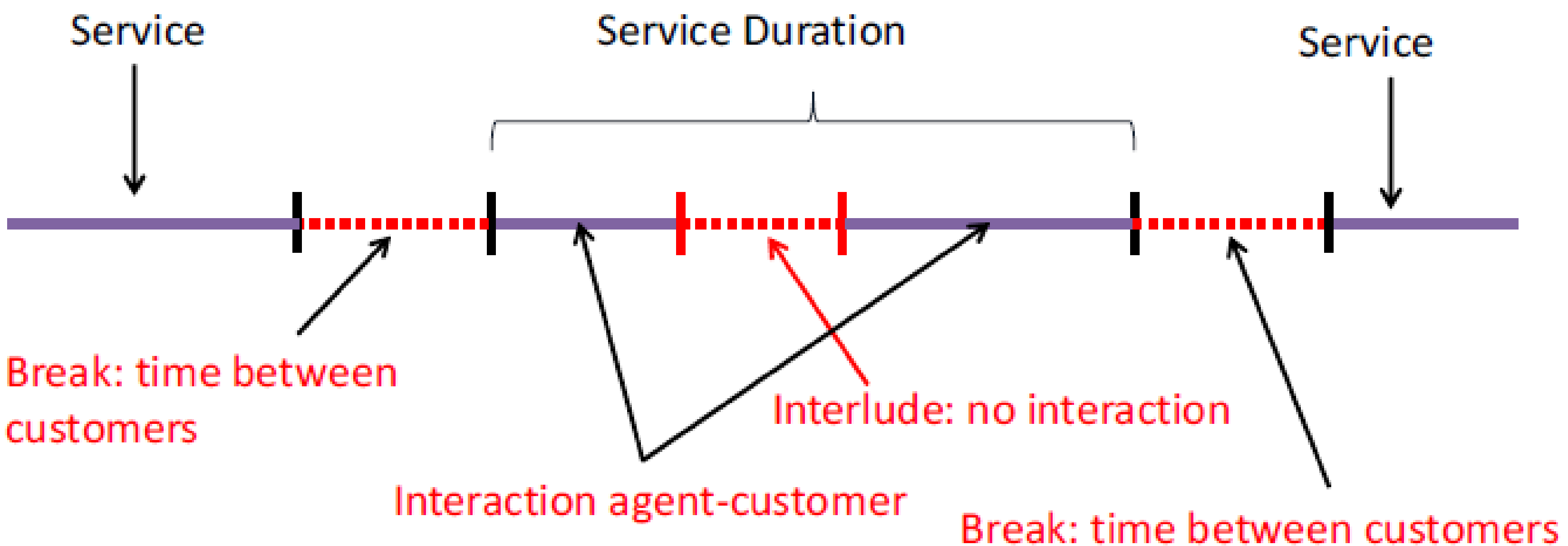
Front-office multi tasking between service encounters and back-office tasks

Zeynep Aksin, Koc University
Benjamin Legros, Paris School of Business
Oualid Jouini, Centrale Supelec
Ger Koole, VU University Amsterdam

Service encounters consist of multiple customer-agent interactions and interludes



HP: high priority

Ready

LP: low priority

Switching time

Switching time

The problem: how to use break and interlude times

- Maximize the expected proportion of time spent on LP
- While satisfying total expected waiting time constraint for HP
- Assuming switching times for HP-LP-HP transitions

$$\begin{cases} \text{Maximize } E(T) \\ \text{subject to } E(W) \leq \overline{w}. \end{cases}$$

Case manager systems



Campello, Ingolffson, Shumsky 2016
KC 2013
Dobson, Tezcan, Tilson 2013
Cui and Tezcan 2016

Our model assumes

- caseload=1
- switching times between tasks
- blend of front and back office tasks

