

Improving Throughput Across the Factory Life-Cycle

Karl G. Kempf, TMG/TME Decision Support Technology, Intel Corp.

Index words: constraints, capacity, supply

Abstract

A semiconductor factory goes through many phases in its life cycle including design, build, various ramps, and many levels of production. Maximizing the profitability and return on investment across this life-cycle is a critical component of Intel's approach to financial success. We have been applying the concepts of Goldratt's Theory of Constraints across the factory life-cycle and have realized improved performance in many of these phases, as well as in the integration of the phases.

The Problem

Within Intel Corporation, there are at least three identifiable supply lines (Figure 1). The most obvious from outside the company is the product supply line. This supply line includes planning to schedule production, materials to supply the ingredients, manufacturing to produce the products, and logistics to deliver them. This is the supply line that springs into action when you place an order with Intel and find it being delivered a short time later. Another supply line for which Intel is famous is its technology supply line, which has two major branches. One is product design, delivering a stream of ever faster and more capable product designs for manufacturing to build. The other is process design, providing a sequence of ever finer and more capable processes for manufacturing to follow in building products. Together they form the supply line that responds to the insatiable market demand for faster semiconductor devices with higher functionality. Perhaps the least obvious supply line from outside the company is the capacity supply line. This is the supply line that manages manufacturing resources. It is always trying to supply the most cost-effective manufacturing capability synchronized with market demand. This supply line involves at least the selection and layout of equipment (design), construction of buildings (build), startup of production (ramp), and operation (Mfg) of Intel factories. Since managing this capacity supply line is the primary focus of this paper, we discuss how this supply line is driven, how its components work together, and what problems it must overcome.

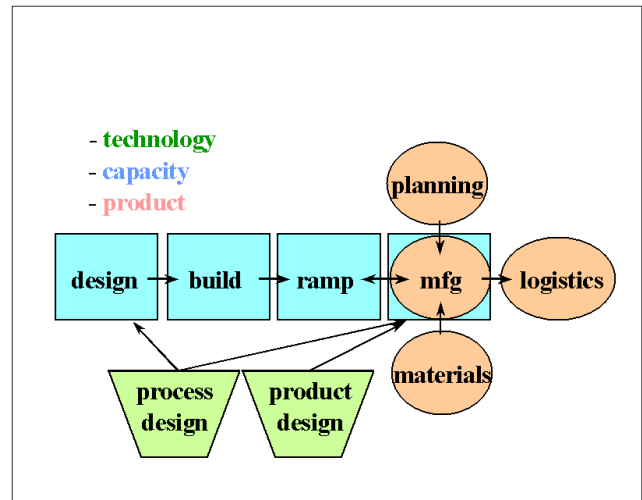


Figure 1: Intel supply lines

Clearly the capacity supply line is driven by market demand. More interestingly, it is also driven by advances in process technology. As semiconductor process design moves to finer line widths, new cleaner factories with improved equipment are needed to produce cutting-edge products commanding premium prices. This results in older factories being relegated to run products that are no longer at the cutting edge and which the market treats as commodities. Both these scenarios have a profound effect on the capacity supply line. Because the expense of building new factories is rising rapidly, there is great pressure to keep costs as low as possible and get productivity as high as possible. Because the older factories are now producing commodity products where every penny counts, there is also great pressure on them to keep costs as low as possible and get productivity as high as possible. One of the major themes of this paper, therefore, is “doing more with less” in the capacity supply line.

Another major issue for the capacity supply line is integration. It is surprising to note that the four main tasks in the design-build-ramp-run sequence of the capacity supply line can be executed almost independently with very little information flowing between them. While blindly following this factory life cycle sequence will result in a running manufacturing facility, if the tasks are not well integrated, the facility

might require extra time and/or money to complete. And while the resulting manufacturing facility will produce products, it might not do so very efficiently from a time or cost perspective. Integrating the design-build-ramp-run tasks provides a dual benefit for the capacity supply line: supplying the capacity as efficiently as possible and applying that capacity to supply products as efficiently as possible. Therefore, another major theme in this paper is “efficiency through integration” in the capacity supply line.

Finally, almost every step in the design-build-ramp-run sequence that makes up the capacity supply line involves the variable availability of resources. Most of the activities in the build, ramp, and run tasks require the simultaneous availability of equipment, materials, and skilled personnel to progress. The absence of any one resource stops activity, and the availability of all such resources is variable. Equipment breaks and needs to be maintained. Materials are supplied by vendors with imperfect resources. People take breaks and even when working diligently can only be in one place at a time. Even the design task requires detailed data about the variability in the availability of the build, ramp, and run resources. The third major theme of this paper, therefore, is “managing variability.”

