The optimal time to adapt the processing rate in a Make-to-order production system

SMMSO 2017

Jannik Vogel, Raik Stolletz

Lecce, June 2017

When to react to increasing demand? Optimization of service systems with discretionary task completion

SMMSO 2017

Jannik Vogel, Raik Stolletz

Lecce, June 2017

Introduction

- Many tasks are completed according to subjective completion criteria
  * Discretionary tasks (Hopp et al., 2007)
  * Customer-intensive services (Anand et al., 2011)
- Health care, personal care, legal or financial consultancy, software engineering, call centers
- Tradeoff **Quality vs. Speed**
  * Fast service $\Longrightarrow$ Low quality ☹, Low waiting times ☺
  * Slow service $\Longrightarrow$ High quality ☺, High waiting times ☹
- Time-dependent setting
  * How are decisions influenced by demand changes in the future?

Agenda

1. Literature overview
2. Problem description & model formulation
3. Stationary solutions
4. Time-dependent results
   - Deterministic fluid approach
   - Stochastic SBC approach
5. Numerical results
6. Summary

Literature overview

1. Service rate decisions **without impact on quality**
   1.1 Based on current work-in-process (George and Harrison, 2001; Stidham and Weber, 1989)
   1.2 Based on demand over finite horizon (Parlar, 1984; Alam, 1979)

2. Service rate decisions **with impact on quality**

Paper	Queue	Objective		Demand	Dynamics
		Quality	Cong.		
Hopp et al. (2007)	$M/D/1$	Exponential	L^S	Ex.	Stat.
Wang et al. (2010)	$M/G/c$	Error prob.	W^Q	End.	Stat.
Anand et al. (2011)	$M/M/1$	Linear	W^S	End.	Stat.
Kostami and Rajagopalan (2014)	$M/M(t)/1$	-	L^S	End.	T-d.
Our model	$D(t)/D(t)/c$ $M(t)/M(t)/c$	Exponential	W^S	Ex.	T-d.

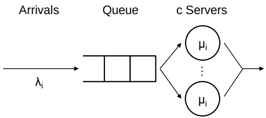
Cong. = Congestion measure; Ex. = Exogenous; End. = Endogenous; Stat. = Stationary; T-d. = Time-dependent

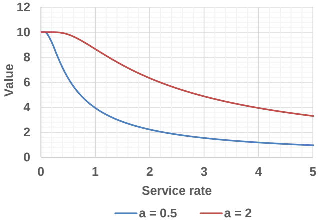
No publication considers time-dependent demand!

Agenda

Problem description

- $D(t)/D(t)/c$ and $M(t)/M(t)/c$ systems
- **Input:**
  - Arrival rates $\lambda_i \quad \forall i = 1, ..., n$
- **Decisions:**
  - Service rates $\mu_i \leq \overline{\mu} \quad \forall i = 1, ..., n$
- **Objective:**
  - Value: $v(\mu_i) = \bar{v}(1 - e^{-\frac{a}{\mu_i}})$ per served costumer (Hopp et al., 2007)
    * $\bar{v}$: Maximum value
    * a: Sensitivity to service rate
  - Waiting cost w per time unit spent in the system





$(\bar{v} = 10)$

Model formulation

1. **Time-dependent model:**

$$\max\ Z = \ell \sum_{i=1}^{n} \left[\bar{v}(1 - e^{-\frac{a}{\mu_i}}) \mathrm{E}[Th_i(\mu_1, ..., \mu_i)] - w\,\mathrm{E}[W_i^S(\mu_1, ..., \mu_i)] \right] \quad (1)$$

$$\text{s.t.} \qquad 0 < \mu_i \leq \overline{\mu} \quad \forall i = 1, ..., n \quad (2)$$

Remark:

- Numerical solution
- Analytical solution

Model formulation

1. **Time-dependent model:**

$$\max Z = \ell \sum_{i=1}^{n} \left[\bar{v}(1 - e^{-\frac{a}{\mu_i}}) \mathrm{E}[Th_i(\mu_1, ..., \mu_i)] - w\, \mathrm{E}[W_i^S(\mu_1, ..., \mu_i)] \right] \quad (1)$$

s.t.

