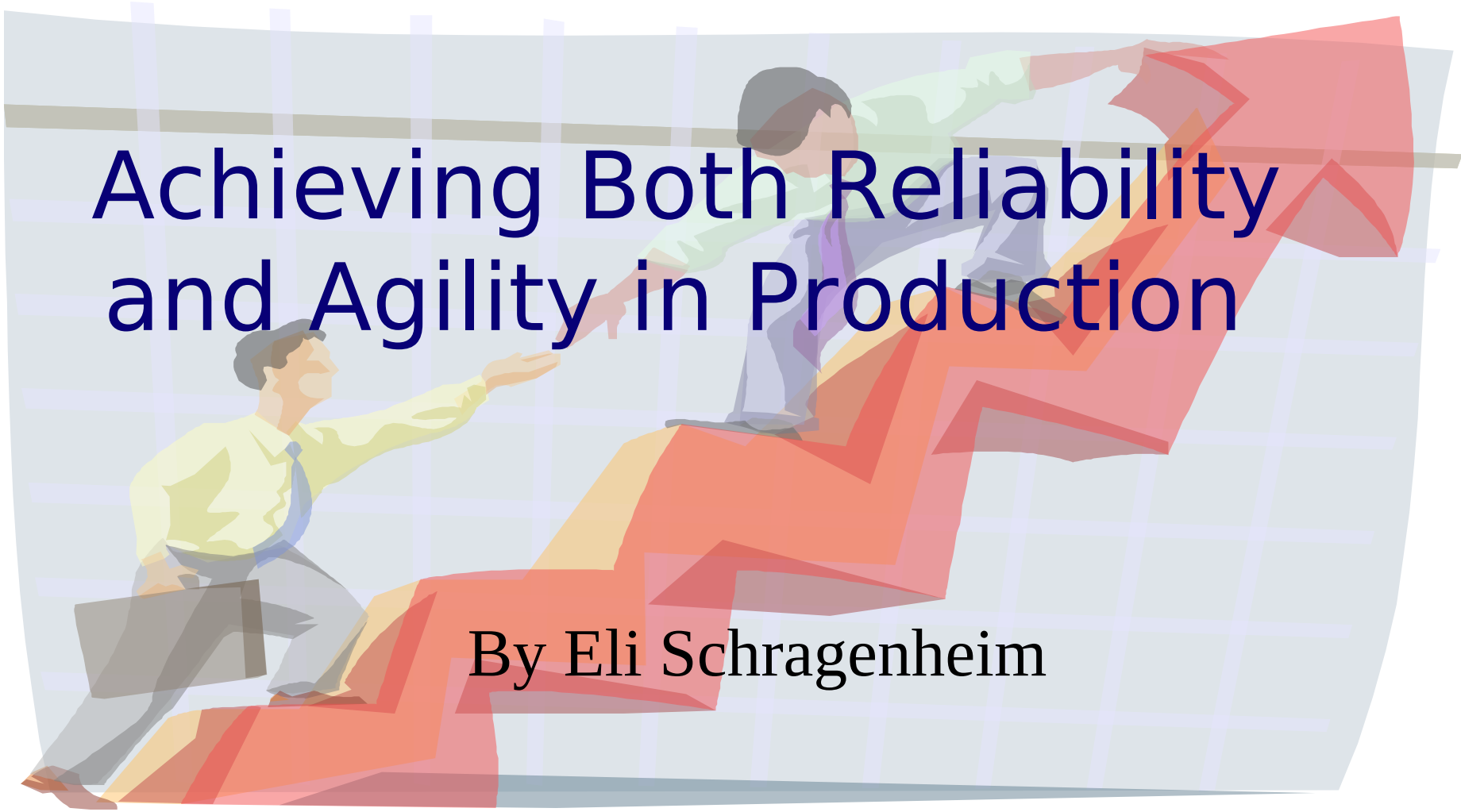


Achieving Both Reliability and Agility in Production

By Eli Schragenheim

Powerful Solutions Start With The Right Questions.

The universal problems in production

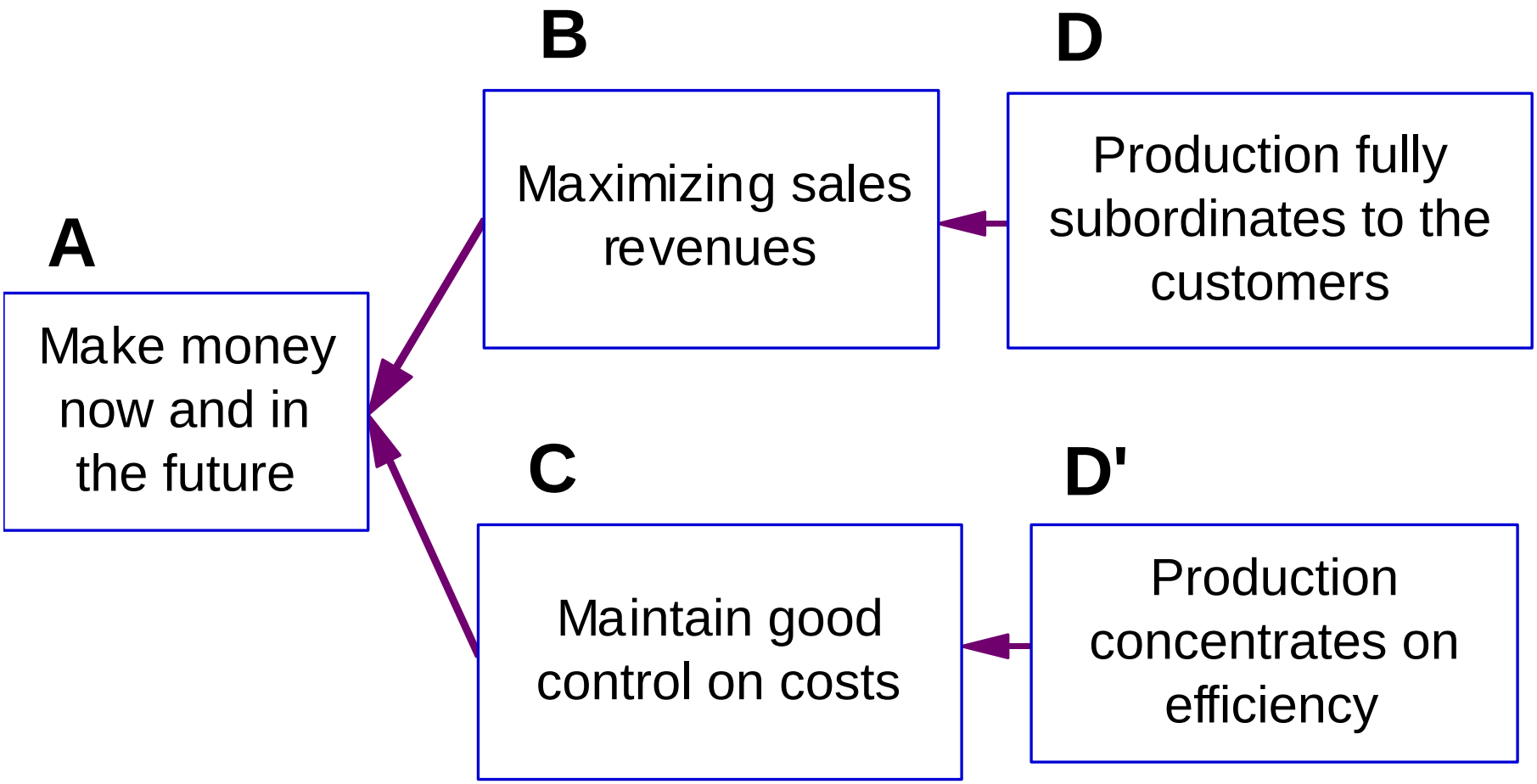
- Most production floors are very **complex.**
  - It takes huge amount of data to describe the operational system.
- Production floors are subject to very high level of **uncertainty.**
- Production is under constant pressure to improve its **efficiency.**
  - As it generates the vast majority of the expenses.
  - And the common paradigm is

