Time is Money: The Value of "On-Demand"

Working Paper, January 7, 2011

Joe Weinman[1]

Permalink: http://www.JoeWeinman.com/Resources/Joe_Weinman_Time_Is_Money.pdf

Abstract

Cloud computing and related services offer resources and services "on demand." Examples include access to "video on demand" via IPTV or over-the-top streaming; servers and storage allocated on demand in "infrastructure as a service;" or "software as a service" such as customer relationship management or sales force automation. Services delivered "on demand" certainly sound better than ones provided "after an interminable wait," but how can we quantify the value of on-demand, and the scenarios in which it creates compelling value?

We show that the benefits of on-demand provisioning depend on the interplay of demand with forecasting, monitoring, and resource provisioning and de-provisioning processes and intervals, as well as likely asymmetries between excess capacity and unserved demand.

In any environment with constant demand or demand which may be accurately forecasted to an interval greater than the provisioning interval, on-demand provisioning has no value.

However, in most cases, **time is money**. For linear demand, loss is proportional to demand monitoring and resource provisioning intervals. However, linear demand functions are easy to forecast, so this benefit may not arise empirically.

For exponential growth, such as found in social networks and games, *any* non-zero provisioning interval leads to an exponentially growing loss, underscoring the critical importance of on-demand in such environments.

For environments with randomly varying demand where the value at a given time is independent of the prior interval—similar to repeated rolls of a die—on-demand is essential, and generates clear value relative to a strategy of fixed resources, which in turn are best overprovisioned.

For demand where the value is a random $\pm$ delta from the prior interval—similar to a Random Walk—there is a moderate benefit from time compression. Specifically, reducing process intervals by a factor of n results in loss being reduced to a level of $1/\sqrt{n}$ of its prior value. Thus, a two-fold reduction in cost requires a four-fold reduction in time.

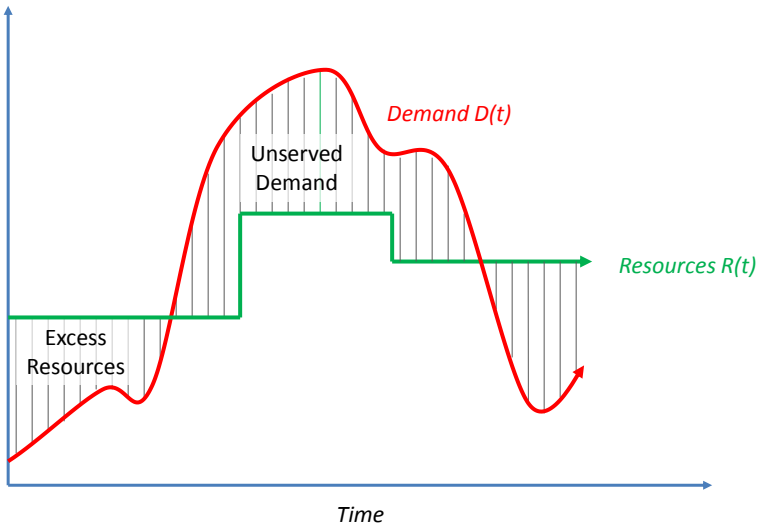
Finally, behavioral economic factors and cognitive biases such as hyperbolic discounting, perception of wait times, neglect of probability, and normalcy and other biases modulate the hard dollar costs addressed here.

---

[1] Joe Weinman leads Communications, Media and Entertainment Industry Solutions for Hewlett-Packard. The views expressed herein are his own. Contact information is at http://www.joeweinman.com/contact.htm

1. Introduction

In virtually any business, it is important to match capacity to demand. When there is a mismatch, costs arise in two ways; first, from unutilized or underutilized resources, and second, from having insufficient resources to meet demand. Excess resources incur unnecessary costs, and insufficient resources can cause lost revenue.



Imperfect Capacity

Consequently, capacity planning for production management and operations / service management is a standard topic in operations research and industrial engineering (OR/IE). How many sheet metal presses are required to produce 50,000 SUVs each year? How many check-out lanes should a grocery store have to optimize the trade-off between customer wait-time and experience vs. labor costs (cashiers), capital expenditures (cash registers), and operating costs (electricity for the registers)?

However, there are several key assumptions typically made for such environments. One is the ability to buffer inputs and/or outputs to decouple demand and resources in the short term but create alignment the long term. On the output side, manufacturers may produce inventory which is held at the plant, in the logistics and warehousing supply chain, or at distributors and retailers. On the input side, there may also be buffers, even in “just-in-time” manufacturing. Demand may be smoothed via standard order intervals, or service queues, i.e., waiting lines.

Related assumptions and variations include time and cost to construct or equip manufacturing facilities, techniques and ability to forecast demand, inventory holding costs and accounting methods (first in, first out; last in, first out; or weighted average), economic order quantities, queuing system behavior (balking, reneging, arrival and service time distributions), and so forth.

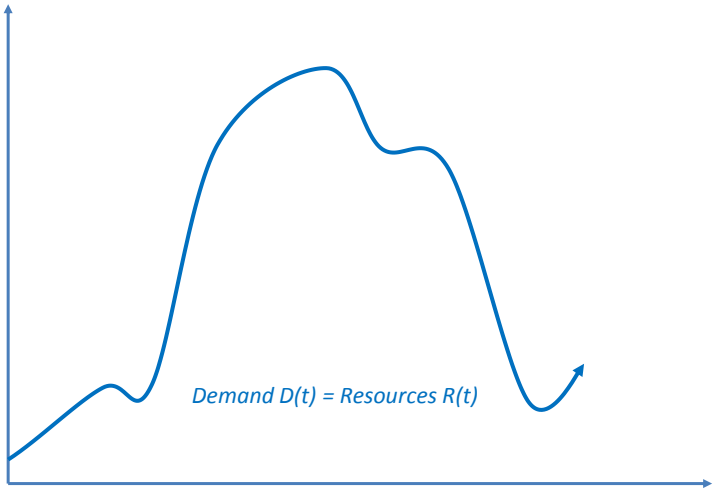
Today’s world of information technology in some ways does not cleanly fit those classic assumptions. Instead of order intervals, users demand instant gratification. While waiting two weeks or longer for a car matching your exact requirements as to exterior color, seat fabric, sunroof, and sport traction package may make sense, one does not expect to wait as long for, say, an online search query to be processed. Waiting even two seconds for the channel to change in an IPTV service is perceived as too long, and AJAX (Asynchronous Javascript and eXtensible Markup Language) applications such as “instant” searches require responses—across a wide area network, no less—in less time than it takes to type the next letter (100 milliseconds or so), or else have little use.

Worse yet, in aggregate, these workloads that demand instant response can have a high degree of variability and unpredictability due to millisecond fluctuations up to daily, weekly, seasonal, and multi-year (e.g., business cycles, fads, and fashions) cyclical factors, special events, openings, deadlines, or flash crowds, or exponential growth of popular sites and online games or communities[2]. Beyond the web, IT functions such as order processing for brokerages, design simulation for new products, compute intensive rendering and other post-production tasks for media and entertainment companies, can cause major shifts in aggregate resource requirements. Not only is it a challenge to defer work or buffer outputs, there is a strategic advantage to acceleration: enabling directors and producers to view movie “dailies,” accelerating drug discovery for pharmaceuticals, optimizing equity portfolios, including algorithmic and proprietary trading, etc.

Even manufacturing has moved beyond the traditional model to fabless production and manufacturing *services*. Whether a commercial service provider offers services to multiple commercial customers such as enterprises and consumers, or an internal service provider such as an Information Technology organization offers services to business units within the same corporation, multiplexing of demand from multiple sources enables greater flexibility and enhanced resource utilization vs. captive silos with fixed resources.

In the real world, it often seems that demand and resources are two time-varying functions that only intersect by accident, if at all. Given the losses associated with excess capacity or unserved demand, in an ideal world of what I’ll call “perfect capacity,” resources would exactly mirror demand, so there would be no such losses, as illustrated below.