The problem for the capacity supply line is to supply the most cost-effective manufacturing capability synchronized with market demand. Three themes interact to complicate any approach to managing the tasks involved in solving this problem. The pressure to do more with less is never ending and takes different forms over time. The individual tasks are complex enough that it is tempting to try to divide and conquer them, but not tackling them as an integrated whole will prove very expensive in time and money. The availability of resources for each of the tasks is always variable.

The Basic Solution

Over the past several years, we have been able to employ the concepts of Goldratt’s Theory of Constraints (ToC) to improve the performance of our capacity supply line. The most abstract version of Goldratt’s ToC has to do with making money. The most concrete version has to do with managing individual resources. Both versions are summarized here and then applied to tasks in the capacity supply line.

One of Intel’s corporate goals, supported in different ways by each of the supply lines, is to make more money now and in the future. Moneymaking is usually measured with two parameters: net profit (how much did we make) and return on investment (relatively, how much did it cost us). Ideally Intel maximizes profit while minimizing the investment required.

Translating these ideas into capacity supply line terms, we can use this corporate goal to drive supply-line decision making. Throughput (T) is money generated by manufacturing that is directly related to quality product shipped on time resulting in sales. Some expenditures are required to make T. Inventory (I) is money inside the capacity supply line such as equipment and spares and in-process materials. Operating Expense (OE) is money required by the capacity supply line such as overhead and personnel expenses to turn inventory into sales. These terms are related to profit (P) and return on investment (RoI) as:

$$P = T - OE$$

$$RoI = (T - OE) / I$$

These equations explain the pressure to reduce inventory and operating expenses while increasing throughput in all stages of the capacity supply line as an approach to doing more with less.

ToC derives its name from the key observation that in any system, the resource with the lowest capacity constrains throughput. The key process in ToC is aimed at improving throughput as the best way of driving up profit and return on investment. Step 1 involves identifying the system constraint. Step 2 focuses on understanding all means to exploit the constraint and maximize its throughput. This almost certainly includes protecting the constraint from the variability of other resources. Step 3 subordinates all other resources to the constraint, supporting all means of exploiting its capacity. Step 4 advises that whenever possible the constraint should be broken or removed, raising the throughput of the system, and the improvement process rejoined at Step 1.

The rest of this paper describes how we have used these simple ideas to develop powerful techniques to integrate and optimize the Intel capacity supply line.

Manufacturing

Our first and so far most successful application of ToC to the capacity supply line has been in manufacturing. This was an obvious place to start since manufacturing is included in all of the supply lines shown in Figure 1. Consider the simple factory shown in Figure 2. There are three processing steps, each with a machine, an operator, and an average run rate in units per shift. Since Step 2 has the lowest capacity of all of the resources in the system, it is identified as the factory throughput constraint. The factory cannot produce any more than this step can run, and any time this step is idle, factory capacity is irreversibly lost. As part of the exploitation process, its rate is identified as the “drumbeat” with which to synchronize the rest of the production line. To fully exploit

the capacity of the constraint, it must have three things available at all times: material to work on or work-in-progress (WIP), a machine to load the product into, and a skilled operator to perform the work. Subordination of the rest of the resources of the factory involves ensuring the constraint has its requirements satisfied at all times. If the factory capacity is to be raised, the capacity at Step 2 must be raised by improvement projects or equipment acquisition. And if it is raised beyond 900 units per shift (ups), then Step 2 is broken as the constraint and Step 3 takes its place.

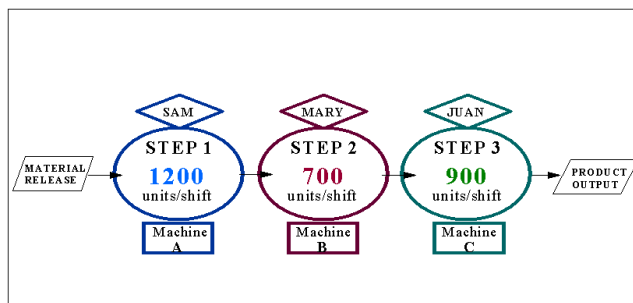


Figure 2: A simple factory

The first requirement for exploitation of the constraint is WIP, and other resources must be subordinated to ensure that the constraint is always fed. One cause of the constraint starving is the inevitable breakdown of Machine A (variable availability). One way to protect the constraint is to place a WIP buffer between Machine A and Step 2. The size of the buffer is based on the historical distribution of times to repair Machine A. Machine A is subordinated to Step 2 by always being run in such a way as to maintain the correct level in the buffer. Too much WIP in the buffer raises factory throughput time (TPT) but does not raise output. Too little WIP in the buffer risks factory capacity. Any other way of running Machine A fails to optimize constraint performance and therefore factory performance.

