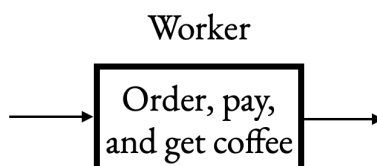


# Measuring process performance

## Class discussion

RADU: I see many of you have your coffee again. That's good. I want to return to this example.  
Can I simplify the story a bit? [*Draws a picture.*]



MING: Why so simple? You just got through teaching us about triangles and diamonds, after all.

RADU: Excellent question. I want to “zoom out” a bit on the various steps and just think about these actions together as like one action. You know, we have been doing that a lot. In class 2, we had this action “Resolve” in the OneWay discussion. Of course, that involved many little steps that we ignored because we were not interested in that level of detail.

MING: But how do we know what the right level of detail is?

RADU: It depends on your inquiry. So, can I say where we are going to inquire about here? [*Ming nods.*] Our question is: How should we *measure* the performance of this process?

[*Students take some time to realize Radu is asking them to come up with measures. Afterwards, they look around for someone else to venture an idea. Eventually, Haoyu gains the courage to talk.*]

HAOYU: Well, let me speak from my perspective. I would want to know how long I need to wait to get my tea.

RADU: Yes, of course, tea. Haoyu, when do you start the clock on that time? Like when you wake up in the morning? From the time you wake up until you get your tea?

HAOYU: That would be silly. I would think I start the clock when I walk into the tea shop. From then until I get my tea.

RADU: And why is that?

HAOYU: That is the time I need to consider adding to my trip to arrive at class. If I decide to get tea, that is how much longer my trip will be to class.

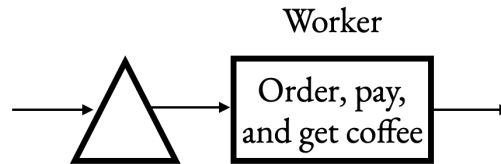
RADU: Makes perfect sense to me. Can I ask, is this time different from the elapsed time from when you *order* your tea to the time you get it?

HAOYU: Yes, it's different. Sometimes I need to wait before I can order.

RADU: So I am realizing there is something is not quite right about our picture, right Farah?

FARAH: I was going to say that earlier. It is missing a “wait” action before you can order.

RADU: Like this? [*Draws picture and Farah nods*]



RADU: [To Camila.] Why didn't I include "wait" in this square box here Camila? Why would I separate them? I made a big deal about putting all the other actions in the single box, but why not "wait"?

CAMILA: I feel the "wait" action is special. It should be separated.

RADU: Special, how?

[Camila shuffles around in her chair, reconsidering her answer.]

BINESI: [Speaks after waiting for Camila for a moment.] I think it's because you don't always have to wait. You only wait when the server is busy.

RADU: Excellent. That's what you had in mind, Camila?

CAMILA: Yes, but I also think this "wait" activity is strange in the sense that it does not actually change the input at all. I think of activities as actions that change the inputs.

RADU: Hmm, I didn't think about that. I guess it really depends on what we mean by "change", does it not? The input is certainly "older". [The class laughs.] Maybe in some contexts that matters, like if it is a depreciating asset.

CAMILA: Ah yes, true.

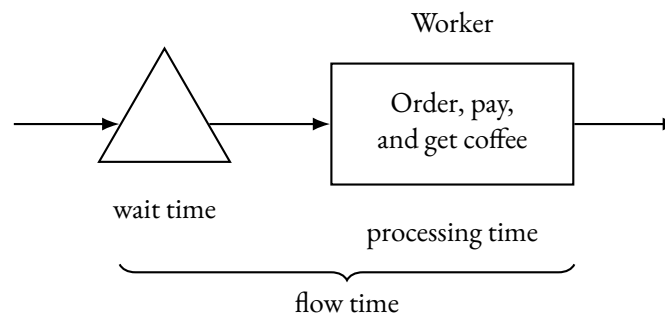
RADU: But Camila's instincts are right here. This wait "activity" is indeed special. We will learn more about this later. [Turns to the whole class.] So, we have this wait time and then the time it takes from ordering to receiving your tea. You wanted to include both, right Haoyu?

HAOYU: Certainly, that is the total time that gets added to my walk to class, so that is the time I worry about.

RADU: As you should. So let me introduce a couple of terms here. The total time it takes from the time you enter the "operation" until you leave is your *flow time*. It consists of two parts, waiting time and processing time. [Writes the following on the board:]

$$\text{flow time} = \text{waiting time} + \text{processing time}$$

I like to draw that in the figure like this:



And how would you measure the flow time Haoyu?

HAOYU: With a watch? Start the watch when I enter the line and stop the watch when I finally get my tea.

RADU: Very good. flow time is measured in a unit of time, like seconds or minutes. Say it takes 5 minutes for Haoyu to get his tea. Good. *[Looking to the rest of the class.]* Any other measures?

