System-wide Management by Criticalities: Hierarchical balancing of stochastic resources

“Make everything as simple as possible, but not simpler”
(Albert Einstein).

Dan Trietsch
College of Engineering, American University of Armenia, Yerevan, Armenia

Abstract. In a companion paper I argued that Management by Constraints (MBC) should be improved by replacing Step 4 (elevate the constraint) by a revised version that seeks to match the criticality of each constraint (defined as the probability it will be the binding constraint in any period) with its economic value. The result is Management by Criticalities (MBC II). In this paper I show how the principles of MBC II can be pursued in hierarchical systems. Instead of asking subsystems to meet rigid objectives (as in Management by Objectives or Policy Deployment), we ask them to meet the demand placed on them with a specified service level, defined as the complement of the criticality. In effect, the result is a combination of MBC and of policy deployment aimed to improve both.

Keywords: Focusing, policy deployment, stochastic balance, theory of constraints

1. Introduction

This paper is about managerial focusing in realistic systems. Many believe that humans are only capable of focusing on one or very few goals. Therefore, they preach focusing on one or very few “wildly important goals” (WIGs) [5, p. 281]. Modeling an organization as if it has one or very few binding constraints, and as if this is how it should always remain, is based on the very same interpretation of reality. Similarly, approaches such as “zero defects” (ZD) or “zero inventories” (ZI) offer WIGs that are supposed to be effective for systems of any size. Because they promise big benefits, we refer to such global WIGs as BIGWIGs. However, although in a descriptive sense it may be true that most individuals cannot maintain a wide focus – come to think of it, “wide focus” is an oxymoron – in a normative sense applying a narrow focus naively in an organization simply cannot work. An organization, as any other viable system, is subject to myriad threats and may have myriad opportunities. Neglect even one threat, and that may be the fatal error. Miss too many opportunities (by excessive focusing that leads to tunnel vision), and death will be slower, but certain. Thus, BIGWIGs can cause big damage. How can an organization focus on an uncountable number of important goals or issues and yet allow each human participant to focus on precious few? Natural evolution has one tried and true response – and only one – hierarchy! For example, while consciously we may be focusing on one WIG (or, for that matter, indulging in utterly frivolous activities instead), parts of our body are focused on totally “unimportant” goals (or activities) such a main-
taining blood pressure within viable limits. People who face death due to problems with this minor detail may call it wildly important, but clearly it had been delegated to subconscious parts of the system in order to be performed routinely. Furthermore, the objectives of the blood pressure maintenance system are not directly related to the objectives of the human “owner-operator” of the particular body, but when this human is considered as a system, they are as vital as any other objective the system may possess. In other words, the issue is how to focus within hierarchical systems in such a way that each part will have few local objectives to focus on and yet, by the parts pursuing their local objectives, the needs of the whole will be met. For simplicity, we will generally assume a basic hierarchy with a tree structure, but the essence of the hierarchical solution is not in the tree structure but rather in distributing responsibilities and powers, such that each part may have its own WIGs and its own autonomy in pursuing them but there is still coordination between the parts. This implies delegation (i.e., partitioning responsibilities and powers) with the correct balance between autonomy and coordination. A basic premise of the paper is that any focusing method that cannot be seamlessly embedded within such a hierarchical framework is defective. Conversely, as part of validating any method we must show that it can be embedded within a hierarchy. We do so for one method: Management by Criticalities (MBC II) – which had been presented in a companion paper [13]. MBC II is a refinement of Management by Constraints (MBC) – a term that we reserve for the correct aspects of the heuristic known as “Theory of Constraints”.

MBC involves one initial step and five main steps, the fourth of which is to elevate the system’s binding constraints. MBC II replaces this step by one that seeks economic balance between resources: when there is a distinct bottleneck (BN), it is often economical to elevate it, so there is no contradiction between the two objectives, but the new one is more explicit, and we need this explicitness for the hierarchical application. For the single throughput with multiple inputs case (where the minimal input dictates the throughput), economic balance is approximately achieved when the criticality of each part of the system – i.e., the probability it will actively limit the system – is proportional to its economic value. We also observed that when the system has multiple outputs this principle requires a minor adjustment, namely to match the criticality of each resource to its economic value as a fraction of the total throughput that depends on it. (See [11, §4] for a more thorough derivation of this adjustment, but interpret the partial derivatives as criticalities.) To continue, there are actually two parallel cycles in MBC which we made explicit in MBC II. The final structure of MBC II is thus given by:

**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
2. Maximize net present value by matching long-term criticalities to the marginal economic value of resources as a fraction of the throughput that depends on them
3. Subordinate other investments as necessary to support Step 2
4. Return to 1.