A resource standing idle is a major waste

The main chronic conflict in operations

- Any operations manager is facing several chronic conflicts:
  - Changing the priorities due to immediate problems and thus causing instability.
    - Or keeping the priorities but risking missing a shipment.
  - Splitting a batch because part of it is urgently needed – or keeping the batch as is.
  - Adding overtime and/or adding manpower – or keeping the current capacity tight.
- The generic conflict is between satisfying the customers and keeping the costs down.

The conflict between Sales and Operations



Understanding the Production problems

- **How can a production facility be reliable?**
  - When it is both complex and uncertain?
  - Without wasting **huge** amount of very expensive capacity, both in machines and manpower?
- Suppose we offer very long delivery times – can we then be truly reliable?
  - Production is provided with a lot of time to process any order, but the total amount of work-in-process is very large.
    - Which could be used to increase the efficiency by saving setups.
    - But, also causing a lot of confusion regarding the priorities.
    - And the queue time is very long, causing the production lead time to be very long up to the limit of possible missing the date.
  - And how would the client react to a very long quotation?

Understanding the Production problems

- ADV200 is a virtual manufacturing organization.
  - Selling 6 different standard products to 6 different market segments.
  - The performance and business rules are in par with the competition.
  - There are two valid ways to add value to the clients
    - Reduce the selling price.
    - Reduce the quoted lead-time.
      - Now the delivery time is 6-7 weeks.
- What can you do, without taking huge risk?

Some lessons learned

- Producing in large batches causes:
  - Customer orders due farther away in time are merged with orders that are due shortly.
  - Orders for specific clients are inflated with make-to-stock for future orders.
- The key difference between process and transfer batch.
  - Process batch: the quantity processed before setting up to the next job.
  - Transfer batch: the quantity transferred to the next work center.
  - MRP assumes both batches to be equal.

Understanding the Production problems

- More telling facts of life in production:
  - There is usually huge amount of WIP (work in process) in the shop floor.
  - Thus there is a need to control carefully the where about of every batch.
    - Thus transfer batch is still equal to process batch in the vast majority of the cases.
    - Which also causes production lead time to be very long.
  - Priorities are constantly changed.
    - Only small part of the batches are really required now.
  - Due-date performance is lousy in most cases.

Understanding the Production problems

- **Planning and Execution.**
  - **Too often the planning cannot be executed as is.**
    - Murphy is the common term to describe the impact of uncertainty on the planning instructions.
  - There is a constant struggle between the planners and the production people in charge of the execution.
- Master Production Schedule (MPS) is the primary planning procedure.
  - It plans what end products would be produced, quantity and time.
  - It normally couples firm orders with stock orders.
  - It is the main source for “available to promise”.
- **There is no real execution methodology.**
  - Except look everywhere and do everything, where priorities are: hot, super hot and do-it-now!!!

Some generic TOC insights

- **Plan only what is absolutely necessary.**
  - Identify and exploit.
  - And add buffers to protect the planning from Murphy hitting the non-constraints.
  - The planning is highly impacted by the capacity of the weakest link.
- **Maintain execution control.**
  - The execution is given some flexibility to maneuver the less sensitive operations.
    - provided the planned instructions are executed as is.
  - It is a distinct methodology to planning.
    - It is based on the actual utilization of the buffers.
    - Monitoring the capacity at hand sends signals from the execution to the planning.

The importance of focusing

- Management has to focus their attention on and around the variables that really impact the goal of the organization.
  - Basic assumption: human managers cannot control too many critical variables.
  - If too many variables truly impact the performance of the organization then the performance is bound to be chaotic!
- Focusing on a few critical variables enables the human manager to control a complex system
- Definition: A **constraint** is anything that significantly limits the performance of the organization.

The Classical 5 Focusing Steps
the 3 tactical ones

Clarification: A constraint is something that if we'd have just a little bit more of it – we'd get more of the goal and if we'd have less we'll get less of the goal.

The first three steps:

1. **Identify** the system constraint.

2. Decide how to **exploit** the system constraint(s).

3. **Subordinate** everything else to the above decision.

The Classical 5 focusing steps
the strategic steps

4. **Elevate** the system constraint.

5. **Go Back** to step 1. Warning: beware of **inertia.**

The warning applies mainly to identifying the new constraints and the exploitation and subordination that result from that.

The general recommendation is to have stability even through the growth by elevation, and that means keeping the constraint in place.

This could be achieved only by elevating those resources that would become constraints after the elevation of the system constraint.

TOC insights regarding production

- The main TOC insight for Production:

Complex as the operational system may be, the output is dominated by one resource with the least capacity

  - It is sometimes referred to as “the weakest link”.
  - More common name is: capacity-constraint-resource (**CCR**).
    - Even though not always the CCR really constrains the company.
    - But, when the market grows we expect it to become a constraint, unless its capacity is elevate.

