**Technical Appendix to accompany**
**Price and leadtime differentiation, capacity strategy and market competition**

### *Appendix A: Proof of Proposition 1*

It is well known that at optimality, the two leadtime reliability constraints ($7^{DC}$) and ($8^{DC}$) must be binding (Boyaci and Ray 2003). This implies that the two service rates will be given by:

$$\mu_c^i = -\frac{\ln(1-\alpha)}{L_c^i} + \lambda_c^i \quad c \in \{l, h\}$$

As a result, $[PLDP_{DC}]$ reduces to maximizing (3) with $\mu_c^i$ as given above. The system stability conditions ($6^{DC}$) are automatically satisfied by the expressions for $\mu_c^i$. Upon substituting the expressions for $\mu_c^i$ into (3), and taking it partial derivatives with respect to $p_h^i$ and $p_l^i$ gives the following Hessian for a fixed $L_h^i$:

$$\begin{pmatrix} -2(\beta_p^h + \theta_p + \gamma_p) & 2\theta_p \\ 2\theta_p & -2(\beta_p^l + \theta_p + \gamma_p) \end{pmatrix}$$

Clearly, the Hessian is negative definite. This shows that the objective function $\pi^i(L_h^i)$ is strictly concave for a fixed $L_h^i$, and, therefore, has a unique pair of optimal prices $p_h^{i*}(L_h^i)$ and $p_l^{i*}(L_h^i)$, which can be obtained by solving the following system of equations:

$$\frac{\partial \pi^i(L_h^i)}{\partial p_h^i} = 0; \ \frac{\partial \pi^i(L_h^i)}{\partial p_l^i} = 0$$

Substituting the optimal prices given by (10) and (11) into the objective function, and differentiating it with respect to $L_h^i$ gives:

$$\frac{\partial \pi^i(L_h^i)}{\partial L_h^i} = -\left(\beta_L^h + \theta_L + \gamma_L\right)\left(p_h^{i*}(L_h^i) - m^i - A^i\right) + \theta_L\left(p_l^{i*}(L_h^i) - m^i - A^i\right) - \frac{A\ln(1-\alpha)}{(L_h^i)^2} \tag{A1}$$

$$\frac{\partial^2 \pi^i(L_h^i)}{\partial (L_h^i)^2} = -\left(\beta_L^h + \theta_L + \gamma_L\right)\left(\frac{\partial p_h^i(L_h^i)}{\partial L_h^i}\right) + \theta_L\left(\frac{\partial p_l^i(L_h^i)}{\partial L_h^i}\right) + \frac{2A^i ln(1-\alpha)}{(L_h^i)^3} \tag{A2}$$

$$\frac{\partial^3 \pi^i(L_h^i)}{\partial (L_h^i)^3} = -\frac{6A^i\ln(1-\alpha)}{(L_h^i)^4} \tag{A3}$$

The the first three derivatives of $\pi^i(L_h^i)$ suggests that it has the following properties: (i) As $L_h^i \to 0^+$, $\pi^i(L_h^i) \to -\infty$. (ii) $\pi^i(L_h^i)$ is increasing concave in $L_h^i$ in the vicinity of $L_h^i = 0^+$. (iii) As $L_h^i$ increases from 0, $\pi^i(L_h^i)$ changes from concave to convex for some $L_h^i \in (0, +\infty)$, and never becomes concave again. It is clear from the above properties of $\pi^i(L_h^i)$ that it has a unique maximum and at most one minimum in $[0, +\infty)$. The stationary points are given by the roots of (A1) in $[0, +\infty)$, and the maximum is always the smaller of the two. Further, $\frac{\partial \pi^i(L_h^i)}{\partial L_h^i}\big|_{L_h^i = L_l^i} < 0$ is sufficient to guarantee that (A1) has only one root in the interval $[0, L_l^i)$, and that it is the point of maximum. The condition simplifies to:

$$\frac{K_1 a^i + K_2 L_l^i + K_3 A^i + K_4 m^i}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p + \beta_p^h\gamma_p + \beta_p^l\gamma_p + 2\theta_p\gamma_p + \gamma_p^2)} - \frac{A^i\ln(1-\alpha)}{(L_l^i)^2} < 0 \tag{A4}$$

where $K_1$, $K_2$, $K_3$, $K_4$ are functions only of the market parameters ($\beta_p^c$, $\beta_L^c$, $\theta_p$, $\theta_L$, $\gamma_p$, $\gamma_L$), and hence are constants. Further,

$$K_1 = -\left\{(\beta_p^l - \beta_p^h)\theta_L + (\beta_L^h + \gamma_L)(\beta_p^l + 2\theta_p + \gamma_p)\right\}$$

Since $\beta_p^h < \beta_p^l$, a necessary condition for (A4) to hold is $a^i$ to be high. A sufficiently high value of $a^i$ also guarantees $p_c^i > 0$, $p_h^i > p_l^i$ and $\lambda_c^i > 0$.

