From Management by Constraints (MBC) to Management By Criticalities (MBC II)

Dan Trietsch

Abstract. The five-step "Theory of Constraints" as articulated and explained in Goldratt's books, is touted as "not only beneficial but mandatory". However, although it is indeed a useful focusing heuristic methodology with an impressive track record, it is not really a theory and it is certainly not mandatory. Furthermore, it involves a serious internal inconsistency that must be "faced courageously": To make drum-buffer-rope (DBR) work, Goldratt forbids balance, and yet Step 4 involves steps that tend to balance the system. I restrict the term *Management by Constraints* (MBC) to the correct aspects of the methodology and argue that successful MBC applications never follow the "official version" with complete faithfulness. MBC implicitly rejects DBR and *allows* balance. We should go further and plan the capacity (criticalities) taking into consideration the variability and the economic costs of the resources. This entails balancing the criticalities of resources based on their economic value relative to the value of the throughput they serve.

Keywords: Please, supply, keywords

Dan Trietsch, PhD (Summa Cum Laude, Tel Aviv University), MBA (Cum Laude, Tel Aviv University), BSME (Technion) is an associate professor in the Department of Information Systems and Operations Management at The University of Auckland Business School. In the past he has authored and co-authored research on optimal network design (generalized Euclidean Steiner Trees), highway and highway network design models, lot splitting, project purchasing under stochastic lead times, scheduling hub airports, and stochastic scheduling subject to optimal service level. The last three subjects are related to his current research interest in hierarchically balanced service levels – designed to manage hierarchical systems or projects of any size effectively. This involves economically balancing buffers of time, of cost, of capacity and – for projects – of scope.

> "... no exceptional brain power is needed to construct a new science or expand on an existing one. What is needed is just the courage to face inconsistencies and to avoid running away from them just because 'that's the way it was always done' ". (Eliyahu M. Goldratt. Introduction to the Second Edition of "*The Goal"*)

1. Introduction and mission statement

The five focusing steps (5FS) methodology known as the "Theory of Constraints" is associated with its developer, Eliyahu M. Goldratt – who had articulated it under the second title but later, in [5], addressed it as "the five focusing steps". Armed with more than 20 years worth of 20/20 hindsight, in my opinion some important issues remain unaddressed in the literature.[1] One result is that not enough had been done to show formally that 5FS is theoretically flawed, specifically where it differs from JIT. Elsewhere [20], I discuss several such points – one of which is central to this paper and will be repeated later.

Ronen and Starr [14] refer to 5FS as *Management by Constraints* (MBC), and I propose to adopt this name but explicitly restrict it to the useful aspects of 5FS, for distinction. As with any other useful methodology, it is the duty of the academic Operations Management (OM) community to utilize and develop MBC. As academics, it is also incumbent upon us to eradicate any errors accepted by the community and to make sure that the academic credit that surrounds MBC is correctly attributed. Very few papers serve these causes, and most of those that do are concerned with Goldratt's venture into the project management field – e.g., [8,13]. But less had been done in the traditional op-

[1] With apologies, (i) I will often use the first person singular; (ii) I will reference my previous work more than is usually acceptable in polite society (see http://staff.business.auckland.ac.nz/staffpages/dtriets/HBC.htm for the relevant files). The paper is based on my personal take of the "Goldratt phenomenon" and my previous writings, so it would be difficult to avoid these transgressions.

0167-2533/05/$17.00 © 2005 – IOS Press and the authors. All rights reserved

erations management area to which MBC was originally addressed. In [20] I have already expressed my opinion how we should proceed in the project management area, and I have also discussed extensively the availability of former academic sources for virtually every single pivotal idea attributed to Goldratt – with the original PERT/CPM methodology leading the list as a prime previous example of the MBC focusing approach. [8] and [14] also provide earlier academic sources for key MBC ingredients in operations and in projects. This paper, although it is not focused on project management, has the same mission as [20]: To put 5FS in perspective, highlight the inconsistencies, and build positively on the legitimate contributions. The paper combines results from [19,23] and [21] showing how to balance stochastic systems to maximize the expected economic throughput.

Section 2 critiques DBR, a scheduling method that, unfortunately, cannot operate in a truly stochastic environment and is inferior even with deterministic inputs. Section 3 discusses MBC, i.e., the useful aspects of 5FS, loosely defined as 5FS minus DBR. In Section 4 we extend MBC to MBC II, where criticalities are balanced economically. Section 5 provides illustrations and mini-case studies, from the literature and from my own experience, demonstrating that MBC II is practical and useful. Section 6 is the conclusion.

2. DBR – a critique

As discussed in [14], the 5FS articulation was preceded by the nine principles of OPT. OPT, in turn, was based on a commercial scheduling software package (also called OPT), touted to be highly effective in a very wide range of applications (see [24] for a different point of view). At the heart of OPT, really the feature that captured the attention of the profession, is what I call the *focusing principle*, articulated by OPT rules 4 and 5 (and central in OPT rules 2 and 6): “An hour lost at a bottleneck [BN] is an hour lost for the entire system; an hour saved at a non-BN is a mirage”. Except for the focusing principle, OPT essentially summarized just-in-time (JIT) principles, but the focusing principle is not part of JIT. Drum-Buffer-Rope is the technique by which the focusing principle was used for scheduling [4,6].

Unfortunately, DBR is not a correct general approach to scheduling even if the system is deterministic. Evidence for this is incorporated within Adams et al. [1], who achieved remarkable results by focusing on bottlenecks (also subject to deterministic assumptions). They demonstrate that in the presence of scheduling conflicts the BN is not clearly defined – which explains their title. See also [11], which provides a very thorough coverage of such methods. Shortly after his joint publication with Fox [6], Goldratt expressed his own reservations about DBR [4]. Subsequently, he embarked on a new scheduling software venture called DISASTER^{TM} [16]. This software allows for BN shifts, in the same sense that the title of [1] refers to and by a method that is also similar. So DBR is no longer part of modern scheduling software. But it can also be used as an inventory control method, and this does not require software. To show that such use is contraindicated, we must discuss DBR in some detail.

