NjP *Operations Management Education Review 18.*

Promo 1st Version

A Monte Carlo Demonstration of the Implications of Workload Variability in Capacity Planning

Timothy S. Vaughan
University of Wisconsin – Eau Claire, USA

Abstract. This paper addresses the implications of workload variability in capacity planning, and the relationships between nominal capacity, effective capacity, and workload. An Excel-based Monte Carlo model provides a graphical depiction of the implications of workload varying relative to some given capacity. A guided exercise demonstrates that workload variability necessitates some amount of "buffer" or "capacity cushion" relative to average workload. In the face of variable terminology that may impair students understanding of the issue, the demonstration also helps to clarify that the buffer of concern exists for reasons other than possible sources of planned or unplanned process downtime.

Keywords: capacity planning, utilization, efficiency, Monte Carlo simulation.

1. Introduction

Capacity planning and scheduling are inter-related activities that occupy a central role in operations and supply chain management. Scheduling generally represents decisions regarding how to best arrange a given workload, in order to satisfy objectives within the constraints of existing capacity. Scheduling problems may however raise the issue of capacity adjustments, particularly those related to short-term, flexible capacity alternatives.

Capacity planning involves decisions regarding the quantity of various resources that should be in place, in anticipation of future workloads. Short-term planning might address adjustment of flexible capacity alternatives, relative to projected workloads as computed from some firm production schedule or short-term demand forecast. Longer term capacity planning addresses decisions regarding less flexible resources that are not economically varied in the short term. Longer term capacity planning typically addresses a greater degree of uncertainty relative to short term.

As a separate concept, both short-term and long-term capacity planning must address predictable and/or random variability in workload. In the face of such variation, capacity planning and scheduling involves the options of (a)

varying capacity in response to variation in workload and/or (b) leveling the workload to fit within some given capacity. Under option (b), workload leveling may take the form of pulling work into an earlier period (if variations in workload can be predicted in advance), and/or allowing the work to backlog into a later period. In either case, the degree of workload leveling required depends on the actual capacity provided *vis-à-vis* the variable workload pattern. As such, setting capacity levels requires attention to the fundamental trade-off between resource utilization vs. system responsiveness in the face of variable workloads.

This paper is concerned with developing students' understanding of this fundamental issue. The following section presents a simple conceptual model designed to clarify and distinguish between separate concepts within the capacity planning discussion. A discrete-time Monte Carlo simulation is then introduced, supporting a guided in-class exercise that provides a compelling depiction of the implications of workload variability when setting capacity levels. The demonstration clarifies that it is workload as a percentage of *effective* capacity, (as opposed to workload as a percentage of *nominal* capacity) that is of primary interest when deciding on capacity levels in the face of a variable future workload. The simulation model provides a readily accessible visualization of the effects of workload variability, in an introductory course that does not allocate time to developing student understanding of queuing models or more complex discrete-event simulation.

2. A Capacity Planning Framework

Capacity planning requires a clear understanding of two related but very distinct concepts, namely workload vs. capacity. *Capacity*, in short, is some measure of how much work could be done within a specific time period. *Workload*, in contrast, is a measure of how much work has been done or will be done within the same time period. Capacity planning thus requires measures that support meaningful comparison between workload and capacity.

Nominal vs. Effective Capacity

In turn we have two different measures of capacity that typically enter the capacity planning discussion. Most capacity discussions begin, appropriately, with some discussion of the difference between what might be termed *nominal* or theoretical capacity, vs. some reduced measure referred to as *effective* capacity. Although the specifics of the definitions vary across textbooks, nominal capacity generally refers to some output rate that could be realized under ideal conditions over some short period of time. As a simple example, a production process capable of producing one unit every 2 minutes would have

nominal or rated capacity of 30 units per hour. Note that this represents the production rate that is possible *while the process is operating* as designed.