## *Appendix B: Matrix Geometric Method*

*Joint Stationary Queue Length Distribution:* If we define $N_h(t)$ and $N_l(t)$ as state variables representing the number of high and low priority customers in the system at time t, then $\{\mathbf{N}(t)\} := \{N_l(t), N_h(t), t \geq 0\}$ is a continuous-time two-dimensional Markov chain with state space $\{\mathbf{n} = (n_l, n_h)\}$. The key idea we employ here is that $\{\mathbf{N}(t)\}$ is a *quasi-birth-and-death* (QBD) process, which allows us to develop a matrix geometric solution for the joint distribution of the number of customers of each class in the system. A simple implementation of the matrix geometric method, however, requires the number of states in the QBD process to be finite. For this, we treat the queue length of high priority customers (including the one in service) to be of finite size $M$, but of size large enough for the desired accuracy of our results. Since high priority customers are always served in priority over low priority customers, it is reasonable to assume that its queue size will always be bounded by some large number.

In the Markov process $\{\mathbf{N}(t)\}$, a transition can occur only if a customer of either class arrives or a customer of either class is served. The possible transitions are:

Table 1. Transition rates for the priority queue

| From | To | Rate | Condition |
|---|---|---|---|
| $(n_l, n_h)$ | $(n_l, n_h+1)$ | $\lambda_h^i$ | for $n_l \geq 0$, $n_h \geq 0$ |
| $(n_l, n_h)$ | $(n_l+1, n_h)$ | $\lambda_l^i$ | for $n_l \geq 0$, $n_h \geq 0$ |
| $(n_l, n_h)$ | $(n_l, n_h-1)$ | $\mu^i$ | for $n_l \geq 0$, $n_h > 0$ |
| $(n_l, n_h)$ | $(n_l-1, n_h)$ | $\mu^i$ | for $n_l > 0$, $n_h = 0$ |

The infinitesimal generator Q associated with our system description is thus block-tridiagonal:

$$Q = \begin{pmatrix} B_0 & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

where $B_0$, $A_0$, $A_1$, $A_2$ are square matrices of order $M+1$. These matrices can be easily constructed using the transition rates described above.

$$A_0 = \begin{pmatrix} \lambda_l^i & & & \\ & \lambda_l^i & & \\ & & \ddots & \\ & & & \ddots & \\ & & & & \lambda_l^i \end{pmatrix}; \quad A_2 = \begin{pmatrix} \mu^i & & & \\ & 0 & & \\ & & \ddots & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix}; \quad B_0 = \begin{pmatrix} * & \lambda_h^i & & & \\ \mu^i & * & \lambda_h^i & & \\ & \mu^i & * & \lambda_h^i & \\ & & \ddots & \ddots & \ddots \\ & & & \mu^i & * \end{pmatrix}$$

where $*$ is such that $A_0\mathbf{e} + B_0\mathbf{e} = \mathbf{0}$. $A_1 = B_0 - A_2$.

We denote $\mathbf{x}$ as the stationary probability vector of $\{\mathbf{N}(t)\}$:

$$\mathbf{x} = [x_{00}, x_{01}, \ldots, x_{0M}, x_{10}, x_{11}, \ldots, x_{1M}, \ldots, \ldots, x_{n0}, x_{n1}, \ldots, x_{nM}, \ldots, \ldots]$$

The vector $\mathbf{x}$ can be partitioned by levels into sub vectors $\mathbf{x}_n$, $n \geq 0$, where $\mathbf{x}_n = [x_{n0}, x_{n1}, \ldots, x_{nM}]$ is the stationary probability of states in level $n$ ($n_l = n$). Thus, $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \ldots]$. $\mathbf{x}$ can be obtained using a set of balance equations, given in matrix form, by the following standard relations (Latouche and Ravaswami 1999; Neuts 1981):

$$\mathbf{x}Q = \mathbf{0}; \quad \mathbf{x}_{n+1} = \mathbf{x}_n R$$

where $R$ is the minimal non-negative solution to the matrix quadratic equation:

$$A_0 + RA_1 + R^2 A_2 = \mathbf{0}$$

The matrix $R$ can be computed using well known methods (Latouche and Ravaswami 1999; He 2014). A simple iterative procedure often used is:

$$R(0) = 0 \ ; \quad R(r+1) = -\left[A_0 + R^2(r) A_2\right] A_1^{-1}$$

The probabilities $\mathbf{x}_0$ are determined from:

$$\mathbf{x}_0(B_0 + RA_2) = \mathbf{0}$$

subject to the normalization equation:

$$\sum_{n=0}^{\infty} \mathbf{x}_n \mathbf{e} = \mathbf{x}_0 (I - R)^{-1} \mathbf{e} = 1$$

where $\mathbf{e}$ is a column vector of ones of size $M+1$.
*Estimation of $S_l^i(\cdot)$:* The leadtime $W_l^i$ of a low priority customer is the time between its arrival to the system till it completes service. It may be preempted by one or more high priority customers for service. So it is difficult to characterize the distribution $S_l^i(\cdot)$. Ramaswami and Lucantoni (1985) present an efficient algorithm based on *uniformization* to derive the complimentary distribution of waiting times in phase-type and QBD processes. We adopt their algorithm to derive $S_l^i(\cdot)$, the distribution of the waiting time plus the time in service of low priority customers.