Blending systems



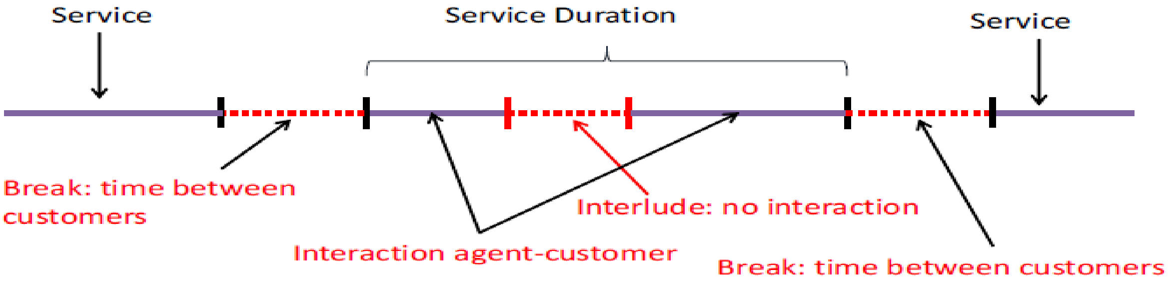
Bhulai and Koole, 2003
Gans and Zhou, 2003
Armony and Maglaras, 2004
Legros, Jouini, Koole 2016

Our model assumes

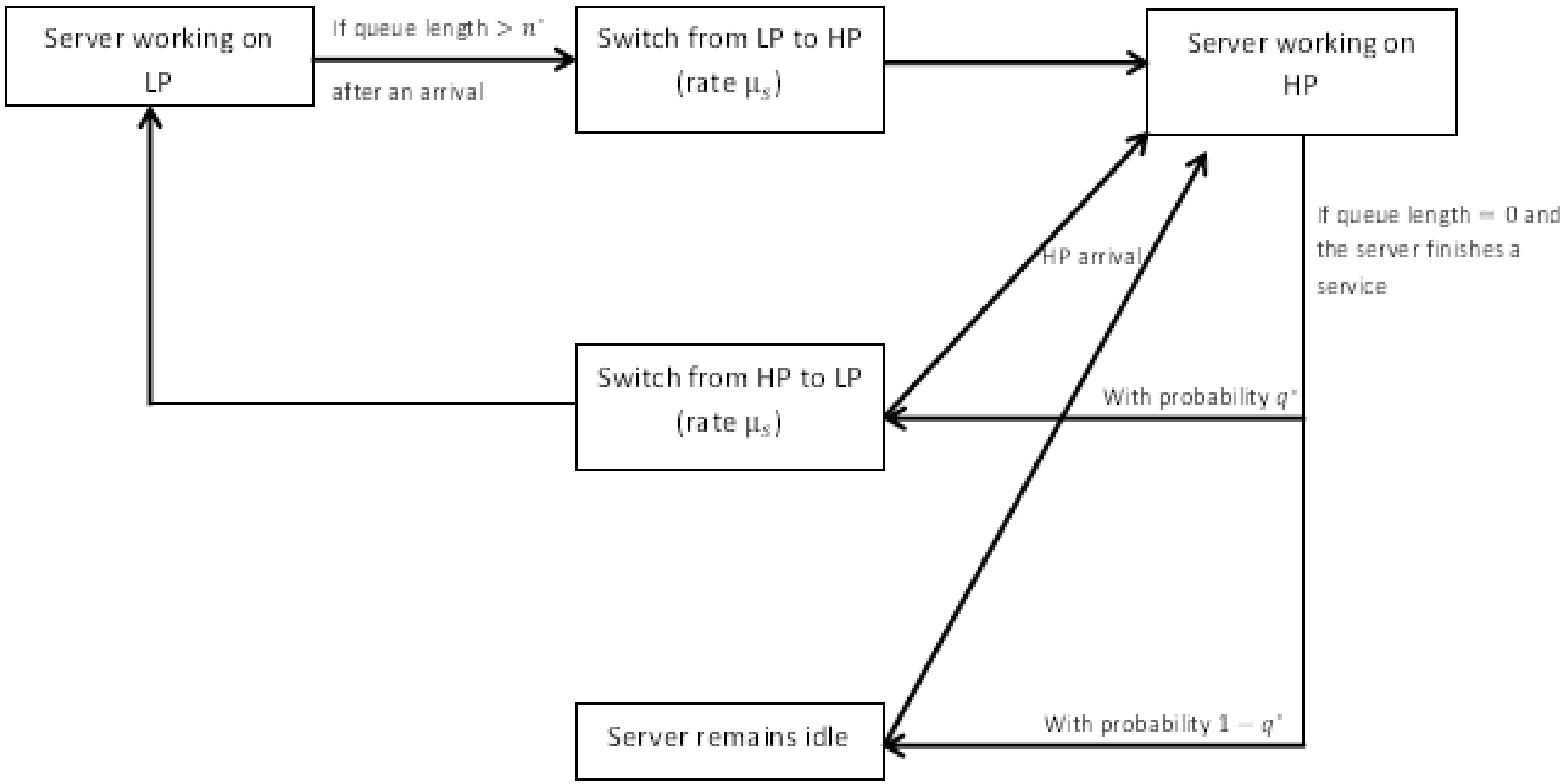
- switching times between tasks
- blending of multi-stage service encounters and back office tasks ⇨ imbrication and blending