DBR’s basic premise is the focusing principle and it starts by partitioning the processing resources (e.g., machines) to three mutually exclusive and exhaustive parts: (i) the drum set, D, which includes the BN itself (or the BN set if more than one resource is limiting) and all the assembly resources downstream from the BN that have at least one other input not controlled by another resource in D; (ii) B(D) – where B is a mnemonic for “before” – defined as the set of all resources that feed D, directly or indirectly; and, (iii) A(D), where A is a mnemonic for “after”, the set of resources that follow D. All the resources in B(D) require inputs from outside the system, and these enter through *gates*, which are the most upstream resources in B(D). Each gate can be traced to at least one resource in D. The very first resource in D that is encountered during this trace is the drum that controls this particular gate. As a result, each resource in D has one or more buffer/gate combinations associated with it. Ropes connect drums to their gates, acting both as information transmission lines and as work in progress inventory (WIP) limits. The “length” of each rope is the amount of WIP that is allowed between the gate and the drum. If a rope is tight, the gate is blocked and must stop supplying materials. Therefore, the sum of the lengths of all ropes is the amount of WIP that is controlled by the DBR system, and the intention is to keep this WIP within tight limits. In contrast, the WIP in A(D), ahead of the drum, is not subject to any control whatsoever! The rationale is that if the BN is identified correctly, and if it is indeed guaranteed not to shift, then it can be shown that not much WIP will tend to collect in A(D) for long. Note that, in contrast to A(D), B(D) cannot be an empty set (because at the very least it includes the gates). If there are no resources in A(D), DBR becomes CAPWIP – a valid methodology that we discuss soon. But if there are resources in A(D), and if for some (random) reason any of them manifests as the true BN, then DBR provides no control against unnecessary inventory buildups. It follows that it is risky to place the drum ahead of any resource: *all* the resources involved should be within the feedback control loop that the drum provides; i.e., the drum should be placed at the output gates of the system.

When D is located at the output gates of the system (i.e., sales), the result can be called CAPWIP – see [18] and [9]. In contrast to the related CONWIP, CAPWIP does not assume that the WIP level will be kept *con*stant, although this would be desirable, but rather that it will never exceed a *cap*. For example, in a make-to-order environment, the WIP can fall below the cap if there are no jobs waiting to be inducted. Incidentally, [9] assume that the system under control is a line, but [18] allows it to have any structure, and so shall we here. If there are multiple gates, CAPWIP requires multiple ropes – one per gate – even if there is only one drum. In spite of being technically a special case of DBR, CAPWIP is fundamentally different: it does not require the focusing step of identifying the BN and it does not use the BN for scheduling. In my opinion, CAPWIP is really a JIT variant, while DBR is not. But, unlike DBR, CAPWIP works very well. Personally I have served as a consultant to two plant managers who were also accredited Goldratt Associates, and they have been using CAPWIP. A third manager I met, an avid Goldratt Institute customer, told me that Goldratt actually added CAPWIP ropes to the DBR ropes, thus imposing a dual control on the system. In fact, [4] seems to corroborate the statement, although not very clearly and certainly without detail. Furthermore, in a more recent development, an essentially identical framework has been presented as *Simplified DBR* (S-DBR), and reportedly it is very successful [3]. The advent of S-DBR provides further evidence that DBR required modification. In conclusion, DBR – that is, the un-simplified original version – should be rejected: It has serious theoretical flaws, it does not work well in practice, and a proven better option is available.

3. MBC = 5FS − DBR

As mentioned already, in this paper I reserve the term MBC for the legitimate aspects of the five-step focusing methodology. [14] analyzed both OPT and MBC, and showed that they have been grounded in previous operations research (OR) theory. Because they had emphasized the focusing principle, however, OPT and MBC were not perceived as an adaptation of JIT. Another reason for the perceived distinction was that at the time JIT was often misinterpreted as limited to the kanban technique. In reality, one can say that MBC is analogous to the more general aspects of JIT – to which [14] refer as BIG JIT, but I will simply continue to call JIT – while kanban had been replaced by DBR. As discussed before, however, the focusing principle is correct only with deterministic capacities, deterministic demands, and without scheduling conflicts (i.e., never). Since the limitations of the focusing principle are associated with DBR, loosely speaking, I will treat MBC as 5FS minus DBR.

MBC is based on 6 steps, but the first one is implicit and not counted usually:

0. Select an objective function and decide how to measure it.
1. Identify the binding constraints that limit the objective.
2. Manage the binding constraints to improve the objective maximally.[2]
3. Subordinate everything else to the needs of the binding constraints.
4. Elevate the binding constraints.
5. Return to Step 1, since the binding constraints may have changed after Step 4.

This defines an improvement cycle with four steps (Step 0 is not repeated and Step 5 is merely an arrow closing the loop). This may look identical to 5FS, but there is a difference in the way Step 4 is interpreted. If we just follow the cycle repeatedly, and thus go through Step 4 repeatedly, with prudent steps (i.e., without elevating the bottleneck more than necessary from an economical point of view), it is likely that the system will become more balanced with time. 5FS, however, involves additional teachings that forbid such balance. Our purpose here is to highlight the difference and thus show that MBC is preferable.