Given such a measure as a starting point, effective capacity essentially accounts for a percentage of time the resources in question might be inoperative over an extended period. As any capacity measure must be independent of workload, this measure must account for loss of productivity for reasons *other than lack of work*. Reid and Sanders (2020, p. 296) for example state "Effective capacity is the maximum output rate that can be sustained under normal conditions. These conditions include realistic work schedules and breaks, regular staff levels, scheduled machine maintenance,..."

Note the above definition recognizes the effect of planned, intentional sources of process downtime, and appears to exclude unplanned or unintentional downtime. In a similar vein, Stevenson (2021, p. 195) states "Actual output cannot exceed effective capacity and is often less because of machine breakdowns, absenteeism, shortages of materials, and quality problems, as well as factors that are outside the control of the operations managers."

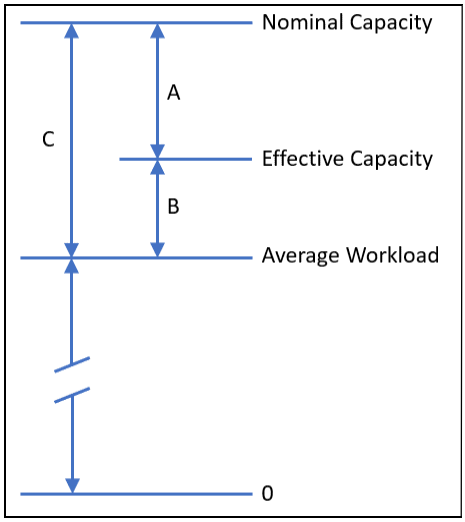
Given these definitions, managers must additionally account for recurring sources of unplanned process downtime when deciding how much capacity to put in place. To the extent that such events can realistically be assumed to occur in the future as they have in the past, failure to account for such events will result in actual capacity less than planned. As a separate issue, managers may implement process improvements designed to reduce unplanned downtime. Unless and until such improvements are realized, realistic capacity planning requires that we provide some allocation for both planned and unplanned process downtime when making capacity decisions.

Recognition of unplanned sources of process downtime could of course be used to refine the definition of effective capacity, or perhaps used to define a separate ("realistic") forward-looking capacity measure. Alternatively, and in the interest of maintaining the focus of this paper, here we will assume that the effects of unplanned downtime have been incorporated as a component of the total workload (capacity consumption) to be placed on the system.

Conceptual Model

Figure 1 provides a simple conceptual model of the relationship between nominal capacity, effective capacity, and average workload. Planned downtimes (and any other components of the difference between nominal capacity vs. effective capacity) are represented as "Gap A" in Figure 1. As noted above, Gap A does not include process idle time due to lack of work.

Figure 1: Conceptual Diagram of Capacity Planning Components



Clearly, nominal capacity needs to be greater than average workload in anticipation of the components of "Gap A" in Figure 1. Moreover, (*and as critically distinct concept*) effective capacity needs to be some amount greater than average workload, thus giving rise to the conceptually distinct "Gap B" in Figure 1. This gap exists to address variability in workload and/or effective capacity. In teaching capacity planning we should not allow students to confuse accounting for planned and/or unplanned process downtime with attending to the conceptually distinct need for "Gap B". This basic concept is demonstrated via the simulation model introduced in the following section.

Utilization

The preceding discussion has intentionally avoided reliance on terms such as "capacity cushion", "utilization" and "efficiency", as these terms are defined and used differently across different textbooks and contexts. Many textbooks (e.g. Heizer, Render, & Munson 2023, p. 305, Stevenson 2021, p. 195, Venkataraman & Pinto 2020, p. 277), use the term "utilization" to refer to average workload as a percentage of *nominal* capacity, and "efficiency" to refer to average workload as a percentage of *effective* capacity. Jacobs and Chase (2023, p. 97) define utilization as capacity used as a percentage of "best operating level", where best operating level is the level at which average cost per unit is minimized. Krajewski and Ritzman (2019, p. 140) define utilization as average output rate as a percentage of "maximum capacity", where maximum capacity is "the greatest level of output that a process can achieve for a longer period, using realistic employee work schedules and the equipment