The model

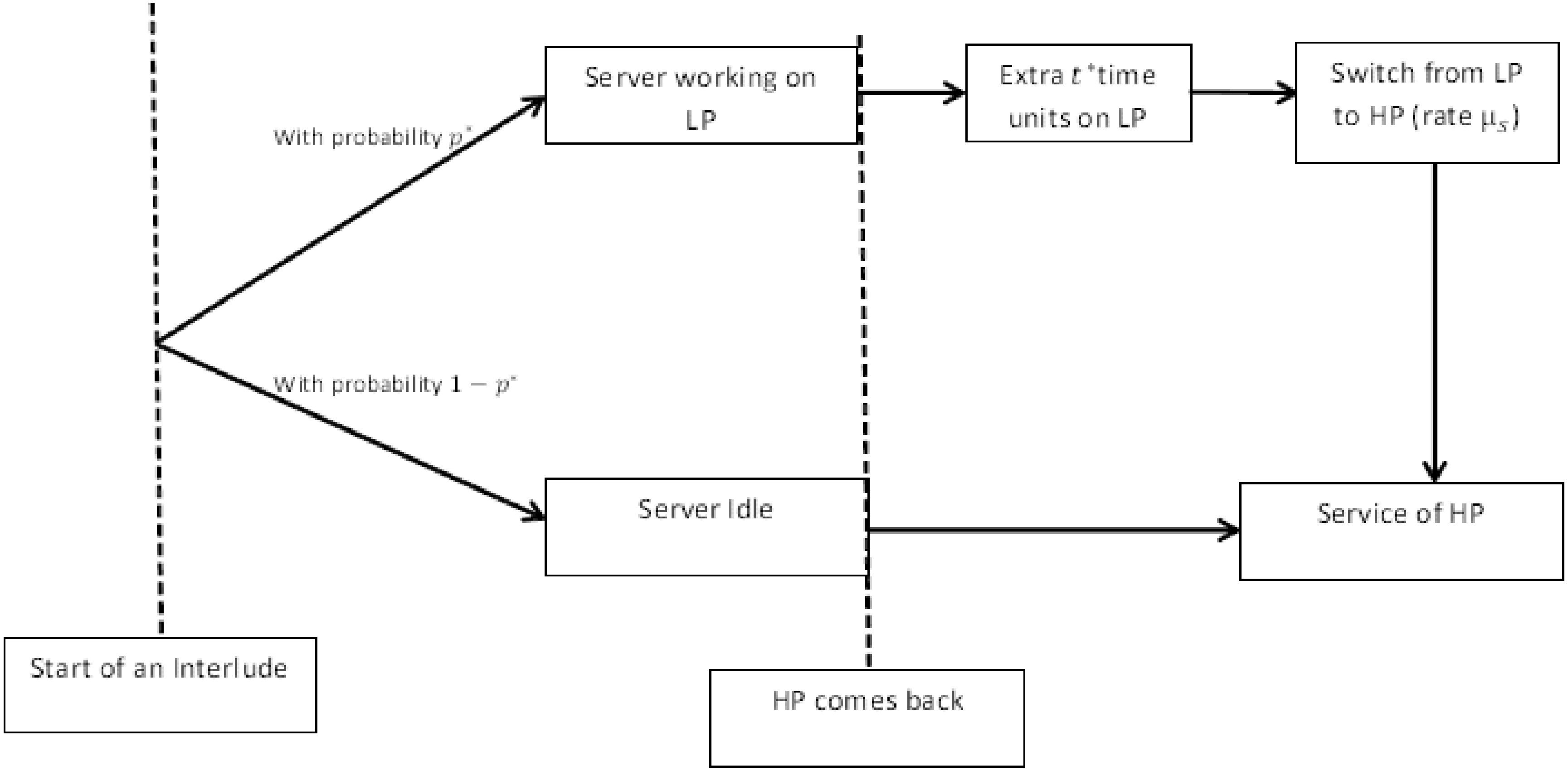
- Single server queue
- Poisson arrivals of HP jobs
- Infinite amount of LP tasks
- Service time of HP: succession of general service time interaction periods with exponential interludes
- Exponential switching times



