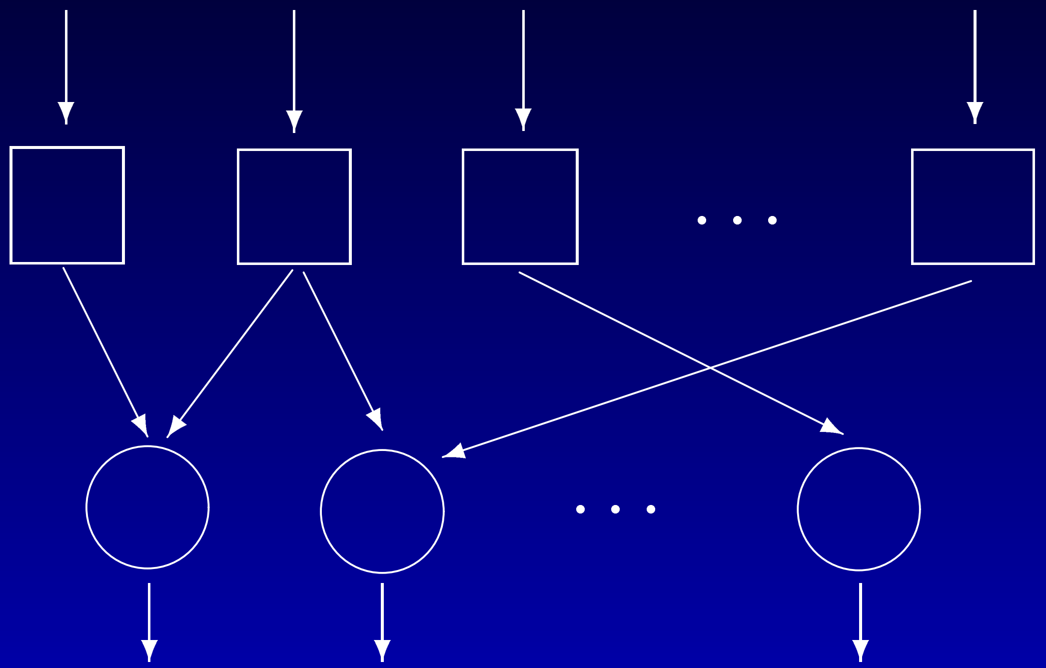
Dynamic Scheduling of a Parallel Server System in Heavy Traffic with Complete Resource Pooling: Asymptotic Optimality of a Threshold Policy