Ideally, capacity balance implies fully utilizing all the resources all the time. Because this is simply impossible, we may refer to this as *naive balance.* More feasible is *balanced utilization*, where all resources are equally loaded but below 100%. In general, however, focusing on utilization neglects important stochastic

[2] I am not quoting Goldratt verbatim. In this case his words were “Decide how to exploit the constraints”, but I assume he had not intended a restriction of this to planning only. My version includes planning, execution, and control.

aspects of system behavior, such as on-time performance. Considering randomness more explicitly, we define *simple balance* as allowing each resource to be equally likely to limit the system; i.e., all parts of the system have the same *criticality,* defined as the frequency at which the resource limits the system. When all resources are symmetric, simple balance is also utilization balance, but in general under simple balance the utilization depends on the coefficient of variation of the resource capacity: high CV implies low utilization and vice versa. We also define *permission to fail* (PTF) of a resource as its desired level of criticality. PTF answers the question: how often should this resource be a bottleneck? For a while, assume the system has a single output such that the throughput is determined by the most limiting resource. Therefore, the sum of all PTFs (and all criticalites) of all parts of the system is 1 (or 100%). Henceforth, as a default, we discuss desired (actual) balance in terms of PTFs (criticalities). We can measure the desired (actual) imbalance by the difference between the highest PTF (criticality) and the lowest PTF (criticality) in the system. Thus if a system is balanced in the simple sense, the imbalance measure is 0. In contrast to simple balance, if all resources except the BN have zero permission to fail and the BN is expected to limit the system 100% of the time (i.e., it has a PTF of 100%), then the desired imbalance measure is 1 (or 100%). With this in mind, MBC does not include any rules about imbalance and therefore it is likely to lead to some arbitrary (and not necessarily high) level of actual imbalance. As we go through the cycle repeatedly, and each time Step 4 is performed with prudent steps (without waste), more resources become likely to limit the throughput. Thus, over time, the imbalance measure under MBC tends to decrease. But unless the desired imbalance measure is 1, by definition more than one resource is allowed to limit the system, so the "de-facto" BN is allowed to shift from one period to the next, depending on the stochastic demand pattern and on the stochastic capacity fluctuations (e.g., due to breakdowns). In my opinion, successful applications of MBC in practice always use this interpretation, at least implicitly. That is, no attempt is made to achieve an imbalance measure of 1 and therefore no consistent BN really exists in the system.

However, 5FS explicitly rejects balance and random BN shifts. Surprisingly, 5FS *does* require an imbalance measure of 1. Step 4 can change the BN, but the new BN is then supposed to remain consistent until the next elevation, so random BN-shifts are forbidden. In order to utilize the BN 100% 5FS necessarily requires an imbalance measure of 100%. To wit, in [5, p. 139] Goldratt provides a technically correct proof demonstrating that balance hurts throughput. More specifically, the proof shows that non-BN resources must have "more than infinitesimal excess capacity" relative to the BN, or the BN will be starved occasionally and throughput will be decreased. The proof starts with the focusing principle (the very same principle that distinguishes 5FS from JIT), adds stochastic queueing considerations, and concludes that in a balanced system the [designated] BN cannot achieve 100% utilization. According to the focusing principle, this necessarily implies a reduction in throughput. After all, the clearest dictum that follows from the focusing principle is that the BN should be utilized 100% of the time. Thus, there is a contradiction between the throughput maximization objective and balance. Goldratt concludes that balance is counterproductive. He then asserts that one must therefore "accept the resulting five focusing steps" [5, p. 141] and that "following these steps in not only beneficial, it's mandatory"; i.e., unless the five focusing steps are followed "good results will not occur" [5, p. 145]. Unfortunately, in spite of the technically correct proof, the whole argument is misleading and rooted in error. Although the presentation does not say so explicitly, it seems to suggest that it is possible to achieve 100% utilization on the BN. The proof, however, only shows that imbalance is *necessary* for 100% utilization, not that it is *sufficient*. In reality, the relative cost of the last iota of utilization from the BN can exceed the benefit infinitely! So there is an inconsistency between the assertion and the economic objective to maximize profit. To "evaporate this cloud" we must examine the assumptions. One way to resolve the issue is to examine and reject the focusing principle. Indeed, as we discussed before, in the presence of stochastic variation (or even predictable scheduling conflicts), the focusing principle fails, and the system is *not* equivalent to the nominal BN. Alternatively, we can critically examine, and reject, the objective of maximizing throughput. Throughput, by Goldratt's definition, excludes operating expenses and inventory cost [6]. But although these may be minor sometimes, they can become very important if we intentionally ignore them while pursuing uneconomical investments in excessive buffers.

4. MBC II: adjust criticalities proportionally to their economic value

Both MBC and 5FS, correctly, promote buffers as protection against stochastic problems. The main 5FS error is treating the nominal BN as an exception to the rule that *all* resources require buffers. The correct balance that is required is between all the buffers. If we wish to maximize our economic throughput, defined as net sales minus *all* economic costs of sales (including buffer costs), then we will obtain *economic balance*. Economic balance should not be confused with simple balance, but it certainly involves an imbalance measure below 100%. Because MBC allows any imbalance measure it implicitly *accepts* economic balance. In this aspect it is similar to regular JIT. 5FS, by requiring an imbalance measure of 100%, *precludes* economic balance. In this section we characterize economic balance in a way that is potentially useful for managers.

There are several types of interacting buffers in any production system, and all of them should be economically balanced. These include stock buffers, excess capacity, and safety time. In project management there are also implicit *scope* buffers [20]. A basic queueing model, M/M/1/K (with Poisson arrivals, exponential service time, a single server, and a WIP cap of K) can demonstrate that balance is required between WIP and capacity buffers. Under M/M/1/K, the probability the system will be empty is a strictly positive (but decreasing) function of K [12]. So if we want to achieve 100% utilization on the server, we must increase K to ∞. Furthermore, if the server is the BN at least 50% of the time, then the average WIP will indeed approach ∞ as K approaches ∞. Clearly this is unacceptable and economic balance should be struck between WIP and utilization. Specifically, we should select the server capacity and the WIP cap together in such a way that the desired output is achieved at minimal economical cost.