TOC insights regarding production planning

- Hence, the production planning should be focused on only three areas:
  - **The actual market demand.**
    - The default is the original customer orders as requested by them.
  - **The capacity limitations imposed by the CCR.**
    - Both areas ensure exploitation of the constraint(s).
  - The required **raw materials release** – quantity and time.
    - This area is critical for the subordination as we'll see later.
- **We do NOT need to plan ahead of time the sequence of work of the non-constraints.**

A key observation

- The **market demand is a major constraint** even when a capacity constraint resource is active.
  - Failing to subordinate to the market could lead to a significant dive down of the market demand.
  - Elevating the internal constraint won't improve anything unless the market demand is elevated as well or clear unsatisfied demand is just waiting for the company to be able to supply.
- In such a case how do we exploit both constraints?
  - Without disappointing the market.
  - Without losing too much capacity of the CCR.

The concepts of Drum-Buffer-Rope

- The **Drum** is the means for exploiting the system constraint (the market demand and the CCR).
  - It dictates **the pace of the shop floor** and **matches it to the constraint's capabilities.**
  - The actual drum is the **master production schedule** (MPS).
  - Possibly also the detailed finite capacity schedule for the CCR (this is not a part of SDBR).
- **Buffer** is the time assigned for the material to move to the completion (or the CCR).
  - Time is the obvious means for protecting the due-dates.
- **Rope** is the schedule for material release.

The important concept of the BUFFER

- Any planning should contain **buffers** against the natural “noise” in the system.
- Buffers should be accumulated and placed only in **areas that are truly critical**.
- The **system constraints** are the critical areas that need protection from disruptions in non-critical areas.
- The buffer size represents the estimation of the amount of (reasonable) uncertainty that might impact the execution.
- Once the buffer is set, **do not allow any material release that is not inline with the Rope schedule.**

The concept of choking the release

- When orders are released based on the Rope then any order in the shop is truly required within buffer time.
- This is **NOT** the situation in the vast majority of the shop floors.
  - Material is released whenever the gaiting operation (the first operation to work on raw materials) has some spare time.
  - The belief is: early release improves the chance of delivering on time.
  - And achieving better efficiency is also good.

Simplified Drum-Buffer-Rope (SDBR)

- SDBR uses only one time buffer, called **the shipping buffer.**
  - The time buffer means the whole time given to an order from material release until completion.
  - That means, the time buffer includes not only the “protection”, but also the net processing times (touch time).
    - In the vast majority of manufacturing environments the net processing times are very short relative to the lead time.
  - Materials are released based on due-date minus shipping buffer time.
  - Thus choking the release, allowing only orders that should be delivered within shipping buffer time to be processed.

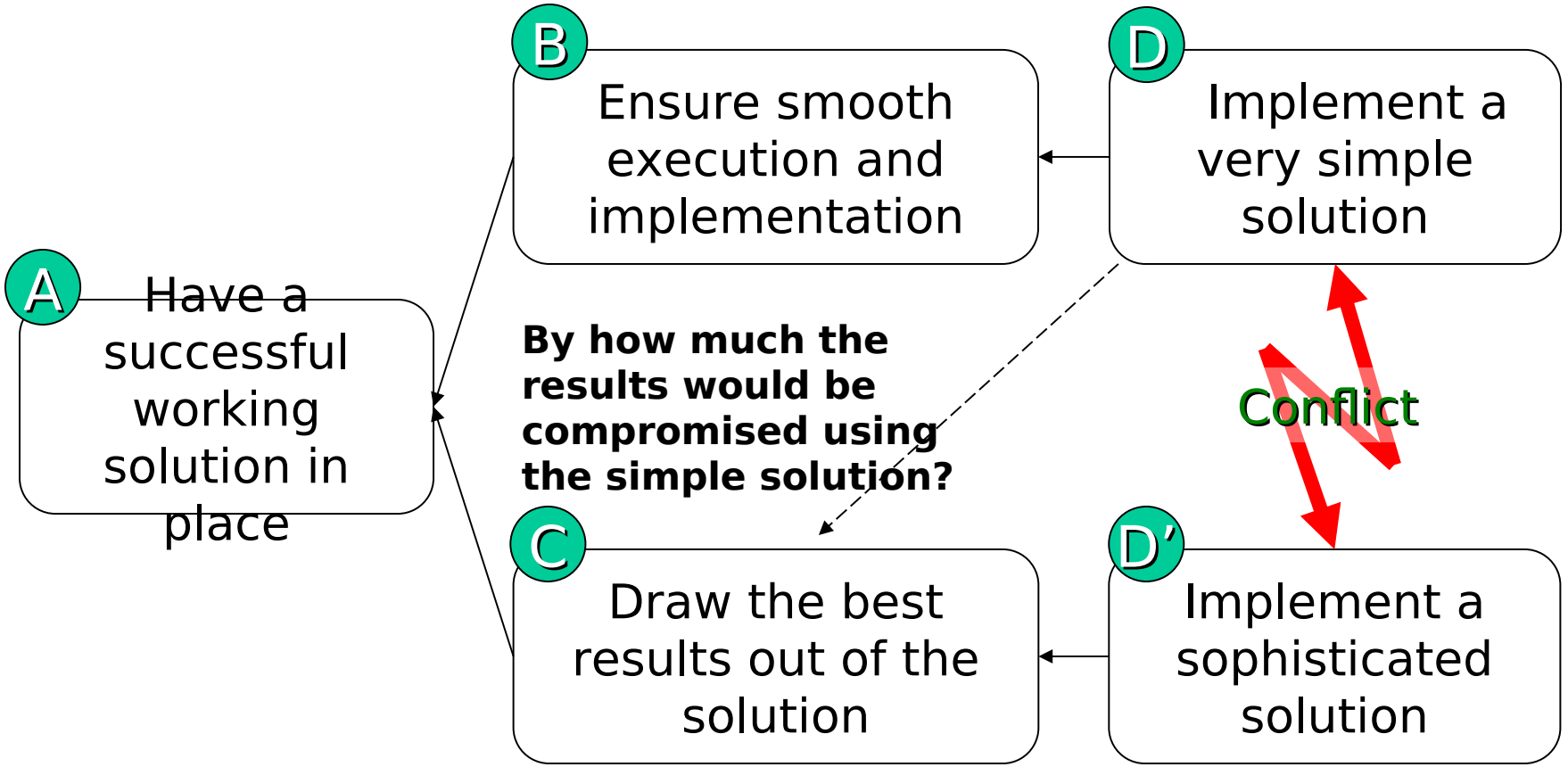
Translating the ideas to the virtual case

- Let's run the virtual case again. What changes we now put?
  - Concentrating on the market orders – let's minimize the batch size (say: ¼ of what it was).
  - Allow partial shipments (subordinating to the market).
  - Choking the release: the production buffer could be ½ of the previous lead-time: that means 3 weeks!
  - Let's promise 5 weeks of delivery.
    - We already know the market would grow by about 10% and this does not threaten the capacity.

