Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, eds

TUTORIAL: ILLUSION OF CAPACITY - CHALLENGE OF INCORPORATING THE COMPLEXITY OF FAB CAPACITY (TOOL DEPLOYMENT & OPERATING CURVE) INTO CENTRAL PLANNING FOR FIRMS WITH SUBSTANTIAL NON-FAB COMPLEXITY

Kenneth Fordyce
John Fournier

IBM
River Road
Essex Junction, VT 05452 USA

R. John Milne

Clarkson University
8 Clarkson Ave
Potsdam, NY 13699 USA

Harpal Singh
Arkieva
5460 Fairmont Drive
Wilmington, DE 19808 USA

ABSTRACT

Since the early 1990s, organizations have focused on making smarter decisions in their integrated supply chain central planning, but the representation of capacity and cycle time has remained static and linear in contrast to its complex nature. This includes central planning for firms with semiconductor fabrication facilities (FABS) as a component of a complex demand supply network(DSN) where much of the complexity is non-FAB. Developing more intelligent solutions for capacity in central planning within computational and process limitations is a critical challenge. For DSNs with FABS, twin challenges are tool deployment and the operating curve. Many in the FAB community are aware of these complexities; options proposed and some implemented within "aggregate FAB planning," rarely within central planning. This tutorial reviews the current state of central planning with respect to capacity and cycle time, outlines the challenges these complexities place on central planning structures, and indicates possible solutions

1 INTRODUCTION

For years the consulting mantra was "lack of "executive level buy-in" was a major impediment to a successful planning process. Often this is not the primary barrier, since most executives now realize a disciplined planning process will help the bottom line. What are the top barriers? A Hitachi Consultant's study of the barriers to successful planning identified the number one barrier is the lack of suitable software tools—which includes more intelligent modeling of the complex nature of capacity (Elmaghraby 2007). Tool Deployment and the Operating Curve are FAB complexities at direct odds with the simple linear methods to model capacity and cycle time that still dominate central planning models or engines (CPE) to support integrated supply chain management of a complex demand supply networks (DSN).

This includes CPEs for firms that produce semiconductor based packaged goods (SBPG) where semiconductor fabrication facilities (FABS) are a component of the DSN that contains substantial non-FAB complexity (such as capacity for post-FAB manufacturing, rich set of demand priorities, client requirements at multiple locations of the network, alternative bill of materials, substitution, sourcing, fair share, partial shipments, minimum starts, date effective cycle times, inventory policy, dynamic pegging, inventory policy and a need to support a comprehensive report and analysis system for central planners

978-1-4673-4781-5/12/$31.00 ©2012 IEEE

(Fordyce et al. 2011). The "non-FAB" complexity generates a substantial computational and process burden.

Many in the FAB community are aware of these complexities - various options have been proposed; new ones are emerging; and some have been implemented. Any reasonable review of this work makes clear there is awareness, but not consensus. This work falls within "aggregate FAB planning," not central planning for the integrated supply chain for SBPG firms. In many cases (even for firms with "advanced" central planning and FAB planning models) the representation of FAB capacity remains "stuck in time" often stating FAB capacity in terms of wafer starts (Appendix 1). Central planners often rely on simple trade-off rules and a once a month "sizing" from the aggregate FAB model(s).

Ignoring these complexities is not sustainable as the burden on responsiveness resulting in underutilization or late delivery of products becomes increasingly unacceptable. Reducing the illusion of FAB capacity in CPEs is a component of the ongoing challenge to improve responsiveness with a tighter coupling between central planning and factory planning (Fordyce et al. 2012).

The tactical goal of this paper is to (a) briefly review the current state of central planning with respect to capacity and cycle time, (b) outline the structures and challenges the twin complexities place on capacity available in traditional CPE structures, and (c) indicate possible solution approaches. It is outside the scope of this paper to discuss possible solutions in detail. Clearly, solutions will need to meet this criteria: (a) initially only require "tolerable and manageable" change to current business practices and "organizational structures" while at the same time providing an opportunity for additional (meaningful) business value from the suggested new technical approach, (b) naturally extend current analysis methods for planners for better direct insight into FAB capacity, (c) support a tighter coupling between central planning and aggregate FAB planning, (d) facilitate the introduction of new models and methods into current practice that ultimately improve effectiveness and responsiveness which require "upsetting the social order," and (e) recognize the plan is the starting point to manage the demand supply network, not the end point.

Section 2 reviews central planning, aggregate FAB planning, deployment, the OPCurve, and wafer start limits with a focus on capacity available (CAPAVAIL) and introduces what influences its value:

- Central Planning – traditional fixed value of CAPAVAIL magically appears as an input parameter
- Aggregate FAB planning – what output can we expect and what tool sets have insufficient capacity
- Deployment – lack of tool uniformity forces a decision to overestimate or underestimate CAPAVAIL
- OPCurve – a portion of CAPAVAIL must be allocated to be idle to meet the cycle time commitment and this varies based on the cycle time and the nature of the OPCurve in a non-linear fashion

The remainder of the paper explores the complexity in more detail moving from the "part" level (central planning) to the operation level (FAB) to identify the lost opportunity generated by current methods, the challenge to improve this representations, and identifies options to better handle these complexities.

The approach starts with a simple FAB focused one period central planning example at the part level (section 3) and incrementally extend it to explore the impact of the OPCurve (section 4) and deployment at the operation level (section 5) on CAPAVAIL. Section 6 extends the model in section 3 to include the OPCurve and deployment generating a one period model which identifies all of the decision points and illustrates the complexity. The more detailed paper is posted at www.arkieva.com. whitepapers

2 BACKGROUND AND BASIC

2.1 Central Planning & Aggregate FAB Planning

Central planning (Fordyce et al. 2011) is defined as matching assets with demand across each manufacturing site in the firm and across time to determine a feasible supply that best meets a set of prioritized customer requests and opportunities to chase. The starting point of central planning is the client requirements (demand). The creation of this plan requires a central planning engine (CPE) with these features:

1. Method(s) to represent the (potential) material flows in production according to business policies, constraints, demand priorities, locations of assets, etc., and relate all this to customer requirements (demand). Typically the lead time or cycle time is a component of the flow description.
2. Methods to represent capacity available (CAPAVAIL) at key resources and the capacity consumption or requirement rate (CAPREQ) to produce parts.
3. Search mechanism(s) to generate "best can do" (BCD) match between demand and supply.