LP treatments between HP



LP treatment during an interlude



Overview of Analysis

- Given n, q, p, t performance evaluation
- Recursive formula for steady state probabilities using a residual service time state definition
- Given p,t optimize n and q

  Proposition 1: characterizes n* and q* as a function of parameters (switching times, slack in SL constraint)
- Optimize p and t numerically

Performance measures of interest

$$\phi_0 = \frac{1 - \lambda\overline{x}}{n^* + a_s\left(1 + \frac{1}{q^*}\right) + \frac{1}{q^*}}.$$

Expected proportion of time spent on LP during interlude

$$E(T) = E(T_e) + \frac{(1 - \lambda\overline{x})(n^* + 1)}{n^* + a_s\left(1 + \frac{1}{q^*}\right) + \frac{1}{q^*}}.$$

$$E(W) = E(W_e) + \frac{\lambda\overline{x}^2(1 + cv^2)}{2(1 - \lambda\overline{x})} + \frac{n^*\left(n^* + 1 + 2a_s\right)}{2\lambda\left(n^* + a_s\left(1 + \frac{1}{q^*}\right) + \frac{1}{q^*}\right)} + \frac{1}{\mu_s} \cdot \frac{1 + a_s}{n^* + a_s\left(1 + \frac{1}{q^*}\right) + \frac{1}{q^*}}.$$

