Managing Distribution According to TOC Principles

By Amir Schragenheim

The current practice of managing supply chains

It is Wednesday afternoon. I'm entering the grocery store and want to purchase some green peppers. However, they don't have any on stock. I can't find any good looking tomatoes either. I'm continuing to the office depot store. I heard great reviews about a new mouse that Microsoft issued and I would like to get one. However, I come to an empty shelf with only the item description stating "out of stock".
How many times did you go to a shoe store, tracked a wonderful pair of shoes you wanted to purchase but they didn't have any in your size?
Why do stores don't keep the right stocks to fulfill the demand? Why can't they do anything right?

Supply chains in our modern age operate in a way that seems to make a lot of sense. Manufacturers have robotic machinery to automate processes; many manufacturers operating nowadays have already installed new state-of-the-art ERP systems to help them manage their shop-floors.
Distributors and manufacturers have very sophisticated forecasting software to predict exactly how many items will be sold of each product or SKU (Stock Keeping Unit). Therefore, they should know how many units they would like to send the consumption points (retail stores) and when.

How is it that organizations still experience problems in managing the supply chains? Is technology not enough?

The natural tendency for push behavior

What is the manufacturer/distributor point of view when he's deciding on how much stock to keep at each location? He has two main questions in mind:

- How much to keep upstream the supply chain?
- How much to keep downstream the supply chain?

The natural tendency is to keep the stock as close to the consumers as possible - if a product is not at the consumption point, then there is a (much) smaller chance the item will be sold. Only a few consumers would let their vendors ship the product to them in a few days instead of taking it right away – immediate consumption is the name of the game. Therefore, it is only logical that the manufacturer/distributor should keep most of the stock as close to the consumer as possible – as far downstream as he can manage – usually at the retail level.
This is a typical push behavior: pushing the products downstream in order to increase consumption. However, the push behavior requires a good forecasting model, in order to predict where and when the stocks will be needed at the stock locations.

Why is it impossible to find a good forecasting model?

The advanced forecasting modules existing today try to model the demand and create a good answer to the availability question: What to hold at which place and when. However, the forecasting mechanism, no matter how good it is, cannot really predict what the demand would be like. Doing very accurate market researches might give

some answers, but one must at all times consider some facts of life regarding statistics.

The first fact is that the narrower the aggregation, the worse the answer becomes – meaning that the question of "how much will be sold from the product overall?" will yield a much better answer than the question: "How much will we sell from the product at this specific location?" This phenomenon stems from the fact that fluctuations average out on the aggregated events (assuming they are independent events). If we predict the sales at 100 different locations, we might get an answer that sales in an average location will range from 10 to 25 units a day. If we ask the same question on the overall quantity that we need to manufacture, we will get a much more accurate answer – probably something like ranging from 1650 to 1850. If we would just take the lows and highs of each consumption point and aggregate them we will get a much worse answer – from 1000 to 2500.

The second phenomenon is the wrong interpretation of data - people using statistics must have good understanding of the aggregation mechanism. There are some large mistakes being carried out on a daily basis all over the world, because of lack of understanding of statistics. For example – a clever man but not experienced in statistics might deduct from the example above that the consumption will be between 1650 to 1850 for all consumption points that each consumption point will have a consumption between 16.5 to 18.5 – keeping 18 units for each location and running out of stock in a fairly large number of them, while others will be left with a lot of stock they can't sell. The fact that we got an aggregated sum does not mean that it can be applied to the points that make out this sum. Another man might suggest protecting availability by putting 25 units at each location – increasing substantially the investment and increasing substantially the number of consumption points in which we'll have excess stock – taking unnecessary space and investment. The more sophisticated the algorithm, the more sophisticated the end user has to be in order to use correctly this algorithm

Another problem is that no forecasting model can take into account sudden change in consumption patterns. An example might be a very enthusiastic article in a paper (or vice versa) that suddenly changes the consumption pattern in a whole region. In today's dynamic market such event are becoming quite frequent.

As the forecast of a single SKU at a specific location is subject to the above mentioned impacts of fluctuations and uncertainty, it is a very poor base for determining the required stock level of that SKU at that specific location. It's clear that another mechanism is needed in order to reach this decision.

The TOC way – pull distribution

The Theory of Constraints (TOC) analyzes the impact of supply together with the demand over the management of the supply chain stocks, with an emphasis on the supply side. If it is possible to respond in an instant to demand, there is no need to rely on forecast at all! While this situation is of course unattainable in almost all business environments, a step in this direction should be considered. In the case of keeping the right amount of stock in the supply chain, the objective is having very good availability of the items at all the consumption points. This objective is limited by the

availability of cash and space, which means that it's impossible to keep high stocks of all items at all locations, even when obsolescence is not an issue. No only that, but also as will be explained later in this article, keeping too high stocks of low demand SKUs will lower the sales overall.