In most CPEs, the core solver does not make decisions about WIP within a manufacturing unit, but focuses on start decisions. A separate method estimates when each lot in WIP will exit the current manufacturing stage (Fordyce et al. 2012).

Despite a series of advancements in other areas, the core representation of capacity and cycle time in CPEs has not changed since the 1980s. The current practice is summarized as:

- Capacity is modeled with simple linear equations where the two primary fixed inputs are
  - Statement of capacity or resource required (CAPREQ) per unit of production for a part
  - Statement of capacity or resource entity available (CAPAVAIL)
- Capacity is consumed at the start of the manufacturing activity
- The cycle time or lead time is fixed and disconnected from capacity from the CPE's point of view.

For example, as illustrated in Table 1, part 111 has a cycle time of ten days and consumes two units of resource or tool group AA and three units of resource BB for each unit produced. Part 222 has a cycle time of twelve days and consumes four units of resource AA and two units of resource BB for each unit produced. The capacity available per unit time for resource AA is 90 units and BB is 120 units.

Table 1: Traditional capacity required (CAPREQ) and capacity available (CAPAVAIL) in CPE

		Resource		
		AA	BB	Cycle Time
Part	Part 111	2	3	10
	Part 222	4	2	12
CAPAVAIL		90	120	

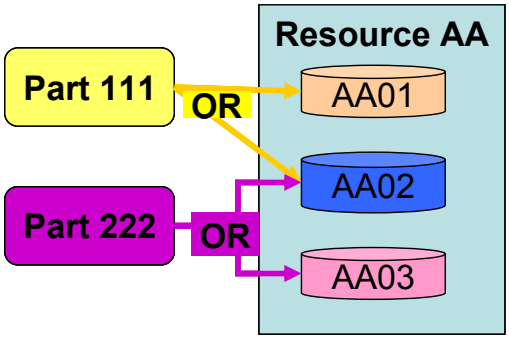
If the decision variables are X_1 is the number of units of Part 111 produced and X_2 is the number of units of Part 222 produced. The traditional capacity constraint equations would be $2X_1 + 4X_2 \le 90$ for resource AA and $3X_1 + 2X_2 \le 120$ for resource BB.

Aggregate FAB capacity or tool planning is also focused on matching assets with demand but with a different orientation. Central planning considers all manufacturing plants in the enterprise; FAB planning focuses on a single FAB at a time. Its starting point is often a wafer starts profile (sometimes current WIP), it works with operations, and it results in an estimated output, identification of tool groups with insufficient capacity to meet demand (load), or some combination of these two. There a variety of "simple" methods that work with a fixed daily start rate and a steady state consumption of tools that handle the twin complexities with differing levels of precision or project WIP output using routes, cycle times, and a limited number of high level capacity constraints. A more sophisticated method found in some FABS uses optimization to allocate capacity at key toolsets within a daily output planning model. A variety of exciting methods have been proposed and some implemented using techniques such as: queuing equations and networks, discrete event simulation, optimization (column generation), and clearing functions. Information on these methods can be found in Kacar et al. 2012, Bermon et al. 1999, Zisgen et al. 2010, Tibbits 1993, Bagch et al. 2008, Schelasin 2011, and Dobson and Karmarkar 2011. It is outside this paper to review FAB planning in detail; it is an exciting and pivotal area.

2.2 Tool Deployment

Many FABs are "so over-run" with tool deployment complexity, that the natural human tendency is to ignore it or handle it with simple rules. Deployment refers to partially shared manufacturing operations be-

tween tools in a tool group or resource. Extending the example from section 2.1 (Figure 1) resource AA might be a work center consisting of three pieces of equipment or tools: AA01, AA02, and AA03 and each has a capacity of 30(=90/3). Although resource AA services parts 111 and 222, only a subset of the tools process each part. For example, part 111 is serviced by tools AA01 and AA02 (and not serviced by AA03). Part 222 is serviced by tools AA02 and AA03 (and not serviced by AA01). The right hand side of figure 1 has a Boolean tabular method to express deployment. There is one row for each part and one column for each tool. A cell value of 1 indicates the tool can service the part. A value of 0 (zero) indicates the tool can not service the part.



Boolean Tabular Representation of Deployment. A 1 indicates tool (column) services this part (row), else 0.				
		Tool		
		AA01	AA02	AA03
Part	Part 111	1	1	0
	Part 222	0	1	1

Figure 1: Example of deployment of part operations to resources

This additional level of detail creates challenges in defining the capacity constraint equations and defining capacity available (CAPAVAIL). If we continue the illusion that all three tools for resource AA support part 111, then we risk over committing the business. For example, there are 60 units of capacity available for part 111 (AA01 and AA02). Since CAPREQ (table 1) is 2, the maximum number of units of part 111 the FAB can start per day is 30(=60/2). However, if we continue to use the equation for resource AA in the original model, then the maximum number of units of part 111 is 45(=90/2). This can over commit the business by 15 units! How do we avoid this? We might "split" AA02's 30 units between part 111 and part 222 and fix the CAPAVAIL for each part to 45(=30+15) and replace the equation for resource AA with the two equations $2X_1 \le 45$ (limiting 111) and $4X_2 \le 45$ (limiting 222). The maximum output of Part 111 would be 22.5(=45/2) creating a risk of underutilization.

2.3 Operating Curve

When variability exists either in arrival times or service times a trade-off exists between server (tool) utilization and the lead or cycle time to complete a service → the higher the utilization the longer the cycle time. Alternatively, the price for shorter cycle times is lower tool utilization. In FAB terms, this is called idle time without WIP (tools being idle due to the absence of WIP to work on). Since planned utilization times nominal capacity translates into effective capacity available, the trade-off can be reframed as effective capacity available (CAPAVAIL) versus cycle time. To achieve a lower cycle time for a certain level of variability requires planning on idle time without WIP. Since effective capacity directly influences the wafer starts that can be supported, the trade-off can be reframed for central planning as "increased output stated as wafer starts per day" versus "lead time."

The curve that describes this trade-off is called the **operating curve** which dynamically relates cycle time with effective capacity available. Typically the curve has a long flat period and then spikes sharply upward forming the steep part of the curve. Where the curve spikes "up" is determined by the level of variability in the system.