Consider a tagged low priority customer entering the system. The time spent by the tagged customer depends on the number of customers of either class already present in the system ahead of it, and also on the number of subsequent high priority arrivals before it completes its service. All subsequent low priority arrivals, however, have no influence on its time spent in the system. The tagged customer's time in the system is, therefore, simply the time until absorption in a modified Markov process $\{\tilde{\mathbf{N}}(t)\}$, obtained by setting $\lambda_l^i = 0$. Consequently, matrix $\tilde{A}_0$, representing transitions to a higher level, becomes a zero matrix. We define an *absorbing* state, call it state $0'$, as the state in which the tagged customer has finished its service. The infinitesimal generator for this process can be represented as:

$$\tilde{Q} = \left(\begin{array}{c|cccccc} 0 & 0 & 0 & 0 & 0 & \cdots & \\ \hline b_0 & \tilde{B}_0 & 0 & & & & \\ 0 & A_2 & \tilde{A}_1 & 0 & & & \\ 0 & & A_2 & \tilde{A}_1 & 0 & & \\ \vdots & & & \ddots & \ddots & \ddots & \end{array}\right)$$

where, $\tilde{B}_0 = B_0 + A_0$; $\tilde{A}_1 = A_1 + A_0$; and $b_0 = [\mu^i \quad 0 \quad \cdots \quad 0]^T_{M+1}$. The first row and column in $\tilde{Q}$ corresponds to the absorbing state $\acute{0}$. The time spent in system by the tagged customer, which is the time until absorption in the modified Markov process with rate matrix $\tilde{Q}$, depends on the prices ($p^i_h$ and $p^i_l$), through the arrival rates ($\lambda^i_h$ and $\lambda^i_l$), and the service rate $\mu^i$. For given prices ($p^{ik}_h$, $p^{ik}_l$) and service rate $\mu^{ik}$, the distribution of the time spent by a low priority customer in the system is $S^{ik}_l(y) = 1 - \overline{S^{ik}_l}(y)$, where $\overline{S^{ik}_l}(y)$ is the stationary probability that a low priority customer spends more than $y$ units of time in the system. Further, let $\overline{S^k_{ln}}(y)$ denote the conditional probability that a tagged customer, who finds $n$ low priority customers ahead of it, spends a time exceeding $y$ in the system. The probability that a tagged customer finds $n$ low priority customers is given, using the PASTA property, by $\mathbf{x}_n = \mathbf{x}_0 R^n$. $\overline{S^{ik}_l}(y)$ can be expressed as:

$$\overline{S^{ik}_l}(y) = \sum_{n=0}^{\infty} \mathbf{x}_n \overline{S^{ik}_{ln}}(y)\mathbf{e} \tag{B1}$$

$\overline{S^{ik}_{ln}}(y)$ can be computed more conveniently by uniformizing the Markov process $\{\tilde{\mathbf{N}}(t)\}$ with a Poisson process with rate $\gamma$, where

$$\gamma = \max_{0 \le m \le M} (-\tilde{A}_1)_{mm} = \max_{0 \le m \le M} -(A_0 + A_1)_{mm}$$

so that the rate matrix $\tilde{Q}$ is transformed into the discrete-time probability matrix:

$$\hat{Q} = \frac{1}{\gamma}\tilde{Q} + I = \left(\begin{array}{c|ccccc} 1 & 0 & 0 & 0 & 0 & \cdots \\ \hline \hat{b_0} & \hat{B}_0 & 0 & & & \\ 0 & \hat{A}_2 & \hat{A}_1 & 0 & & \\ 0 & & \hat{A}_2 & \hat{A}_1 & 0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{array}\right)$$

where $\hat{A}_2 = \frac{A_2}{\gamma}$, $\hat{A}_1 = \frac{\tilde{A}_1}{\gamma} + I$, $\hat{b_0} = \frac{b_0}{\gamma}$. In this uniformized process, points of a Poisson process are generated with a rate $\gamma$, and transitions occur at these epochs only. The probability that $r$ Poisson events are generated in time $y$ equals $e^{-\gamma y}\frac{(\gamma y)^r}{r!}$. Suppose the tagged customer finds $n$ low priority customers ahead of it. Then, for its time in system to exceed $y$, at most $n$ of the $r$ Poisson points may correspond to transitions to lower levels (i.e., service completions of low priority customers). Therefore,

$$\overline{S^{ik}_{ln}}(y) = \sum_{r=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^r}{r!}\sum_{v=0}^{n} G^{(r)}_v \mathbf{e}, \qquad n \ge 0 \tag{B2}$$

where, $G^{(r)}_v$ is a matrix such that its entries are the conditional probabilities, given that the system has made $r$ transitions in the discrete-time Markov process with rate matrix $\hat{Q}$, that $v$ of those transitions correspond to lower levels (i.e., service completions of low priority customers). Substituting the expression for $\overline{S^{ik}_{ln}}(y)$ from (B2) into (B1), we obtain:

$$\overline{S^{ik}_l}(y) = \sum_{r=0}^{\infty} d_r e^{-\gamma y}\frac{(\gamma y)^r}{r!} \tag{B3}$$

where, $d_r$ is given by:

$$d_r = \sum_{n=0}^{\infty} \mathbf{x}_0 R^n \sum_{v=0}^{n} G_v^{(r)} \mathbf{e}, \qquad r \geq 0 \tag{B4}$$