This queueing model can also demonstrate that the roles of marketing and sales (the arrival generation) and of production (the "server") are completely symmetric. For this purpose view the server and the arrival generator as two resources connected by a cycle, and K is the total number of units allowed anywhere in the enhanced system. The server only generates departures when at least one unit is in the service subsystem and similarly the arrival generator only operates when there is room in the service subsystem (so there is at least one unit between the server and the generator). The units will then shift between the two possible locations according to the random service/arrival pattern. But the system is identical to the original M/M/1/K. With this in mind we see that we need good balance between marketing and sales efforts, service capacity, and WIP. Note also that the number of units in the service subsystem is K or less – thus demonstrating that under CAPWIP there is no guarantee that the production system will always have the full desired level of WIP. During periods in which production is the BN, WIP will tend to collect in the server subsystem, but when marketing & sales are limited, there will be less physical WIP.

We can gain insight for the economic balance problem by a basic model introduced in [21] (where further details and proofs may be found). The model involves n parallel inputs and a single output. The single output represents a composite product that may comprise manufactures, software, and services. The resources include traditional production resources (such as machines, inventories and labor), but also the capacity to sell, e.g., marketing. Throughput is only realized when we can make and sell the product, so marketing is modeled as one of the n resources. Equivalently, sales = min{supply, demand}. The capacity of the resources is stochastic and the exact structure of the composite product is subject to variation between periods (which translates to variations in terms of capacity relative to demand). Resource capacities can be changed by investments (including continuous improvement projects and purchasing capacity outright). Investments are medium- or long-term, but the exact system requirements and availabilities can change each period. Therefore, the system bottleneck is likely to shift between periods.

If we could maintain a consistent BN that is utilized 100% as 5FS requires, its criticality would be 100% and its service level would be 0. In such a case no other resource could ever be critical. In this case criticality and utilization are the same (100%), but in general utilization may be higher than criticality. To find the correct balance (approximately), consider that the economic value of the future net throughput of the system is equal to (or at least not far from) the economic investment necessary for such a system, amortized according to the return that the market expects from investments of the same type. Our key assumption is that we can increase any part of the system at roughly the same rate that the system value indicates. To illustrate, if a subsystem is economically assessed at \$1 000 000, we can increase this capacity by 1% at a cost of \$10 000. Suppose now that the subsystem is a distinct BN (i.e., its criticality is close to 100%) and that it represents 40% of the total value (i.e., the full system is valued at \$2 500 000). Suppose further that if we increase the capacity of the subsystem by 1% then its criticality will be reduced to 97%. Then the expected system throughput will increase by approxi-

Table 1

Net return on 1% investment (10 000)

Scenario	Criticality after investment	Approximate throughput gain	NPV of investment	Net return on investment
1	97%	24 250	14 250	142.5%
2	80%	20 000	10 000	100%
3	60%	15 000	5000	50%
4	50%	12 500	2500	25%
5	40%	10 000	0	0
6	20%	5000	−5000	−50%
7	1%	250	−9750	−97.5%

mately \$24 250 (97% of 1% of \$2 500 000) at least, in spite of a decrease in the utilization of the subsystem. To see this consider that 3% of the time the throughput will not increase or will increase by less than 1%, but with probability 97% it will increase by about 1%. This is a net present value (NPV) gain of \$14 250, or 142.5% of the investment. But if, after a small investment, the subsystem would be critical only 40% of the time, then NPV would increase by 0 – i.e., the investment would be neutral. We can then say that the system is balanced economically.

Table 1 presents some scenarios of this type. In all cases we assume that the criticality before investment is not much higher than after. Cases 1 and 5 have been discussed before. Case 4 is simple balance, and we can see that it is different to economical balance. Cases 6 and 7 show the effect of investing too much in elevating a constraint. Generalizing the insight that 40% criticality is the correct balance for a resource that represents 40% of the whole (Case 4), we obtain our approximate economic balance rule:

The optimal criticality of a resource is approximately equal to its economic value expressed as a fraction of the whole system.

In other words, the PTF should be proportional to economic value. It can be shown that if there are multiple outputs this principle remains almost unchanged, but instead of measuring the resource as a fraction of the whole system it should be measured as a fraction of the total value of the outputs that require it. That is, the criticality of a resource should be proportional to its economic value as a fraction of the total throughput that depends on it. This result is related to optimal feeding buffers for assembly projects [10,15]. It had also been discussed in a multi-product, multi-period context by [19]. While it is only a heuristic, and indeed it is more elaborate than MBC, it is still simple and intuitive. Expensive resources should often be the BN, cheap ones rarely, and no resource should be allowed to be a consistent BN.

How to focus on the most beneficial opportunities? To implement this rule in the most economical way we must focus on the resources that are farthest from their correct criticalities. For resources whose criticality is excessive, we must match the criticality of the resource with its economic value by prudent investments. For this, we need to monitor the frequency at which the resource is the BN [21,23]. The economic value is calculated by multiplying the nominal capacity of a resource by the marginal cost of increasing its capacity. It is then expressed as a fraction of the total economic value of the system (or the total economic value of the outputs that requires this resource). If the frequency is too high, the resource justifies expansion. To focus correctly, the highest investment priority goes to the resource that maximizes $(p_i - v_i)/v_i$ where p_i is the frequency, i.e., the estimator of the criticality, and v_i is the normalized economic value (such that, for the single output case, $\sum v_i = 1$). Resources with very high $(v_i - p_i)/p_i$ values are under-utilized, however, and it is useful to find ways to increase their criticalities economically. Technically speaking, this may call for disinvestment. But considering the full implications, it is usually more attractive to achieve the same objective while growing economically by finding marketable products that can use the under-utilized capacity without loading the other resources too much. Every new product has a profile of loading it imposes on the system and it is possible to compare alternatives. One way to approach this problem is by linear programming [19].

