HAL open science

Practical Guidelines for Production Planning and Control in HVLV Production

Erik Gran, Erlend Alfnes

▶ To cite this version:

Erik Gran, Erlend Alfnes. Practical Guidelines for Production Planning and Control in HVLV Production. IFIP International Conference on Advances in Production Management Systems (APMS), Sep 2019, Austin, TX, United States. pp.596-603, 10.1007/978-3-030-30000-5_73 . hal-02419192

HAL Id: hal-02419192

https://inria.hal.science/hal-02419192v1

Submitted on 19 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Practical guidelines for production planning and control in HVLV production

Erik Gran[1] and Erlend Alfnes[2]

[1] SINTEF Digital, SP Andersens vei 5,7465 Trondheim,Norway
[2] Norwegian university of science and technology, 7465 Trondheim

Abstract. In this paper we will propose a set of considerations and guidelines we believe are critical to achieve efficient flow in non-repetitive production. The proposed factors are insights gained through application of lean principles to two cases of non-repetitive manufacturers.

Keywords: Make-to-order, planning and control, coordination for assembly.

1 Introduction

High-variety, low volume (HVLV) manufacturing aims at providing manufacturing services tailored towards a variety of customised products. The level of product customisation offered by HVLV manufacturers can vary significantly ranging from standard products made to order to pure customisation where the entire product is designed and manufactured according to customer specification [1]

There is no clear consensus in extant literature as to the actual characteristics that constitute a HVLV manufacturer. Many definitions of HVLV production environments are related to supply chain structures and the customer decoupling point. [2], however, propose decoupling of the engineering and production dimensions with a separate engineering dimension ranging from Engineer-To-Order, where a new product is designed, to Engineer-To-Stock, where a specific design is already "in stock". Between ETO and ETS engineering modifications are used in varying degrees [3]). In accordance with this, a definition of HVLV as non-repetitive manufacturing [4] is used here without distinguishing different degrees of engineering change or redesign. This to avoid the ambiguous characterisation of an ETO producer depending on the level of engineering content and the confusion surrounding the amount of design standardisation [3].

In this paper we will propose a set of considerations and guidelines we believe are critical to achieve efficient flow in non-repetitive production. The proposed factors are insights gained through application of lean principles to two cases of non-repetitive manufacturers.

Literature on non-repetitive manufacturing tend to focus on creating flow through either Work Load Control (WLC) or a card-based system, such as Conwip [5], Polca [6] and latest Cobacabana [7]. We find however that the challenge in producing products with complex structures, often is in coordinating the different flows of parts in

order to assemble finished products. We have found little support in existing literature on how to synchronize the flow of unique parts from fabrication to assembly with the possible exemption of Polca [6]. By synchronization we understand the necessary coordination of fabrication lines for customized parts in order to have all parts available at assembly. The presence of synchronization constraints at assembly stations results in intractable analytical models for throughput estimation of fabrication/assembly systems [8]. This paper however proposes some practical guidelines for achieving better synchronization.

2 Cases

Case A is a manufacturer of ETO equipment for offshore and maritime industries. The company offers a wide range of tailor-made packages of different equipment for propulsion, manoeuvring and control systems for medium speed configurations. Their main products and spare parts for maintaining earlier installed equipment, are all manufactured at the same plant. The plant produces less than 500 units a year in a large variety of sizes and configurations.

In recent years the company has experienced a shift in demand towards increased physical size and configurations with more and larger machined parts. As a result, the workload of the plant per delivered unit has increased significantly as a result of more machined parts that also require more time for machining due to their size.

Case B produces customized hydraulic cylinders and dampers in a variety of configurations and sizes. The customization ranges from relatively simple modifications of rod and cylinder mountings to complete design of new cylinders according to customer technical specifications. About 50% of the revenue of the factory comes from producing large cylinders of more than 200 mm diameter, but this segment only makes up about 15% of the total number of cylinders. The company has quite recently invested in new machinery for machining large size cylinders and other parts and aims at creating more business in this segment.

Both case companies produce customized products that need to be managed like projects. The degree of customization in a product might vary, but all products contain some parts that are either unique to the customer order due to customization or some parts/components where demand is too low for stock-keeping, ie. made-to-order.

The typical product structures for both case companies are complex with multiple level BOMs. The parts for several different products/projects are produced simultaneously utilizing the same production resources. All products include both fabrication of parts and assembly of finished products. The fabrication process is a mix of ETO/MTO (project specific) parts and more generic MTS parts. The product structures usually contain subassemblies with similar synchronization constraints as the main assembly. The necessary synchronisations have so far been achieved by routing all parts to stock before (sub-)assembly. This practice is not very lean.

