VECTOR Consulting Group | Leveraging the Potential

How bad are local **efficiency measures?**

by **Satyashri Mohanty** and **Shailesh Rajan**

One of the predominant paradigms of managing shop floors is maximizing work center efficiencies. Many managers tend to believe that it is the best way to reduce the product costs and get the best out of the plant. The origin of the paradigm can be traced to the allocation rules of cost accounting. The rule of allocating overheads of a cost center on to the products by using machine or labor hours tends to create scenario where-in, increased output leads to overheads distributed over more output units and hence consequently lesser cost per unit. Hence the need to maximize work center efficiencies particularly for high investment resources is almost ingrained in every plant manager and guides his daily decisions.

This article tries to raise and answer the following questions:

1. Is maximizing local work center efficiencies good for the plant as a whole?
2. Does the paradigm of maximizing local efficiencies, at least, help the specific work center?
3. If we establish that maximizing work center efficiencies is not good for anyone, then should they be abolished?

To understand more, let us analyze the environment of custom manufacturing plants. Typically most custom manufacturing plants have two types of resources or operations - some operations (let us call them type A) where people have a choice to not focus on orders like machine shop producing parts for servicing many orders and some work centers (call them type B) where there is no choice but to focus on order completions. Typically they are assembly operations, site installation, packing or dispatch. In most plants the type B resources are usually towards the last stage of operations and type A are in the earlier stage.

In such environments, the need to focus on utilizing work centers, forces managers to batch components across orders and when many type A work centers do that, the level of de-synchronization goes up as the order progresses in the shop. However when the order reaches close to type B resources like the assembly or dispatch or site installation, the de-synchronization is no longer tolerated. The type B resources trigger the expediting signals (or flow backs) to complete the remaining parts of an order. These flow backs usually come in as urgencies of specific single orders and as a result all supplying work type A work centers go to the other extreme – leave everything and attend to supplying the remaining parts of a specific order – they break their queues and take unplanned extra setups to service a single order. So from batching across many orders to minimizing the setup, they go to the other extreme of taking set up to service the remaining parts of one order. Typically this happens close to month end for plants which have to dispatch complete assembled units when the output of type A resources or the feeding lines to assembly drops to complete orders. This period of correction is payment for the sin of excess batching done in the initial part of the month. For organizations which supply independent parts to site installations, the capacity killer flow back requirement comes after few months depending on site installation lead time. It is not surprising for such companies to have an output which fluctuates widely across months. If in some month, the output goes up by selective cherry picking for dispatches, then after some time, the output drops when the plant tries to complete the mess of many open sites.

This painful experience is so ingrained in the minds of the plant people that they tend to believe that the focus on order completion causes plant output to go down.

But what we just analyzed is that the pressure point to focus on one order (and corresponding loss of capacity) is triggered because of delays. The delays are primarily due to increased de-synchronization. The increased de-synchronization is due to actions to maximize local efficiencies. So by trying to maximize efficiencies we end up eventually reducing it!

We know for sure that trying to expedite one order leads to waste of capacity but at the same time cherry picking across orders to fully utilize the non-bottlenecks is also the other extreme. If we operate any one extreme we will soon be forced to the other extreme. So the solution is very simple – do not operate at any extreme and you will be better off in capacity utilization as well as order completion. So we can operate at the middle of these two extreme which should be the stable zone. What is the middle zone?

De-synchronization was because of “cherry picking” of components across orders. “Cherry picking” was because orders were there to be cherry picked. So if we can reduce visibility of orders, we can find the middle zone. TOC’s Operations solution suggests that we cut the current buffer size (production lead time) to half and not allow material before the material release date as determined by the reduced buffer. Cutting the current buffer size by half means there will be fewer orders (than before) in the shop floor (WIP reduces by half). There is much less orders to cherry pick from but it is not one order - There is enough orders (though lesser than before) which allows for batching and the plant doesn’t have to process order by order. This coupled with the elapsed time based color priority system allows the work centers to play within a zone and still focus on moving many orders out from their department.

