*Quality Technology & Quantitative Management*
Vol. 6, No. 2, pp. 87-106, 2009

QTQM


# Balancing Stochastic Resource Criticalities Hierarchically for Optimal Economic Performance and Growth*

Dan Trietsch[1] and Francisco Quiroga[2]

[1]College of Engineering, American University of Armenia, Yerevan, Armenia
[2]Villacero, Monterrey, Mexico
(*Received* October 2005*, accepted* December 2006)

**Abstract:** Traditionally, resource balance has been measured by utilization. Because 100% utilization of all resources all the time is impossible, some advocate intentional imbalance so all resources except one—the bottleneck (BN)—have enough excess capacity to enable 100% BN utilization. Instead of focusing on utilization, we demonstrate that economic balance requires allocating stochastic criticality, i.e., *permission to fail* (PTF), to resources correctly. Exact and approximate economic balance models for a simple system with multiple inputs (resources) and one output (product) are presented. The approximation allocates PTF in proportion to a measure of the economic value of each resource. Both models can be used to achieve balanced growth and improvement in hierarchical systems. Numerical results demonstrate the robustness of the approximation. We also study the merit of pursuing 100% BN utilization.

Keywords: Management by constraints (MBC), newsboy problem, risk, stochastic economic balance, theory of constraints (TOC).

## 1. Introduction

Ideally, capacity balance implies fully utilizing all the resources all the time. We refer to this as *naive balance.* Naive balance works very well in perfectly designed systems that have no variation and no scheduling conflicts. That is, it doesn't. A related approach is *balancing utilization* perfectly at some level below 100%, e.g., 80%. In general, however, focusing on utilization neglects important stochastic aspects of system behavior, such as on-time performance. Considering randomness more explicitly, we define *simple balance* as letting each resource be equally likely to limit the system; i.e., all parts of the system have the same *permission to fail* (PTF). In sharp contrast to both simple balance and balanced utilization, some authors advocate pursuing 100% utilization on one resource, the *bottleneck* (BN) (Goldratt [6]). Equivalently, this allocates 100% PTF to the BN (because no other resource is allowed to limit the system). This requires subordinating the whole system to the needs of the BN to make sure that it will neither starve nor get blocked. We refer to this as *BN-subordination*. BN-subordination is touted to improve upon JIT, and adopts the JIT principle of seeking balance in terms of a smooth flow of materials in the system (Schonberger [14]). To address stochastic issues, the JIT approach combines proactive variance reduction and protection by appropriate buffers. These include [often large]

* "Make everything as simple as possible, but not simpler" (Albert Einstein).
"Celestial navigation is based on the premise that the Earth is the center of the universe. The premise is wrong, but the navigation works. An incorrect model can be a useful tool" (Kelvin Throop III, as quoted in FamousQuotations Network). Paraphrased as: "All models are wrong, some models are useful."

capacity buffers and [ostensibly very small] WIP inventory buffers. For example, Toyota treats machines as a cheap resource that should rarely be allowed to limit the system: their labor cost is roughly five times higher, so they prefer idle machines over idle workers, and machine utilization is not even measured. In contrast, their combative attitude to WIP is well known—motivated, *inter alia,* to support variance reduction (Shingo [15]). Under BN-subordination, WIP inventory buffers are specified, but only in front of the BN. Furthermore, allowing large capacity buffers on all non-BN resources, as BN-subordination does, also fits Toyota's model. But BN-subordination departs from JIT by espousing an intuitively attractive but flawed focusing principle (Trietsch [20]). For this reason, to the extent that BN-subordination differs from JIT, it is fallacious. Mukherjee and Chatterjee [11] suggest that a careful definition of the term "bottleneck" can help clarify the issue. They assume a deterministic environment with discrete capacity decisions and scheduling conflicts and define a bottleneck as a resource whose *average shadow price* is positive. Earlier academic papers that seek to balance the utilization of non-BN resources while adopting BN-subordination (under the assumption that the bottleneck is simply the most constrained resource) include Atwater and Chakravorty [1] and Ronen and Spector [13].

The main focus of this paper is on correct balance *given* the variation in the system. We seek economic balance between resources, to maximize profit while accounting for their cost. Although—unlike Mukherjee and Chatterjee [11]— we focus on stochastic variation and allow continuous capacity increases, our approach—like theirs—can be described as adjusting shadow prices to match economic expansion costs. A secondary focus is on BN-subordination. Although we present some analytic results, our main purpose is to find useful approximations. We believe this to be of significant value because all useful management science models involve approximations, at the very least in the degree to which they model reality and in the estimation of parameters. In stochastic analysis, approximations have been used extensively, e.g., using diffusion models based on high traffic assumptions to analyze queueing systems. This is also true for papers related to our subject. For instance, Bradley and Glynn [4] use diffusion to study the correct balance of production capacity and finished goods inventory (FGI) in a make-to-stock system with a single machine and a single product. They obtain asymptotically optimal approximate closed-form expressions. The approximation remains valid even for moderate traffic, and only fails at low traffic volumes (i.e., when FGI is relatively expensive). Trietsch [17] addresses a wider production system with multiple resources and several products subject to long-term demand variation. Harrison and Van Mieghem [7], independently, present a similar model. Trietsch and Buzacott [18] discuss the results of Trietsch [17] in a hierarchical context (and much of this paper is based on it). Van Mieghem and Rudi [23] generalize the results of Harrison and Van Mieghem [7] to multi-period multi-process networks. Van Mieghem [24] surveys such results and provides further references. The last six sources, in different ways, yield newsboy-type results: Optimal decisions involve optimal probabilities of shortage (i.e., PTFs) determined by long- and short costs. The long costs are associated with safety buffers, and the short cost is the penalty for not having enough safety.

Whereas our focus is on capacity and capacity buffers, there is a dual problem that is concerned with safety time, seeking to minimize the total expected delay penalty and holding costs of an assembly. This problem has been solved by an exact generalization of the newsboy model. For completeness, we will summarize this solution later. Our first capacity model is a straightforward extension of the same generalized newsboy model. It assumes that capacity can be increased by a constant without changing the distribution. For example, this assumption is correct when we can outsource some load at a fixed price per

unit, in which case the net effect is equivalent to adding a constant to our true capacity. Our second model is approximate (except in special cases), and based on a different assumption about stochastic expansion. However, numerical evidence suggests it is more robust than the analytical model when we don't know which assumption to use. It is also easier to estimate the necessary parameters for it. Thus our main new result is this second model. Our results are more complete than former contributions, because we explicitly include more types of resources—not just those that are conventionally recognized as production resources. Specifically, we include both supply and demand as high level resources—seeking to maximize min{supply, demand} subject to a budget constraint on resources. Including demand in the model is justified because for any organization the throughput is determined by both supply and demand and every organization has tools to influence both supply and demand. In addition, unlike Bradley and Glynn [4] and Harrison and Van Mieghem [7], we do not require a high traffic assumption and we use a richer model for stochastic expansion that considers the effects of expansion on the variance and not just on the mean. Last but not least, our models are designed for hierarchical systems: any group of resources can be combined to one resource that is then subject to the same balance conditions. Thus we obtain a focusing tool for continuous improvement efforts and for growth, all within hierarchical systems. However, although the first model can be extended to the multi-product multi-period case (by interpreting results previously obtained by Trietsch [17]), we limit our scope here to the single product.

Section 2 presents examples illustrating how and where our models can be applied, and that they may be useful in spite of their simplicity. Technically, this section may be skipped without loss of continuity. Section 3 presents the dual model. Section 4 defines the problem formally and develops the models. Section 5 discusses the potential gain associated with moving towards balance. Section 6 addresses risk and how to account for negative investments. It also shows how to solve for tight resources that cannot be expanded in the short term and the question is how much capacity to provide on adjustable resources (i.e., how to balance a system with an immutable BN). Section 7 presents numerical results demonstrating the quality and robustness of the models, and compares the results to simple balance and to BN-subordination. Section 8 illustrates the use of our main models for designing a balanced new facility. Section 9 is the conclusion.

## 2. Example

Our first example is loosely based on a classic application of the *machine interference model*. When a thread breaks on an automatic loom, it stops and a light (andon) indicates that the machine requires service by a *sider*. The sider rethreads the yarn in the spindle and restarts the machine. Suppose there are $s$ siders and $m \geq s$ machines that may require service at random. This defines a queueing system with $s$ servers. Let $\pi_k$ denote the long-term frequency of observing $k$ machines in the queueing system, i.e., $(k - s)^+$ machines in queue and $(s - k)^+$ idle siders (where $y^+ = \max\{y, 0\}$). Let the periodic cost of increasing the machines capacity by one throughput unit be $P$, and let the analogous cost of siders be $C$. For large $m$ and $s$ (i.e., in the limit as they approach infinity), we can show that the conditions of our generalization of the newsboy model (Section 3, Theorem 1) are satisfied. If so, the optimal frequency at which we should observe a queue is $C / (P + C)$. Theoretically, it may be possible to calculate the optimal number of siders and machines analytically, but it is more useful to think about this as a system that can be managed by trial and error. Just monitor the frequencies and add or subtract siders (or machines) as necessary. This approach, perhaps in conjunction with simulation, is more practical for more complex systems. If we limit ourselves to positive investments (e.g., adding machines

or siders), the result is *balanced growth.*

A machine interference instance with high stakes (billions of dollars), is balancing the capacity of shipyards and the number of ships in a large navy. Here, shipyards are analogous to siders, and ships to machines. If repair (and refurbishing) capacity is reduced too much, repair time increases and more ships are tied up, thus reversing any savings and turning them to losses. Trietsch [22] reports examples of this sort that involved very high levels of waste due to lack of balance. These examples also demonstrate that it is important to devise balancing methods that can be implemented seamlessly in hierarchical systems, as we suggest here.