Now,

$$\begin{aligned}
&\sum_{n=0}^{\infty} R^n \sum_{v=0}^{n} G_v^{(r)} \mathbf{e} \\
&= \sum_{n=0}^{r+1} R^n \sum_{v=0}^{n} G_v^{(r)} \mathbf{e} + \sum_{n=r+2}^{\infty} R^n \sum_{v=0}^{r} G_v^{(r)} \mathbf{e} && \left(\text{since } G_v^{(r)} = 0 \text{ for } v > r\right) \\
&= \sum_{v=0}^{r+1} \sum_{n=v}^{r+1} R^n G_v^{(r)} \mathbf{e} + (I-R)^{-1} R^{r+2} \mathbf{e} && \left(\text{since } \sum_{v=0}^{r} G_v^{(r)} \mathbf{e} = \mathbf{e}\right) \\
&= \sum_{v=0}^{r+1} (I-R)^{-1} (R^v - R^{r+2}) G_v^{(r)} \mathbf{e} + (I-R)^{-1} R^{r+2} \mathbf{e} \\
&= \sum_{v=0}^{r} (I-R)^{-1} R^v G_v^{(r)} \mathbf{e} + (I-R)^{-1} R^{r+1} G_{r+1}^{(r)} \mathbf{e} && \left(\text{since} \sum_{v=0}^{r+1} G_v^{(r)} \mathbf{e} = \mathbf{e}\right) \\
&= \sum_{v=0}^{r} (I-R)^{-1} R^v G_v^{(r)} \mathbf{e} && \left(\text{since } G_v^{(r)} = 0 \text{ for } v > r\right) \\
&= (I-R)^{-1} H_r \mathbf{e} && r \geq 0
\end{aligned}$$

where, $H_r = \sum_{v=0}^{r} R^v G_v^{(r)}$. Therefore,

$$S_l^{ik}(L_l^i) = 1 - \overline{S_l^{ik}}(L_l^i) = \sum_{r=0}^{\infty} e^{-\gamma L_l} \frac{(\gamma L_l)^r}{r!} \mathbf{x}_0 (I-R)^{-1} H_r \mathbf{e} \tag{B5}$$

$H_r$ can be computed recursively as:

$$H_{r+1} = H_r \hat{A}_1 + R H_r \hat{A}_2; \quad H_0 = I$$

Therefore, for given prices ($p_h^{ik}$, $p_l^{ik}$) and service rate ($\mu^{ik}$), $S_l^{ik}(\cdot)$ in (16) can be computed using (B5).

### *Appendix C: Estimation of the Gradient of $S_l^i(\cdot)$*

There are several methods available in the literature to compute the gradients of $S_l^i(\cdot)$. We use a *finite difference* method as it is probably the simplest and most intuitive, and can be easily explained (Atlason, Epelman and Henderson 2004). Using the finite difference method, the gradients can be computed as:

$$\frac{\partial S_l^{ik}(\cdot)}{\partial p_h^i} = \frac{S_l^{i}(\cdot)\big|_{(p_h^{ik}+dp_h^i, p_l^i, \mu^i)} - S_l^{ik}(\cdot)\big|_{(p_h^{ik}-dp_h^i, p_l^i, \mu^i)}}{2dp_h^i}$$

$$\frac{\partial S_l^{ik}(\cdot)}{\partial p_l^i} = \frac{S_l^{ik}(\cdot)\big|_{(p_h^i, p_l^{ik}+dp_l^i, \mu^i)} - S_l^{ik}(\cdot)\big|_{(p_h^i, p_l^{ik}-dp_l^i, \mu^i)}}{2dp_l^i}$$

$$\frac{\partial S_l^{ik}(\cdot)}{\partial \mu^i} = \frac{S_l^{ik}(\cdot)\big|_{(p_h^i, p_l^i, \mu^{ik}+d\mu^i)} - S_l^{ik}(\cdot)\big|_{(p_h^i, p_l^i, \mu^{ik}-d\mu^i)}}{2d\mu^i}$$

where $dp_h^i$, $dp_l^i$ and $d\mu^i$ (referred to as step sizes) are infinitesimal changes in the respective variables.

### *Appendix D: The Cutting Plane Algorithm*

We now describe the cutting plane algorithm to solve $[PDP_{(K)}]$. The algorithm fits the framework of Kelley's cutting plane method (Kelley 1960). It differs from the traditional description of the algorithm in that we use the matrix geometric method to generate the cuts and evaluate the function values instead of having an algebraic form for the function and using analytically determined gradients to generate the cuts. Figure 1 shows a flowchart of the cutting plane algorithm. The