The TOC solution is based on constant renewal of the consumed stocks, and is comprised of several steps:

- Aggregating as much as possible at the source – the plant or central warehouse – setting a high inventory target there (called Stock Buffer Size)
- Determining inventory targets at all stock locations (Stock Buffer Sizes)
- Enabling the transfer of real consumption data from all stock locations
- Shortening the replenishment time as much as possible
- Replenishing as frequently as possible from the main (plant or central) warehouse to the consumption points – units are shipped only in order to replenish to real consumption (or to readjusting of buffer sizes)
- Monitoring the buffer sizes according to consumption and readjust them accordingly

Aggregation: Building a Plant/Central Warehouse

The important part of the proposed model for managing a supply chain is to keep the stocks at the divergent point – where the stocks can be used to serve many different destinations, and using a pull mechanism from the destination to replenish. This method guarantees we keep the lowest stock possible to support the demand of the various consumption points.

In order to have the product available at different locations – it is recommended to aggregate the stocks at the source and build a plant or central warehouse (PWH/CWH). If the organization is a manufacturer, the entity is called a Plant warehouse (PWH) as this is the finished goods warehouse of the plant. If the organization is a distributor, the entity is called a central warehouse (CWH).
In this warehouse we keep **most of the stock**. According to the principles of statistics, this aggregation guarantees a more stable system than keeping it at the different consumption points. At the consumption point the amount of stock is very limited. Once a certain consumption point sells a unit – the consumed unit will be replenished as soon as possible from the PWH/CWH.

When the transportation time from the PWH/CWH to the consumption points is very long – a regional warehouse (RWH) might be needed between the PWH/CWH and the consumption points. A regional warehouse will behave as a consumption point to the PWH/CWH and as a central warehouse to the consumption points which it is serving. This is just an extension of the TOC model and all of the assumptions and considerations remain the same – the idea is still to pull from the PWH/CWH only based on consumption from the RWH.

The Replenishment Lead Time and how it can be managed

The size of the needed stocks at the different locations is dependant upon two totally different factors:

- Demand – this is the factor that affects the rate at which the stock is depleting from the different consumption points

- Supply – this is the factor that affects how quickly the consumed units can be replenished

Amazingly enough, the supply factor is usually ignored in tactical and strategic decision making. Most efforts for improvement are directed at the demand side – especially trying to come up with more sophisticated forecast algorithms.

The replenishment lead time (RLT) is defined as: The time it takes from the moment a unit is consumed until it is replenished from the previous link in the supply chain. The RLT is comprised of 3 different parts:

- Order Lead Time – this is the time it takes from the moment a unit is consumed until an order is issued to replenish it. In other words, this is the **frequency of ordering of the same SKU**
- Production Lead Time – this is the time it takes the manufacturer/supplier from the moment he decides to issue the order until he finishes producing it
- Transportation Lead Time – this is the time it takes to actually ship the finished product from the supplying point to the stock location

TOC suggests challenging all of these 3 elements in order to cut the Replenishment LT to a bare minimum. By cutting the RLT, the supply side factor is becoming less dominant, and the following is achieved:

- The needed stock levels at the consumption points (and at the WHs) is lower – since it needs to cover for less demand days
- The fluctuations in supply time become smaller as the supply time decreases
- The needed forecast for new product's sales is much more accurate – since the forecasting error becomes larger as we need to forecast more into the future (the trajectory becomes wider)
- The ability to respond much quicker to actual demand is apparent

The TOC principles direct us to find ways to trim the different elements of the RLT. These are the general guidelines:

- Order Lead Time – if possible, cut it to be 0 – usually meaning trying to replenish daily from each consumption point what was consumed that day. more considerations involved will be covered later in the article
- Production Lead Time –Simplified DBR (the TOC methodology for managing production shop floor) should be implemented and the priority of the manufactured parts should be tied to their stock level at the plant WH – this will be elaborated further later in the article
- Transportation Lead Time – try to see alternatives for transportation – for example daily trains or ships instead of weekly, or flying some parts by airplanes. Finding closer suppliers for RM or purchased parts is also a possibility in a lot of cases. Usually this is the part of the RLT that one can do the least about, so every possibility needs to be checked

Frequency of replenishment versus shipment costs

When applying the TOC solution for managing distribution in the supply chain, some factors are relevant when considering how high the frequency of delivery should be.