[2] “Peaking Through the Clouds,” http://gigaom.com/2009/06/25/peaking-through-the-clouds/



Perfect Capacity

One hidden assumption in such a world is that the granularity (degree of quantization) of resource allocation quantity matches that of demand quantity: if I have a squeaky hinge, I not only need a drop of WD-40 *right now*, but I need a *drop*, not a barrel or supertanker’s worth. In computing, virtualization and multi-tenancy enable extremely fine-grained allocation of resources, so we will not address this topic beyond surfacing the assumption. In other scenarios, however, resourcing delays can cause loss, and so can quantization errors.

Another assumption is that we are taking the customer’s or user’s view of on-demand resources. Resources don’t materialize magically out of thin air. A private implementation may re-allocate existing resources on demand or procure them the old-fashioned way. But for the economics to be more than internal accounting transfers requires a service provider (commercial or internal) that can dynamically allocate, de-allocate, and re-allocate capacity among different customers. The service provider then has a separate challenge to maximize pool size and target different customer segments with different demand profiles to achieve benefits of statistical multiplexing, while managing aggregate capacity.[3] Otherwise, costs would just shift and one would potentially trade moderately-priced fixed resources for on-demand resources priced at a premium to enable elasticity, subject to other gives and takes such as learning curve effects, scale economies, competencies, technology refresh cycles, and the like.

I’ve previously[4] defined CLOUD as an acronym: a Common, Location-Independent Online Utility on-Demand service. This is roughly equivalent to the U.S. National Institute of Science and Technology’s[5] definition of a cloud as “a model for enabling convenient, **on-demand** network

[3] “The 10 Laws of Cloudonomics,” http://gigaom.com/2008/09/07/the-10-laws-of-cloudonomics/
[4] http://www.greentelecomlive.com/2009/03/16/full-interview-att%E2%80%99s-joe-weinman/
[5] http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc

access to a shared pool of configurable computing resources…that can be **rapidly provisioned and released** with minimal management effort or service provider interaction." Each of the parts of the definition creates economic value, but three are particularly important for the current discussion. *Common* pooled resources enable statistical multiplexing, reducing total resource requirements; *Utility*, or pay-per-use pricing, ensures no payment when resources are not used; and *on-Demand* is an enabler of business agility and cost reduction in meeting unexpected demand variation. From here on we will focus on this last attribute.

2. Formal Preliminaries

We will characterize a demand function $D(t)$: $\mathbb{R} \rightarrow \mathbb{R}$ for resources as a mapping from time to a real quantity of (needed) resources. Typically, $D(t) \geq 0$.[6] Similarly, we can define the resource allocation function over time as $R(t)$: $\mathbb{R} \rightarrow \mathbb{R}$.

Sometimes, demand and resources are expressed in the same units, for example, the demand for hotel rooms and the number of hotel rooms, or the demand for oil (in barrels) and the available supply. In other domains, such as computing, demand may be expressed in different units, e.g., the demand for a web site in user visits per day, page views per hour, hits per month, or transactions per second, vs. resources such as compute or storage. We will assume that we can translate one into the other, so that, say, "5 page views per second" of demand is equivalent to "1 server" worth of capacity.

We can define a *perfect capacity strategy* $\hat{R}(t)$, given a demand function $D(t)$, as one where $D(t) = \hat{R}(t), \forall t$, that is, one where the resources allocated at all times *exactly* match the requested demand. Moreover, if resources are sold in accordance with a usage-sensitive linear tariff, where $P(r) = k \times r$ is the price paid for r resources, then a utility with perfect capacity and linear tariff is one where $\frac{P(t)}{k} = \hat{R}(t) = D(t)$, or, more colloquially, one pays exactly in accordance with the resources allocated, which in turn exactly match the offered demand. Today's cloud computing services essentially match this definition.

This simple equivalence captures the notion of an agile, "infinitely" elastic (i.e., no *a priori* bound on either demand or resources allocated) pay-per-use utility. As I've previously explored[7], such a utility can offer compelling benefits, either on its own, or in a hybrid architecture when used together with dedicated resources with fixed capacity (and thus without on-demand resourcing).

[6] We can imagine scenarios where $D(t) < 0$ as ones where an agent has excess resources and would like to sell, rather than consume them.

[7] "Mathematical Proof of the Inevitability of Cloud Computing," http://cloudonomics.wordpress.com/2009/11/30/mathematical-proof-of-the-inevitability-of-cloud-computing/

Such benefits are so compelling that it can be shown that, in a duopoly with rational optimizing participants, virtually all users will migrate to the pay-per-use utility.[8]

It is tempting to interchange the terms "perfect capacity" with "on-demand." However, as we will see, while an on-demand strategy offers perfect capacity, there are perfect capacity approaches for some demand functions that don't require on-demand provisioning, for example, when demand is perfectly flat, or when demand shifts can be accurately forecasted in terms of both timing and magnitude. In practice, however, such situations are rare. For example, there was no question as to the timing of the Beijing Olympics (08/08/08), but the number of viewers that tuned in was subject to a number of factors. Benjamin Franklin suggested that nothing is certain but death and taxes—the magnitude of the latter is unclear and the timing of the former typically is as well.

Suppose, however, that $\exists \mathrm{i}, R(\mathrm{t_i}) \neq D(\mathrm{t_i})$, i.e., there is at least one point in time where the quantity of resources differs from the level of demand. If $R(\mathrm{t_i}) < D(\mathrm{t_i})$, then there are insufficient resources (or too much demand) at time $\mathrm{t_i}$. On the other hand, if $R(\mathrm{t_i}) > D(\mathrm{t_i})$, there is an excess of resources (or, one might argue, insufficient demand) at time $\mathrm{t_i}$.

We may associate costs with both scenarios. In practice, these costs may be difficult to determine, and may be non-linear. For example, costs associated with *excess* resources for a given duration may be evaluated as pro-rated lease costs or depreciation associated with those resources, considering the weighted average cost of capital for the firm deploying the resources, the time value of money given a projected discount rate, the opportunity cost associated with other applications of that capital, operations, administration and maintenance costs, overhead, and so forth.

However, in this simple, preliminary analysis, we will assume that all these factors are aggregated, so that there is merely a fixed cost associated with each resource per unit time, c_r, so the cost of r unemployed resources for a duration of time t is merely $c_r \times r \times \mathrm{t}$.

The costs of *insufficient* resources may also be challenging to determine. In general, however, we can assume that the resources are utilized to accomplish some margin-enhancing revenue-generating or cost-avoiding function. For example, plane seat resources are used to enable the sale of air transportation services, hotel room and bedding resources are used to enable the sale of lodging services, and compute and storage resources are used to sell plane tickets as well as hotel rooms as well as other ecommerce services, deliver advertising impressions, and so forth. Instead of generating revenue, resources may be used to indirectly maximize revenue, avoid cost or both, e.g., calculating optimal routes or running yield management algorithms, providing a back-up to avoid the probability-weighted (risk-adjusted) cost of compliance penalties or operating loss associated with loss of data, and so forth.

Having r *insufficient* resources for time t is then *not* symmetrically $c_r \times r \times \mathrm{t}$ but rather a margin delta associated with the opportunity cost of lost revenue or non-avoided cost. These costs also may be non-linear, for example, one too few seats on a plane may result in a lost ticket sale, but

[8] "The Market for Melons," http://www.joeweinman.com/Resources/Joe_Weinman_The_Market_For_Melons.pdf

on a train, merely a congestion externality as someone is inconvenienced somewhat by overcrowding. They also may be time-dependent and application-dependent—consider insufficient servers for a tax preparation firm at 11PM on April 15th.[9] In practice, we will be interested in the after-tax margin, not the revenue: spending a million dollars in resources to generate a hundred million dollars in revenue may make sense, but not if the hundred million dollars in revenue is generated at a net loss or such a low margin as to make it wiser to avoid the whole enterprise together.

Again, we will simplify, and assume that the opportunity cost associated with insufficient resources is c_d, so therefore the cost of r insufficient resources is $c_d \times r \times$ t. A more sophisticated analysis could include additional costs due to customer defections due to poor service, bad word of mouth driving additional customer acquisition costs or reducing margin on customer acquisitions to overcome poor service, etc. All these are certainly real costs as well.

We can also expect that in most cases $c_d \gg c_r$, otherwise the resources cost more than the value that they generate, or they cost the same, in which case, why bother?