Another way to ensure that WIP is fed to the constraint is to control material release. Naively releasing 700 ups is not adequate. Subordinating material release to the constraint involves allowing the constraint to pull in the amount of work it requires. When the constraint is undergoing maintenance, less material is released. When the constraint exceeds its average output, more material is released. This concept is described in ToC as tying the “rope” between the constraint and material release so that the constraint can “pull in” the work it needs as it needs it.

The ToC-based approach to WIP management gets its name

from the combination of these ideas: drum-buffer-rope or DBR (Figure 3). Consistent use of these ideas drives the factory towards maximum throughput at minimum throughput time in the face of any variability in the availability of equipment. We have used these basic ideas, complemented with our own extensions, in process technology development facilities (TD fabs), high-volume manufacturing fabrication facilities (HVM fabs), and factories where die are assembled with their packaging and final testing is done (A/T). In all cases, factory throughput has gone up by 10% to 20% while inventory has gone down, with no capital outlay or increase in operating expenses. Simply getting the same equipment and people to work more effectively is the key.

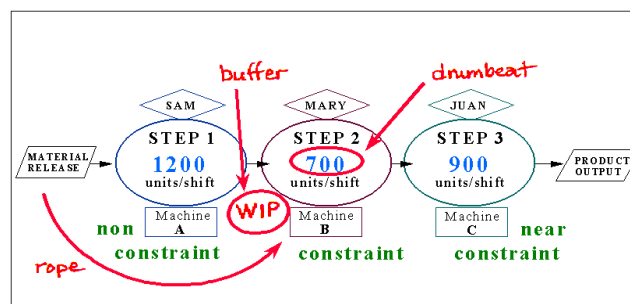


Figure 3: The Drum-Buffer-Rope factory

The second requirement for exploitation of the constraint is that the constraining equipment be up as much of the time as possible. The subordination required here is that maintenance of the constraint equipment takes priority over other equipment. In general, the more the equipment constrains the factory, the higher its maintenance priority. In the simple factory, Machine B has priority over Machine C, and Machine C has priority over Machine A. Since Machine B being down is equivalent to the factory being down, we have extended this basic thinking to further specify the number and training level of personnel dispatched to perform maintenance as a function of the constrained-ness of the tool(s) involved. This approach can yield as much as 5% more throughput with no change in inventory or operating expense.

The third requirement for constraint exploitation is that the constraint be staffed at all times. The subordination here can come in many forms. For example, machine operators assigned to the constraint must collaborate to cover for each other at breaks. Furthermore, breaks for machine operators should be coordinated with the activities of repair technicians. Another possibility is to cross-train operations and maintenance personnel on multiple tasks. The subordination here is to bias cross-training more toward the constraint and less toward

the non-constraints. In the simple factory, all three operators would be trained for Step 2, two operators for Step 3, and one for Step 1.

These ideas can be applied to many other facets of manufacturing to further manage variability and get more throughput from the same (or less) inventory and operating budget. For example, consider line yield (the scraping of WIP at intermediate positions in the process flow). With finite resources to address line yield problems in the simple factory, it is important to address line yield losses at Step 3 before Step 1. This is true because losses at Step 1 can be recovered by releasing more raw material into the factory and using the excess capacity at Step 1. Losses at Step 3 are much more costly since each involves irreversibly discarding the constraint capacity invested in the WIP that is lost, and running new material through the constraint again to replace it.

Another example involves the prioritization of engineering projects (assuming finite engineering resources). Given the choice of increasing the capacity of Machine A through an elegant and interesting engineering effort, or decreasing the preventive maintenance time on Machine B through a mundane and uninteresting effort, the latter should have priority. The reason for this is that the mundane project will increase factory capacity by increasing the availability of the constraint to do productive work. Investing effort in the former project on a non-constraining tool will have little (if any) positive impact on factory performance.

Design

Our second major application of ToC to the capacity supply line has been in the area of factory design (Figure 1), specifically in selecting the number of pieces of equipment to purchase. The goal of “doing more with less” in the capacity supply line starts here. The naive approach to design would simply be to build a balanced line, that is, a line that has the same capacity at each process step and is equal to the desired output of the factory (Figure 4). Aside from being very difficult because of the integer nature of equipment, ToC argues that this would be a very difficult factory to operate. Although there is no obvious constraint from the point of view of inspection of the average run rates of the three processing steps, there would be a constraint on the floor of such a factory. It would be the last machine that had an unscheduled breakdown, and so would move around. This would make it very difficult to instruct Sam, Mary, and Juan on how to run shift by shift, to release the right volume of material, to optimize cross-training, and to prioritize equipment maintenance, line yield improvement, and engineering projects.