The current practice of managing a supply chain is to ship in large bulks. The main reasons are:

1) Usually a discount is offered to large quantities of each item ordered. This discount might be negotiated to be offered for large quantities ordered over a period of time – this way one can order frequently and still enjoy the discount, but it is not always possible (although becoming quite standard nowadays)
2) There is a certain effort in listing all available inventories and issuing orders even for a small quantity
3) Some items can only be shipped in bulks because of transportation issues – fragile items sometimes can be better protected if shipped in a whole container

There is a tradeoff between the additional cost one might invest in raising the frequency of shipments and the cost of having lower availability – by making the frequency of delivery higher – a better availability is created whereas the cost of shipments is increasing. By making the frequency lower – one will have to pay with either lower availability than possible or with higher inventory levels kept at the consumption points. In most cases the extra cost will be dwarfed by the additional revenue produced.

What are buffers and buffer penetration in a distribution environment

The TOC logic is to define a safety and constantly monitor how the safety is being used. This safety is called a **buffer**. In a distribution environment, the quantity we would like to keep at the stock locations (including the PWH and RWHs) is defined as **buffer size**, and this is a **stock** type buffer. The buffer size in a distribution environment (Make to stock Buffer Size) is the number of units one would like to keep overall in the supply chain for this stock location from this SKU. For example – if the stock buffer size is 100 units and currently at the stock location we have 40 units, we expect 60 units to be on order or on the way from the feeding stock location to this one (the feeding stock location for the PWH is the plant). If those 60 units are not on the way, a replenishment order of 60 units should be issued immediately.
Note: different stock locations will have different buffers for the same SKU, since the supply and/or demand pattern might be different between them.

Buffer penetration is defined to be the number of missing units from the buffer divided by the buffer size. For the above example the buffer penetration for the stock at site is 60% ((100 – 40) / 100). The buffer size is divided into 3 equal zones. The buffer penetration sets the color of the buffer according to the different zones:

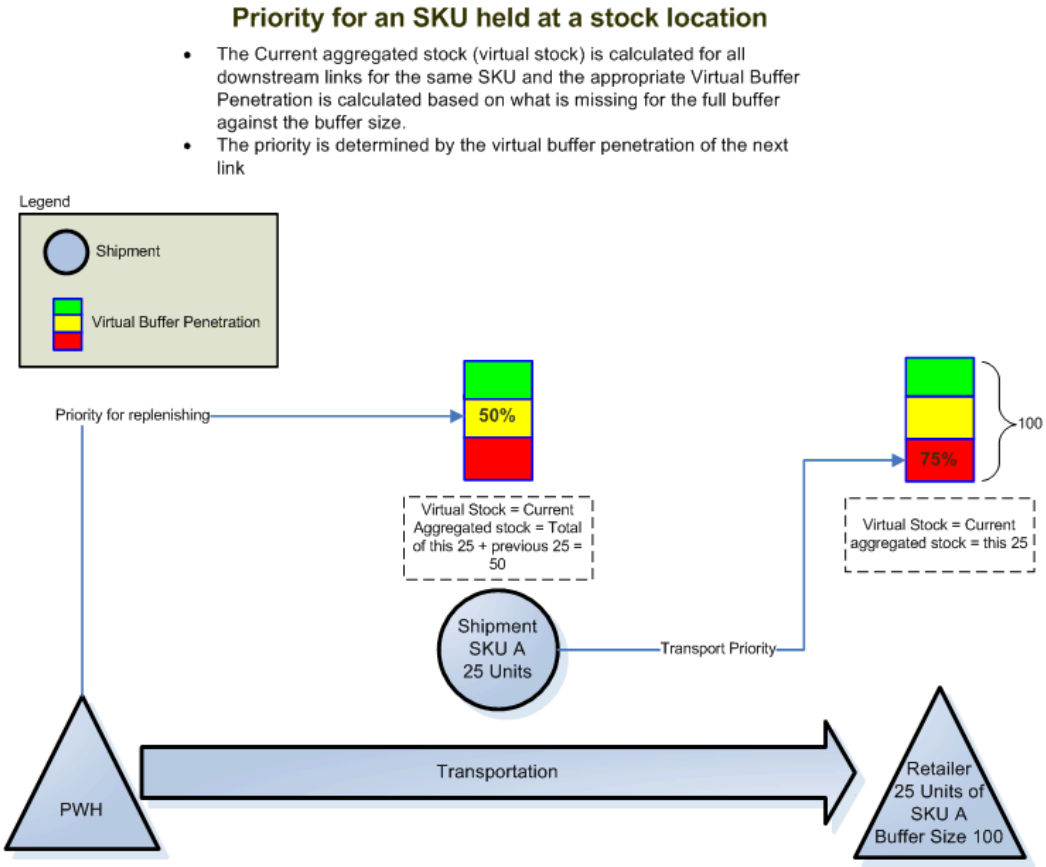
- less than 33% buffer penetration: Green
- Between 33% and 67% buffer penetration: Yellow
- Between 67% and 100% buffer penetration: Red
- 100% buffer penetration: Black