> ***Definition: The Loss Function:*** The *loss* L associated with resource allocation function $R(t)$ intended to fulfill demand $D(t)$ over a time interval $[t_1, t_2]$ is the sum of the losses due to unused resources and unserved demand weighted by their respective costs, which may be expressed as the definite integral:

$$L = \int_{t_1}^{t_2} [D(t) - R(t)] \times c_d \, dt \mid D(t) > R(t) \; + \int_{t_1}^{t_2} [R(t) - D(t)] \times c_r \, dt \mid R(t) > D(t)$$

With that as a definition of loss, we can evaluate a perfect capacity strategy.

> ***Proposition 1:*** A perfect capacity strategy $\hat{R}$, where $\hat{R} = D(t), \forall t$, has a loss of 0.
>
> ***Proof:*** By the definition of the loss function L,
>
> $$L_{\hat{R}} = \int_{-\infty}^{\infty} \left[D(t) - \hat{R}(t)\right] \times c_d \, dt \mid D(t) > \hat{R}(t) + \int_{-\infty}^{\infty} \left[\hat{R}(t) - D(t)\right] \times c_r \, dt \mid \hat{R}(t) > D(t)$$
>
> Since, for a perfect capacity strategy, $\hat{R} = D(t)$,
>
> $$L_{\hat{R}} = \int_{-\infty}^{\infty} [0] \times c_d \, dt \mid D(t) > \hat{R}(t) + \int_{-\infty}^{\infty} [0] \times c_r \, dt \mid \hat{R}(t) > D(t) = 0 \; \blacksquare$$

With this in mind, the cost of imperfect capacity is the sum of the costs of unused resources not employed to service demand, together with the opportunity costs driven by unserved demand. Different capacity plans or strategies may have a different mix of unused resource costs and opportunity costs. The value of perfect capacity $\hat{R}$ versus an imperfect capacity plan I is then the net difference in costs between $\hat{R}$ and the costs of the imperfect capacity plan I. If we denote the value as V and the losses associated with the imperfect capacity plan L_I, and negate

[9] In the United States, personal income taxes must be postmarked or e-filed by April 15th.

each to denote them in terms of gains to determine relative gains, then the value V is the loss 0 less the avoidance of the loss L_I, or simply:

$$V = -L_{\hat{R}} - -L_I \;\; = 0 - -L_I = L_I$$

In other words the value of perfect capacity is that of not having losses associated with other approaches, the way that the value of an umbrella is in avoiding the cost of getting drenched.

In a prior analysis[10], I've analyzed the value of services with usage-sensitive pricing models. An underlying assumption in the analysis was that it is a utility with pay-per-use pricing (a linear tariff) and on-demand capacity, in other words, one pays only for resources used, and those resources are immediately available for use at a constant unit price.

In this analysis, we will particularly focus on the value of "on-demand," in effect, the value of a perfectly-capacitated environment versus other capacity strategies. The way we shall think of the value of on-demand is in comparison to the best possible alternate strategy. Sometimes, "on-demand" is loosely used as a substitute for "utility," because often, pay-per-use utilities such as electricity are available on demand. However, I prefer to clearly distinguish between them. A hotel where reservations need to be made a year in advance is hardly "on-demand," but may still be a "utility," i.e., have usage sensitive pricing: the total charge is the number of rooms times the number of nights times the rate. Conversely, having sufficiently large fixed capacity can ensure that any demand can be met, but may not be priced on a usage–sensitive basis.

Instead of adjusting resources to fit demand, one might attempt to shape demand to fit resources. For example, hospitality businesses, such as hotels, and transportation businesses, such as airlines, use yield management techniques to charge more for resources when demand is high and charge less or otherwise run promotions to encourage resource use when demand is low. Cities use congestion pricing to disincent traffic from the city center. Doctors arrange patient visits for open slots in their schedule.

In today's increasingly interactive and real-time world, however, this is difficult for businesses in many industries, consequently, rather than shaping demand (what one might call "on-resource demand"), the challenge is to ensure resource availability to meet offered demand, so we will assume that $D(t)$ is a given.

Although time is continuous, we will sometimes simplify our analysis by assuming that demand and resources may vary in discrete steps. These may be weeks, hours, minutes, seconds, or milliseconds, but it lets us simplify some of the math and illustrate more clearly certain effects.

In the next 9 sections, we'll explore a variety of scenarios: constant demand, linear growth and decline, exponential growth, uniform stochastic demand, and Random Walks. The value of on-demand varies across those scenarios from none, to sublinear, to linear, to exponential.

[10] "Mathematical Proof of the Inevitability of Cloud Computing," http://cloudonomics.wordpress.com/2009/11/30/mathematical-proof-of-the-inevitability-of-cloud-computing/

3. Constant Demand

As we might guess, in an environment where the demand function is constant, on-demand provisioning has no value compared to the best alternative, fixed capacity.

> ***Proposition 2:*** For constant demand where $\forall t, D(t) = k$, k a constant, the optimal fixed capacity is $F = \overline{D} = k$, which is perfect.
>
> ***Proof:*** We show that the optimal fixed capacity is $F = k = \overline{D}$ and that the value of the loss function under this scheme is identical to the perfect capacity strategy of zero. Let the fixed capacity be $F = k = \overline{D}$. The loss function, by definition, is
>
> $$L = \int_{t_1}^{t_2} [D(t) - R(t)] \times c_d \, dt \mid D(t) > R(t) \; + \int_{t_1}^{t_2} [R(t) - D(t)] \times c_r \, dt \mid R(t) > D(t)$$
>
> Let the resource strategy be $R(t) = F$. Since $F = \overline{D}$, the loss associated with this strategy is:
>
> $$L = \int_{t_1}^{t_2} [D(t) - F] \times c_d \, dt \mid D(t) > F \; + \int_{t_1}^{t_2} [F - D(t)] \times c_r \, dt \mid F > D(t)$$
>
> $$= \int_{t_1}^{t_2} [k - F] \times c_d \, dt \mid k > F \; + \int_{t_1}^{t_2} [F - k] \times c_r \, dt \mid F > k$$
>
> $$= \int_{t_1}^{t_2} [0] \times c_d \, dt \mid k > F \; + \int_{t_1}^{t_2} [0] \times c_r \, dt \mid F > k$$
>
> $$= 0 = L_{\hat{R}} \; \blacksquare$$

Although this proof was straightforward—and arguably more formal than was necessary—there are several observations to be made here. The most obvious is that if a given demand level is fixed at a given level, the correct amount of resources to serve it is also fixed at that level[11]. Although we didn't use this approach, we also might have shown that if $F > \overline{D}$, for example, $F = \overline{D} + \delta$, or $F < \overline{D}$, e.g., $F = \overline{D} - \delta$, then the loss associated with either $F = \overline{D} + \delta$, or $F = \overline{D} - \delta$ is greater than the loss associated with $F = \overline{D}$. In one case, the loss is $T \times \delta \times c_r$, and in the other it is $T \times \delta \times c_d$. The best way to minimize L is at $F = \overline{D}$.

But perhaps the most important point is that this proof illustrates that **"on-demand" differs from "perfect capacity."** An on-demand resourcing approach will ensure perfect capacity, but simply because there is perfect capacity does not mean that there is on-demand resourcing.

In fact, it is not only the trivial case of fixed demand and equivalent fixed resources that can enable perfect capacity. Indeed, if wildly variable demand can be exactly forecasted at least as

[11] Excluding the need for $n + 1, n + 2, n \times 2$, or other sparing and business continuity requirements.

far in advance as the provisioning/deprovisioning interval, resources may be deployed or deallocated just in time for the demand to vary.

However, in the real world it is difficult to achieve a high level of forecast accuracy for very long. Consequently, on-demand capabilities are proving to be important.

4. Forecasting and Provisioning Intervals

For the remaining analyses, we need to take into account four time intervals. The first is our forecast visibility. For weather, for example, we know not only whether it is raining now or not, but have a pretty good idea a few days out. For other events, we also have a good ability to forecast even years out—as I write this in 2010, the World Cup locations have been set for Russia 8 years from now and Qatar 12 years from now in 2022. Other events are often surprises, e.g., the death of a celebrity and the corresponding flood of hits at a microblog or news web site. To the extent that events drive demand for resources, our ability to forecast resource requirements is an important factor to consider.