## 3. The Dual Problem

Consider the assembly coordination model (ACM) of optimizing the schedule of ordering parts for an assembly to minimize the expected completion time or an expected linear delay penalty, while accounting for the holding costs associated with early deliveries. Our primal capacity balance problem looks at throughput per time unit, while the ACM concerns processing time for a given throughput. Maximizing the former is dual to minimizing the latter, which is why we refer to the ACM as the *dual problem*. For completeness, we summarize here results for the dual problem obtained independently by Chu *et al.* [5], Kumar [9] and Ronen and Trietsch [12].

Let an assembly have $n$ parallel inputs with independent stochastic supply lead times $Y_i$ ($i$ = 1,…, $n$) with marginal cumulative distribution function (CDF) $\mathrm{F}_i(y)$ and probability density function (PDF) $f_i(y)$. Assume that once all inputs are in the assembly is instantaneous. Let $\boldsymbol{T} = \{T_i \mid i = 1,\ldots, n\}$ be a vector of the ordering times of the inputs ($T_i$ is also known as the *gate* of input $i$). Assume that $Y_i$ is *stationary*; i.e., it is not a function of $T_i$. Therefore, input $i$ arrives at time $T_i + Y_i$, which is a random variable. If the assembly has a due date, it is useful to model it as a gate, $T_0$, and set $Y_0 = 0$. The gates are our decision variables but $T_0$ may be dictated exogenously. Specifically, by setting $T_0 = 0$ we can communicate a desire to finish the assembly as early as possible while a larger $T_0$ implies that the assembly is not considered complete until $T_0$ even if all the physical inputs have arrived. That is, the completion occurs with the last arrival, at time $\max_{i=0,1,\ldots,n}\{T_i + Y_i\}$. Let $c_i \geq 0$ ($i$ = 1,…, $n$) be the economic cost (gain) of decreasing (increasing) $T_i$ by one time unit; e.g., incurring holding costs on input $i$ if it arrives too early. $c_i$ is restricted only to marginal costs that can be influenced by scheduling: as a rule, fixed charges should not be included. $c_0 \geq 0$ is the assembly tardiness penalty per time unit; i.e., the tardiness penalty is $c_0[\max_{i=0,1,\ldots,n}\{T_i + Y_i\} - T_0]^+$. Define $s = \sum_{i=0,1,\ldots,n} c_i$ : $s$ represents the assembly time-unit value. The objective function is to minimize the expected total holding cost of all $n$ inputs from their start until the assembly is complete plus the expected tardiness penalty, i.e., to minimize $Z = Z = \mathrm{E}[\sum_{i=0,1,\ldots,n} c_i (\max_{i=0,1,\ldots,n}\{T_i + Y_i\} - (T_i + Y_i))]$, where $Y_0 = 0$. It can be shown that

$$Z = \sum_{i=1}^{n} c_i(T_0 - T_i - E(Y_i)) + s\int_{T_0}^{\infty}\left[1 - \prod_{1}^{n}\mathrm{F_i}(w - T_i)\right]dw.$$

Taking partial derivatives by $T_i$,

$$\frac{\partial Z}{\partial T_i} = -c_i + s\int_{T_0}^{\infty}\mathrm{f_i}(w - T_i)\prod_{k\neq i}\mathrm{F_k}(w - T_k)\,dw.$$

Observe now that the integral in the partial derivative gives the probability that input $i$ will be last, or, the *criticality* of this input (denoted by $p_i$): $\mathrm{f}_i$ is the density of the probability

that input $i$ will arrive at time $w$ and the product of the remaining CDFs is the probability that the other inputs will have arrived before (Chu *et al.* [5]). If we use stars to denote optimal values, then we can say that for $\boldsymbol{T}^*$ this integral is $p_i^*$. For $i = 1,\ldots, n$, it follows that if $T_i^* > 0$ then $p_i^* = c_i/\boldsymbol{s}$ (to set the partial derivative to zero) and otherwise we must have $p_i^* \geq c_i/\boldsymbol{s}$: if $p_i^* < c_i/\boldsymbol{s}$ would occur at a constrained $T_i^* = 0$ gate, we could open the gate later and increase the derivative towards zero—a contradiction. For constrained gates, the true economic time value of input $i$ ($i = 1, \ldots, n$) is not $c_i$ but $v_i^* = p_i^*\boldsymbol{s} \geq c_i$. Thus, the optimal gates satisfy the following generalized newsvendor result: $p_i^* = v_i^*/\boldsymbol{s}$ (so if a gate is not constrained its optimal criticality is $c_i^*/\boldsymbol{s}$ ). The implication is that if a gate is constrained its criticality must be increased at the expense of the optimal criticality of gate 0, and not at the expense of the criticality of any other gate. Consider that the service level of the assembly's on-time performance is $1 - p_0^*$, and if no gate is constrained it should be given by $SL_0 = 1 - c_0/\boldsymbol{s}$. Thus if one or more gates are constrained, the optimal assembly service level is smaller than $1 - c_0/\boldsymbol{s}$. When no gates are constrained it is easy to see that this result is a straightforward generalization of the newsboy model for $n$ inputs, suggesting that the optimal criticality of an input should be proportional to its marginal time unit cost. In Trietsch and Quiroga [19] we demonstrated how this optimality condition can be used for an efficient simulation-based search that converges to the optimum very quickly. (Subsequently, Trietsch [21] extended this generalized newsboy optimality condition for project feeding buffers without requiring statistical independence between inputs.)

## 4. Optimal Economic Resource Balance

Henceforth, we will use the term *production* to refer to manufacturing or to providing service. For our purpose, the *product* represents an aggregate mix of individual sub-products, and the latter may include manufactures, software, services, etc. We define the *total production system* (or just *system*) as the collection of all the resources that are required to produce, sell and service the product. Emphatically, marketing and sales compose such a resource. Conceptually, we can produce demand, not just supply, so both are endogenous. But our inclusion of demand as an endogenous product, mathematically isomorphic to any other part of the total product, is in sharp contrast to most sources; e.g., Atwater and Chakravorty [1], Goldratt [6], Harrison and Van Mieghem [7], Lawrence and Buss [10], Mukherjee and Chatterjee [11] and Ronen and Spector [13] all treat demand as exogenous. In general, balance may be achieved by a (possibly negative) investment or expenditure (e.g., adding machines or hiring more operators). But it may also involve an adjustment (e.g., setting prices). For convenience, henceforth we use the term "investment" for investments, expenditures, and adjustments; and we'll interpret the term "to invest" accordingly. For example, the "investment" involved in price discounting (to increase selling capacity) is the reduced income on items that could have been sold at full price. Also, we will express all investments in amortized terms; e.g., in $ per month.

Let a total production system comprise $n$ resources, $R_i$ ($i = 1,\ldots, n$), each with a random periodic capacity, $X_i$, such that the output of the system is $\min\{X_i\}$. We note now that if there is a quota defining the maximal desired throughput we could model is as another resource, $R_0$, which is analogous to the due date, $T_0$, in the dual problem. Nonetheless, although we will occasionally use the index 0 to denote the total net throughput, regardless of whether it is subject to a quota or completely random, our default is to exclude $R_0$. Let $F_i(x)$ and $f_i(x)$ be the CDF and density function of $X_i$ with mean $\mu_i$ and standard deviation $\sigma_i$. We assume statistical independence between periods and between resources. Because periods are independent, it suffices to focus on one period only (Harrison and Van Mieghem [7]). Let $c_i$ be the marginal cost of expanding $\mu_i$ by one unit

per period. As in the dual problem, in our first model we assume that this expansion will not affect $\sigma_i$, but we will release this assumption for the second model. (In the dual problem the assumption is true when the distribution is stationary but in our primal problem it implies that we can increase capacity without increasing the variance, which is rarely true in practice. Let $\Delta_i$ be the amount by which $\mu_i$ is expanded. Because changes in $\mu_i$ are captured by $\Delta_i$, for the purpose of analysis we will treat $\mu_i$ as a constant. Accordingly, let $F_{i|\Delta}(x)$ denote the CDF of $X_i$ after increasing $\mu_i$ by $\Delta_i$. In general, $\Delta_i \geq -\mu_i$, but often either $\Delta_i \geq 0$ is required or $c_i$ is much lower for $\Delta_i < 0$; i.e., some investments are difficult or impossible to reverse. We ignore this issue for a while. Let $q_i = c_i / \Sigma c_j$ (we may omit summation limits, indices, etc. when they are implicitly clear; e.g., here, $j = 1,\ldots, n$). That is, $q_i$ is the normalized marginal cost of increasing $\mu_i$ such that $\Sigma q_i = 1$. Notice that $q_i$ is analogous to the unconstrained $p_i^*$, or $c_i / s$, of the dual problem. Similarly, let $Q_i = c_i\mu_i / \Sigma c_j\mu_j$; i.e., $Q_i$ is the relative marginal value of $R_i$, and $\Sigma Q_i = 1$. The model we obtain is

$$\text{maximize } Z = E(\min_{i=1,\ldots,n} X_i)$$

$$\text{s.t. } \Sigma_{i=1,\ldots,n} c_i \Delta_i \leq b$$

$$X_i \sim F_i(x|\Delta_i);\ i = 1,\ldots, n.$$

That is, we seek to maximize the expected throughput, defined as the minimum of $n$ inputs, subject to a budget constraint. When the system is potentially profitable, we need the budget constraint to avoid unbounded solutions. Otherwise, zero capacity is the trivial optimum. Henceforth, we assume that the system is profitable (when balanced). Therefore the budget constraint is binding. The pursuit of maximal profit subject to the budget leads to an optimal tradeoff between resource capacities. We refer to this tradeoff as *economic balance* (Lawrence and Buss [10]).