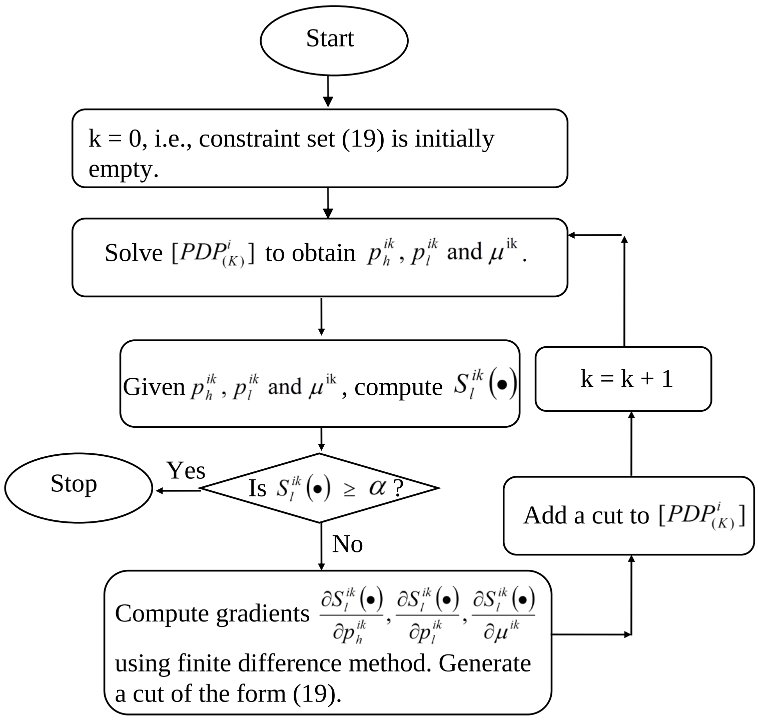


Figure 1. Cutting Plane Algorithm

algorithm works as follows: We start with an empty constraint set (19), which results in a simple QPP, and obtain an initial solution $k_0 := (p_h^{i0},\ p_l^{i0},\ \mu^{i0})$. We use the matrix geometric method to compute the distribution $S_l^{ik_0}(\cdot)$ of $W_l^i$. If $S_l^{ik_0}(\cdot)$ meets the leadtime reliability constraint $\alpha$, we stop with an optimal solution to $[PDP^i_{(K)}]$, else we add to (19) a linear constraint/cut generated using the finite difference method. The new cut eliminates the current solution but does not eliminate any feasible solution to $[PDP^i_{(K)}]$. This procedure repeats until the leadtime reliability constraint is satisfied within a sufficiently small tolerance limit $\epsilon$ such that $\left|S_l^i(\cdot) - \alpha\right| \leq \epsilon$. The method has been proved to converge (Atlason, Epelman and Henderson 2004).

The success of the cutting plane algorithm relies on the concavity of $S_l^i(\cdot)$. We have already demonstrated, using computational results obtained by the matrix geometric method, that $S_l^i(\cdot)$ is concave in $(p_h^i,\, p_l^i)$ and separately concave in $\mu^i$. However, it is difficult to establish the joint concavity of $S_l^i(\cdot)$ in $(p_h^i,\, p_l^i,\, \mu^i)$. If the concavity assumption is violated, then the algorithm may cut off parts of the feasible region and terminate with a solution that is suboptimal. We include a test to ensure the concavity assumption is not violated. This is done by ensuring that a new point, visited by the cutting plane algorithm after each iteration, lies below all the previously defined cuts, and that all previous points lie below the newly added cut. The test, however, cannot ensure that $S_l^i(\cdot)$ is concave unless it examines all the points in the feasible region. Still, it does help ensure that the concavity assumption is not violated at least in the region visited by the algorithm. Details of the test can be found in Atlason, Epelman and Henderson (2004).

## *Appendix E: Proof of Proposition 3*

The equilibrium prices are given by the simultaneous solution of the 4 linear equations given by (10) and (11) for $i \in \{1, 2\}$. The system of equations in matrix notation is given by $\mathbf{Ax} = \mathbf{b}$.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \frac{-(\beta_p^l+\theta_p+\gamma_p)\gamma_p}{2D} & \frac{-\theta_p\gamma_p}{2D} \\ 0 & 1 & \frac{-\theta_p\gamma_p}{2D} & \frac{-(\beta_p^h+\theta_p+\gamma_p)\gamma_p}{2D} \\ \frac{-(\beta_p^l+\theta_p+\gamma_p)\gamma_p}{2D} & \frac{-\theta_p\gamma_p}{2D} & 1 & 0 \\ \frac{-\theta_p\gamma_p}{2D} & \frac{-(\beta_p^h+\theta_p+\gamma_p)\gamma_p}{2D} & 0 & 1 \end{pmatrix} \tag{E1}$$

where $D = \beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p + \beta_p^h\gamma_p + \beta_p^l\gamma_p + 2\theta_p\gamma_p + \gamma_p^2$

$$\mathbf{x} = \left(p_h^{1*}\ p_l^{1*}\ p_h^{2*}\ p_l^{2*}\right)^T$$

and $\mathbf{b}$ is a 4x1 matrix of constants. $\mathbf{A}$ is symmetric and strictly diagonally dominant since we have $A_{ij} = A_{ji}\ \forall i, j$ and $\sum_{j\neq i} |A_{ij}| < A_{ii}\ \forall i$. Hence, $\mathbf{A}$ is positive definite (Horn and Johnson 1985). This implies that $\mathbf{A}$ is full-rank, and hence the system of linear equations $\mathbf{Ax} = \mathbf{b}$ has a unique solution. This proves the uniqueness of the equilibrium.