A second key interval is the amount of time it takes to provision resources when requested. Sometimes this is a contractually guaranteed time, for example provisioning a managed server. Other times it is a distribution, such as acquiring a cab on a rainy day. This distribution may have a long, unbounded tail. No matter how long you’ve already waited, it may take a little longer before you get one.

The third key interval is the amount of time it takes to de-provision or deallocate resources. In the case of a cab, it is fairly quick: pay and hop out. In the case of selling a resource such as a house after the collapse of the housing bubble, it may take a while.

The fourth key interval is the demand monitoring cycle: how frequently are we evaluating demand? When this is zero, we are monitoring continuously, otherwise we are “checking in” periodically.

In the real world, all of these intervals may have some non-trivial probability distribution. This distribution may be due to underlying random processes, or may be caused by deterministic factors, e.g., all equipment leases are exactly three years long and begin on the first day of the quarter.

Let us call the forecast visibility t_f, the provisioning interval t_p, the deprovisioning interval t_d, and the monitoring cycle time (periodicity) t_m. These are all important intervals. When the forecast visibility $t_f \geq t_p$ and $t_f \geq t_d$, then we have plenty of advance warning to allocate and deallocate resources. If the forecast visibility $t_f < t_p$ or $t_f < t_d$, then we may have a surprise in store for us if there is a variation that occurs beyond where the forecast period reaches where

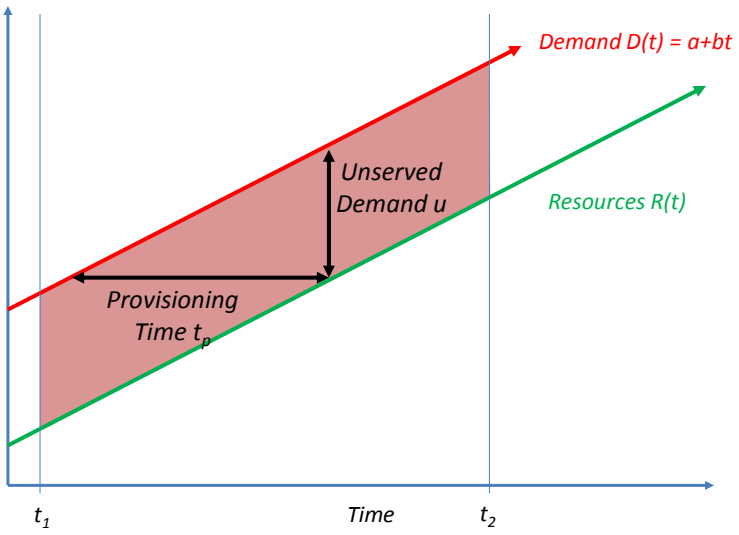
we don't have enough time to react. The size of the surprise and therefore the cost of excess resources or unserved demand will vary based on the underlying process driving the demand.

If we wanted to be even more formal than we have been, we could characterize an environment as an 8-tuple $(D(t),\ R(t),\ t_f,\ t_m,\ t_p,\ t_d,\ c_r,\ c_d)$. We won't go that far, but will examine various permutations of these and use English to differentiate between them.

5. Linearly Increasing Demand, No Forecasting, Continuous Monitoring, Non-Zero Provisioning Interval

The simplest interesting case is linearly increasing demand, where $D(t) = a + bt$. We will assume for now that $t_f = 0$, that is, that we have no visibility into the future, to understand what happens when we reduce provisioning intervals, including all the way to $t_f = 0$. This is a somewhat artificial example, in that we aren't postulating a forecasting process that simply extrapolates linear demand, but will be helpful for a first analysis. Also, it is not as artificial at it seems, after all, housing prices and equity prices often rise…until they don't; and a conservative capital expenditure strategy might dictate no resource investment until there is provable demand.

In this simple case, we don't need to worry about deprovisioning or t_d, since the demand is monotonically increasing. Let us first consider a continuous monitoring scenario.



Linearly Increasing Demand with Continuous Monitoring
And Non-Zero Provisioning Interval

***Proposition 3*:** Excluding edge effects, when $t_f = 0$ and $D(t) = a + bt$, the loss is proportional to the provisioning interval t_p.

Proof: In this scenario, we look at the current demand, and use it to provision future capacity, without attempting to extrapolate the demand curve. Since we continuously evaluate demand—which is linearly increasing—and (upwardly) adjust capacity accordingly, the resources also are linearly increasing. However, the quantity of resources is delayed by the provisioning interval t_p, that is, $R(t) = D(t - t_p)$, so we constantly are under-capacity by the amount u of unserved demand as shown in the diagram. In fact, u, which is $D(t) - R(t)$ is simply $b \times t_p$. The loss associated with this unserved demand is

$$L = \int_{t_1}^{t_2} [D(t) - R(t)] \times c_d \, dt \mid D(t) > R(t) \; + \int_{t_1}^{t_2} [R(t) - D(t)] \times c_r \, dt \mid R(t) > D(t)$$

Which, since demand is always greater than resources simplifies to

$$L = \int_{t_1}^{t_2} [D(t) - R(t)] \times c_d \; dt$$

Therefore,

$$L = \int_{t_1}^{t_2} b \times t_p \times c_d \, dt \; = (t_2 - t_1) \times b \times t_p \times c_d$$

Thus, the loss L is proportional to the provisioning delay t_p .

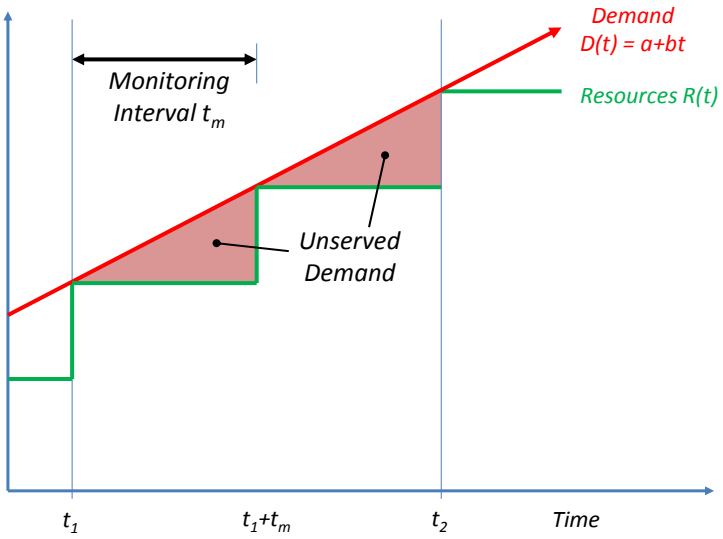
Another way to look at this is that $D(t) = a + bt$ and since $R(t)$ is displaced by t_p we have $R(t) = D(t - t_p) = a + b(t - t_p) = a + bt - bt_p$. Thus we know that $D(t) > R(t)$ everywhere and that $D(t) - R(t) = a + bt - (a + bt - bt_p) = b \times t_p$ which leads us to the loss above. ■

Note that there may be edge effects when $D(t) < u$. Before demand has had a chance to grow very much, the actual difference between $D(t)$ and $R(t)$ may not yet have reached u. We can ignore this for the general case.

6. Linearly Increasing Demand, No Forecasting, Periodic Monitoring, On-Demand Provisioning

An alternate scenario is one where there is periodic monitoring with on-demand provisioning. In other words, we check demand levels, say, once a day, and then instantaneously provision

capacity based on the determination of demand level. In the prior scenario, we were checking demand requirements, say, every second, but it took a while to provision capacity. In this scenario, we only check, say, every day, but can instantaneously provision capacity. One might think that the scenarios are identical, but as can be seen from the two illustrations, they are different. In the first scenario, the resource function mirrors the demand function, except for time delay. In the second scenario, the demand function is linear, but the resource function is stairstep.



Linearly Increasing Demand with Periodic Monitoring And On-Demand Provisioning

How does the monitoring interval relate to the loss function?

Proposition 4: Assuming a monitoring event occurs at time t_1 and periodically every t_m thereafter until time t_2, that $t_m \mid (t_2 - t_1)$, and $t_p = 0$, then the loss is proportional to the monitoring interval.