Steven L. Bell

University of California, San Diego

The Parallel Server System

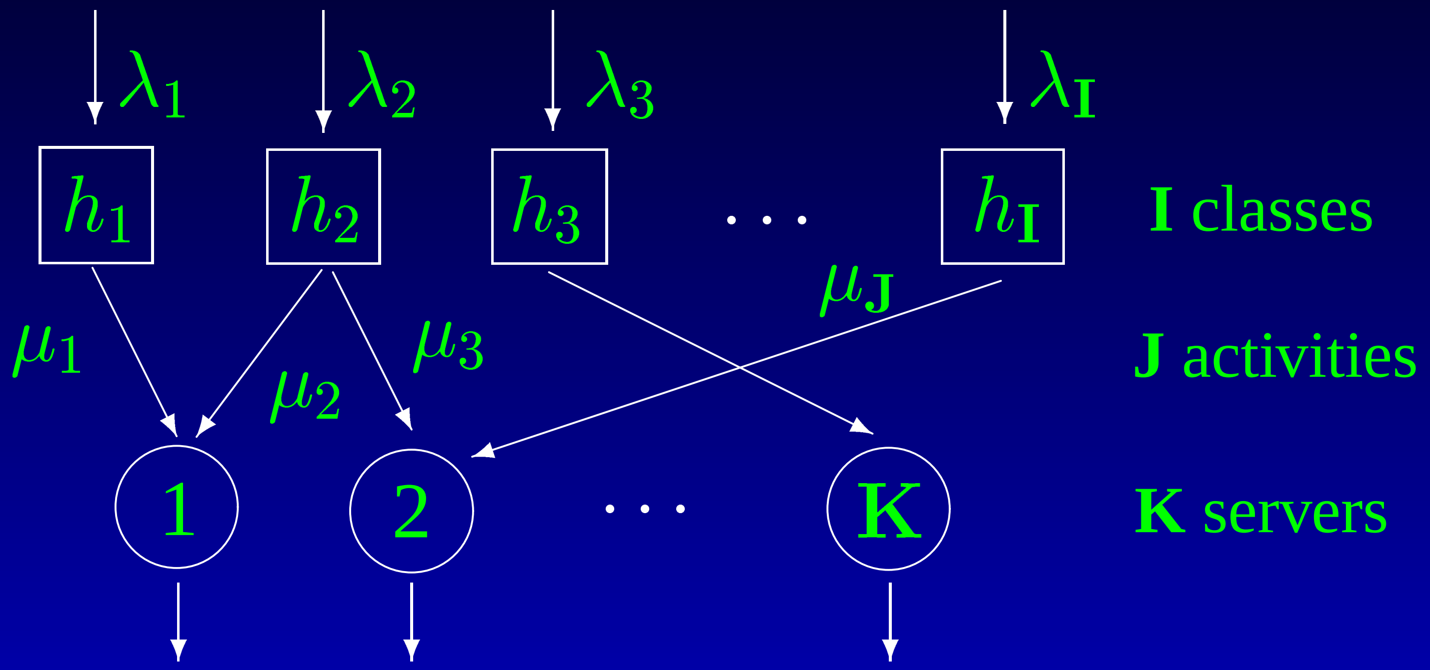


I classes

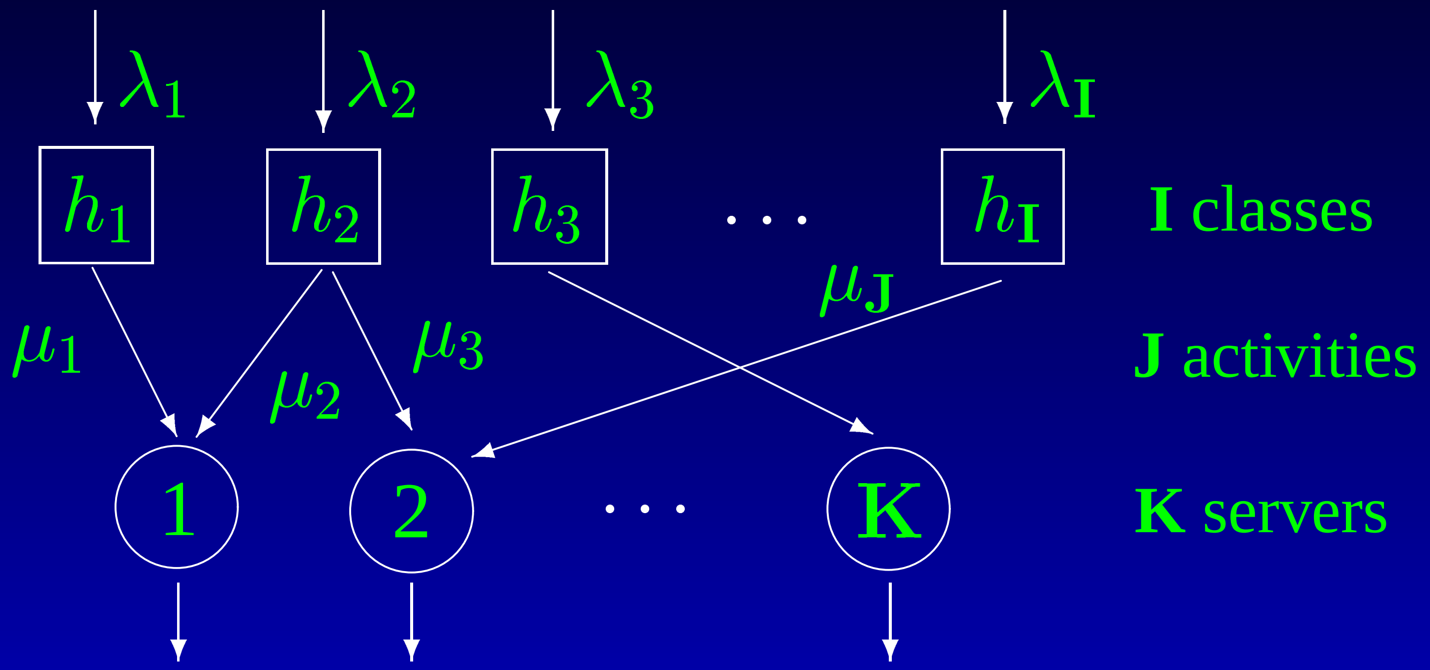
J activities

K servers

The Parallel Server System



The Parallel Server System



J-dimensional Scheduling Control Process $T(\cdot)$

Brownian Methodology

Harrison (1988,2000), Harrison-Van Mieghem (1997)

- Formulate a queueing system control problem

Brownian Methodology

Harrison (1988,2000), Harrison-Van Mieghem (1997)