Further, when the firms are identical, they have the same operating parameter values ($a^1 = a^2$; $m^1 = m^2$; $A^1 = A^2$; $\alpha^1 = \alpha^2$; $L_l^1 = L_l^2$; $L_h^1 = L_h^2$). Denote the equilibrium solution by the 2-tuple $(s^{1*}(L_h), s^{2*}(L_h))$, where $s^{i*}(L_h) := (p_h^{i*}(L_h), p_l^{i*}(L_h))$. Assume the contrary that the equilibrium solution is not symmetric, i.e., $s^{1*}(L_h) \neq s^{2*}(L_h)$. Since the two firms are identical, this implies that $(s^{2*}, s^{1*})$ must also be a Nash Equilibrium, which contradicts the uniqueness of the Nash Equilibrium. Hence, $s^{1*}(L_h) = s^{2*}(L_h)$. Substituting $p_h^{1*}(L_h) = p_h^{2*}(L_h) = p_h^*(L_h)$ and $p_l^{1*}(L_h) = p_l^{2*}(L_h) = p_l^*(L_h)$ in the expressions for the best response prices, given by (10) and (11), and solving the resulting system of 2 equations in 2 unknown gives (E2) and (E3).

$$\begin{aligned} p_h^*(L_h) = &\frac{(2\beta_p^l + 4\theta_p + \gamma_p)a - \{\beta_L^h(2\beta_p^l + 2\theta_p + \gamma_p) + (2\beta_p^l + \gamma_p)\theta_L\}L_h}{D_1} \\ &\frac{+\{(2\beta_p^l + \gamma_p)\theta_L - 2\beta_L^l\theta_p\}L_l}{D_1} \\ &\frac{+(2\beta_p^h\beta_p^l + 2\beta_p^h\theta_p + \beta_p^h\gamma_p + 2\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2)(A + m)}{D_1} \end{aligned} \tag{E2}$$

$$p_l^*(L_h) = \frac{(2\beta_p^h + 4\theta_p + \gamma_p)a + \{(2\beta_p^h + \gamma_p)\theta_L - 2\beta_L^h\theta_p\}L_h}{D_1} \\ \frac{-\{\beta_L^l(2\beta_p^h + 2\theta_p + \gamma_p) + (2\beta_p^h + \gamma_p)\theta_L\}L_l}{D_1} \\ \frac{+(2\beta_p^l\beta_p^h + 2\beta_p^l\theta_p + \beta_p^l\gamma_p + 2\beta_p^h\theta_p + 2\beta_p^h\gamma_p + 4\theta_p\gamma_p + \gamma_p^2)(A+m)}{D_1} \tag{E3}$$

where $D_1 = 4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 2\beta_p^h\gamma_p + 4\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2$.

### *Appendix F: Proof of Proposition 4*

The duopoly prices under pure price competition are given by (E2) and (E3). The monopolist prices can be obtained from (E2) and (E3) by substituting $\gamma_p = \gamma_L = 0$. Comparing the monopolist prices with the duopoly prices, we get:

$$p_h^{DD*}(L_h)\Big|_{duopoly} - p_h^{DC*}(L_h)\Big|_{monopoly} \\ = \frac{-\gamma_p\left\{K_1^h a + K_2^h L_h + K_3^h L_l + K_4^h(A+m)\right\}}{4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 4\beta_p^l\theta_p + 2\beta_p^h\gamma_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2} \tag{F1}$$

$$p_l^{DD*}(L_h)\Big|_{duopoly} - p_l^{DC*}(L_h)\Big|_{monopoly} \\ = \frac{-\gamma_p\left\{K_1^l a + K_2^l L_h + K_3^l L_l + K_4^l(A+m)\right\}}{4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 4\beta_p^l\theta_p + 2\beta_p^h\gamma_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2} \tag{F2}$$

$$\left(p_h^{DD*}(L_h) - p_l^{DD*}(L_h)\right)\Big|_{duopoly} - \left(p_h^{DC*}(L_h) - p_l^{DC*}(L_h)\right)\Big|_{monopoly} \\ = \frac{-\gamma_p\left\{K_1^d a + K_2^d L_h + K_3^d L_l + K_4^d(A+m)\right\}}{4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 4\beta_p^l\theta_p + 2\beta_p^h\gamma_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2} \tag{F3}$$

where, $K_i^d = K_i^h - K_i^l$, and $K_i^h$ and $K_i^l$ for $i \in \{1, 4\}$ are some functions of system parameters, and hence Clearly, when $\gamma_p = 0$, $p_h^{DD*}(L_h)\Big|_{duopoly} = p_h^{DC*}(L_h)\Big|_{monopoly}$ and $p_l^{DD*}(L_h)\Big|_{duopoly} = p_l^{DC*}(L_h)\Big|_{monopoly}$. For, $\gamma_p > 0$, (F1), (F2) and (F3) are dictated mainly by $K_1^h$ and $K_1^l$ and $K_1^d$, respectively since $a$ is assumed to be large. Further,

$$K_1^h = 2(\beta_p^l)^2 + 2\beta_p^h\theta_p + 6\beta_p^l\theta_p + 8\theta_p^2 + \beta_p^l\gamma_p + 2\theta_p\gamma_p > 0$$

$$K_1^l = 2(\beta_p^h)^2 + 6\beta_p^h\theta_p + 2\beta_p^l\theta_p + 8\theta_p^2 + \beta_p^h\gamma_p + 2\theta_p\gamma_p > 0$$

$$K_1^d = (\beta_p^l - \beta_p^h)\gamma_p + 2\{(\beta_p^l)^2 - (\beta_p^h)^2\} + 4(\beta_p^l - \beta_p^h)\theta_p > 0$$