Proof: Recall that the loss in the interval t_1 to t_2 is:

$$L = \int_{t_1}^{t_2} [D(t) - R(t)] \times c_d \, dt \mid D(t) > R(t) + \int_{t_1}^{t_2} [R(t) - D(t)] \times c_r \, dt \mid R(t) > D(t)$$

Since $R(t)$ is strictly not greater than $D(t)$, this reduces to:

$$L = \int_{t_1}^{t_2} [D(t) - R(t)] \times c_d \, dt \mid D(t) > R(t)$$

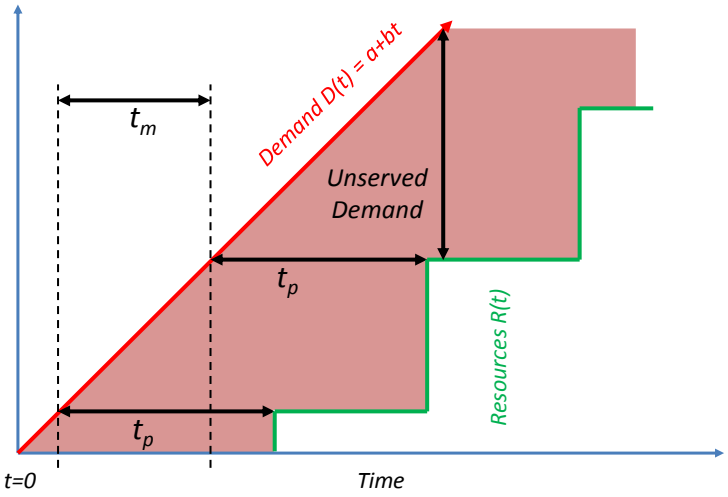
Each right triangle shown has a base of length t_m, and a height of $b \times t_m$, consequently its area is $\frac{1}{2}\, b \times t_m^2$. Since t_m divides $(t_2 - t_1)$, let $t_2 - t_1 = k \times t_m$. Then

$$L = k \times \tfrac{1}{2}\, b \;\times\; t_m^2 \times c_d = \frac{(t_2 - t_1)}{t_m} \times \frac{1}{2} b \;\times\; t_m^2 \times c_d = \frac{(t_2 - t_1)}{2} \times b \times t_m \times c_d \;\blacksquare$$

What this shows is that, in an on-demand environment, the loss is proportional to the monitoring interval. As we might expect, if the monitoring interval t_m drops to zero, that is, there is continuous monitoring with on-demand provisioning, the loss drops to zero as well. If the slope b drops to zero, then the loss is zero as we showed in the flat demand case earlier, regardless of the monitoring interval t_m.

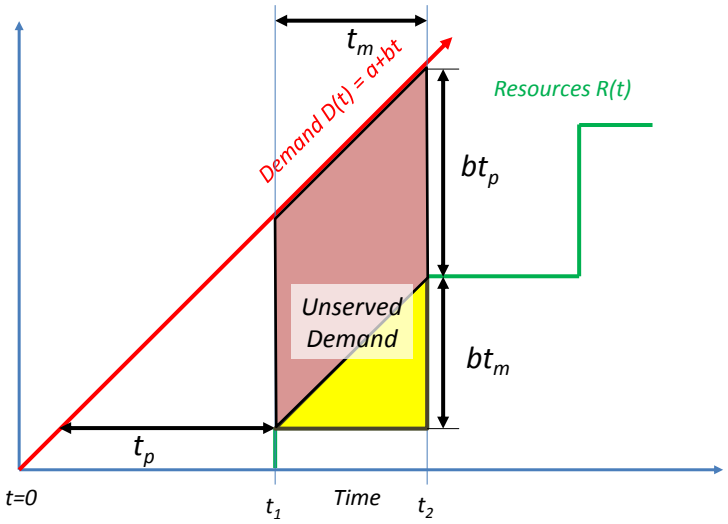
7. Linearly Increasing Demand, No Forecasting, Periodic Monitoring, Non-Zero Provisioning Interval

Monitoring interval and provisioning interval delays can drive unserved demand separately, as we have seen above, as well as together, as the chart below illustrates. Monitoring occurs periodically, and once a new "snapshot" of demand is taken, the provisioning delay causes a further gap. The largest such gap occurs just before resources are incremented.



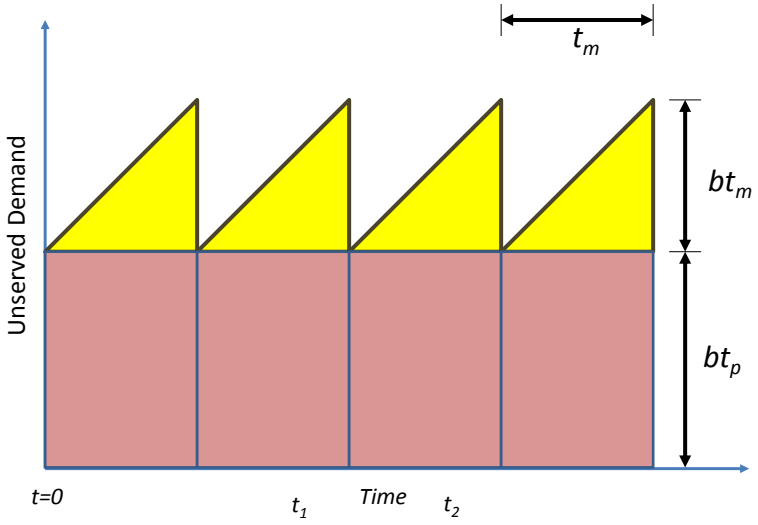
Linearly Increasing Demand with Periodic Monitoring And Non-Zero Provisioning Interval

It will be appreciated that the quantity (i.e., area) of unserved demand is now the sum of the two prior results. We can look at the unserved demand in the period between resource adjustments in this way:



Linearly Increasing Demand with Periodic Monitoring And Non-Zero Provisioning Interval

It will be appreciated then that the unserved demand over time is a sawtooth of height $b \times t_m$ and wavelength t_m sitting on a base of constant height $b \times t_p$.



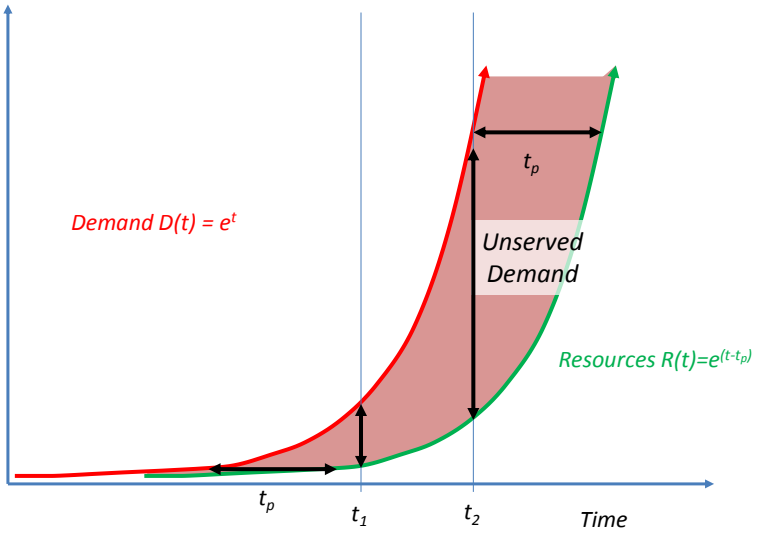
Linearly Increasing Demand with Periodic Monitoring And Non-Zero Provisioning Interval

If t_p is small relative to t_m, this is dominated by t_m. Conversely, if t_m is small relative to t_p, this is dominated by t_p. If we restrict our intervals to those that are multiples of t_m, the loss is strictly

proportional to the time. In any event, however, the longer the interval, the closer to exact proportionality to time the loss gets.

8. Exponential Growth

Suppose demand growth is exponential, as a number of services seem to be these days. Growth of many systems such as social networks, especially in the early days is proportional to the size of the network. This can't continue forever, so normally growth slows and then flatlines in accordance with the logistic function, but the timing on this is hard to predict, as any number of investors have found. Let's assume that $(t) = e^t$, that monitoring is continuous ($t_m = 0$), and that the resource provisioning interval as always is t_p. It appears as though the width of the difference is narrowing, but it is in fact constant



Exponential Growth with Continuous Monitoring And Non-Zero Provisioning Interval

As may be seen, even though the interval remains constant, as time unfolds, the gap between demand and resources continues to grow. That gap is $e^t - e^{(t-t_p)}$. The implication is that in an environment of exponential growth, on-demand becomes increasingly important, because no matter how short the provisioning interval (assuming it is greater than zero), the quantity of unserved demand, and thus the loss associated with that demand, will grow without bound.