In both cycles, Step 1 requires related analysis. Repeatedly performing Step 1 of Cycle 1 (which we denote by C1/S1) provides input for C2/S1. But other than that, the cycles are separate (parallel to each other). The new Step 4 is now given by C2/S2, but sometimes it also has short-term ramifications, as in C1/S2; e.g., as described in [13], Toyota workers adjust capacities frequently by rebalancing the line upon changes in the product mix. Similarly, revenue management pricing can be interpreted as C1/S2 balancing [3]; technically this involves adjusting prices to influence demand. Notably, the need for two cycles in MBC II (and in MBC, for that matter) underscores the fact that the time horizon of Cycle 2 is necessarily longer, if only because we cannot collect information about our current criticalities without data spanning several periods. In contrast, as long as we believe that our BN is consistent, one period is sufficient to identify it. Note that the need to deal with different time horizons by itself already calls for a hierarchical approach.

Inherently, MBC II is a focusing method, calling for focus on resources whose criticality is excessive relative to their marginal value. In our present context, it is important to note that MBC II is just one of several ways to focus, and it is often possible to use several of
them concurrently and synergistically. Some, including MBC II (and MBC), provide leverage: when we focus on resources that are too critical, we utilize them better and we also utilize better all the other resources in the system. Not every focusing approach shares this strength. For example, the TQM movement is associated with focusing on customer needs. This is predicated on the idea that everybody has customers, internal or external, and their needs provide a focus. Unfortunately, this does not provide focus for the organization as a whole. For example, sometimes it is necessary to focus on a partial set of customers, and this must be done by some other approach (e.g., MBC II). Furthermore, our present customers may not be the most important potential future customers – another reason why focusing on customers is not sufficient; in fact, it is often fatal [4].

Another approach is to focus on reducing waste. If we explicitly consider both waste of resources and waste of opportunities, this casts a wide net. Indeed, my own definition of quality is the pursuit of waste reduction, including both waste types. Implicitly focusing on waste of opportunities, some preach that we should avoid the so-called “cost world” mentality, and opt for another paradigm: “the throughput world” [7]. Believers in the “throughput world” paradigm do not attempt to reduce costs but look for opportunities to produce and sell more with the resources they have. But the waste reduction paradigm includes both “worlds”. For example, neglecting the needs of potential future customers is a waste of an opportunity that zealous “cost world” practitioners can suffer from. Spending too much to meet the potential needs of each and every conceivable future customer is a waste of resources that can easily bankrupt zealous “throughput world” practitioners. Rather than promote one and neglect the other, the two must be balanced. We will revisit the need to consider the needs of the present as well as the needs of the future in Section 3. Within this framework of waste reduction, however, it is often useful to focus on processes that waste time. This, again, provides leverage, and it does so in two ways: (i) many types of waste cause waste of time as a by-product (e.g., rework wastes time), so to reduce waste of time we must often discover and eliminate other waste; (ii) organizations (e.g., production systems) are prone to harmful system dynamics that increase with delays. The Beer Game demonstrates such harmful dynamics, and the role of delay in it is well known. So, by combating waste of time we are likely to reduce at least some of these complexity-causing delays, and thus achieve leverage. When asked to describe Toyota’s approach, Taiichi Ohno said: “All we are doing is looking at the time line, from the moment a customer gives us an order to the point when we collect the cash. And we are reducing that time line by removing the non-value-added wastes” [8, p. ix].

Finally, under JIT, it is recommended to withdraw kanbans to expose the processes that cannot deliver reliably without the current buffer protection (“exposing the rocks”). This generates improvement projects because it tends to expose resources that suffer from excessive variation. But such resources are likely to manifest as bottlenecks too often, so this is just a way to predict excessive criticality. We can generalize this idea by observing that the amount of work in progress (WIP) in the system is a constraint (controlled by the number of kanbans) that is relatively easy to change, and by changing it experimentally we can study and improve the system. The generalization, then, would be by experimenting with any relevant constraints that are easy to change (e.g., the number of workers assigned to a cell). This option is implicit in our model: we employ continuous adjustment of the service levels based on the current variation causes, and changes of service levels are really changes in constraints.

In conclusion we have several potentially overlapping focusing tools: customer needs, constraints (including reducing variation on resources that cause disruptions too often), and prevention of waste of time. All should be assumed to operate. Various participants can generate candidate improvement projects by any or all of them. Selecting and prioritizing such projects is the essence of managing change and improvement, so the generation of a (continuously updated) list of candidate projects is vital. Appropriate focusing increases the quality of this list and prioritizes it. This also involves a culture where improvements that do not require significant monetary investments or the attention of higher levels are allowed on an ongoing basis [14].