Since "effective" capacity can be translated into maximum wafer starts, this curve connects two critical business decisions that are not linked in planning: cycle time and wafer starts. Connecting them adds complexity in two phases: (a) modifying capacity available based on cycle time decisions using cy-

cle time tax and (b) reformulation/restructuring of core models to dynamically link the cycle time decision with the capacity available which influence the wafer starts decision

While some FAB folks are familiar with this trade-off, for a central planner fresh out of an elite supply chain undergraduate program who sees the FAB as a black box that never seems to deliver on time, this is counter intuitive – wait longer and get more output – and the relationship is not linear. When the FAB is lightly utilized, one can ask for more output without any impact on cycle time. When the FAB is full, extra output costs a lot of cycle time, never mind variability. The reader can see the *comfort provided by the illusion of a fixed cycle time and traditional capacity constraints.*

3 SIMPLE CENTRAL PLANNING MODEL USING AGGREGATED RESOURCES

We start our adventure into the wilds of the twin complexities and their tributary streams with a simple single FAB focused one period part level central planning model with three product or part families, four key resources, and two "feature" resources specific to a subset of the part families. The FAB has three part families: Antelope, Gazelle, and Lion. A part family is a set of specific unique parts that have similar manufacturing characteristics in terms of manufacturing sequence and resources consumed. Some level of aggregation is typical in FAB planning. The FAB has four major resources or tool groups that are shared by all part families and two "feature" resources that are only used by a subset of the part families in smaller quantities. The major resources are Mid Ultraviolet Light photolithography (MUV), Deep Ultraviolet Light photolithography (DUV), Ion Implant (ION), and metal etch (ETCH). The feature resources are ANT/GAZ and GAZ. Table 2 provides the CAPREQ and CAPAVAIL information.

Table 2: Consumption rate for per wafer for family at resource and capacity available

		Resource Entity					
		shared by all part families				feature	
		MUV	DUV	ION	ETCH	ANT/GAZ	GAZ
Part Family	Antelope	5	5	6	4	2	0
	Gazelle	8	4	5	7	1	1
	Lion	6	10	10	6	0	0
CAPAVAIL		100	100	150	130	30	15

This table informs us each Antelope wafer start consumes 5 units of MUV and 5 units of DUV. In contrast, Lion consumes 6 units of MUV and 10 units of DUV. For this single period, there are 100 units of MUV to allocate across all part families. Lion does not use ANT/GAZ. The decision variables are:

- X_A = number of wafers of Antelope to start this period
- X_G = number of wafers of Gazelle to start this period
- X_L = number of wafers of Lion to start this period

The following 6 equations capture the capacity constraints. The notation (eq 1-X) has the following meaning: the first digit indicates this is a capacity constraint equation for a resource. The second digit tells us which resource: 1 for MUV, 2 for DUV, .etc.

$$5X_A + 8X_G + 6X_L \le 100 \quad MUV \quad (eq\ 1-1)$$

$$5X_A + 4X_G + 10X_L \le 100 \quad DUV \quad (eq\ 1-2)$$

$$6X_A + 5X_G + 10X_L \le 150 \quad ION \quad (eq\ 1-3)$$

$$4X_A + 7X_G + 6X_L \le 130 \quad ETCH \quad (eq\ 1-4)$$

$$2X_A + 1X_G + 0X_L \le 30 \quad ANT/GAZ \quad (eq\ 1-5)$$

$$0X_A + 1X_G + 0X_L \le 15 \quad GAZ \quad (eq\ 1-6)$$

To complete our example, Table 3 provides information about the demand for each part family stated in terms of wafer starts. The "must make" column tells us the required starts for the part family to meet firm commitments to clients. These wafer starts must be satisfied before any other wafer starts are contemplated. The "max demand" column indicates the maximum number of wafers of this part family to start, even if extra capacity exists – since there are no customers beyond it.

Table 3: Demand information stated as starts

		Demand Information		
		must make	max dmd in market	profit per wafer
Part Family	Antelope	20	25	$3.00
	Gazelle	10	30	$8.00
	Lion	10	60	$12.00

This information generates three "must make" (eq 2-X) and three "max demand" constraints (eq 3-X).

Must make generates these equations
$X_A \geq 20 \Rightarrow Antelope \quad (eq \quad 2-1)$
$X_G \geq 10 \Rightarrow Gazelle \quad (eq \quad 2-2)$
$X_L \geq 10 \Rightarrow Lion \quad (eq \quad 2-3)$

Max demand generates these equations
$X_A \leq 25 \Rightarrow Antelope \quad (eq \quad 3-1)$
$X_G \leq 30 \Rightarrow Gazelle \quad (eq \quad 3-2)$
$X_L \leq 60 \Rightarrow Lion \quad (eq \quad 3-3)$

The objective function is maximize $3X_A + 8X_G + 12X_L$ (eq 4)

This model illustrates a typical CPE model. This formulation ignores the complexity generated by the deployment, the operating curve, and moving from a part level to an operation level.

4 LINKING CAPACITY AND CYCLE TIME WITH THE OPERATING CURVE

In section 2.3 we established that a function could be identified that directly related cycle time and resource. While there are many such functions, the curve below is from Morrison and Martin (2006):

$$CTM = 1 + \text{offset} + \alpha\left[\frac{util^M}{1-util^M}\right] (\text{eq } 5-1) \qquad util = \left(\frac{CTM - (\text{offset}+1)}{CTM - (\text{offset}+1) + \alpha}\right)^{\frac{1}{M}} (\text{eq } 5-2)$$

- **CTM** is the cycle time multiplier of raw process time (RPT) – measure of cycle time
- **util** is tool utilization of the entity (expressed as a percentage) – facility, tool set, checkout clerks, etc.
- **offset** represents several of aspects of the process that generate wait time that cannot be eliminated.
- **M** is the number of identical parallel machines or servers. Typically this value ranges from 1 to 4
- **α** represents the amount of variation in the system (arrival times, service times (including machine outage, raw process time (RPT), and operator availability)) and controls how long the curve stays flat. The lower the value of **α** the less variation and the longer the curve stays flat.

How do we use these equations? Using the section 3 example, assume the business has decided to only start LION wafers, and therefore a quick evaluation of table 2 makes its clear DUV will be the bottleneck tool. Assume the business decides to start 9 wafers a day for LION. Since LION consumes 10 units (CAPREQ) of capacity per wafer start for the resource DUV, then 90 (=9x10) units of capacity for DUV are needed in total for DUV. Since there are only 100 units of DUV capacity available (CAPAVAIL), the tool is 90% utilized. What cycle time will the FAB be able to achieve? Equation 5-1 will enable us to estimate this CTM, but we first need an estimate of offset, alpha, and M appropriate for DUV. For this example assume offset=1, alpha=0.5, and M=1. When util=90%, CTM = 6.50.