FARAH: Maybe it's the same, but I would be interested in how efficient the workers are.

RADU: Ah yes, this word "efficient". I noticed you like this word Farah. The challenge I have with that word is that people use it in all kinds of ways. I tend to avoid the word for that reason, we will introduce more precise words as we go along here. But before we do that, how would you measure the efficiency of the workers Farah, as you understand the term?

FARAH: I guess I would measure how long it takes them to serve a customer.

RADU: Ah, so efficiency is a time? Like measured in seconds with a stopwatch?

FARAH: Yes, I think so. And the shorter is this time, the more efficient is the worker.

RADU: How is different than flow time?

FARAH: Right, that is why I wondered if they were the same thing.

BINESI: *[Cutting in after a brief pause.]* Well, I actually think they are different.

RADU: *[Turning to Binesi.]* How so?

BINESI: Well, if I wanted to measure the performance of a worker, I don't think I should include the waiting time. It is not really the fault of the worker that there is waiting. If a bunch of people arrive in a short period of time, it doesn't mean that the workers did not work hard. So it seems this "efficiency" *[Makes air quotes when saying "efficiency"]* is different than flow time, which includes waiting time.

RADU: Very astute point Binesi. We should maybe look more to the "processing time" to understand how effective our workers are.

MING: I am not sure I agree. Binesi said that the waiting time is not the fault of the worker, but if the worker were faster, wouldn't there be shorter of a line?

*[Ming looks to Radu who turns to Binesi, gesturing her to answer.]*

BINESI: Ah, I see your point. I was thinking that wait comes from the customers arriving, which is not the fault of the worker. But I think maybe I was only thinking about half of it. The line is also there if the worker is too slow.

RADU: Yes, there are differing contributing factors to the wait. Can we avoid this complexity for now? I wanted to tackle it later. Instead, can I give you a related concept? *[All students nod.]* I want to define the *throughput rate* of the worker. It measures how many jobs the worker completes in a certain period of time. Are you comfortable with that? *[All nod.]* So let me ask this. What is the throughput rate of a worker that finishes 4 jobs every 15 minutes? *[Writes on the board:]*

4 jobs in 15 minutes  
throughput rate = ?

FARAH: [*Excitedly.*] Yes! Finally a calculation! A rate should have a unit something like “job per unit time”. So in this case, the rate is 4 jobs every 15 minute or 16 jobs per hour.

[*Radu writes on the board without saying anything.*]

$$\begin{array}{l} \text{4 jobs in 15 minutes} \\ \text{throughput rate} = \frac{4 \text{ jobs}}{15 \text{ minutes}} = 16 \frac{\text{jobs}}{\text{hr}} \end{array}$$

RADU: Like this?

FARAH: Yes, precisely.

RADU: So I just want to make sure everything is comfortable with these “units”. The number 16 has the “unit” *jobs per hour*. Why didn’t we stick with the unit *jobs per minute*?

BINESI: Because that would be too easy?

[*Everyone laughs*]

RADU: Farah, why did you state it in the jobs per hour?

FARAH: I just felt 4/15 jobs per minute was just messier. I like whole numbers.

RADU: It is a matter of taste and convenience. Regardless, you should be really comfortable with changing units, like from minutes to hours. I will give some practice on this in the email I send you later today.

[*Some of the students looked relieved, others looked bored.*]

Okay, so I want one last measure of performance. It is neither a time nor a rate. Suppose I just took a picture of the coffee shop. I don’t have a stopwatch at all. What could this picture tell me? What would we hope to see in the picture?

MING: Wouldn’t we want the coffee shop to be full of customers?

RADU: Who is *we* here? “We” as in us as coffee buyers or “we” as in the managers?

MING: I would personally prefer being the manager, so let’s go with that. The more customers in the store, the more money you are making.

CAMILA: But even if we were the manager, wouldn’t it depend on whether the people were *waiting in line* or *enjoying coffee*?

MING: Ah, right. More people waiting in line might not be so good, people might start to complain about the service. More people enjoying their coffee after being served would be better.

RADU: Actually, it can get even more complicated than this, but let’s not digress too much. Can I simply interject here with some terminology? What we can learn from the picture is something called *inventory*. I know that sounds a little strange when you first hear it because we tend to associate inventory with boxes of goods sitting on a shelf. So let me define it more clearly: inventory refers to jobs that have come into our operation but have yet to be completed. [*Turns to the board to write.*]

inventory = number of jobs that arrived but not yet completed.

In other words, they are inputs that have not been fully transformed yet into outputs. [*Turning to Camila*] So Camila, you talked about two types of customers in the picture, those waiting to be served and those enjoying their coffee. Are both of them inventory?