$$0 < \mu_i \leq \overline{\mu} \quad \forall i = 1, ..., n \quad (2)$$

Remark:

- Numerical solution
- Analytical solution

2. **Stationary model:**

$$\max Z = \bar{v}(1 - e^{-\frac{a}{\mu}})\lambda - w\, \mathrm{E}[W^S(\mu)] \quad (3)$$

s.t.

$$0 < \mu \leq \overline{\mu} \quad (4)$$

Remark:

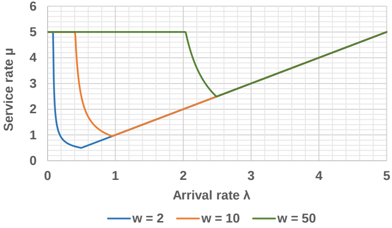
- A solution exists iff $\begin{cases} c\mu \geq \lambda & \text{deterministic model} \\ c\mu > \lambda & \text{stochastic model} \end{cases}$
- Analytical solution exists for $\begin{cases} c \geq 1 & \text{deterministic model} \\ c = 1 & \text{stochastic model} \end{cases}$

Agenda

Optimal solutions in the stationary system

***D/D/c*:**

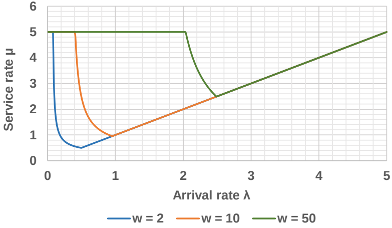
$$\mu^D = \begin{cases} \min\left(\max\left(\frac{a}{\ln\frac{a\bar{v}\lambda}{w}}, \frac{\lambda}{c}\right), \overline{\mu}\right) & \text{for } a\bar{v}\lambda > w \\ \overline{\mu} & \text{otherwise.} \end{cases}$$



Optimal solutions in the stationary system

D/D/c:

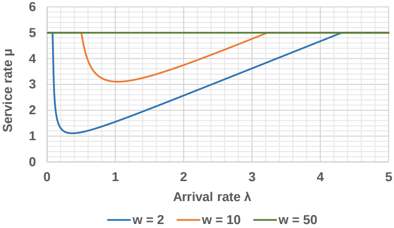
$$\mu^D = \begin{cases} \min\left(\max\left(\frac{a}{\ln \frac{a\bar{v}\lambda}{w}}, \frac{\lambda}{c}\right), \overline{\mu}\right) & \text{for } a\bar{v}\lambda > w \\ \overline{\mu} & \text{otherwise.} \end{cases}$$



M/M/1:

$$\mu^S = \begin{cases} \min\left(\mu^*, \overline{\mu}\right) & \text{for } a\bar{v}\lambda > w \\ \overline{\mu} & \text{otherwise} \end{cases}$$

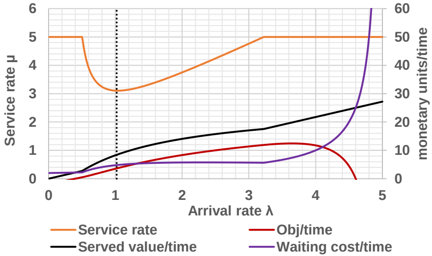
with μ^* determined by $\left(\frac{\mu^* - \lambda}{\mu^*}\right)^2 e^{-\frac{a}{\mu^*}} = \frac{w}{a\bar{v}\lambda}$.

