The Ongoing Challenge for a Responsive Demand Supply Network: The Final Frontier—Controlling the Factory

Kenneth Fordyce and R. John Milne

Abstract Over the past 20 years organizations have put significant energy into making smarter decisions in their enterprise wide central planning and "available to promise" processes to improve responsiveness (more effective use of assets and more intelligent responses to customer needs and emerging opportunities). However, firms have put only limited energies into factory floor decisions and capacity planning and almost none into generating a tighter coupling between the factory and central planning. The bulk of the work to make "smarter factory decisions" has focused on two simple metrics: increasing output and reducing cycle time—often without accommodating the need to run lots at different velocities and without recognizing how the operating curve (trade-off between lead time and tool utilization—Appendix 3) links them. In fact, many of the recent Lean initiatives have focused on eliminating variability to induce simplicity to achieve improved output or cycle time without concern for the impact on responsiveness or capacity. The purpose of this paper is to (a) make clear the critical, and often overlooked, role of factory responsiveness with respect to central planning; (b) explain how traditional factory planning and the current application of Lean can severely impact the firm's responsiveness; (c) elaborate on touch points between central and factory planning demonstrating simple tactical methods that can improve responsiveness and protect the factory from churn; (d) explain why smarter dispatch scheduling is critical to successful responsiveness; and (e) outline the basics of smarter dispatch scheduling. Although the focus of this

Tighter Coupling with Central Planning, Smarter Near Term Tactical Planning, and More Intelligent Dispatch.

K. Fordyce (✉)
IBM Systems and Technology Group, 2455 South Rd, Poughkeepsie, NY 12601, USA
e-mail: fordyce@us.ibm.com

R. J. Milne
Clarkson University School of Business, P.O. Box 5790, Potsdam,
NY 13699-5740, USA
e-mail: jmilne@clarkson.edu

D. Armbruster and K. G. Kempf (eds.), *Decision Policies for Production Networks*,
DOI: 10.1007/978-0-85729-644-3_3, © Springer-Verlag London 2012

paper is the factory, many of the core concepts apply to a wide range of industries from restaurants to health care delivery.

1 Positioning the Factory Within an Enterprise Wide Demand-Supply Network

Organizations, from healthcare facilities to manufacturing giants to small restaurants, can be viewed as an ongoing sequence of loosely coupled activities where current and future assets are matched with current and future demand across the demand-supply network.

These planning, scheduling, and dispatch decisions across a firm's demand-supply network are best viewed as a series of information flows and decision points organized in a decision hierarchy or tiers and further classified by the type of supply chain activity creating a grid for classification. The row dimension is the decision tier and the column is the responsible unit (Fig. 1). Observe the decisions in each tier limit and the options in the tiers below it.

The time frame for the first decision tier, *strategic planning*, is typically driven by the lead time required for business planning, resource acquisition, new product development and introduction, and to produce a product. Depending on the actual lead times for these activities, decision makers are concerned with a set of problems that are 3 months to 7 years into the future even with the same industry. For example, acquiring and validating a new tool may only take 3 months and re-orientating the product line takes 1 year. In both cases these decisions are removed from the production and delivery of current product. Issues in this tier include, but are not limited to, what markets the firm will be in, general availability of equipment and skills, major process changes, risk assessment of changes in demand for existing products, required or expected incremental improvements in the production or delivery process, and the lead times for adding additional equipment and skills.

The second tier, *tactical planning*, deals with aggregate level plans, estimates, and commitments. The time frame can range from 1 week to 6 months and is typically based on production lead times and the pace of change for demand and factory performance. Estimates are made of yields and cycle times (lead times), the likely profile of demand, productivity and reliability of equipment, etc. Decisions are made about scheduling releases into the manufacturing line or staffing levels. Delivery dates are estimated for orders or response times for various classes of patients are estimated. Deployment of equipment and staffing is adjusted. The order release plan is generated or regenerated, and (customer-requested) reschedules are negotiated.

The third tier, *operational scheduling*, deals with the daily execution and achievement of a weekly, biweekly, or monthly plan. Shipments are made, patients receive treatments, customers are waited on, serviceability levels are measured, and recovery actions are taken. Optimal capacity consumption and product output are computed. The time frame is again dependent on the production lead time and the rate of change in the factory.

Demand-Supply (DS)Network Planning, Scheduling, and Dispatch (PSD) Activity Areas and Decision Tiers			
		Enterprise Wide. global view - central planning	**Enterprise Subunits (manufacturing, distribution, retail) factory planning**
Decision Tiers	**Tier 1 Strategic**	**Enterprise wide Central Plan once or twice a year for 2–5-year horizon** at aggregate level with forecasted demand focused on business scenarios. Net result strategic direction established and financial commitments made	**Capacity Analysis t**ypically at tool family level and overall manpower to support forecasted demand, creation of production flow and capacity information for central plan, determining new production processes to introduce and estimated learning curve
	Tier 2 Tactical	**Enterprise wide central planning weekly/biweeky/monthly** > create demand statement (current orders, forecasts) > capture capacity, WIP, BOM, business policy > central planning engine to match assets with demand > estimate supply line linked to demand, early warning, production requirements, chase situations	**Capacity (tools and manpower) analysis** to gauge impact of changing product mix, identify challenges, review and modify deployment decisions and manufacturing engineering requirements, and **create capacity constraint information for central planning and WIP status**. Monitor tool level performance and take appropriate actions. Establish rules and metrics to set global lot importance - example, how many priority classes, algorithm to set lot importance within a class, limits on number of exppedites.
	Tier 3 Operational "daily"	**Enterprise Wide central planning reduced focus / what if** > what if commit on large orders > what if on major asset change > status of key WIP and actions to take if needed > cross factory signals	**Provide information to central plan and daily factory adjustments** > establish target outs, due dates on lots > maintenance priorities > short term changes in deployment > review key lot status and change priority (up or down) based on progress (eithermanually or dynamically) > one time changes in lot importance guidance > establish mfg lot vs development lot preference > revised projected outs for enterprise planning
	Tier 3.5 sub daily guidance	Change in Priorities, updated supply projections based on updated WIP or capacity status; change in customer reserved supply	**As needed Updates to Guidance to support response decisions** > regular updates to lot status based its progress, entering a time process window, status of short term manufacturing targets, WIP position and tool status > regular updates to tool status based on manufacturing engineering requirements, tool events, etc
	Tier 4 - Response	Available to Promise or Automated Order Commit process, cross factory signals	**Dispatch Scheduling & Tool Response** > assign sequence of lots to a tool > change status of a lot (for example on or off hold) > monitor signals from tools and respond as needed

Fig. 1 Decision grid for demand-supply networks

Tier "3.5" *straddles operational and real time response*. For example, a monitoring system might observe a lot has entered a "process time window" and its "urgency" to be assigned has "increased." A process time window is a sequence of activities that must be accomplished within a certain time limit or the lot might need to be scrapped due to some type of contamination. A non-factory example would be the

"triage" system that occurs regularly in an emergency room where a patient is placed in one of four or five categories based on urgency. Although this decision does not directly assign the patient to a healthcare provider, it has a strong influence over the type and urgency of the assignment.

The fourth tier, *real-time response system*, addresses the problems of the next hour to a few weeks by responding to conditions as they emerge in relevant time. Within the demand-supply network, relevant time response is often found in two areas: manufacturing dispatch (assign lots to tools) and order commitment (available to promise, or ATP). For the emergency room setting it would the initial assignment of the patient to a health care professional and then a sequence of assignments based on the initial review (for example go to X-ray, immediately call in the senior resident, and run a blood test).

Within manufacturing, the decisions made across the tiers are typically handled by groups with one of two responsibilities: maintaining an enterprise-wide global view of the demand-supply network and ensuring that subunits (such as manufacturing location, vendor, and warehouse) are operating efficiently. Ideally all planning would be centralized; in practice complexity precludes this. Capacity planning is a good example. At the enterprise level, capacity is modeled at some level of aggregation, typically viewing a key tool set as a single capacity point. At the factory level, each tool, or potentially each chamber within a tool, is modeled.

2 Challenges and Opportunities

Over the past 20 years organizations have put significant energy into making smarter decisions in their enterprise-wide central planning and "available to promise" process to improve responsiveness (more effective use of assets and more intelligent responses to customer needs and emerging opportunities) [9]. However, firms have put limited energy into factory floor decisions and capacity planning, and almost none into a tighter coupling between the factory and central planning. Much of the recent work to improve factory performance has attempted to implement Lean planning [6] concepts of (a) elimination of variability, (b) establishing uniform flow (every part every interval), (c) supermarket-like goods flow (kanbans), and (d) elimination of due dates and on time delivery metrics. Clearly, every factory will run "better" with steady output and predictable lead times—however, the real world always injects variability that sets the price of implementing such methods as reduced responsiveness and/or excess capacity.

The net result is that many factories still operate with the mind set: "establish a set of starts for the month; set a fixed schedule with target outs; and measure actual outs versus target outs." For this approach to work, demand must be accurately forecasted over an extended period of time and uniformly spread across time; all lots must travel at the same speed; tool sets should operate with clockwork precision (never suffering "surprises"); and the flow of parts in the line (even with stable capacity) must never create "piles" or "gaps" due to the variations (for example batch versus single lot tools) intrinsic in the manufacturing process. In today's world, accurate detailed

forecasts of demand remain an illusion; even the best factories which have "tool set surprises" (breakdowns and quality excursions), product mix introduces variability in speeds, and the competitive nature of the market precludes carrying excess capacity and insists on responsiveness. Those demand-supply networks that can get their factories more engaged in responsiveness while recognizing the importance of "tools" and "output" will flourish. They will eliminate the variability that matters—a failure to deliver a part on its committed date and the inability to capture a market opportunity that could be handled with "intelligent" factory decisions.

Accomplishing this goal requires "retooling" the approaches for interaction with central planning, near-term tactical planning, and dispatch scheduling to be more adaptive without a loss in productivity. Bob Bixby observed—the optimization time horizon is ever shrinking. The purpose of this paper is to (a) make clear the critical, and often overlooked, role of the factory within central planning; (b) explain how traditional factory planning and the current application of Lean methodology can severely impact the firm's responsiveness; (c) elaborate on touch points between central and factory planning demonstrating simple tactical methods can improve responsiveness and protect the factory from churn; (d) explain why smarter dispatch scheduling is critical to successful responsiveness; and (e) outline the basics of smarter dispatch scheduling.