The buffer penetration color gives an indication regarding the urgency of replenishing this stock:

- Green – the inventory at the consumption point is high – providing more than enough protection for now
- Yellow – the inventory at the consumption point is adequate – there is a need to order more units from the upstream supply chain

- Red – the inventory at the consumption point is at risk of a depletion – units in transport/manufacturing (depending on which consumption point it is) should be considered for expediting efforts and an urgent replenishment order must be put to the supplying source if nothing is available on the way to the consumption point
- Black – the stock has run out at the consumption point, meaning every hour passed at this stage is lost sales opportunities – this situation must be resolved ASAP as it represents real damage, especially at the most downstream links in the supply chain

There could be several buffer views on the same buffer - we at Inherent Simplicity call it the **Virtual Buffer** concept. Let's examine the following:



The buffer size for this SKU at the stock location is 100 units. We have 25 units in stock at the stock location, and we have a shipment on the way from the PWH to the stock location for 25 units. You can see above the stocks on the way their virtual buffer penetration – taking into account the aggregated stock of downstream stocks. Their priority is determined by the Virtual Buffer Penetration of the next downstream stock.

The virtual buffer penetration gives us a very powerful tool – we have only one measurement with different views, but all the decision makers involved in the supply chain can get their priority according to their need:

- The manager of the stocks at the stock location can see clearly that the priority of this SKU is red (75% Buffer Penetration) – he needs to find out how to get more stock of this SKU ASAP

- The transportation manager can get the priority of the shipments – what shipments need to be expedited – in this case the only shipment need to be expedited (75% buffer penetration)
- The Plant WH manager can get the replenishment priority of this SKU at this stock location – in this case he needs to replenish 50% of the buffer size of SKU A in this stock location and the priority of this replenishment shipment is yellow (50% buffer penetration)

Dynamic Buffer Management

TOC aims at very simple straightforward methods to use, in order for the people using it to really understand it, and therefore does not want to use very sophisticated modules of forecasting. Earlier in this article the problems such modules create were presented – it requires deep understanding of statistics in order to use them correctly. The TOC logic dynamically measures the actual usage of the stocks and readjusts the inventory levels accordingly. This method is referred to in current TOC literature as **Dynamic Buffer Management (DBM)**.

By monitoring the buffer penetration at each stock location for each product, we can identify whether the buffer size that we keep to this product at this stock location is about right. The Dynamic Buffer Management (DBM) approach argues that by monitoring and adjusting the buffer sizes we can easily come to the "real" stock we need to keep at the site in order to cover for the demand, taking into consideration the supply side (how fast we can deliver to the stock location).
The DBM looks on two different occurrences – one is whether the buffer size is too large and the other is when the buffer size is too small.
When trying to measure whether the buffer size is too high, the indication is when the buffer penetration at site of a stock keeping unit (SKU) in a certain stock location has been Too Much in the Green (TMG) – meaning being in the green for several consecutive days (green check period – usually equal to the replenishment time). This means that we have a too high buffer to support the demand, at least for this time period, which suggests several alternatives:

- Demand has gone down
- The supply side has gone a major improvement
- The initial buffer size was too high
- Demand fluctuates severely (and then the green check period should be enlarged rather than the buffer be decreased)

The default recommendation for handling the too much green is to decrease the buffer. The basic principle says that we decrease the buffer by 33% when we need to, but this is a guideline and depends on several factors:

- How fast we want to lower inventories once we see that demand is going down
- How risky/important do we think this SKU is
- How risky/important do we think this stock location is

A very similar mechanism is used for determining whether the buffer is too low – determining whether this SKU in this stock location has been Too Much in the Red (TMR). However, the algorithm is usually different, since in this case we would like the algorithm to be very responsive to depletion of stock, not like in the too much

green case in which we would like to take his time and play it safe. The most basic algorithm for the TMR is to determine whether an SKU is in the red for several days (usually using the replenishment time). The more advanced ones take into consideration also how deep into the red the inventory at site dropped to.
The reasons for being in the TMR:

- Demand has gone up (the preferred reason)
- The supply side has gone through a deterioration
- The initial buffer size was too low
- Demand fluctuates severely

The general treatment in a TMR indication is to increase the buffer, the default being by 33%, and again this is just guidelines and each case is different then the rest.