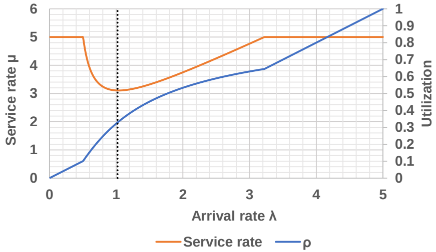


(Parameters in the figures: $c = 1, \bar{v} = 30, a = 1, \overline{\mu} = 5$)

Optimal solutions in the stationary system (II)

Setting: $M/M/1$, $c = 1, \bar{v} = 30, a = 1, w = 10, \overline{\mu} = 5$





Findings:

- $\mu^S \geq \mu^D$
- High sensitivity in λ
- μ^S and μ^D no monotone function in λ

Agenda

Optimal solutions in the deterministic model

Fluid assumptions (Newell, 1971)

1. **Numerical solution**
   Determined by solving the MINLP

2. **Analytical solution** for increasing demand λ_i

$$\mu_i^{TD} = \begin{cases} \min\left(\max\left(\frac{a}{\ln \frac{a\bar{v}\lambda_i}{w}}, \frac{\lambda_i}{c}\right), \overline{\mu}\right) & \text{for } a\bar{v}\lambda_i > w \\ \overline{\mu} & \text{otherwise.} \end{cases} \quad (5)$$

Optimal solutions in the deterministic model

Fluid assumptions (Newell, 1971)

1. **Numerical solution**
   Determined by solving the MINLP

2. **Analytical solution** for increasing demand λ_i

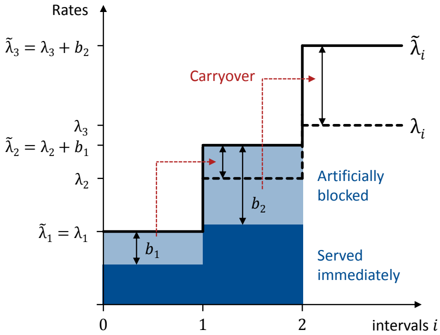
$$\mu_i^{TD} = \begin{cases} \min\left(\max\left(\frac{a}{\ln \frac{a\bar{v}\lambda_i}{w}}, \frac{\lambda_i}{c}\right), \overline{\mu}\right) & \text{for } a\bar{v}\lambda_i > w \\ \overline{\mu} & \text{otherwise.} \end{cases} \quad (5)$$

Finding: Optimal solution only depends on the current period!

Stochastic SBC approach (Stolletz, 2008)

For every period i with given length ℓ:

1. **Stationary loss system** $(M/M/c/c)$
   * Input: Artificial arrival rate $\tilde{\lambda}_i$
   * Output: Utilization $\mathrm{E}[U_i]$ and backlog rate b_i
2. **Stationary waiting system** $(M/M/c)$
   * Input: Modified arrival rate λ_i^{MAR} which results in $\mathrm{E}[U_i]$
   * Output: System performance



Good approximations for $\ell = \mu^{-1}$ $\Longrightarrow$ Calibration of the period length

NLP for the SBC approach

$$\max Z = \ell \sum_{i=1}^{n} \left[\bar{v}(1 - e^{-\frac{a}{\mu_i}}) \mathrm{E}[Th_i(\mu_1, ..., \mu_i)] - w\, \mathrm{E}[W_i^S(\mu_1, ..., \mu_i)] \right] \tag{6}$$

$$0 < \mu_i \leq \overline{\mu} \qquad \forall i = 1, ..., n \tag{7}$$

1. Step

$$b_0 = 0 \tag{8}$$

$$b_i = \tilde{\lambda}_i P_i^B \qquad \forall i = 1, ..., n \tag{9}$$

$$\tilde{\lambda}_i = \lambda_i + b_{i-1} \qquad \forall i = 1, ..., n \tag{10}$$

$$P_i^B = \frac{(\tilde{\lambda}_i/\mu_i)^c}{c! \sum_{k=0}^{c} \frac{(\tilde{\lambda}_i/\mu_i)^k}{k!}} \qquad \forall i = 1, ..., n \tag{11}$$

$$\mathrm{E}[U_i] = \frac{\tilde{\lambda}_i(1 - P_i^B)}{c\mu_i} = \frac{\lambda_i + b_{i-1} - b_i}{c\mu_i} \qquad \forall i = 1, ..., n \tag{12}$$

2. Step

$$\lambda_i^{MAR} = \mathrm{E}[U_i] c\mu_i = \lambda_i + b_{i-1} - b_i \qquad \forall i = 1, ..., n \tag{13}$$