3 Basics of Enterprise-Wide End-to-End Central Planning

To understand how factory floor decisions can limit responsiveness in central planning, we need to review the key elements of central planning which are given in the list below and in Fig. 2 [9, 31].

1. Create a demand statement
2. Gather and collect key supply information from the factory
    - 2.1. Project the completion of WIP to a decision point (often completion of the part).
    - 2.2. Statement of capacity consumption rates and capacity available.
    - 2.3. Statement of lead time or cycle time to complete a new start.
3. Create a model that captures key enterprise relationships of the demand-supply network (Central Planning Engine—CPE).
4. Create an enterprise-wide central plan by matching current and future assets with current and future demand using the CPE to create a future projected state of the enterprise and the ability to soft peg the current position of the enterprise to the projected future position. Information from the CPE model includes
    - 4.1. Projected supply linked with exit demand
    - 4.2. Identification of at risk customer orders either to a commit date or request date
    - 4.3. Synchronization signals across the enterprise
    - 4.4. Capacity utilization levels

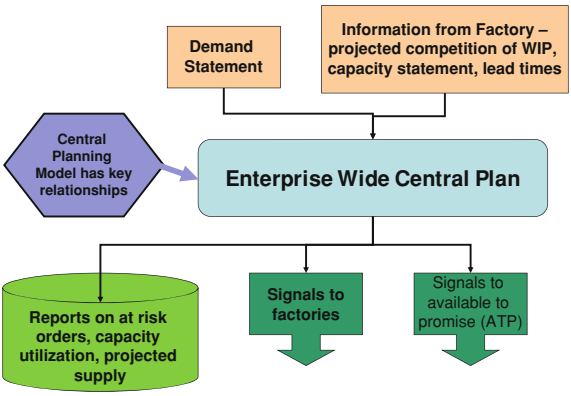

Fig. 2 Basic steps in central planning

4.5. Ability to trace each production and distribution activity that supports meeting a demand.

5. Typically this is an iterative process where each iteration is done with different assumptions and conditions, for example

   5.1. Different assumptions about capacity available
   5.2. Different business policies for protective stock
   5.3. Different commit dates and/or demand priorities for orders

6. Execute the plan, that is,

   6.1. Send signals to each manufacturing facility and distribution center
   6.2. Send projected supply to available to promise (ATP)

4 Basics of Factory Planning and Dispatch

For central planning organizations, the plan (developing the plan, making the customer commits, and monitoring conformance to the plan) is their primary end product. For the factory, planning and dispatch are always secondary in importance to successfully making the parts. Factories are foremost concerned about making a product that works (yields), second keeping their tools operational, and third keeping their output levels high (either operational outs or exit outs).

Dispatch refers to assigning a lot to a tool and requires balancing effective tool utilization with stable delivery (either to the commit date on the lot or to the number of exit lots per day or week). Factories see dispatch as important since it generates the assignment of lots to tools and therefore impacts output. Typically there are different camps with substantially different views on how to make this decision. For example, manufacturing is looking to maximize output while the business team is as concerned about the lots for key clients. Both groups are always suspicious of the applications that do automated dispatch combining rules and models—either

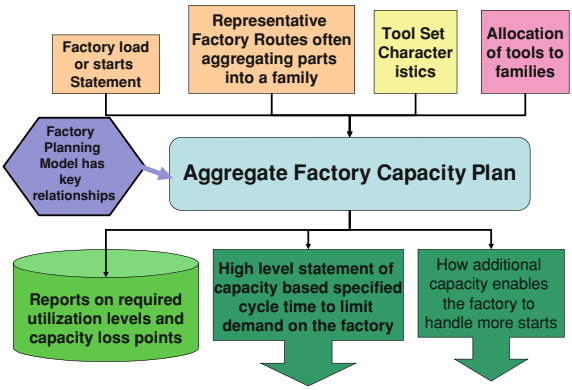

Fig. 3 Aggregate tool planning steps

thinking the automated methods are too simple and they know best how to balance complex trade-offs or "complaining" the automated decisions are too complicated to understand.

Aggregate tool planning [1, 34, 38] is typically focused on assessing the ability of the factory to satisfy certain demand (demand is stated as manufacturing starts a.k.a. planned manufacturing releases) and creating the capacity inputs required by central planning. The basic steps of aggregate factory planning are given below and in Fig. 3:

1. Capturing representative product routes—sequence of operations, raw process time for each operation, and tool set consumed for each key operation.
2. Capturing a specific factory load—typically given in starts.
3. Gathering data on tool set characteristics: number of tools, tool availability, which operations the tool set handles, overlapping tool sets (shared operations between toolsets called a cascade [1], and its operating curve (Appendix 3, [2, 5, 12, 14, 15]) which establishes the trade-off between cycle time and tool utilization/capacity available.
4. Allocation of tool sets to product parts or families either as user input or based on history
5. A model that captures key relationships—often a spreadsheet based single iteration model
6. Executing the model to determine how this load impacts toolsets
   6.1. Required utilization levels
   6.2. Capacity loss points (planned maintenance, high raw process time, etc.)
7. Providing information for central planning
   7.1. High level statement of capacity based on starts or a few key tool sets fixed for the specified cycle time to limit demand on the factory
   7.2. How additional capacity enables the factory to handle more starts

Near term tool planning (deployment) refers to determining which operations each tool will be qualified to handle over the short term. Typically a tool can service many operations, but a factory will limit the number of operations it is "allowed" to service to reduce workload on manufacturing engineering and make dispatching simpler by reducing options.

Observe the lack of influence of an order book on factory planning and decisions.

5 Current Interaction Between Factory and Enterprise: Factory Limits Responsiveness—Opportunities Abound

Steps 3 and 4 in Sect. 3 on the central engine planning process (CPE) are often viewed as the planning "hub" and the focus of making a firm more responsive through "smarter" engines and better (quality and timeliness) data [9]. However, in Step 2 the factory sets the boundaries of "responsiveness" [29]. The CPE relies on the factory to provide

- Estimated completion date for each lot in the line (either to completion or staging point)
- Statement of capacity available and required for each manufacturing start (typically at an aggregate level)
- Estimated lead or cycle time to complete a start fixed for some time interval

Additionally,

- The central planning process cannot change the due date on the lot or the lot's priority without extensive manual negotiations with manufacturing
- Central planning has no control (and typically no knowledge) of the lot importance metric used by the factory or how it balances utilization and delivery.
- Each piece of information supplied by the factory to the central planning process is "fixed"—stripped of all of the information that enables trade-offs to be made. For example, the following possibilities are invisible to central planning:
  - Slowing one lot down to enable another lot to go faster
  - Trade-offs between cycle time and capacity available based on the operating curve
  - Redeploying tools to handle a different mix of manufacturing processes or products

Additionally, limiting "change or variation" within the "factory black box" to improve responsiveness fits the factory culture and is reinforced by Lean principles. Factories "dislike" change. Factories "like" steady rates of production referred to as smooth flow. This has been reinforced with Lean initiatives that view variation as evil and desperately attempt to create smooth demand and production flow with aggregation and kanbans or "super markets" that essentially serve as inventory replenishment decision points to absorb variability and generate smooth (every part every interval)

production flow in the factory [6]. They try justifying this by claiming all variability can be eliminated and higher productivity will occur. The reality is this view

- Requires excess capacity to facilitate the "smoothing."
- Is completely divorced from client needs, variability in production flows, and tool availability.
- Has no inherent ability to allocate scarce capacity or project a supply line.
- Fails to account for the operating curve.

Despite the substantial forces to limit change, constant pressure from emerging market opportunities to manufacturing quality excursions to inaccuracies in planning (deviations between the plan and the actual) drive an ongoing sequence "off line one of a kind" negotiations between the central planner and the factory planner to make adjustments that rely on quasi-manual decision support tools with limited function. For example

1. A client may need three lots 4 days earlier than committed and this can be accommodated by placing these lots on expedite.
2. The demand for product A requires 30 units of capacity from Tool Set A1 on average each day. Tool Set A1 only has 25 units of capacity available. Tool Set A2, which is not listed as a capacity option for product A, can service product A, but it runs slower. A review of capacity utilization for Tool Set A2 indicates it will be underutilized. A decision is made to qualify Tool Set A2 to handle product A.
3. A client has had a steady order for 10 units daily of product W with a cycle time of 15 days where the constraining tool set is Tool Set W3. The business has been able to achieve an on time delivery rate of 97%. The client would like to increase its standing order from 10 to 12 units. Central planning initially rejects this opportunity since the stated maximum daily capacity in their model for Tool Set W3 is 10. However, when the two planners look at the details of the tool set and its operating curve, the business decides it can commit to 12 per day if the cycle time is increased to 16 days (or if the OTD commit percentage is lowered).
4. Assume a client has placed an order for five lots of "part A" per day with a cycle time of 10 days. On average there are 50 lots of "part A" in WIP and the factory completes five lots per day. The due date posted on each lot is the start date plus 10 days. For example, lots started on day 6 are due on day 16. Due dates on the lots can only be changed manually by a factory planner. The factory has an abnormal set of tool outages and goes 3 days without delivering any lots—it is past due 15 lots ($= 3 \times 5$). It has continued to start five lots per day. At the start of the fourth day the number of lots in the line is 65 (=50 normal + 15 past due). On the evening of the third day the client and Central Planning meet about a recovery strategy. The client determines demand has been soft for this part and agrees to "forgive" five lots and have the remaining ten lots "caught" up at a pace of one per day (in addition to the regular five per day). Therefore the new order book for this client is six lots per day for the next 10 days and then returns to five per day. Without changes to the due date on the lots in WIP, the factory continues to see 15 lots past due and will drive to "catch up" as quickly as possible. The factory

may decide to delay lots for a second client to catch up all 15 past due lots for the first client in 5 days. Therefore the factory planner has to manually change the due dates on the lots to insure the factory floor has the correct guidance.

These examples make it clear that when central planning can make effective use of the flexibility within the factory that is hidden from its traditional view—good things can happen. The opportunity for improved responsiveness simply needs to "widen and straighten" this trail with appropriate planning and dispatch tools, processes, and protocols. Each one would be considered muda (wasteful) by Lean which would say to eliminate them, not to build tools to make doing this more intelligence and efficient. We contend a goal of any firm is to eliminate unnecessary complexity, but ignoring the complexity that remains is like tackling snow storms in the north with bald tires.