- Formulate a queueing system control problem
- Formulate a notion of heavy traffic

Brownian Methodology

Harrison (1988,2000), Harrison-Van Mieghem (1997)

- Formulate a queueing system control problem
- Formulate a notion of heavy traffic
- Formulate a Brownian control problem (BCP) and reduce the BCP to an equivalent workload problem (EWP)

Brownian Methodology

Harrison (1988,2000), Harrison-Van Mieghem (1997)

- Formulate a queueing system control problem
- Formulate a notion of heavy traffic
- Formulate a Brownian control problem (BCP) and reduce the BCP to an equivalent workload problem (EWP)
- Analyze the BCP (or EWP) and “interpret” its solution to generate a candidate control policy for the original queueing system

Brownian Methodology

Harrison (1988,2000), Harrison-Van Mieghem (1997)

- Formulate a queueing system control problem
- Formulate a notion of heavy traffic
- Formulate a Brownian control problem (BCP) and reduce the BCP to an equivalent workload problem (EWP)
- Analyze the BCP (or EWP) and "interpret" its solution to generate a candidate control policy for the original queueing system
- Prove that the candidate policy is asymptotically optimal and that it achieves the same cost as the solution to the BCP

Queueing Model

$$Q_i(t) = A_i(t) - \sum_{j=1}^{\mathbf{J}} \mathbf{C_{ij}} S_j(T_j(t))$$

$$I_k(t) = t - \sum_{j=1}^{\mathbf{J}} \mathbf{A_{kj}} T_j(t)$$

Queueing Model

$$
\begin{aligned}
Q(t) &= A(t) - \mathbf{C}S(T(t)) \\
I(t) &= \mathbf{1}t - \mathbf{A}T(t)
\end{aligned}
$$

Assumption: (A(t) , S(t)) are independent renewal processes satisfying exponential moment condition.

Queueing Model

$$\begin{aligned} Q(t) &= A(t) - \mathbf{C}S(T(t)) \\ I(t) &= \mathbf{1}t - \mathbf{A}T(t) \end{aligned}$$

- $T_j(\cdot)$ is Lipschitz continuous (Lip. constant ≤ 1), non-decreasing, $T_j(0) = 0$,
- $I_k(\cdot)$ is continuous, non-decreasing, $I_k(0) = 0$,
- $Q_i(\cdot) \geq 0$.

Heavy Traffic

LP formulation
(Harrison–1999, Williams–2000):

minimize ρ

subject to:

$$\mathbf{R}x = \lambda, \quad \mathbf{A}x \le \rho\,\mathbf{1}, \quad x \ge 0$$

Heavy Traffic

LP formulation
(Harrison–1999, Williams–2000):

minimize ρ

subject to:

$$\mathbf{R}x = \lambda, \quad \mathbf{A}x \leq \rho\,\mathbf{1}, \quad x \geq 0$$

LP has a unique solution (ρ^*, x^*), and that solution satisfies $\rho^* = 1$, $\mathbf{A}x^* = \mathbf{1}$
x^* is called the nominal allocation vector

Assumption: $x_j^* > 0$ for $j \leq \mathbf{B}$

Queueing Control Problem

Diffusion Scaling:

$$
\begin{aligned}
\hat{A}^r(t) &= r^{-1}\left(A(r^2t) - \lambda r^2 t\right) \\
\hat{S}^r(t) &= r^{-1}\left(S(r^2t) - \mu r^2 t\right) \\
\hat{Q}^r(t) &= r^{-1}\hat{Q}^r(r^2t) \\
\hat{I}^r(t) &= r^{-1}\hat{I}^r(r^2t)
\end{aligned}
$$

Fluid Scaling: $\bar{T}^r(t) = r^{-2}\,\bar{T}^r(r^2t)$

Queueing Control Problem

Diffusion Scaling:

$$\hat{A}^r(t) = r^{-1}\big(A(r^2t) - \lambda r^2 t\big)$$

$$\hat{S}^r(t) = r^{-1}\big(S(r^2t) - \mu r^2 t\big)$$

Fluid Scaling: $\bar{T}^r(t) = r^{-2}\,\bar{T}^r(r^2t)$

$$\hat{Q}^r(t) = \hat{A}^r(t) - \mathbf{C}\hat{S}^r(\bar{T}^r(t)) + \mathbf{R}\hat{Y}^r(t)$$