> ***Proposition 5:*** In a continuous monitoring environment with resource provisioning interval t_p, if $D(t) = e^t$ then $R(t) = e^{(t-t_p)}$, and the loss in the interval t_1 to t_2 is $k(e^{t_2} - e^{t_1})$ where k is a constant equal to $(1 - e^{-t_p}) \times c_d$.

Proof: We note that $t_p \geq 0$, therefore $D(t) \geq R(t)$ everywhere. Consequently, we only need worry about unserved demand, i.e., the loss function is

$$L = \int_{t_1}^{t_2} [D(t) - R(t)] \times c_d \, dt$$

which is

$$L = \int_{t_1}^{t_2} \left[e^t - e^{(t-t_p)} \right] \times c_d \, dt$$

We can rewrite this as

$$L = \int_{t_1}^{t_2} e^t \times (1 - e^{-t_p}) \times c_d \, dt$$

The terms to the right are just a constant, so let's call them $k = (1 - e^{-t_p}) \times c_d$, and then we have

$$L = \int_{t_1}^{t_2} k e^t \, dt$$

Since the integral of $c \times e^x$ is $c \times e^x$ we see that

$$L = k e^t \Big]_{t_1}^{t_2} = k(e^{t_2} - e^{t_1}) \blacksquare$$

What does this actually say? It means that not only does the *demand* grow exponentially, but that if we set t_1 to be, say zero, then as time progresses (we let t_2 get larger and larger), the *loss* function is *also* essentially exponential. The only way to prevent this is to have $k = 0$, which happens only in an on-demand environment:

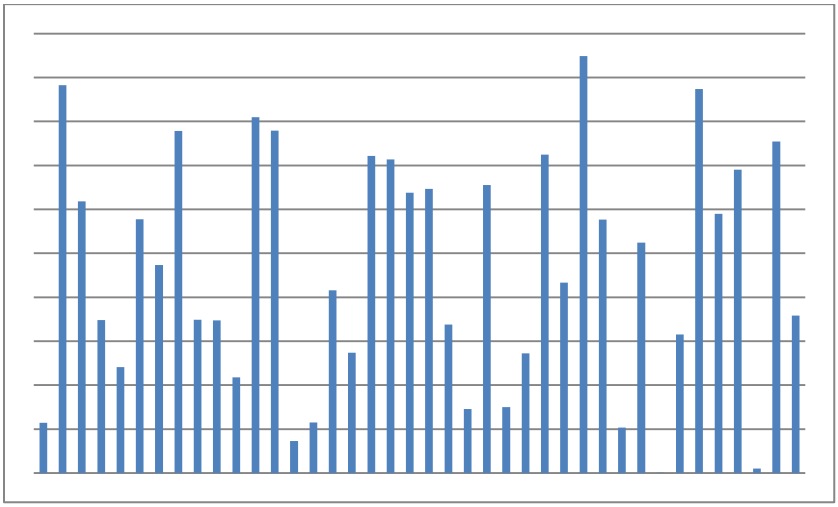
$$t_p = 0 \Rightarrow k = (1 - e^{-t_p}) \times c_d = (1 - e^0) \times c_d = (1 - 1) \times c_d = 0$$

9. Declining Demand Scenarios with No Forecasting

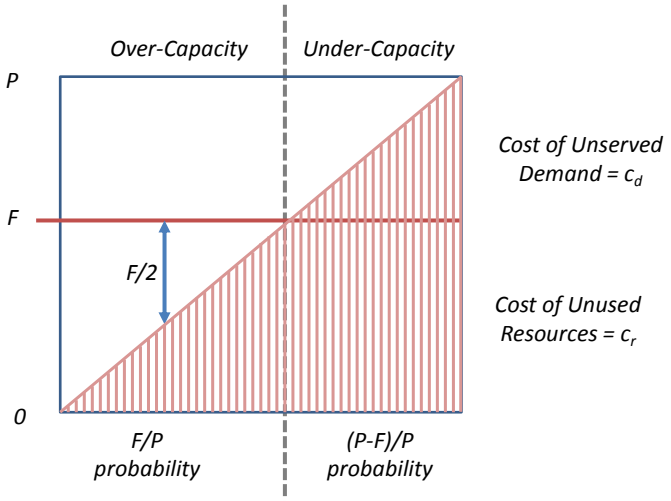
We could spend another few pages formally repeating the last few proofs for the cases where demand *decreases*, but this is hardly necessary. Rather than go through the details, it should be clear that in an environment of declining demand, the same principles hold, except that every c_d (cost of unserved demand) needs to be replaced with a c_r (cost of (excess) resources), and every t_p needs to be replaced with a t_d. After all, we are no longer worried about how long it takes to provision *new* resources, but rather with how long it takes to *get rid of* excess resources that we no longer need. However, the proportionality—"time is money"—remains, with the best possible solution being on-demand.

10. Random Demand I: On-Demand vs. Fixed Capacity for Uniformly-Distributed Demand

In a uniform distribution, all values are equally likely:



For the analysis, without loss of generality we will sort the values from lowest to highest, leading to a canonical sorted distribution that looks like a right triangle.



While the distribution of actual trial results will not exactly match this idealized triangle, for our purposes here this is “close enough” and correct in the limit. Let D be distributed uniformly on a

range from 0 to P (for peak). Although the average demand is then $P/2$, that is not necessarily the optimal fixed capacity if there is an asymmetry between c_d and c_r, as is likely in the real world. Since c_d should be greater than c_r, the optimal fixed capacity is greater than $P/2$, because it is more costly to not serve demand than to pay for unused resources.

Proposition 6: The optimal fixed capacity F for uniformly distributed demand is $(P \times c_d) / (c_r + c_d)$.

Proof: Let the optimal fixed capacity be F. Since D is uniformly distributed, F/P of the time, the demand will be less than F, $(P - F)/P$ of the time, the demand will be greater than F.

In the first case, that of overcapacity, we have $R(t) > D(t)$. The expected value of $R(t) - D(t)$, given that $D(t) < F$, is $F/2$. That is, $(D(t) - R(t)) = F/2 \mid 0 \leq D(t) \leq F$. In this case, the cost associated with such overcapacity is $F/2 \times c_r$.

In the second case, that of undercapacity, we have $R(t) < D(t)$. The expected value of $D(t) - R(t)$, given that it is greater than F, is $(P - F)/2$. In other words, $E(D(t) - R(t)) = (P - F)/2 \mid F \leq D(t) \leq P$. Now however, the expected cost associated with undercapacity is $(P - F)/2 \times c_d$.

Given the likelihood of either case, we see that the expectation of the cost C at any given time t is:

$$E(C_t) = \frac{F}{P} \times \frac{F}{2} \times c_r + \frac{(P - F)}{P} \times \frac{(P - F)}{2} \times c_d$$

Rearranging some terms, we have:

$$E(C_t) = \frac{F^2 c_r + [P^2 - 2PF + F^2]\, c_d}{2P}$$

This is at a minimum when the first derivative with respect to F is 0, which is when

$$E'(C_t) = \frac{F^2 c_r + [P^2 - 2PF + F^2]\, c_d}{2P}$$

$$= 0 = \frac{1}{2P} \times [2F c_r + 0 - 2P c_d + 2F c_d]$$

Cancelling the "2"'s and multiplying both sides by P gives us:

$$E'(C_t) = \frac{F^2 c_r + [P^2 - 2PF + F^2]\, c_d}{2P}$$

$$= 0 = F c_r - P c_d + F c_d$$

From which we can deduce that:

$$F = P\frac{c_d}{c_r + c_d} \blacksquare$$

This matches our intuition. The smaller the peak is, the smaller F should be. If the cost of over- and under-capacity are identical, the loss is minimized when F is at the average demand level, but the larger c_d is relative to c_r, the closer F should be to the peak demand P. In other words, it's better to be safe than sorry.