Additionally, such tools and processes can keep bad things from happening. For example, if many lots are being "expedited" already there is no room for an additional expedited lot. Tool W3 may be needed for engineering lots not in the central planning data or it may have a history of time consuming qualifications making it too large a risk. Just as the factory prefers "steady" and conservative, central planning often fall preys to an overly optimistic mindset that is fine with constant churn. Tools for improving factory/enterprise coordination fall into three groups

1. Direct interaction with central planning tools (for example WIP projection, expedite decisions, demand pegging, and specialized capacity planning models for flexibility in manufacturing)—Sect. 6
2. Tactical decision models (tool deployment, allocation of cycle time, or tool capacity referred to as operational outs or moves)—Sect. 7
3. Dispatch scheduling (assigning lots to a tool)—Sect. 8

The following sections outline methods in each of these three areas that can provide additional flexibility to the factory without destroying factory output and cycle times. The differences in the length of the sections reflect the amount of detail required to convey the core issues, rather than the relative importance of each area.

6 Dynamic Interaction Between Central Planning and Factory Planning

As previously described in Sects. 3 and 4, the contact points between the factory and central planning include:

- Accurate projection of when lots already in the factory (WIP) will arrive at stock
- Setting due date for lots.
- Changing the committed date or speed for lots.
- Capacity and cycle time information that influence planning manufacturing start decisions and customer commits.

We will explore examples in each of these four areas.

6.1 Smarter WIP Projections by Considering Capacity

Typically each part moves sequentially through a set of manufacturing steps (route) that can be characterized by a raw process time (RPT) at each step, a cycle time multiplier (CTM) that adjusts for the average wait time, total cycle time (TCT) which is RPT × CTM, and the tool set that handles this manufacturing step or activity. A sample route is provided in Table 1.

A factory planner typically uses one of two methods to project when a lot will finish

- Use the commit date for the lot
- Add the remaining cycle time for the lot to the current date. For example if lot 101 was at step 04, we would project its completion date to be NOW + 91 h (= 40 + 41 + 10).

The following methods have proven effective in improving the quality of this projection:

1. *Status of the lot at the current step*: Instead of solely using the TCT to estimate the time a lot will spend at its current step, directly examine the number of lots that are expected to be processed ahead of this lot at this step and adjust for their processing times.
2. *Different CTM estimators*: Typically the CTM is based on a planned value. The quality of this estimate can sometimes be improved using recent manufacturing history to create an estimated CTM for the next 7–14 days.
3. *WIP Projector on a Parcheesi Game Board*: In this method [9] we project the movement of each lot step-by-step according to its cycle time, but incorporate capacity constraints by limiting the number of moves (i.e. number of operation completions) (or time) per day allowed at certain tool sets and allocating these moves based on lot importance. For example, assume LOT201 and LOT202 are at Step 01 and LOT301 and LOT302 are at Step 05. All four lots are serviced by the lion tool set for their present steps. LOT 201 is an expedited lot, LOT 302 is three days behind schedule, and the other two lots are on time. Additionally, the lion tool set has a daily capacity limit of two lots per day. Only LOT201 and LOT302 would move to the next manufacturing step on the game board today (LOT201 to Step 02 and LOT302 to Step 06). The other two lots would have a chance to move tomorrow based on how their priorities rate relative to the competition from other lots.
4. *Queuing Network Equations*: The most complicated, but also the most accurate and flexible is to represent the route in its entirety as a system of queuing network equations. This has been used successfully in some situations [5, 37, 38].

The caveat in each option is to avoid "thrashing" the estimate by over reacting to the normal day-to-day variations in manufacturing flow. When a manufacturing line has sufficient buffer capacity and is appropriately managed (e.g. not too many

Table 1 Route or process steps to manufacture a part

Manufacturing step	Raw process time (RPT) for this step	Cycle time multiplier (CTM) applied to RPT	Total cycle time (TCT) for this step = RPT* multiplier	Tool set
Step 01	10.0	3.3	33.0	Lion
Step 02	15.0	3.8	57.0	Tiger
Step 03	8.0	4.0	32.0	Apple
Step 04	20.0	2.0	40.0	Furnance
Step 05	10.0	4.1	41.0	Lion
Step 06	5.0	2.0	10.0	Squirrel
Total / average	68.0	3.1	213.0	

jobs being expedited, commitments are reasonable), most lots will complete at or near their original commit date by reallocating capacity from lots that are ahead of schedule to lots that are behind. The goals of the projection mechanisms are: to identify when attempts to reallocate capacity so all due dates are met is not likely to succeed; find lots that have fallen too far behind to finish by their commit date; and issue an early warning when it is clear that the assumptions in the planning model are at variance with reality—resulting in many lots finishing late (or early). The goal is not to overreact to normal fluctuations, but catch systemic issues—especially since factories are notorious for "convincing" themselves they will catch up next week once tool availability stabilizes.

6.2 Dynamically Resetting Due Dates

In Sect. 5 we described a situation where the factory fell behind in meeting its commitment to a client, the central planning organization worked with the client to reset the demand, and then the factory planner needed to manually recalculate and reset the due dates on the lots. This is one example of many where the actual "need" date for a lot is different from the due date posted to the lot when the lot starts. A second example occurs when a lot (LOT51) started on day five is placed on hold for 2 days and another lot (LOT61) started on day six "passes" it. If we assume a cycle time of 10 days, then LOT51 has an initial due date of 15, but is behind (further from finishing than) LOT61 with a due date of day 16. This is called "leapfrogging."

Typically when the demand driving that starts on a factory is a complex combination of fixed orders, loose orders, build to forecast, line stock replenishment, etc, and the manufacturing process is long and complex the eventual need date will often be different than the initial due date. A tool to dynamically reset the "due date" on the lot to match real need improves responsiveness. The algorithm works essentially as follows [9, 25].

1. The factory must maintain a "demand" statement or order book on what the business currently expects it to produce (part, quantity, and date). This is the hardest part.
2. The lots of the same part are sorted according to raw process time remaining (low to high).
3. The demand with the nearest (earliest) due date is assigned to the lot with the least remaining raw process time remaining and the pattern continues.
4. Some adjustments need to be made if the lots have different quantities, some lots are manufacturing expedites, etc.

This is a standard MRP algorithm that can and does enable the factory to appropriately focus energy on lots and facilitates early warning when a demand will not be met on time. This is the same MRP that is constantly maligned by Lean advocates. It would appear that having quality need dates would be an asset in any factory concerned about on time delivery.

6.3 Committing Some Lots to Run a little Faster: Collateral Impact

A common, yet manual, practice is for central planning to negotiate with factory planning to "speed" up certain lots to meet a customer request, overcome a manufacturing delay, or compensate for a planning failure. Typically, the analysis is limited and ad hoc with no comprehensive process as seen in central planning or tool planning. There are rules of thumb such as:

- The number of "expedite" lots cannot exceed some fixed number *N* or a maximum percentage of total lots.
- The fastest an expedite lot can run is some CTM less than that of the normal lots. If normal lots have a CTM of five, expedite lots might have a CTM of three.

A closer look makes it clear the core of this decision is a reallocation of either wait time or factory moves [23] that enables some lots to run faster by having others run slower over some subset of the manufacturing line over some time duration. For example, assume the factory has five lots (LOT01 . . . LOT05) in the last stage of production; each lot requires four moves (manufacturing actions) to complete; the maximum number of total moves (capacity) per day is five; and the most moves a lot can have in 1 day is two. If capacity is allocated "fairly," then each lot gets one move per day and each lot finishes in 4 days. Now assume the business decides LOT01 and LOT02 must finish in 2 days, then each needs two moves per day for 2 days, and therefore on each of these days two of the other three lots sit "idle" during these 2 days to enable this expedite.

Appendix 1 develops this allocation concept in more detail focusing on wait time allocation instead of moves. In each case lots that look essentially the same are required to run at different speeds on the factory floor. The planned speedup is

worthless without successful factory execution. This places a substantial burden on dispatch and precludes the use of simple methods (and Lean favorites) such as first in first out (FIFO) and elapsed time. Again we see that Lean and responsiveness are not in sync.

This topic is part of an area called *General Plan Repair Process*. In this process central and factory planners identify actions to take that will enable orders that are currently flagged as "late" to be met on time. Fordyce et al. [9] reviews this challenge from the central planning perspective—only through increased intelligence on both sides of the fence can responsiveness be improved.

6.4 Smarter Central Planning Through Better Modeling of Factory Capacity

As we outlined before, the central planning process requires as critical inputs from the factory: capacity (consumption rates and availability) and cycle times. Since the 1980s manufacturing resource planning (MRP) and material balance equations (MBE) in optimization formulations have been the two dominant methods used in central planning [30]. In these methods the factory representation is "static" and linear. The cycle times and capacity information are fixed across some time period and handled with linear relationships. For detailed information about central planning, the reader is referred to Refs. [9, 13, 20, 28, 31, 35, 36].

Historically intricacies of factory tool planning (availability, deployment decisions, cascading, setup times, batching, et al.) and the dynamic interaction between equipment utilization (effective capacity) and cycle time through the operating curve have for the most part been ignored. This will not be sustainable in the future as the burden on responsiveness resulting in under utilization or delivering products late is increasingly unacceptable.

In Sect. 5 we described a situation where the client needed 30 units per day and the initial central planning analysis determined the maximum the factory could make was 25. Appendix 2 elaborates on the method to improve responsiveness by capturing alternative deployments of tools to manufacturing operations.

In the same section we outlined that we could trade longer cycle time for more tool capacity (and hence output) based on the operating curve. Appendix 3 contains a simple example that makes it clear the assumption in typical central planning processes that cycle time and capacity are independent is not correct—the two are clearly coupled. We can view this as classical planning meets its uncertainty principle. It is a rich ground for improved responsiveness and a headache for classical planners. Since Lean advocates believe that all variation can be eradicated, it has no awareness of an operating curve, and no methods to capture this opportunity. It is like attempting to ignore special and general relativity and still produce GPS locations [26].

For additional information about work that pushes beyond traditional methods for handling capacity in central planning and factories see: [1, 7, 19, 22, 27, 38].

7 Tactical Decisions in the Factory: Only the Shadow Knows

There are series of ongoing tactical decisions in factories that fall well below interaction with central planning and are not part of dispatch—but have a strong influence on dispatch by constraining the available options to assign lots to tools. We refer to these as the "shadow" decisions—powerful, but difficult to find and capable of substantially restricting responsiveness.