Using the new rule, one might think that if a resource is extremely expensive or even impossible to expand, 100% utilization should be pursued on it. This is intuitive because such a resource can be viewed as having a value of ∞. But unless we can *sell* the resource for ∞, it is still wasteful to protect it fully against idling. The economic value of a resource is not really determined by the potential cost of increasing it but rather by the value of the sales it can generate. There is no resource whose output can be sold for an infinite return, so even completely rigid capacity resources should not be overprotected. This has also been demonstrated by simulation [2]. Furthermore, the owners of a resource with a truly infinite value do not need optimization models: they can better utilize their time enjoying, or perhaps protecting, their already infinite wealth.

Define a *composite resourc*e as an aggregation of parts such that the throughput of the aggregation is de-

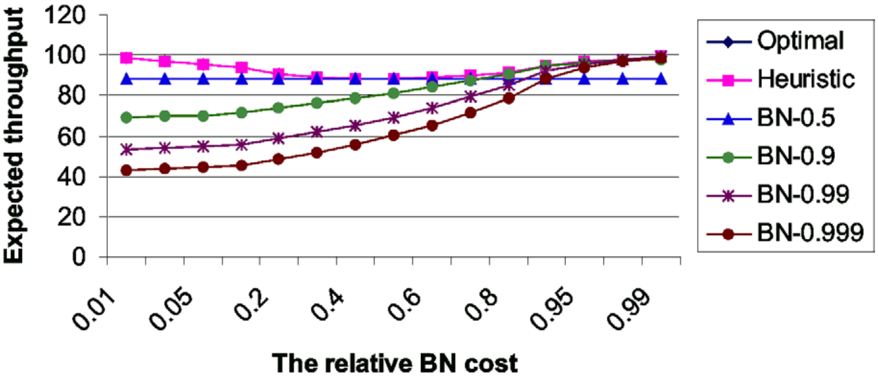

Fig. 1. Comparing BN-subordination and regular balance with economical balance.

termined by the minimum throughput of the parts. This can be used to describe large systems hierarchically. Also, it allows balancing any single resource with the composite resource representing the remainder of the system. Figure 1 is reproduced from [21], and it involves two composite resources with normally distributed capacities that are subject to change by investments. One of the resources is designated as the (nominal) BN. A budget limits the total investments, so whatever we invest to protect the BN will not be available for the BN itself. We normalize the scale such that the cost of each capacity unit of the BN plus that of the other resource is one, and there is a budget of 100. E.g., if the cost of each capacity unit of the BN is 0.8 the cost of the other resource is 0.2, and we can choose to buy 100 capacity units of each, or 95 units of the BN resource and 120 units of the other, etc. It is possible to achieve any desired BN-criticality strictly below 1, e.g., 0.999, 0.99, 0.9, or 0.5, marked in the figure as BN-0.999, BN-0.99, BN-0.9, and BN-0.5. The figure gives the expected throughput as a function of the marginal cost of each capacity unit of the BN resource for various budget allocations. Five results are visible and a sixth – the optimal balance – is practically covered by the balancing heuristic results. By the figure, when the costs are equal and we invest the same in each of the two resources (BN-0.5), the utilization is 88% and this is also optimal. But when the costs are far from equal (e.g., 0.01 or 0.99) the optimal solution yields a higher utilization of the investment as a whole, approaching 100% at the limits. This is achieved by protecting the expensive resource abundantly (but not excessively) by the cheap resource, at progressively negligible cost. For this reason, when the cost of BN capacity approaches 1, very high criticality (and therefore a very high utilization) is not far from optimal. But it is only better than regular balance for very high BN values, where it is approximately correct by the economic balance rule. For lower BN values, there is an increasing loss involved, and the more faithfully one follows 5FS, the higher the loss. To the extent that 5FS set out to combat the waste associated with naive balance (i.e., the notion that all resources should be fully utilized), the cure is far worse than the disease!

Therefore, we must limit our endorsement of 5FS to what we call MBC, which allows balance. But it is now clear that we should actively pursue balance. My suggestion is to reinterpret MBC as Management by Criticalities (MBC II), and to change Step 4 accordingly.

4.1. Balance resource criticalities to match their marginal economic value as a fraction of the throughput that depends on them

Once we do this, however, it becomes apparent that in Step 1 we are considering the short-term constraints, based on our current product structure, while in Step 4 we are looking for a long-term balance. This is also true for the original version (elevate), but it is now clearer. Therefore, Step 4 operates on a different schedule than Steps 2 and 3. So it is not correct to include all the steps in the same cycle. A better approach is to view MBC II as two parallel cycles:

Cycle 1 (routine operations)

1. Identify the binding constraints that limit the objective currently.
2. Manage the binding constraints to improve the objective maximally (including balancing other than by medium- and long-term investment).
3. Subordinate other resources to the needs of the binding constraints.
4. Return to 1.

Cycle 2 (investments for the medium and long term)

1. Identify the constraints whose medium- or long-term criticalities are excessive or (when disinvestment is allowed) too low.
2. Maximize net present value by matching long-term criticalities of resources to their economical levels; i.e., the criticality of each resource should equal its economical value as a fraction of the throughput that depends on it.
3. Subordinate other investments as necessary to support Step 2.
4. Return to 1.