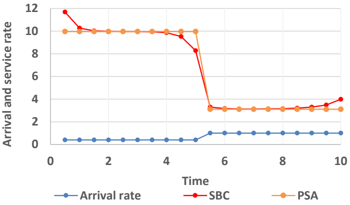
$$P_i^0 = \left(\sum_{n=0}^{c-1} \frac{(\lambda_i^{MAR}/\mu_i)^n}{n!} + \frac{(\lambda_i^{MAR}/\mu_i)^c}{c! \cdot (1 - \frac{\lambda_i}{c\mu_i})} \right)^{-1} \qquad \forall i = 1, ..., n \tag{14}$$

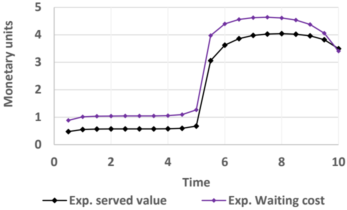
$$\mathrm{E}[W_i^S] = \frac{(\lambda_i^{MAR}/\mu_i)^c}{(c-1)!\mu_i(c - \lambda_i^{MAR}/\mu_i)^2} P_i^0 + \frac{1}{\mu_i} \qquad \forall i = 1, ..., n \tag{15}$$

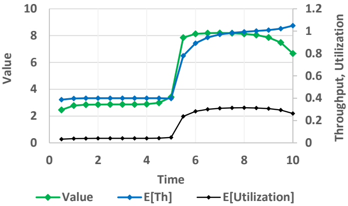
Agenda

Anticipation of demand changes: λ_i low

Setting: $c = 1$, $n = 20$, $\ell = 0.5$, $w = 10$, $\bar{v} = 30$, $a = 1$,
$\lambda(t) = 0.4$ for $t \leq 5, 1$ otherwise.

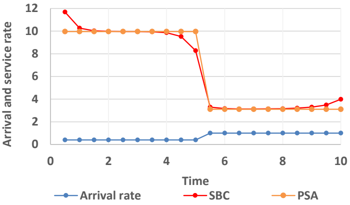


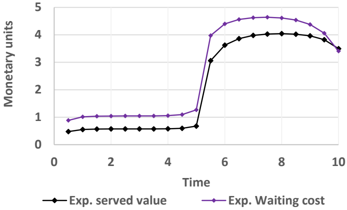


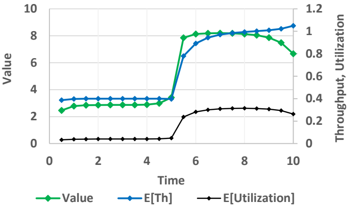


Anticipation of demand changes: λ_i low

Setting: $c = 1$, $n = 20$, $\ell = 0.5$, $w = 10$, $\bar{v} = 30$, $a = 1$, $\lambda(t) = 0.4$ for $t \leq 5, 1$ otherwise.



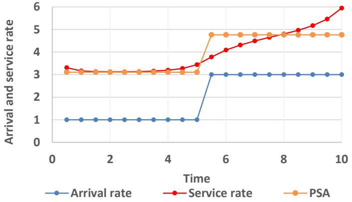


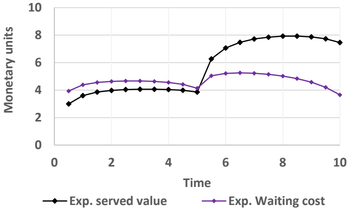


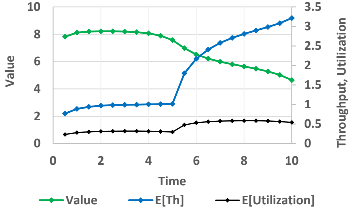
Finding: Service rate changes before the demand changes!

Anticipation of demand changes: λ_i high

Setting: $c = 1$, $n = 20$, $\ell = 0.5$, $w = 10$, $\bar{v} = 30$, $a = 1$,
$\lambda(t) = 1$ for $t \leq 5, 3$ otherwise.