$$\hat{I}^r(t) = \mathbf{A}\hat{Y}^r(t)$$

$$\hat{Y}^r(t) = r^{-1}\,(\,x^*\,r^2t - T^r(r^2t)\,)$$

Queueing Control Problem

$$\hat{J}^r(T^r) = \mathbf{E}\left(\int_0^\infty e^{-\gamma t} h \cdot \hat{Q}^r(t)\, dt\right)$$

$$\begin{aligned}
\hat{Q}^r(t) &= \hat{A}^r(t) - \mathbf{C}\hat{S}^r(\bar{T}^r(t)) + \mathbf{R}\hat{Y}^r(t) \\
\hat{I}^r(t) &= \mathbf{A}\hat{Y}^r(t) \\
\hat{I}_k(\cdot) &\quad \text{is} \quad \text{continuous, non-decreasing,} \\
\hat{I}_k(0) &= 0, \quad \hat{Q}_i(\cdot) \geq 0.
\end{aligned}$$

Brownian Control Problem

$$\text{minimize} \qquad \mathbf{E}\left(\int_0^\infty e^{-\gamma t} h \cdot \tilde{Q}(t)\, dt\right)$$

by choosing $\tilde{Y}(\cdot)$ such that

$$\begin{aligned}
\tilde{Q}(t) &= \tilde{X}(t) + \mathbf{R}\tilde{Y}(t) \\
\tilde{I}(t) &= \mathbf{A}\tilde{Y}(t) \\
\tilde{I}(\cdot) &\text{ is non-decreasing and } \tilde{I}(0) > 0 \\
\tilde{Y}_j(\cdot) &\text{ is non-increasing and } \tilde{Y}_j(0) \leq 0, \; j > \mathbf{B}, \\
\tilde{Q}(t) &\geq 0
\end{aligned}$$

Solution to BCP (Harrison-López, 1999; Williams, 2000)

The following are equivalent (Complete Resource Pooling Condition):

- The dual LP has a unique solution $(y^*, z^*) > 0$
- $\mathbf{B} = \mathbf{I} + \mathbf{K} - 1$
- the buffer-server graph linked by basic activities is a tree

Solution to BCP (Harrison-López, 1999; Williams, 2000)

$$\tilde{Q}^*_{i^*} = \tilde{W}^* \;/\; y^*_{i^*} \qquad \tilde{Q}^*_i = \mathbf{0} \ \text{ for all } \ i \neq i^*$$
$$\tilde{I}^*_{k^*} = \tilde{V}^* \;/\; z^*_{k^*} \qquad \tilde{I}^*_k = \mathbf{0} \ \text{ for all } \ k \neq k^*$$

$$i^* \in \operatorname*{argmin}_i \left\{ \frac{h_i}{y^*_i} \right\}$$

$$\tilde{W}^*(\cdot) = y^* \cdot \tilde{X}(\cdot) + \tilde{V}^*(\cdot)$$
$$\tilde{V}^*(t) = \sup_{0 \le s \le t} \left(-\, y^* \cdot \tilde{X}(s) \right)$$

Solution to BCP (Harrison-López, 1999; Williams, 2000)

$$\tilde{Q}^*_{i^*} = \tilde{W}^* \,/\, y^*_{i^*} \qquad \tilde{Q}^*_i = \mathbf{0} \ \text{ for all } \ i \neq i^*$$

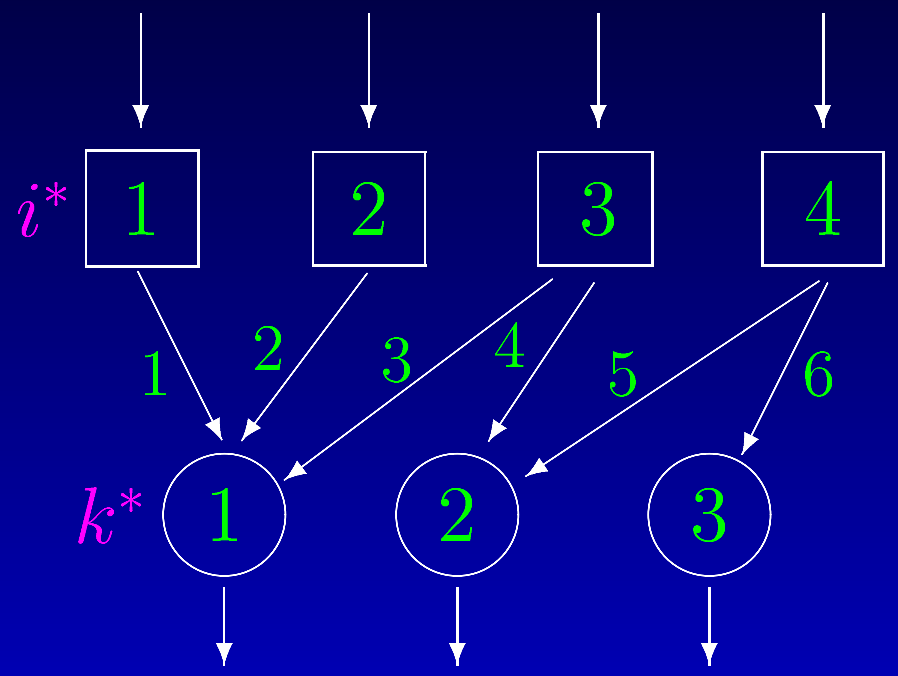
$$\tilde{I}^*_{k^*} = \tilde{V}^* \,/\, z^*_{k^*} \qquad \tilde{I}^*_k = \mathbf{0} \ \text{ for all } \ k \neq k^*$$