Because we address a single aggregate product, some of the variation in $X_i$ is due to periodic variation in the product composition. Hopp and Spearman [8] identify such variation as the cause of the *floating bottleneck* phenomenon (or *shifting* bottleneck). The purpose of BN- subordination is to prevent BN-shifting. In contrast, both JIT and the related CONWIP and CAPWIP (ibid, but with a system-wide cap) expect BN-shifting and make no attempt to control it. Indeed, this is the main distinction between BN-subordination and CAPWIP. (Our study of BN-subordination here does not include the elements that it shares with CAPWIP.)

***Modeling Hierarchical Systems***

A *basic* system has a single hierarchical level with $n$ resources. But realistic resources can comprise lower level components and compose higher level aggregates; e.g., manufacturing belongs to the total production system at the higher level and comprises many resources at several lower hierarchical levels. To enable hierarchical analysis, we allow partitioning the $n$ resources to $K \leq n$ non-empty sets, denoted by $\boldsymbol{R}_k$; $k = 1,\ldots, K$. Each such set yields a *composite* resource. We use bold letters to denote values or functions of composite resources; e.g., $\boldsymbol{X}_1$ is the minimum of the resources in $\boldsymbol{R}_1$. Composites of composites are also allowed. By the stochastic independence assumption, $\mathbf{F}_k(x) = 1 - \Pi(1 - F_i)$ $(\forall i \in \boldsymbol{R}_k)$, and

$$\boldsymbol{\mu}_k = \int_0^\infty \left(1 - \mathbf{F}_k(x)\right) dx = \int_0^\infty \prod_{i \in R_k} \left(1 - F_i(x)\right) dx.$$

When we invest in a composite resource, there are infinitely many ways to allocate the

investment to the parts. We refer to any rule governing such allocations as a *set-allocation rule.* Specifically, the *simple* set-allocation rule is $\Delta_i = \Delta_j$ ($\forall i, j \in \boldsymbol{R}_k$). It implies $\boldsymbol{q}_k = \Sigma q_i$ ($\forall i \in \boldsymbol{R}_k$). The *proportional* set-allocation rule is $\Delta_i / \mu_i = \Delta_j / \mu_j$ and $\boldsymbol{Q}_k = \Sigma Q_i$ ($\forall i, j \in \boldsymbol{R}_k$). In practice, we will often invest in the single most attractive resource, which is yet another set-allocation rule. For any set-allocation rule, given either $\boldsymbol{q}_k$ or $\boldsymbol{Q}_k$ we can compute $\boldsymbol{Q}_k$ or $\boldsymbol{q}_k$, respectively. Given a set-allocation rule, for all intents and purposes each composite resource behaves as a single resource. Accordingly, we will consider $R_i$ as either a single resource or a composite resource, and we will use the notation $\boldsymbol{R}_k$ only where it is important to show the distinction. For convenience we refer to the whole set of resources by $\boldsymbol{R}_0$. For example, the utilization of $R_i$, denoted by $U_i$, is $\boldsymbol{\mu}_0 / \mu_i$ (since $Z = \boldsymbol{\mu}_0$). Finally, we define the $R_i$-*partition* by setting $\boldsymbol{R}_1 = R_i$ and $\boldsymbol{R}_2 = \boldsymbol{R}_0 - R_i$, where some set-allocation rule applies to $\boldsymbol{R}_2$. We will use the $R_i$-partition to generalize $n = 2$ results to any $n$.

***Modeling Stochastic Expansion***

***Lemma 1:*** There exists an expansion function $g_i(x, \Delta)$ (more precisely $g_i(x, \Delta_i)$) such that,

1. $g_i(x,0) = x.$
2. $F_{i|\Delta}(x) = F_{i|0}\left(g_i(x,\Delta)\right) = F_i\left(g_i(x,\Delta)\right).$
3. $\int_0^\infty \left(1 - F_{i|\Delta}(x)\right)dx = \int_0^\infty \left(1 - F_i\left(g_i(x,\Delta)\right)\right)dx = \int_0^\infty \left(1 - F_i(x)\right)dx + \Delta.$

(All proofs are given in the Appendix.)

By specifying $g_i(x, \Delta)$ we effectively prescribe the exact way in which a distribution is changed during capacity expansions. If $F_{i\,|\,\Delta}(x) \geq F_{i\,|\,\Delta+\varepsilon}(x)$ ($\forall\ \varepsilon > 0$, $\Delta > -\mu_i$), the expansion is *proper*. By definition, after a positive (negative) proper expansion, capacity increases (decreases) in the strict stochastic sense. A function $g_i(x, \Delta)$ is *proper* if, for any admissible $x$, $g_i(x, \Delta) \leq x$ for any $\Delta_i > 0$ and $g_i(x, \Delta) \geq x$ for any $-\mu_i < \Delta_i < 0$. Proper expansion implies proper $g_i(x, \Delta)$ and vice versa. If $g_i(x, \Delta)$ is proper and differentiable, then $\partial g_i / \partial \Delta \leq 0$ ($\forall \Delta > -\mu_i$). Let $g_i^\kappa(x, \Delta)$, or simply $g^\kappa$, denote the following proper expansion function,

$$g_i^\kappa(x,\Delta_i) = \kappa_i\, x\, \frac{\mu_i}{\mu_i + \Delta_i} + (1-\kappa_i)(x - \Delta_i);\ 0 \leq \kappa_i \leq 1\ ;\ \Delta_i > -\mu_i.$$

Taking $\partial g_i / \partial \Delta_i$ and evaluating it for $\Delta_i = 0$,

$$\frac{\partial g_i^\kappa(x,\Delta_i)}{\partial \Delta_i} = -\kappa_i\, x\, \frac{\mu_i}{(\mu_i + \Delta_i)^2} - (1-\kappa_i) \Rightarrow \left.\frac{\partial g_i^\kappa(x,\Delta_i)}{\partial \Delta_i}\right|_{\Delta_i = 0} = -\kappa_i \frac{x}{\mu_i} - (1-\kappa_i). \qquad (1)$$

$g^0$ yields *simple* expansion, constituting a lateral shift of the distribution by a constant. $g^1$ yields *proportional* expansion, which involves scaling the argument. Proportional expansion satisfies Condition 3 of Lemma 1 because the scaling applies to the mean too. Likewise, $g^\kappa$ conforms to the condition for any $0 \leq \kappa_i \leq 1$. Let $CV_i = \sigma_i / \mu_i$ be the initial coefficient of variation, and suppose that after expansion $\sigma_i$ is increased by $\kappa_i CV_i \Delta_i$. $g_i^\kappa$ achieves this. The limits $0 \leq \kappa_i \leq 1$ are justified as follows. $\kappa_i < 0$ would imply a decrease in the standard deviation during expansion, which is rarely likely. $\kappa_i = 1$ implies constant $CV_i$,

and it applies if we essentially scale every important ingredient of the capacity. If, however, we increase capacity by adding incremental capacity units in parallel to existing ones, then $CV_i$ should decrease. Considering the practical difficulties involved in estimating $F_i(x)$ even without attempting to estimate $g_i(x, \Delta)$, the use of $g^\kappa$ is an appropriate approximation. Although limiting $g_i(x, \Delta)$ to $g^\kappa$ is a strong restriction, practically all sources just use $g^0$!

***Criticality and Optimal Permission to Fail (PTF)***

Analogously to the dual case, we define $p_i = \Pr\{X_i \le X_j;\ \forall j \ne i\}$—the probability that $X_i$ will be the minimum—as the *criticality* of $R_i$. With continuous distributions, $\Sigma p_i = 1$ and $\boldsymbol{p}_k = \Sigma p_i$ ($\forall i \in \boldsymbol{R}_k$). Below, we use "*" to denote optimal values and "a" superscripts for approximate values; e.g., $T^* \approx T^a$. By definition, $p_i^*$ ($p_i^a$) is the optimal (approximate) PTF that should be allocated to $R_i$.

**Lemma 2:** For $n = 2$, a necessary condition for optimality is given by

$$\frac{p_1^*}{p_2^*} = \frac{c_1 \mathrm{E}\left(\frac{\partial g_2(x,\Delta_2)}{\partial \Delta_2} \mid X_2 \le X_1\right)}{c_2 \mathrm{E}\left(\frac{\partial g_1(x,\Delta_1)}{\partial \Delta_1} \mid X_1 \le X_2\right)} \Leftrightarrow \frac{q_i}{p_i^*} \propto \mathrm{E}\left(\frac{\partial g_i(x,\Delta_i)}{\partial \Delta_i} \mid X_i \le X_{3-i}\right)\ ;\ i = 1, 2\,.$$

**Lemma 3:** For any $n$, if the expansion function is simple (proportional) and the simple (proportional) set-allocation rule is used for all $j \ne i$, then

$$\frac{q_i}{p_i^*} \propto \mathrm{E}\left(\frac{\partial g_i(x,\Delta_i)}{\partial \Delta_i} \mid X_i \le X_j\ ;\ \forall j\right)\ ;\ \forall i\,.$$

By Lemma 3 and Equation 1, for $g^\kappa$ ($\kappa = 0, 1$) with $\mu_i = \mu_i^*$, $q_i / p_i^*$ is proportional to $\kappa_i \mathrm{E}(X_i \mid X_i \le X_j; \forall j) / \mu_i + (1 - \kappa_i)$. This leads to two simple optimal expressions as follows.