Therefore, $K_1^h > 0$, $K_1^l > 0$ and $K_1^d > 0 \Rightarrow$ (F1) $< 0$, (F2) $< 0$ and (F3) $< 0$, respectively if $\gamma_p > 0$. This shows that pure price competition decreases both the express and regular prices as well as the price differentiation. Further, it clearly follows from the expressions for $K_1^h$, $K_1^l$ and $K_1^d$ that the effects are more pronounced when $\theta_p > 0$, i.e., in presence of product substitution.

### *Appendix G: Proof of Proposition 5*

Given the strategy of firm $j \in \{1, 2\}$, the best response express leadtime of firm $i = 3 - j$ satisfies:

$$\frac{\partial \pi^i}{\partial L_h^i} = 0$$

Taking the total derivative of the above relation with respect to the express leadtime $L_h^j$ of firm $j$, we get:

$$\begin{aligned}\frac{d}{dL_h^j}\left(\frac{\partial \pi^i}{\partial L_h^i}\right) &= \frac{\partial}{\partial L_h^j}\left(\frac{\partial \pi^i}{\partial L_h^i}\right) + \frac{\partial}{\partial p_h^j}\left(\frac{\partial \pi^i}{\partial L_h^i}\right)\frac{\partial p_h^j}{\partial L_h^j} \\ &+ \frac{\partial}{\partial p_l^j}\left(\frac{\partial \pi^i}{\partial L_h^i}\right)\frac{\partial p_l^j}{\partial L_h^j} + \frac{\partial}{\partial L_h^i}\left(\frac{\partial \pi^i}{\partial L_h^i}\right)\frac{dL_h^i}{dL_h^j} = 0\end{aligned}$$

$$\Rightarrow \frac{dL_h^i}{dL_h^j} = \frac{-\left[\frac{\partial^2 \pi^i}{\partial L_h^j \partial L_h^i} + \frac{\partial^2 \pi^i}{\partial p_h^j \partial L_h^i}\frac{\partial p_h^j}{\partial L_h^j} + \frac{\partial^2 \pi^i}{\partial p_l^j \partial L_h^i}\frac{\partial p_l^j}{\partial L_h^j}\right]}{\frac{\partial^2 \pi^i}{\partial (L_h^i)^2}}$$

For a DD setting, the above relation simplifies to:

$$\frac{dL_h^i}{dL_h^j} = \frac{-\left[\gamma_p\left\{\left(\frac{\partial p_h^j}{\partial L_h^j}\right)^2 + \left(\frac{\partial p_l^j}{\partial L_h^j}\right)^2\right\} + \gamma_L\left(\frac{\partial p_h^j}{\partial L_h^j}\right)\right]}{\frac{\partial^2 \pi^i}{\partial (L_h^i)^2}} \tag{G1}$$

We know that for $L_h \leq L_h^*$:

$$\frac{\partial^2 \pi^i}{\partial (L_h^i)^2} < 0$$

The numerator in RHS of (G1) consists of terms that are functions only of the market parameters, and hence is a constant for a given parameter setting. Further,

$$\gamma_p\left\{\left(\frac{\partial p_h^j}{\partial L_h^j}\right)^2 + \left(\frac{\partial p_l^j}{\partial L_h^j}\right)^2\right\} > 0 \; and \; \gamma_L\left(\frac{\partial p_h^j}{\partial L_h^j}\right) < 0$$

Therefore, we have:

$$\frac{dL_h^i}{dL_h^j} \geq 0 \, if \, \gamma_p \left\{ \left( \frac{\partial p_h^j}{\partial L_h^j} \right)^2 + \left( \frac{\partial p_l^j}{\partial L_h^j} \right)^2 \right\} \geq \gamma_L \left( \frac{\partial p_h^j}{\partial L_h^j} \right) \tag{G2}$$

$$\frac{dL_h^i}{dL_h^j} < 0 \, if \, \gamma_p \left\{ \left( \frac{\partial p_h^j}{\partial L_h^j} \right)^2 + \left( \frac{\partial p_l^j}{\partial L_h^j} \right)^2 \right\} < \gamma_L \left( \frac{\partial p_h^j}{\partial L_h^j} \right) \tag{G3}$$

This suggests that if the market parameters are such that (G2) holds, firm $i$ always increases (decreases) its express leadtime $L_h^i$ in response to a corresponding increase (decrease) in firm $j's$ express leadtime $L_h^j$. We let $p_h^i(n)$, $p_l^i(n)$ and $L_h^i(n)$ be the best response decisions of firm $i$ at the $n^{th}$ iteration of the procedure. If $L_h^i(0) = 0$, then $L_h^i(n) \geq L_h^i(0)$ for all $n$. We will show that if (G2) holds, $L_h^i(n)$ is increasing in $n$ for $i \in \{1, 2\}$. As $L_h^i$ is bounded above ($L_h^i < L_l$), for $i \in \{1, 2\}$, this will establish that the iterative procedure converges. We prove the convergence by induction as follows:

(1) (Step $n = 1$): We know that $L_h^i(1) \geq L_h^i(0)$ for $i \in \{1, 2\}$.
(2) (Step $n - 1$): Assume that $L_h^i(n-1) \geq L_h^i(n-2)$ for $i \in \{1, 2\}$.
(3) (Step $n$): Given the inductive assumption from Step $n-1$, (G2) implies that $L_h^i(n) \geq L_h^i(n-1)$ for $i \in \{1, 2\}$.