Minimum cost:

$$J^* \equiv \mathbf{E}\left(\int_0^\infty e^{-\gamma t} h \cdot \tilde{Q}^*(t)\, dt\right)$$

Threshold Policy (Williams, 2000)

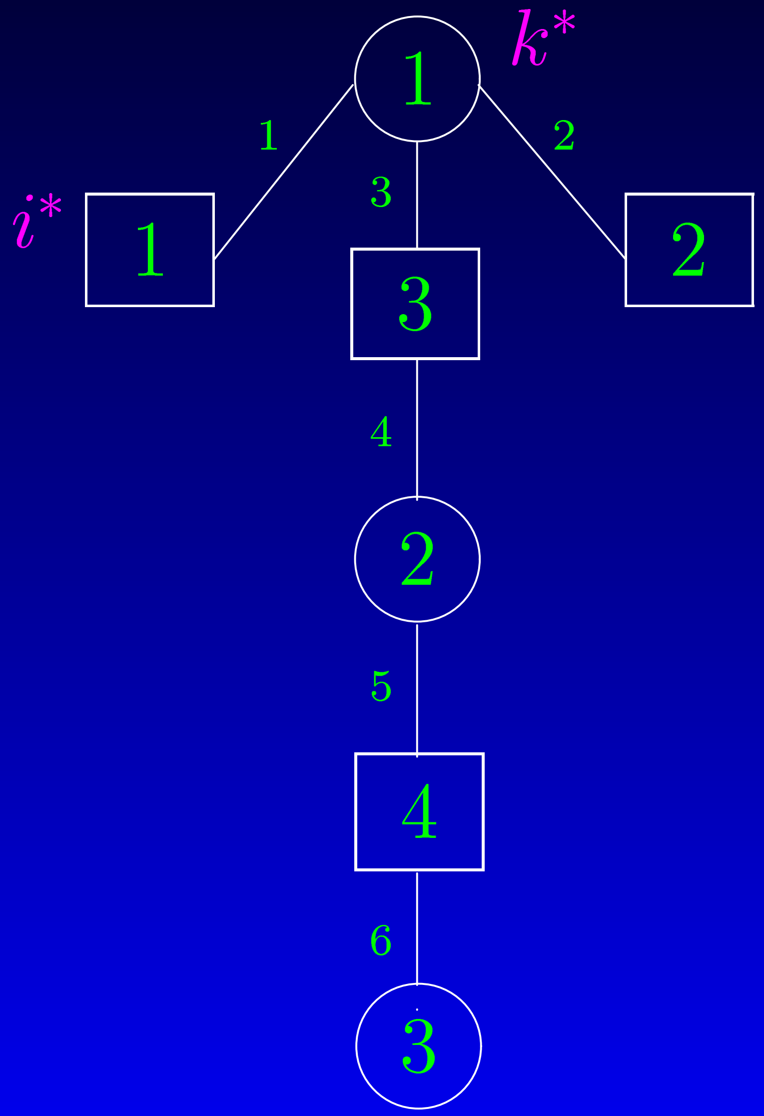
Example:



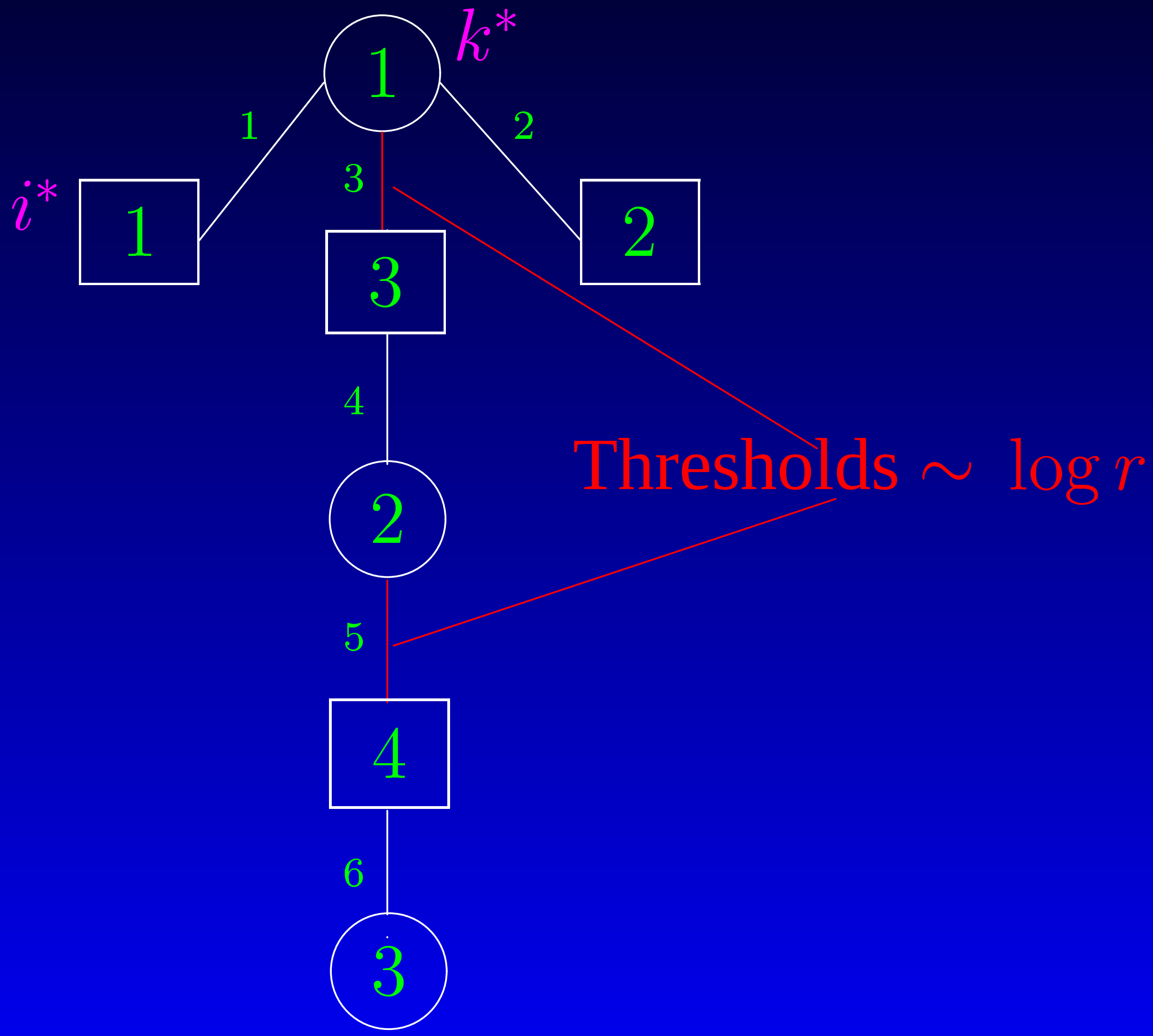
B = 6, **I** = 4, **K** = 3

The Threshold Policy uses only basic activities

Threshold Policy (Williams, 2000)



Threshold Policy (Williams, 2000)



Asymptotic Optimality

Theorem: The Threshold Policy is asymptotically optimal for the general parallel server system under the complete resource pooling condition:

$$
\begin{aligned}
& \varliminf_{r\to\infty} \mathbf{E}\left(\int_0^\infty e^{-\gamma t} h\cdot\hat{Q}^r(t)dt\right) \\
\geq \quad & \varliminf_{r\to\infty} \mathbf{E}\left(\int_0^\infty e^{-\gamma t} h\cdot\hat{Q}^{r,*}(t)dt\right) \\
= \quad & J^* < \infty
\end{aligned}
$$

State Space Collapse

Residual Process:

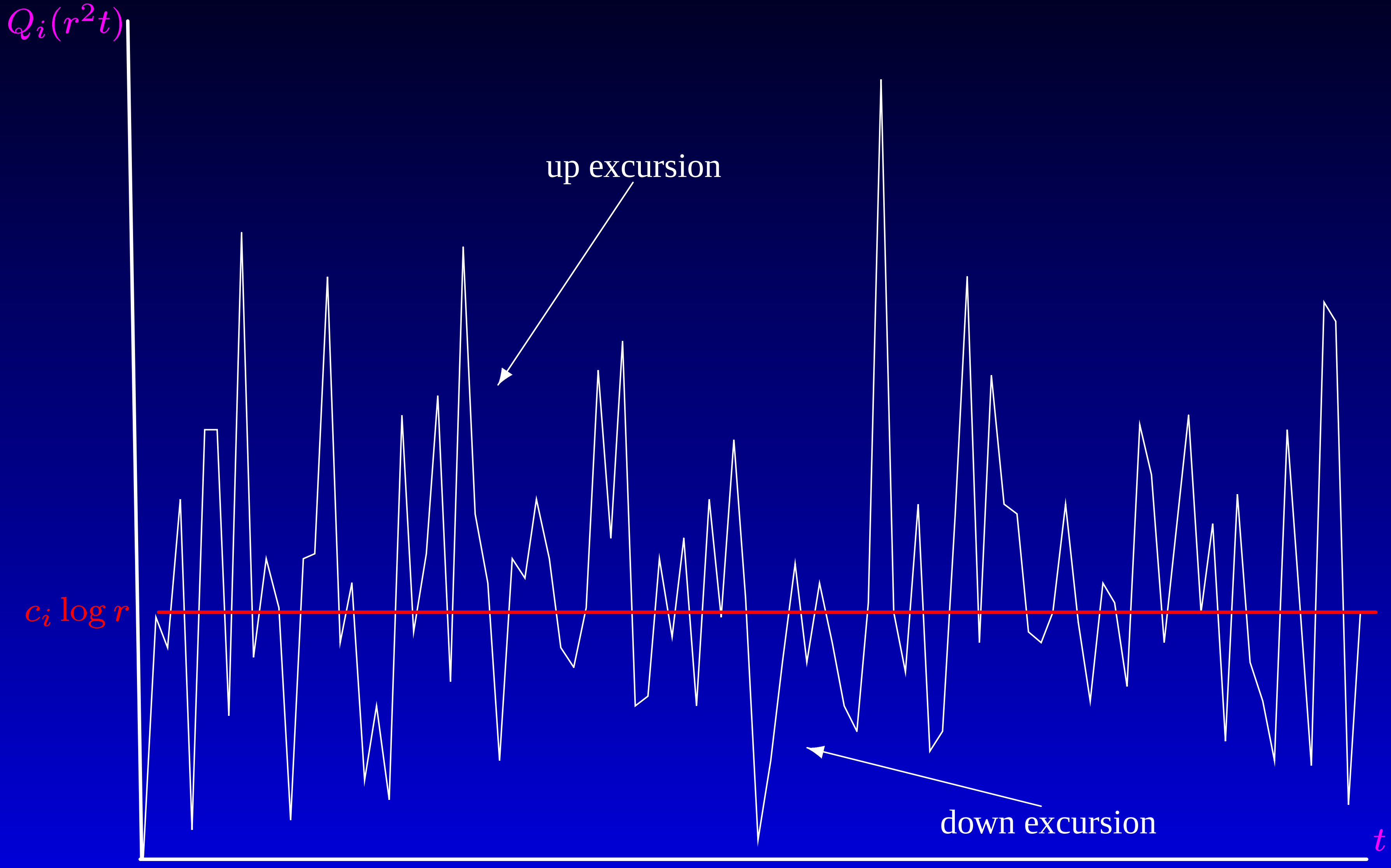
$$R_i^r(s) = \begin{cases} Q_i^r(s) - c_i \log r, & \text{if } i \text{ has a threshold} \\ Q_i^r(s) & \text{otherwise} \end{cases}$$

Lemma:

For $i \neq i^*$,

$$\mathbf{P}\Big(\sup_{s \in [0, r^2 t]} |R_i^r(s)| \geq c_i \log r \Big) \to 0 \text{ as } r \to \infty.$$