Wait during service

M/G/1 wait

When q=1
n/2λ

Only switching
dependent term

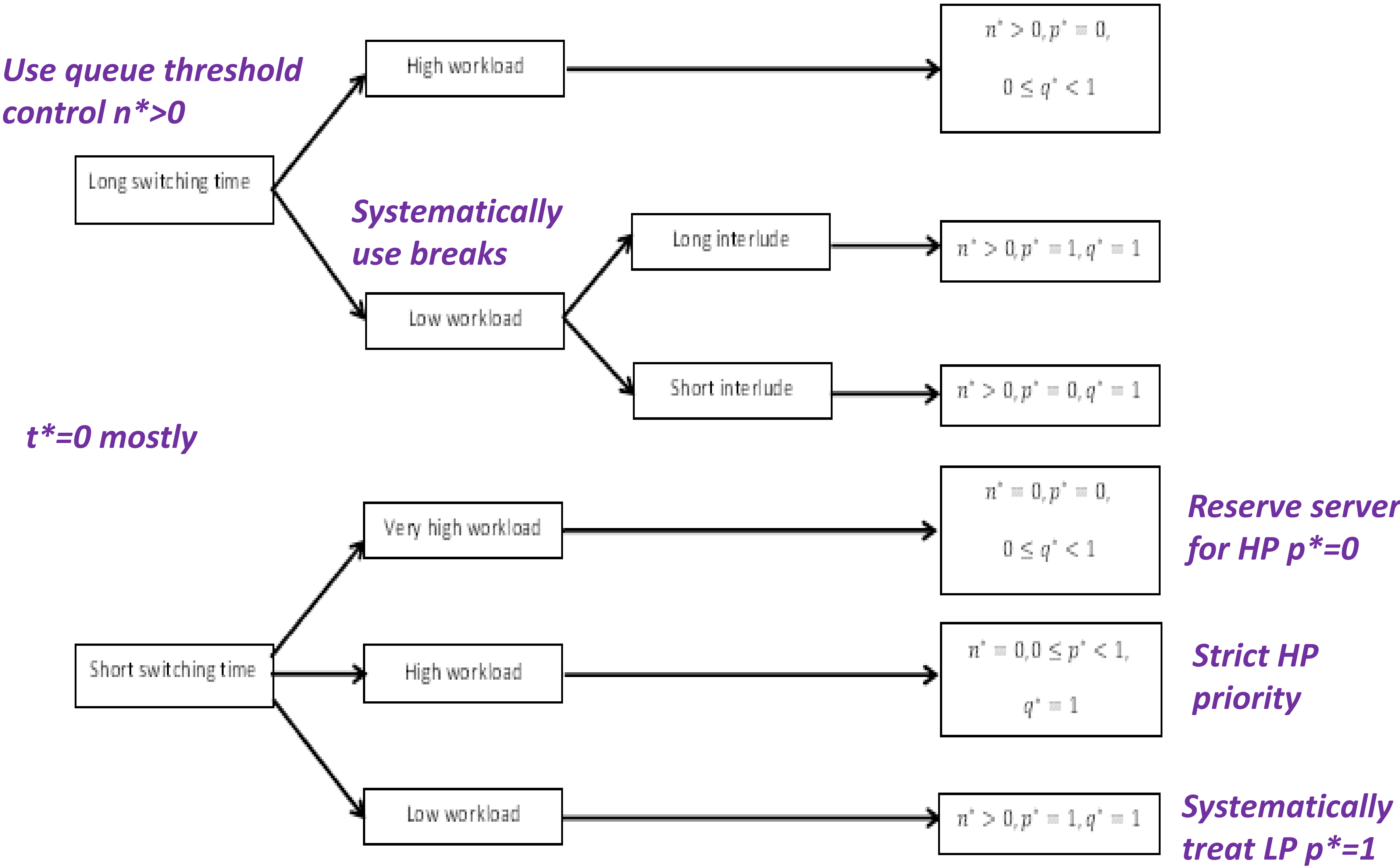
Optimal n* and q* given p and t

Slack in SL constraint

Proposition 1 *The following holds.*

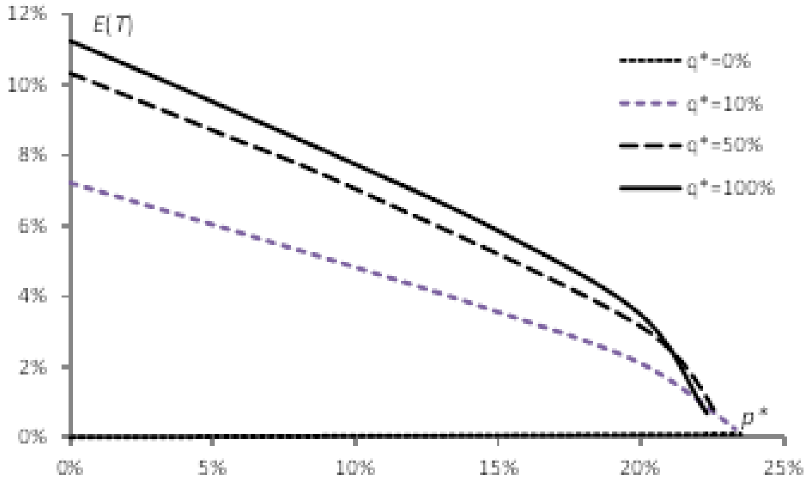
1. *If* $a_s \geq \frac{1}{\sqrt{2}}$ *and* $K \leq a_s - \frac{\sqrt{2}}{2\sqrt{2}+4}$, *then the optimal couple* (n^*, q^*) *is* $n^* = \sqrt{2}a_s - 1$ *and* $q^* = \frac{2(a_s+1)K}{2\sqrt{2}a_s^2+4a_s^2-\sqrt{2}a_s-K(2\sqrt{2}a_s+2a_s-2)}$.
2. *If* $a_s < \frac{1}{\sqrt{2}}$ *and* $K \leq \frac{a_s(1+a_s)}{1+2a_s}$, *then the optimal couple* (n^*, q^*) *is* $n^* = 0$ *and* $q^* = \frac{(a_s+1)\lambda K}{a_s(a_s+1)-\lambda K a_s}$.