**Theorem 1:** When $\kappa_i = 0$ ($\forall i$), with general $F_i(x)$, $p_i^* = q_i$.

**Theorem 2:** Let the resources have exponential distributions with $\kappa_i = 1$ ($\forall i$), then $p_i^* = Q_i$.

Theorem 2 implies allocating the budget to resources in proportion to $\sqrt{q_i}$. This yields $\mu_i^* \propto 1/\sqrt{q_i}$ and $U_i^* \propto \sqrt{q_i}$. The average utilization is $1/n$ (so much for naive balance!). Although the exponential distribution is a special case, we can also conclude that it is impossible to achieve balanced utilization with arbitrarily prescribed $U_i$.

**Theorem 3:** When $g^\kappa$ applies, the objective function, $Z$, is concave.

***Heuristic Approximations for $\boldsymbol{\kappa \ne 0}$***

When $\kappa_i = 0$ ($\forall i$), by Theorem 1, $p_i^* = q_i$ (as in the dual case). Otherwise, either to implement exact models or to obtain accurate simulation results, we need exact information on all distributions and expansion functions. The availability of such information in practice is not likely, however. To see this, consider that every period we can find the true capacity of one resource only, the minimal one. All we know about the remainder is that they could have delivered more, potentially. That is, the data is censored (unlike the dual problem case). As for estimating $g_i(x, \Delta)$, by limiting ourselves to $g^\kappa$ we can simplify the task considerably, but it is still difficult in practice. Therefore, (i) we need heuristics that do not require so much information, and (ii) robustness to $g_i(x, \Delta)$ is important.

With this in mind, when we know that $\kappa_i \approx 0$ ($\forall i$), $p_i^a = q_i$ is suitable. When we know that $\kappa_i \approx 1$ ($\forall i$), we recommend $p_i^a = Q_i$. More generally, we could propose the *mixed heuristic,* $p_i^a \propto \kappa_i Q_i + (1 - \kappa_i)q_i$, where the proportionality becomes an equality if $\kappa_i = \kappa_j$ ($\forall i, j$) (or if we just select a single value $\kappa$ for all resources). Nonetheless, for simplicity we propose the use of $p_i^a = Q_i$ as the default, to be replaced by $p_i^a = q_i$ only when we know that all $\kappa_i$ are very low (say 0.25 or below) for all $i$. Numerical experiments suggest that $p_i^a = Q_i$ is a robust choice when $\kappa_i$ is not known. It is the best approximation we have for $\kappa_i = 1$ and performs well for $\kappa_i = 0$, while $p_i^a = q_i$ is not robust for $\kappa_i > 0$ (see Section 6). Furthermore, consider that small $\kappa_i$ is not likely in practice except in large systems. However, if we have many resources—a case likely in relatively large systems—then the gradual addition of more resources to increase capacity leads to low $\kappa_i$. For example, suppose we have $m$ identical machines in parallel and we add one more in parallel, the total capacity will increase by $1 / m$ and when $m$ is large $\kappa_i \approx 1 / 2m$. (This is why, when $s$ and $m$ are large, Theorem 1 applies to the machine interference example in Section 2.)

**Theorem 4:** For $\kappa_i = 0$, $Q_i \rightarrow p_i^*$ as $b \rightarrow \infty$ (i.e., as the budget grows, $p_i^a = Q_i$ is asymptotically optimal for $\kappa_i = 0$).

If we stipulate that $p_i^a = Q_i$ is a good heuristic for $\kappa_i = 1$, it can only improve with size for $\kappa_i < 1$ (as it does for $\kappa_i = 0$ by Theorem 4). Thus, the use of $p_i^a = Q_i$ (rather than $q_i$ or a mixture of $Q_i$ and $q_i$) is supported. Furthermore, typically, the model will be applied within a subsystem, say $\boldsymbol{R}_1$, and it is important to assign the correct PTF to the subsystem as a whole (at least approximately). The use of $\boldsymbol{Q}_1$ is feasible: if all else fails, we can simply use estimates of the value of the subsystem and of $\boldsymbol{R}_2$ (the rest of the system) in lieu of $\boldsymbol{c}_1\boldsymbol{\mu}_1$ and $\boldsymbol{c}_2\boldsymbol{\mu}_2$. In contrast, the information required for estimating $\boldsymbol{q}_2$ is rarely available. Within the subsystem, however, we have the option to use any $p_i^a$ (e.g., $p_i^a \propto q_i$) and employ any set-allocation rule (e.g., the simple set-allocation rule). The following theorem does not require the conditions of Theorem 3 (or it would be redundant due to concavity).

**Theorem 5:** For any vector $\{p_i^a\}$ and budget there is a unique vector $\{\mu_i^a\}$ that satisfies it.

In summary, our heuristics require monitoring adherence to $\{p_i^a\}$. We consider adjusting $R_i$ only upon evidence that $p_i$ does not match $p_i^a$. By Theorem 5, they lead to unique results regardless of the exact sequence of expansions we use. Our main heuristic, $p_i^a = Q_i$ is robust for any $g_i$. Since $\{p_i\}$ is the only required input, both heuristics are parsimonious.

## 5. The Economic Return on Balance

For n=2, assume that $\Delta_1$ is measured relative to $\mu_1^*$ and $\Delta_2$ relative to $\mu_2^*$. Therefore, for $\Delta_1=\Delta_2=0$ the budget constraint is satisfied as an equality. To maintain the budget constraint intact we must have $\Delta_2 = -\Delta_1 c_1/c_2 = -\Delta_1 q_1/q_2$. Substituting this for $\Delta_2$, the objective function is,

$$Z(\Delta_1) = \int_0^{\infty} \left(1 - F_1\left(g_1(x, \Delta_1)\right)\right)\left(1 - F_2\left(g_2\left(x, \frac{-q_1}{q_2}\Delta_1\right)\right)\right) dx.$$

and

$$\frac{\mathrm{d}\,Z(\Delta_1)}{\mathrm{d}\,\Delta_1} = -\int_0^{\infty} \frac{\partial g_1}{\partial \Delta_1} f_1\left(g_1(x,\Delta_1)\right)\left(1 - F_2\left(g_2\left(x, \frac{-q_1}{q_2}\Delta_1\right)\right)\right) dx\,.$$

$$-\frac{q_1}{q_2}\int_0^{\infty} \frac{\partial g_2}{\partial \Delta_2} f_2\left(g_2\left(x, \frac{-q_1}{q_2}\Delta_1\right)\right)\left(1 - F_1\left(g_1(x,\Delta_1)\right)\right) dx.$$

Using $R_i$-partitioning, $\Delta_1$ may be replaced by any other $\Delta_i$. Until further notice, assume simple expansion, where $\partial g_1^0/\partial\Delta_1 = \partial g_2^0 / \partial\Delta_2 = -1$. Therefore, dZ / d$\Delta_1 = -p_1 + p_2q_1/q_2$ (and—since n = 2 — $p_2 = 1-p_1$ and $q_2 = 1-q_1$). Away from the optimal solution, this yields $(q_1 - p_1)(1+q_1/q_2) = (p_1^*-p_1)/p_2^*$.

We now consider two special cases. First, in *Case* 1, $R_1$ is deterministic while $F_2(x)$ is general. Case 1 is very similar to the classical newsboy model—but to change the deterministic capacity we must also shift the stochastic one (to avoid violating the budget constraint). *Case* 2 involves two normal variables. In both cases, except for a simple lateral shift, the distribution of the *difference variable* $X_2 - X_1$ is invariant to $\Delta_1$. In Case 2, $X_2 - X_1$ is a normal variable with $\mu = \mu_2 - \mu_1$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. If we increase (decrease) $\mu_1$ by $\Delta_i$ (while maintaining the validity of the budget constraint) the mean of this distribution decreases (increases) by $\Delta_1(1+q_1/q_2)$, but the variance, and therefore the shape, does not change. The same result applies for Case 1, where the CDF of the difference variable is identical to the CDF of $R_2$, except for a lateral shift. Let $h(\Delta_1)$ denote the density function of the difference variable, in both cases. We then obtain $dZ/d\Delta_1 = \Delta_1(1+q_1/q_2)h(\Delta_1) = \Delta_1 h(\Delta_1) / p_2^*$ (because this expression measures the change of $p_1$ as a function of $\Delta_1$). When we take the integral of this derivative from 0 to $\Delta_1$ we obtain the loss associated with missing the optimum. If h() is fairly constant or monotone between h(0) and $h(\Delta_1)$, the loss for small $\Delta_1$ is given approximately by $((p_1^*-p_1)(1+q_1/q_2))^2/2h(2\Delta_1/3) \approx (\Delta_1/p_2^*)^2h(2\Delta_1/3)/2$ (because $p_1^*-p_1 \approx \Delta_1 h()$).

In the general $g^0$ case, there must exist a function $h(\Delta_1)$ that measures the rate of change of $p_1$, and the loss will still be given approximately by $(\Delta_1/p_2^*)^2h(2\Delta_1/3)/2$. But the density of the difference variable will only be an approximation of h(). Likewise, if we generalize this expression to $g^1$, by a similar analysis, the loss is proportional to $((p_1^*-p_1)(1+Q_1/Q_2))^2/2h(2\Delta_1/3)$ for a similar positive function h(). Arthanari and Trietsch [2] gives a graphical representation of the economical gain as calculated above. This can help managers focus on the best opportunities.