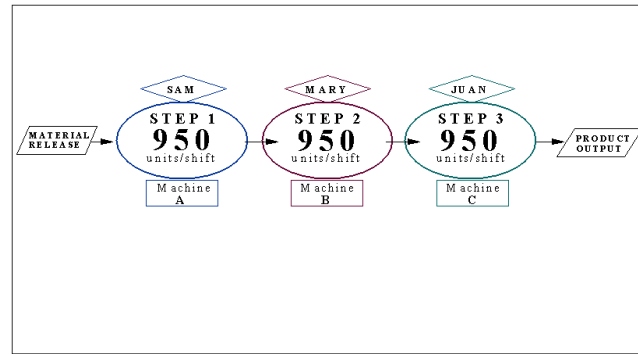


Figure 4: A balanced line

ToC would argue for an unbalanced line. Such a line might require a small increase in equipment inventory to provide the imbalance that identifies the constraint, but would supply increased throughput due to its ability to be efficiently managed. The interesting question that we have answered over the past few years is “how much imbalance?” Too little imbalance in the design of factory capacity leads to inefficient operation and lost throughput. Too much imbalance leads to wasted capital expenditure and too much equipment inventory, not to mention too much of an operating expense to run and maintain it.

The first key to answering the imbalance question has to do with the cost of the equipment. Reasoning from the inventory perspective, the most expensive equipment set should be the constraint, the next most expensive equipment set the near constraint, and so on. This means that we can imbalance the factory with the least expensive equipment sets making them the non-constraints.

The second key to the imbalance question has to do with the manner in which the variability in the availability of the equipment, people, and WIP stacks up across the factory. Since it is important to protect the constraint from starving due to the variability of upstream resources, more of those highly variable resources can be purchased. This imbalances the line, and does so in such a way as to reduce the variability of the troublesome equipment set by having more members in the set.

The third key to the imbalance question has to do with the WIP management ideas described in the previous section. Remembering that the goal of drum-buffer-rope is to maximize throughput at minimum throughput time in the face of any variability in the availability of the resources, it is important to include the WIP policy in the factory design process. Selection of the equipment sets to be the constraint and near-constraints must include considerations of how easily they can be exploited and how easily the other equipment sets can

be subordinated. As a contingency for future success and market upside, the design process must also consider the possibilities for breaking the constraint.

After practicing on two previous process technologies, we have applied this approach to our three latest process technologies as they rolled out and forced existing factories to be refit with new equipment or else new factories had to be built. The savings compared to our previous design methods have been in the range of 3% to 8% in capital cost (or I) for equal or improved throughput. While these percentages may seem small, given the multi-hundreds of millions or billions of dollars spent on equipment in each of our factories, the absolute savings have been substantial.

Ramp

Our most recent application of ToC to the capacity supply line has been in the area of factory ramp (Figure 1). Once again, the theme of “doing more with less” arises, and once again ToC points a way forward. Ramping a factory means going from one level of production to another. In an existing factory, this might mean ramping volume up or down on the current process, or ramping down an old process and ramping up a new process. In a green field situation, this means going from zero production to full-volume production. Since semiconductor production equipment is so expensive, it is normal in all these cases for equipment installation and equipment operation to be going on simultaneously. Once one (or a few) pieces of each type of equipment has been installed, raw materials are released into the line and production proceeds. As more tools are added, more raw materials are released.

ToC can be used in two ways to produce the fastest, cheapest ramp. One is to determine the identity and utilization of the constraint. On the one hand, since production is going on during the ramp, all of the ideas described above in the section on manufacturing should be applied for maximizing throughput. This would require that the constraint is known and does not move around. On the other hand, equipment is being installed daily and the capacity of the factory is dynamic. This means that the identity of the constraint could change from day to day. Goldratt's ToC argues that the installation of the equipment should be choreographed so that the identity of the constraining equipment set is constant and throughput can be maximized.

The other way in which ToC can be applied is to use the same thinking described in the section on manufacturing, but this time, transform the entities being discussed. In manufacturing, the process flow is the physical/chemical transformations being made on the wafers in fab or the die in A/T. The WIP is the product that is moving across the flow. The constraints are usually the processing equipment or the person-

nel operating or maintaining the processing equipment. During a ramp, the process flow is the installation and qualification steps required for each of the equipment types. The WIP is the pieces of processing equipment being installed including supporting materials. The constraints are the electricians, piping specialists, mechanical contractors, qualification technicians, and so on who are executing the installation and qualification steps.