Load distribution between products varies greatly. Many factors influence the choice of which resources to apply for a task, but these choices are often made during

the engineering phase. The typical choice of resource is the most technically efficient in terms of processing time. In both companies there are few opportunities to reroute production through other resources with lesser load depending on the current situation although alternative routes are described in some cases.

3 Scheduling challenge

With the fall in oil prices both case companies have experienced serious changes in the demand for their products. Most of their customers used to be in the oil exploration industry so the subsequent decline in activity forced the case companies to look for other markets at a time when many other companies were in a similar situation. Even for highly customised products/producers the typical trends today are:

- Increasing need for shorter delivery times
- Increasing product variety
- Capacity utilisation is still important for cost efficiency

Meeting these requirements implies a balancing act for scheduling: How to create shorter throughput time for all projects, and still maintain satisfactory capacity utilisation from a cost perspective?

4 Considerations

4.1 Uncertainty and variations

A common feature of the production in both case companies is that many parts are Made-To-Order. Parts are made for a specific product/customer and there is only limited opportunity for using stocks as buffers. Non-repetitive manufacturing is inherently uncertain as there will always be limited experience in making parts that are sufficiently similar for comparison and guiding towards production time estimates. That some of the products will even be Engineer-To-Order will only add to this uncertainty.

Production buffers against variations are necessary but are limited to either time buffers or slack capacity (underutilisation of machinery) for non-repetitive production. Traditionally this type of industry has solved this by defining slack time either directly as extra time in the manufacturing process or indirectly by adding to process times.

According to [6] the use of slack time is a self-increasing spiral where long lead times imply the need for expediting jobs which in turn delay "normal" jobs and thereby increase the need for additional defined slack time. [9] proposes the use of capacity buffers and points to complex series of dysfunctional interaction that results from focus on the 100% capacity utilization. Adequate slack capacity is also an important condition for lean production.

The start/release of the assembly process should ensure the availability of all parts and the assembly operations thus require high inbound delivery precision. A special

characteristic of this type of join operations is that the resulting variation on start-up time for the assembly will depend on the worst outcome of the preceding tasks and the likelihood of a long delay increases with the number of parts that should be assembled [10]. In practice this means that a buffer/slack time is required. The challenge is then to keep control of this buffer without having to resort to standard stock keeping practises. After all, these are non-standard parts assigned to a specific project.

4.2 Shifting bottlenecks and capacity utilization

Customization implies that both product structure and routing as well as workload will change from project to project. This creates shifting bottlenecks in fabrication depending on the product mix. These bottlenecks determine the possible throughput of the production system and thus determine the response-time for the project.
Ideally you would want to plan and release workorders to the shop floor that balances workload between the available resources [11] The result of a balanced workload will be the overall fastest throughput of the production system, but the lack of synchronization between the different value streams may lead to a lot of parts and no finished product.

The sequence of a given mix of project orders has a major impact on the workload distribution on resources and the location of the bottleneck(s). The Cobacana card-based system for workload control [7] claims to achieve a better balance of workloads through managing workloads for the different work-centres when orders are released to the shop floor. Orders with workloads that doesn't fit the current reservations of capacity on the shop floor will sit in a pre-shop pool until the situation changes or the order must be released for due date adherence. Coordination between the fabrication of different parts are however not addressed directly and seems to be dependent on the order release and priority in dispatching jobs between work-centres.

4.3 Critical time path/critical chain of production activities

For all products there exists a critical time path indicating the value stream that takes the longest time to finish. All activities on this path must be completed in time or the throughput time of the product will increase. All other activities on the other hand have slack time.

A machine resource is often used to make more than one part for a project. Finding the critical time path thus involve mapping the production as an activity network with the extra restriction that each resource can only process one part at a time.

When several projects are combined in a workflow, the critical paths do not necessarily coincide with the critical chain of activities. The critical chain of production activities will all be located at the present bottleneck resource while the critical time path by definition is related to the throughput time of the project. The sequence and mix of projects will determine which potential bottleneck is active.

Most non-repetitive manufacturers that do both fabrication and assembly, produce modifications on a small number of basic designs. Each product is thereby unique but contain many of the same parts and subassemblies as other projects based on the same

basic design. The different basic designs can be thought to generate product families with enough similarities as to routings and estimated workloads. A structured method to define and use product families in designing production control is found in [12]. Using the routings and workload-norm of the product families we can investigate the limitations and resulting bottlenecks for possible production mixes. Changing the product family mix will give rise to shifts in bottleneck resource. Templates based on product families may in other words form the basis for rough-cut capacity planning.

