
Managing constrained capacity: a simulation study

Eldon Y. Li

Department of Management Information Systems
College of Commerce, National Chengchi University
64, Sec. 2, Zhi-nan Rd., Wenshan, Taipei 11605, Taiwan
Fax: +8862-2939-3754
E-mail: eli@calpoly.edu

Joseph R. Biggs*

Orfalea College of Business
California Polytechnic State University
San Luis Obispo, CA 93407, USA
Fax: +1-805-756-1473
E-mail: jbiggs@calpoly.edu
*Corresponding author

Emil A. Thies

College of Business Sciences, Zayed University
P.O. Box 4783, Abu Dhabi, United Arab Emirates
Fax: +971-2-443-4847
E-mail: emil.thies@zu.ac.ae

Abstract: This research focuses on the use of constrained capacity availability as a managerial decision variable in production systems. A simulation model was used to evaluate production system performance at various levels of constrained capacity before additional incremental levels of capacity were added. The results indicate that as levels of available capacity approach planned demand levels, or as demand approach capacity levels, the system performance declines. This deterioration of system performance begins at some capacity level higher than average production requirements. That is, at 100% planned capacity utilisation, system performance is considerably degraded, suggesting the need for local managerial manipulation of short-term available capacity levels. In addition, this paper reviews several issues related to constrained capacity and bottleneck management that need clarification.

Keywords: capacity management; capacity constraint; inventory level; service level; production management; bottleneck.

Reference to this paper should be made as follows: Li, E.Y., Biggs, J.R. and Thies, E.A. (2005) 'Managing constrained capacity: a simulation study', *Int. J. Internet and Enterprise Management*, Vol. 3, No. 4, pp.365-384.

Biographical notes: Eldon Y. Li is University Chair Professor of the College of Commerce at the National Chengchi University in Taiwan. He was Professor and Dean of College of Informatics at Yuan Ze University in Taiwan during 2003-2005. He was Professor and Coordinator of MIS Program at the College

of Business, California Polytechnic State University, San Luis Obispo, California, USA. He visited the Department of Decision Sciences and Managerial Economics at the Chinese University of Hong Kong during 1999–2000. He was Professor and Founding Director of the Graduate Institute of Information Management at the National Chung Cheng University in Chia-Yi, Taiwan. He holds a PhD from Texas Tech University. His current research interests are in human factors in Information Technology (IT), strategic IT planning, software engineering, quality assurance, and information and systems management. He is Editor-in-Chief of *International Journal of Electronic Business*, *International Journal of Information and Computer Security*, *International Journal of Information Policy and Law*, *International Journal of Internet and Enterprise Management*, and *International Journal of Internet Marketing and Advertising*.

Joseph R. Biggs is Professor Emeritus of Operations Management in the College of Business, California Polytechnic State University, San Luis Obispo, California. He received a PhD degree from The Ohio State University. His research interests include purchasing, supply chain management, outsourcing, operations management, and quality management. He has published articles in the following: *Journal of Purchasing and Materials Management*, *Guide to Purchasing*, *Management Accounting*, *Human Resources Planning*, *International Journal of Production Research*, *Decision Sciences*, and *Journal of the American Production and Inventory Control Society*. He is past President of the Western Decision Sciences Institute.

Emil A. Thies has been Assistant Dean of the College of Business Sciences at Zayed University since August 2001. From 1991 to 2001, Dr. Thies was the Programme Director of the Fort Smith University Centre for Arkansas State University. Prior to that, he was Associate Professor of management at California Polytechnic State University in San Luis Obispo, California. Dr. Thies received his PhD in 1984 from the University of Missouri, Columbia. He has published numerous articles and proceedings in the areas of service operations, capacity planning and inventory control.

1 Introduction

Capacity has several definitions (Krajewski and Ritzman, 2001; Meredith and Shafer, 2001; Heizer and Render, 2003) such as design capacity, effective capacity, excess capacity, short-term capacity, long-term capacity, and constrained capacity. Most researchers agree that capacity is the ability to produce products or services, and *effective* capacity results after considering the available factors of production and limitations imposed by product design and process design.

Process constraints limit the actual effective capacity of a production process. Included in these process constraints are raw material availability, plant location, plant layout, design and method of processing machinery, skill levels and training of the process operators relative to the learning curve or manufacturing progress function (Meredith and Shafer, 2001), system or preventive maintenance strategy (or lack thereof) (Nicholas, 1998; Steudel and Desruelle, 1992; Bell and Burnham, 1991), and replacement plans for obsolete and inefficient processes (Bell and Burnham, 1991;

Schonberger, 1997). Although we usually find only the cost of the operator being considered in capacity trade-off problems, most process constraints are related to the final cost of production. Constrained capacity is effective capacity which is less than or equal to demand, and which limits the production output of the system.

The limits on constrained capacity are caused by disruptions in product or service design (Umble and Srikanth, 1990) including the product design itself, product quality requirements, required volume, location, price, or any possible combination of these. Since our goal is to make a profit, and since the market determines the price, then price minus expected profit is equal to the cost, which can be attributed to the production function. Thus, the expected price can be a significant factor in the design of the process capacity.

The capacity problem

The common capacity problem is not the grossly over- or under-constrained capacity system, but rather the managerial problem of trying to fine-tune capacity; *i.e.*, planning for a small increment more, or a small increment less, than anticipated demand. A better plan would be to have some small increment of excess capacity.

The issues of capacity availability and the impact on system performance caused by aggregate differences from demand requirements have been explored only peripherally. This limited research has suggested that the capacity utilisation level at which production systems operate may influence system performance. Dixon and Silver (1981) found that in a single level, multi-item production system, as the difference between the production schedule time requirements and production time available decreased, the system performance also decreased. Collier (1980), in studying a single item system, reported that capacity constraints had an effect on lot-size heuristic performance. Biggs (1979; 1985), focusing on a multi-item system, acknowledged that the capacity utilisation level has an effect on system performance. Schonberger (1997) reported that manufacturers usually have a 'herky-jerky' production schedule, which causes high capacity management costs. Nicholas (1998) notes that any operation can become a bottleneck if a large batch is released to it in a short time, while Chase *et al.* (2003) note that a capacity constrained resource could become a bottleneck if scheduled improperly. McClelland (1988) explored the use of finite system loading through the order promise as a method for better managing the situation of lumpy production requirements. Her work included the use of inventory to smooth production and prevent long cycle times; *i.e.*, the period of time elapsed beginning with receipt of an order and ending with the time the order is promised.