Section 2 presents practical mini-case studies demonstrating the need for correct focusing within hierarchies. In one of them, MBC was implemented locally with great success, but the system as a whole failed to utilize the results, squandering potential benefits of several hundred million dollars per annum. Section 3 discusses hierarchical balancing of criticalities (HBC), which is essentially a hierarchical application of MBC II based on organizational cybernetics principles developed by Stafford Beer. Section 4 discusses measurement and control issues. Section 5 is the conclusion.
2. Mini-case studies

The following mini-case studies took place in the early 1990s, while I was employed by the US Navy (at Naval Postgraduate School). 50% of my time then was devoted to internal consulting. MBC was central to my teaching and consulting (although I had not yet realized quite how far MBC was from the official version of TOC). MBC was a very popular approach in the US Navy in those days and the Goldratt Institute had provided extensive training and consulting services. The Navy was also committed to total quality leadership (for which they adopted Deming’s teachings as guidelines) and utilized policy deployment (as taught by juran). At least in contemporaneous terms, it is difficult to think of a better combination of methods that the Navy could have selected. It should have worked very well.

**The cost of pursuing the “Zero Defects” BIGWIG.** Navy Aviation Depot (NAD) North Island had a strong total quality program – in fact, the acronym TQM had been coined there. The NAD had been a client of Deming, but earlier they had a federally mandated ZD program, and the mindset survived. One process that was required there was electro-plating worn parts with chrome, and then grinding them to the original design dimension. The plating department counted as a defect any case of plating too little, which would cause the grinding department to return the part for rework. In their earnest pursuit of zero defects the plating department intentionally plated too much and indeed they came very close to the ZD ideal. But although in those days quality was free, this ZD achievement cost quite a bit. The load on the plating shop was higher (because it took longer to plate), and the load on the grinders necessitated work around the clock. In fact, the grinders became a system constraint, and limited the throughput of these parts. Later I found out that management was aware of the problem in the grinding department, but by the time I became involved it had persisted for years. The way I did become involved was almost by pure chance. While walking around the shop looking for waste to reduce (a clear “cost world” approach, I must admit), I saw a run chart of the amount of plating displayed on a board (as part of the total quality program), so I asked why there was so much excessive plating. I discovered that the grinding department was a bottleneck only after raising the question! It was clear that the plating-and-grinding subsystem had to be improved, but the two functions were managed separately and nobody had attempted to balance them with each other. The average amount of plating required was about 0.023″ but because of process variation the actual average plating was about 0.040″, i.e., on average, it was required to grind 0.017″ off each part. Studying the minimal grinding required showed that (in the sample) it was 0.005″. The immediate and obvious improvement was to reduce the plating, while allowing occasional rework to happen, though rarely (i.e., re-treating from zero defects). We reduced the average plating by 0.006″, and still got very rare instances of rework. As a result, the load (including rework) on both the plating and grinding operations was reduced. For plating this saved about 15% (0.006/0.04), but for grinding (the bottleneck) the average load went down by about 35% (0.006/0.017) – a very welcome relief to a department that was working around the clock and still not meeting its desired throughput. Most of the remaining grinding necessary was due to plating variation, so the next step was to reduce variation. To this end, the plating setup was redesigned to improve the positioning of the anodes. This was done using single minute exchange of die (SMED) techniques [9]. The new plating setup, it should be noted, aimed to reduce several types of waste at one stroke, including setup time and anodes wear and tear. This is typical of many SMED applications. The existence of techniques that are capable of reducing several types of waste together is the key to the synergy that such methods often display. Technically, further improvements could also have been achieved by statistical experimental design methods (as inspired by Taguchi) applied to the various plating parameters (e.g., chemical concentration levels, current, and temperature). At this stage, however the whole project was stopped because the technical specifications were changed: instead of plating and grinding the use of sleeve inserts was authorized. This authorization was the culmination of a three years process which the shop had already given up on. As such, it is an example of the common phenomenon whereby waste of time causes other types of waste.