Suppose the business decides this is not acceptable and asks what utilization is needed to achieve a cycle time multiplier of 4.0. Then using equation 5-2, the required utilization is 80%. This means 20% of the available capacity has to be reserved (idle without WIP) to meet this cycle time target! If alpha (amount of variation) increases to 0.8, then the utilization rate to achieve this cycle time decreases to 71%. If alpha decreases to 0.4, then the required utilization is about 83%.

5 DEPLOYMENT AND ROUTES – COMPLICATING CAPAVAIL AND CAPREQ

The reentrant flow characteristic of the manufacturing route ensures that each wafer will visit a tool group or resource multiple times. "Visit" is also called a "pass." However, the detailed manufacturing activity or operation at each pass may be the same or different and operations may be shared across routes.

5.1 Expanding the MUV CAPREQ and CAPAVAIL to MUV Operations – The Transition

In the previous example the consumption rate in table 2 is the number of passes that part family has with each resource. In this section we will drill down on the resource MUV. We switch from viewing MUV as a single resource servicing part families to viewing MUV as a collection of operations (MUV operations) served by a set of tools (MUV tools) where each part family has a route which includes a sequence of MUV operations. If we ignore the order of operations, it can be viewed as mix of operations. The mix varies between part families, but operations are shared. Each tool services only a subset of the operations. From a central planning perspective, this is a critical transition and generates a good bit of complexity – the illusion of the clean link between a part and a resource has vanished.

Table 4 breaks down each pass for each part family to a sequence of specific MUV operations handled by the tools which service MUV. For now, assume we do not know anything about these "MUV tools." Antelope has 5 MUV passes and the sequence is muvop01, muvop02, muvop03, muvop01, and muvop05. Observe operations are repeated within a part family and "shared" between part families. Table 5 takes the information in table 4 and summarizes it by eliminating the pass sequence and focusing on the operation count. There is one row for each MUV operation and one column for each part family where the cell value is the number of passes the part family has through each operation. Gazelle has two passes through MUV operations 01, 04, and 05 and it has one pass at operations 02 and 03. With this information we can expand the MUV capacity column in table 2 to include operation level granularity (table 6). We see the MUV consumption rate of 5 for Antelope is composed of 2 units for MUV operation 01, 1 unit for operation 02, 1 unit for operation 03, and 1 unit for operation 05.

Table 4: Detailed flow sequence of each part through MUV

	Part Family		
pass	**Antelope(5)**	**Gazelle(8)**	**Lion(6)**
pass 1	muvop01	muvop01	muvop01
pass 2	muvop02	muvop02	muvop02
pass 3	muvop03	muvop03	muvop06
pass 4	muvop01	muvop01	muvop06
pass 5	muvop05	muvop04	muvop07
pass 6	na	muvop04	muvop05
pass 7	na	muvop05	na
pass 8	na	muvop05	na

Table 5: Number times part family invokes a specific MUV operation

	Part Family		
operation	**Antelope(5)**	**Gazelle(8)**	**Lion(6)**
muvop01	2	2	1
muvop02	1	1	1
muvop03	1	1	0
muvop04	0	2	0
muvop05	1	2	1
muvop06	0	0	2
muvop07	0	0	1

Table 6: CAPREQ from table 1 for MUV resource expanded to granular level of MUV operations

		MUV	→	muvop01	muvop02	muvop03	muvop04	muvop05	muvop06	muvop07
Part Family	Antelope	5	→	2	1	1	0	1	0	0
	Gazelle	8	→	2	1	1	2	2	0	0
	Lion	6	→	1	1	0	0	1	2	1
CAPAVAIL		100	→	cap01?	cap02?	cap03?	cap04?	cap05?	cap06?	cap07?

We placed a "cap0N?" in the CAPAVAIL cells. "cap0N?" represents the amount of capacity available from the tools to support the MUV operations specific to MUV operation N. If we knew the values of "cap0N?", we could replace the single capacity constraint for MUV (eq 1-1) in the original model with seven constraint equations – one for each MUV operation. This is called equation set MUV which we divide into two components – equation set MUV-RE1 and MUV-RE2. The reason for this split will be clear in the next section. The individual equations are number 1-1-X. The first digit (1) indicates a constraint equation, the second (1) is the index for MUV, the third (X) is for the specific MUV operation.

$$2X_A + 2X_G + 1X_L \le cap01?\ \ muvop01\ \ (eq\ 1-1-1)$$
$$1X_A + 1X_G + 1X_L \le cap02?\ \ muvop02\ \ (eq\ 1-1-2)$$
$$1X_A + 1X_G + 0X_L \le cap03?\ \ muvop03\ \ (eq\ 1-1-3)$$

Equation Set MUV-RE1

$$0X_A + 2X_G + 0X_L \le cap04?\ \ muvop04\ \ (eq\ 1-1-4)$$
$$1X_A + 2X_G + 1X_L \le cap05?\ \ muvop05\ \ (eq\ 1-1-5)$$
$$0X_A + 0X_G + 2X_L \le cap06?\ \ muvop06\ \ (eq\ 1-1-6)$$
$$0X_A + 0X_G + 1X_L \le cap07?\ \ muvop07\ \ (eq\ 1-1-7)$$

Equation Set MUV-RE2

Equation Set MUV – the 7 MUV Operations split into two Equation Sets (MUV-RE1 and MUV-RE2)

However, we don't know the value of "cap0N?". For that information, we need the second piece of the puzzle → the deployment information that links the MUV operations to the MUV tools.

5.2 Deployment Table – Linking MUV Operations to Tools which Support MUV Operations

In this example we have five tools (MUVTL01 to MUVTL05) in the MUV resource to provide coverage for the seven MUV operations. Table 7 contains the core template for the deployment table linking MUV operations to MUV tools. The "?" indicates we have yet to decide if this tool can service this operation. "???" refers to the raw capacity available for this tool after accounting for various factors (Martin 1999). In the cases that follow, we will review the impact of different deployment options on the ability to estimate capacity available (CAPAVAIL) accurately for the model presented in section 3.