We used a uniform distribution to simplify the exposition, but it should be appreciated that for other distributions, e.g., Gaussian (normal), roughly similar results should hold: there is a single fixed level of demand that minimizes loss, and that level is likely to increase as c_d/c_r increases.[12]

11. Random Demand II: Random Walks and Similar Stochastic Processes

Suppose we flip a coin, and I win a dollar if it lands heads and you win a dollar if it lands tails. How rich will either of us be after, say, a million coin flips? Assuming a fair coin with a probability of exactly $1/2$ of coming up heads and exactly $1/2$ of coming up tails, and independent tosses, in one sense we can "expect" to be no better or no worse off at the end of the process. However, that expectation conceals a wide range of outcomes, including the (remote) possibility that you or I end up millionaires. Moreover, at any given time, one of us or the other may be ahead, and if we were to play the game an infinite number of times, we can expect to be back to the "break-even" point (or any other point, for that matter) an infinite number of times due to a mathematical result called the "level-crossing phenomenon." The process we describe is a "simple random walk," which is more generally a type of stochastic process called a "martingale."

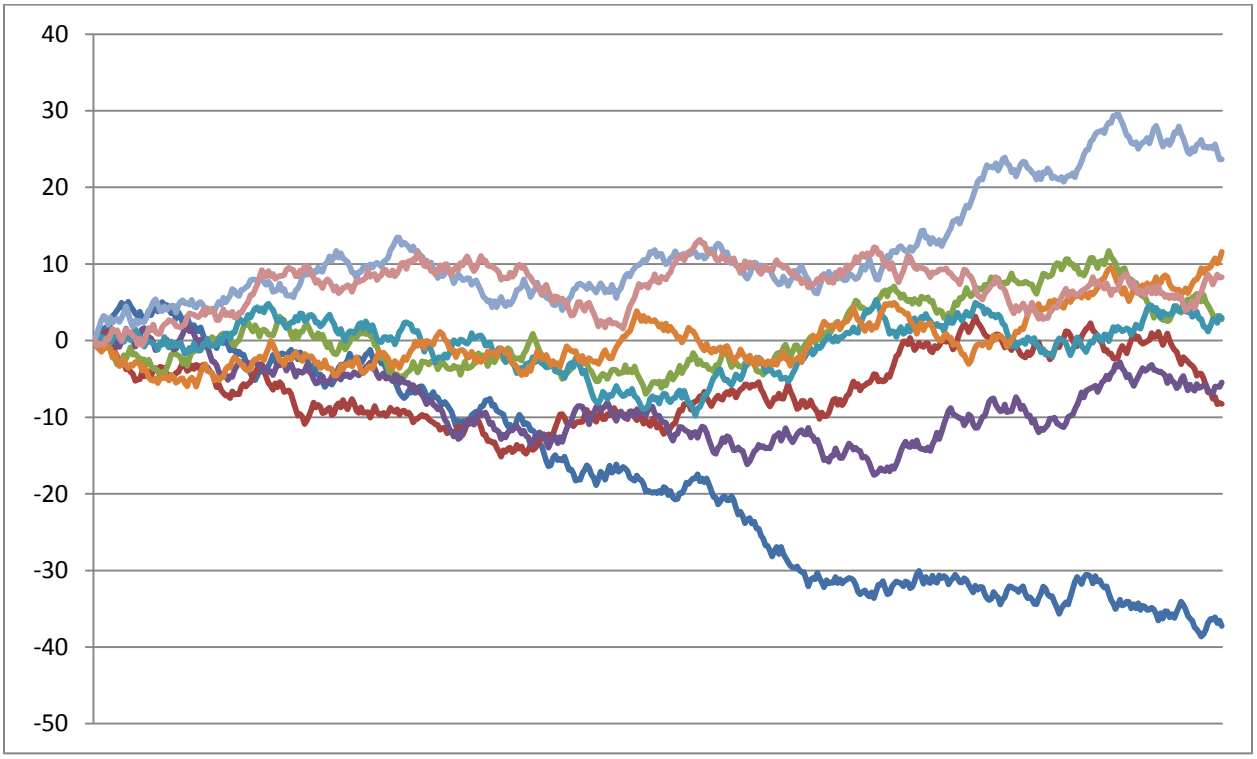
Some examples of this type of process are illustrated below, where the value at step n differs from the value at step $n-1$ by an amount uniformly distributed on the interval $[-1,1]$.[13] It may be observed that the "spread" widens over time, but for each of the runs, much less than the theoretical maximum of slope 1 or slope -1 which would lead to values at the right side of the chart of either 500 or -500 respectively.

The chart looks something like a smoke plume in a breeze. The coincidence is not accidental, as the position of smoke particles emanating from a point fire when there is a constant breeze

[12] Not always, e.g., for Bernoulli trials, once $\frac{c_d}{c_r} > 1$, setting $F = P$ will remain the best strategy.

[13] This chart was generated in a popular spreadsheet program, using eight columns each with 500 cells differing from the prior one by "2*RAND()-1"

provides similar[14] underlying math: Random Walk translation in the vertical dimension of the chart relates to the three-dimensional Brownian diffusion of smoke particles, coupled with a horizontal time dimension (particles are birthed at the fire and “age” as they travel downwind).



Such processes have been used to characterize everything from coin-tossing games to stock market behavior. They also may reasonably characterize web traffic or aggregate continuous streams such as live video, as in any short interval, a random net number of users may begin or end sessions. For our purposes, we will attempt to characterize the value of on-demand in such processes, which may reflect, say, the number of search or online auction or news story readers at any given time. There are a few important characteristics to note of such processes in general, and simple random walks (where each step has a delta of either +1 or -1.)

1) For simple random walks, after n steps, a total change of $+n$ or $-n$ is possible (though of probability $½^n$ and thus increasingly unlikely as n gets larger.

2) Although the expected value of the delta is zero, that expectation alone does not effectively characterize the nature of the distribution, as it is the expected value of the “translation” distance that is of interest (i.e., how far did the random walk “get” after taking one step forward, two steps backward, and so forth. This translation distance over time is illustrated by the value of the curves shown here, which is the sum of the trial results for all prior steps.

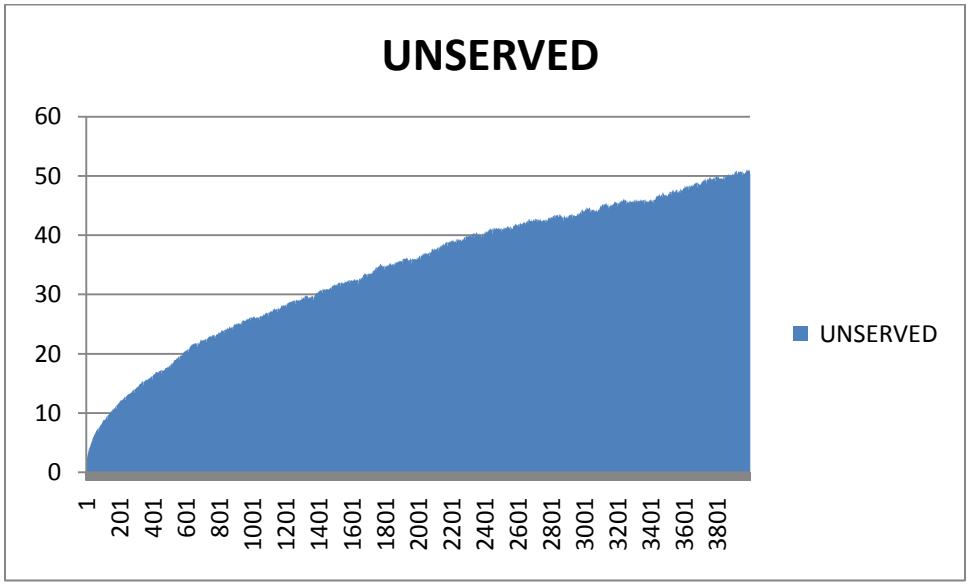
[14] Technically, simple random walks, which occur on an n-dimensional grid or lattice differ from Brownian motion, in which any direction of movement is possible.

3) The expected translation distance is $O(\sqrt{n})$, specifically, when the step size is a function with distribution 0 and standard deviation σ, it asymptotically approaches $\sqrt{\frac{2}{\pi}}\sigma\sqrt{n}$. For a simple random walk, since $\sigma = 1$, this reduces to $\sqrt{\frac{2}{\pi}}\sqrt{n}$.