One area is manufacturing engineering requirements (MER). Manufacturing Engineering's (ME) first concern is producing quality products (keeping high yields) and in their zeal can create collateral damage. For example, assume tool A01 is being qualified to run a new process called "yellow tiger"; ME might put in place two rules:

- Only 25% of the total widgets produced over a 24 h period can run on tool A01 (the other 75% has to run on other tools in the tool set) in case tool A01 has quality issues.
- Most of the other processes that can run on tool A01 are soft coded as not available to tool A01 to insure enough widgets for the "yellow tiger" process visit this tool.

On the surface, this sounds logical. In practice, especially when the factory is busy, most simple dispatch decisions systems (automated or human) will initially drive "yellow tiger" widgets to tool A01 and place other widgets on the other tools in the toolset. However, quickly the ME "police" will shutoff assigning these widgets to tool A01 since the 25% limit is met. Typically, there is no method to increase the importance of running "yellow tiger" widgets on the other tools or dynamically alter either rule. Some simple tactical models and dynamic guidance (defined in the next section) will catch this imbalance before it becomes an issue that can, in the heat of the "battle," take days to find without the appropriate diagnostic tools. As Gary Sullivan [33] observed—it is usually better to blow out the lighted match before it gets to the gasoline unless you are measured by putting out fires as opposed to preventing them!

A second area is called deployment decisions. Here the tools that make up a group of similar tools (toolset) are allocated to the operations covered by this toolset. Again this limits the dispatch options. Appendix 4 describes an approach that helps us gauge near term the effectiveness of the deployment decisions for the WIP currently waiting to be processed. A third area is the deployment of manufacturing operators. Again, nothing in the Lean literature tackles these tough questions that live in the shadows.

8 Fundamentals of Dispatch Scheduling for Better Factory Performance

As we observed in the prior sections, for the factory to be responsive, simple dispatch applications are insufficient to ensure planned actions are executed on the floor. In addition, simple dispatch cannot handle the ever increasing complexity and variability factories face on a daily basis—from manufacturing equipment whose throughput is very sensitive to batch sizes and the sequence lots are placed on the tool; diversity in the product mix and quantity which eliminates the ability to run a fixed quantity per day and still meet client expectations; ever tighter boundaries on quality control; competitive pressures that require factories to run at higher utilization rates without an increase in cycle time, more specialty and design parts, etc. Smarter dispatch is required to offset increases in variability and keep the operating curve from shifting in an unfavorable direction.

Essentially, the factory is constantly balancing effective tool utilization with stable delivery against a complex demand statement. This drives the requirement for intelligent dispatch scheduling applications to optimally achieve these goals and limit the quantity of variability the factory introduces into the system. This leads to simple applications for dispatch scheduling being replaced with applications that make the "complex" manageable. Some of the essential components of dispatch scheduling are given in the next subsection. For a comprehensives review of this topic the reader is referred to: [3, 8, 11, 16–18, 32, 33].

8.1 Basics of Dispatch Scheduling

The key inputs to dispatch scheduling can be broken down as follows:

1. Tool—Lot affinity (usually linked by the operation)
    - 1.1. What lots can run on this tool? What tools can handle this lot?
    - 1.2. What are preferred tools? What are preferred lots?
    - 1.3. Manufacturing engineering requirements
        - 1.3.1 Count limits (avoid too many lots on certain tools)
        - 1.3.2 Time limits (tool requires re-qualification after a specified amount of activity)
        - 1.3.3 Process time windows (lot must finish a sequence of steps within a time limit)
        - 1.3.4 Special customer specifications
2. Global importance of the lot to the supply chain or business—priority
3. Pacing lot movement: fluctuation smoothing, flow balance cycle time allocation, delta schedule, critical ratio

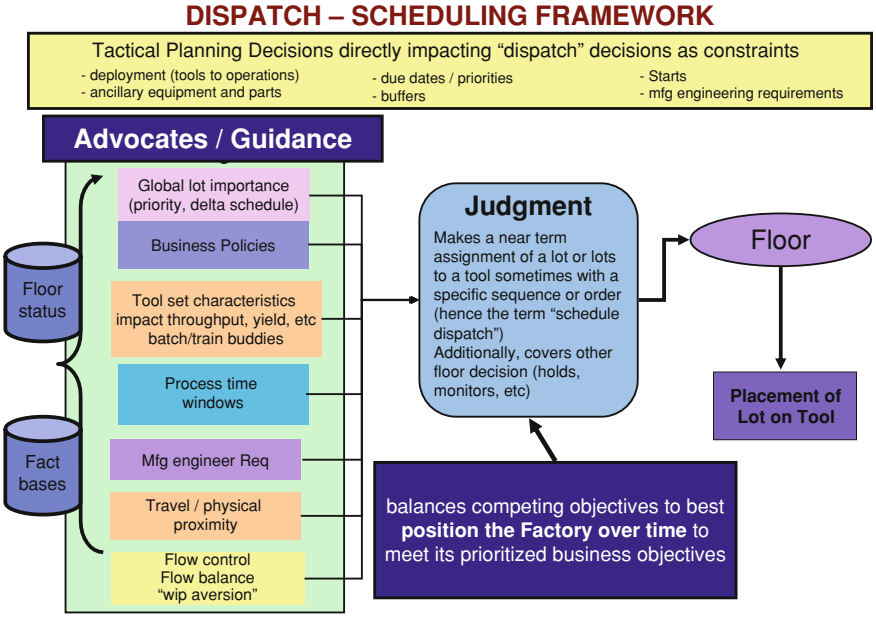

Fig. 4 Dispatch scheduling framework

4. Local tool characteristics and performance
    - 4.1. Batching and operational trains
    - 4.2. Setup times dependent on previous job run at the tool
    - 4.3. Parallelization opportunities
    - 4.4. Differences in raw process time
    - 4.5. Multiple chambers within a tool
    - 4.6. There may be ancillary equipment required at an operation in addition to the core toolset and labor
5. Upstream and down stream requirements
    - 5.1. Sending wafers to tools with limited WIP in queue in front of them
    - 5.2. Avoiding tools with large piles of WIP in queue
    - 5.3. Balancing across repeated levels which use the same tool set

The core dispatch decision-making activities can be divided into two primary components: guidance and judgment (Fig. 4). Appendix 5 has additional details on both components.

9 Conclusion: Slow Steady Progress in Extending the Borders of Bounded Rationality

Herbert Simon [29] observed, "As humans, we have 'bounded rationality' and break complex systems into small manageable pieces." The challenge for organizations is to integrate information and decision technology to push boundaries out and improve performance. Nick Donofrio [4], retired IBM Senior Vice President, observed, "Access to computational capability will enable us to model things that would never have believed before." The challenge reaches beyond coding algorithms, linking to data, and turning it on. Each decision-science team must execute its role as "intelligent evolutionist" to ensure the organization adopts complex decision technology in a sustained incremental fashion. Each management must be willing to push their organization beyond its comfort zone.

Little [21] observed: "Manufacturing systems are characterized by large, interactive complexes of people and equipment in specific spatial and organizational structures. Because we often know the sub units already, the special challenge and opportunity is to understand interactions and system effects. There are certainly patterns and regularity here. It seems likely that researchers will find useful empirical models of many phenomena in these systems. Such models may not often have the cleanliness and precision of Newton's laws, but they can generate important knowledge for designers and managers to use in problem solving."

Improving responsiveness in the factory is one of the most difficult challenges in the near-term horizon, but clearly one of the most important. For many firms substantial gains in end-to-end supply chain responsiveness is limited by the modeling tools and approaches in factories for matching assets with demand and flowing production and the false illusion from Lean advocates that variability and complexity can be eliminated.

Appendix 1: Committing Some Lots to Run Faster—Collateral Impact

Deciding Which Lot are Candidates to Speed Up

Prior to any allocation decisions, the planners must first decide candidate lots to speed up. This requires two critical pieces of information:

- An assessment of whether the lot is currently behind or ahead of schedule
- The exit demand supported by the lot

The second requirement places a burden on the central planning process to be able to link each lot to a specific exit demand(s) and trace all intermediary manufacturing steps. Since this linkage can and will change, creation of the linkage must be dynamic.

This is called *demand pegging* or *coverage analysis* [9] and responsibility for this foundation of factory responsiveness belongs to central planning.

Model 1: Expediting a Set of Lots from Release into the Line

Assume the factory makes two parts (A and B) with the following routes (Tables 2 and 3).

Each lot for Part A (Table 2) goes through six manufacturing steps and takes 20.7 time units to complete. Of this 20.7 units, 6.8 represent actual processing time and 13.9 is wait time. Part A spends 67.1% (= 13.9/20.7) of its time waiting and 32.9%(= 6.8/20.7) of its time being processed. Each lot for Part B (Table 3) goes through four steps (different than Part A) and takes 15.8 time units. Of this 15.8 units, 4.0 represent actual processing time and 11.8 is wait time. Part B spends 74.7% (= 11.8/15.8) of its time waiting and 25.3% (= 4.0/15.8) of its time being processed.

A move is defined as the completion of one manufacturing step. Part A accomplishes six moves in 20.7 units of time. Therefore it averages 0.290(= 6/20.7) moves per unit time in the factory. Part B lots average 0.253 (= 4/15.8) moves per unit time. The flow information for the average lot for each part is summarized in Table 4.

Table 4 also contains information on Part B* which is exactly the same as a Part B lot, but travels faster ("fast track" or expedite lots). Each B* lot goes through the same steps as a regular Part B lot and incurs the same RPT. The difference is average wait time is smaller generating a smaller total cycle time (CT) and CTM. In this example B* has a CTM of 2.50 (= 10/4). Since it moves faster than regular Bs, its move per unit time is higher 0.400 (= 4/10) compared to 0.253. We will refer to this set of parts with these cycle times as Case 1.

Of particular importance in Table 4 is the last column—average wait time per unit time. It is the amount of wait time on average a lot sees per unit time. A lot is in one of two states—waiting to be processed or being processed (a move). Therefore for each part "actual process time (RPT) per unit time" (+) "wait time per unit time" equals 1. In Table 4 we observe the sum of the last two columns for each part is 1.

Now assume the factory starts 10 Part A lots, 20 Part B lots, and 0 Part B* lots per unit time. Then it has a total of 207(= 10 × 20.7 = starts per day × total cycle time) Part A lots, 316 (= 20 × 15.8) Part B lots, and 0 (= 0 × 10) Part B* lots in the line. The total accumulated wait time per unit time for Part A lots is 139 (= 207 × 0.671 = total lots in the line for this part times average wait time per lot for this part per unit time) and 236 (= 316 × 0.747) for Part B lots. Wait time for Part B* is 0, since no Part B* lots were started. This information is summarized in Table 5.