Table 7: MUV deployment table core structure link 7 operations with 5 tools

		MUV Tools				
		MUVTL01	MUVTL02	MUVTL03	MUVTL04	MUVTL05
MUV Operations	muvop01	?	?	?	?	?
	muvop02	?	?	?	?	?
	muvop03	?	?	?	?	?
	muvop04	?	?	?	?	?
	muvop05	?	?	?	?	?
	muvop06	?	?	?	?	?
	muvop07	?	?	?	?	?
Capacity Avail		???	???	???	???	???

Case 1: Simplest Deployment Table – Easy Street for Navigating Capacity. Table 8 illustrates a deployment pattern which eliminates the need to drill down below the aggregate statement of capacity found in Table 2 and equation (1-1) for additional accuracy. Here all tools can handle all operations with equal effectiveness (same raw process time) and the capacity available for each tool is the same (20=100/5). For simplicity we assumed each pass has the same raw process time and used this as the consumption rate. In practice this is not true, but scaling factors resolve this complication. With this deployment pattern equation set MUV can be transformed into a single equation by adding all equations together and combining like terms since all tools can service all operations generating equation 1-1.

Table 8: MUV deployment case - all tools handle all operations

		MUV Tools				
		MUVTL01	MUVTL02	MUVTL03	MUVTL04	MUVTL05
MUV Operations	muvop01	1	1	1	1	1
	muvop02	1	1	1	1	1
	muvop03	1	1	1	1	1
	muvop04	1	1	1	1	1
	muvop05	1	1	1	1	1
	muvop06	1	1	1	1	1
	muvop07	1	1	1	1	1
Capacity Avail		20	20	20	20	20

When case 1 is reality, equation set MUV can be replaced with equation 1-1 without loss of accuracy. *However this case is rarely the actual deployment!*

Case 2: Two independent MUV Groups – Divide and Conquer. Table 9 illustrates a deployment pattern where we can divide the operations and tools into two independent groups. In this case the MUV tools TL01, TL02, and TL03 exclusively (and identically) support MUV operations op01, op02, and op03. These tools do not support any other operations and there are no other tools which service these operations. Second, MUV tools TL04 and TL05 exclusively support MUV Operations op04, op05, op06, and op07. In this case we have two independent resource entities within MUV that identically service the MUV operations "belonging" to them. They are

1. MUV Resource Entity 1 – MUVRE1 – tools 1, 2, and 3 servicing operations 1, 2, and 3
2. MUV Resource Entity 2 – MUVRE2 – tools 4 and 5 servicing operations 4, 5, 6, and 7

Table 9: Case 2 two independent subgroups each with identical coverage

		MUV Tools				
		MUVTL01	**MUVTL02**	**MUVTL03**	**MUVTL04**	**MUVTL05**
MUV Operations	muvop01	**1**	**1**	**1**	0	0
	muvop02	**1**	**1**	**1**	0	0
	muvop03	**1**	**1**	**1**	0	0
	muvop04	0	0	0	**1**	**1**
	muvop05	0	0	0	**1**	**1**
	muvop06	0	0	0	**1**	**1**
	muvop07	0	0	0	**1**	**1**
Capacity Avail		20	20	20	20	20

Therefore equation set MUV can be separated into two independent equations sets MUV-RE1 and MUV-RE2; one set for each MUV resource entity. The fact that all tools are exclusive and fully shared within an MUVRE, enables us to reduce equation set MUV-RE1 and MUV-RE2 each to a single equation by adding component equations and combining like terms:

- Equation Set MUV-RE1 becomes $4X_A + 4X_G + 2X_L \le 60$ (eq MUV-RE1-Single). The "-Single" stands for single equation to represent all of the equations in this equation set.
- Equation Set MUV-RE2 becomes $1X_A + 4X_G + 4X_L \le 40$ (eq MUV-RE2-Single)

In this case equation set MUV can be replaced by two equations. We observe these key points:

1. With this deployment scheme (where the tools and operations can be divided into two mutually exclusive groups within which coverage is identical), it is straightforward to create a set of constraint equations that provide full and accurate coverage for MUV in the CPE model structure
2. Even in this simple deployment, the original aggregate level MUV capacity constraint equation (1-1) is a poor approximation for the real constraint equations (MUV-RE1-Single and MUV-RE2-Single) and will always overstate capacity since it incorporates more flexibility than actually exists creating the risk the FAB will commit to more starts than it can handle.

Case 3: complexity increases: non-uniform deployment requires the introduction of a new decision variable – capacity allocation. Table 10 and Figure 2 have a slight modification of the deployment of MUV tools to MUV operations from the one presented in Table 9 For the independent sub group MUVRE1, each tool no longer can support each operation, but is restricted to a subset of the operations.

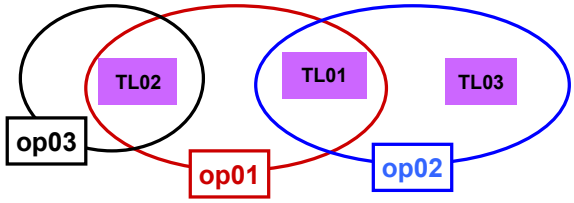


Figure 2: Non Uniform Deployment for MUVRE1

Table 10: Case 3 MUVRE1 non-uniform coverage

		MUV Tools				
		MUVTL01	MUVTL02	MUVTL03	MUVTL04	MUVTL05
MUV Operations	muvop01	**1**	**1**	**0**	0	0
	muvop02	**1**	**0**	**1**	0	0
	muvop03	**0**	**1**	**0**	0	0
	muvop04	0	0	0	**1**	**1**
	muvop05	0	0	0	**1**	**1**
	muvop06	0	0	0	**1**	**1**
	muvop07	0	0	0	**1**	**1**
Capacity Avail		20	20	20	20	20

How does this impact identifying the capacity available (cap01?, cap02?, and cap03?) for equation set MUV-RE1? What is the risk in using e.q MUV-RE1-Single instead of all three equations? It creates a situation that requires a careful balance between solution accuracy, model complexity, model performance, and stressing the social order. We have the following solution options:

1. Make the assumption all tools can handle all operations and continue to use equation MUV-RE1-Single. The risk is we overstate capacity flexibility potentially committing the FAB to produce more than it is able to produce. If the "deviation" from uniformity is low, this is a reasonable option.
2. Establish a fixed allocation of tool capacity for tools 01, 02, and 03 to each operational constraint enabling us to estimate cap01, cap02, and cap03 and therefore replace equation set MUV-RE1 with three constraint equations (one for each operation) and accept the risk of turning away business.
3. Create a projected wafer start profile that would include demand priorities translated to the starts and use optimization to simultaneously make start decisions and allocate tools between operations to maximize supply against prioritized demand. The downside is we do not know the wafer start pattern until after the CPE runs to determine the best can do (BCD) start profile to meet prioritized demand. Therefore we would need to partition the CPE into an explode run and an implode run and loosely couple the FAB model with the rest of the CPE – probably in an iterative approach.
4. Modify the traditional method of stating capacity restrictions in CPE models to include the "or" conditions. This similar to, but not identical with, handling alternative operations in CPEs.
5. Introduce a capacity allocation decision set to the model which is described in the next section. There is a rich history of this type of approach in FAB tool planning dating back to at least the early 1980s in practice and literature. This decision set is close to the "M" variable in Hung and Cheng (2002).
6. Institute a combination of these options guided by some heuristics to combine tools where reasonable into a single resource for the CPE and establish other structures to handle "problem tool sets."

It is outside the scope of this paper to address these solution options in detail.

Next, we introduce the decision set capacity allocation (option 5) to the model using equation set MUV-RE1. This is the allocation of a fraction of each tool to each operation to create the CAPAVAIL.

Introduction of the Capacity or Tool Allocation Decision – Complexity Grows. The values for cap01?, cap02?, and cap03? (CAPAVAIL) in equation set MUV-RE1 are limited to being some combination of capacity allocated to each operational constraint from tools 01, 02, and 03 that does not violate deployment restrictions and does not allocate more than 100% of each tool. The decision on what percentage of each tool to allocate to each constraint (cap0N?) determines the CAPAVAIL for each capacity constraint equation which directly influences the wafer start profile the FAB can support.

Table 11 which has sections a and b illustrates this decision. In table 11a there is one row for each tool and one column for each operation (constraint equation). "??" indicates a decision to be made by the model or some entity to allocate a fraction of this tool to service this operation to best meet some criteria. "NA" indicates the assignment of this tool to this operation is not allowed by the deployment table. The last column is the percentage of a tool allocated across all operations and must be ≤ 100%.

Table 11a: Percentage of each tool allocated to each operation

CAPAVAIL	MUV Tool	muvop01	muvop02	muvop03	total
20	**MUVTL01**	??	??	**NA**	0.0%
20	**MUVTL02**	??	**NA**	??	0.0%
20	**MUVTL03**	**NA**	??	**NA**	0.0%
60	**<-- total capacity available**				

Table.11b has a specific allocation decision. 100% of TL01 is assigned to op01. TL02 has a 40/60 split between op01 and op03. 85% of TL03 is assigned to op02 and 15% is unassigned.

Table 11b: Specific allocation of each tool to each operation

CAPAVAIL	MUV Tool	muvop01	muvop02	muvop03	total
20	**MUVTL01**	100.0%	0.0%	**NA**	100.0%
20	**MUVTL02**	40.0%	**NA**	60.0%	100.0%
20	**MUVTL03**	**NA**	85.0%	**NA**	85.0%
60	**<-- total capacity available**				

Table 12 demonstrates how this allocation decision is "translated" to creating specific CAPAVAIL values (cap01?, cap02?, and cap03?) for each constraint. For example, the value of 8 in cell(TL02,muvop01) is the result of 40% of MUVTL02 being allocated to MUVop01 and the available capacity for TL02 is 20. "cap01?" is 28, a contribution of 20 units from TL01 and 8 units from TL02.

Table 12: Actual CAPAVAIL based on allocation of raw tool capacity

	MUV Tool	muvop01	muvop02	muvop03	total
	MUVTL01	20.0	0.0	0.0	20.0
	MUVTL02	8.0	0.0	12.0	20.0
	MUVTL03	0.0	17.0	0.0	17.0
CAPAVAIL		cap01? =	cap02? =	cap03? =	
Capacity Available		28.0	17.0	12.0	57.0

We have introduced a new decision to the model – capacity allocation between tools and operations to create CAPAVAIL. As with all decision variables in the model, the objective functions remains the same: to allocate material and capacity assets to best meet a prioritized set of demands.

6 ILLUSTRATING HOW THE PIECES FIT TOGETHER – A MORE COMPLICATED NETWORK TO INFLUENCE CAPACITY AVAILABLE (CAPVAIL)

As we saw in the model in section 3 the network of relations in a typical central planning model (right side of Figure 3 below) are simple. The primary FAB decision variable is wafer starts. Customer requirements (demand), cycle time, and capacity (CAPREQ and CAPAVAIL) are input parameters to the CPE model. In this network there are no actions the typical CPE model can take to influence CAPVAIL.

As we have seen, there are a series of decisions incorporated into cycle time and capacity: deployment, the allocation of tools to operational constraints, estimating raw capacity, and the cycle time / utilization trade-off that influence effective capacity available. In this section, we extend our model for MUVRE1 to include demand requirements and the "trade-off" to present a comprehensive one period steady state model demonstrating the complex interactions. We frame this from the point of view of the

decision variables whose values would be determined by an optimization solver. To make clear that each of these tables are related and in aggregate form one model, we will number the equations 13-X.

Decision Variables. In this model there are five decision variable sets: wafer starts, cycle time, raw capacity available, deployment, and capacity allocation. Table 13-1 contains the wafer start decision. The wafer start values of 2, 3, and 6 were arbitrarily selected for illustrative purposes.

Table 13-1: Decision set 1 - atarts for each part family

Part Fam	# wafers		profit per wafer	total profit
Antelope	2.0		$3.00	$6.00
Gazelle	3.0		$8.00	$24.00
Lion	6.0		$12.00	$72.00
total	11.0			$102.00

Table 13-2 has the set of decisions involving cycle time and its impact on effective tool utilization. The first four values are "implicit" decisions and represent an estimate of this tool set's operating curve performance characteristic. Cycle time is a business decision that is converted to the utilization required.