Concerns regarding Simplified DBR

- At this point three concerns should be raised:
  1. **How can we ensure that the limited capacity of the CCR would not be wasted by unplanned idle time?**
     - The material release is dictated by the customer's orders only, and might not be inline with the need to utilize the CCR capacity.
  2. **How can we ensure that the CCR is not overloaded?**
     - Note, if we get many orders to be delivered at the same date, this could cause a temporary overload of the CCR.
     - And if the demand would go up – the CCR could become a real bottleneck and then the due-date performance would collapse.
  3. **While the time buffer could be adequate for most orders, how could we make sure every single order is on time?**

A common basic conflict to consider



Not wasting the CCR capacity

- When the CCR is not a real constraint, then it should have some idle time here and there.
- When it is active as a constraint, then most of the time enough WIP is accumulated at the CCR site, thus preventing starvation (forced idle time due to lack of parts).
- Still, the ability of S-DBR to exploit the capacity of an active constraint depends on:
  - Changes in the **sequence** of the CCR won't cause major waste of capacity.
  - The total load on the CCR is closely monitored and kept within the shipping buffer time.

Uncertainty and Control
The TOC Approach

The Concept of the Buffer:

- TOC planning uses buffers as protective mechanisms against common variations of uncertainty.
- Nevertheless, in the execution we need to identify problematic situations and respond by corrective actions.
  - Buffers are used to overcome **common fluctuations**, not cases where the fluctuation is exceptionally large.
  - Buffers reflect the current intuition in the system, however the intuition may be biased or a change in the system might cause the buffer to be not adequate.

Uncertainty and Control
The TOC Approach

Two Definitions for ‘Control’:

1. Comparing actual results with desired results and deciding whether to revise objectives or methods of execution, Fogarty, Blackstone, Hoffmann (1991).
2. A reactive mechanism to handle uncertainty by monitoring information that **points to a threatening situation** and **taking predetermined corrective actions accordingly.**

Buffer Management

- Buffer management is a control mechanism technique.
- The objective of buffer management is to check whether the due-dates or availability are threatened.
- Another objective is to validate whether the buffers used are about right, adequately protecting the system performance.
- The kernel of the idea behind buffer management is to monitor the cases when the protection mechanism is nearly exhausted!

Buffer Management

- In make-to-order environments the buffers are all time-buffers.
  - Even though the full protection system includes also the protective capacity.
    - All decisions are based on: when should the materials be released to ensure the adequacy of both the CCR schedule and the delivery dates?
    - All the questions are time based questions, not stock questions.
    - These questions determine the size of the buffers.
      - But, how can we be certain that the buffers are about right?

Buffer Management

- Every order with a due-date is planned with the appropriate buffer.
- The buffer status is defined as:

The percentage of the buffer consumed relative to the size/length of the buffer

- The higher the buffer status is the more urgent is the order.
- Thus, the buffer status list, sorted in descending order, reflects the real priorities in the shop floor.

Buffer Management

- An order penetrating into the Red Zone should initiate efforts to expedite it, so it'd be still completed on time.
- It should also initiate more actions:
  - Entering the list of **"red orders".**
  - Monitoring the total number of "red orders" to validate there are not too many "reds".
    - Too many means above 10% of the total open orders.
    - When there are too many "red orders" there is a risk to the whole protection system.
      - It might be because of an **emerging bottleneck** or a policy constraint.
    - Noting and recording where the order is stuck now, meaning which resource is holding the last piece.
    - There might be more than one work center that holds the last piece - when the order includes several legs of an assembly.

Running the virtual case under full S-DBR and buffer management

- In the computerized case, buffer management has been partially implemented through a feature called: **The Red-Line Time.**
  - The red-line time is actually a definition of the red zone.
    - Giving the signal that the order is “almost late”.
    - Providing enough time to you to react.
      - By demanding the work center to switch to the red order.
      - Providing high priority to the subsequent work centers.
      - And even considering going onto 2nd shift!

Bridging from simulation to reality

- The most significant missing element in the simulation is:
  - **The lack of human behavior aspects.**
  - Trying to conceal the excess capacity.
    - By matching the pace of work to the load at the site.
  - Pressuring management and even stealing materials in order to create more inventory.
  - Hence, in reality we would have far more inventory than in the initial state of ADV200, and the amounts of excess capacity would be hidden by slow work in slow times.

What truly impacts the lead-time?

- When the flawed policies are eliminated – the biggest impact on lead time is the **amount of available excess capacity.**
  - **The queue time behind the CCR.**
    - Note, the queue is both the WIP up to the CCR and the amount of orders not released yet because of the queue for the CCR.
- Hence, when short lead-times are mandatory, the CCR must have adequate excess capacity.

The Planned Load

- Definition: **The Planned load is the accumulation of the derived load on a relatively loaded resource of all the firm orders that have to be delivered within a certain horizon of time.**
- It is the red bar in the simulator.
- The planned load is expressed as a number of hours/days.
  - it is **NOT** a description of the spread of the load over time.
  - Suppose we have only three orders in the backlog, estimated CCR time are: 50, 125 and 55 hours.
    - The planned load is 230 hours.
    - The sequence is **not determined**.