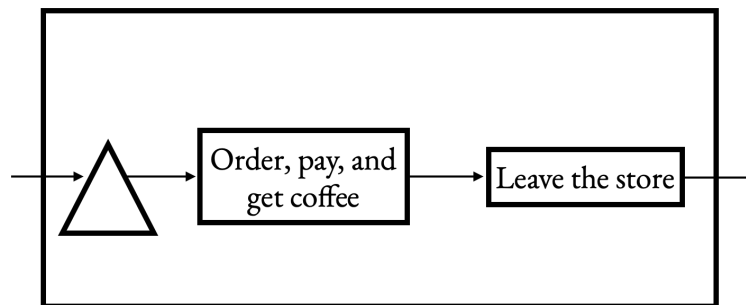
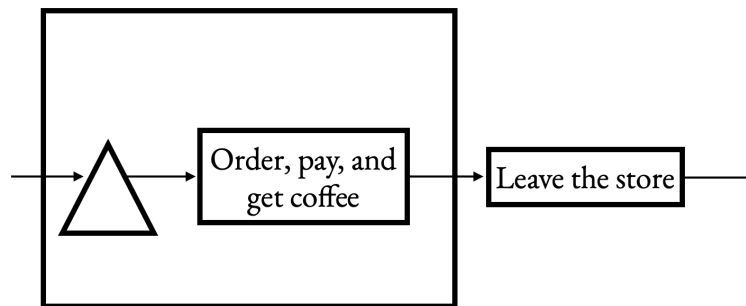
CAMILA: I guess not. The ones who are enjoying their coffee are not inventory since they have their coffee already; they have already been served.

HAOYU: Yes, they have been served, but they are still in the shop, so is the job really complete?

RADU: Very good, this is a subtle distinction. In our process flow we drew on the board, we did not distinguish between them having coffee in their hand and them leaving the shop. But depending on how we define a job being completed—have their coffee versus left the coffee shop—we would include those customers—or not—as inventory.

BINESI: You are saying there is different types of inventory?

RADU: In a way, yes. Let me redraw the picture of our process in light of this. [*Draws two new figures on the board.*]



I am using this larger “box” to represent our reference for the operation. In the first figure, we consider customers who are served as being the final output, in which case inventory is only those customers who are in the store who have yet to get their coffee. In the second figure, we have a larger reference. Now, all customers who are in the store, whether they have their coffee yet or not, are considered inventory. What we count as inventory depends on our point of view.

MING: [*Said in a disappointed and grumpy tone.*] It feels like so many things depend on your point of view. Feels very shaky.

RADU: “Shaky” is an interesting word. What do you like to see, what is less shaky to you, Ming?

MING: I thought this is more of an analytical class. So I was expecting to see things to be defined in a more explicit and fixed way. Now it feels there are too many moving parts in the picture.

RADU: Why is too many moving parts a bad thing, Ming?

MING: [*Ming thought for a second and said*] I don't know if it's a bad thing, it just blurs the picture in my head.

RADU: Ah, I see your confusion. If it confused you, you can just choose one way that you feel comfortable and build your understanding on top of that. This is okay for this class. Let me say another word on this point. The reason I didn't define these concepts in a "fixed" way is because I don't want to give you the impression that there's only one way to describe the system. In reality, depending on different purposes, we care about different parts of the process. If a tool can only cover a fixed way of describing operations, actually it's quite rigid and weak. One of the desiderata for a good methodology is flexibility, i.e., it covers a wide class of cases. Does this make you feel happier about this shaky part, Ming?

MING: Ah, I see. That's a good point.

RADU: [*Turning to Farah.*] Now, Farah, what should the unit of measurement for inventory be?

FARAH: Whatever the unit of the jobs are. In this case, "customers"?

RADU: Very good! We have now described three key performance measures: flow time, throughput rate, and inventory. Each of them has a different unit of measurement. They sound pretty straightforward (hopefully), but honestly, there are quite a few subtleties here. It took me quite a few years to get my head around them. Let's end things here and take this up more in future classes, shall we?

MING: Do we have a choice?

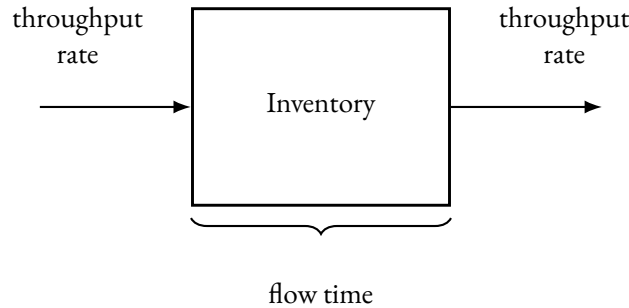
[*All laugh, including Radu. Radu erases the whiteboard as his students slowly leave the classroom with a mixture of relief and boredom.*]

## Email

Dear students,

I didn't cover as much ground as I originally planned today, but that's okay. No need to rush through this foundational material. In fact, I think showing some of these ideas in written form first could be helpful. Then we look at examples in later classes to make sure we are all on the same page. So watch out, because today's email is going to be "theoretical" and involve numbers.