State Space Collapse



State Space Collapse

$$Q_i(s) = A_i(s) - \sum_{j=1}^{\mathbf{J}} \mathbf{C_{ij}}\, S_j(\, T_j(s)\,)$$

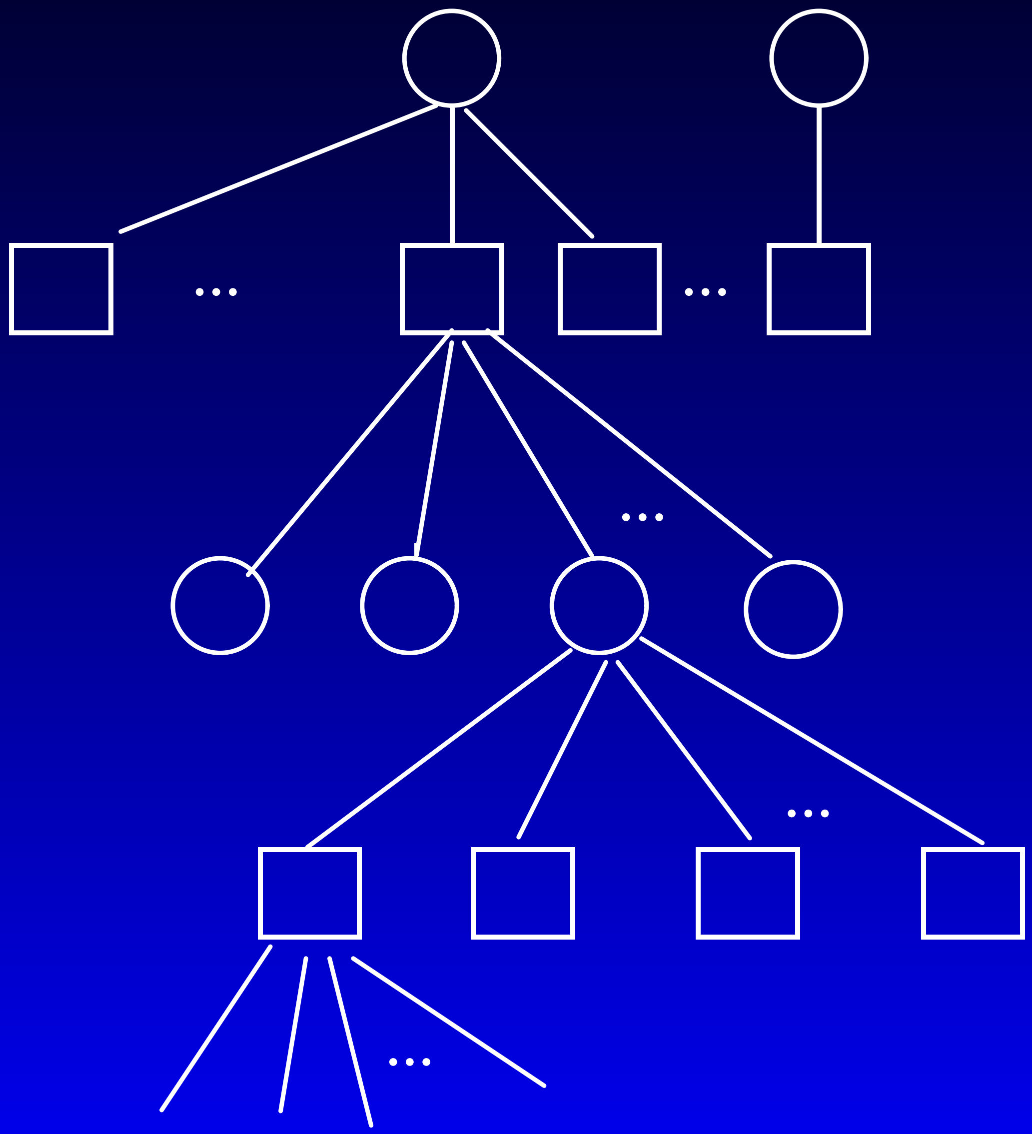
$$I_k(s) = s - \sum_{j=1}^{\mathbf{J}} \mathbf{A_{kj}}\, T_j(s)$$

Related Works

- Ata and Kumar, 2002
- Mandelbaum and Stolyar, 2002
- Stolyar, 2002

FIN

Partial Tree



Testing colors

GreenYellow Yellow Goldenrod Dandelion

Apricot Peach Melon YellowOrange

Orange BurntOrange Bittersweet RedOrange

Mahogany Maroon BrickRed Red

OrangeRed RubineRed WildStrawberry Salmon

CarnationPink Magenta VioletRed Rhodamine

Mulberry RedViolet Fuchsia Lavender

Thistle Orchid DarkOrchid Purple

Plum Violet RoyalPurple BlueViolet

Periwinkle CadetBlue CornflowerBlue MidnightBlue NavyBlue

RoyalBlue Blue Cerulean Cyan

ProcessBlue SkyBlue Turquoise TealBlue

Aquamarine BlueGreen Emerald JungleGreen SeaGreen

Green ForestGreen LimeGreen YellowGreen SpringGreen

OliveGreen RawSienna Sepia

Brown Tan Gray Black White