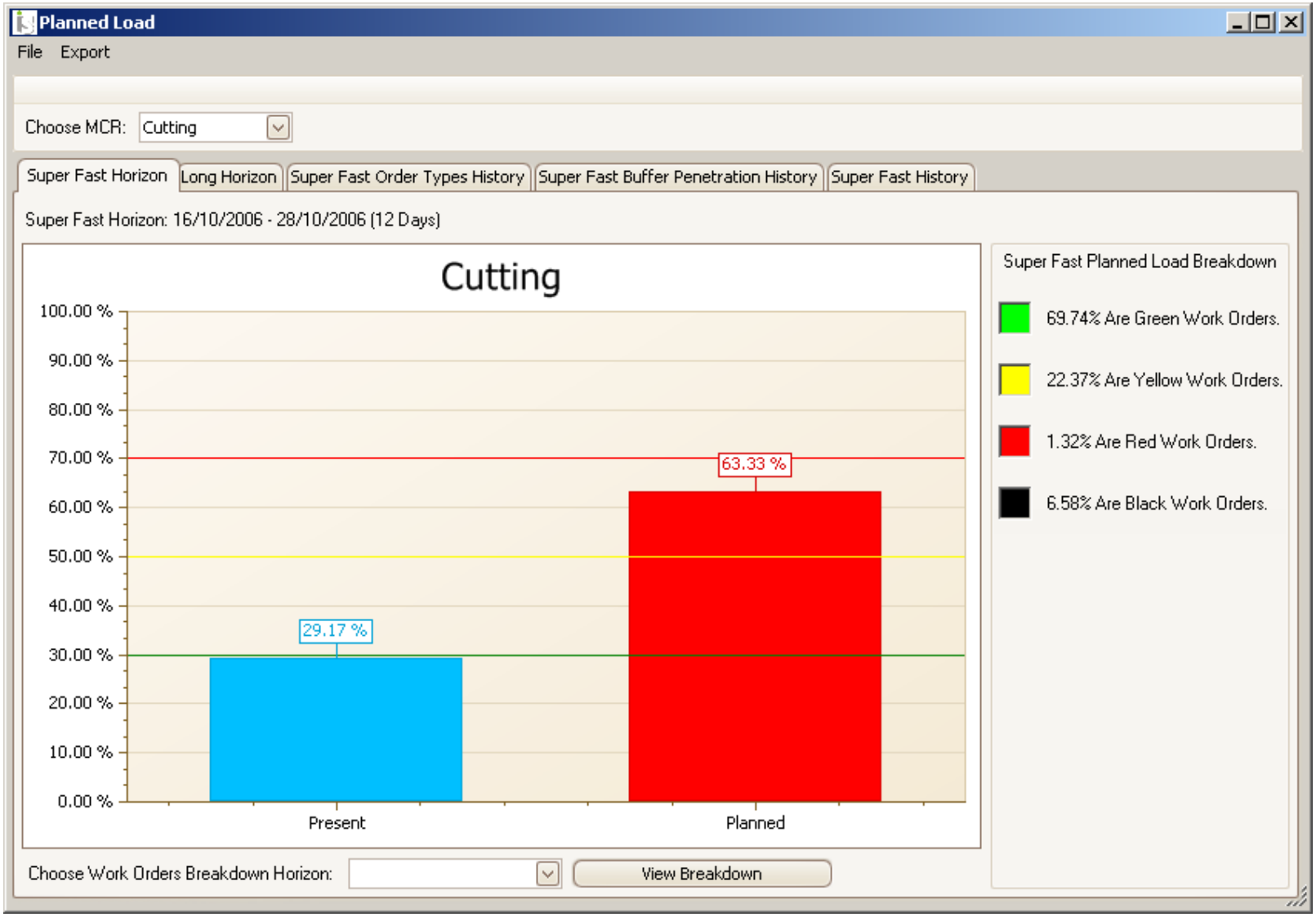
What does the planned load tell us?

- The planned load is a rough estimation of the **time when the specific resource would be able to work on a new order.**
  - Assuming the resource is able to work all the time without starvation.
  - Assuming the new order has no special priority relative to the other orders.
- When we monitor the planned load of the CCR we get a pretty good idea of the lead time.
  - The planned load plus enough safety for the downstream operations.

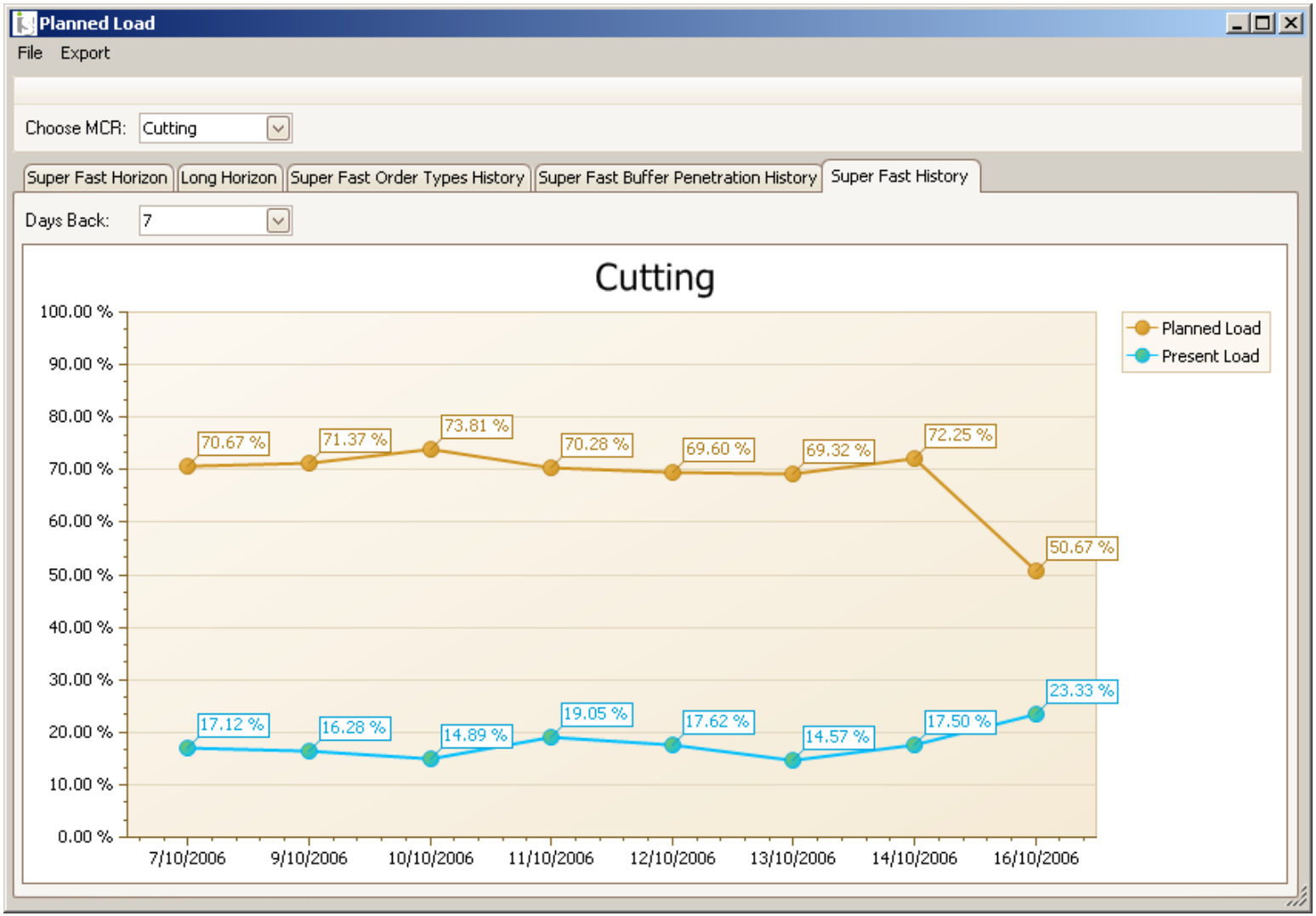
An example: **what do we know on the environment just from what we see?**

- The red bar shows the planned load of 192 hours of CCR work.
- The green bar shows the orders already reside at the CCR site.
  - This information is harder to get.
- We can easily assume the minimum safe lead time for a new order is around 240 working hours from now.
  - Assuming 40 plus hours are more than enough from the CCR to completion.
  - Assuming the new order could easily reach the CCR in 192 hours.