There are many ways to measure different aspects of a process flow. Our focus here is on a few basic metrics to measure the performance of a process flow. The following figure helps visualize them.



I will introduce some terminology to flesh out our description of this figure. A *job* is a unit of input that is destined to be converted to a unit of output. The “box” captures all of the activities that are needed to convert a job from an input to an output. For the sake of this discussion, I will include “wait” activities in this big box. I originally wanted to do this in class today, but you all convinced me that maybe “waiting” should not be included. I will go with my original plan now and put all of the activities “in the box”.

We also got into a bit of trouble defining the “box” during the class today. We had the issue of whether you left the box after you got your coffee or when you left the coffee shop entirely. This is a commonplace confusion. Defining exactly what is in the box and outside the box is important if we want to be precise. There are typically many possibilities.

I will interchangeably refer to the “box” as “the system” or “the operation”. I can’t help myself in using these words interchangeably, so this is a word of warning.

Let me now try to define more precisely the three operational metrics we discussed in today’s class.

**Throughput rate.** This is the average number of jobs that “come and go” from the system in a given period of time, say, one hour. By “come”, we mean arrives as input to the box, and by “go”, we mean leave the box as output.

Let’s say  $N$  jobs “come and go” in a time period that starts at a time we call 0 and ends at a time I will denote by  $\tau$ . Think of  $\tau$  as like one hour or something like this. The throughput rate  $R$  in this time period is  $N/\tau$ . It’s like the “speed” at which the process finishes jobs. For example, suppose I was interested in measuring the throughput rate of a machine in a factory. I could sit there for an hour ( $\tau = 1$  hour) and track how many jobs come to the machine and leave processed by the machine (let’s say 3 job arrive and leave). Then the throughput rate of that machine for the hour was  $R = N/\tau = 3 \text{ jobs}/1 \text{ hour} = 3 \text{ jobs per hour}$ .

**Flow time.** Suppose we start a stopwatch every time a job arrives as input. That stopwatch stops when the job leaves the system as an output. The elapsed time is the flow time of that job.

Of course, not all jobs need to have the same flow time. Suppose, as we had in our throughput rate calculation, that  $N$  jobs “come and go” in an elapsed time of  $\tau$  time units. Each of these  $N$  jobs could have a different flow time, say  $T_1, T_2, \dots, T_N$  are the  $N$  different flow times of these  $N$  different jobs. We could then compute the *average* flow time of these  $N$  jobs in this time period:

$$T := \frac{\sum_{i=1}^N T_i}{N} = \frac{T_1 + \dots + T_N}{N}.$$

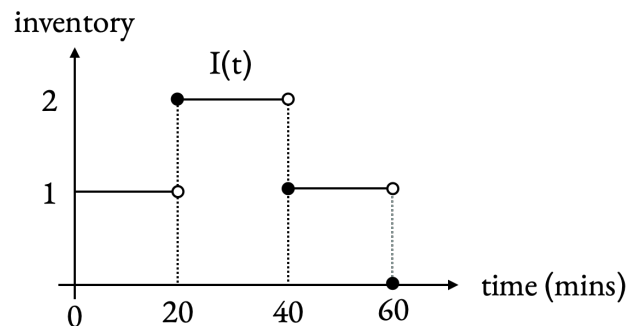
We denote this average flow time  $T$ , with no subscript.

Let's return to our machine example to make this more concrete. We saw three jobs come and go in one hour ( $N = 3$  jobs,  $\tau = 1$  hour,  $R = 3$  jobs per hour). Let's say it took 23 minutes for the machine to finish the first job, 17 minutes to finish the second job, and 18 minutes to finish the third job. That is,  $T_1 = 23$  minutes,  $T_2 = 17$  minutes,  $T_3 = 18$  minutes, and  $T = \frac{23+17+18}{3} = 19\frac{1}{3}$  minutes.

**Inventory.** I think throughput rate and flow time are relatively straightforward concepts that don't require so much explanation. Your intuition about them will probably suffice to reason about them. The third metric "inventory" is a lot more subtle.

Let's think about inventory in this time period from time 0 to time  $\tau$  have been discussing. Let  $t$  be an arbitrary time between 0 and  $\tau$ . The *inventory* at time  $t$  is the number of jobs that came before time  $t$  but did not leave by time  $t$ . As we discussed in class, if you took a picture of the box at time  $t$ , the inventory at time  $t$  is a count of the number of jobs in the box at time  $t$ .