**The cost of pursuing the “Zero Inventory” BIGWIG.** In the late 1980s, the US Congress instructed the Military to pursue a policy of “Zero Inventory” (ZI). More precisely, using policy deployment, they gave the military an aggressive inventory reduction objective. They had good reason to do it, since it was shown by Japanese industry that this was very beneficial. The idea of aggressively reducing inventories as a first step on improvement journeys was also advocated by MBC experts. Indeed, it can be useful when done right, as
part of a “rock exposing” exercise. The US military is simply too large and complex a hierarchy in which to conduct such an exercise on a wholesale basis, however, and certainly not without the basic understanding that variance reduction projects must be identified and pursued as part of the inventory reduction process. Interesting results followed throughout the Navy (and probably in the other services as well). Naval Postgraduate School, for example, experienced a two months shortage of paper clips. Nobody had the combination of authority and initiative to go to the nearest stationary shop and purchase some paper clips to use until the official shipment arrived. Meanwhile, people like me, who were paid the equivalent of about 625 paper clips per minute (they were cheap paper clips though, purchased from the lowest bidder), wasted quite a bit of time due to the shortage. (Needless to say, I learned from the experience and stocked a few years worth of paper clips when they did arrive, but this inventory-increasing side effect of ZI is almost beside the point.) At the same period, according to Vice President Al Gore, about ten percent of the US Navy motor vehicles were queued in repair shops, mostly awaiting standard repair parts [6]. The vehicles themselves, however, were not counted as inventory and so they were not subject to the decree! The same also applied to aircraft and ships [10]. But, with the exception of the air wing personnel, when a ship is under repair all the sailors must stay with it instead of providing the defense readiness they are meant to deliver. With this in mind, the economic value of an average ship-day was approximately $300 000. With respect to submarines of a particular type, each had five valves, valued at $75 000 apiece, which could only be exchanged one at a time (i.e., at least four had to be installed at all times). At Naval Shipyard (NSY) Portsmouth, the standard operating procedure had been to bring five good valves to the ship and exchange them, one by one, taking a few hours. The removed valves were then refurbished within a few weeks and shelved to await the next submarine. When Congress dictated the new policy, the shipyard had to look for inventories it could reduce, and these valves were taken off the roster. Instead, quick refurbishing was prescribed. In turn, each valve was removed, refurbished and reinstalled, which added about a week to the critical repair path. The week was added to the repair time in spite of working around the clock to refurbish the valves as quickly as possible. The holding cost savings were certainly well below $100 000 per annum; the annual penalty was that about five submarines spent one more week each at the shipyard (plus a nickel or two for the overtime expense). This is equivalent to losing 10% of a submarine for a gain that is a fraction of the true value of one submarine for one day! In other words, when viewed in context, the $75 000 valves were comparable to paper clips. Ships in shipyards are very expensive WIP inventory items! As components of a larger system, ships and repair parts should be in balance.

Snatching [Industrial] Defeat from the Jaws of Victory. As in other US Naval Shipyards, at Mare Island NSY typical ship overhaul and refurbishing projects took about 20 months. Of these 20 months, 6 were spent on dry dock. During this period, the critical path was at a particular machine shop (Shop 31). Two other shops worked in parallel to Shop 31 but only required 4 months per ship. Shop 31 employed about 200 machinists, and constituted about 8% of the value of the whole shipyard. Yet it was critical 30% of the repair time. This by itself already demonstrated lack of balance, and suggests that it would have been appropriate to invest in this shop. Fortuitously, at the time I was trying to promote setup reduction methods [9], and I was given the opportunity to do so there. Under the leadership of a very talented foreman, a team of dedicated machinists reduced the fraction of time each machinist spent on setups and setup related activities from 46% to 13%. The monetary investment was miniscule: $250 000 (although it took a year to complete the purchase orders!). This implies a potential throughput increase of about 60% (instead of up to 54% productive time we now had up to 87%). But to translate this to a reduction of the critical path from 6 to 4 months required a change in purchasing practices. This should not be difficult to achieve in theory, but was not trivial at all in reality. It required intervention from the top, which was not forthcoming. Instead, the shipyard utilized the improvement by encouraging more than 50 machinists to leave. Although saving the cost of 50 mechanics is not completely trivial, this was not the intention at all. To put this in perspective, the marginal cost of operating a shipyard is much lower than the value of the ships in it. So using the improvements to reduce the payroll by 50 instead of increasing the readiness of the Navy by 2 months per ship (and several ships were involved annually), meant forfeiting almost a full ship equivalent per year for the economic value of about one ship-week: a fifty fold leverage opportunity on top of the similar leverage that was already achieved! Considering similar opportunities at the other shipyards, the lost opportunity was in the order of magnitude of 0.75% of the 1990
Navy budget (i.e., approximately 0.75% of $100 billion), while the savings were firmly within the round-off error (and the expense completely negligible). One reason for this blatant waste may be because reducing repair time would make the apparent utilization of the shipyards lower, and thus encourage Congress to close some of them. This problem could have been prevented if decision makers both in the Navy and in Congress would understand that high utilization implies high criticality (low service level); instead, they measured utilization only. Eventually, Congress closed three of the eight shipyards anyway, including Mare Island, but the repair time in the other shipyards became even more excessive (higher queueing).

Ignoring the ultimate failure of this case, and focusing instead on its potential, consider how such an improvement opportunity might be identified. In terms of focusing, it could have been generated either by (i) the machine shop personnel, focusing on processes that waste time (which was indeed the case); (ii) the shipyard management level, studying the repair project’s critical path constraint; (iii) the staff function in charge of shipyards (and indeed they were pivotal in facilitating the project and disseminating the technical lessons learned at other shipyards); or (iv) by a yet higher level, focusing on the readiness of the Navy subject to its budget constraint. But unless there is someone at the top who at least understands the implications, the first three are likely to lead to frustration – as was the case indeed. Be that as it may, in all these cases one could also couch the arguments in terms of serving a customer (e.g., readiness is the customer need that the Navy satisfies for the nation). Thus, this example also shows that the different focusing approaches often lead to the same end result.

