \in N_R(k) \text{ and } R(k) \notin \sigma^r(t).
\end{cases}
\]

- **Consider** $j = R(M)$.
  - **First line**: Every associated resource has at least one job-type rated higher with a stocked queue. Rate allocated to job-type $R(M)$ is lower than nominal.
  - **Second line**: There is at least one associated resource such that none of its job-types that are rated higher that $R(M)$ has a stocked queue and the queue for job-type $R(M)$ is stocked. Allocate a flow rate higher than nominal.
  - **Third line**: Allocate Lower than nominal flow rate if the queue $R(M)$ is not stocked.
Resource Allocation Policy: Discussion

Jobs in $S^1$. For $j \in S^1$

$$y_j(t) \doteq [C_{\hat{i}(j)} - \sum_{l \neq j: K_{\hat{i}(j), l} = 1} y_l(t)]1_{\{i(j) \in \omega^r(t)\}} + [\rho_j - \delta]1_{\{i(j) \notin \omega^r(t)\}}$$

- If the resource $\hat{i}(j)$ has some stocked queue, allocate job $j$ all remaining capacity of resource $\hat{i}(j)$.

- If not, assign less than nominal allocation.
Discussion of the Ranking Map

- **Example:** \( I = 4, J = 7, \mu_i = 1 \) for all \( i \) and \( h_1 = h_2 = h_3 = h_4 = 4, h_5 = 6, h_6 = 7, h_7 = 13 \).

- In this example \( S^p = \emptyset, S^m = \{5, 6, 7\} \).

- \( R(1) \) will be the least expensive job among \( \{5, 6, 7\} \)

- In this example \( R(1) = 7 \).
Discussion of the Ranking Map

$h_1 = h_2 = h_3 = h_4 = 4$, $h_5 = 6$, $h_6 = 7$, $h_7 = 13$
Sufficient Conditions for Existence of the Ranking Map

- Any network with $S^m = \emptyset$ trivially satisfies the condition.

- Consider the network with $I = 3$, $J = 6$ and $h_1 = h_2 = h_3 = 1$, $h_4 = h_5 = h_6 = 4$. Here $S^m = \emptyset$. 

![Diagram showing a network with nodes and flow rates labeled $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, $\lambda_5$, and $\lambda_6$.]
Sufficient Conditions for Existence of the Ranking Map

- Any network with $S^m$ a singleton also trivially satisfies the condition.
- In particular any linear network (e.g. 2LLN, 3LLN of HMSY) satisfies the condition.
Sufficient Conditions for Existence of the Ranking Map

- Let $N_j$ denote the set of resources associated with job-type $j$.

- For all $j, k \in S^m$, either $N_j \subset N_k$ or $N_k \subset N_j$ or $N_j \cap N_k = \emptyset$. 

\[
\begin{array}{c}
\lambda_1 \\
\lambda_2 \\
\lambda_3 \\
\lambda_4 \\
\lambda_5 \\
\lambda_6 \\
\lambda_7
\end{array}
\]
Another Sufficient Condition for Existence of the Ranking Map

- All **expensive** jobs have the property that their every **minimal cover** is **efficient**.

- The example C3LN of HMSY satisfies this condition. Here \( I = 3, J = 6 \) and \( h_i = 1 \) for all \( i \). A ranking map is given as \( R(1) = 6, R(2) = 4, R(3) = 5 \).
Another Sufficient Condition for Existence of the Ranking Map

- All expensive jobs have the property that their every minimal cover is efficient.

Example: \( I = 4, J = 8, h_j = 1 \) for all \( j \). Here \( R(8) = 1, R(5) = 2, R(6) = 3, R(7) = 4 \).
A Second Sufficient Condition for Existence of the Ranking Map

- All **expensive** jobs have the property that their every **minimal cover** is **efficient**.

- The ‘negative example’ of HMSY satisfies the sufficient condition. \( I = 6, J = 9, h_j = 1 \).
  \( \mathcal{R}(1) = 7, \mathcal{R}(2) = 8, \mathcal{R}(3) = 9 \).
An Example where Ranking Does not exist

Suppose $I = 3, J = 6, \mu_i = 1$ for all $i$,

$$h_1 = h_2 = h_3 = 5, h_4 = 7, h_5 = 8, h_6 = 11.$$ 

Here $S^m = \{4, 5, 6\}$. Ranking does not exist.
An Example where Ranking Does not exist

- Workload cost and its minimizer. The workload $C$ for this example can be given explicitly as follows. Let for $w \in \mathbb{R}^3_+$, $w_{12} = w_1 \wedge w_2$, $w_{23} = w_2 \wedge w_3$, $w_{123} = w_1 \wedge w_2 \wedge w_3$.

$$C(w) = \begin{cases} 
5w_2 + 2w_1 + 3w_3, & \text{if } w_2 \geq w_1 + w_3 \\
3w_1 + 4w_2 + 4w_3, & \text{if } w_1 + w_3 > w_2 \geq w_1 \lor w_3 \\
5(w_1 + w_2 + w_3) + w_{123} - 3w_{12} - 2w_{23}, & \text{if } w_1 \lor w_3 > w_2
\end{cases}$$

- Optimal $q^*(w)$

$$q^*(w) = \begin{cases} 
(0, w_2 - w_1 - w_3, 0, w_1, w_3, 0), & \text{if } w_2 \geq w_1 + w_3 \\
(0, 0, 0, w_2 - w_3, w_2 - w_1, w_1 + w_3 - w_2), & \text{if } w_1 + w_3 > w_2 \geq w_1 \lor w_3 \\
(w_1 - w_{12}, w_2 + w_{123} - w_{12} - w_{23}, w_3 - w_{23}, w_{12} - w_{123}, w_{23} - w_{123}, w_{123}), & \text{if } w_1 \lor w_3 > w_2
\end{cases}$$

- Note that $C$ is nondecreasing. In particular the HGI performance is also the optimal cost in the associated BCP.