Notice that inventory can change from moment to moment. Jobs can arrive at any time and leave at any time. So, in fact, the level of inventory is a *function* of time. This inventory function is one of the reasons the GMAT has functions as part of its curriculum! Another name for function is "mapping". It's like it takes an input and maps to an output. In this situation, the inventory function takes time as input and maps the inventory value for each time  $t$ . We typically visualize functions by their graph in two-dimensional space. The graph of the inventory function has a special name for some reason. It's typically called the *inventory buildup diagram*.



The graph of the function  $I(t)$ , which gives the inventory level as a function of  $t$ , looks something like in the above figure. I made this example really simple to illustrate the idea clearly. In next class, we will see a "messy" one.

In the situation depicted in this graph, at time 0, there is already one job in the system that arrived at time 0. The end of the time period is  $\tau = 60$  minutes, as shown in the picture. At time  $\tau/3 = 20$  minutes, a second job arrives, taking inventory to 2. This is seen in the "upward" jump of the graph at time 20. At time  $2\tau/3 = 40$  minutes one of the two jobs in the system leaves, bringing the inventory back down to 1. Then, at time  $\tau$ , the second job leaves and the inventory goes to 0. In other words,

another way to represent the function is the following:

$$I(t) = \begin{cases} 1, & \text{for } t \geq 0 \text{ and } t < 20, \\ 2, & \text{for } t \geq 20 \text{ and } t < 40, \\ 1, & \text{for } t \geq 40 \text{ and } t \leq 60, \\ 0, & \text{for } t = 60. \end{cases}$$

That is what is happening to inventory throughout this time period, but it's pretty complicated, frankly. No one wants to think about functions like this for too long. What we often want is a *single number* that captures the inventory level *on average*. But what does average exactly mean here? Some of the time the inventory is 1 and some of the time the inventory is 2. So the average should be somewhere between 1 and 2. But should it be closer to 1 or closer to 2?

Well, notice the 2/3 of the time (from 0 to 20 and 40 to 60) the inventory level is 1, whereas it is 2 only 1/3 of the time (from 20 to 40).<sup>1</sup> So the average should be closer to 1 than 2. How much closer? I hope that your basic mathematical sense suggests to you that the average should be a *weighted* average:

$$\frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 2 = 1\frac{1}{3}.$$

since 2/3 of the time inventory is 1 (and so we multiply 2/3 and 1 together) and 1/3 of the time inventory is 2 (and so we multiply 1/3 by 2) and we add these two scenarios together. If this does not make sense to you, send me a follow-up email and we can discuss more.

We let  $I$  denote the *average inventory* in this time period from 0 to  $\tau$ , in this case,  $I = 1\frac{1}{3}$ . We will discuss how to compute  $I$  more generally next class.

**Why bother?** For those of you who don't like math too much, maybe this email has been a little heavy for you. If you found throughput rate easier to understand than average flow time, and average flow time easier to understand than inventory, then you are like everyone else and your mathematical sense is correct. I wish I could say the notion of average inventory was easy to grasp and calculate, but honestly, it's not. I am still somewhat confused about it myself.

So why are we dealing with things like average inventory that confuse us? Isn't it better to avoid confusing things unless there is a really big payoff to paying attention to them? Let me give you a pitch for why they are important to pay attention to. Operations management is about looking at things we have seen a million times differently. One way to look at it "differently" is to look at things with more granularity than we might otherwise—we make a big deal about things others see as small or don't see at all.

We do this because operations management is about *improving* the way we do things, *even if* we have already done them a million times another way. It is hard to convince people to change how they are doing something, especially when they are in a well-established pattern. The language of process flows provides a vocabulary to describe what's being done, but description often does not suffice for making change. We also need a way to compare two *different* processes to achieve the same (or similar) goals. We want to say that one way is *better* than another. This is how we improve.

---

<sup>1</sup>It is 0 for just an instant at time  $\tau = 60$  so that this does not contribute to the average.

These three metrics—throughput, flow time, and inventory—are ways we can *compare* different processes. They are guiding metrics for *improvement*. Generally speaking, more throughput is better, is it not? A café that can serve more customers in a given period of time makes more revenue. Similarly, a smaller flow time would be good. Customers don’t need to wait in line so long for their coffee. And I think we can all somewhat agree that less inventory sitting around is better than more. After all, incomplete jobs lying around are like costs not converted into benefits. These three operational metrics will form our core language of operational improvement.

That’s all for now. Sorry for the long and detailed email!  
Warmest, Radu.

## Follow up email: Conversion of units

Binesi reminded me in a follow up email that I had promised a short tutorial on converting units. I kept putting this off, but it is genuinely useful background for what is coming in our next few chapters, so let me make good on the promise now.

You have all been converting units casually throughout our discussions—4 jobs in 15 minutes becomes 16 jobs per hour, and so on. This feels intuitive enough that it hardly seems worth dwelling on. But there is a subtlety lurking here that is worth bringing into the open, because it will matter later.