3. *In the remaining cases, the optimal couple* (n^*, q^*) *is* $n^* = \lambda K - a_s - \frac{1}{2} + \frac{1}{2}\sqrt{(1+2a_s-2\lambda K)^2 - 8(a_s(1+a_s) - \lambda K(1+2a_s))}$ *and* $q^* = 1$.

Long switching times

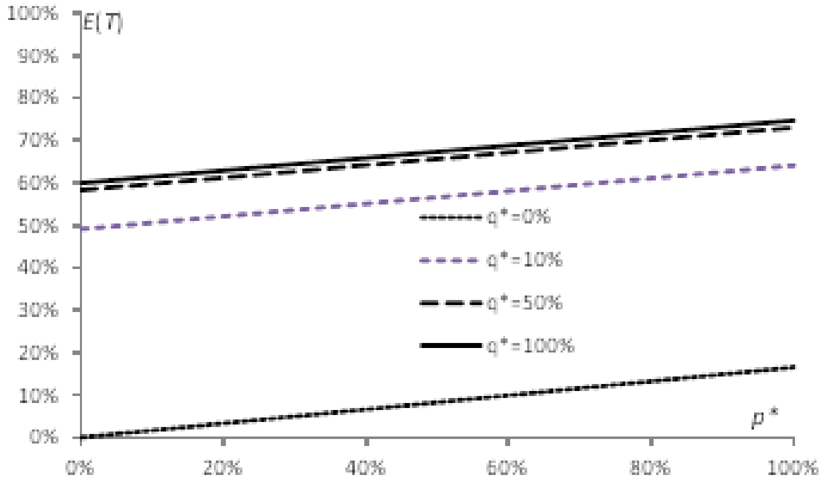

Use queue threshold control n*>0
Long switching time
High workload
n* > 0, p* = 0,
0 ≤ q* < 1
Systematically use breaks
Low workload
Long interlude
n* > 0, p* = 1, q* = 1
Short interlude
n* > 0, p* = 0, q* = 1
t*=0 mostly
Short switching time
Very high workload
n* = 0, p* = 0,
0 ≤ q* < 1
Reserve server for HP p*=0
High workload
n* = 0, 0 ≤ p* < 1,
q* = 1
Strict HP priority
Low workload
n* > 0, p* = 1, q* = 1
Systematically treat LP p*=1

Sensitivity analyses: p* and q*

High workload



Low workload



What to do with multiple interludes?