3. Hierarchically balanced criticalities (HBC)

Consider the root cause of the problems identified in the mini-case studies. In the electro-plating case people knew that there was a tight constraint at the grinding department, but they did not see that the root of the problem was a misplaced commitment to ZD at another department. There was no attempt to achieve balance in the system or to question the definition of quality the plating department used. The (ultimately successful) solution management sought instead – using inserts – was outside their immediate zone of authority to implement. The hierarchy was such that those who could authorize this solution – far removed in the hierarchic structure both in terms of geography and level – did not realize the full cost of the delay in their decision. Although this is speculation, I doubt very much whether they have indeed prioritized their work and took three years because they were dealing with even more urgent cases first. But if my speculation is wrong, then I would have to say that their capacity was certainly too critical. The ZI-related cases were again the result of pursuing an absolute objective without balance, but even more importantly, they were the result of high hierarchical levels (Congress) usurping powers that should be vested far below them and dictating decisions that they should never have interfered with. To meddle in the decisions of more than two levels below is a sure recipe for waste and discontent. One also wonders who gave Congress the idea that ZI was useful. Considering that business management experience is not required for being elected, I strongly suspect that the source was some bright staffer hired straight out of school, perhaps with the help of a professor. To continue speculating, I suspect that those involved don’t even know that they erred, let alone how huge their error was. Neither MBC, nor TQM or policy deployment helped. In fact, what we have here is a case of a policy being deployed in the prescribed manner and based on what must have been considered sound practical knowledge. In the Mare Island case the system proved that it could neither measure the true value of time and other resources nor coordinate any sort of useful balance between its parts. Presented with a goose capable of laying golden eggs, they ate it.

What we need, instead, is a system that uses balance as the guideline. For example, at North Island, the correct balanced number of under-plating defects was much higher than zero, and we achieved a major improvement by allowing about 1% rework. With a bit more effort, the optimal conformance level could be identified, but the real major benefit was by just recognizing that extreme prescriptions tend to cause extreme waste. With balance in mind, the opportunity would have been identified easily. And since the plating-grinding subsystem was easy to identify as too critical, one would assume that a better hierarchical management approach would have identified the need for better plating setups taking less time and yielding lower variation. The conclusion is that we need to deploy policy in a different way, delegate authority and autonomy to the lowest possible level, where it belongs, and yet maintain a way to measure that the decisions we make are good for the system as a whole.

Now consider the extent to which MBC is likely to provide the solution (which, for the US Navy, it cer-
One problem with the original MBC cycle that remains in MBC II is that the subordination step actually refers to a lower level in the hierarchy and does not really belong on the same cycle. For the same reason MBC and MBC II do not emphasize that activities at the level of the “system in focus” [2], i.e., the level we are managing, often have to respond to similar needs that come from above in the hierarchy. Furthermore, by their nature, the activities in Cycle 1 are distinct from those of Cycle 2. So MBC and MBC II – at the very least – require a more explicit approach to delegation of responsibility and authority. This is our challenge, and we approach it by using insights discovered by Stafford Beer [1].

Beer’s framework – namely his viable system model (VSM) [2] – is based on the neurophysiologic model, whose validity and success has been proven by natural evolution. The model is nested, which means that at whichever level we may be looking, we see a similar picture: our system, also known as the system in focus, is always a subsystem of a larger system and always has subsystems of its own, all with the same type of connections and controls. The system in focus is called System 1 and includes the management function and several (typically 2 to 9) isomorphic subsystems also known as System 1, but at their level. Communications and command links connect management and the subordinate subsystems. Within management, Beer distinguished four additional systems – systems 2 through 5 – each responsible for a different aspect of managing System 1. Of these, System 2 is a subsystem of System 3, System 4 operates in parallel to System 3, and System 5 coordinates Systems 3 and 4. Specifically, System 2 is in charge of coordination and prevention of harmful conflicts. For example, scheduling, kanban and CAPWIP are System 2 applications. So are rules of the road and police. System 3 is the operations manager, and in Beer’s words System 3 manages the “here and now”. System 4 provides strategic management functions, focusing on the “outside and then”. That is, System 4 is always on the lookout for new technologies (i.e., new ways to continue doing what we are already doing and serving the customers that we already have) and new markets (new products and new customers). Because Systems 3 and 4 may be in conflict (3 tends to prefer the status quo and resist change, while 4 may be too focused on the future), System 5 is required to make final calls when 3 and 4 fail to agree. For a large firm, System 5 is the board of directors; for the US military, System 5 is Congress (although, as we have seen, Congress sometimes wanted to act as System 3). Finally, System 3 sometimes exhibits a “cost world” attitude while System 4 can be overly committed to the “throughput world”. But according to Beer these two functions should be balanced.