These research results indicate that as the available capacity becomes more constrained, the expected volume of production falls short of expectations. Using a computer simulated production process, this paper intended to bring a production system to levels of constrained capacity in order to determine the causes of less-than-expected output. Of further interest were local managerial measures or responses that could be implemented to better manage this constrained capacity condition. In addition, the relevant literature of constrained capacity was examined to determine where the results reported by this paper fit into the body of knowledge.

2 Literature review

2.1 Time and other capacity dimensions

Capacity is a time-related resource; *i.e.*, if a unit of capacity for this immediate hour is not used this hour, it becomes a forgone, non-retrievable resource, and much of the attendant costs are incurred whether or not the capacity to produce is used. If this process causes a bottleneck, or Capacity Constrained Resource (CCR) (Chase *et al.*, 2003), then the idle time of the processes before and after the bottleneck is also forgone. The only savings attributable to this unused capacity are the deferred wages of process operators and possible 'wear and tear' on equipment.

2.2 Bottlenecks

When two or more processes must be used to manufacture a product, one process is usually slower than the others and becomes the upper constraint on total volume throughput; *i.e.*, a bottleneck. Chase *et al.* (2003) define a bottleneck as being a resource whose capacity is less than the demand placed on it (Meredith and Shafer, 2001), limiting production volume to all other process linked to it. Capacity Constrained Resources (CCRs) are those other processes that are nearly at their capacity limit in relation to demand (Chase *et al.*, 2003). The prudent manager must maintain an awareness that is broad enough to include these processes, as well as the bottlenecks, in order to maximise the output of the system.

2.3 Flexibility

With today's worldwide economy, the profusion of computer-aided manufacturing, and changes in market demand, many feel that the successful firm of the future will need greater flexibility in process methods, production volume, and all facets of demand related to product offerings (Krajewski and Ritzman, 2001; Heizer and Render, 2003; Chase *et al.*, 2003; Meredith and Shafer, 2001). This means that there will be a need for greater flexibility in manufacturing capacity. The strategy implies that we need to move as fast as possible away from the traditional, historical, inflexible, special-purpose-machine end of the continuum. In order to have the capacity flexibility to manufacture low volume products of different designs to respond to market pressures quickly, we need to have general-purpose capacity, which historically has been the higher variable cost, lower quality end of this continuum. One response to this dilemma is the use of general-purpose machines that are controlled by computers, or CAD-CAM (Computer-Assisted-Design/Computer-Assisted-Manufacturing). Thus, computer-controlled, general-purpose machines with relatively cheap software, in many instances, can replace expensive, special purpose machines and tooling to provide some of the flexibility noted above without the higher variable cost and loss of product quality.

2.4 Methods of capacity management

Setup time is the time a processing machine is non-productive because it is being made ready for the next production part schedule. In changing the expensive, time-consuming methods of set-up when implementing Just-In-Time (JIT), one difficult aspect is that it requires the coordination of not only the production department, but that of several functional areas within the plant. Many firms have solved the setup problem by locating all the tools that are to be used in a particular process on a conveyor rack around that machine. All these tools are adjusted by the toolmakers so that the critical dimensions are the same for all. Now it takes only one person to change the tooling in a very short time (five minutes or less). Such process redesign/layout calls for the coordination of the plant and tooling engineers, the toolmakers, and the production departments and overall budget authority to incur the cost of these changes. Instead of incurring high costs due to the loss of the idle time during setups, the production department can now produce more with the same 'design' capacity. Many of these changes have come about at the insistence of those implementing JIT. Nicholas (1998) does an excellent job discussing setups and methods for minimising the time required, as does Shingo (1985) with his four-stage method for setup reduction.

2.5 Maintenance

As noted above, the design capacity of a production function is determined assuming most system 'support functions', or process constraints, are at predetermined levels. One of the major support functions is maintenance of the processing equipment. Advocates of JIT insist that to maintain capacity, preventive maintenance functions are very necessary so that production can be accomplished 'just-in time'. Thus, a reasonable preventive maintenance program would seem to be a good management strategy for constrained capacity by making sure that constrained capacity is available when needed and not sitting idle and non-productive while being repaired.

2.6 Managerial uncertainty

Capacity problems are focused on the fact that we have either too much capacity or we have too little capacity. Seldom does a production manager have exactly enough capacity to meet the production requirements of a particular time period. Sources of managerial uncertainty that cause the inequality between demand and capacity include the uncertainty related to forecasting demand, the uncertainty in predicting the capacity actually available at any given time, and the uncertainty in predicting how much capacity, or standard processing time, will be needed to produce a unit of product or service (Wemmerlov, 1984).

Capacity Requirements Planning (CRP) is the use of the standard time of a single unit to predict the capacity needed to make several. We all know the time study or work study method of determining the standard time to make one unit (Krajewski and Ritzman, 2001; Chase *et al.*, 2003). However, very few time or motion studies are performed in today's industry. The method used to calculate the standard time is the historical average time in the database for a particular component manufactured on a particular process as charged over time by the production department, and this just does not guarantee a very accurate standard production time.

Capacity cannot be readily reduced and sold off; yet, in many instances, a capacity utilisation factor (Krajewski and Ritzman, 2001; Meredith and Shafer, 2001; Heizer and Render, 2003) is calculated in an attempt to better manage capacity. Goldratt and Cox (1984) ignore capacity utilisation except for the constrained or bottleneck process; and for those, they advocate maximum utilisation. For other processes, they recommend producing just enough to keep the bottleneck processes busy. To maximise the utilisation of those processes that are non-bottlenecks is to make useless work-in-process inventory, thus adding to operating cost rather than decreasing it.