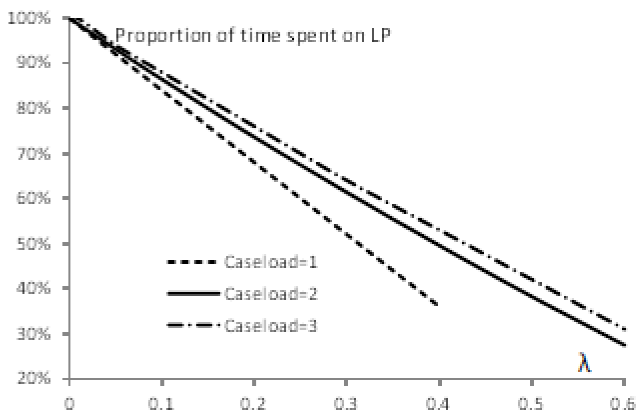
- N interludes, ordered from longest to shortest
- Denote by p_j the control parameter for interlude j=1,….,N
- Then the optimal policy will have the structure

$p_1^* = p_2^* =....= p_j^* =1,$

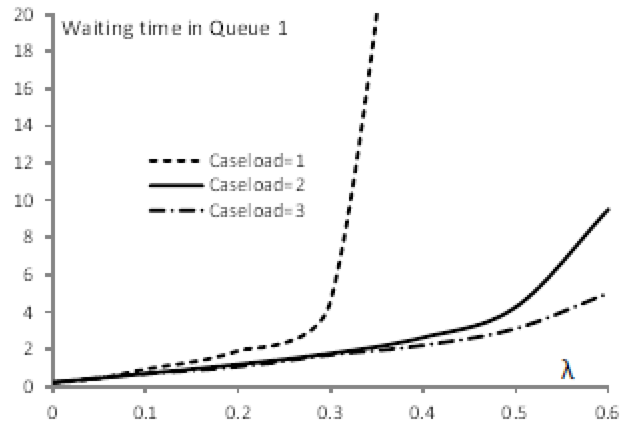
$0<= p_{j+1}^* < 1,$

$p_{j+2}^* = ...= p_N^* =0$

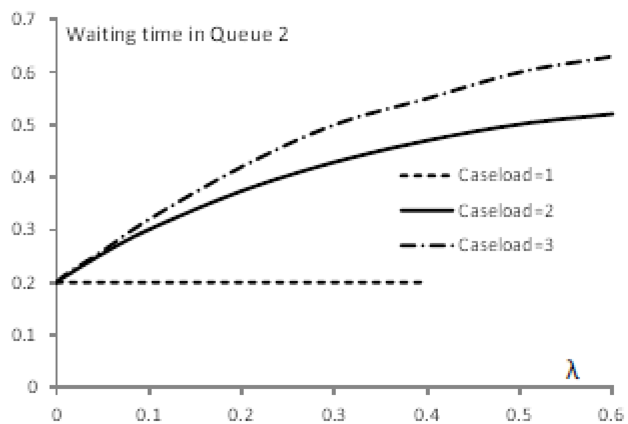
What about allowing for caseload > 1? (analysis for a simplified model)



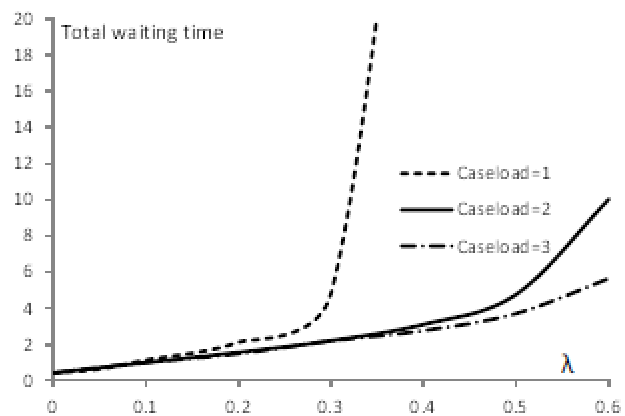
(a) Proportion of time spent on LP



(b) Waiting time in Queue 1



(c) Waiting time in Queue 2



(d) Total waiting time

Conclusions

- Using interlude times in multi-stage services for blending can be beneficial
- The decision of when to blend hinges on the switching times
- First blend during breaks, only then consider blending during interludes
- A control on the number in queue is more effective than a control on the interlude duration
- Concentrate blending effort on fewer and longer interludes