Consider the following two quantities:

$$\frac{1}{2} \quad \text{and} \quad \frac{2}{4}.$$

As numbers, these are equal. We all learned this early in school, and it is so deeply internalized that it barely registers as a fact at all. And yet, in another sense,  $\frac{1}{2}$  and  $\frac{2}{4}$  are *not* the same thing. One is written with a numerator of 1 and a denominator of 2. The other is written with a numerator of 2 and a denominator of 4. If I handed you a pizza cut into two slices and asked for  $\frac{1}{2}$ , I would give you one slice. If the same pizza were cut into four slices and you asked for  $\frac{2}{4}$ , I would give you two slices. These are different physical actions, involving different numbers of slices, that result in the same amount of pizza. In other words, handling half of a pizza to eat is different than handling two quarter slices. (Anyone without a knife nearby certainly knows the difference!) But ultimately you have eaten the same amount of pizza when are you are done,

So which is it? Are  $\frac{1}{2}$  and  $\frac{2}{4}$  the same, or are they different?

I think the honest answer is: both, depending on what you mean by “the same.” They are different as *expressions*—different numerators, different denominators, potentially different physical procedures for arriving at them. But they are the same as *values*—the same point on the number line, the same proportion of the pizza, the same amount of stuff.

This distinction between “different as expressions” and “same as values” is, I think, exactly what is going on when we convert units. When I write

$$16 \frac{\text{jobs}}{\text{hour}} \quad \text{and} \quad \frac{4 \text{ jobs}}{15 \text{ minutes}},$$

these are different expressions. One involves the number 16 and the unit “jobs” and “hour.” The other involves the numbers 4 and 15 and the units “jobs” and “minutes”. And yet they describe the exact same

rate, the exact same underlying reality of how quickly jobs are being completed. Converting between them is not changing anything fundamental about the world, just how we look at it. It is changing how we *describe* the same thing.

Why does any of this matter for us? Because in the weeks ahead, you will frequently need to take quantities that arrive in one set of units and convert them into another set of units before you can meaningfully compare or combine them. If you think of unit conversion as merely a mechanical chore that multiply here and divide there, you will make careless errors, especially when several conversions are happening at once. But if you think of it as finding a different (and more useful) name for the same underlying quantity, you will have a much firmer footing. The quantity itself never changes. Only its name does.

A practical tip, which I will return to: when converting units, write out the units explicitly at every step and treat them algebraically, the same way you would treat variables. If a unit appears in both a numerator and a denominator, it cancels, just as  $x$  cancels in  $\frac{x \cdot y}{x}$ . We saw a hint of this already when computing 16 jobs per hour from  $\frac{4 \text{ jobs}}{15 \text{ minutes}}$ . Let me slow that down, just to make the point crystal clear:

$$\frac{4 \text{ jobs}}{15 \text{ minutes}} = \frac{4 \text{ jobs}}{15 \cancel{\text{minutes}}} \cdot \frac{60 \cancel{\text{minutes}}}{1 \text{ hour}} = 16 \frac{\text{jobs}}{\text{hour}} = 16 \text{ jobs per hour}.$$

## Follow up email: Weight average and area underneath a function

Camila asked me after class for more detail on how we arrived at the average inventory of  $1\frac{1}{3}$  in my earlier email. She made an observation that I think is worth sharing with everyone, because it connects what we are doing to something many of you may have already encountered elsewhere.

Camila noticed that the calculation

$$\frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 2 = 1\frac{1}{3}$$

looked a lot like the calculation for the *expectation* of a discrete random variable. If you have taken a probability course, you may recall that if a random variable  $X$  takes on values  $x_1, x_2, \dots, x_k$  with probabilities  $p_1, p_2, \dots, p_k$  (where these probabilities sum to 1), then the expectation of  $X$  is

$$\mathbb{E}[X] = \sum_{j=1}^k p_j x_j = p_1 x_1 + p_2 x_2 + \dots + p_k x_k.$$

Camila's instinct is exactly right. In our example, the inventory  $I(t)$  is a quantity that, at any given moment  $t$ , takes on some value. In our setting, this value is always a whole number, since it is a count of the number of people in the store. That makes inventory a *discrete* quantity, just like a discrete random variable.

Recall that in our example,  $I(t)$  takes on the value 1 for  $\frac{2}{3}$  of the time period, and the value 2 for  $\frac{1}{3}$  of the time period. If we treat these fractions of time as if they were probabilities, the “probability” that a randomly chosen moment  $t$  has inventory 1 is  $\frac{2}{3}$ , and the “probability” it has inventory 2 is  $\frac{1}{3}$ . Naturally, the average inventory calculation is precisely the expectation formula:

$$I = \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 2 = 1\frac{1}{3}.$$

So Camila's connection is a real one, and a useful one to keep in mind. In essence, this connection is a natural link to what weighted average is about. In the expected value example, we are trying to find one number to represent the value of a random variable. In the inventory function example, we are trying to find one number to represent the value of inventory value. In some sense, we are both looking to find one number to represent something that's changing, and that's when weighted average comes in.