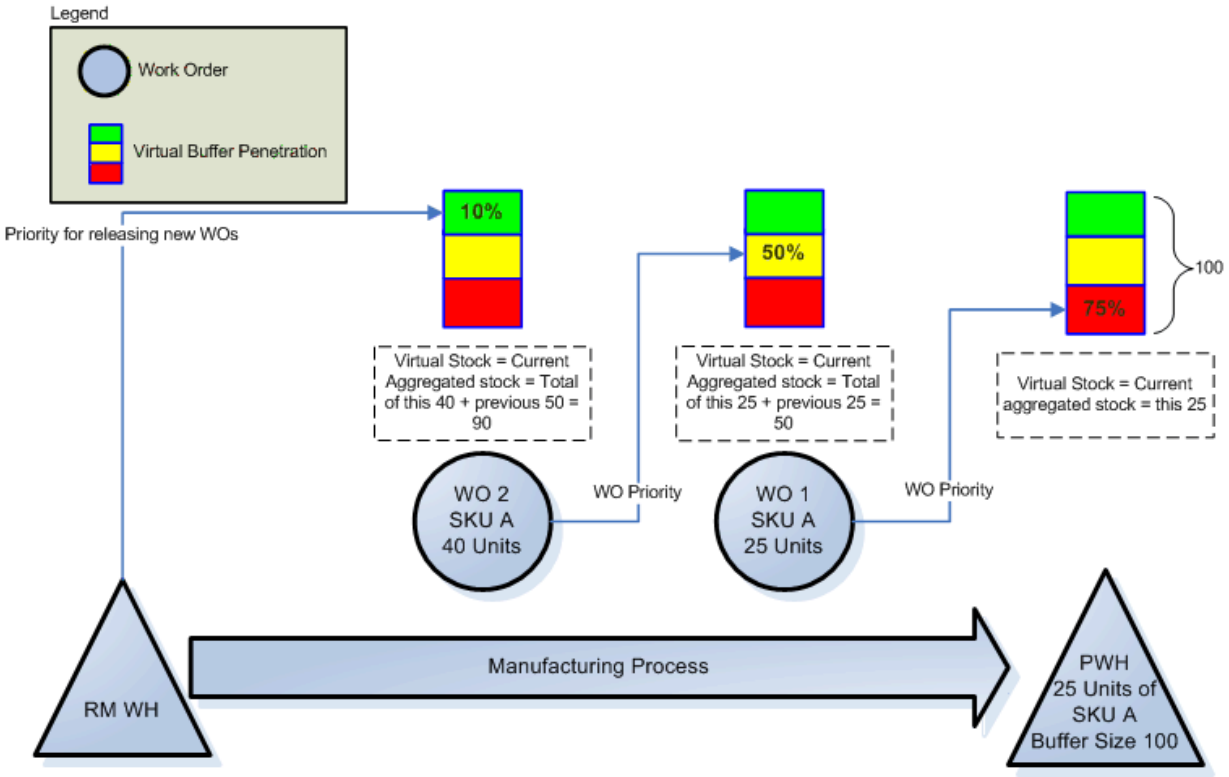
After adjusting the buffer, the SKU needs to get into a "cooling period" in which no buffer suggestions in the same direction are given (until the system adjusts to the revised buffer size). This cooling period should be long enough to let the adjustment take place (the new quantities ordered to arrive to the stock location) yet short enough so that a sudden real change in the market demand will not occur without someone noticing. For the TMR – the cooling period is a full replenishment time, and for the TMG the cooling period is letting the inventory at site cross over to the green from above (since lowering the buffer size probably caused the current inventory at site to be above the buffer size level).

Manufacturing priorities according to urgency in PWH

Usually the manufacturers manufacture to order. That means that each work order on the shop floor is for a specific customer for a given due date. TOC for that environment prioritizes the production orders based on their due dates (for more details please refer to published literature on Simplified DBR).
When manufacturers embrace the TOC solution for distribution, another angle should be thought of. In this case the production orders are not for a specific customer, and are just covering for consumption from the PWH. Therefore, the right priority should be set not according to time, but rather in the same way the priority in the stock locations for the SKUs was defined – the best priority mechanism is to take the buffer penetration at the site as the priority for the Work order that needs to replenish it. If there is more than one WO for the same SKU – the best priority mechanism is to take again the Virtual BP in the following way:

Priority for an SKU held at the PWH

- The Current aggregated stock (virtual stock) is calculated for all downstream links for the same SKU and the appropriate Virtual Buffer Penetration is calculated based on what is missing for the full buffer against the buffer size.
- The priority is determined by the virtual buffer penetration of the next link



Every WO looks at the virtual BP of the next WO in production (the one who was released before it) to get its' production priority. This ensures that the production is in line with the actual usage of the stock – if the stock is depleted fast the WO will be expedited throughout the production and otherwise it will float in production on the excess capacity of the production system. Every entity in the supply chain is fully aligned and synchronized with the goal of the system – to be responsive to the actual consumption of stocks from the next link in order to create availability otherwise unattainable.