5 Proposed guidelines

We propose takt based planning and control as the planning and control principle for non-repetitive manufacturing. Takt defines how many customized products of a certain type or family the manufacturing system will complete in a fixed time period. Each product family has a defined takt based on prior investigation of how the mix will affect sequencing and resource utilization at the resulting bottlenecks.

5.1 All parts and subassemblies must be finished before final assembly of a product.

In our interpretation this means finishing the products one-by-one in each assembly cell. There is no 90% finished when it comes to preparing for assembly. Missing one part will stop the assembly process and it will need to be restarted when the part becomes available later.

Final assembly might not be the final activity in the production process, but it's usually the last point where the separate value chains come together (marriage point). Activities that follow the final assembly may be successfully controlled by FIFO chains.

5.2 Develop a takt-based plan using product family.

We propose that rough-cut capacity plans are made using product families as templates. As pointed out earlier the templates may serve to identify the bottlenecks in producing a certain mix of products. The bottleneck resource defines the maximum output from the production system. Changes in the mix may move the bottleneck to a different resource but the idea is that for a time period the mix of products can be estimated and will also be quite stable. The maximum output for each of the product families will that way be coordinated through the identified bottleneck resource.

The next question is then how to sequence the different production orders. We propose takt time control because takt in addition to levelling activity at the bottleneck also provides opportunities for control and coordination at predictable points in time. For non-bottleneck resources the time leading up to final assembly will include some slack. How much slack depend on the total workload from all products that are produced simultaneously but this excess production capacity may also be utilised for MTS production of generic parts or service parts. Having predictable time limits thus

offer other opportunities for fabrication that will not interfere with production of customised products.

5.3 Insert time buffers before join operations in the production process.

Before assembly all parts must be available, and this will in our opinion require a buffer before assembly. With takt-based planning and control assembly occurs at regular intervals for each product family and the size of the buffer may reflect the size of normal variations in processing times for that family. A buffer for variations in real processing times can be achieved either through establishing time buffers/WIP, the use of capacity buffers (excess/unplanned production capacity) or a combination of both. With takt time control the size of the WIP buffer translates directly to available time for fixing variations in processing times. Major disruptions will still have to (and should) be handled outside the production lines.

5.4 Plan the sequence of projects in production (Operation plan)

Having excess production capacity (slack) for non-bottleneck resources does not ensure that deliveries can be made to the schedule set by Takt. A plan for latest start of all production activities timed to deliver parts according to takt, will often show conflicts in allocation of capacity between the different products. Often this can be fixed by starting some activities earlier but sometimes the routings prohibit earlier start. However, these sequencing problems may be resolved in various ways including extra/slack capacity, rerouting of parts, etc. The point is that many of these options are available at the planning stage but are more limited when executing the plan.

Some of the concerns are shown in Figure 1. The figure shows capacity usage by different machine centres (lines in the fig). The plan is for fabrication of all parts of 8 products belonging to two different product families/assembly lines. Each colour bar represents capacity usage for fabrication of a product at each machine centre. Parts should be delivered to assembly line 1 at the end of each day (16 hrs) and to assembly line 2 each 1,5 day. The plan shows a situation with almost 2 bottleneck resources but the capacity usage at both are within the limit of what can be accommodated. Changes in the production sequence on non-bottleneck machines will however in many cases lead to delayed delivery for assembly.

We propose that an operations plan is made using the real estimates for all products. The objective of this planning activity is to identify these conflict areas and plan how this should be resolved before production is started.

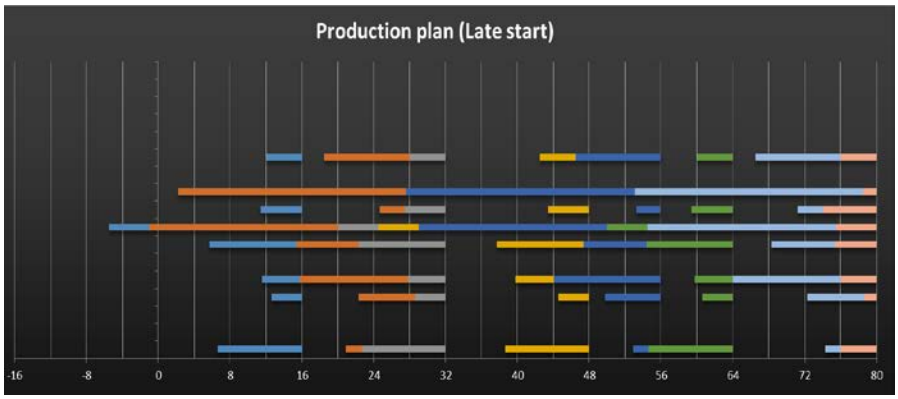

Figure 1 Operation plan for a week

5.5 Execute and evaluate the operations plan

Planned times and resource utilization are estimates and deviations will occur. Disruptive events such as machine breakdowns or scrap production we hold to be beyond the scope of the planning system. These types of problems should be managed as exceptions, maybe using the capacity reserve set aside for those purposes. Deviations are thereby limited to naturally occurring variations and buffers should be used to eliminate the accumulation of delaying effects. The size of the buffers needs to be monitored.