I want to offer you another way of seeing the same calculation, one that I think is, in some sense, even more fundamental and will generalize more easily to the “messier” examples we will see in future classes. This is to think of  $I(t)$  simply as a *function of time*, and to ask: what is the *average value* of this function over the period from 0 to  $\tau$ ?

Those of you who have taken calculus may recall that the average value of a function  $f$  over an interval from 0 to  $\tau$  is defined as

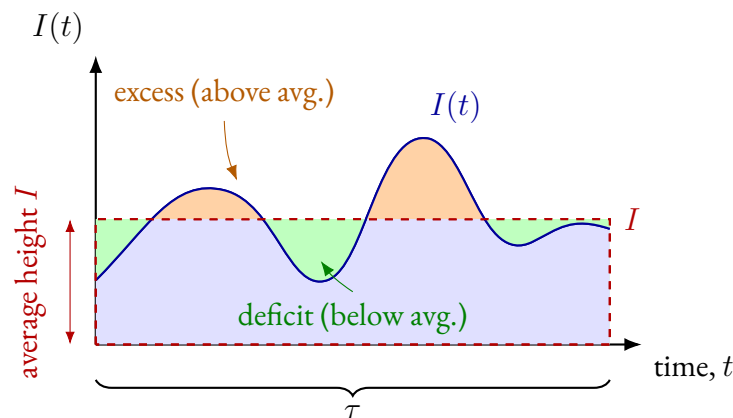
$$\text{average value of } f = \frac{\text{area under the graph of } f \text{ from } 0 \text{ to } \tau}{\tau}.$$

This formula is really just a generalization of something much more basic. Think about the simplest possible shape whose area you know: a rectangle. The area of a rectangle is base times height,  $A = b \cdot h$ . If you know the area  $A$  and the base  $b$ , you can recover the height by dividing:  $h = A/b$ . In other words, the height of a rectangle is its area divided by its base.

Now, here is the key idea. If you have some wiggly, irregular curve, like our inventory function  $I(t)$  which jumps up and down over time, you can ask: what single height, if I drew a rectangle with that height over the same base, would give me a rectangle with the *same area* as the area under my wiggly curve? That height is, by definition, the average height of the curve. And by the same logic as before,

$$\text{average height} = \frac{\text{area under the curve}}{\text{base}}.$$

This is exactly the formula for the average value of a function I wrote above, with “base” being the length of the time interval,  $\tau$ . We can see this idea in the following figure:



So when we compute average inventory, we are doing nothing more exotic than this: we are finding the area under the inventory buildup diagram, and dividing by the total length of time,  $\tau$ . The result

is the height of the rectangle that has the same area as the curve, the “average height” of our inventory function. This is precisely the average inventory.

You can check that this gives the same answer as Camila’s probability-flavored calculation. The area under our step function  $I(t)$ , from 0 to  $\tau = 60$ , can be broken into two rectangular pieces: one of height 1 and width 40 (the times from 0 to 20, plus 40 to 60, both at height 1), and one of height 2 and width 20 (the time from 20 to 40, at height 2). The total area is

$$(1 \times 40) + (2 \times 20) = 40 + 40 = 80.$$

Dividing by  $\tau = 60$  gives

$$\text{average inventory} = \frac{80}{60} = \frac{4}{3} = 1\frac{1}{3},$$

which matches what we found before.

I find it useful to keep both perspectives in mind. Camila’s probability perspective is helpful because it connects to ideas you may already know well from statistics, and it generalizes nicely to settings where inventory takes on many different values with many different “frequencies.” The area perspective is helpful because it connects to a single, simple geometric idea: the area divided by base equals average height. This idea applies no matter how complicated or irregular the inventory function becomes. In the weeks ahead, when our inventory buildup diagrams get considerably messy, it is this area perspective that will do most of the work.

## Extensions

*[Posted as homework on the course’s website for this class session.]*

1. Consider two rates: 120 jobs per hour and 2 jobs per minute. Are these the same rate? Why or why not? If you claim that they are the same, how do you reconcile the fact that the numbers 120 and 2 are so different? Are you somehow suggesting  $120 = 2$ ?
2. What was the flow time of the activity of you reading this question? Describe how you calculated it. Anything that made it challenging?
3. Suppose you run a library. First, you are thinking about your process from a perspective where the unit of inventory is “book”. What makes a job “come” and “go” from this perspective? What would the associated flow time measure? Second, you change to thinking about the process as one where the unit of inventory is “patron”. Does it change your notion of “job”? What would flow time measure now?
4. Suppose  $N$  jobs “came and went” from time 0 to time  $\tau$ , and they had flow times  $T_1, T_2, \dots, T_N$ . Let  $T_{\max} = \max\{T_1, T_2, \dots, T_N\}$  denote the *maximum* flow time of those  $N$  jobs. Why is the average flow time likely to be different than the maximum flow time? As a manager, which concerns you more: a high average flow time or a high maximum flow time? Would your concern change in different scenarios?