2.7 *Safety capacity*

In most JIT installations, average capacity, over the long run, is usually designed to be slightly greater than average demand. Vollmann *et al.* (1997) discuss inventory safety stock as that inventory greater than needed to satisfy gross requirements, and they advocate safety lead time as a substitute. Safety lead time is the procedure whereby the production time schedule is moved up so that the produced goods are available ahead of time.

We would like to offer the concept of 'safety capacity'; *i.e.*, adjusting either the capacity or the scheduled production whereby the anticipated inequality for capacity is to be some small amount greater than that of demand. We cannot take credit for the concept of having a little more capacity than demand or having the master production schedule be somewhat less than anticipated capacity because this strategy is recommended by most of the authors in the literature, with no reason given. The strategy of managing error or intentionally biasing forecast events in a more favourable direction has been explored by Biggs and Campion (1982), and they believe that long-term market share is better served by offering a little more rather than a little less. We offer the 'safety capacity' concept as a label and as a rationale for including a small amount of over-capacity instead of a small amount of under-capacity for potential bottleneck processes when planning capacity.

JIT means that the managerial slack or inventory cushion, by managerial edict, is taken away, necessitating that we learn to manage differently than we have in the past. Decision-makers must respond to unanticipated differences between demand and capacity more effectively, and respond more quickly because they have very little, if any, extra inventory and very little extra capacity with which to respond to unanticipated demand. In addition, these responses to problems occur more frequently, with shorter response time, and with smaller increments and adjustments possible. Typically, first-line supervisors are left with the JIT production scheduling problems, and they behave in a very nervous manner, often verging on panic. The reason for this panic is that they are facing new production problems with no experience to guide them, and with no inventory or capacity cushion to absorb the results of managerial errors between demand and capacity. We would like to point out that this need for more frequent responses to problems is also related to the management of system capacity.

In dealing with capacity uncertainties, the predicted future result, carried to a necessary extreme, will be a computer response to manufacturing environmental stimuli. The computer will be programmed with a 'conditioned response' for these perceived environmental conditions or events. This technique has been successfully demonstrated for many years with the use of computers to fly airplanes such as the F-117 stealth fighter. Human response is much too slow to make the control changes necessary to keep the plane flying. In a like manner, by the time a capacity problem is recognised and a

managerial response developed and implemented to meet the problem, the response is too slow and too late. To respond to production problems without an inventory or capacity cushion, managers need more information faster and need to have predetermined the best managerial response to anticipated possible situations.

2.8 *Managed capacity*

With the exception of macro actions to adjust capacity over the long term, little else is done to manage capacity, other than the use of shop floor-control scheduling rules (Wemmerlov, 1984; Allen *et al.*, 1997; Schonberger, 1997). In this method, the performance criteria are the number of orders processed or the magnitude of the work-in-process inventories. Since the introduction of Material Requirements Planning (MRP) circa 1960–1970 (Orlicky, 1975), there has been an extensive body of research focused on the effects that management decisions can have on system performance. One group of production research assumed, or at least inferred, that the production system had infinite production capacity at the outset, with the primary measures of effectiveness being the minimisation of total set-up costs and inventory holding costs. In contrast, the other group of research assumed the capacity of a production system to be finite or fixed. System performance was evaluated as a result of management decisions within this fixed capacity constraint. Much of the capacity management literature follows the traditional textbook alternatives to capacity, which are hire and fire (layoff and rehire), subcontract (purchase, outsource), and inventory. Stock-outs and backordering, if mentioned at all, are considered a subset of inventory management. In these problems, the relative cost of each alternative is described by Holt *et al.* (1956) in their linear decision rule (also known as the HMMS rule, named for Holt, Modigliani, Muth, and Simon). The only short-term cost attributed to increasing or decreasing capacity is the cost of changing the labour force level by hiring and laying-off. The cost of acquiring new plant and processing equipment or selling off excess and/or obsolete plant and processing equipment is not considered a part of this decision-making process.

It has been noted by purchasing professionals that the purchasing department is the ‘manager of outside manufacturing’ (Dobler *et al.*, 1996). As more manufacturing firms follow the current trends of downsizing and outsourcing (Biggs *et al.*, 1998), the responsibility for insuring that outside capacity is capable of manufacturing to scheduled demand will fall to the purchasing function. With this, we should question whether the outside capacity is truly infinite as historically assumed, or is it also constrained at some point in our planning horizon.

2.9 *Cost of capacity*

In the aggregate planning chapter of most texts there is a section on the timing of the additions of new capacity in relation to increasing demand. These additions are modelled as discrete step functions, while demand is modelled as a continuous function. In these decision-making models the cost of adding capacity is not considered.

If the cost of capacity is treated in the same manner as the cost of inventory, then this new, higher cost should lead to planning for only the bare minimum capacity available to meet the production schedule, as noted above. Carrying this idea a step further, we now believe that the cost to change the volume of production by increasing or decreasing capacity is much greater than the mere ‘hire-fire’ costs of the linear decision

rule (Holt *et al.*, 1956) noted above. At a minimum, the cost of idle processing equipment should be included, part of which should cover the opportunity cost of capital. This reasoning leads us to the conclusion that if capacity costs and the cost to change production volumes are much greater than previously thought, then the best capacity cost minimising strategy would be one of level production, with a minimum number of production changes (Chase *et al.*, 2003; Heizer and Render, 2003). It is a reasonable conclusion that the best way to manage capacity is to run the factory or production function at a level rate most of the time.

In summary, several authors and researchers have noted that as the demand approaches the theoretical limit of the constrained capacity, the expected production throughput becomes less than was expected from that particular level of capacity. One possible reason for this problem in a real world situation is that the standard time value used for capacity planning is inaccurate. Another possibility is that as the available capacity is diminished and the order lot sizes remain at the same level, all capacity is used for only a few components, leaving others not even started and no finished goods completed. If, instead, all order sizes were decreased to suit available remaining capacity so that all of the component parts of any final good were completed, the firm would be ahead by that many units sold. Still another possibility is that other unnoticed, non-bottleneck CCRs become bottlenecks.