Planned Load as shown in Symphony



A graph of the Planned Load



The difference between a finite-capacity schedule and the planned load

- On the face of it the planned load is a schedule.
  - It is similar to “forward loading” process.
  - But, there is no sophistication in the sequence of orders.
  - No efforts are done to validate the material would really arrive to the CCR on the time predicted by the planned load.
  - **The planned load should not be treated as a schedule** – it should not dictate any sequence the operator has to follow.
- The most important asset of the planned load is the **rough determination of the time a new order has to wait for the CCR.**

Dealing with the overloading concern

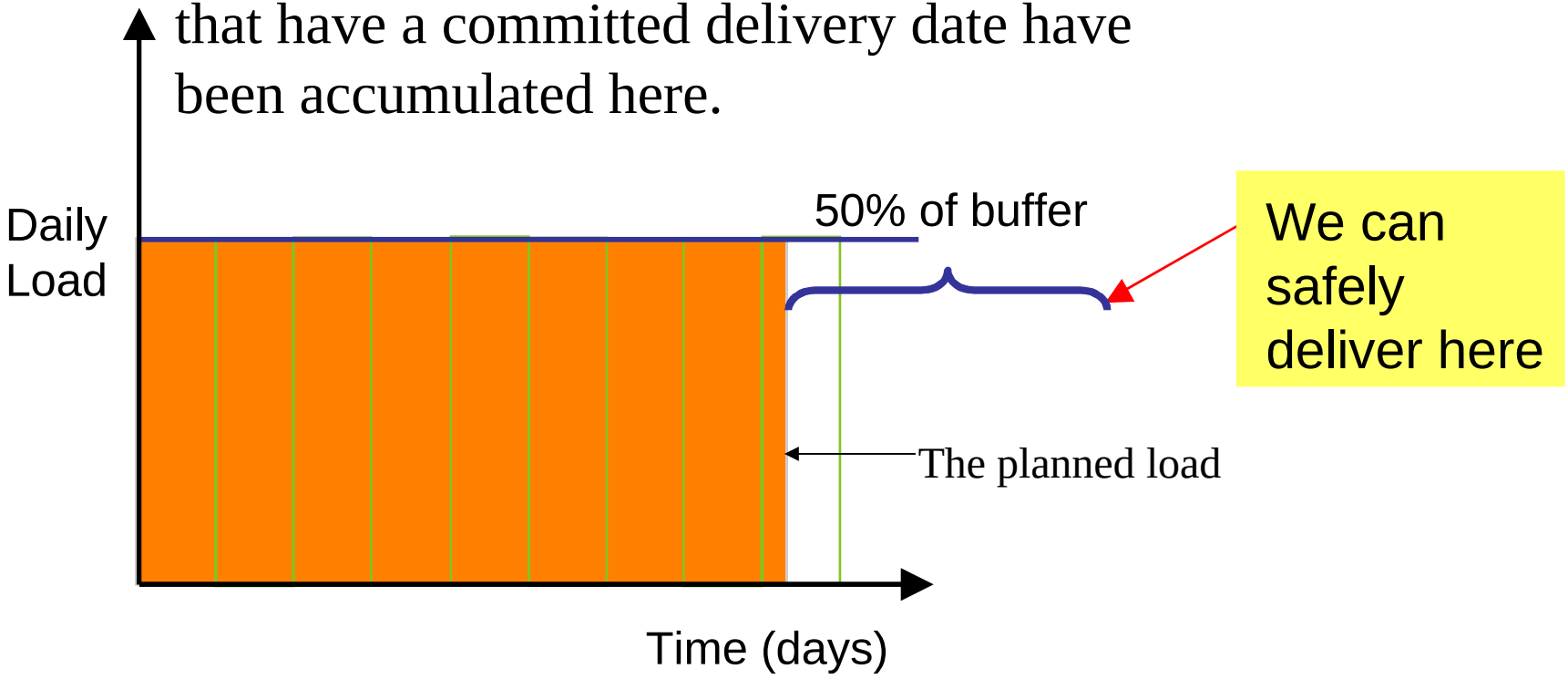
- The **planned load of the CCR**, at any given time, should be compared to the **standard lead time** required by the market.
  - Certainly we expect the planned load to be LESS than the standard lead-time.
- There is, yet, no guaranty that the load on the CCR would be smooth.
  - A cluster of orders to be shipped at the same date could be problematic.
- We need to take another step and **control the promised dates of deliveries** according to the CCR's planned load.