Why does a supply chain based on pull distribution work better?

Let's look at the retail store and the different entities operating in this environment. We can categorize the items in the store to 3 different types:

1) **Cheetah** items – these items are sold very fast, enabling the retailer to reach high inventory turns
2) **Elephant** items – these items are items the retailer just can't get rid of – items which are running very slowly with low inventory turns
3) **Regular running** items – the items which do not fit the above categories

What is bound to happen with the fast running items?

When items are cheetahs, by definition the market demand is high for them relative to the amount of inventory we keep from them. Therefore, they are the ones **most likely to be sold out**. If we go to a retailer and ask him how many shortages he experiences, the most likely answer would be: very few, maybe 2-3%. There is a lot of misconception here – since if we'll ask him: let's say we stand outside your store and ask people whether they found what they were looking for – in how many cases will we get an answer of "no" even though you're supposed to carry what they were looking for? The most probable answer would be: OK – probably 10-15%. This means the level of shortages experienced in shops is much higher than what the retailers think. If the typical buying pattern of customers in the shop is purchasing more than one item at a time, this phenomenon is ten-fold: what is the chance, when having only 10% shortages, for a customer to find all 8 items he's looking for in the shop? The answer is almost zero – affecting the buying experience of almost every customer that buys in the shop.

A very interesting factor comes into play when analyzing those missing items: The 10-15% consists of mainly the cheetah items! If the retailer would have known these items would be sold so quickly, he would have bought a whole lot more. Therefore, the amount of lost sales he experiences is far more than the 10-15% he will actually admit to! This is true especially in the fashion business. Goods are bought by the retailers once for the whole season. Therefore, the fastest running items will be missing almost throughout the season! For example – an item which sells so fast all the inventory is consumed in 2 weeks in an 8 weeks season has lost sales of 3 times as much as was kept of it!

The other side of the coin is the elephant items. These items are not sold as the retailer had envisioned when he bought them, otherwise he would have avoided them. The phenomenon that happens here is absurd – the retailer will invest a lot of efforts to sell these elephant items and block his display space at the expense of the other items in the shop! This behavior, while expected from the psychological side, is counter-intuitive in the business sense – huge efforts that will be invested by the shopkeeper to sell the elephant items could have yielded much higher revenues from the cheetah items.
This phenomenon sometimes dwarves the effect of shortages in the cheetah items!

Some industries have adopted even phrasings to hide the fact they are operating in a counter-intuitive way, because they have become desperate trying to solve these problems. The industry glorifies the stock outs of cheetah items (in TOC it is called lost sales) by calling them "sold out"! The industry simply ignores the elephant items phenomena by marking them as "on sale" and investing huge efforts in selling them.

In a supply chain that is based on pull distribution, these negative phenomenons are cut to minimum. Since the TOC mechanism is based on reacting to the actual market demand, and adjusting the buffers accordingly, if the market demand picks up, the buffers will be increased, creating a mechanism that allows stock-outs only for very limited time periods. That means lost sales due to stock outs of cheetah items are minimal with the TOC methodology. Due to the fact that lower inventories of all items are kept, and the quantities are further decreased when consumption is low, elephant items are much less of a problem as their quantities are minimal. Therefore, using pull distribution is very effective in eliminating lost sales.

Some of the finer points in implementing TOC distribution

Setting good criteria for variety decisions

To differentiate between cheetah items, regular running items and elephant items a simple criterion exists: the inventory turns – meaning how many items are sold from a specific SKU at a specific stock location relative to the inventory level of this SKU. However, it's not enough to know the quantity in which several items are sold, it's important to know also their financial value. Just knowing from the items which are the cheetah items and which are the elephant items will not help much in driving any operational decision. There are other criteria that must be considered. It's important to know the financial value of such items.
The goal of setting such criteria is obviously relevant when the shop owner needs to choose which items he would like to keep and which ones not to keep. This is only relevant when the variety of SKUs is very large and the ability of each stock location to keep a large amount of SKUs is limited. Just taking into account the inventory turns will not help – some items are sold at such a low margin that even if they are cheetah items they are not giving much to the bottom line, and a certain item can be sold only once every year (an obvious elephant item), but the margin is so high relevant to the investment that it's a great item to have. For the manufacturer/distributor, a measurement like that can be used to determine which products he would rather not have in the supply chain at all, signaling a new product design is needed.