3 Research methodology

A large-scale computer simulation model of a multi-product, multi-stage system was used as the vehicle for this study. By using the simulation approach, long experimental time periods can be shortened into relatively short actual times, and the problem of changing only desired variables between replications is minimised. Solutions from a simulation are only as globally applicable as the simulation model is similar to its real world counterparts. However, the same problem exists when applying and transferring results from one real world situation to another. For the simulated factory, Material Requirements Planning (MRP) was used as the ordering structure since many real world companies use MRP for this purpose (Rom *et al.*, 2002).

The material requirements planning model has four types of inventory, namely, raw materials, component work-in-process, subassembly work-in-process, and finished products. There are eight purchased raw materials. The model internally manufactures or assembles 15 components, seven subassemblies, and ten end items (finished products). The simulated factory has one purchasing function and three departments: component manufacturing, subassembly, and final assembly. It has the ability to forecast sales, make aggregate and detailed planning decisions, and accumulate statistical data for the analysis of results.

Two independent variables were singled out as having an effect of interest on the system performance; *i.e.*, forecast/demand variation, and constrained capacity. The purpose was to cause some controllable problems within the constrained system, and then analyse the outcomes. The end item demand is deterministic; *i.e.*, the actual demand for the ten end items matches the forecasts exactly. Further, the demand pattern for each end item was set at an average level with a specified amount of variation. Three levels of demand variation, 10%, 20%, and 30% (the ratio of standard deviation to mean demand), were assumed. There is no distinction between capacity and labour for this study.

Therefore, the simulation was replicated over 14 distinct levels of labour force or capacity, representing incremental increases in available capacity of approximately 5% between levels.

The research was designed to set the capacity of the system at some reasonably constrained level at which the system would perform without having many problems, and then to allow the capacity to be increased by regular increments. As the capacity was increased by these small amounts, various important system outputs were monitored.

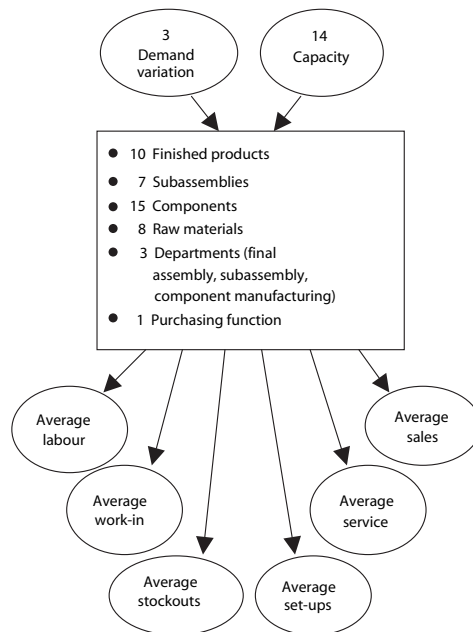
The labour force variable, the variable that was used to control the system constrained production capacity, was initially set at low levels and was increased until the system was able to operate without incurring major problems. The labour force level of capacity started at 248 and was then reset at increments of four additional workers between treatments, up to 300 workers to insure that the system had been given an adequate range to investigate problems of constrained capacity.

This being a simulation, any variable or constraint could be tightly controlled. Two such variables were forecast and demand wherein there is no forecast error. The sample size was set at 30, and the difference between replications was the forecast/demand variable. Thus, there were 30 different forecast/demand patterns that were repeated exactly for all the samples.

Having some initial data, the simulation does not start empty. It is allowed to run for 13 weeks to achieve stochastic convergence, and then it is allowed to run for a 48-week year, with performance data collected at the end of each week. In addition, the simulation is allowed to run four weeks more to preclude ‘plant closing’ anomalies from being included in the statistics.

Thus, the research design was (three forecast variations) × (14 labour force levels) × (30 sample size years) × (48 weeks) which is equal to a total of 60 480 weeks. Each labour force level-forecast variation is a sample of 4320 weeks or 90 years. Figure 1 depicts the simulation model used in this study.

Figure 1 Simulation model



Performance measures

From other research experiences, it has been determined that the only useful global measures are as follows:

- Workforce or capacity – that which the system could use if the capacity were, indeed, immediately adjustable. The system has no new sources of capacity, but if it were in trouble, this is the additional capacity it has calculated it would need to get out of trouble. Thus, it is a measure of the system's production problems with constrained capacity, and in the long run, additional system operating costs.
- Inventory – the Work-In-Process (WIP) inventory at all levels, except the finished goods level. It is a measure of money needed to operate the system under a particular set of conditions, matching that of Goldratt (Goldratt and Cox, 1984; Goldratt and Fox, 1994).
- Stock-outs – they occur at all production levels, including finished goods. The stock-outs indicate the production problems the system is having under a particular level of constrained capacity. This system allows total backorders, so stock-outs do not represent lost sales except at the 48th week.
- Set-ups at all levels – Set-ups are a measure of the problems within the factory (Samaddar, 2001) and are a measure of the amount of constrained capacity that is used for set-ups. They are relative measures of additional system costs and nearly fit Goldratt's (Goldratt and Cox, 1984; Goldratt and Fox, 1994) cost measure.
- Service level – measured at the finished goods or customer level only. This is the best measure of throughput under capacity constraints and best matches Goldratt's (Goldratt and Cox, 1984; Goldratt and Fox, 1994) measure for throughput.
- Sales at the finished goods level – Sales data were collected to assure the simulation was performing per expectations.

Workforce, set-ups, and stock-outs are the best measures of operating cost or production difficulties that the system is experiencing under a particular set of conditions and constraints, and these agree with Goldratt's (Goldratt and Cox, 1984) measures of operating costs. These three, together with the remaining two measures, inventory and service level, accurately reflect the behaviour of the production system. Other researchers have used 'costs' as their major statistic, but our feeling is that costs, as applied to a simulation, are at least one step beyond reality. We leave it to the real world users to look at the performance criteria of this study to determine which are important or of high cost in their individual production system.

4 Results and discussion

The results are displayed in Tables 1–6 and Figures 2–5. It was anticipated that as more capacity was added, the system would perform better. Further, as greater degrees of forecast/demand variation were added, the outputs would be worse or not as good.