***Optimal Growth***

Alternatively, suppose the system starts in-or out of balance, and we wish to invest a small positive incremental amount, $b$, in *one* of the resources, say $R_1$. This model is natural when $\Delta_i \geq 0$ applies and when it is not feasible or desirable to divide the budget to many resources. Then,

$$v_1 \equiv \frac{\partial Z(\Delta_1,\Delta_2)}{\partial \Delta_1} = -\int_0^{\infty} \frac{\partial g_1}{\partial \Delta_1} f_1\left(g_1(x,\Delta_1)\right)\left(1 - F_2\left(g_2(x,\Delta_2)\right)\right) dx\,, \qquad (2)$$

where $Z$ is now a function of both $\Delta_1$ and $\Delta_2$ because we do not trade them off directly. We are primarily interested in the immediate net benefit by investing $b$ in $R_1$, under the assumption that the increment is very small. The net marginal benefit is given by

$$\frac{v_1}{c_1}-1=\frac{1}{c_1}\frac{\partial Z}{\partial \Delta_1}-1=-1-\frac{1}{c_1}\int_0^{\infty}\frac{\partial g_1}{\partial \Delta_1}f_1\left(g_1(x,\Delta_1)\right)\left(1-F_2(x)\right)dx. \quad (3)$$

By $R_i$-partitioning, Equation 1 actually defines a value $v_i$ for any $i$. $v_i$ is the marginal benefit associated with increasing $R_i$. The best single resource is selected by maximizing $v_i/c_i$, as per Equation 2 (when a balanced system is profitable, $\max\{v_i/c_i\} \geq 1$, so this investment will be attractive). If we do this in a balanced system, with really small increments, then in terms of first order effects it should not matter which resource we invest in. But otherwise the resource whose net marginal return is maximal will receive the investment, thus leading to better balance. A succession of small incremental investments, each in the best single available resource, will tend to keep the system approximately balanced while it grows.

***Optimal Adjustment***

Suppose we are still interested in the effects of a small change in a single resource, but we also allow a negative budget, i.e., we allow disinvesting. In general, the marginal gain by reducing capacity (i.e., the salvage value), denoted by $c_{-i}$, should satisfy $c_{-i} \leq c_i$ (or we could make unbounded profits by buying capacity cheap and selling it dear). $v_i/c_i - 1 \leq 0$ is necessary for disinvestment, but we also require $c_{-i}/v_i - 1 \geq 0$. Because $c_{-i} \leq c_i$, the combination $v_i/c_i - 1 \leq 0$ and $c_{-i}/v_i - 1 \leq 0$ is possible, in which case $R_i$ is *stable:* its criticality is not high enough to justify investment and not low enough to justify disinvestment; Harrison and Van Mieghem [7] identified an indifference region that speaks to the same issue. Therefore, maintaining balance on an ongoing basis does not require adjusting all resources periodically.

## 6. Negative Investment and Risk

Common financial practice imposes a minimal expected return on investment, i.e., a hurdle, to justify risk-taking. Conceptually, this is just a buffer, but here we assume the hurdles are determined exogenously. When we take into account the ability to sell capacity, if $c_{-i} = c_i$, there is little or no risk: we can reverse the investment at zero cost. But if $c_{-i} < c_i$, risk should be accounted for. Since different investments entail different risks and different hurdles, we should modify each $c_i$ separately. For example, if for a positive investment in $R_i$ the hurdle is 10%, we multiply the true, or "raw," $c_i$ by 1.1 before using it. $v_i$ is measured in expected throughput units, so $v_i/c_i \geq 1$ (where $c_i$ is modified) *iff* the hurdle is satisfied. When disinvesting, there are two risks: (1) the return may be less than expected, and (2) we may regret the decision later. So we have to reduce $c_{-i}$ relative to its raw value (in contrast to increasing $c_i$). Again, $c_{-i}/v_i \geq 1$ *iff* the hurdle is satisfied. Thus risk increases stability.

Now consider the effect of a budget constraint in a profitable system. Profitability implies meeting or beating all hurdles. We assume binding budget constraints, so the Lagrange multiplier associated with the budget should exceed 1 (money is "expensive"). To achieve appropriate balance this multiplier inflates all $c_i$ and $c_{-i}$ values in such a manner that the budget becomes binding. Here, in contrast to the former result, $c_{-i}$ should be inflated too, because it is a source of funds and the binding budget constraint increases their value (we can disinvest in one resource and invest the proceeds in another). Henceforth we assume that $\{c_i\}$ and $\{c_{-i}\}$ are first modified to incorporate appropriate hurdles and then adjusted by the multiplier.

***Adapting the Heuristics for $c_{-i} < c_i$***

Let $\{c_i^a\}$ be a set of expansion costs such that $c_{-i} \leq c_i^a \leq c_i$ $(\forall i)$, and define

$q_i^a = c_i^a / \sum c_j^a$ and $Q_i^a = c_i^a \mu_i / \sum c_j^a \mu_j$. Suppose we partition the resources to three sets, $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ and $\boldsymbol{R}_3$ such that for all $R_i \in \boldsymbol{R}_1$, $p_i$ [as measured] $< p_i^a$; for all $R_i \in \boldsymbol{R}_3$, $p_i > p_i^a$; and $R_i \in \boldsymbol{R}_2$ *iff* $R_i$ is stable. By construction, $\boldsymbol{R}_1$ ($\boldsymbol{R}_3$) calls for negative (positive) investments, and thus should be associated with $c_i^a = c_{-i} (c_i^a = c_i)$. $\boldsymbol{R}_2$ should have a $c_i^a$ value between $c_{-i}$ and $c_i$, such that $p_i^a = p_i$. We can now replace our main (secondary) heuristic, $p_i^a = Q_i(q_i)$ by $p_i^a = Q_i^a(q_i^a)$. The following linear program yields $\{c_i^a\}$ and identifies the partition. $c_i^a$, $x_i^+$ and $x_i^-$ are the decision variables. Constraint 3 is for $p_i^a = Q_i^a$ and constraint 3a is for $p_i^a = q_i^a$. (We can merge the two alternative constraints by a convex combination as per $\kappa_i$, to make possible using the mixed heuristic, $p_i^a \propto \kappa_i Q_i + (1 - \kappa_i) q_i$.)

$$\textit{mininize} \quad \sum_{i=1}^{n} \left( x_i^+ + x_i^- \right)$$

$$\textit{subject to} \quad (\forall i):$$

$$\textit{either} \quad x_i^+ \geq c_i^a \mu_i - p_i \sum_{j=1}^{n} c_j^a \mu_j \geq -x_i^- \quad [3]; \quad \textit{or} \quad x_i^+ \geq c_i^a - p_i \sum_{j=1}^{n} c_j^a \geq -x_i^- \quad [3a]$$

$$c_{-i} \leq c_i^a \leq c_i; \quad x_i^+, x_i^- \geq 0,$$

to interpret the output, $x_i^+ > 0$ ($x_i^- > 0$) in $\boldsymbol{R}_1$ ($R_3$); $x_i^+ = x_i^- = 0$ in $\boldsymbol{R}_2$.

***The Balance Coefficient***

For $p_i^a = Q_i^a$ ($q_i^a$), dividing the objective function by $\Sigma c_j^a \mu_j$ ($\Sigma c_j^a$) yields a *balance coefficient* ranging from 0 to 1, where 0 indicates perfect compliance with the prescribed $p_i^a$. Such a coefficient may help investors compare systems in terms of the quality of their balance and provide an indication of the relative size of balancing opportunities.

***Optimal Balance with a Fixed-Capacity Resource***

A truly rigid resource, say $R_i$, can be identified by possessing $c_i = \infty$ and $c_{-i} = 0$ (at least in the short term). Informally—because $c_i = \infty$—if $p_i$ is high then $R_i$ is likely to be nominated as the BN. Atwater and Chakravorty [1] posed a research question: given such a BN, how much capacity should we provide on the other resources? But there is no conceptual difference between such a BN and any other resource, say $R_j$, for which $c_{-j} \leq v_j^* \leq c_j$. That is, the resource is stable regardless of its criticality. As a direct consequence, unless all other $c_i$ are 0, its optimal utilization must be below 100%. Thus, our analysis answers Atwater and Chakravorty [1]'s research question. The linear program will place such resources in $\boldsymbol{R}_2$, so the heuristics can resolve it too.

When all resources are amenable to expansion the term "bottleneck" has no clear meaning. We agree with Mukherjee and Chatterjee [11] that a resource with positive shadow price limits throughput and is thus conceptually a bottleneck, but the term bottleneck also has a connotation of "the single most binding constraint" and in this sense there is no single resource that fits the bill all the time. We can call the current minimal resource "the [shifting] bottleneck," but we should certainly not consider increasing the long-term criticality of this resource (at the expense of others) more than optimal balance calls for. Our numerical results support this conclusion as well, and indeed Atwater and Chakravorty [1] also found numerical evidence that some idleness of the "bottleneck" is beneficial. Thus, the case against BN-subordination includes both theoretical and empirical evidence. The bad news is that not *all* incorrect models are useful, and this particular one is

simply too simple. The good news is that it is already well understood that BN-subordination is flawed—e.g., see Hopp and Spearman [8], Spearman [16] and Trietsch [20]. Nonetheless, to our knowledge, ours is the first mathematical demonstration that BN-subordination is always wrong. Perhaps the proof was overlooked because usually BN-subordination is expressed in utilization terms, which masks the conflict with optimal criticality.

***A Note about Stochastically Dependent Inputs***

Trietsch [21] considered the dual problem embedded within a project network and implicitly assumed that $g_i^0$ prevails. For that case he demonstrated that there is no need for the stochastic independence assumption. Whereas the question whether stochastic independence is necessary for our results with respect to $g_i^1$, in the $g_i^0$ case the proof offered by Trietsch [21] can be used to show that our results apply even when the relevant distributions are dependent.