Table 13-2: Decision set 2 - OP curve parameters and cycle time

parameters equation calculate utilization for given cycle time	alpha	**0.5**		
	offset	**1**		
	max utility	**1**		
	numb mach	**1**		
Cycle Time		**4.00**		
utilization required (eq 5.2)		0.80		
cycle time check		4.00		

Table 13-3 column 2 has the decisions (estimates) for raw capacity available for each tool. Columns 4 and 5 apply the utilization required (0.80 from Table 13-2) to allocate the raw capacity available (20) between cycle time reserve (4=20x(1.00-0.80) and effective capacity available (16=20x0.80) Effective CAPAVAIL is eventually distributed to operations and ultimately wafer starts. Table 13-4 has the FAB deployment decisions. Table 13-5 has the capacity allocation decisions. In this example, each tool's capacity was distributed equally between each operation the deployment table permitted it to service.

Table 13-3: Decision set 3 - raw capacity & calculation of effective capacity

MUV Tool	raw capacity available	util req meet cycle time (table 13-2)	cycle time reserve	effective capacity available
MUVTL01	20	0.80	4.0	16.0
MUVTL02	20	0.80	4.0	16.0
MUVTL03	20	0.80	4.0	16.0
total	60		12.0	48.0

Table 13-4: Decision set 4 - deployment decisions

MUV Tool	muvop01	muvop02	muvop03	total
MUVTL01	1.0	1.0	0.0	2.0
MUVTL02	1.0	0.0	1.0	2.0
MUVTL03	0.0	1.0	0.0	1.0
total	2.0	2.0	1.0	

Table 13-5: Decision set 5- capacity allocation

MUV Tool	muvop01	muvop02	muvop03	total
MUVTL01	50.0%	50.0%	0.0%	100.0%
MUVTL02	50.0%	0.0%	50.0%	100.0%
MUVTL03	0.0%	100.0%	0.0%	100.0%

Constraints on the decisions belong to the following major groups: demand, capacity allocation, capacity needed versus available, and deployment requirements. Typical demand constraints would be minimums (must make) and maximums (saturated market).

Table 13-6 contains the constraints that limit the options for capacity allocation (Table 13-5) based on current deployment and inability to allocate more than 100% of a tool. The relationship between these tables is summarized as (Table 13.5) ≤ (Table 13.8) by cell. For example the 0 value in cell(MUVTL03,muvop01) reflects the 0 value in equivalent cell in the deployment table (13-4). The value of 100% in each cell in the total column insure we do not allocate more than 100% of this tool. An example of another limit would be requiring certain tools to service certain operations at least 20% of the time to remain qualified to handle that operation.

Table 13-6: Constraint set 3 - capacity allocation constraints

MUV Tool	muvop01	muvop02	muvop03	total
MUVTL01	100.0%	100.0%	0.0%	100.0%
MUVTL02	100.0%	0.0%	100.0%	100.0%
MUVTL03	0.0%	100.0%	0.0%	100.0%

Table 13-7 has the calculation of actual capacity available based on capacity allocation (Table 13-5) and effective CAPAVAIL (Table13-3). The value in each cell for the operation columns is the effective CAPAVAIL allocated from a tool to this operation. The last row is actual CAPAVAIL for each operation (constraint). Table 13-8 is the traditional calculation of capacity needed to support starts. Table 13-7 is combined with Table 13-8 to create the traditional capacity constraints.

Table 13-7: Calculation CAPAVAIL from capacity allocation and effective capacity

effective capacity available	MUV Tool	muvop01	muvop02	muvop03	total
16.0	MUVTL01	8.0	8.0	0.0	16.0
16.0	MUVTL02	8.0	0.0	8.0	16.0
16.0	MUVTL03	0.0	16.0	0.0	16.0
actual capacity available		16.0	24.0	8.0	48.0

Table 13-8: Calculation of capacity needed from starts and CAPREQ

part family		CAPREQ capacity required per unit start for each part for each operation			
Wafer Starts	Group	muvop01	muvop02	muvop03	total
2.0	Antelope	2	1	1	
3.0	Gazelle	2	1	1	
6.0	Lion	1	1	0	
calculation capacity needed		16	11	5	32

The previous paragraphs review core decisions and constraints. What about the solver? The goal remains to find a set of decisions which are feasible and optimize the ability to meet a prioritized set of demands. However, the number of decisions is larger and the relationships more complex (Figure 3).

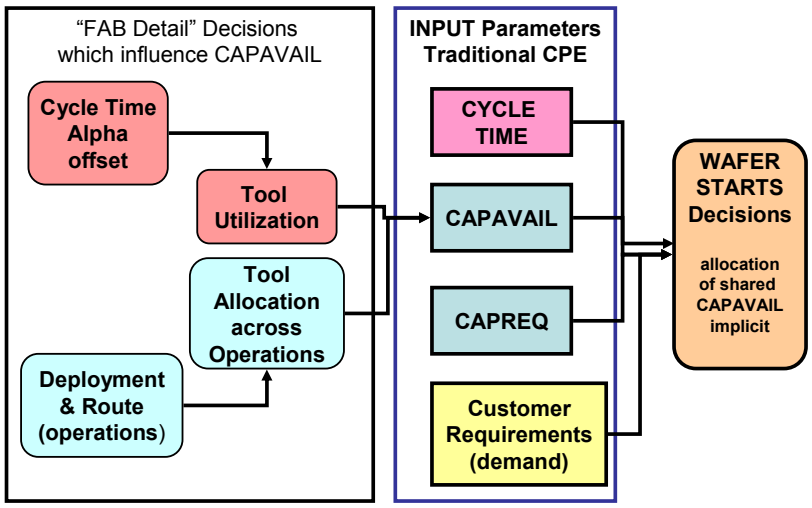


Figure 3: Illusion of Capacity Relationships and Complexity of the Network

7 CONCLUSION: DIFFERENT WORLD VIEWS COLLIDE

The core of the complexity is integrating the different world views of CPE and the FAB. The CPE has parts, resources, and demand: a part (e.g. Antelope) is linked to a process to build that part where the "process" has build instructions which include cycle time, component parts, and which resources are consumed at what rate (CAPREQ). The resource (MUV) is *a mythical entity* with a fixed amount of capacity available. Given a set of demands, the CPE finds a solution to best meet demand. This view is different than the FAB view composed of operations, tools, and starts which interact as follows:

1. collection of operations that are similar (but not identical) in nature (e.g. MUV operations).
2. set of tools (MUV tools) which service these operations - typically all tools do not handle all operations; which tools handle which operations is called deployment.
3. Part or Part Family (Antelope, Gazelle, Lion) have routes which are sequences of operations across different collections of operations (MUV, DUV, ION, etc). If we ignore the order of operations, we can view each part as consuming some mix of operations (which operations in what quantity). The mix of operations varies between parts, but often operations are shared between parts.