The best measurement for determining how much a certain SKU is worth keeping at the stock location is simply Return on Investment (ROI) – how fast this SKU is bringing value. Since the ROI measurement was created in order to help in decisions concerning choosing between different projects, it's a perfect fit here. The distributor and shop owner are always limited by the amount of cash and/or space, so they should be focused on the items that would contribute the most to the bottom line.

In TOC financial terms, this is the way this return is measured: How much Throughput (in short T – meaning margin – selling price minus truly variable cost) does one gain from this SKU over a period of time. The best time period to look at is a year, to take into account the effect of sudden peaks in demand (usually stemming from seasonality).
To calculate the Investment, consider the following:

- The inventory kept at the stock location is the one covering the demand
- The inventory kept on the way is also an investment in order to protect from the fluctuations in demand
- There is almost always something on the way as in the pull distribution replenishment solution stocks are replenished on a daily basis (and sometimes more frequently)

Taking these into account, the best number to represent the Investment needed to generate the Throughput this SKU generates is the **buffer size**. By multiplying the buffer size in the Truly Variable Cost of this SKU the real Investment needed to generate the sales of this SKU is realized.
Therefore, the formula is very simple – to calculate the ROI, all that is needed is taking the yearly T from this SKU and divide it by the TVC per unit from this SKU multiplied by the (average) buffer size throughout the year.

The ROI measurement enables differentiating between 3 different groups of SKUs:

1) **Star** items – these items represent a very quick ROI – meaning keeping them is very good for business – and for the manufacturer/distributor this is a kind of product he would like to keep at all the stock locations he services
2) **Black Hole** items – these items take a very long period in order to return the investment done in them. For the manufacturer/distributor and item in this group signals a possible item to stop producing/purchasing. However, this is not conclusive, as some items (usually referred to as **Strategic**) are a must to have even though their margin is so low that it puts them in this group
3) **Regular ROI** items - these items are not in either category

It's obvious there is a correlation between the Cheetah items and the Star items, but this is in no way a 1:1 correlation, as is clearly demonstrated by the extreme cases discussed earlier.
The decision how to set the limit between the different groups is up to the specific environment, but the general guidelines are taking the high 10% as stars and the low 20% as black holes. One of the possibilities to treat black hole items is to try and change the price tag of some of those products – making them more lucrative if they can be sold at the higher prices.

Rules for setting up initial buffer sizes

The first step in moving from push distribution to pull distribution is setting up the plant warehouse (PWH) and starting to build inventories to fill the initial stock buffers.
The decision of what the initial stock buffer should be might seem a very complex decision – the amount of uncertainty is huge, so fear is very natural – being afraid of making the wrong decision, as if difficulties will be encountered the TOC logic might get blamed and the whole effort of moving to pull distribution might be regarded as stupid.
There are not enough words in the dictionary to emphasize the difference here between being exactly wrong and about right. At Inherent Simplicity, we've encountered several cases in which determining the initial buffer targets took more than 3 months! This amount of time would have been enough to reach the right buffer from almost ANY initial buffer size and we would have reached some results on top of it.
Since the DBM mechanism will adjust the buffers according to real consumption, all the initial estimate needs to be is in the neighborhood of the right buffer, and even that is not a must.

Inherent Simplicity suggests its clients to start with an initial guesstimate: taking the replenishment time from the source to the destination and multiplying it by the average consumption and by a factor of 1.5.
The replenishment time to use should be:

- For a production environment (plant warehouse) – taking the current quoted production lead time for this item (after implementing TOC in the manufacturing environment the lead time will usually be cut in half and then the DBM will automatically suggest lowering the buffer)
- For a transportation environment (central warehouse, regional warehouse and consumption points): mainly transportation time plus something to account for shipping a limited times each week

A simple rule can be used also to determine whether an SKU has some yearly seasonality effects: if looking back on last year's consumption (and the year before if possible) one month's sales are more than twice the monthly average of the total sales (somewhere between 15-20%), this SKU should be defined as seasonal in that month. For the seasonal SKUs, a different initial buffer can be defined for the seasonal months and for the regular months (using the same rules stated above). The difference in the buffer sizes can be calculated and fed into a seasonality model – in which the buffers are set manually or automatically by software before the season starts/ends. The monthly/weekly seasonality can be detected using a similar mechanism.

Implementing the TOC distribution model – how can software help and is it really needed?