Table 1 Desired labour force

<i>Capacity</i>	<i>Forecast deviation</i>		
	<i>10%</i>	<i>20%</i>	<i>30%</i>
248	227	228	229
252	225	225	224
256	223	223	224
260	221	222	223
264	219	220	221
268	217	219	217
272	217	217	216
276	217	215	213
280	215	215	214
284	214	214	211
288	212	211	212
292	210	211	211
296	209	210	210
300	209	209	209

Table 2 Work-in-process inventory (000's)

<i>Capacity</i>	<i>Forecast deviation</i>		
	<i>10%</i>	<i>20%</i>	<i>30%</i>
248	227	226	225
252	224	223	220
256	223	222	221
260	222	220	220
264	221	220	219
268	219	218	216
272	218	217	215
276	218	216	213
280	216	215	214
284	214	214	213
288	212	212	212
292	212	212	212
296	211	211	211
300	211	211	210

Table 3 Stockout level

<i>Capacity</i>	<i>Forecast deviation</i>		
	<i>10%</i>	<i>20%</i>	<i>30%</i>
248	7995	7913	8217
252	7084	7344	6306
256	6382	6617	7039
260	6091	5972	6289
264	6022	5462	6023
268	5261	4861	4211
272	4012	4052	3351
276	4187	3457	2495
280	3440	3079	3063
284	3216	2753	1708
288	1868	1512	1938
292	1093	1314	1624
296	717	734	1057
300	516	548	643

Table 4 Set-up level

<i>Capacity</i>	<i>Forecast deviation</i>		
	<i>10%</i>	<i>20%</i>	<i>30%</i>
248	32.65	32.60	32.49
252	32.67	32.68	32.60
256	32.69	32.61	32.60
260	32.64	32.59	32.66
264	32.61	32.59	32.51
268	32.52	32.52	32.40
272	32.46	32.42	32.31
276	32.44	32.41	32.27
280	32.36	32.35	32.28
284	32.36	32.33	32.22
288	32.28	32.23	32.21
292	32.22	32.22	32.17
296	32.19	32.18	32.13
300	32.17	32.16	32.09

Table 5 Service level

<i>Capacity</i>	<i>Forecast deviation</i>		
	<i>10%</i>	<i>20%</i>	<i>30%</i>
248	96.47	96.29	96.01
252	97.44	97.39	97.47
256	97.75	97.73	97.28
260	98.02	97.90	97.84
264	98.18	98.20	97.92
268	98.45	98.48	98.59
272	98.79	98.73	98.74
276	98.72	99.03	99.24
280	98.89	99.05	99.11
284	99.04	99.16	99.44
288	99.50	99.54	99.37
292	99.69	99.57	99.43
296	99.83	99.81	99.65
300	99.82	99.85	99.78

Table 6 Sales level

<i>Capacity</i>	<i>Forecast deviation</i>		
	<i>10%</i>	<i>20%</i>	<i>30%</i>
248	12687	12661	12631
252	12699	12668	12649
256	12700	12667	12643
260	12700	12675	12648
264	12700	12674	12647
268	12702	12673	12650
272	12698	12676	12653
276	12700	12676	12654
280	12697	12678	12653
284	12702	12675	12655
288	12701	12679	12656
292	12703	12677	12656
296	12705	12679	12656
300	12703	12680	12657