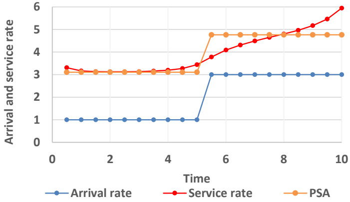


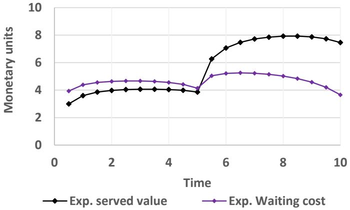


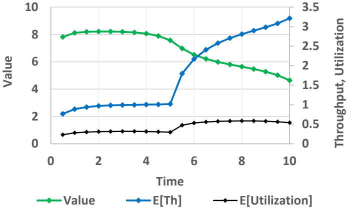


Anticipation of demand changes: λ_i high

Setting: $c = 1,\ n = 20,\ \ell = 0.5,\ w = 10,\ \bar{v} = 30, a = 1,$
$\lambda(t) = 1$ for $t \leq 5, 3$ otherwise.







Finding: Demand change leads to several service rate changes.

Summary

Conclusion:

- **Model:**
  - Quality-speed tradeoff
  - New model: Service rate optimization with time-dependent demand
- **Method:** Iterative procedure for the SBC-approach
- **Managerial insights**
  * Optimal service rate not monotone in λ
  * Deterministic model: Increasing demand does not influence service rates beforehand
  * Stochastic model: Later demand influences decisions

Summary

Conclusion:

- **Model:**
  - Quality-speed tradeoff
  - New model: Service rate optimization with time-dependent demand
- **Method:** Iterative procedure for the SBC-approach
- **Managerial insights**
  * Optimal service rate not monotone in λ
  * Deterministic model: Increasing demand does not influence service rates beforehand
  * Stochastic model: Later demand influences decisions

Future research:

- Include state-dependent information

References I

Alam, M. (1979). An application of modern control theory to a time-dependent queuing system for optimal operation. *International Journal of Systems Science 10*(August 2013), 693–700.

Anand, K. S., M. F. Paç, and S. Veeraraghavan (2011). Quality–Speed Conundrum: Trade-offs in Customer-Intensive Services. *Management Science 57*(1), 40–56.

George, J. M. and J. M. Harrison (2001). Dynamic Control of a Queue with Adjustable Service Rate. *Operations Research 49*(February 2015), 720–731.

Hopp, W. J., S. M. R. Iravani, and G. Y. Yuen (2007). Operations Systems with Discretionary Task Completion. *Management Science 53*(1), 61–77.

Kostami, V. and S. Rajagopalan (2014). Speed–Quality Trade-Offs in a Dynamic Model. *Manufacturing & Service Operations Management 16*(1), 104–118.

Newell, G. F. (1971). *Applications of Queueing Theory*. London: Chapman and Hall.

Parlar, M. (1984). Optimal dynamic service rate control in time dependent M/M/S/N queues. *International Journal of Systems Science 15*(August 2013), 107–118.

Stidham, S. and R. R. Weber (1989). Monotonic and Insensitive Optimal Policies for Control of Queues with Undiscounted Costs. *Operations Research 37*(4), 611–625.

Stolletz, R. (2008). Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research 190*, 478–493.

Wang, X., L. G. Debo, A. Scheller-Wolf, and S. F. Smith (2010). Design and Analysis of Diagnostic Service Centers. *Management Science 56*(11), 1873–1890.

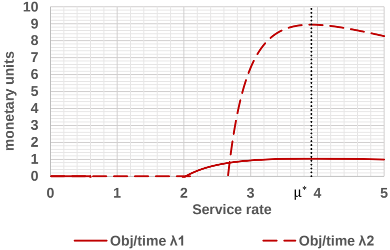
Objective value and performance measures depending on μ

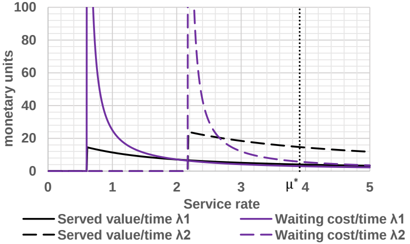
Setting: $c = 1, \bar{v} = 30, a = 1, w = 10, \overline{\mu} = 5, \lambda_1 = 0.6, \lambda_2 = 2.17, \mu^* = 3.91$