Assume this load on the factory (10 starts per unit time of Part A at CTM = 3.04 and 20 starts per unit time of Part B at CTM = 3.95) leaves little unused capacity—putting the factory on the steep part of the operating curve (Appendix 3 [14, 23]). Therefore any attempt to make some lots run faster requires other lots to run slower. The slower lots have to absorb wait time from the faster lots.

Table 2 Route or process steps to manufacture a Part A

Manufacturing step	Raw process time (RPT) for this step	Cycle time multiplier (CTM) applied to RPT	Total cycle time (TCT) fot this step = RPT* multiplier
Step 01 (for Part A)*	1.0	3.3	3.3
Step 02	1.5	3.8	5.7
Step 03	0.8	4.0	3.2
Step 04	2.0	2.0	4.0
Step 05	1.0	3.5	3.5
Step 06	0.5	2.0	1.0
Total / average	6.8	3.04	20.7

* Steps for Part A and Part B are not the same

Table 3 Route or process steps to manufacture a Part B

Manufacturing step	Raw process time (RPT) for this step	Cycle time multiplier (CTM) applied to RPT	Total cycle time (TCT) fot this step = RPT* multiplier
Step 01 (for Part B)*	0.8	3.4	2.7
Step 02	1.4	4.5	6.3
Step 03	1.2	4.0	4.8
Step 04	0.6	3.3	2.0
Total / average	4.0	3.95	15.80

* Steps for Part A and Part B are not the same

One way to capture this constraint is to assume the total wait time for all lots in the line per unit time is fixed ($375 = 139 + 236$). When we sum the wait time for all lots in the line, it must be equal to this value. If I reduce wait time on one lot by two units, another lot has to gain wait time by two units. With this constraint we can use a model to gauge the impact of a decision to have some lots run faster. For example:

Assume Central Planning decides it needs five out of the 20 Part B lots started each day to complete production in ten time units and makes them "fast track" Part B* lots (Table 6). This is called Case 2. The total wait time burden per unit time for the Part B* lots is 30. Five starts per day for 10 days puts 50 lots in the factory. Each has an average wait time burden per unit time of 0.600 which results in a total wait time burden for the 50 Part B* lots of 30 ($= 50 \times 0.6000$). The remaining 15 B lots have a wait time burden of 177 ($= 15 \times 15.8 \times 0.747$). The total wait burden on B (regular and fast track) is 207 ($= 30 + 177$) compared to the prior burden of 236 units. We are now "short" 29 time burden units ($= 236–207$)—some lots have to gain waiting time.

Cycle time balance and wait time conservation require other lots in the line to absorb the 29 units of wait time. There are many possible solutions; one is the normal CTM for regular Part B lots increases from 3.95 to 4.44, which increases

Table 4 Flow information for average lot for each part Case 1

Part	Total RPT	Total CT	Total wait time	Number of steps	Ave RPT per step	Average CTM	Average wait time per step	Moves per time unit	Actual process time (RPT) per unit time	Wait time per unit time
Part A	6.8	20.7	13.9	6.0	1.13	3.04	2.31	0.290	0.329	0.671
Part B	4.0	15.8	11.8	4.0	1.00	3.95	2.95	0.253	0.253	0.747
Part B*	4.0	10.0	6.0	4.0	1.00	2.50	1.50	0.400	0.400	0.600

Table 5 Flow information on all lots in the line Case 1

Part	Starts per time unit	Total lots in line	Total moves per time unit	Total process time (RPT) per unit time	Total wait per unit time
Part A	10	207	60	68	139
Part B	20	316	80	80	236
Part B*	0	0	0	0	0
Total	30	523	140	148	375

Table 6 Flow information on all lots in the line Case 2

Part	Starts per time unit	Total lots in line	Total moves per time unit	Total process time (RPT) per unit time	Total wait per unit time
Part A	10	207	60	68	139
Part B	15	237	60	60	177
Part B*	5	50	20	20	30
Total	30	494	140	148	346

the normal cycle time to 17.8 from 15.8. This is called Case 3 and the details are in Tables 7 and 8.

From this base the collateral impact can be estimated through trial and error with a spreadsheet model or with an optimization formulation. It should be noted, this approach makes some simplifying assumptions that are not a problem when considering the impact of minor adjustments; major ones would require more elaborate models. *The goal is to get the central planner and factory planner to develop a process and formally recognize waiting time constraints and trade-offs.*

Model 2: Expediting Lots Close to the End of the Manufacturing Line

Often the decision to expedite does not occur for all of the lots of a certain part or group in the manufacturing line, but only for selected lots in the final third of their route. At this juncture planners have a reasonable sense of the likelihood these lots will complete on time and their importance to clients. Planners will look to selectively pick key lots to push faster (expedite) for four reasons: (a) attempt to finish on time; (b) attempt to finish early to meet a customer request; (c) meet quarter end revenue targets; or (d) build some buffer to insure the lots finishes on time. The following example illustrates a simple methodology to organize this decision process.

Table 7 Flow information on individuals Case 3

Part	Total RPT	Total CT	Total wait time	Number of steps	Ave RPT per step	Average CTM	Average wait time per step	Moves per time unit	Actual process time (RPT) per unit time	Wait time per unit time
Part A	6.8	20.7	13.9	6.0	1.13	3.04	2.31	0.290	0.329	0.671
Part B	4.0	17.8	13.8	4.0	1.00	4.44	3.44	0.225	0.225	0.775
Part B*	4.0	10.0	6.0	4.0	1.00	2.50	1.50	0.400	0.400	0.600

Table 8 Flow information on all lots in the line Case 3

Part	Starts per time unit	Total lots in line	Total moves per time unit	Total process time per unit time	Total wait per Total wait per
Part A	10	207	60	68	139
Part B	15	266	60	60	206
Part B*	5	50	20	20	30
Total	30	523	140	148	375

In this example (Table 9), the factory has six lots with only about 80 h of raw process time remaining (REMRPT) until completion. At this juncture the planners review these lots, assess whether are they are ahead or behind schedule to finish on their commit date, and decide if certain lots should be "sped" up—which means other lots have to slow down.

The "Due Time" column is the number of hours left between now and when the lot is committed to be finished. D2_CTM (drive to cycle time multiplier) is how fast (given as a multiple of REMRPT) the lot must travel to finish on time. For lot A01, its D2_CTM is 3.08 (= 240/78) where 240 is "Due Time" and 78 is REMRPT. FA_CTM is the average factory cycle time multiplier or the average factory speed based on the factory load and its point on the operating curve. For this example its value is 3.5. "Factory Time" is the time the lot will remain in the factory if it travels at FA_CTM. For lot A01, this is $273 = 3.5 \times 78$. "Delta Schedule" is an assessment of whether a lot is ahead or behind schedule. It is "Due Time" minus "Factory Time." Positive values indicate the lot is ahead of schedule and negative values behind schedule. For lot A01, the delta schedule is $-33 (= 240 - 273)$.

The next to last column in Table 9 is "wait time burden." Conceptually the expected completion date for the lot depends on the "wait time burden" allocated to the lot plus its remaining raw process time (REMRPT). The lot's estimated exit (last column) is REMPRT + Wait Burden. In Table 9, the wait burden initially assigned to each lot is based on the lot traveling at speed of FA_CTM, where "Wait Burden" = "Factory Time"—REMRPT. For lot A01, "wait burden" is $195 (= 273 - 78)$. However, the *wait time burden assigned to each lot is a decision controlled* by the planners based on business needs and subject to some constraints—three are:

(1) System balance requires the sum of the wait time burden across all lots match the current factory operating curve performance point which requires a fixed amount of total wait time
(2) There is a minimum wait time burden no lot can avoid
(3) There are limits in diversity of wait time burden across lots

How do we calculate the total required wait time? Since the average factory velocity stated at CTM is 3.5 and the RPT remaining for this group of lots is 478, this establishes a fixed amount of total wait time that must be burdened across all lots. This total wait time burden is 1195 $(= (3.5 \times 478) - 478 = 2.5 \times 478)$. Observe this is the total for the "Wait Burden" column.

Table 10 provides an example of a simple "what if" spreadsheet-based tool, where a planner can try out various wait burden allocations to lots and gauge the impact of these decisions on the estimated completion time for each lot. The seventh column (wait time allocation, shaded) is the decision variable. As planners try different values, columns 8–10 (estimated finish, delta schedule, and required velocity) are automatically updated. The total for wait time allocation must be greater than or equal to 1195.

It is straightforward to enhance this modeling method to provide more automated support for "what if" analysis and incorporate constraints (2) and (3). Additionally, the model can be adapted to an optimization application.

Table 9 Example for allocating wait for lots almost completed

Lot id	REMRPT remaining RPT until lot is finished	"Due time" time remaining between now and the lot's committed due date	D2_CTM speed required for lot to finish on time	FA_CTM average CTM for lots in the factory	"Factory time" remaining cycle time for each lot if it ravels at FACCTM for REMRPT	"Delta schedule" estimate of late (negative) or ahead (positive) for each lot comparing "due time" with "factory time"	"Wait burden" is the total wait time allocated to lot initial burden if lot runs at FA_CTM speed	"Estimated exit" is REMPT + Wait burden is the estimate of when the lot finishes
A01	78	240	3.08	3.5	273	−33	195	273
A02	80	248	3.10	3.5	280	−32	200	280
A03	82	280	3.41	3.5	287	−7	205	287
A04	75	220	2.93	3.5	263	−43	188	263
A05	85	270	3.18	3.5	298	−28	213	298
A06	78	231	2.96	3.5	273	−42	195	273
Total	478	1489			1673	−184	1195	

Table 10 Example for allocating wait for lots that generates equal lateness

Lot id	REMRPT remaining RPT until lot is finished	"Due time" time remaining between now and the lot's committed due date	D2_CTM speed required for lot to finish on time	FA_CTM average CTM for lots in the factory	"Factory time" remaining cycle time for each lot if it ravels at FACCTM for REMRPT	**Wait time allocation decision by planners to allocate wait time based on business needs**	"Estimated exit" is REMPT + Wait burden is the estimate of when the lot finishes	"Delta schedule" due time– estimated exit	RQ_CTM required lot velocity to finsh at estimated time
						Planner Decision			
A01	78	240	3.08	3.5	273	**192.7**	270.7	− 30 (7	3.5
A02	80	248	3.10	3.5	280	**198.7**	278.7	− 30 (7	3.5
A03	82	280	3.41	3.5	287	**228.7**	310.7	− 30 (7	3.8
A04	75	220	2.93	3.5	263	**175.7**	250.7	− 30 (7	3.3
A05	85	270	3.18	3.5	298	**215.7**	300.7	− 30 (7	3.5
A06	78	231	2.96	3.5	273	**183.7**	261.7	− 30 (7	3.4
Total	478	1489			1673	**1195**		− 184	

The Planned Speedup is Worthless Without Execution

Both previous examples require lots that look essentially the same to run at different speeds on the factory floor. In the first example most of the Part B lots will run at a velocity of 4.44 (the CTM), while a few will need to run substantially faster at 2.50. In the second example, we observe six lots now must run at their required velocity (RQ_CTM) instead of a standard factor velocity to meet the new business objectives.