We propose that the sequence of projects/customer orders is maintained in production of all parts and subassemblies, unless there has been a disruptive event that renders the plan impossible to implement. Using slack capacity to produce a part for another product with a later due date will not speed up the finishing of that product unless all the other parts for that product also can be accommodated. Changing the sequence of products in parts or subassemblies production thus seldom has positive impact on throughput times and will increase work in progress and confusion about priorities. The objective should be to find a sequence of products that optimizes the use of production capacity in the first place.

6 Experience and further work

The considerations and guidelines evolved trying to apply lean principles to the production of case A. Their products are complex with a lot of fabricated parts and subassemblies. Several methodologies for lean production control where considered. Inspired by the work of Duggan [12] we arrived at using the assembly subprocess as pacemaker and two distinct product families were identified. The process time of the assembly activities did not vary much within each product family which suggested the use of Takt to coordinate fabrication.

The guidelines we propose is not all that different from standard practice in hierarchical planning with ERP. The major differences are: (1) the investigation of how production mix affects production bottlenecks and (2) the use of Takt to regulate deliveries to the chosen pacemakers.

Initially the guidelines were quite successful by reducing the throughput time of a whole unit by 50%. However, the change in demand brought on by the reduction of oil prices has changed the dynamics of the production system to the extent that our current model for rough-cut planning doesn't work.

Without a working model for rough-cut capacity planning we are not be able to predict how the mix of products may create other bottlenecks than our assigned pacemaker; the assembly process. We are currently working on a new model for rough-cut planning that can accommodate more product families and describe the capacity constraints of different product mixes. We still believe that template BOM/WBS based on product families are adequate estimates for this purpose.

Investigating product families in case B has shown that 70% of the number of cylinders they make are sufficiently similar to organize their production in separate dedicated lines. In other words, the work involved in applying the guidelines has led to better knowledge about the production system that suggested other improvements.

Other areas that need more research are among several; priorities in release of workorders and in execution, a model for operations planning that supports different routings for different parts, joining the routings at assembly and the fact that a machine normally only can process one at a time. We expect that such a model will be a hybrid between a project model and MRP.

References

[1] Amaro, G., Hendry, L. and Kingsman, B., "Competitive advantage, customization and a new taxonomy for non-make-to-stock companies," *Internasjonal Journal of Operations & production management,* pp. 349-371, 1999.

[2] Wikner, J. and Rudberg, M., "Integrating production and engineering perspectives on the customer order decoupling point," *International Journal of Operations & Production Management, Vol. 25 Issue: 7, pp.623-641,* pp. 623-641, 2005.

[3] Gosling, J. and Naim, M., "Engineer-to-order supply chain management: A litterature review and research agenda," *International Journal of Production Economics,* pp. 741-754, 2009.

[4] Katic, M. and Agarwal, R., "The flexibility paradox: Achieving ambidexterity in high-variety, low-volume manufacturing," *Global Journal of flexibility systems management,* pp. S69-S86, 2018.

[5] Spearman, M., Woodruff, D. and Wallace, J. H., "Conwip; a pull alternative to kanban," *International Journal of Proudction Research,* pp. 879-894, 1990.

[6] Suri, R., The practitioner's guide to polca: The production control sysstem for high-mix low-volume and custom products, Boca Raton: Taylor & Francis CRC Press, 2018.

[7] Thürer, M., Land, M. and Stevenson, M., "Card-based workload control for job shops; Improving COBACABANA," *International Journal of Production Economics*, pp. 180-188, 2014.

[8] Rao, P. C. and Suri, R., "Performance analysis of an assembly station with input from multple fabrication lines," *Production and operations management*, pp. 283-302, 2000.

[9] Suri, R., "QRM and Polca: A winning combination for manufacturing enterprises in the 21st century," Center for quick response manufacturing, Madison WI 53706, 2003.

[10] Baccelli, F. and Makowski, A., "Queueing models for systems with synchronisation constraints," in *Proceedings of the IEEE*, 1989.

[11] Germs, R. and Riezebos, J., "Workload balancing capability of pull systems in MTO production," *International Journal of Production Research*, pp. 2345-2360, 2009.

[12] Duggan, K., Creating mixed model value streams, Boca Raton: Taylor & Francis CRC press, 2013.