Figure 2 Work-in-process inventory level

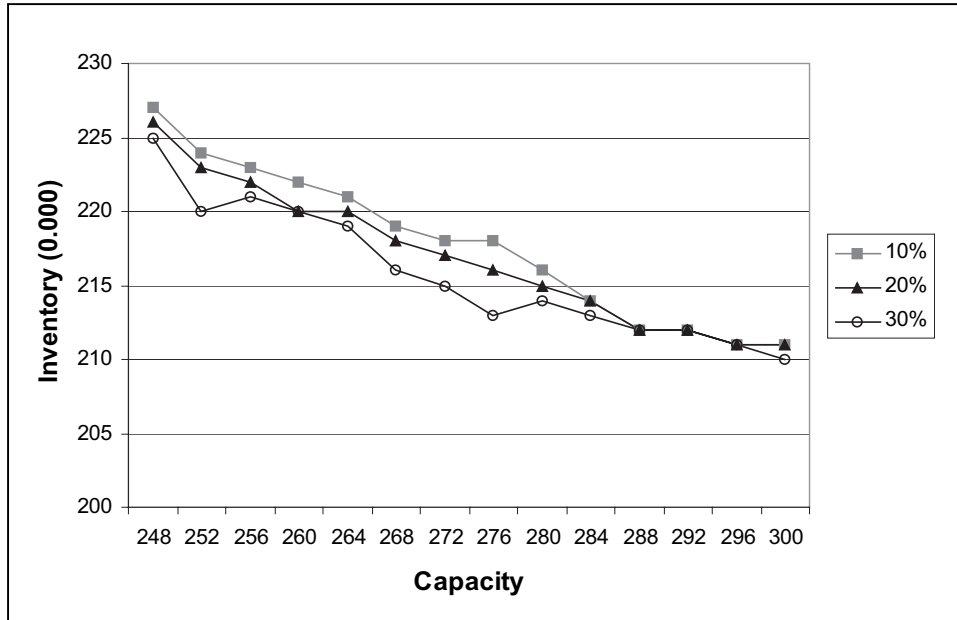


Figure 3 Stockout level

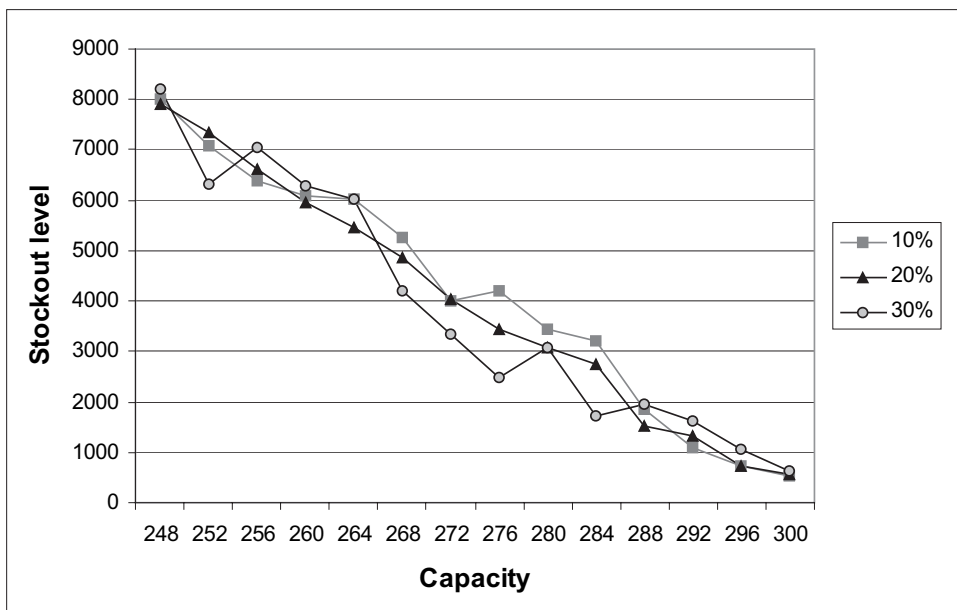


Figure 4 Set-up level

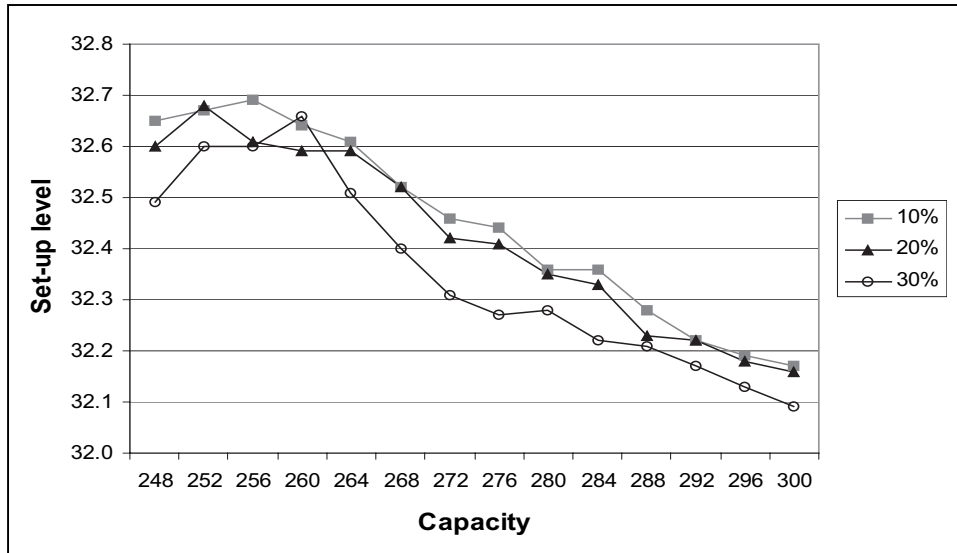
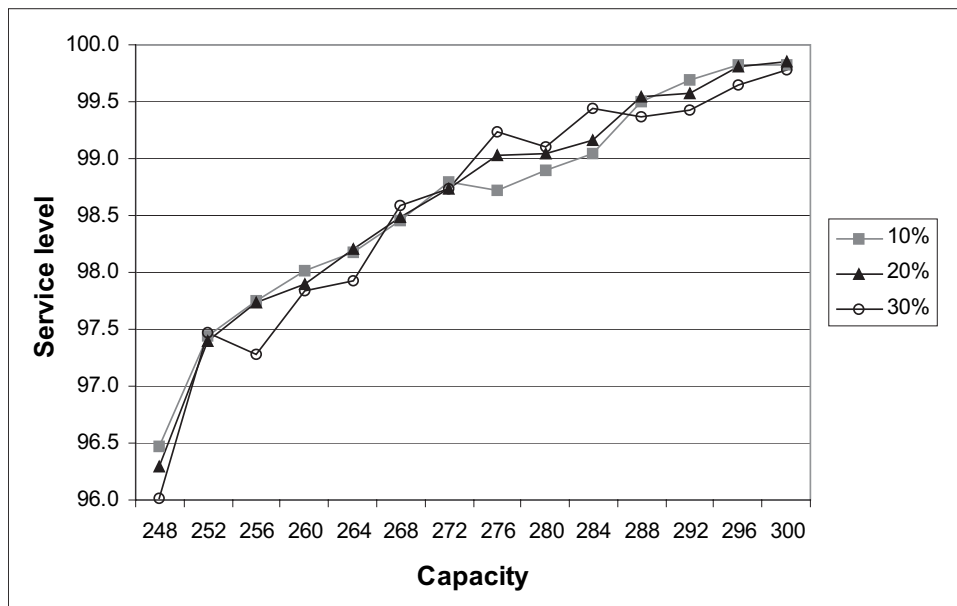


Figure 5 Service level



Per Table 1, in most instances the system called for additional capacity at the 20% demand variation level and for an even greater amount at the 30% level. Although the mean demand does not change, the variation in demand causes scarce capacity, which will be needed in a future period, to go unused. Since the system has no source for the needed additional capacity in the future period, the problems will multiply at an increasing rate. This indicates that the sooner the constrained capacity is enhanced, the fewer future problems there would be.

This phenomenon may explain why production systems, as reported above, perform less than expected when they approach their capacity limits (McClelland, 1988). As the system comes closer to its remaining capacity limits, and order sizes are not downsized, the unfinished portion of the orders impede the production of other needed components. It would have been much better to scale back all orders to a size whereby all could be completed.

Chase *et al.* (2003) advocate increasing order sizes for bottleneck operations to avoid using up constrained capacity as idle time for setups, and the above in no way disagrees with their conclusion. Their assumptions are that other interdependent manufacturing processes in the system can produce to match the large lot sizes used on the bottleneck processes. If most of the production system is constrained, as with the simulated factory, all lots sizes should be scaled back to fit the remaining capacity so that the use of the capacity is maximised and there is no remaining WIP. In this way, the production flow would be balanced to maximise the total flow of production from the system (Goldratt and Cox, 1984; Goldratt and Fox, 1994). Setups should indeed be minimised, but with the goal of matching production volumes of components and subassemblies or balancing the flow so that all value added production is completed as finished goods and not kept on the shop floor as WIP.