The order completion rate from the work center goes up dramatically and as a result the flow back from downstream operations (e.g. assembly/packing, installation site) diminishes because the entire order (in full kit) is being delivered. This drastically reduces the number of set-ups due to flow backs. Even though some set-ups may increase to lower batch size (after cutting the current buffer size to half) but the overall impact on number of set-ups is much lower due to reduced flow backs. So if one looks at the number of set-ups taken in a period (say one month) before & after TOC you will not be surprised to see that the total number of set-ups have actually gone down. Therefore, the output actually goes up.

Our experience of deploying TOC solution in a wide variety of environments has clearly established that the output of the plant as well as every work center goes up. Infact, the local efficiencies of all workcenters also improve after the transition phase. In one of our implementations, there was a fear that the cost at one of the non-bottleneck work centers (paint shop) would go up due to more set-ups (each setup in paint shop leads to some waste of paint). However, after the solution was implemented it was found that the cost went down. It was found that the number of expediting requests coming from downstream work centers (i.e. Assembly & Installation site) came down drastically. So the number of set-ups that the paint shop was taking after the implementation was far lower than prior to implementation. This brought down the cost at the paint shop. Infact all other local measures of the paint shop such as carrier efficiency actually improved. Moreover, on-time delivery in this plant went up from 30-40% to 98% coupled with an increase in output by 25%.

We have well established the fact that focusing on maximizing local work center efficiencies actually leads to too much work–in progress, incorrect prioritization, reduced capacity utilization, increased lead time and poor due date performance. Does this mean that the non-bottlenecks are to be ignored? If you need not focus, you can ignore it as well! Do we have to fight the battle of removing the local efficiency measures?

Let us try to understand the message “Do not focus”. It does not mean “ignore it or “neglect it”. It is the availability of protective capacity in the non-bottlenecks which allows us the luxury to focus on the bottleneck machine. What if the protective capacity drops down? Don’t we want to know if the protective capacity of some other machine is actually eroding? Many times these local efficiency measures provide us with the feedback. So why discard it?
But we learnt that bad measures lead to bad behaviors!

So we have a conflict, on one hand local measures provide us with crucial information on protective capacity but on the other hand local efficiencies drive wrong behaviors.

To understand the conflict we need to understand what we are calling as measures? Are these the numbers which are incorporated in the formal end of period MIS reports or are these questions that the top management frequently asks in meetings? In most cases what is captured in the MIS does not matter much unless it is backed by top management questioning. So measures which impact behavior are only the ones which are reflected in the day to day top management questioning. Rest all are just numbers in some papers.

So if we teach top management to stop asking the local efficiency questions, we would have removed the practice of local efficiency decisions? Is this the full proof method to prevent the damage of local efficiency decisions?

Well measures or no measures, we as human beings are hard wired to work efficiently. Give an exercise of creating 50 envelopes to any person without any performance measures. The immediate tendency is to cut

all 50 envelopes, then fold all 50 in the shape and then glue all 50. Set up reduction to minimize efforts comes natural to us. This is almost same as in a shop. Give any worker 3 orders, 2 similar and 1 dissimilar which involves a set up. The immediate instinct would be to batch the similar orders with or without a 'measurement system". So actually it is not the measurement system which is causing the problem, it is our inherent instincts. The exercise of removing the old measurement system and replacing with the new measurement system is not going to help anyone. The best approach is just to choke the release of material – there will not be enough material to batch in the first place. In case of envelopes, just give the person to work on first 25 envelopes, let him batch within that and then give the next 25. The person does not have 50 envelopes to batch in the first place – he had only 25 with him and as a result we have the first 25 coming out much earlier.

The antidote against local efficiency measures is the step of chocking the release material in the shop and the rule of following the color priorities. These steps, when implemented give significant results and take away the negatives of the local efficiency measures. So if we are having an enabling process which counters the negative impact of local measures and we ensure that top management do not ask the wrong questions – we could as well live with formal measures and observe it from time to time to validate the availability of protective capacity along with the buffer signals.

Satyashri Mohanty and Shailesh Rajan are founding Directors' of Vector Consulting Group.

Vector Consulting Group (www.vectorconsulting.in) is the leader of 'Theory of Constraints' consulting in India. Vector has been working closely with some of the well known retail chains, FMCG, fashion products, custom manufacturing industry and auto after market companies to improve their overall profitability through supply chain effectiveness.

Satyashri Mohanty can be reached at satya@vectorconsulting.in
and Shailesh Rajan can be reached at shailesh@vectorconsulting.in