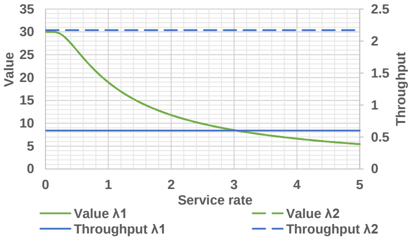
Condition for optimal μ:

$$ServedValue(\mu)' - WaitingCost(\mu)' = 0 \qquad (16)$$

$$\frac{a}{\mu^{*2}} \bar{v} \lambda e^{-a/\mu^*} = \frac{w}{(\lambda - \mu^*)^2} \qquad (17)$$







Optimal solutions for a $D/D/c$-system: Proof

$$\max Z = \bar{v}(1 - e^{-\frac{a}{\mu}})\lambda - \frac{w}{\mu} \tag{18}$$

Setting the first order partial derivative $\frac{\partial Z}{\partial \mu} = \frac{w}{\mu^2}(1 - \frac{a\bar{v}\lambda}{w}e^{-\frac{a}{\mu}})$ to zero gives $1 - \frac{a\bar{v}\lambda}{w}e^{-\frac{a}{\mu}} = 0$. Let us first assume that $a\bar{v}\lambda > w$. The zero of the function is found at $\mu' = \frac{a}{\ln \frac{a\bar{v}\lambda}{w}}$. Notice that for $\mu < \mu'$, $\frac{\partial Z}{\partial \mu}(\mu) > 0$ and for $\mu > \mu'$, $\frac{\partial Z}{\partial \mu}(\mu) < 0$. Thus, there is a maximum at μ'. Furthermore, if $\mu' < \frac{\lambda}{c}$, the objective function (18) is maximized at $\frac{\lambda}{c}$. If $\mu' > \overline{\mu}$, (18) is maximized at $\overline{\mu}$.

Let us now consider the special case $a\bar{v}\lambda \leq w$. For those parameters the zero of the first order partial derivative does not lie in A. Notice that $\frac{\partial Z}{\partial \mu}(\mu) > 0$ for $\mu \in A$. Thus, the maximum is attained for $\mu = \overline{\mu}$.

Optimal solutions for an $M/M/1$-system: Proof

$$\max Z = \bar{v}(1 - e^{-\frac{a}{\mu}})\lambda - \frac{w}{\mu - \lambda} \tag{19}$$

Setting the first order partial derivative equal zero gives $\left(\frac{\mu-\lambda}{\mu}\right)^2 e^{-\frac{a}{\mu}} = \frac{w}{a\bar{v}\lambda}$. Notice that the left hand side of the equation, $f : (\lambda, \infty) \to (0,1), \mu \mapsto \left(\frac{\mu-\lambda}{\mu}\right)^2 e^{-\frac{a}{\mu}}$ is continuous and strictly monotonically increasing. Therefore, for all $a\bar{v}\lambda > w$, a solution can be found. However, this solution does not need to lie in A and then the optimal solution is $\overline{\mu}$, because the first order partial derivative in μ is positive on the set $(\lambda, \overline{\mu})$.

Let us now consider the special case $a\bar{v}\lambda \leq w$. For those parameters the zero of the first order partial derivative does not lie in A. Notice that $\frac{\partial Z}{\partial \mu}(\mu) > 0$ for $\mu > \lambda$. Thus, the maximum is attained for $\mu = \overline{\mu}$.

