IESE CIIL
Business School
University of Navarra

PN-464-E
January 2010

Capacity Management in Operations Systems

Alejandro Lago
Philip G. Moscoso
Marc Sachon

In this technical note we discuss how to analyze and manage the capacity of an operations system. The main reason for analyzing capacity is to find the answer to the question: what is the maximum amount of money an operations system can generate (via products or services)? In fact, many of the decisions associated with "operations" have to do mainly with capacity (e.g., number of factories, whether or not to install a new production line, number of employees in a call center). Moreover, capacity decisions affect all departments (R&D, commercial, financial, legal), and not just production and services, as all these departments need to be sized appropriately.

Rightsizing capacities in a company is, therefore, a fundamental aspect of management and, as practice shows, by no means a trivial task.

In this note we discuss the following issues:

- The capacity of a system measures how many items (products, customers, services) the system is able to process (i.e., how much money our operations system will be able to generate) per unit of time.
- The capacity of a system is always determined by the *bottleneck*, i.e., the processor or resource that has the lowest individual capacity. The capacity of an operations system must therefore always be considered as a whole. Capacity decisions (investments, designs, etc.) based on a consideration of only part of the process may well be inefficient or even wrong.
- The capacity of a process, and therefore the bottleneck, is affected by: 1) the mix of items, 2) the setup time and production batches, 3) the assignment of tasks to processors and 4) quality problems. Generally speaking, managers can influence all these aspects and discover "hidden sources of capacity," reducing the need for additional investments.

This technical note was prepared by Professors Alejandro Lago, Philip G. Moscoso and Marc Sachon. January 2010.
This case was written with the support of the CIIL (Centro Internacional de Investigación Logística), IESE.

Last edited: 3/6/20

- As soon as one bottleneck is removed, a new bottleneck will appear in some other part of the system or process. Capacity improvement is therefore an ongoing process of eliminating bottlenecks, one after another.

1. Throughput and Capacity

Before we define the concepts of throughput and capacity, let us briefly review the elements that make up an operations system[1]:

- *Items* are all the elements that are processed and flow through the system. They can be customers, products or parts, or simply information.
- *Activities* are the basic units through which items are transformed or processed. Each item is associated with certain activities in the process through which value is added, using processors. The times taken by the activities are what we call "consumptions" of processor time.
- *Processors* or *resources*[2] (labor, machines, plant, etc.) are the elements of the operations system that carry out the activities on the items. The processor working times available for processing items are what we call "availabilities."

A good analogy for an operations system is that of the "operations pipeline," through which the items to be produced flow.

Throughput is (colloquially) the rate of production of our process. Formally, we can define it as the quantity of items the system processes per unit of time. In the pipeline analogy, throughput may be visualized as the quantity of items that come out of the end of the operations pipeline in a given period. In a factory, for example, this would be the number of finished units transferred to the finished goods store per day; in a call center, the number of calls answered per hour.

Besides actual throughput, it is also useful to know the maximum throughput or capacity of an operations system. In other words, the capacity is the maximum number of items a system can process per unit of time. Traditionally, capacity is expressed in items per unit of time. For example, an operative may assemble five chairs per hour, a call center may answer one hundred calls per hour or a hospital may attend to 500 customers per day[3].

Throughput and capacity seem simple to understand, but in practice grasping their details and accurately calculating their value for an operations system as a whole is more complicated. This is because, as explained below, the *capacity* of a process depends on:

[1] Readers interested in knowing more about the fundamental concepts of operations are advised to read technical note PN-458-E, "Fundamental Concepts and Parameters of Operations Management."

[2] Sometimes they are referred to as "resources," while in services they are also known as "servers" (be careful not to refer to items that are to be processed, e.g., energy, as resources).

[3] There is a problem with this way of understanding capacity: if capacity is defined as items per unit of time, the capacity of a system will depend on the type of item processed, i.e., if the type of item changes, then so will capacity. For example, if the calls received by a call center become more complex and take longer to handle on average, the call center's capacity will decrease. Some authors (e.g., see Muñoz-Seca, B. and J. Riverola, *Del buen pensar y mejor hacer*, McGraw-Hill, 2003, chapter 5) propose, as an alternative, that the "absolute" capacity of a processor be defined as the processor's available working time. This is what we call availability. In the case of a call center, for example, availability would be the number of "hours of operators' working time available per day or week," i.e., the time available to process calls of any kind.

- How activities are assigned to processors, what processors we actually have at our disposal and the processors' availability of working time.
- The type of items processed and how they are processed (batch, series, etc.).

To better understand what capacity is and how it is calculated, we shall start by looking at the simplest cases (sequential systems processing just one type of item), before moving on to more complex examples.

2. Capacity in Single Product Systems and the Bottleneck

The fundamental idea for understanding the capacity of an operations system is as follows: *each processor has a certain capacity when it operates in isolation and processes a certain type of item; however, the capacity of the system as a whole depends on all the processors in the system*.

Our first example will show how the individual capacities of the various processors interact to determine the capacity of the system as a whole.

Example 1. Procer Furniture 1 (a sequential system). A woodworking company manufactures a single type of chair. For simplicity, let us assume that the manufacturing consists of three basic activities, each carried out by a different operative (processor): 1) cutting the pieces of wood, 2) shaping and assembling the pieces to make the chair and lastly, 3) varnishing the chairs. In this simple case, we have a sequential process: the items visit each processor only once. The time it takes to carry out the activities (i.e., the *consumption* of processor time for each activity) is, respectively, 12, 30 and 24 minutes per item and the amount of time each operative works (i.e., the processor's *availability*) is eight hours per day.

In this example it is easy to calculate how much the system can produce, without having to use any special method. If the first worker (cutting) works the maximum number of hours, he will process 60/12 = 5 chairs/hour (40 chairs/day); therefore, the second (shaping and assembling) will process 2 chairs/hour (16 chairs/day) and the third (varnishing), 2.5 chairs per hour (20 chairs/day).

In more complex systems the calculation becomes more complicated. Fortunately, however, there is a general, systematic method for doing it, known as *capacity analysis*. Capacity analysis focuses on *how much* the system can process in theory, without worrying about *how* the items actually flow through the system over time. In other words, it tells us the average capacity over the long run, ignoring any temporary fluctuations that may occur. First, we will calculate *the capacity of each processor* and then we will see how the capacity of the system as a whole depends on the processors' individual capacities.

The capacity of a processor. A processor reaches capacity (or maximum throughput) if it processes items continuously, i.e., if its entire *availability* is used productively and no time is wasted waiting for items to arrive or for repairs to be made. The capacity of a processor will therefore be given by the ratio of processor *availability* (in working time) to the total consumption of that availability by items to be processed. The capacity of a processor is therefore calculated in three basic steps (Figure 1):