MBC II fits in here as an explicit model for the optimization of desired levels of the criticalities, or, equivalently, the service levels within the VSM, i.e., at the embedded System 1 units. More importantly, the model should be used to prioritize and motivate improvement projects at any part of the organization where gaps exist between the desired service level and the actual one. Beer’s own solution for that need involved monitoring utilization figures, but we suggest monitoring criticalities (or service levels) instead. To be sure, service levels and utilization are closely related, but measuring balance by utilization may lead to suboptimization, while the optimal criticality is robust.

For example, MBC II involves monitoring criticalities of parts of a system. It is then natural to apply the Pareto principle and focus on the most critical ones. But if we focus on the most critical constraints we necessarily must delegate the management of the less critical ones and this creates a hierarchy. In terms of the VSM this hierarchy involves embedded System 1’s within one’s system in focus. For example, take project management. The project manager always focuses on the activities that are most likely to be on the critical path, so she must delegate the other activities. As a result, with or without a formal hierarchy, the project is managed hierarchically. To prevent the need to transmit excessive amounts of information, data is strongly filtered (attenuated) before being sent up. This can be done, say, by the use of control charts to report only signals that are out of control; i.e., management by exception. Conversely, directives should be fleshed out (amplified) at the lower levels when they are executed. For example, using the model of this paper, the management of the system in focus may direct one of its subsystems (at the lower level) that inventory availability service level (SL) should be 92%, but the lower level amplifies this directive by specifying service levels of, say, 90%, 91%, and 95% to the three main types of inventory, such that the average SL will be 92% overall. A similar procedure is repeated at the sub-subsystems level, dealing, perhaps, with actual inventory items in the various inventory types. In this case, the overall service level is the weighted average of the parts – a case to which [14] refer as additive. Using the economic balance principle of MBC II, we would assign the higher service level to items that are cheaper to hold and allow lower service lev-
els in more expensive items (so paper clips would justify very high service levels). In contrast, the model of [12] is multiplicative because the system service level is the product of the parts. As a multiplicative example, suppose we require from a machine a service level of 0.91, then we have to allocate the “permission to fail” (PTF) – i.e., desired criticality of 9% – to the subsystems correctly. For instance, the PTF may be divided to 6% allowance for having too little capacity (i.e., capacity SL = 0.94) and 3% for being down for repairs (i.e., uptime SL = 0.97). To achieve the latter we may then restrict the criticalities of maintenance replacement parts for the machine. Thus, the hierarchical allocation of PTF helps determine maintenance replacement parts stock levels. Similarly it drives decisions at other subsystems. Note that, at least for the multiplicative case, smaller subsystems always receive less PTF than the bigger ones that are hierarchically above them, and as we go down the hierarchy the optimal service levels typically rise to the high nineties – but this is not the case for higher levels. Thus, the intuition that the same service level should prevail everywhere is undermined by this model.

The management of the system in focus will rarely, if ever, be informed of the status of individual inventory items or repair parts: This can happen only if there is a major problem associated with one of them that cannot be handled at the lowest or next-to-lowest level. Thus, the lower level has autonomy over the decisions under its control, as long as its performance does not imperil the performance of the whole. Proper autonomy is one of the main issues that Beer emphasized: Lack of autonomy prevents the lower levels from carrying out their function effectively; and the higher levels simply cannot handle the decisions for the lower ones. The allocation of PTF is useful for defining the boundaries of the autonomy. If the lower level does not conform to their PTF, the upper level may intervene. Since the upper level has to prioritize its own actions, it may take a quite substantial deviation for such an intervention to occur. But because the system is not sensitive to small deviations, this approach is sound both for the whole and for the parts.

Essentially, we can always group related activities or resources together and manage them as a subsystem. This by itself creates a hierarchy. The essence of the ideas above, to which we refer as hierarchically balanced criticalities (HBC), is that each subsystem can be assigned its optimal or near-optimal permission to fail (PTF) by the level above it, and it can in turn assign PTF levels to the parts that compose it. This process can be repeated at the lower level in the same manner. The result, HBC, is an application of MBC II to the hierarchy as a whole. Note now that this can be done much more explicitly than would be the case with MBC. Finally, this constitutes an improved version of policy deployment.