## 7. Numerical Experiments

All our numerical experiments are for $n = 2$ [composite] resources (e.g., $\boldsymbol{R}_1 = R_i$ and $\boldsymbol{R}_2 = \boldsymbol{R}_0 - R_i$), with $\boldsymbol{q}_1$ ranging from 0.01 to 0.99, and with a budget of 100. Thus, it is always possible to balance utilization by setting $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 100$, and $Z$ is $\boldsymbol{U}_0$ in percents. We compare $\boldsymbol{p}_i^a = \boldsymbol{Q}_i$ and $\boldsymbol{p}_i^a = \boldsymbol{q}_i$ against $\boldsymbol{p}_i^*$ (in terms of $Z$). We also study the performance of simple balance and of BN-subordination. Initially, several $g^1$ examples were run on Excel®, using normal and uniform random variables. Another $g^1$ example was run for the dual problem (minimizing the expected completion time of an assembly plus holding costs) with exponential random variables (Theorem 2 does not hold for the dual problem). These distributions were selected because it is possible to calculate the minimum of two of them by Excel® (without cumbersome numerical integration). The results indicate that for $\boldsymbol{c}_i = 0.01$, $\boldsymbol{Q}_i / \boldsymbol{p}_i^*$ was typically between 0.8 and 1.25, and invariably closer to 1 for larger $\boldsymbol{c}_i$. In some cases, however (e.g., where the distributions involved were very dissimilar), results farther from 1 were obtained. Nonetheless, in terms of the objective function, the relative error was usually well below half a percent, and often it was completely negligible (e.g., 0.03%). The worst case was obtained in the dual problem, where the loss due to the approximation is up to 2.3%. Furthermore, in all our runs, the relative error achieved a maximum for $\boldsymbol{q}_1$ within the range 0.01 to 0.99. That is, although the discrepancy between $\boldsymbol{Q}_i$ and $\boldsymbol{p}_i^*$ is maximized for the lowest $\boldsymbol{c}_1$, the importance of the exact specification of $\boldsymbol{\mu}_1$ is not high there, so the relative error was maximized within the range $0.01 < \boldsymbol{c}_i < 0.99$.

Following these exploratory experiments, additional tests were run on Mathematica® (which supports highly accurate numerical integration) to test the quality of the approximations. These runs included comparisons between the two heuristics for both $g^0$ and $g^1$ (and other $g^\kappa$ values not reported). We compared the performance of the heuristics with the "wrong" assumption about $g$ (and while doing it we obtained numerical corroboration for Theorem 1). For these experiments we selected the beta distribution to represent arbitrary capacity distributions. The beta distribution also makes possible the approximate representation of the minimum of several other beta random variables, preserving the correct minimum, maximum, mean and variance. The traditional use of the beta distribution in PERT may be criticized for being optimistic, because it assumes a maximum that cannot be exceeded while some project activities do not seem to respect any such limit. The same reservation holds with respect to the dual problem, which is really a project scheduling case. In our present context, however, such a maximum is not optimistic,

because we are trying to maximize throughput. As for the minimum, we can use 0 if we want absolute confidence. For $X_i$, assume we can estimate the minimum, $a_i$ (with a default of 0) and the maximum, $b_i$ (subject to $b_i > \mu_i > a_i$). Then it is usually possible to find $\alpha_i$ and $\beta_i$ such that a beta distribution with the parameters $a_i$, $b_i$, $\alpha_i$ and $\beta_i$ will have the correct mean $\mu_i$ and standard deviation $\sigma_i$.

Figures 1 to 3 show some results (note that the $x$-axis is symmetric with respect to 0.5 but not linear). Figure 1 is a typical case where the distributions involved are similar to each other, each with a mode strictly within the range, but one of them is more likely to represent the minimum of several others. We compare the error of using $p_i^a = Q_i$ for $g^0$ (where $q_i$ is optimal) to the error of using $p_i^a = q_i$ for $g^1$ (where $Q_i$ is the best approximation we know). Clearly, $p_i^a = Q_i$ is much more robust. The error associated with it is 0.026% of $Z^*$, as compared to 1.85% for the other. The figure also demonstrates that $Q_i$ is better than $q_i$ for $g^1$, with a maximal error of 0.26%. We see that $g^1$ is less amenable to solution by any heuristic, but $Q_i$ is clearly better.

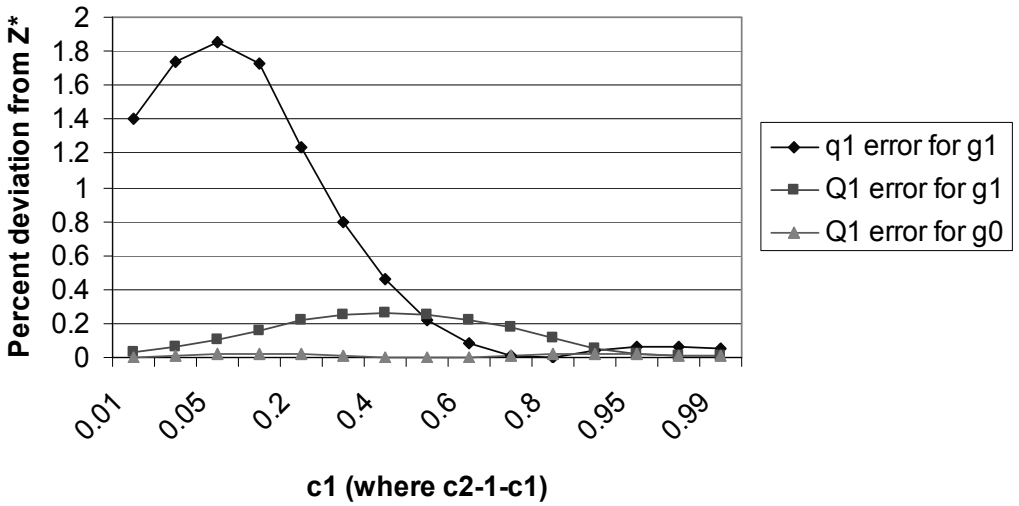


Figure 1. X1~beta(2.5, 1.8), X2~beta(6, 1.2).

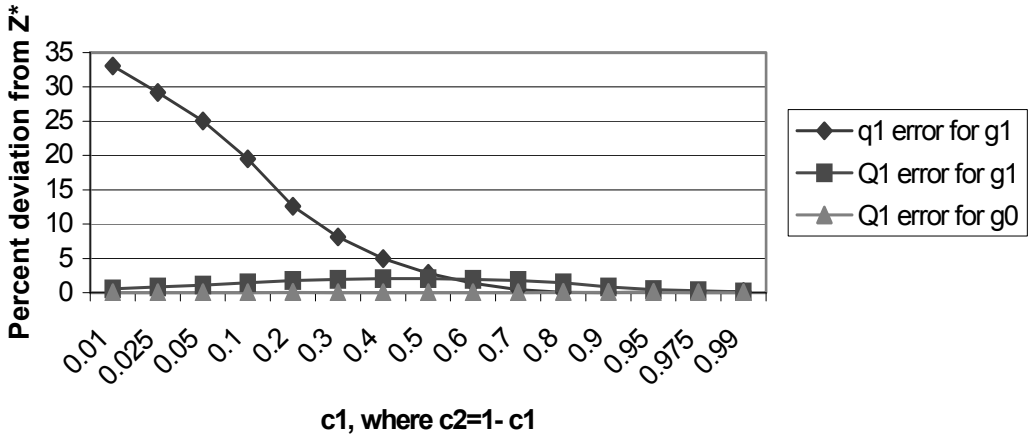


Figure 2. X1~beta(1, 2), X2~beta(4.5, 1.5).

Figure 2 compares the performance of the two heuristics for dissimilar distributions: one is left triangular (i.e., beta(1,2)), and the other is similar to the former case. Here, the maximal error of $p_i^a = Q_i$ for $g^0$ was a negligible 0.044%, but both heuristics performed less spectacularly for $g^1$—actually this was the worst case in the series, perhaps as a result of the distributions not being similar to each other. Nonetheless, $p_i^a = Q_i$ was more robust than $p_i^a = q_i$, again, with a maximal error of just above 2%, as compared to 33%. 2% may look like a very large error, but consider that it would likely be smaller for $\kappa < 1$ (based on the excellent $\kappa = 0$ performance), not to mention that there is no known viable alternative.

Figure 3, which is based on a $g^1$ case with normal variables, compares the performance of our heuristic to BN-subordination and to simple balance. BN-subordination requires 100% BN-criticality, but this is impossible for the normal distribution. We use the notation BN-PTF to denote the PTF that may be achieved instead of 100%, for example, BN-0.999 means that other resources are critical with a frequency of 1 / 1000. The graph includes results for BN-0.999, BN-0.99, and BN-0.9. Furthermore, simple balance (PTF = 0.5) also involves balanced utilization here ( $\mu_1 = \mu_2$ ), so we refer to it as *regular balance*. But simple balance can also be interpreted as BN-0.5 (although the essence of BN-subordination is that PTF should be much higher). To continue, the heuristic result is so close to the optimum that it completely masks it! Simple balance achieves a constant Z which is optimal for $c_1 = 0.5$. Indeed, it is tangential to the optimal curve there. Similarly, for high $c_1$ (at the top right of the figure), the various BN-PTF graphs are tangential to the optimum (although it is difficult to see this due to the scale). Including balance as BN-0.5, we obtain a clear picture: BN-subordination is only better than regular balance for very high $c_{BN}$ values, where it is approximately correct by our main heuristic. To the extent BN-subordination is supposed to improve upon "regular" balance, the remedy is usually worse than the disease! And as PTF grows, BN-PTF is more damaging. In the limit, when BN-100% is achieved, the damage is 100% loss of investment. Our heuristic, in contrast, is the clear winner everywhere in the range.

