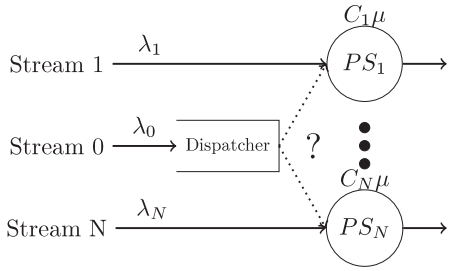
Poster: Optimal dispatching policies for parallel processor sharing nodes with partial information

Joost Bosman[1,2], Rob van der Mei[1,2] and Gerard Hoekstra[1,3]

[1]CWI, Probability and Stochastic Networks, Amsterdam, The Netherlands
[2]VU University Amsterdam, Department of Mathematics, The Netherlands
[3]Innovation Research & Technology, Thales Nederland B.V., Huizen, The Netherlands

Many of today's wireless networks have already closely approached the Shannon limit on channel capacity. A powerful alternative to increase the overall data rate then becomes one in which multiple, likely different, networks are used concurrently. The concurrent use of multiple networks simultaneously has opened up enormous possibilities for increasing bandwidth, improving reliability, and enhancing Quality of Service (QoS) in areas that are covered by multiple wireless access networks.

We study a model consisting of N non-identical parallel networks that are modeled as processor sharing (PS) nodes that serve $N+1$ streams of flows. Stream 0 is called the *foreground* stream, and streams $1,\ldots,N$ are called the *background* streams. From each stream flows arrive according to a Poisson process with arrival rate λ_i, $i=0,1,\ldots,N$. Flows from background stream i are served exclusively at PS node i. Each flow from the foreground stream has to be dispatched to one of the PS nodes on the basis of information on the *total* number of flows at each of the nodes, such that the expected sojourn time $E[S_0]$ for an arbitrary foreground flow is minimized. Flow sizes are exponentially distributed with rate μ, and each node has processing speed C_i, so that server i can handle $C_i\mu$ flows per time unit. For each stream, the offered load is given by $\rho_i := \lambda_i/\mu$ $(i=0,1,\ldots,N)$. The total offered load is given by $\rho := \rho_0+\rho_1+\cdots+\rho_N$. Finally, we define $\hat{\rho} := \rho/(C_1+\cdots+C_N)$.



We develop a heuristic approach for near-optimal dispatching in the case of partial information, i.e. the dispatcher only knows the *total* numbers of (foreground plus background) jobs at each node. The approach is based on the combination of two policies that perform well on complementary sets of parameter combinations: (1) the Weighted Join the Shortest Queue (WJSQ) policy, and (2) the Conditional Sojourn Time (CST) policy. WJSQ routes a foreground flow to the node where the total number of (foreground and background) flows n_i, normalized by the node speed, is minimal:

$$\underset{i\in\mathcal{I}}{\operatorname{argmin}}\left\{\frac{n_i}{C_i}\right\}. \tag{1}$$

CST routes an incoming flow to the node for which the expected sojourn time, conditioned on the fact that there are n_i other flows at node i at that moment, is minimal:

$$\underset{i\in\mathcal{I}}{\operatorname{argmin}}\left\{\frac{n_i+2}{2\mu C_i-\lambda_i}\right\}. \tag{2}$$

Both policies generate a switching curve given the total number of flows n_i on each node. To develop a method that works well for the whole range of values of ρ_0/ρ between 0 and 1, we propose the so-called convex combination (CC) method, defined as follows:

$$\underset{i\in\mathcal{I}}{\operatorname{argmin}}\left\{(1-\alpha)\frac{n_i}{C_i}+\alpha\frac{n_i+2}{2\mu C_i-\lambda_i}\right\}, \tag{3}$$

where α $(0\le\alpha\le 1)$ is given by $\alpha := \rho_0/\sum_{i=1}^{N}(C_i-\rho_i)$.

In our numerical experiments we fix $N=2$ and vary ρ_0, while keeping the total offered load ρ at a fixed level, such that the ratios ρ_i/C_i are kept fixed for $i=1,\ldots,N$. Furthermore, $C_1=1$, $C_2=4$, and $\mu=1$. Figure 1 shows the expected sojourn time $\mathbb{E}[S_0]$ as a function of ρ_0/ρ for each of the three dispatching heuristics, using the optimal solution, based on a MDP using full state information, as benchmark.

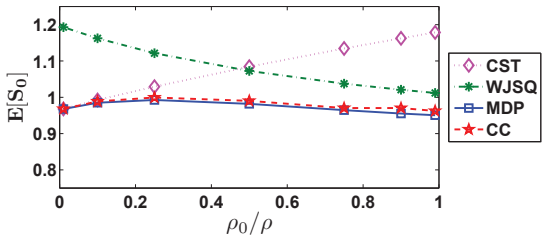


Figure 1: Comparison of policies for $\hat{\rho}=0.75$

The results illustrate that the CC-heuristic performs extremely well over the whole range of ρ_0/ρ-values, and strongly outperforms WJSQ and CST. Most remarkably, despite the fact that CC is based on the partial information, its performance is extremely close to the full-information MDP.

Acknowledgment: This work has been carried out in the context of the IOP GenCom project Service Optimization and Quality (SeQual), which is supported by the Dutch Ministry of Economic Affairs, Agriculture and Innovation via its agency Agentschap NL.