When the total system capacity is constrained, as in these experiments, and the system does not have perfect information, the flow of components and subassemblies is not balanced, thus causing unnecessary accumulation of WIP. The constrained capacity has been operated to make components that cannot be used, since the components with which they are to be combined are delayed at another machine. So the system asks for more setups, again using up scarce time on the constrained capacity. Table 2 and Figure 2 exemplify this behaviour of a constrained capacity system, as the capacity increased, the WIP inventory decreased.

Table 3 and Figure 3 show the number of internal stock-outs that were incurred in manufacturing a higher level subassembly or final assembly at each production level, again reflecting the problems within the system and the added cost of operations. As was expected, there were greater stock-outs when the capacity was most constrained. There were fewer stock-outs at the highest levels of capacity, but they did not clear up entirely. This data is consistent with other results that indicate as the demand approaches the constrained capacity limits, the system has undetermined problems which it cannot solve. Given its current level of information and decision-making power, as noted above, the system cannot get out of difficulties.

In reviewing the data in Table 4 and Figure 4, there is some correlation between stock-outs and setups, the reason being that higher WIP and setups are a result of system problems with constrained capacity. As the demand variation becomes greater, the missed production will be produced in following periods because this system allows backorders. Inadvertently, the system is increasing order size, which will also lead to fewer setups. Had the system not allowed backorders, there would be much greater imbalance between the master production schedule and the available capacity.

Table 5 and Figure 5 show the service level at the end item level only; and since total backorders are allowed, the system eventually manages to fill most orders. The internal stock-outs of Table 3 and Figure 3 indicate how many products would have come up short without backorders. So we have a relative measure of both situations. From the lowest capacity level to the highest capacity level within the same demand variation, the service level differences are what some might consider insignificant at about 3% total difference between highest and lowest capacity levels. But consider how this 3% difference in service level might affect market share if the productive capacity were constrained at the lower level, or managed in such a fashion for the next 10–15 years.

Table 6 shows the resulting sales figures. Here, again, the sales do not differ much between high and low levels of constrained capacity, but enough to make a significant difference in market share over the long term. Certainly, sales are better at the lower levels of demand variation, as are the other performance criteria above. This strongly suggests that master production scheduling to level load the constrained capacity would help the system produce at a better quantity. That is, allowing the demand variation to be incorporated into the master production schedule for a constrained capacity production system will increase the production problems, decrease the amount actually produced, increase the costs of production, and, over the long term, cause the loss of market share. A managerial interpretation of these results is that the system performance improves with more capacity, but the costs of the inputs will eventually reach the point of diminishing returns.

5 Conclusion and recommendation

At this point the most interesting findings are that the system did have problems, as the constrained capacity and the demand were nearly equal. The problem that results with this situation is that production volume is less than would have been anticipated for that constrained capacity level. Most successful growing firms at one time or another will face this problem at least once, and perhaps continually, in their corporate history. As a firm expands in a growing market, there will be several instances over time when demand and capacity will be nearly equal before the firm acquires additional capacity. If these problem situations could be anticipated, capacity could be better managed to alleviate the production problems as the firm progresses smoothly through these transitory stages of growth.

Capacity management, from a micro point of view, has been shown to have merit in that system outputs can be improved by the manipulation of capacity inputs. This research demonstrates that if a manager has a source of short-term additional capacity, it may be advantageous to implement this alternative source of capacity as soon as the demand/capacity gets out of balance. In a poorly managed production system, one is never certain whether the problem is constrained capacity or bad scheduling. In a well managed production system, the typical constrained capacity situation is one in which the imbalance between available capacity and immediate demand is of short time cycle and of limited excess demand. Therefore, in order to capture even a part of this momentary excess demand, managerial response must be timely, but not necessarily massive. If a source or sources of small, available, unused capacity could somehow be made available at the time of constrained capacity, many of the problems would cease to exist. Thus, a better strategy would be not to allow stock-outs and WIP to accumulate and multiply.

In the best-scheduled and best-managed production systems, there are unanticipated problems that could be helped by some small amount of safety capacity. But if the system is working at full capacity, we need to investigate possible sources for this safety capacity. The traditional sources were described earlier in this paper:

- Use hire-fire method of acquiring additional workers to run the processes if there is idle processing equipment.
- Overtime, or the use of the current workforce and processes to start work earlier than usual and work a longer day or work additional time at the end of the work week. We know this is not possible in a plant that runs 24 hours per day, seven days per week. Most plants do not work all 21 possible shifts in a week.
- Use inventory to shift production to other time periods, and then use inventory or backorder to meet the demand.
- Subcontract, or outsource, to vendors that needed increment of production capacity.

There are other possible, non-traditional sources. For example, stock-outs are a possible alternative we seldom mention as a solution to the demand/capacity ratio. However, stock-outs and backorders are the method used by custom order shops. Consider the 'No Vacancy' sign at a motel; does that not mean stock-out to the next customer after the 'No' part of the sign is turned on? At the point where demand becomes nearly equal to capacity and production problems arise, one alternative may be to allow stock-outs or to refuse new orders, rather than incur the larger than expected additional costs of production.

Alternative processes may be used to make up the production gap for the constrained capacity. We would like the most efficient process to be used, but there may be other, less efficient processes that could be used to alleviate the constrained capacity. Perhaps an obsolete machine that has not yet been sold can be pressed into service.

It may be advisable to delay planned preventive maintenance downtime until another time period, using the current maintenance time to relieve constrained capacity. As firm believers in preventive maintenance instead of breakdown maintenance, we do not advocate this as a long-term solution. However, in most systems, the processes can operate a bit longer without maintenance, much as our automobile can go a few extra miles until it is convenient to have the oil changed.

If we look around us, there are instances of managers being able to change their production system so that additional capacity is available in the short term. At the local grocery store we see many seemingly unused check-out lanes with their computerised cash registers sitting idle. When the waiting lines on those check-out lanes that are staffed become too long or become bottlenecks, 'part time' cashiers are called from other duties that can be delayed until later, such as shelf restocking, and are pressed into temporary checkout service.