In both cases, Step 1 requires related analysis. Repetitively performing Step 1 of Cycle 1 (which we denote by C1/S1) provides information for C2/S1. But other than that, the cycles are separate (parallel to each other).[3]

5. Examples

This section includes examples aimed to illustrate the practical relevance of MBC-II. The first example is based on the literature and observations during plant visits in the late 1980's and early 1990's; the other two are based on my personal experience.

Mixed-Model Paced Line Operations at Toyota: Toyota is a pioneer of mixed model paced assembly line operations. Periodically, the product mix is changed to match the demand pattern. Whenever this happens, employee teams must rebalance the line because the load profile is a function of the model mix. In terms of MBC II, this activity belongs to C1/S2, which is why this step includes balance other than by investment. Toyota also has special teams tasked with lending help where it is required sporadically. This is a generalized buffer similar to a CAPWIP buffer in the sense that it can focus on current trouble spots. To see the similarity consider that under CAPWIP the safety stock tends to accumulate consistently ahead of resources whose capacity is tightest and sporadically ahead of resources whose variance is highest or which follow high variance resources. This happens naturally and can be demonstrated by basic queueing rules [9]. So, by rushing to trouble spots, the support team helps the same resources that would attract the WIP buffers in an un-paced line. Yellow andon lights are used to show where buffers are stretched (so the support team knows where to go), and the line is considered balanced if they appear sporadically all over the floor. That is, balance means that the identity of the most critical resource shifts randomly over time. The speed of the line is also the result of balance, measured by the frequency of red andon lights. This supports our measurement of balance in terms of criticality.

Balancing Inspection and Machining: In a jet engines refurbishing plant at Naval Aviation Depot (NAD) Alameda, managed by a dedicated Goldratt Associate, the first major step in an extended MBC journey involved implementing CAPWIP. Following an initial period when lead times actually increased – because workers could no longer cherry-pick the easiest jobs and postpone the tough ones indefinitely – lead time decreased and throughput was improved [7]. After this step, it was observed that a disproportionate amount of WIP was consistently queued in front of a non-destructive inspection (NDI) department. NDI identified parts with hairline cracks, and condemned them. In contrast, the machining department was visibly under-loaded, and assembly also had sufficient capacity. Rough analysis revealed that all parts had to visit NDI twice, once before machining and once before assembly into a final refurbished engine. The first inspection was necessary to prevent machining parts that would have to be condemned eventually, and the second inspection was absolutely required for safety reasons. By identifying the parts that had the least frequency of being condemned on the first NDI round relative to the amount of redundant machining risked, and letting these parts skip the first inspection, NDI and machining became much more balanced while throughput of good parts increased further by about 9%. At this stage, final assembly became the most critical resource, which was deemed satisfactory. In addition to demonstrating that economic balance increases net throughput, this is also a rare case where C2/S2 was achieved with practically no investment (although there was some marginal machining cost in-

[3] If negative investments are allowed, we must pay explicit attention to the difference between investment costs and salvage value. When salvage values are lower than investment costs, resources are naturally partitioned to three mutually exhaustive sets: (i) those whose criticality is excessive so they justify positive investment; (ii) those whose criticality is too low so, in the economical sense, they may justify disinvestment; and (iii) those whose criticality is too high to justify disinvestment at the low salvage value but too low to justify investment at the high investment cost (for more details and earlier references, see [21]). If we limit ourselves to positive investments (growth), only the first set should be considered. Otherwise, we can achieve balance faster by transferring investments from the second set to the first set.

volved which is mathematically identical to any other amortized investment cost). In the final analysis, machinists often found cracked parts to condemn even without the NDI, so the wasted machining was quite minimal.

Stocking Repair Parts at SONY New Zealand: At the SONY NZ Service Centre (repair facility), in spite of an inventory of parts valued at $900 000, the service level (SL) was only 68% (i.e., 32% of repair parts required were not available immediately). Items had different frequencies of use, different holding costs, and different lead-times. By adjusting the safety stock levels such that their criticalities were proportional to their holding cost, the total value of inventory dropped to $450 000 and yet SL increased to 87%. Later, actions were taken to reduce the lead-time of many expensive items and as a result the total value of inventory dropped further to $300 000 but SL increased to 94%. Table 2 summarizes this progression.

The last column gives an upper bound on the repair parts criticality, namely 1-SL. To see this and to present a relatively tight lower bound (which is not included in the table) it is necessary to look at the system in some detail (Fig. 2). Jobs enter the diagnosis part of the service through an induction queue. The main function here is to identify the necessary repair part (or kit of parts). Once identified, with probability SL the part (or kit) is on the shelf, and the job is completed and released after the minimal service time required. For this fraction of our throughput repair parts are not critical, so 1-SL is an upper bound on their combined criticality. If a repair part is missing, the job enters an interim stocking point (physically, a highly visible shelf) where it awaits the part. This can be interpreted as a queue for the service of the repair part subsystem. When the part arrives, the job enters the repair phase of the service (with a high priority). Because parts are essential, technicians idle whenever the induction queue is empty even if some jobs are awaiting parts. Hierarchically, the repair part subsystem is best viewed as part of the service system and indeed, from the point of view of customers (who bring in the repair jobs) it does not make a difference if their job is queued outside the service system or inside. Therefore, the repair part performance should be in balance with the technicians' capacity. The frequency at which at least one technician is free but there are jobs awaiting parts is a lower bound on the criticality of the repair parts. It is only a lower bound because jobs that lack repair parts have to be processed twice and this wastes some repair capacity (i.e., some of the observed technicians' criticality is really due to repair parts). From the facility manager's point of view, however, the frequency of this visible event is a good approximation for parts criticality. The test of balance was that this measure was roughly close to the relative value of inventory in the system. (For more on observing queues and other ways to measure criticality see [21].)