Fluid approximation (Newell, 1971)

- **Key idea:** Replace discrete stochastic arrivals by a deterministic continuum
- Queue length at the end of period i:

$$\mathsf{E}[L_i^{Q,end}] = \max\{\mathsf{E}[L_{i-1}^{Q,end}] + \ell(\lambda_i - c\mu_i), 0\} \tag{20}$$

- Percentage of period i that has a positive queue length:

$$\pi_i = \begin{cases} 0 & \mathsf{E}[L_{i-1}^{Q,end}] = 0, \mathsf{E}[L_i^{Q,end}] = 0 \\ \frac{\mathsf{E}[L_{i-1}^{Q,end}]}{\ell(c\mu_i - \lambda_i)} & \mathsf{E}[L_{i-1}^{Q,end}] = 1, \mathsf{E}[L_i^{Q,end}] = 0 \\ 1 & \mathsf{E}[L_i^{Q,end}] > 0 \end{cases} \quad \forall i = 1, ..., n \tag{21}$$

- Expected average queue length in period i: $\mathsf{E}[L_i^Q] = \pi_i \frac{\mathsf{E}[L_{i-1}^{Q,end}] + \mathsf{E}[L_i^{Q,end}]}{2}$
- Expected average cycle time:

$$\mathsf{E}[W_i^S] = \begin{cases} \frac{1}{\mu_i} & \mathsf{E}[L_{i-1}^{Q,end}] = 0, \mathsf{E}[L_i^{Q.end}] = 0 \\ \mathsf{E}[L_i^Q]/\lambda_i + \frac{1}{\mu_i} & \mathsf{E}[L_{i-1}^{Q,end}] = 1, \mathsf{E}[L_i^{Q.end}] = 0 \\ \mathsf{E}[L_i^Q]/\mu_i + \frac{1}{\mu_i} & \mathsf{E}[L_i^{Q,end}] = 1 \end{cases} \tag{22}$$

MINLP for the fluid approach

$$\begin{aligned}
\mathrm{E}[L_0^{Q,end}] &= 0 && && (23)\\
\mathrm{E}[L_i^{Q,end}] &= \mathrm{E}[L_{i-1}^{Q,end}] + \ell(\lambda_i - \mathrm{E}[Th_i]) && \forall i && (24)\\
\mathrm{E}[Th_i] &\le c\mu_i && \forall i && (25)\\
c\mu_i - \mathrm{E}[Th_i] &\le M(1-\beta_i) && \forall i && (26)\\
\mathrm{E}[L_i^{Q,end}] &\le M\beta_i && \forall i && (27)\\
\beta_i &\le M\mathrm{E}[L_i^{Q,end}] && \forall i && (28)\\
\mathrm{E}[L_i^Q] &= \pi_i \frac{\mathrm{E}[L_{i-1}^{Q,end}] + \mathrm{E}[L_i^{Q,end}]}{2} && \forall i && (29)\\
\ell c\mu_i\pi_i &\ge (\mathrm{E}[L_{i-1}^{Q,end}] + \ell\lambda_i\pi_i)(1-\beta_i)\beta_{i-1} && \forall i && (30)\\
\pi_i &\ge \beta_i && \forall i && (31)\\
\pi_i &\le \beta_{i-1} + \beta_i && \forall i && (32)\\
\mathrm{E}[W_i^Q] &= \beta_{i-1}(1-\beta_i)\frac{\mathrm{E}[L_i^Q]}{\lambda_i} + \beta_i\frac{\mathrm{E}[L_i^Q]}{\mu_i} && \forall i && (33)\\
\mathrm{E}[W_i^S] &= \mathrm{E}[W_i^Q] + \frac{1}{\mu_i} && \forall i && (34)\\
\beta_i &\in \{0,1\} && \forall i && (35)\\
0 \le \pi_i &\le 1 && \forall i && (36)\\
\mathrm{E}[L_i^Q], \mathrm{E}[L_i^{Q,end}], \mathrm{E}[W_i^S], \mathrm{E}[Th_i] &\ge 0 && \forall i && (37)
\end{aligned}$$

Optimal solution in the deterministic system: Proof idea

Let j be the smallest period with $c\overline{\mu} < \lambda_j$, $n+1$ if non-existing.

(i) Consider periods $i = 1, ..., j-1$ sequentially.
- Assume $\mu_i < \frac{\lambda_i}{c} \implies$ Costumers left in the queue at the end of i
  * Leaving costumers in the queue (even tough they could have been served) cannot be optimal
  * Reducing additional queue later on is not optimal, because of the concave value-rate function
- Finding an optimal $\mu_i \geq \frac{\lambda_i}{c}$ leads to the same solution as in the stationary model.