This requirement places a substantial burden on factory floor execution, specifically dispatch scheduling—assigning lots to tools. Simple methods such as FIFO (first in first out) and elapsed time will not work (the lot that has been waiting at the tool the longest or the lot whose wait time exceeds a certain threshold goes next), since each inherently assumes equal wait time for all lots waiting to be processed at a tool set.

Why? Let us look at the case referenced in Table 10. To achieve the planner's goal of equal lateness, lot A03 needs to travel at a speed of 3.8 (CTM) and lot A04 at 3.3. A04 must travel faster than A03, which means it needs to absorb less wait time than A04.

Assume for a moment A03 and A04 are both waiting to be processed at tool LION, A03 has been waiting at the tool for 185 h and A04 has been waiting at the tool for 176 h. If jobs are processed FIFO, then based on its longer elapsed time—A03 would be selected for processing first. However, if we look at the allocation of wait time burden, we see A04 at 176 is past its burden point of 175.7, while A03 at 180 is well below its burden point of 228.7.

Simple dispatch rules such as FIFO and elapsed time worked when factories made only a few products in large quantities with steady demand—a rare environment today. Therefore we observe factory responsiveness requires not just smarter planning, but the ability to execute which requires smarter dispatch.

Appendix 2: Revisiting Capacity Allocation: a Rabbit Out of the Hat

The core elements of resource allocation in central planning engines (CPE) are: (a) linking a manufacturing activity to one or more resources; (b) establishing a consumption rate for each unit of production by that manufacturing activity for the selected resource; (c) providing the total available capacity for the resource; and (d) connecting manufacturing releases (starts) to resource consumption with a linear relationship. In Table 11, we see operations 101 and 151 can be handled by Tool A or B. Operations 201, 202, 301, and 302 can be handled only by Tool A. Table 12 tells us the available capacity for Tool A and Tool B is 1152 working minutes per time unit (for example per day).

Assume, we have a uniform start rate of one lot per day and each lot goes through each operation (101, 151, 201, 202, 301, and 302) once. In steady state, each operation

Table 11 Operation resource linkage

Manufacturing activity	Resource	Consumption rate
Operation 101	Tool A	10
Operation 101	Tool B	10
Operation 151	Tool A	10
Operation 151	Tool B	10
Operation 201	Tool A	15
Operation 202	Tool A	15
Operation 301	Tool A	15
Operation 302	Tool A	15

Table 12 Available capacity

Resource	Consumption rate
Tool A	1152
Tool B	1152

would need to process one lot per day. The optimal way to allocate operations to tools is to assign operations 101 and 151 to Tool B and the remaining four operations (201, 202, 301, and 302) to Tool A. This creates a load on Tool B of 20 ($= 10+10$) minutes of processing and 60 ($= 15 + 15 + 15 + 15$) minutes on Tool A. Since Tool A has 1152 units available, the maximum number of lots per day is 19.2 ($= 1152/60$).

If the demand rises to 30 lots, the CPE would indicate this is not feasible and most likely would push some production out in time showing the pieces being delivered late.

However, the CPE lacks access to the tactical deployment detailed information and therefore has no method to identify better solutions. That is a solution that enables the enterprise to deliver lots on time. If the delivery delay is large and the customer is important, the central planner will contact the factory planner and a review of the detailed deployment decision will occur to "mine for capacity." That is look for opportunities to reallocate capacity to satisfy the demand on time. In many industries a tool can potentially handle many different operations, but at a given point in time is only actively deployed (linked) to a small subset of these operations. This "reduced" deployment occurs for a number of reasons including:

- It is physically impossible for the tool to be "actively" ready for more than a small number of operations. If we want the tool to handle an operation different than those currently selected, the tool has to be brought down for a while, "reconfigured," and brought back up.
- There are manufacturing performance advantages to limit the number of operations a tool is currently deployed to handle.
- The manufacturing team often uses deployment decision to attempt to "balance" tool load by estimating future workload.
- The manufacturing team prefers to deploy its fastest tools to certain operations to keep total cycle time low.
- Habit or prior practice.

Table 13 Operation resource linkage

Manufacturing activity	Resource	Consumption rate
Operation 101	Tool A	10
Operation 101	Tool B	10
Operation 151	Tool A	10
Operation 151	Tool B	10
Operation 201	Tool A	15
Operation 201	Tool C	30
Operation 202	Tool A	15
Operation 301	Tool A	15
Operation 301	Tool C	30
Operation 302	Tool A	15

Table 14 Available capacity

Resource	Consumption rate
Tool A	1152
Tool B	1152
Tool C	2000

The reality is the tactical deployment decision made by Manufacturing when reflected in the capacity information sent to Central Planning understates the flexibility of Manufacturing to produce parts to meet an increase in demand. This flexibility can only be uncovered through manual intervention when central planning presses factory planning.

In our example, it might be Tool C can, after being retooled, be switched from working on "gadgets" to "widgets"—specifically it could be re-configured to handle operations 201 and 301. This change is reflected in Tables 13 and 14.

Additionally, the model in the central planning engine can be enhanced to accommodate:

- Differences in how effectively different tools can process work at a specified set of operations. Typically these differences are speed (Tool A is faster than Tool B), cost (Tool B's unit cost is lower than Tool A), or yield (a manufacturing error). For example, this might result in a view that operation 301 will be assigned to Tool C only as a last resort.
- Minimum tool usage (if operation 301 is assigned to Tool C then it needs least two units of work per day to remain qualified)
- Limitations of tool-operation pairings (Tool C can do work from Operation 301 or 201, but not both)

We might be tempted to state the old adage—garbage in garbage out—but that would be wrong and fail to understand the environment that generated the "limited" but accurate capacity information. In fact, the capacity information provided by Manufacturing to the central supply chain model as reflected in Tables 11 and 12 was accurate, but limited. It was limited to the current near-term production requirements and the near-term ability to use the tools. Tool C can not be used for operations 201

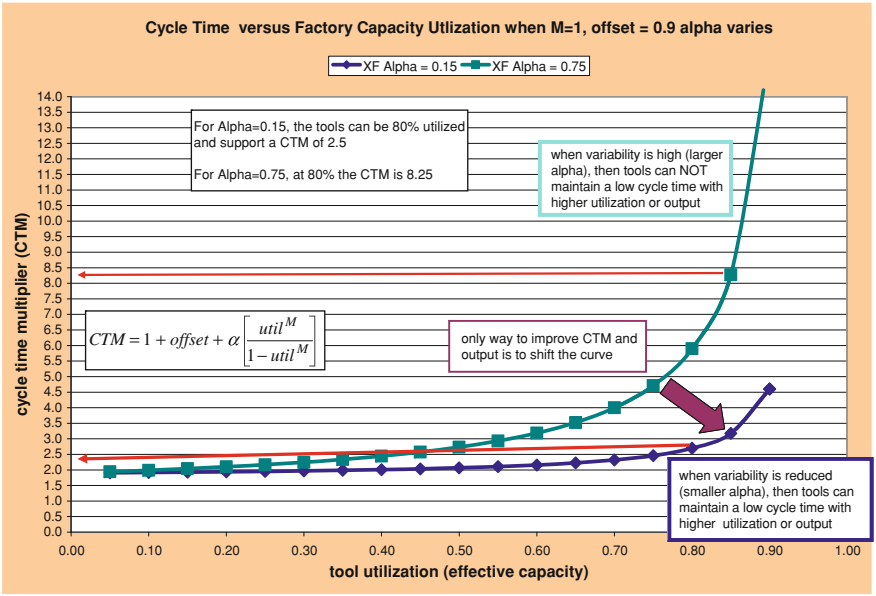

Fig. 5 Cycle time versus factory capacity for Morrison and Martin equation (2006)

and 301 until it is "retooled" and qualified. This may take a day or may take a week—but Manufacturing cannot afford to "retool" on a daily basis.

Appendix 3: Recognizing the Trade-Off Between Capacity and Cycle Time

Review of the Operating Curve

When variability exists either in arrivals or in services there is a trade-off between server (tools, people) utilization and the lead time or cycle time to complete an activity or service. The higher the server utilization, the longer the cycle time. Since higher planned utilization translates to more effective output the trade-off can be reframed as effective capacity available or ouput versus cycle time. The curve that describes this trade-off is called the **Operating Curve** (OC). Lead time or Cycle time is often measured as a cycle time multiplier (CTM), where CTM equals total elapsed cycle time divided by raw process time (RPT). Typically the curve is almost flat for low utilization levels, then spikes sharply upward from the steep part of the curve Fig. 5.

There are a number of equations that can generate this curve, but the one we will use is $\text{CTM} = 1 + \text{offset} + \alpha \left[\frac{\text{util}^M}{1 - \text{util}^M} \right]$ [24].

- **CTM** is the cycle time multiplier of RPT—measure of cycle time as a function of RPT.
- **util** is the fraction of utilization in the entity—facility, tool set, checkout clerks, etc.
- **Offset** represents several aspects of the process that generate wait time and cannot be eliminated. For example: travel time, hold hours, and post-processing hours relative to total RPT. A common value for offset may be about 0.9. When offset is 0.9 this sets the minimum CTM at 1.9.
- **M** is the number of identical parallel machines or servers. Typically this value ranges from 1 to 4 (even when the number of tools or servers exceeds 4) work best.
- $\boldsymbol{\alpha}$ represents the amount of variation in the system and controls how long the curve stays flat. The lower the value of $\boldsymbol{\alpha}$ the less variation and longer the curve stays flat. Common values for α range between 0.35 and 0.65 [10].