Table 2

The progression of balance and improvements in repair parts inventory

	Repair parts inventory budget	SL	Criticality bound
Initial state	900 000	68%	32%
After balance	450 000	87%	13%
After lead time reduction	300 000	94%	6%

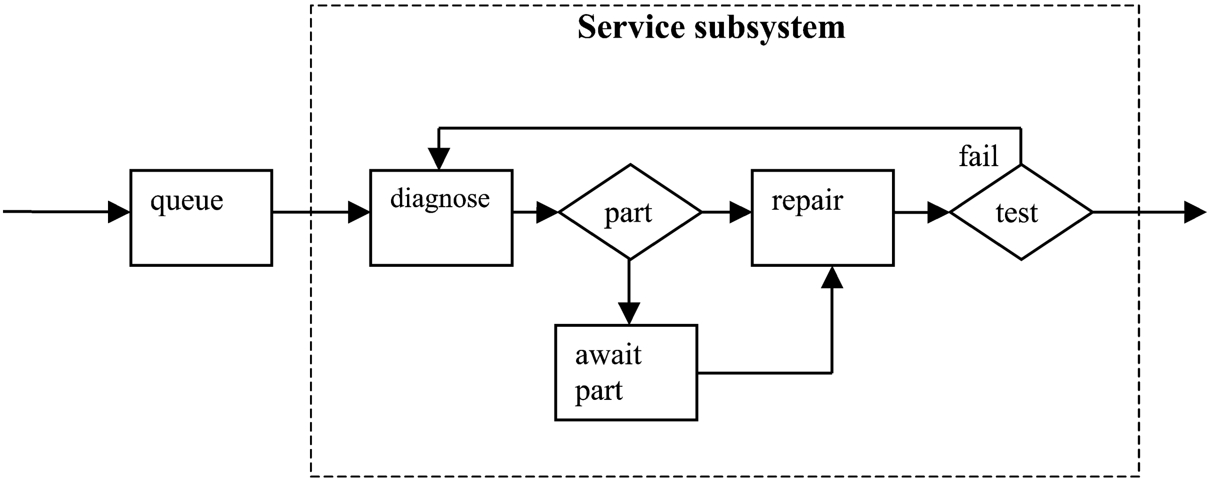

Fig. 2. Flow of materials at SONY (New Zealand) service centre.

It may also be useful to discuss the way that the criticalities of individual repair parts should be balanced within the total budget. This was actually accomplished by slightly adapting classical stochastic inventory models, but here we limit ourselves to the conceptual measurement involved. Each part number (i.e., each specific repair part) has an expected number of shortages per period, a function of the base stock we use (our decision variable). The part number also has an expected delay time per shortage, which is a function of the lead time and of the base stock. The product of these two is a measure of the *combined shortage* of this part number (per period). When divided by the sum of the combined shortages of all part numbers, we obtain the relative weight of this part number in the total repair part criticality, as a fraction. Multiplying this fraction by the repair part budget we obtain the correct amount of the budget that should be allocated to this part number. However, rounding may force deviations. Also, some high value low usage items are best not stocked at all, especially if their lead times are tolerable (which may be interpreted as "rounded to zero"), so they are not subject to the budget. Indeed, a sizable part of the improvement was achieved by excluding such parts from the stock, and the lead time reductions made possible the exclusion of even more such items.

The same environment also provides an example of balance associated with changing the system rather than just adjusting it. While pursuing the process of balancing the system, it was recognized that fast customer service was essential, especially with respect to customers whose unit was still under factory warranty. This led to a new marketing approach. It involved going beyond the standard that prevailed in New Zealand and replacing products under warranty with new ones immediately. The defective ones were then repaired and sold at discount out of a refurbished goods outlet situated alongside the service reception. This facility became popular enough to turn over an annual sale of $1.5Million in FY03 and a tidy profit. As an additional example of balance, the prices at the refurbished goods shop were set so that their average shelf life was in balance with the throughput. On the one hand, when prices are too high, the turnaround time is high. In addition, shelves are full and off-site storage may be required. Since consumer electronics products lose value as soon as a new model is introduced, overstocking can be very costly. On the other hand, when prices are too low there are not enough units on the shelf to attract customers to visit the shop in the first place, not to mention the reduced income per unit sold.[4]

6. Conclusion

Based on the literature and anecdotal evidence, 5FS works. I argue, however, that it is really a subset of 5FS, namely MBC, that deserves this commendation. More specifically, following the focusing principle religiously is very counterproductive; e.g., DBR is inherently a manifestation of excessive reliance on the focusing principle. Instead, we need balance. The fact that MBC – i.e., 5FS but without taking the focusing principle beyond its usefulness – *can* work probably contributed to the wide-spread myth that 5FS is meritorious. Those who think about 5FS as a theory may even believe that these successes constitute a proof!

In order to move forward, I suggested starting with MBC, and introducing active economic balancing into it, thus obtaining MBC II (Management by Criticalities). Practical examples were presented demonstrating that MBC II is useful. Personally, I also believe that the adoption of neutral terms is vital. "5FS" is a neutral term that can be used to describe the method as explained by Goldratt. Indeed, it's a term that Goldratt himself has used [5]. In contrast, "TOC" has controversial connotations and it acts as a lightning rod for some professionals. Specifically, they include those who do not think it is a theory and those who do not think that Goldratt deserves academic credit for it [20]. The term "MBC" has been introduced in the very early days of 5FS and it describes the valid aspects of the methodology very clearly and correctly [14]. Because it had not been universally accepted as synonymous with "5FS", the term "MBC" can serve the purpose of denoting the correct aspects of 5FS. MBC II is not intended to invalidate MBC but rather to build on its proven success.