Promising delivery dates based on the planned load

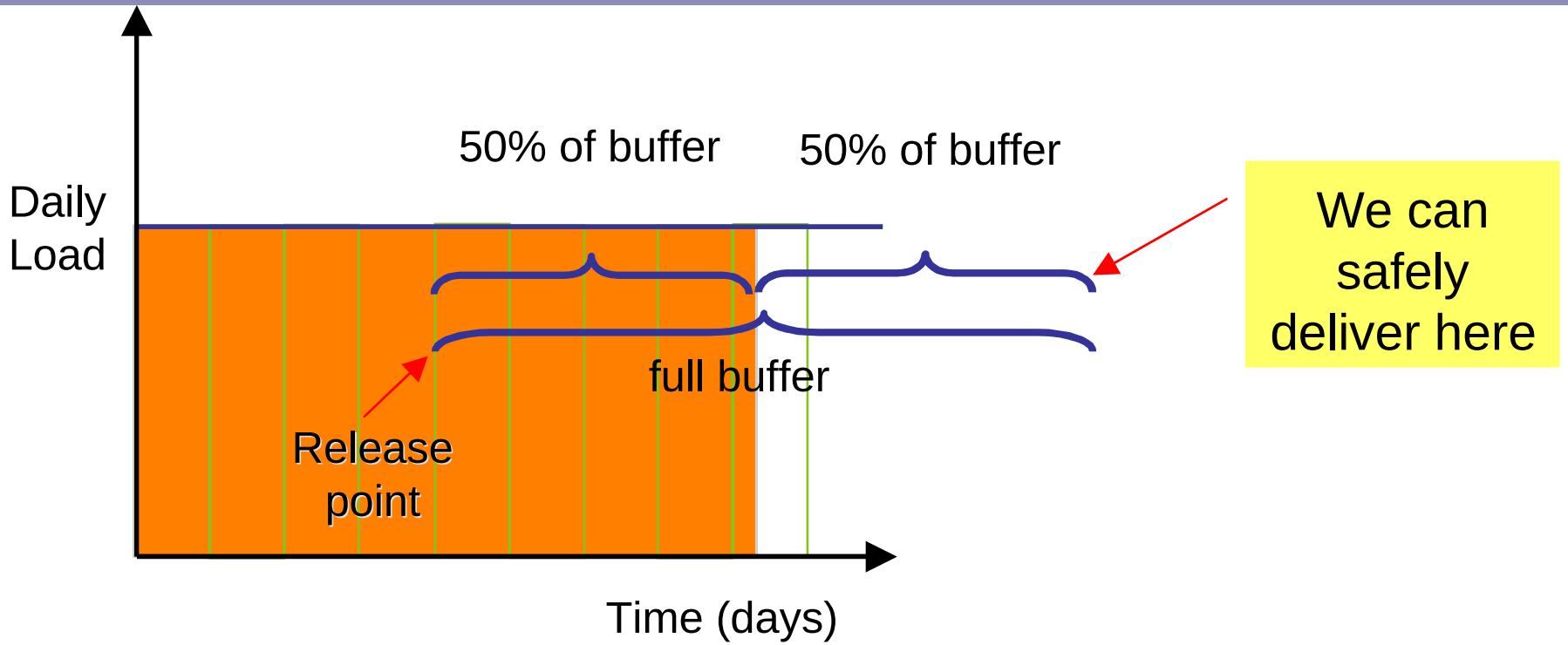
- In order to achieve reliable delivery dates **Sales and Operations must synchronize their actions!**
- Instead of Sales promising delivery dates and then Operations trying to meet as many of the dates as possible, let's build a **mechanism to determine the dates that Operation can commit to deliver.**
- As the planned load is a rough estimation of the time a new order would be processed by the CCR, we should add a time buffer to the planned load and get a "safe delivery date" for the order.
  - Let's decide to add ½ of the total shipping buffer.

This is how it looks like

All the required CCR load for the orders that have a committed delivery date have been accumulated here.



And this is how we tie the Rope to the Planned Load



The same approach, assuming the new order would be processed roughly at the current planned load time, leads us also to release the raw materials at half of the total shipping buffer before that time.

All in all we have a full buffer between the release and delivery.

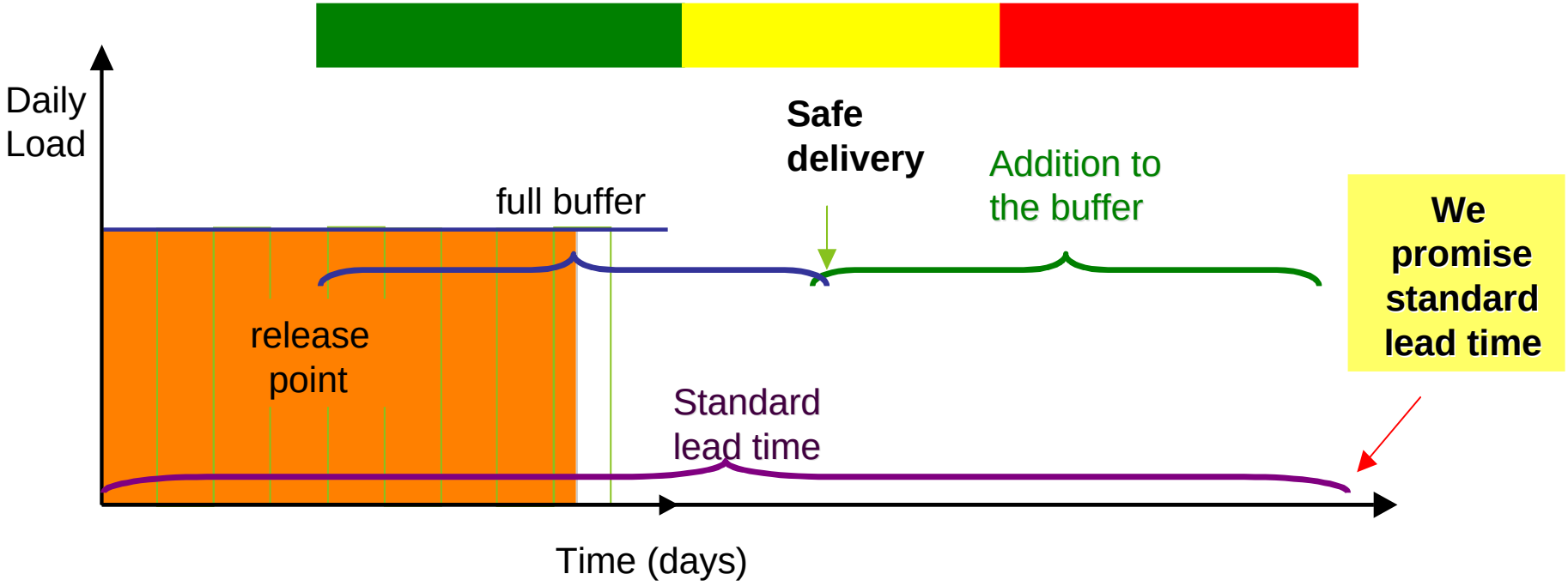
The dilemma of quoting short lead-times

- When the planned load is close to the standard lead-time or longer, the safe-date is later than the standard lead-time.
- But, what happens when the safe-date is much earlier than the standard quoted lead-time?

Should we quote short delivery times whenever possible?

- On one hand we could attract more clients now.
- On the other hand, the clients would expect those short deliveries as the regular service also in the future and for the same price.

The shipping buffer when the safe date is earlier than the quoted lead time



The actual shipping buffer for this order is expanded from the release date, dictated by the planned load minus half of the original buffer, until the formal date for the regular price

The dilemma of quoting short lead-times

- A good solution is to offer a special service for very fast deliveries for a considerable markup, while keeping the regular price for standard times that are fully reliable.
- Then we have to deal with orders with very short lead-time, while the regular orders have much longer lead-times.