(ii) Consider periods $i = j, ..., n$ sequentially. Objective function is increasing in the service rate μ_i. $\implies \overline{\mu}$ is optimal.

Iterative procedure for the SBC-approach

- Performance approximation quality depends on period length (Stolletz, 2008)
  - * Period length ≈ Processing time
- Optimization of processing rates $\implies$ Good period length not known a priori
- **Key idea:** Evaluation periods with period length $\ell_{eval} = \frac{\ell}{N_{eval}}$
  - * divide decision period into multiple periods ($N_{eval} > 1$) or
  - * unify multiple decision periods to a single larger period ($N_{eval} < 1$)
- Number of decisions remains the same!
- Choose ℓ_{eval} such that

$$\frac{\ell}{N_{eval}} = \ell_{eval} \approx \frac{1}{\bar{\mu}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\right)^{-1} \tag{38}$$

- **Iterative procedure**

  $N_{eval} \leftarrow 1$

  do

  $\mu_i \leftarrow$ solve problem using N_{eval}

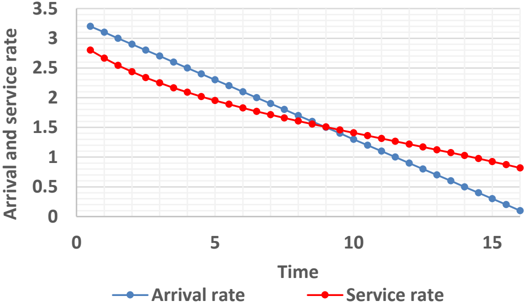
  $N_{eval} \leftarrow \ell\bar{\mu} = \frac{\ell}{n}\sum_{i=1}^{n}\mu_i$

  while N_{eval} different that in previous loop

Deterministic system with decreasing demand

Setting: $c = 1, n = 32, \ell = 0.5, w = 1, \bar{v} = 30, a = 1$

Optimal solution:

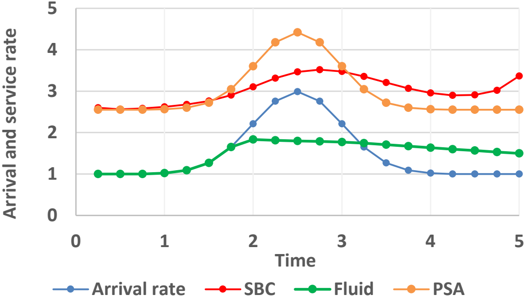


Comment: Anticipation of demand

The future does matter!

Impact of time-dependent decision

Setting: $\ell = 0.25$, $n = 20$, λ_i shows single peak, $w = 10$, $\bar{v} = 40$, $a = 1$

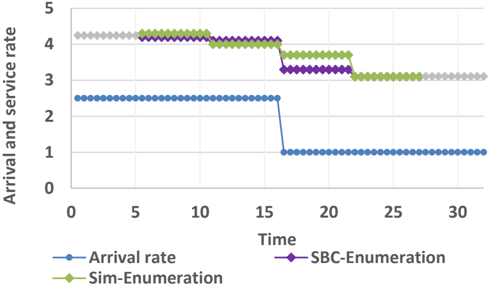


Objective value Z	Integrated model	Simulation	Difference
PSA	48.09	43.69	10.08%
Fluid	139.83	45.83	205.07%
SBC	48.35	46.26	4.52%

Optimal solution via simulation-enumeration

Setting: $\ell = 0.5$, $n = 64$, $\lambda(t)$ decreases at $t = 16$, $w = 10$, $\bar{v} = 30$, $a = 1$

(a) 4 decisions ($t = 5.5, 11, 16.5, 22$) (b) service rates: 2.8, ..., 4.6 (step size 0.1)



	Z	E[Th] total	E[W^S] total
Sim-Enumeration	216.05	55.64	16.35
SBC-Enumeration	217.85	55.76	16.85
SBC-Enumeration simulated	216.01	55.64	16.81
Difference (rel.)	0.02%	0.01%	2.75%