This completes our induction. In case (G3) holds, convergence of the algorithm can proved similarly by letting $L_h^1(0) = L_l$ and $L_h^2(0) = 0$ and by showing that $L_h^1(n)$ is decreasing in $n$ while $L_h^2(n)$ is increasing in $n$.

### *Appendix H: Proof of Proposition 6*

The effect of competition on the express leadtime when firms use dedicated capacities is given by:

$$\left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{duopoly} - \left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{monopoly}$$
$$= \frac{-\{K_1 a + K_2 L_h + K_3 L_l + K_4(A+m)\}}{2(4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 4\beta_p^l\theta_p + 2\beta_p^h\gamma_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2)(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{H1}$$

where, $K_1$, $K_2$, $K_3$ and $K_4$ are some functions only of the system parameters, and hence are constants. For large $a$, (H1) is dictated mainly by $K_1$, which is given by:

$$\begin{aligned} K_1 = & \left\{ 4\beta_p^h(\beta_p^l)^2 + 4(\beta_p^l)^2\theta_p + 8\beta_p^h\theta_p^2 + 8\beta_p^l\theta_p^2 + 12\beta_p^h\beta_p^l\theta_p \right\} \gamma_L \\ & - \left\{ \beta_L^h\beta_p^l + 2\beta_L^h\theta_p + [(\beta_p^l)^2 - (\beta_p^h)^2]\theta_L \right\} \gamma_p^2 \\ & - \left\{ 2\beta_p^h\beta_L^h\theta_p + 2(\beta_p^l)^2\beta_L^h + 4\beta_p^l\theta_L + 8\beta_L^h\theta_p^2 \right\} \gamma_p \\ & - \left\{ [6\beta_p^l\beta_L^h - 4\beta_p^h\theta_L]\theta_p + 2[(\beta_p^l)^2 - (\beta_p^h)^2]\theta_L \right\} \gamma_p \\ & + 2\left\{ \beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p \right\} \gamma_L\gamma_p \end{aligned} \tag{H2}$$

Clearly, the effect of competition on $L_h$, and hence on leadtime differentiation, depends on the relative intensities of price competition ($\gamma_p$) and leadtime competition ($\gamma_L$), as well as other demand parameters. $\gamma_p = 0$ and $\gamma_L > 0$ results in (H2) $> 0$, and hence (H1) $< 0$. Further, $\pi(L_h)$ is increasing concave in $L_h$ for $L_h \leq L_h^{DC*}$ (see Appendix A). This, together with (H1) $< 0$, implease that:

$$L_h^*|_{duopoly} := \{L_h|_{duopoly} : \left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{duopoly} = 0\} < L_h^*|_{monopoly} := \{L_h|_{monopoly} : \left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{monopoly} = 0\}$$

This implies that $L_h$ is smaller under competition when $\gamma_p = 0$. Further, (F1), (F2) and (F3) suggest that for a given $L_h$, the equilibrium prices as well as the price differentiation coincide with the monopolist prices and price differentiation for $\gamma_p = 0$. However, a smaller $L_h$ under monopoly compared to duopoly results in a larger price differentiation for $\gamma_p = 0$. $\gamma_p > 0$ and $\gamma_L = 0$, on the other hand, results in (H2) $< 0$, and hence (H1) $> 0$. Thus, $L_h$ is larger under competition. A larger $L_h$ results in a smaller price differentiation.

## References

Atlason J., Epelman M.A., Henderson S.G., 2004. Call center staffing with simulation and cutting plane methods. Annals of Operations Research 127, 333–358.

Boyaci, T., Ray, S., 2003. Product differentiation and capacity selection cost interaction in time and price sensitive markets. Manufacturing and Service Operations Management 5, 18–36.

He, Q. M., 2014. Fundamentals of matrix-analytic methods. New York: Springer.

Horn, R.A., Johnson, C.R., 1985. Matrix Analysis. Cambridge University Press, Cambridge, UK.

Kelley, J.E. Jr., 1960. The cutting plane method for solving convex programs. Journal of the Society for Industrial and Applied Mathematics 8, 703–711.

Latouche, G., Ramaswami, V., 1999. Introduction to matrix analytic methods in stochastic modeling. The American Statistical Association and the Society for Industrial and Applied Mathematics, Philadelphia, USA.

Neuts, F.M., 1981. Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach. Dover Publications, Mineola, NY, USA.

Ramaswami, V., Lucantoni, D.M., 1985. Stationary waiting time distribution in queues with phase type service and in quasi-birth-and-death processes. Stochastic Models 1, 125–136.