The net is CPEs are part / resource centric. FABs are operation / tool centric. Their respective planning is linked because each part is produced by a set of operations and each resource is a set of similar but non-identical tools. The difficulty in linking the FAB and CPE stems from the difficulty of aggregating non-identical operations into an overall part view and non-identical tools into a single resource. This leads to errors (over committing or missed opportunities) when utilization is high. To improve responsiveness, the challenge is to seamlessly incorporate important details of the left (FAB) side of Figure 3 with the right (CPE) side recognizing the CPE must create a plan for the entire enterprise which drives the requirement to identify the critical components of FAB capacity to incorporate into the model.

REFERENCES

Bagch S, Ching-Hua C, Shikalgar, S. and Toner, M. 2008. "A full factory simulator as a daily decision support tool for 300mm wafer fabrication productivity," *WSC'08 (MASM 2008)*, pp 2021-2029

Bermon S. Hood, S. 1999. "Capacity Optimization Planning System," Interfaces, vol.29, no.5, pp.31 – 50.

Dobson, G. and Karmarkar, U. 2011. "Production Planning under Uncertainty with Workload-Dependent Lead Times: Lagrangean Bounds and Heuristics," chapter 1 in *Planning Production and Inventories in the Extended Enterprise*. Vol. 1. Springer. editors Kempf, Keskinocak, Uzsoy.

Elmaghraby, S. 2007. "Production Capacity: Its Bases, Functions and Measurement," North Carolina State University, Raleigh, NC elmaghra@eos.ncsu.edu

Fordyce, K., Wang, C., Milne, R.J. et al 2011. "Ongoing Challenge: Creating an Enterprise-Wide Detailed Supply Chain Plan for Semiconductor and Package Operations," chapter 14 in *Planning Production and Inventories in the Extended Enterprise*. Springer, ed: Kempf, Keskinocak, Uzsoy.

Fordyce, K. and Milne, R.J. 2012. "The Ongoing Challenge for a Responsive Demand Supply Network, The Final Frontier – Controlling the Factory," chapter 3 in *Decision Policies for Production Systems* edited by Karl Kempf and Dieter Armbruster, Springer-Verlag,

Hung, Y. and Cheng, G. 2002. "Hybrid Capacity Modeling for alternative machine types in linear programming production planning," *IIIE Transactions*, 34, 157-165

Martin, D. 1999. "Capacity and cycle time-throughput understanding system (CACTUS)," Advanced Semiconductor Manufacturing Conference and Workshop Proceedings, pp. 127 - 131

Morrison, J. and Martin, D. 2006. "Cycle Time Approximations for the G/G/m Queue Subject to Server Failures and Cycle Time Offsets with Applications," ASMC 2006 Proceedings, pp. 322-326

Kacar N, Irdem, D, and Uzsoy, R. 2012. "An Experimental Comparison of Production Planning Using Clearing Functions and Iterative Linear Programming-Simulation Algorithms," *IEEE Transactions on Semiconductor Manufacturing.*

Schelasin, R. 2011. "Using Static Capacity Modeling and Queuing Theory Equations To Predict Factory Cycle Time Performance In Semiconductor Manufacturing," MASM 2011.

Tibbits, B. 1993. "Flexible simulation of a complex semiconductor manufacturing line using a rule-based System," *IBM Journal Research and Development*, Vol. 37, No. 4, pp. 507-521

Appendix 1: A Cultural Preference to State Capacity as Wafer Starts

An additional challenge to linear structures in CPEs is the cultural preference to state FAB capacity as a nested set of wafer starts or wafer exits limits. Table 14 has one example.

Table 14: Example Stating FAB Capacity Limits as a Nested Set of Start Limits

Group	Time frame 1	Time frame 2	Time frame 3
Wiring Group 1	600	675	675
Technology Group A	400	425	450
Technology Group B	300	325	350
Option set W	100	100	100
Option set X	210	300	300
Wiring Group 2	500	525	550
Technology Group D	350	350	375
Technology Group E	250	275	275
Option set Y	100	100	100
Option set Z	200	200	200
Total Fab Limit	1000	1100	1150

In this example a part that maps to option set W also maps to Technology Group B. For example, a part consuming some of the 100 units of Option set W capacity simultaneously consumes some of the 300 units of Technology Group B. The same applies to Option set X. Similarly a part that maps to Option set Y or Z also maps to Technology Group E. A part can belong to at most one option set, at most one technology group, and at most one wiring group. All parts belong to "Total FAB limit." This method allows the CPE to start up to, but not exceed any limit to which products are mapped.

AUTHOR BIOGRAPHIES

KENNETH FORDYCE joined IBM in 1977 and is a senior computational decision scientist and holds a Ph.D. from Union University. His email address isfordyce@us.ibm.com.

R. JOHN MILNE is the Neil ’64 and Karen Bonke Assistant Professor in Engineering Management at Clarkson University. Prior to this, he spent 26 years with IBM mostly in the Micro-Electronics Planning and Scheduling group as senior technical staff. His work has focused on the research and application of operations research to decision problems in supply chain management. This work was recognized by INFORMS with the Franz Edelman Finalist Award for Achievement in Operations Research and the Management Sciences and the Daniel H. Wagner Prize for Excellence in Operations Research Practice. He holds a Ph.D. from Rensselaer Polytechnic University. His email address is jmilne@clarkson.edu.

JOHN FOURNIER is a senior industrial engineer for the IBM FAB in Burlington, Vermont where he is the technical leader for all aspects of planning, scheduling, and dispatch. Before hold this position, he was a senior manufacturing engineer with direct responsibility for tools. He holds an undergraduate degree in chemical engineering, a masters in material science, and has taken graduate courses in OR and statistics. His email address is fourniej@us.ibm.com.

HARPAL SINGH is the Chief Executive Officer and co-founder of Arkieva. He is recognized as a thought leader, teacher, writer, software designer, and consultant for process change. In addition to years of industry and consulting experience, he has taught graduate and undergraduate courses and run numerous executive management seminars. He holds a Ph.D. in operations research from Cornell University. His email address is hsingh@arkieva.com.