The plant manager of a local manufacturer noted that the productive capacity available to him was being 'diluted' by set-ups for various product items. Not only was this making certain processes bottlenecks, but there was the additional cost of expensive raw material lost as scrap as part of the setups. This plant manager decided it was cheaper to acquire more of the least expensive process machines so that each machine could be semi-permanently set-up for a particular production component and allowed to sit idle when that component was not needed.

This paper has shown that greater capacity results in better system performance and small improvements, over the long term, may enhance the firm's market share. It has reaffirmed the phenomenon that as demand approaches capacity constraints, the problems in the production system multiply. As noted, this research indicates that as the demand approaches the limits of capacity, order sizes or lot sizes should be decreased to balance the product flow. The constrained parts of the system should be treated as bottlenecks per Goldratt (Goldratt and Cox, 1984; Goldratt and Fox, 1994).

On the other hand, this study has not spotlighted the ratio of system capacity to scheduled production required to achieve the goals of the JIT practitioners, as noted at the outset. Obviously, additional research using parameters of each real system is needed to determine the point at which enough capacity is available relative to cost-benefit analysis. Research is also needed to determine the sources and types of incremental capacity that are available to practitioners other than those noted above. Each practitioner will have to carefully weigh the costs of the incremental additional sources of capacity available and make decisions based on both the long-term and short-term effects of implementing any additional capacity. The concept of short-term safety capacity introduced in this paper is not intended as a permanent solution to constrained capacity, but as a temporary, 'stop-gap' measure. When a production problem is perceived, it is recommended that safety capacity be brought into play quickly to forestall an accumulation and multiplication of WIP and unfinished orders. The frequency with which it is needed should act as an indicator that long-term, permanent capacity should be acquired for that machine or process.

References

- Allen, S.J., Martin, J.L. and Schuster, E.W. (1997) 'A simple method for the multi-item, single-level, capacitated scheduling problem with setup times and costs', *Production and Inventory Management Journal*, Vol. 38, pp.39–47.
- Bell, R.R. and Burnham, J.M. (1991) *Managing Productivity and Change*, Cincinnati, OH: South-Western Publishing Company.
- Biggs, J.R. (1979) 'Heuristic lot-sizing and sequencing rules in a multi-stage production inventory system', *Decision Sciences*, Vol. 10, pp.96–115.

- Biggs, J.R. (1985) 'Priority rules for shop floor control in material requirements planning systems under various levels of capacity', *International Journal of Production Research*, Vol. 23, pp.33–46.
- Biggs, J.R. and Champion, W.C. (1982) 'The effect and cost of forecast error bias for multiple stage production-inventory systems', *Decision Sciences*, Vol. 13, pp.570–583.
- Biggs, J.R., Li, E.Y. and Rogers, J.C. (1998) 'The effect of backordering on performance: marketing and production cooperation in supply chain management', *Journal of Marketing: Theory and Practice*, Vol. 6, pp.26–40.
- Chase, R.B., Aquilano, N.J. and Jacobs, R.F. (2003) *Operations Management for Competitive Advantage*, 10th edition, New York: McGraw-Hill/Irwin, 6 May.
- Collier, D.A. (1980) 'A comparison of MRP lot sizing methods considering capacity change costs', *Journal of Operations Management*, Vol. 1, pp.23–29.
- Dixon, P.S. and Silver, E.A. (1981) 'A heuristic solution procedure for the multi-item, single-level, limited capacity, lot-sizing problem', *Journal of Operations Management*, Vol. 2, pp.23–39.
- Dobler, D.W., Burt, D.N. and Lee, L. (1996) *Purchasing and Materials Management*, 5th edition, New York: McGraw-Hill.
- Goldratt, E.M. and Cox, J. (1984) *The Goal: Excellence in Manufacturing*, New York: North River Press.
- Goldratt, E.M. and Fox, R. (1994) *The Race*, New York: North River Press, 21 September.
- Heizer, J. and Render, B. (2003) *Principles of Operations Management*, 5th edition, Upper Saddle River, NJ: Prentice Hall, 19 February.
- Holt, C.C., Modigliani, F., Muth, J.F. and Simon, H.A. (1956) 'Derivation of a linear decision rule for production and employment', *Management Science*, Vol. 2, pp.1–30.
- Krajewski, L.J. and Ritzman, L.P. (2001) *Operations Management*, 6th edition, Reading, Mass: Addison-Wesley, 16 July.
- McClelland, M.K. (1988) 'Order promising and the master production schedule', *Decision Sciences*, Vol. 19, pp.858–879.
- Meredith, J.R. and Shafer, S.M. (2001) *Operations Management for MBAs*, 2nd edition, 10 October New York: John Wiley and Sons.
- Nicholas, J.M. (1998) *Competitive Manufacturing Management*, New York: McGraw-Hill/Irwin.
- Orlicky, J. (1975) *Materials Requirements Planning: The New Way of Life in Production and Inventory Management*, New York: McGraw-Hill.
- Rom, W.O., Tukul, O.I. and Muscatello, J.R. (2002) 'MRP in a job shop environment using a resource constrained project scheduling model', *Omega*, August, Vol. 30, No. 4, pp.275–286.
- Samaddar, S. (2001) 'The effect of setup time reduction on its variance', *Omega*, June, Vol. 29, No. 3, pp.243–247.
- Schonberger, R.J. (1997) 'Rate-based scheduling at OPW fueling components', *Production and Inventory Management*, Vol. 38, pp.6–10.
- Shingo, S. (1985) *A Revolution in Manufacturing: The SMED System*, Cambridge, MA: Productivity Press.
- Stuedel, H.J. and Desruelle, P. (1992) *Manufacturing in the Nineties: How to Become a Mean, Lean, World-Class Competitor*, New York: Van Nostrand Reinhold.
- Umble, M. and Srikanth, M.L. (1990) *Synchronous Manufacturing and Integration*, Cincinnati, OH: South-Western Publishing Co.
- Vollmann, T.E., Berry, W.L. and Whybark, D.C. (1997) *Manufacturing Planning and Control Systems*, 4th edition, Homewood, Ill: Irwin, 1 March.
- Wemmerlov, U. (1984) *Capacity Management Techniques*, Falls Church, VA: The American Production and Inventory Control Society.