Once this transformation has been made, it is simply a matter of using the identify-exploit-subordinate-break process that is the core of ToC. The constraining resource is identified, and all ways to exploit it for maximum throughput are identified. All other resources are subordinated to the constraint. If higher throughput is desired, that is, the ramp needs to be done faster, the constraint must be broken.

This leads to the interesting question that we are currently answering, that is, “How fast should a ramp be done?” At one extreme, a large number of resources could be utilized (high OE), and the ramp could move along very quickly. But, if the result is more product at a faster rate than the marketplace can absorb (low throughput), then the RoI on the ramp is not very good. (Remember T is product sold, not just product produced.) At the other extreme, a very low-speed ramp could be executed with few resources (low OE). But, if the result is less product at a slower rate than the marketplace can absorb (low T), again the RoI is not very good. This is magnified by the fact that inventory would also go up since the equipment would be WIP for a longer period.

And of course, the ever-present variability in the availability of the resources has an impact on the ramp rate just as it did on the design and manufacturing phases. However, we expect that applying Goldratt's ToC principles to this balancing problem in the face of variability will increase our performance during a ramp by as much as 15%.

Integration

Each of the previous sections has described how we have used ToC to do more with less while managing variability across the capacity supply line. The topic that has not been mentioned since the problem statement is “integration,” and ToC has helped in many ways on this important topic.

Using ToC in the design phase has decreased the amount of equipment we purchase (I) to deliver the same throughput (T). This means that even if we didn't use ToC in the ramp phase, there would be less equipment to install and so the ramp could be faster and cheaper. But since we do use ToC to ramp, we can better manage the constraint and apply all of our manufacturing ToC ideas earlier to realize higher throughput sooner from the reduced equipment set.

Using ToC in the design phase forces one to include the WIP policy that the resulting factory will use, and that in turn forces one to carefully consider how the factory will run at high volume. The fact that the WIP policy has already been designed means that it can also be applied during the ramp instead of waiting until all of the equipment has been installed.

Last but not least, we use modeling and simulation to try different scenarios around the ideas of ToC over the factory life cycle. And in some cases, we embed ToC ideas into our automation systems. The fact that one model can be implemented and used during the design phase, and the same model can be used in the ramp phase, and used again in the manufacturing phase saves a tremendous amount of effort and provides a very large boost to continuity. For example, we are now in a position based on our work with ToC to design a WIP policy during the design phase, and to plug it into our automation system to be used in the ramp and manufacturing phases. This is a markedly different approach from that of considering the factory life cycle phases as separate minimally-communicating events.

Conclusions

The application of ToC to the design, ramp, and manufacturing components of Intel's capacity supply line has significantly benefited each individual component financially as well as benefiting the integration of the components. ToC has been a practical way to continuously improve the "more for less" mentality that pervades our capacity supply line, and it has enabled us to manage the inherent variability of availability of all of the resources that the capacity supply line depends upon.

The most obvious missing component of this story is the application of ToC principles to the build component of the capacity supply line. Thinking even more broadly, one might speculate about the magnitude of the benefits of applying ToC to the non-manufacturing components of the product supply line and the technology supply line. Pushing outside of the supply lines, one might inquire about using ToC in finance, human resources, or marketing. Given the steady increase in our rate of applying ToC based on our successes, it should not be too long before a description of our work in this area appears in this journal.

Acknowledgments

There are far too many contributors to this work to list here. The intellectual leaders are Bruce Sohn, Eamonn Sinnott, Bob Rodgers, Steve Notman, Ray Dudonis, Ken Gray, Greg Mazenko, John Bean, and Dane Parker. Management sup-

port has come from Mike Splinter, Bob Baker, David Marsing, Chuck Roger, Don Rose, and Gene Meieran.

References

- 1) E. Goldratt and J. Cox, "The Goal," North River Press, 1984.

Author's Biography

Karl Kempf is currently the Principal Scientist for Manufacturing Systems for Intel Corporation in Chandler, Arizona. He holds degrees in physics, chemistry, and applied mathematics/computer science and is interested in performance optimization in complex man-machine systems. While working at SEFAC Ferrari, he was a member of three world championship teams. At Pinewood Movie Studios, he was on the team that won an Academy award for Special Cinematic Effects for the Superman series of movies. His e-mail is karl.g.kempf@intel.com.