5. Suppose you arrive at an airport security checkpoint and find a line of 40 people. Every 30 seconds, someone is processed and leaves the checkpoint. Roughly how long will it take you to reach the front of the line? What assumptions are you making? What would you guess is your total flow time?
6. In Zen Buddhism there are two major schools of thought for reaching enlightenment. The *sudden school* holds that enlightenment is realized instantaneously once delusion falls away, while the *gradualist school* understands enlightenment as something cultivated through continuous practice over time. Recall that flow time is often decomposed into processing time (value-adding work) and waiting time (delays, queues, and idle periods). How might a sudden versus a gradualist Zen perspective lead to different interpretations of what flow time “is”, particularly with respect to the distinction between processing and waiting?
7. Binesi and Ming disagreed about whether waiting time is “the worker’s fault.” Binesi first said the wait comes from customers arriving; Ming countered that a slower worker also lengthens the line. Take a side. Then describe a concrete situation where the same amount of waiting should be blamed mostly on *arrivals*, and a second situation where it should be blamed mostly on the *worker*. What observable difference between the two situations lets you tell them apart?
8. Convert each of the following, writing out every unit and cancelling explicitly in the style of Radu’s follow-up email: (a) a barista who makes 3 drinks every 4 minutes, into drinks per hour; (b) a help desk that closes 90 tickets per 8-hour day, into tickets per hour; (c) a rate of 0.05 jobs per second, into jobs per hour. For one of the three, state the same rate in a way that uses a deliberately ugly, non-whole number—and explain why Farah would dislike it even though it is equally correct.
9. Radu drew two different “boxes” for the coffee shop: one that ends when a customer gets their drink, and one that ends when they leave the shop. Pick a real operation you know and draw two legitimate boxes for it that disagree about where the process ends. For each box, state what inventory counts, what the flow time measures, and what the throughput rate measures. Then name one management decision that would come out *differently* depending on which box you had chosen.
10. Camila objected that “waiting” is a strange activity because it does not change the input. Radu partly conceded, then hinted that waiting sometimes *does* change the input—“the input is certainly older.” Give two examples where the mere passage of waiting time genuinely changes the input in a way a manager should care about, and one example where it truly changes nothing. What distinguishes the two kinds of case?
11. The email computes an average flow time of  $19\frac{1}{3}$  minutes by averaging over the three jobs, and separately computes an average inventory of  $1\frac{1}{3}$  by averaging over time. Notice that these are two different *kinds* of average—one is taken “per job,” the other “per unit of time.” Invent a small example of your own (three or four jobs, with arrival and departure times you choose) and compute both averages for it. Then explain, in your own words, why averaging over jobs and averaging over time are not the same operation, and describe a situation where confusing the two would mislead a manager.

12. Should we always compute throughput rate, flow time and inventory for all processes? In which case do you think considering one of them is enough? Why that one?
13. Radu insisted that “more throughput is better,” “smaller flow time is better,” and “less inventory is better.” Each of these is usually true and occasionally false. For each of the three metrics, construct a realistic operation in which moving the metric in the “better” direction actually makes the operation *worse* for someone who matters. What do your three counterexamples have in common?

## Reflections

*[Posted as a reflection exercise on the course’s website for students to write about in a journal due at the end of semester.]*

1. Radu admitted that after many years he is “still somewhat confused” about average inventory. Think of something you rely on daily but have never fully understood. Have you made peace with that confusion, or do you hide it?
2. How do you understand that the average inventory of  $1\frac{1}{3}$ ? What does  $1/3$  mean here? Is it saying there’s  $1/3$  person in the store?
3. Farah lit up at the first calculation and preferred whole numbers to “messier” fractions, even when they meant the same thing. Do you have similar preferences? Has a preference for the clean figure ever led you slightly astray?
4. Haoyu starts his clock the moment he enters the tea shop, not the moment he wakes up. When you complain that something “took forever,” where do you secretly start the clock, and would the person who kept you waiting agree with where you started it?
5. The class discovered that whether a served customer still counts as “inventory” depends entirely on where you draw the box. Think of a place where you have technically been served but are not yet free to leave. For example, a waiting room after the appointment, or a table after a check. Who benefits from your being counted as “done” while you are still sitting there?
6. Ming found it “shaky” that so much depends on your point of view. When someone insists there is one correct way to see a situation you’re in, do you interpret that as a sign of clarity?