Solving for Util, we have $\text{Util} = \left(\frac{\text{CTM}-(\text{offset}+1)}{\text{CTM}-(\text{offset}+1)+\alpha}\right)^{\frac{1}{m}}$

Linking Capacity and Cycle Time

Assume that the product XYZ is processed five times by tool set AAA during its production route. Each time a widget goes through tool set AAA it is referred to as a "pass." In this case product XYZ has five passes on tool set AAA. Additionally, assume 100% process yields and that the average RPT for each XYZ widget on tool set AAA is two units. In steady state, this makes for a total RPT required per day of 10 ($= 5 \times 2$ units per widget of XYZ on tool set AAA. Assume we start one widget per day and we have ten units of capacity, what would the cycle time be? The CTM equation makes it clear the cycle time would be infinite since the capacity required matches capacity available making tool utilization 100%.

The business states it wants to run this product with a CTM of 4.0. This requires some portion of time that the tool set is available to produce, but does not have WIP. How do we incorporate that into the planning process?

We calculate a burden or uplift factor (ULF) per widget based on the target CTM and the specific characteristics of the Operating Curve for this tool set. Assume the Operating Curve for this tool set has offset $= 1$, alpha $= 0.5$, and m $= 1$. Using the above equation to solve for UTIL, the required utilization to achieve the CTM target of 4.0 is 0.80 (80%). For each unit of raw capacity required, we need 1.25 units available to meet the CTM target. The value 1.25 is the uplift factor (ULF) and determined by:

$ULF = 1/tool\_utilization\_meet\_cycle\_time\_target\_from\_opcurve.$

If we have 250 units of capacity available per day, how many widgets can we start per day at committed cycle time? The answer is 20 ($= 250/(1.25 \times 10)$).

If the business wanted to achieve a cycle time of 3.5, how many widgets could it start per day? Using the same equation for UTIL, the utilization required to achieve a

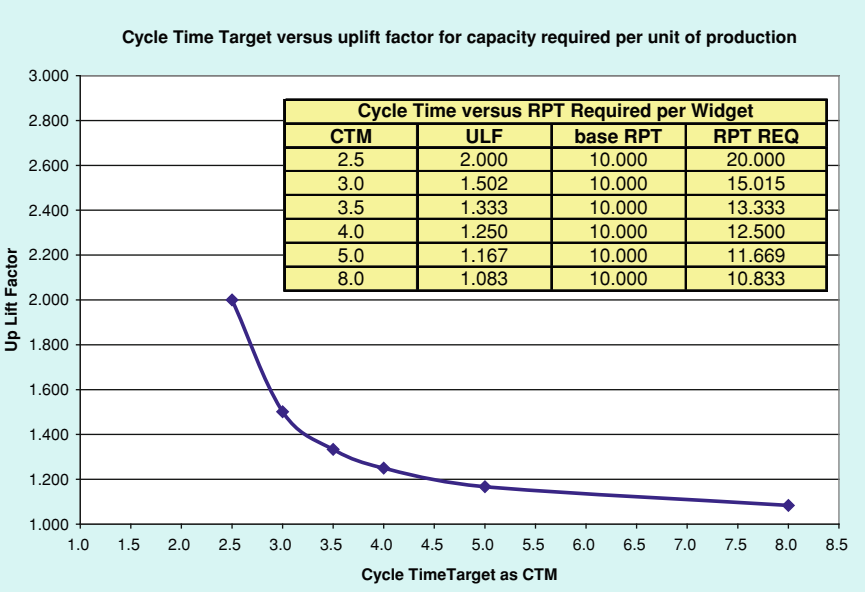

Fig. 6 Uplift in capacity consumption (required) based on cycle time commit

cycle time of 3.5 is 0.75 (75%). The ULF is 1.333 ($= 1/0.75$). The maximum number of widgets per day at this cycle time commit is 18.75 ($= 250/(1.333 \times 10)$). Shorter cycle time equates to reduced widget starts.

Alternatively, instead of decreasing the capacity available, we can uplift the capacity required per unit of production. For the 4.0 cycle time commit, we uplift the capacity consumption rate of ten units per day per widget to 12.5 ($= 10 \times 1.25$). For the 3.5 commit, the ten units is uplifted to 13.33 ($= 10 \times 1.33$). Figure 6 provides the uplifted capacity consumption rates to input to the central planning process based on the cycle time the business wishes to achieve.

This example makes it clear that the assumption in a typical central planning process that cycle time and capacity are independent is not correct—the two are clearly coupled. We can view this as classical planning meets its uncertainty principle [10]. It is a rich ground for improved responsiveness and a headache for classical planners. Since Lean advocates believe that all variation can be eradicated, they have no awareness of an operating curve and no methods to capture this opportunity. It is like attempting to ignore special and general relativity and still produce GPS locations [26].

Table 15 Deployment decision

	Tool A	Tool B	Tool C	No tools covering oper
Operation 1	1	1	1	3
Operation 2	1	1	0	2
Operation 3	0	1	0	1
Operation 4	0	0	1	1
Operation 5	1	1	1	3
Operation 6	1	0	0	1
Operation 7	1	0	0	1
Number opers tool covers	5	4	3	

Table 16 RPT, lots, available capacity

	Tool A	Tool B	Tool C	# lots
Operation 1	4	20	5	40
Operation 2	15	20	9999999	30
Operation 3	9999999	15	9999999	10
Operation 4	9999999	9999999	20	60
Operation 5	5	5	5	10
Operation 6	8	9999999	9999999	200
Operation 7	10	9999999	9999999	200
Capac avl	720	1152	1296	

Appendix 4: How Well Does the WIP Match the Tool Deployment?

In this example we have three tools (A, B, and C) and seven operations (operation 1–7) to handle the current WIP at this tool set. Table 15 shows the current deployment decisions made by the factory planner. In the table, a 1 means the tool (column) is able to service the operation (row). For now these decisions are fixed.

Table 16 provides all of the key pieces of information for the model. The value in operation / tool cell is the raw process time (RPT) that the tool requires to process one lot at this operation. For example, the RPT for Tool A to process one lot at operation 1 is 4 time units. A value of 9,999,999 indicates this operation/tool combination is not currently active, corresponding to a 0 in Table 15. The last column (# lots) provides the number of lots requiring service of that operation at this point in time. For example, there are 40 lots at operation 1 waiting to get on either Tool A, B, or C. The last row ("capac avl") provides the number of time units of capacity available for that tool over the service period. Tool A has 720 units available.

Table 17 has the basic "what if" model. The value in each operation/tool cell is the *business decision*- the number of lots the tool is assigned to handle for this operation over some time period. For example, Tool B will handle 30 of the 40 lots that require

Table 17 Allocation decision and results

	Tool A	Tool B	Tool C	Lots served	Goal	# lots	Delta
Operation 1	0.0	30.0	10.0	40.0	=	40	0.0
Operation 2	3.0	0.0	0.0	3.0	=	30	−27.0
Operation 3	0.0	10.0	0.0	10.0	=	10	0.0
Operation 4	0.0	0.0	60.0	60.0	=	60	0.0
Operation 5	0.0	10.0	0.0	10.0	=	10	0.0
Operation 6	90.0	0.0	0.0	90.0	=	200	−110.0
Operation 7	0.0	0.0	0.0	0.0	=	200	−200.0
Cap used	765.0	800.0	1250.0		Total unmet dmd		−337.0
Constraint	<	<	<				
Cap avl	720.0	1152.0	1296.0				
Delta	−45.0	352.0	46.0				

service at operation 1, ten lots for operation 4, and ten lots for operation 5. These 21 values (cells) (7 operations by 3 tools) represent allocation decisions.

The results of these decisions are found in column 5 (lots served) and row 10 (cap used). The "lots served" column is the total number of lots served for this operation across all tools. For example 40 lots at operation 1 will be served—0 on Tool A, 30 on Tool B, and 10 on Tool C ($40 = 0 + 30 + 10$). The row "cap used" tells us how much capacity is used for each tool. This is the sum of the product of the allocation times RPT (Table 16). For tool A, the cap used is 765 $(=(0\times 4)+(3\times 15)+(0\times 9999999)+(0\times 9999999)+(0\times 5)+(90\times 8)+(0\times 10))$.

The last component of the model is comparing the results of the business (allocation) decision made with the goals of the factory. The factory has two goals: service as many lots as possible and use all available capacity.

The "lots served" goal comparison information is in columns 6 (goal type), 7 (target), and 8 (delta). Our goal is to service all lots waiting (=). The target is the number of lots that are currently waiting (last column in Table 16). The result is posted in the "delta" column which is simply "lots served" minus "target." For example, for operation 2 the value is −27, since our target was 30 and the actual number of lots served was 3 ($-27 = 3 - 30$). This tells us the current allocation decision leaves 27 lots waiting at operation 2 "still waiting." The last value in the delta column is the total unmet demand based on the current allocation decision. The value −337 is simply the column sum.

The "capacity goal" information is in the last three rows (goal type, cap maximum, and delta cap goal). It is a constraint—do not make allocation decisions that exceed available capacity. The target is the capacity available for each tool (last row in Table 16). The result is posted in the last "delta" row which is simply "cap used" minus "cap maximum." For example, for Tool B the value is 352, since the actual capacity used was 800 and the maximum capacity was 1152 ($352 = 1152 - 800$). This tells us the current allocation decision leaves 352 units of capacity at Tool B idle.

Table 18 Revised allocation decision and results

	Tool A	Tool B	Tool C	Lots served	Goal	# lots	Delta
Operation 1	0.0	37.0	3.0	40.0	=	40	0.0
Operation 2	0.0	10.0	0.0	10.0	=	30	−20.0
Operation 3	0.0	10.0	0.0	10.0	=	10	0.0
Operation 4	0.0	0.0	60.0	60.0	=	60	0.0
Operation 5	0.0	10.0	0.0	10.0	=	10	0.0
Operation 6	90.0	0.0	0.0	90.0	=	200	−110.0
Operation 7	0.0	0.0	0.0	0.0	=	200	−200.0
Cap used	720.0	1140.0	1215.0			Total unmet dmd	−330.0
Constraint	<	<	<				
Cap avl	720.0	1152.0	1296.0				
Delta	0.0	12.0	81.0				

This simple model enables planners to manually assess the quality of their tentative deployment decisions and estimate the maximum number of lots that can be serviced with this deployment. Typically a planner will try different allocation decisions (leaving the deployment decision unchanged) to determine how to best allocate WIP to tools to meet prioritized demand and then send guidelines to Manufacturing. Table 18 shows an improved allocation plan eliminating overusing capacity on tool A and reducing overall unmet demand from 337 to 330.