Figure 1 depicts HBC by showing the MBC II cycles of the system-in-focus level and the subordinate level. Step 1 of both cycles is in the center of the figure and the numerals indicate which VSM systems are involved. Thus, Cycle 1, depicted at the bottom half, is identified as a System 3 activity (“here and now”), while the upper half is associated both with System 4 (since it involves planning future activities) and System 3 since these are relatively routine investments and also because it is System 3 that actually manages them. System 4 may also use the same information to come up with “out of the box” strategic projects that may remedy some lack of balance, e.g., by introducing an innovative technique with a higher capacity (such as improving the plating setups or making it redundant by using inserts instead), or a new product with an additional market. Depending on the circumstances, System 4 issues may be crucial or tangential within this particular framework. Note also that the short term (Cycle 1) and the long term (Cycle 2) have a System 5 conflict resolution inserted between them. Horizontal arrows and double arrows indicate exchanges of information, negotiations, and directives. These horizontal arrows start higher up in the hierarchy, i.e., from the left of the figure, and continue to the lower levels, towards the right in the figure. Thus the structure includes the two cycles within a hierarchical structure.

This is the essence of HBC. Finally, it might be more intuitive to depict the levels vertically instead of horizontally, but the horizontal depiction stresses the fact that the hierarchy in question does not necessarily reflect rank or importance. For example, the same type of hierarchy obtains in a supply chain regardless of the status of the vendors involved. The orientation of the figure helps convey this fact. We need a detail-handling hierarchy here, not a pecking order.

MBC II dictates how to actually decide the correct PTF levels; i.e., by adjusting criticality proportionally to the marginal cost of capacity as a fraction of the economic value of the throughput that depends on the resource. In the next section we discuss how to monitor criticalities to ensure the desired PTF levels are achieved (at least approximately), and how to use deviations between criticalities and PTF levels to focus on improvement projects.
4. Managing criticality for machines and inventories

Assume we have $n$ resources in our hierarchical structure. The basic idea of MBC II is that Resource $i$ (which we denote by $R_i$) has a prescribed permission to fail, denoted by $PTF_i$ (where $i = 1, \ldots, n$). $R_i$ also has a measurable criticality, $p_i$. When $p_i \approx PTF_i$, $R_i$ requires no changes (we should not focus on it). Thus, we should measure and monitor $\{p_i\}$. Due to scope limitations, we ignore the question how much evidence is required to determine that $p_i$ is indeed far enough from $PTF_i$ to justify adjustment. Nonetheless, we should note that if $PTF_i$ is low – which is likely for large $n$ – then it is easier to identify cases where $p_i > PTF_i$ than where $p_i < PTF_i$. Therefore, our heuristics are more robust for balanced growth decisions: the necessary signal for a positive investment is $p_i > PTF_i$. The measurements should cover the criticality of machines (or similar) resources, workers, raw materials and WIP inventories (to avoid starving critical machines too often), physical space for WIP between machines, maintenance resources, financial resources to support these activities, and so on. Similarly, in a produce-to-stock environment, sales require FGI as a WIP to produce completed sales. All these can be modeled as resources at various hierarchical levels. We discuss two specific resource types. Most machines, as well as raw material stocks, process or serve items one at a time. For example, we must draw stock items for production and, depending on stock capacity, the next item will be available or not. We refer to such resources as one-on-one. In contrast, WIP, flexible workers (which are similar to WIP), financial resources, etc’, may shift to different locations and may serve several operations concurrently.

4.1. The criticality of one-on-one resources

A practical way to measure $p_i$ is by monitoring queues, estimated in time-to-clear units, operationally
defining the critical resource as the one with the longest queue. This requires taking a random sample of system observations and nominating the resource with the longest queue as the critical one on each observation. The frequency of observing a resource as critical this way is an estimator of $p_i$. If an observation reveals more than one queue with roughly equal length, we can allocate a fraction of the count to each resource based on their $PTF_i$ values. Such an allocation reduces the likelihood of investing in the wrong resource. This, however, is most suitable for a system with a single output. If there are several outputs than it is possible that distinct resources will be critical for each of them at the same time. Thus we need to monitor criticality for each output separately. Weighing each output by its economical value, the criticality of a resource is the weighted average of its criticality for each output. And it is this criticality that should be matched with the PTF.

Usually, queues are easy to monitor, e.g., by sampling methods. But two cautionary notes are in order. First, due to volatility, we must be prudent and avoid excessive or abrupt responses. A typical example was reported in May 2003, concerning the public health system in Christchurch, New Zealand. (New Zealand has a public health system supported by optional private insurance. *De-facto*, public service is rationed: patients, especially with minor ailments requiring elective treatment, may have to wait very long periods. But the government is committed to providing prompt service where it counts. This example concerns such a case.) Suspected cancer patients have been waiting about three months to see oncologists for diagnosis and treatment specification. In response, two new oncologists were hired, and the queue time dropped significantly. However, the queue in the radiology department quickly grew by almost as much as the oncologists’ queue was reduced so the net improvement was very small. The balanced approach calls for increasing the two resources—which evidently had comparable capacity initially—in concert. Nonetheless, initially the queues ahead of diagnosticians were larger, so the evidence is that they probably had a slightly smaller capacity and/or higher variability. The conclusion is that queueing evidence is only valid for small expansions. (A simulation model could have helped in this case.) Second, we must include invisible or conceptual queues, e.g., lost sales (unmet demand), unsatisfied purchasing orders, or a queue of investments waiting for funding. For example, one of the queues that could have identified grinding as a bottleneck in

the North Island case consisted of lost sales. We must also take into account that in the presence of blocking, a queue may extend across several resources, and we must be careful not to count all of them as causing it—only the one resource that is causing these consecutive blocking queues is truly critical. It may be downstream or on another line. In general, a constrained resource can produce parts whose shortage causes starvation and blocking elsewhere in the system. This is revealed by queues of incomplete kits awaiting assembly with missing critical parts.