Both MBC and MBC II lack an explicit approach to hierarchical implementation. [21–23] discuss hierarchical balancing of criticalities, which is a combination of MBC II and policy deployment designed to improve both. In subsequent work (based on these sources) I intend to show that without hierarchical balancing of criticalities highly successful MBC II efforts can be squandered. My own experience includes a case where the US Navy failed to utilize an opportunity to leverage

[4] The facility manager who achieved all this was subsequently appointed as General Manager of SONY India – a five-fold larger operation – where he proceeded to implement similar methods.

the results of a successful MBC application at naval shipyards by about fifty fold [17]. In fact, aided and abetted by Congress, they allowed an annual savings opportunity of about 750 millions slip through their fingers. The investment necessary for this savings was completely negligible – it mainly involved making the right decision. But making the right decisions in hierarchical systems is not a negligible challenge.

Acknowledgement

The author is grateful to David Doll (formerly Head of the TQM Office at NAD Alameda), Jerry Ghiselli (formerly T56 Engine Division Head at NAD Alameda) and Aldon Hartley (General Manager, SONY India).

References

[1] J. Adams, E. Balas and D. Zawack, The shifting bottleneck procedure for job shop scheduling, *Management Science* **34**(3) (1988), 391–401.

[2] B. Atwater and S.S. Chakravorty, A study of the utilization of capacity constrained resources in drum-buffer-rope systems, *Production and Operations Management* **11**(2) (2002), 259–273.

[3] H.W. Dettmer and E. Schragenheim, *Manufacturing at Warp Speed: Optimizing Supply Chain Financial Performance*, CRC Press, 2000. (See also http://www.goalsys.com/HTMLobj-269/S-DBRPaper.PDF.)

[4] E.M. Goldratt, Computerized shop floor scheduling, *International Journal of Production Research* **26**(3) (1988), 443–455.

[5] E.M. Goldratt, *Critical Chain*, North River Press, 1997.

[6] E.M. Goldratt and R.E. Fox, *The Race*, North River Press, 1986.

[7] V.D.R. Guide, Jr. and G.A. Ghiselli, Implementation of drum-buffer-rope at a military rework depot engine works, *Production and Inventory Management Journal* **36**(3) (1995), 79–83.

[8] W. Herroelen and R. Leus, On the merits and pitfalls of critical chain scheduling, *Journal of Operations Management* **19** (2001), 559–577.

[9] W.J. Hopp and M.L. Spearman, *Factory Physics*, 2nd edn, Irwin, 2001.

[10] A. Kumar, Component inventory costs in an assembly problem with uncertain supplier lead-times, *IIE Transactions* **21**(2) (1989), 112–121.

[11] T.E. Morton and D.W. Pentico, *Heuristic Scheduling Systems with Applications to Production Systems and Project Management*, Wiley, 1993.

[12] S. Nahmias, *Production and Operations Analysis*, Irwin, 1989.

[13] T. Raz, R. Barnes and D. Dvir, A critical look at critical chain project management, *Project Management Journal* (December) (2003), 24–32.

[14] B. Ronen and M.K. Starr, Synchronized manufacturing as in OPT: from practice to theory, *Computers and Industrial Engineering* **18**(8) (1990), 585–600.

[15] B. Ronen and D. Trietsch, A decision support system for purchasing management of large projects, *Operations Research* **36**(6) (1988), 882–890.

[16] J.V. Simons and W.P. Simpson III, An exposition of multiple constraint scheduling as implemented in the goal system (formerly DISASTER$^{\mathrm{TM}}$), *Production and Operations Management* **6**(1) (1997), 3–22.

[17] D. Trietsch, Focused TQM and synergy: a case study, *AS* Working Paper 92-06, Naval Postgraduate School, 1992. Available for non-commercial use at http://staff.business.auckland.ac.nz/staffpages/dtriets/HBC.htm.

[18] D. Trietsch, JIT for repetitive and non-repetitive production, in: *Quality Management for System Optimization: Leadership, Focusing, Analysis and Engineering*, Chapter 21 (Draft textbook), 1995. Available for non-commercial use at http://staff.business.auckland.ac.nz/staffpages/dtriets/HBC.htm.

[19] D. Trietsch, Economic resource balancing in plant design, plant expansion, or improvement projects, in: *Proceedings of the 32nd Annual Conference of ORSNZ*, 1996, pp. 93–98. Available for non-commercial use at http://staff.business.auckland.ac.nz/staffpages/dtriets/HBC.htm.

[20] D. Trietsch, Why a critical path by any other name would smell less sweet: towards a holistic approach to PERT/CPM, MSIS, *Project Management Journal* **36**(1) (2005), 27–36.

[21] D. Trietsch, Balancing resource criticalities for optimal economic performance and growth, ISOM, University of Auckland, Working Paper No. 256, 2004. Available for non-commercial use at http://staff.business.auckland.ac.nz/staffpages/dtriets/HBC.htm.

[22] D. Trietsch, From the flawed "Theory of Constraints" to Hierarchically Balancing Criticalities (HBC), ISOM, University of Auckland, Working Paper No. 281, 2004. Available for non-commercial use at http://staff.business.auckland.ac.nz/staffpages/dtriets/HBC.htm.

[23] D. Trietsch and J. Buzacott, Managing change and improvement by balancing service levels hierarchically, MSIS, University of Auckland, Working Paper No. 255, 1999. Available for non-commercial use at http://staff.business.auckland.ac.nz/staffpages/dtriets/HBC.htm.

[24] B. Wilkins, Judge orders software firm to hand over source code, *Computerworld* **9** (1984), 2.