To successfully implement the TOC methodology to manage a distribution environment, two major requirements need to be fulfilled:

1) Replenishment – meaning replenish to the different locations according to consumption
2) DBM – Dynamic buffer management to change the buffer size constantly and keep it at the right size to support the current consumption from the consumption points

These requirements are not the only ones that need to be implemented, but these two are the most basic – they will be needed in any distribution environment.

Even considering only these two requirements – the conclusion must be that no organization can manage it without software, unless this is a really small distribution chain (anything more than 50 buffers to manage requires some kind of software). The question is: what kind of software can be used?

First – define how many buffers are likely to be kept under the TOC distribution model:

- The first number that needs to be figured out is the number of SKUs that are planned to be managed – this is the number of SKUs the company right now offers the market
- The second number is the number of stock locations that the SKU will be managed in – all warehouses (regional, plant warehouses) as well as distributor warehouses and client shops in which in the future the SKUs will be kept to stock

The estimate on number of buffers that will need to be managed stems from the multiplication of the two numbers above.

In general, there are three options to choose from regarding software:

1) Develop the needed software components within the existing ERP system used by the organization
2) Develop the needed software components as excel sheets – external to the ERP system
3) Purchase an external TOC specialized software

The answer on the question which of the three options should be the one to choose mainly depends on the scale. For any environment in which less than 500 buffers need

to be managed, using an internal software is a possibility (whether an excel sheet or a development of the current IT system).
For any environment that contains more than 500 managed buffers – the recommended solution is to get external software, which is fully focused on the TOC processes and decision making.

Why not use internally developed software instead of investing a large amount of money and effort in external TOC software?

1) Quality Assurance - ensuring that the internally developed software module is doing what it should be doing is very problematic – the good TOC add-on software vendors are investing most of their efforts on checking the validity of the modules they program
2) Reliability - ensuring that now and in the future, no changes or additions are done to the modules (causing negative ramifications) by people who "think they know"
3) Development - The TOC knowledge is right now in the beginning of the process. New insights are developed continuously by TOC consultants and software companies, and the TOC software companies invest a lot in order to incorporate the latest knowledge of TOC into their software. An internally developed system will never keep up with the developments
4) Proper know how – There are a lot of fine details that are not within the public knowledge domain. When considering companies with special needs – such as seasonal products, groups of similar products or large numbers of buffers – only a TOC software company can incorporate software modules to correspond with those needs. Developing them internally in the company will take huge amounts of time and effort without promising any results
5) Long Lead Time – from a lot of experience in trying to develop internal TOC software modules, the time needed exceeds even the most pessimistic estimations. Inherent Simplicity had a lot of experience trying to consult companies to build their own internal modules to support the TOC processes. Even when internal IT capabilities were not an issue, huge amounts of time have gone to waste waiting for the solution to be incorporated into the system, with a lot of uncertainty when the modules will be ready
6) An excel sheet, despite its relative ease of use, is especially not recommended: An excel sheet is very easy to change, and therefore cannot really be used in order to enforce the correct use of the tool. On top of that – an excel sheet is **very** hard to debug. In other words, the first two entries apply very strongly to the use of excel sheets

Pilot and software

Before launching the distribution solution on a full scale, many companies would like to experiment with a pilot to see whether the TOC distribution solutions makes sense and brings results. A pilot phase is not always possible (for example – keeping stock at the PWH just for a few retailers that use the TOC distribution solution will yield much lower results than they would for a full supply chain), but where it is possible, the question of software again presents itself.

A few TOC software companies offer today a model in which the software can be used for a pilot phase in a relatively low price.

Since the solution is not meant to be a long term one – it provides the option to take short-term solutions on purpose, meaning:

- For a short-term solution, if choosing to go with a TOC add-on software, the interfaces can be managed manually and therefore implementation can start right away without too much hassle
- An excel sheet is very easily used – since this is a very short term solution and the goal is not a perfect process but a proof of concept (there is no need for a process-driven software), this option is actually better than trying to program the internal software to try and cope with the requirements of the pilot

Results from using TOC based distribution

From Inherent Simplicity's experience in implementing the pull distribution TOC solution, it is safe to say that the results are remarkable. If using the rules of thumb as listed here, especially to set initial buffer sizes, remarkable results were encountered within a period of 3 months since the start of implementation. The average (yes – average) results of implementing the TOC solution using Inherent Simplicity's software (Symphony) are 40% increase in sales, coupled with a reduction of 50% of the overall stock in the stock locations.
Taking these results into Inventory Turns, from the pure mathematical perspective normally Inventory Turns are improving by a **factor of 2.8**.
The results are still without considering the criteria that was introduced here (star and black-holes items) – that criteria will start being used only in Q2 of 2007 (getting more due to the better availability of star items). The expectation is that this criterion will bring results even higher.