### 4.2. The criticality of WIP

Because multiple one-on-one resources interact with WIP (or with similar resources such as flexible workers, financial budgets, etc.), it is not clear whether starvation of any particular one of them at a particular time will translate to throughput loss. To outline a possible measurement for WIP, assume a make-to-stock environment with a total cap on the WIP and FGI combined (similarly to CAPWIP—[13]). We treat these two as one because FGI is simply WIP ahead of the sales resource. Thus, the cap defines a single WIP inventory resource, which we denote by the index wip. If we measure the frequency of starvation of any one-on-one resource, the result is an upper bound on $p_{\text{wip}}$. We propose using the minimum of these upper bounds as our estimate of $p_{\text{wip}}$. The rationale is that this minimal bound measures the frequency of starvation of the current one-on-one BN, and that’s when the WIP resource is critical.

In practice, we cannot adjust the capacity of all parts of a system by fractional amounts at all times. We must select one or few resources to elevate (focusing). Limiting ourselves to expansions only, it is most useful to focus on the resource that maximizes $p_i/PTF_i - 1$. For example, Shop 31 at NSY Mare Island had $p_i = 0.3$, $PTF_i = 0.08$, and thus $p_i/PTF_i - 1 = 3.75 - 1 = 2.75$; i.e., the net present value of profit by investing in it was 275%. (Because we did not actually increase capacity but instead eliminated waste, we achieved a much higher return, more akin to 5000%. As for the opportunity that the Navy squandered, it was roughly 3600%, so the combination should have yielded 180 000%.)

If we allow disinvestment, a resource with high $PTF_i/p_i - 1$ is most attractive to reduce. Now, consider that it may be possible to find new products to sell that require more than their fair share of resources with high $PTF_i/p_i - 1$ and not much on those with high $p_i/PTF_i - 1$. Such products increase balance in
the system; in effect, selling such products is equiva-
lent to “selling” the abundant resource without having
to actually downsize, thus avoiding the political diffi-
culties that downsizing entails.

5. Conclusion

MBC II allocates the optimal criticality (permis-
sion to fail – PTF) to subsystems and – thus achieves
economic balance. Subsystems, in turn, have sub-
subsystems and the criticality of a subsystem can be
calculated as a function of the criticalities of sub-
subsystems. Therefore, the PTF of a subsystem can
be allocated to the sub-subsystems by MBC II prin-
ciples. This creates a hierarchical framework for bal-
ancing criticalities, and thus provides a rational focusing
mechanism. The result is a data-based policy de-
ployment framework, replacing arbitrary top level de-
cisions or equally arbitrary participatory negotiations.

References

[3] P.C. Bell, Revenue Management for MBAs, ORMS Today (Au-
[6] Gore, Vice President Al, From Red Tape to Results: Creating
a Government that Works Better and Costs Less, Report of the
National Performance Review, US Government Printing Of-
the Theory of Constraints, St. Lucie Press, Boca Raton, FL,
1998.
[8] T. Ohno, Toyota Production System: Beyond Large-Scale Pro-
[9] S. Shingo, A Revolution in Manufacturing: The SMED System,
Working Paper 92-06, September, Naval Postgraduate School,
auckland.ac.nz/staffpages/dtrietsch/HBC.htm.
Plant Expansion, Or Improvement Projects, in: Proceedings of
the 32nd Annual Conference of ORSNZ, 1996, pp. 93–
auckland.ac.nz/staffpages/dtrietsch/HBC.htm.
[12] D. Trietsch, “Balancing Resource Criticalities for Opti-
mal Economic Performance and Growth”, MSIS, Uni-
for non-commercial use at http://staff.business.auckland.ac.nz/
staffpages/dtrietsch/HBC.htm.
[13] D. Trietsch, From Management by Constraints (MBC) to Man-
agement by Criticalities (MBC II), Human Systems Manage-
[14] D. Trietsch and J. Buzacott, “Managing Change and Improve-
ment by Balancing Service Levels Hierarchically”, MSIS, Uni-
for non-commercial use at http://staff.business.auckland.ac.nz/
staffpages/dtrietsch/HBC.htm.