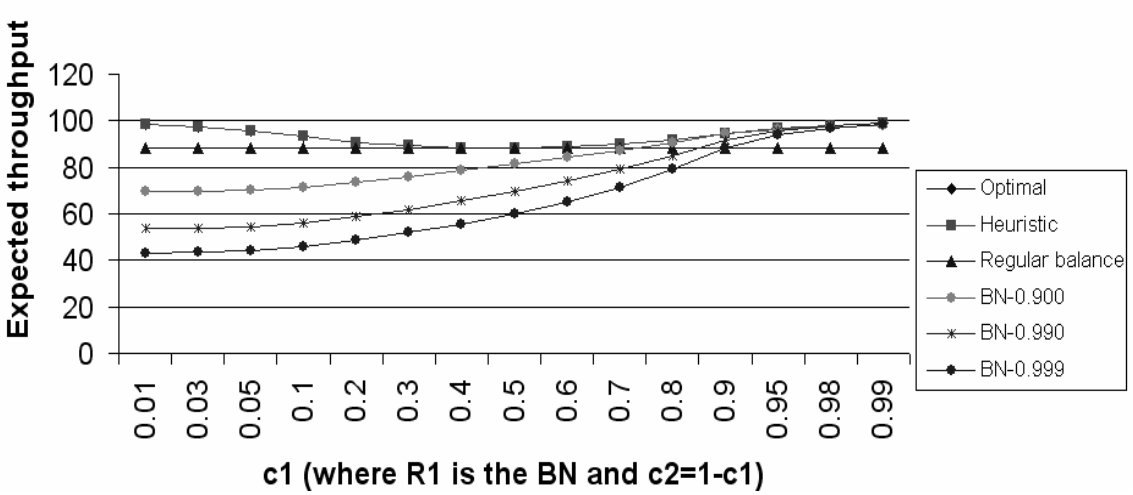


Figure 3. Comparing the heuristic with BN-subordination and with regular balance.

## 8. An Illustration

To illustrate the use of our approach, we constructed an example involving the design of a new restaurant. Balancing problems always require iterations so we must make some assumptions to derive an initial solution, and then iterate to converge to the optimal

solution. In this example we start with utilization balance as the initial solution; i.e., we set the initial capacity of all parts to the same value (150). We assume all distributions are lognormal and as such they are fully defined by their mean and coefficient of variation. We assume the coefficient of variation does not change with capacity; i.e., $g_i^1$ prevails. We aim to maximize the return on the total capitalized value by optimizing the capacity in diners per day per money unit (MU). There is a budget of 975 on the total capitalized value.

1. Without loss of generality assume that the land value is 1 MU per area unit and initially assume the use of 100 area units. Assume further that by regulation it is required to landscape 20% of the land area (at a cost that is included in the land value).

2. A building covering 20% of the ground (well below an upper bound of 25% imposed by the municipality) can serve 150 diners per day with a coefficient of variation of 10%. The randomness is due to the different sizes of parties (e.g., parties of 3 may take tables for 4) and differences in the time they take. The structure can be built on a fixed cost contract for 287 MU (i.e., construction cost is 14.35MU per area unit).

3. Together with landscaping a structure of this size leaves 60 area units to be paved for parking, which also suffices on average for 150 diners, with a coefficient of variation of 15% (this variation is due to the fact that different parties use different number of cars per diner for different durations). The cost of paving this area is 3 MU (or 0.05 MU per area unit).

***Solution:*** Considering the landscaping requirement each usable land area unit costs 1 / 0.8 = 1.25 MU. Because the constraint on the building size is not binding we can say that each area unit of the structure costs $14.35+1.25=15.6$. 20 such units are sufficient for 150 diners so this translates to $15.6\times 20/150=2.08$ per capacity unit. A similar calculation for the parking capacity cost yields $(0.05+1.25)\times 60/150=0.52$ per capacity unit. For the balanced utilization case, this implies that the relative criticality of the building should be $2.08/(2.08+0.52)=80\%$ and that of the parking 20%. This relative criticality value is an example of the hierarchical approach because there may be other critical inputs that we ignored so far, such as demand, ingredients, workers etc. Assume now that all these other inputs are incorporated into a single additional input with a capacity of 150 on average and coefficient of variation 25%. Let the target criticality for the structure as a whole be 40% (i.e., 8% for parking and 32% for the building). This is equivalent to stating that the capitalized unit capacity price of the remainder of the system is 3.9 (because 3.9 / (3.9 + 2.08 + 0.52) = 0.6). At this stage we can test by simulation whether the current plan is balanced. To this end we simulated 1000 repetitions in an Excel® spreadsheet (available from the authors), each including three random values for parking, building capacity and the combination of the other inputs. We then compared the frequency each input is minimal to its PTF and adjusted the capacities subject to the budget constraint until the desired criticality was obtained. At that stage, however, the ratio among the inputs was no longer the same as that used to calculate the PTF, which required a new adjustment, and so on.

Table 1 traces the process. The utilization balance case yields criticalities of 0.25, 0.311 and 0.439 instead of the PTF of 0.32, 0.08 and 0.6. Note that in spite of the fact that the nominal capacity of all resources is 150 the true capacity is only 127.61. After one iteration the criticalities approximately match the PTF and the capacity increased by about 2.4% to 130.70. The main change necessary for that purpose was to increase the capacity of the parking area to 178 mainly at the expense of input 3. This changes the PTF and led to one more iteration that actually reduced the capacity to 130.66. This reduction reflects the

fact that we are using a heuristic. The optimal solution as identified by Solver yields 130.74. Note that the difference between the optimal solution and the result of the heuristic is about 0.06%, which is totally negligible in practice. If we would choose to reject the last step—since it was detrimental—the difference would be only 0.03%.

To illustrate the hierarchical approach consider that the building is divided to parts such as the dining room and the kitchen, each of which can be broken down further; e.g., the number of tables of each size has an effect on the utilization of the dining area. Similarly, input 3 represents several resources such as staff, equipment, ingredients, etc.

Table 1. Performance of the heuristic

| | Input 1 | Input 2 | Input 3 | Minimum |
|---|---|---|---|---|
| Initial PTF | 0.32 | 0.08 | 0.60 | |
| Initial Capacities | 150.00 | 150.00 | 150.00 | 127.61 |
| Criticalities | 0.25 | 0.311 | 0.439 | |
| First capacity adjustment | 153.04 | 180.33 | 144.33 | 130.70 |
| New PTF | 0.326 | 0.096 | 0.577 | |
| 2nd capacity adjustment | 152.2821 | 177.96 | 145.05 | 130.65 |
| Optimal capacities | 154.7656 | 181.11 | 143.31 | 130.74 |
| Optimal criticalities | 0.295 | 0.086 | 0.619 | |

## 9. Conclusion

We extended the newsboy model to the balance of n parallel resources and a single output—with the objective of maximizing the expected economic return (net throughput). In addition to an analytic result, we obtained a deceptively simple approximation: the optimal criticality (probability of shortage) of a resource should be directly proportional to the marginal cost of increasing its relative capacity. This, in turn, is a measure of the true economic value of the resource. The model is designed to steer change and improvement in hierarchical systems. Our analytical model is a simple generalization of the newsboy model, but numerical experience demonstrates that the analytic basis of this model is sensitive to deviations from the simplistic assumption that the model is based on—namely that expansion simply shifts the capacity distribution by a constant ($g^0$) —while our approximation model is robust. Thus we can only recommend the exact method where the assumption is known to be reasonable. ($g^0$ is the implicit standard in the literature, however.) Finally, we defined a balance coefficient that may be used to compare subsystems or even independent systems. The higher the coefficient, the more urgent it is to balance the system.

Our model is simple, but it's up to the readers to decide whether it is too simple or useful. However, we demonstrated that BN-subordination (i.e., focusing on one resource and ensuring that it will not starve) is indeed too simple and does not work as advertised. We proved theoretically and demonstrated numerically that it is very wasteful. We believe that consultants and practitioners, especially those who teach or accept that throughput maximization is important, should abandon BN-subordination forthwith. Our approximation can help focus on the best balancing steps required to move away from BN-subordination while maintaining optimal growth. Of course, our model is equally useful when lack of balance is due to any other cause. Often the culprit is lack of proper coordination between hierarchical levels—which is why the hierarchical aspects of our model are important Trietsch [22].

There are myriad potential extensions of the model. On the technical front, we might wish to consider statistical dependence; this is easy for $g^0$ but is still an open research question for $g^1$. An interesting result that could be useful for the technical analysis is given by Bertsimas *et al*. [3], who use mean and variance information to find a bound on the distribution of any order statistic, and the minimum is therefore a special case their bound can provide an approximation for. This approach can handle statistically dependent variables. An extension of the model itself is to consider multiple products, multiple alternative processes and multiple periods. It can be shown that if we assume $g^0$ then the results of Bradley and Glynn [4] can be interpreted as a direct generalization of our model where the optimal criticality of a resource is determined by the marginal cost of increasing it divided by the value of the total throughput that depends on it. Trietsch [20] discusses this approach in qualitative terms. A multi-period model may focus on the dynamics of the system and its statistical estimation aspects. Similarly, a more detailed study considering specific input/output structures (layouts) in detail is required. This may also be a proper framework to study how to measure the criticality of resources. An important extension is to include discrete expansions—a case that can only be approximated by our continuous model. Yet another research direction is a more detailed study of the hierarchical implications of the model. Last but not least, empirical research may focus on the size of typical practical balancing opportunities that are caused by management either not using any model or using an inappropriate one.