When capacity is highly utilized and tensions are high, most planners will welcome an upgrade to a small optimization model where the decision variables are the allocation values, the constraints are not exceeding the capacity available, and the objective is to minimize unmet prioritized demand. The extensions to handle integer values, demand priorities, multiple periods, and partial deployments (which occur during a phase in) are straightforward.

To enable the planner to model the impact of deployment decisions, we need to couple the decisions made in Tables 15 and 17. He or she can change the deployment decision (Table 15); then revisit the allocation decision (Table 17); then assess the impact on the WIP waiting to be serviced. Again, optimization methods can be used to reduce the workload on the analyst and handle demand priorities and multiple time periods [1, 34].

The Planned Deployment Decision is Worthless Without Execution

As with changing lot velocities, finding an allocation of WIP to tools is only the first part of success for the factory. The second part is execution. This requirement places a substantial burden on factory floor execution, specifically dispatch schedule decision making—assigning lots to tools. Simple methods will not work.

In this example (Table 18), the plan requires all of the lots for operation 5 to run on Tool B even though all three tools are equally proficient (Table 16) at processing operation 5. This decision was made because lots at operation 6 can only run on Tool A and lots at operation 4 can only run on Tool C. However, on the factory floor doing this type of analysis at best would be very difficult. Second, from the floor's point of view, the RPT for lots at operation 5 is the same (5) for all three tools. Therefore a casual analysis would make the floor indifferent to which tool handled lots at operation 5—with dire consequences to the factory! What dire consequences? If the operator places the lots for operation 5 on tool A, then there is no tool available to run lots for operation 6. The result would be lower utilization of Tool B and additional delays for the lots at operation 6.

Appendix 5: Dispatch Scheduling Details on Guidance and Judgment

Referring to Fig. 4, *Guidance* or advocate logic is the set of computational activities (which may be a computer program or manual) to create information posted to some location (often a table structure) that the assignment logic accesses or to trigger an assignment module to execute. The most common example is a calculation to determine whether a lot is ahead or behind its planned pace. Another example is the updating of a fact base that may contain operation—tool preference based on static information (such as difference in raw process times between tools executing the same manufacturing action) or dynamic information (the amount of time will take to set the tool up to handle this manufacturing action). Other types of guidance include flow balance (avoid starving a tool set), manufacturing requirements (avoid running all lots of a certain type on a single tool, but distribute them across three tools), and process control time windows (lot must complete the next three steps within 5 h or it will need to be scrapped due to contamination).

Judgment or assignment is the set of computational activities that when completed, result in a change of state or action on the manufacturing floor. The judgment logic must balance competing requirements such as meeting on time delivery, demand priorities, improving throughput with batches and trains, current WIP position, and tool status.

The real goal of any judgment application is make a sequence of decisions over time that in aggregate improve the future position of the factory relative to its role in the total supply chain or demand supply network. The decisions are based on impact on the future state of the factory, not based on prior events. The sum total of the prior events has resulted in the current state of the factory. All other measures are attempts to create an interim goal that can be measured and decisions made against that is a reasonable approximation of the ultimate goal. Additionally, under some circumstances, these goals can be at odds with each other and the overall good of the demand-supply network.

Simple Judgment typically

- Does not consider the assignment of lots to other tools in the tool group
- Does not consider the assignment of lots over time
- Does not consider upstream or downstream conditions (WIP level and Tool status and near term throughput rates)
- Uses simple rules of thumb for complex trade-offs
- Written with decision tree one iteration logic
- Generates a single decision, through a series of filters and if-then conditions
- Gives a reasonable (though myopic) decision
- Relies on manual intervention for process time windows
- Typically have to be rewritten for different WIP levels (static adjustment)

Advanced Judgment typically

- Looks across the tool set and upstream and downstream
- Handles all process time windows
- Establishes an anticipated sequence of assignments at all tools in a tool group over time for lots at the tool or which will arrive soon
- Measures the quality of a proposed solution and anticipates impact on factory performance
- Uses an iterative search process in judgment logic
- Dynamically adjusts for WIP levels and other business conditions

References

1. Bermon S, Hood S (1999) Capacity optimization planning system (CAPS). Interfaces 29(5): 31–50
2. Bitran G, Tirupati D (1989) Tradeoff curves, targeting and balancing in manufacturing networks. Oper Res 37(4):547–555
3. Bixby R, Burda R, Miller D (2006) Short-interval detailed production scheduling in 300 mm semiconductor manufacturing using mixed integer and constraint programming. *Semiconductorfabtech*, 32nd edn, http://www.fabtech.org, pp 34–40
4. Buchholz J (2005) Interview with Nick Donofrio. IBM on the spot series, posted on http://www.ibm.com. Accessed 9 Aug 2005
5. Chen H, Harrison M, Mandelbaum A, Ackere A, Wein L (1988) Empirical evaluation of a queuing network model for semiconductor wafer fabrication. Oper Res 36(2):202–215
6. Dennis P (2007) Lean production simplified. Productivity Press, New York
7. Denton B, Forrest J, Milne RJ (2006) Methods for solving a mixed integer program for semiconductor supply chain optimization at IBM. Interfaces 36(5):386–399
8. Fordyce K, Bixby R, Burda R (2008) Technology that upsets the social order—a paradigm shift in assigning lots to tools in a wafer fabricator—the transition from rules to optimization. In: Proceedings of the 2008 winter simulation conference
9. Fordyce K, Wang C-T, Chang C, Degbotse A, Denton B, Lyon P, Milne RJ, Orzell R, Rice R, Waite J (2011a) In: Kempf, Keskinocak, Uzsoy (ed). The ongoing challenge—creating an enterprise-wide detailed supply chain plan for semiconductor and package operations. Planning production and inventories in the extended enterprise: a state of the art handbook, Vol 2 (Chapter 14)

10. Fordyce K, Fournier J, Milne RJ (2011b) Basics of the operating curve—classical planning meets its uncertainty principle. Working paper, fordyce@us.ibm.com, jmilne@clarkson.edu
11. Fox B, Kempf K (1985) Complexity uncertainty, and opportunistic scheduling. In: Proceedings of the IEEE second conference on artificial intelligence applications: the engineering of knowledge based systems, Miami, FL, pp 487–492
12. Gross D, Harris C (1998) Fundamentals of queueing theory, 3rd edn. Wiley, New York
13. Hackman ST, Leachman RC (1989) A general framework for modeling production. Manag Sci 35(4):478–495
14. Hopp W, Spearman M (2008) Factory physics, 3rd edn. McGraw-Hill Irwin, New York
15. Horn G, Podgorski W (1998) A focus on cycle time-vs.-tool utilization "paradox" with material. In: Advanced semiconductor manufacturing conference and workshop proceedings, pp 405–412
16. Kempf K, Pape D, Smith S, Fox B (1991) Issues in the design of AI based schedulers. AI Mag 11(5):37–45
17. Kempf K (1989) Manufacturing scheduling: intelligently combining existing methods. In: Working notes of AAAI AI in manufacturing symposium. Fox M (ed.), AAAI, Burgess Drive Menlo Park
18. Kempf K (1994) Intelligent scheduling semiconductor wafer fabrication. In: Mark Fox, Monte Zweben (eds) Intelligent scheduling. Morgan Kaufman Publishers, pp 473–516 (Chapter 18)
19. Kempf K (2004) Control-oriented approaches to supply chain management in semiconductor manufacturing. In: Proceedings of the 2004 American control conference, Boston, MA, pp 4563–4576
20. Leachman R, Benson R, Liu C, Raar D (1996) IMPReSS: an automated production planning and delivery-quotation system at Harris corporation–semiconductor sector. Interfaces 26(1):6–37
21. Little J (1992) Tautologies, models and theories: can we find laws of manufacturing? IEEE Trans 24(3):7–13
22. Liu J, Yang F, Wan H, Fowler J (2010) Capacity planning through queuing analysis and simulation-based statistical methods: a case study for semiconductor wafer FABS. web.ics.purdue.edu/ hwan/docs/IJPR
23. Morrison J, Dews E, LaFreniere J (2006) Fluctuation smoothing production control at IBM's 200mm wafer fabricator: extensions, application and the multi-flow production index (MFPx). In: Proceedings of the 2006 IEEE/SEMI Advanced Semiconductor Manufacturing Conference, Boston, MA
24. Morrison J, Martin D (2006) Cycle time approximations for the G/G/m queue subject to server failures and cycle time offsets with applications. In: ASMC 2006 Proceedings, p 322
25. Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York
26. Pogge R (2009) Real world relativity, http://www.astronomy.ohio-state.edu/~pogge/Ast162/Unit5/gps.html
27. Shirodkar S, Kempf K (2006) Supply chain collaboration through shared capacity models. Interfaces 36(5):420–432
28. Shobrys D (2003) History of APS. Supply chain consultants (www.supplychain.com), Wilmington, DE 19808, USA
29. Simon HA (1957) Administrative behavior, 2nd edn. The Free Press, New York
30. Singh H (2009) Supply chain planning in the process industry. Supply Chain Consultants, Wilmington
31. Singh H (2009) Practical guide for improving sales and operations planning. Supply Chain Consultants, Wilmington
32. Sullivan G (1990) IBM Burlington's logistics management system (LMS). Interfaces 20(1): 43–61
33. Sullivan G (1994) Logistics management system (LMS): integrating decision technologies for dispatch scheduling in semiconductor manufacturing. Intelligent scheduling. Morgan Kaufman Publishers, San Francisco pp 473–516

34. Sullivan G (1995) A dynamically generated rapid response fast capacity planning model for semiconductor fabrication facilities. the impact of emerging technologies on computer science and operations research, Kluwer Academic Publishers, Boston (presented at Winter 1994 computers and operations research conference)
35. Uzsoy R, Lee C, Martin-Vega LA (1992) A review of production planning and scheduling modules in the semiconductor industry, Part 1: system characteristics, performance evaluation, and production planning. IIE Trans Sched Logist 24(4):47–60
36. Uzsoy R, Lee C, Martin-Vega LA (1994) A review of production planning and scheduling modules in the semiconductor industry, Part 2: shop floor control. IIE Trans Sched Logist 26(5):44–55
37. Zisgen H, Ments I, Wheeler B, Hanschle T (2008) A queuing network based system to model capacity and cycle time for semiconductor fabrication. In: Proceedings of the 2008 winter simulation conference
38. Zisgen H, Brown S, Hanschke T, Meents I, Wheeler B (2010) Queuing model improves IBM's semiconductor capacity and lead-time management. Interfaces 40(5):397–407