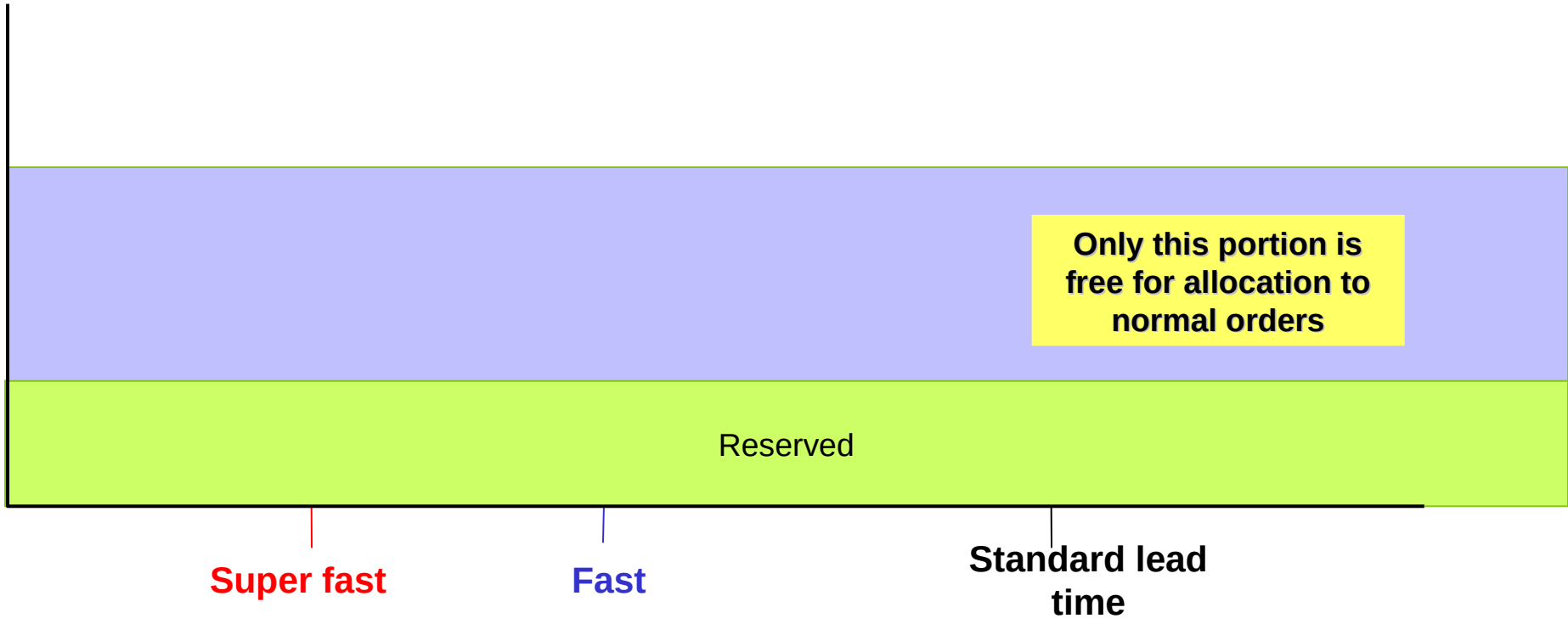
Dealing with more than one standard lead time

- Defining the problem:

When a company has different commitments to lead times, then promising reliable due-dates is problematic because not all the orders are known at the time.

- This problem includes different categories:
  1. The Rapid Response viable vision template.
  2. Commitment to a strategic client to fast response.
  3. Maintaining product mix with highly variable standard lead times in the industry.
  4. Items that are made to stock.

Reserving capacity for the faster items



The proposed approach is to dedicate certain percentage of the capacity for the faster response items.

The calculation of the planned load for determining the safe dates is now based on (100 – ReservedPercentage) of the available capacity.

Managing the reserved capacity

- **How much reserved capacity we need?**
  - Like in the determination of buffer size a good initial guess is all we can hope for.
  - Then, we need to look for a way to assess whether the current reservation level is about right.

What do we lose from determining a too high reservation level?

- **Do we lose precious CCR capacity** when we have reserved too much capacity for the “special” items?
  - We do not really mean that every day the CCR will work X% on the special orders and (100-x)% on regular orders.
    - The CCR should work on **whatever is available**, taking into account the buffer statuses.
  - Thus, when there is not much work for the special orders, then most regular orders would be processed **earlier than predicted.**
    - We are not going to move the due-dates backwards.
    - It is easy to move the release of materials backwards when it becomes evident the CCR would be able to process them earlier.

What do we lose from determining a too high reservation level?

- When we plan with high reservation of capacity the **safe-dates are too conservative.**
- So, when the safe-dates lie beyond the standard lead-time, we might lose market demand.
- Then we lose something substantial, which makes it worth to re-check the amount of reservation.

What happens when the reservation level is too low?

- The high number of special orders would push the completion of regular orders later in time.
  - As long as the calculated safe dates for regular orders are earlier than the standard lead times, the situation is under control.
  - When the safe dates are very close to the standard time, and too many special orders arrive, then regular orders are pushed into the red, and the reliability reputation is compromised.
  - Then there is a need to increase the reservation level, causing the safe dates to be later in time, possibly beyond the standard time and causing losing some sales.

Tools for managing the reservation level

- Monitoring the two parts of the planned load
  - The first part is the regular planned load, consisting of all the regular commitments, and how the front of the planned load stands relative to the standard lead time.
  - The second part contains the special orders assuming only the reserved capacity is available for them and how close they are to the short lead-times references.

Tools for managing the reservation level

- Eventually we need to control the overall capacity and how it is used for both regular and special orders.
- For that we need **full planned load** for the shortest lead-time (like super-fast).
  - **All the orders that need to be worked on within the short lead time.**
  - The planned load needs to show some excess capacity within the short time frame.

Summary

- S-DBR is appropriate for the vast majority of the manufacturing environments.
- We have focused here on S-DBR in **make-to-order environments**, but the approach also covers **make-to-stock and hybrid environments**.
- The **market demand** is the **major constraint** to subordinate to.
- **Capacity monitoring of the CCR** is critical and can be easily done through the **planned load**.

Summary

- **Buffer management**, which is really followed by the operators, is a **necessary condition** for successfully implementation of S-DBR.
- Another condition, which is necessary for maintaining full reliability of delivery, is to have **Sales fully integrated with Operations.**
  - Operations provides realistic safe-dates, which are followed by Sales.
  - Managing reliability for urgent orders is also possible, but requires some capacity reservation.
    - Without truly losing precious capacity from the CCR.