## References

1. Atwater, B. and Chakravorty, S. S. (2002). A study of the utilization of capacity constrained resources in drum-buffer-rope systems. *Production and Operations Management,* 11(2), 259-273.
2. Arthanari, T. and Trietsch, D. (2004). A graphical method for the pursuit of optimal or near-optimal stochastic balance. *Proceedings of the 9th International Conference on Industrial Engineering - Theory, Applications and Practice.* The University of Auckland, November 27-30, 260-266.
3. Bertsimas, D., Natarajan, K. and Teo, C.-P. (2005). Tight bounds on expected order statistics. *Probability in the Engineering and Informational Sciences* (to appear).
4. Bradley, J. R. and Glynn, P. W. (2002). Managing capacity and inventory jointly in manufacturing systems. *Management Science,* 48(2), 273-288.
5. Chu, C., Proth, J.-M. and Xie, X. (1993). Supply management in assembly systems. *Naval Research Logistics,* 40, 933-949.
6. Goldratt, E. (1990). *Theory of Constraints.* North River Press, Croton-on-Hudson.
7. Harrison, J. M. and Van Mieghem, J. A. (1999). Multi-resource investment strategies: Operational hedging under demand uncertainty. *European Journal of Operational Research,* 113(1), 17-29.
8. Hopp, W. J. and Spearman, M. L. (2001). *Factory physics.* 2nd edition. Irwin/ McGraw-Hill.
9. Kumar, A. (1989). Component inventory costs in an assembly problem with uncertain supplier lead-times. *IIE Transactions,* 21(2), 112-121.
10. Lawrence, S. R. and Buss, A. H. (1995). Economic analysis of production bottlenecks. *Mathematical Problems in Engineering,* 1(4), 341-369.
11. Mukherjee, S. and Chatterjee, A. K. (2006). Unified concept of bottleneck. Working Paper No. 2006-05-01, IIM Ahmedabad.
12. Ronen, B. and Trietsch, D. (1988). A decision support system for planning large

projects. *Operations Research,* 36(6), 882-890.

13. Ronen, B. and Spector, Y. (1992). Managing system constraints: a cost/utilization approach. *International Journal of Production Research,* 30(9), 2045-61.
14. Schonberger, R. J. (1982). *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity.* The Free Press.
15. Shingo, S. (1988). *A Study of the Toyota Production System*. Productivity Press.
16. Spearman, M. L. (1997). On the theory of constraints and the goal system. *Production and Operations Management,* 6(1), 28-33.
17. Trietsch, D. (1996). Economic resource balancing in plant design, plant expansion, or improvement projects. *Proceedings of the 32nd Annual Conference of ORSNZ*, August, 93-98.
18. Trietsch, D. and Buzacott, J. (1999). Managing change and improvement by hierarchical balancing of service levels, MSIS, University of Auckland, Working Paper 255.
19. Trietsch, D. and Quiroga, F. (2004). Coordinating n parallel stochastic activities by an exact generalization of the newsvendor model. ISOM Working Paper #282, August (revised July 2005). http://ac.aua.am/dtrietsch/web/RTK.htm.
20. Trietsch, D. (2005). From management by constraints (MBC) to management by criticalities (MBC II). *Human Systems Management,* 24, 105-115.
21. Trietsch, D. (2006). Optimal feeding buffers for projects or batch supply chains by an exact generalization of the newsvendor model. *International Journal of Production Research*. 44(4), 627-637.
22. Trietsch, D. (2007). System-wide management by criticalities (MBC II): hierarchical economic balancing of stochastic resources. *Human Systems Management*, 26, 11-21.
23. Van Mieghem, J. A. and Rudi, N. (2002). Newsvendor networks: inventory management and capacity investment with discretionary activities. *Manufacturing & Service Operations Management*, 4(4), 313-335.
24. Van Mieghem, J. A. (2003). Capacity management, investment and hedging: review and recent developments. *Manufacturing & Service Operations Management,* 5(4), 269-302.

## Appendix

***Proofs of Lemmas and Theorems***

**Proof of Lemma 1:** Conditions 2 and 3 are proved by construction: For any possible investment, by the CDF's before and after, calculate $\Delta$ to satisfy condition 3. Then, for every value x, $g_i(x,\Delta) = \arg_y\{F_{i|\Delta}(x) = F_i(y)\}$. Condition 1 states that a zero expansion has no effect.

**Proof of Lemma 2:** Let $\mu_1$ and $\mu_2$ be any tentative solution such that $c_1\mu_1+c_2\mu_2 = b$. The budget constraint is then $c_1\Delta_1+c_2\Delta_2 = 0$, leading to the Lagrangian,

$$L(\Delta_1,\Delta_2,\lambda) = \int_0^\infty (1-F_1(g_1(x,\Delta_1)))(1-F_2(g_2(x,\Delta_2)))\,dx + \lambda(c_1\Delta_1 + c_2\Delta_2),$$

subject to $\Delta_j \geq -\mu_j$ (j = 1,2). Setting $\partial L / \partial\Delta_1$ to zero yields,

$$\int_0^\infty \frac{\partial g_1(x,\Delta_1)}{\partial\Delta_1} f_1\big(g_1(x,\Delta_1)\big)\Big(1-F_2\big(g_2(x,\Delta_2)\big)\Big)\,dx = \lambda c_1.$$

Notice that

$$\int_0^\infty f_1\left(g_1(x,\Delta_1)\right)\left(1-F_2\left(g_2(x,\Delta_2)\right)\right)dx = p_1.$$

This is true because $f_1(g_1(x,\Delta_1))$ is the density of $X_1$ and $(1 - F_2(g_2(x,\Delta_2)))$ is the probability that $X_1 \le X_2$. Therefore, $f_1(g_1(x,\Delta_1))(1 - F_2(g_2(x,\Delta_2)))$ / $p_1$ is the conditional density function of $X_1 | X_1 \le X_2$, and therefore at the optimum,

$$\int_0^\infty \frac{\partial g_1(x,\Delta_1)}{\partial \Delta_1} f_1\left(g_1(x,\Delta_1)\right)\left(1-F_2\left(g_2(x,\Delta_2)\right)\right)dx = p_1^*\ E\left(\frac{\partial g_1(x,\Delta_1)}{\partial \Delta_1} | X_1 \le X_2\right) = \lambda c_1.$$

A symmetric expression obtains for $R_2$, and division yields the first part of the lemma. The second part follows immediately (negative proportionality factors are allowed).

**Proof of Lemma 3:** In these two cases a linear cost function for each resource implies a linear cost function for any combination of resources. The proof follows by $R_i$-partitioning and Lemma 2.

**Proof of Theorem 1:** For $g^0$, $\partial g / \partial \Delta = -1$, so $q_i / p_i^* = q_j / p_j^*$ $(\forall i,j)$.

**Proof of Theorem 2:** For $g^1$, $\partial g / \partial \Delta_{|\Delta=0} = -x / \mu_i$, so $c_i\mu_i / p_i^*$ must be proportional to $E(X_i | X_i \le X_k; \forall k)$. But $E(X_i | X_i \le X_k; \forall k)$ is the mean of an exponential variable with a rate equal to the sum of the rates of all resources, and thus $E(X_i | X_i \le X_k; \forall k) = E(X_j | X_j \le X_k; \forall k)$ $(\forall i,j)$.

**Proof of Theorem 3:** $g^0$ is linear and leads to a linear expansion. For a small $\delta$, the original probability of falling between x and $x+\delta$ is approximately $f_i(x)\delta$, and after proportional expansion the same probability applies to falling between $x(1+\Delta_i / \mu_i)$ and $(x+\delta)(1+\Delta_i / \mu_i)$. Therefore, the expansion described by $g^1$ is also linear (although $g^1$ itself is a strictly convex function of $\Delta$). For $0 \le \kappa_i \le 1$, the expansion described by $g^\kappa$ is a convex combination of the two linear expansions of $g^0$ and $g^1$, and thus linear too. So Z involves maximizing the minimum of linear functions subject to a linear budget constraint, and must be concave.

**Proof of Theorem 4:** $|\mu_1^* - \mu_2^*|$ is constant, so $|\mu_1^* - \mu_2^*| / \mu_j^* \to 0$ $(j = 1,2)$ as $b \to \infty$. Hence, in the limit, $E(X_1 | X_1 \le X_2) / \mu_1^* = E(X_2 | X_2 \le X_1) / \mu_2^* = 1$. This, with Lemma 3, yields the required result.

**Proof of Theorem 5:** Immediate for $n = 2$ and by induction for $n > 2$.

*Authors' Biographies:*

**Dan Trietsch** is a Professor of Industrial Engineering at American University of Armenia. His current research focuses on management of stochastic systems by balanced safety buffers of stock, capacity and time; e.g., in sequencing and scheduling, capacity management and project scheduling. He is a co-author (with Kenneth R. Baker) of a recent book on sequencing and scheduling, the author of a book on statistical quality control and has published in the fields of O.R., transportation, applied mathematics (with a focus on network design), and quality.

**Francisco Quiroga** is Head of Strategic Development and Research at Villacero. His research focuses on applied stochastic scheduling in project networks and multi-stage decision-making with imperfect information. His publications have appeared in the Applied Engineering and Economics literature.