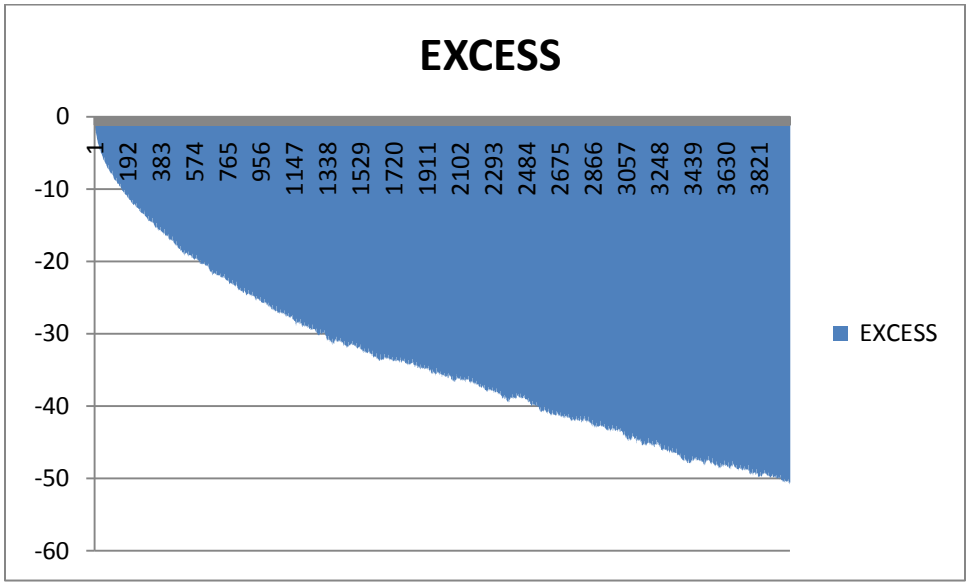
4) Interestingly, *only* the mean and standard deviation of the step function distribution are needed to characterize the translation distance, hence the result applies to a wide range of statistical distributions used to generate the random step.

These results are discussed and proven elsewhere in depth using formal analytic methods, e.g., see *Random Walk: A Modern Introduction* by Greg Lawler and Vlada Limic. However, we may use a simple HTML and Javascript Monte Carlo simulation for statistical computation to illustrate this behavior, documented in the Appendix below.

To run the simulation, use a simple text editor to cut and paste the code into a file, save the file with an ".htm" extension, then open the file in a browser. Select, cut and paste the comma separated values into a spreadsheet program, convert the single column into three data columns as necessary, and then chart them. Apple's Safari and Google's Chrome currently appear to run computations significantly more efficiently than at least one other popular browser. The results look something like this for a simple random walk over 4,000 steps, based on a Monte Carlo simulation with 4,000 trials.



This is the expected value of the translation distance, given that it is positive. As may be expected, the expected value of the translation distance, given that it is negative, is the mirror image. To put it another way, the absolute value doesn't change:



This is nicely in accordance with the theoretical result for the expected translation distance after 4,000 steps, which is $\sqrt{\frac{2}{\pi}}\sqrt{n}$, or $.798 \times \sqrt{4000} = .798 \times 63.246 = 50.462$.

Now let’s consider any stochastic demand curve that follows such a pattern[15], namely,

$$E(|D(t_2) - D(t_1)|) = k\sqrt{(t_2 - t_1)}$$

What this equivalence is stating is that given we know the demand at t_1, the expected difference in $D(t)$ at time t_2 (the translation distance) is proportional to the square root of how much time—or how many steps—have passed. Note that the unit of time doesn’t matter, any conversion of units would just change the value of k. Moreover, let’s assume that the translation distance is equally likely to be positive as negative.

When such an equivalence holds, how much is time compression worth, or to put it another way, what happens to the loss function? Recall that the loss is:

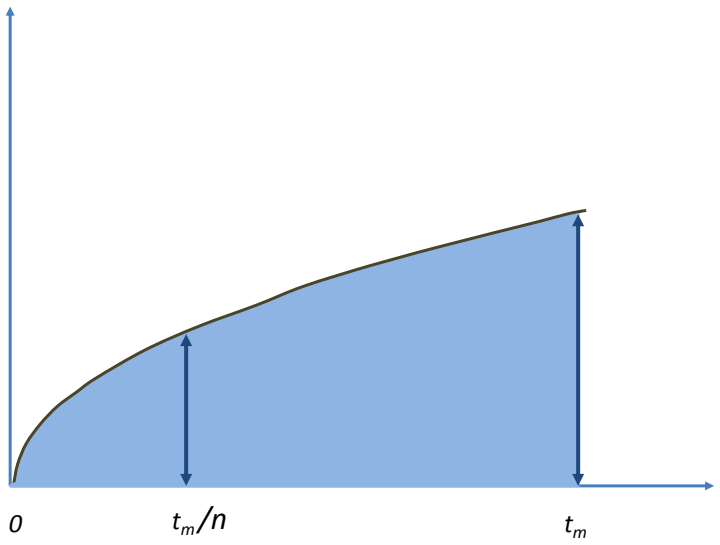
$$L = \int_{t_1}^{t_2} [D(t) - R(t)] \times c_d \, dt \mid D(t) > R(t) \; + \int_{t_1}^{t_2} [R(t) - D(t)] \times c_r \, dt \mid R(t) > D(t)$$

Let us assume periodic monitoring with a provisioning interval of $t_p = t_m$. In other words, we check the demand, say, every morning at 9:00 AM, and place an order to provision or deprovision capacity which is implemented by the next morning at 9:00 AM (in this case, $t_p = t_m = 1$ day). How far off will the demand curve be by then, and how much would we gain by moving to an 8 hour cycle? Or a one hour cycle? To answer this, we need to know how much the demand might have drifted in either of those times, and how much loss we expect to have incurred in that time.

[15] Technically, random walks only asymptotically approach this pattern, but per the empirical simulation results the values are close enough for our purposes. We will work with the “idealized” form of this demand.

Proposition 7: If a demand curve may be characterized by $E(|D(t_2) - D(t_1)|) = k\sqrt{(t_2 - t_1)}$ as described above, and there is synchronized periodic monitoring and provisioning such that $t_p = t_m$, reducing this interval by an integer factor of n to t_m/n, reduces the expected loss to $1/\sqrt{n}$ of the pre-reduction loss.

Proof: Let us consider first the expected loss due to insufficient capacity. We are really asking what the difference is between the area of this curve between 0 and t_m vs. the difference between 0 and t_m/n.



Fortunately, this is easy to determine, since the integral of $k\sqrt{x}$ is $\frac{2}{3}k\sqrt{x^3}$.

Determining the definite integrals for t_m and for $\frac{t_m}{n}$, we have that

$$\int_0^{t_m} k\sqrt{x}\,dx = \frac{2}{3}k\sqrt{{t_m}^3} \text{ and that } \int_0^{\frac{t_m}{n}} k\sqrt{x}\,dx = \frac{2}{3}k\sqrt{\left(\frac{t_m}{n}\right)^3}$$

What we need to do now is to plug in expected loss in both scenarios. We will ignore edge effects, to arrive at a good approximation of the benefit, or we can think of only considering total time frames that are multiples of the least common multiple of t_m and t_m/n.

In the scenario before time compression, the expected value of the loss over the interval $[0, t_m]$ is:

$$E(L_{t_m}) = \frac{1}{2} \times \frac{2}{3}k\sqrt{{t_m}^3} \times c_d + \frac{1}{2} \times \frac{2}{3}k\sqrt{{t_m}^3} \times c_r$$

Note that there is a 50% chance that demand will have "randomly walked" a translation distance that is positive, in which case we pay the loss associated with unserved demand, c_d, and there is a 50% chance that demand will have translated a distance that is negative, in which case we pay the loss associated with excess resources, c_r.

After time compression, in the same interval we have n copies of the loss associated with $\frac{t_m}{n}$, which is:

$$E(L_{\frac{t_m}{n}}) = n \times \left[\frac{1}{2}\times\frac{2}{3}k\sqrt{(\frac{t_m}{n})^3} \times c_d + \frac{1}{2}\times\frac{2}{3}k\sqrt{(\frac{t_m}{n})^3} \times c_r\right]$$

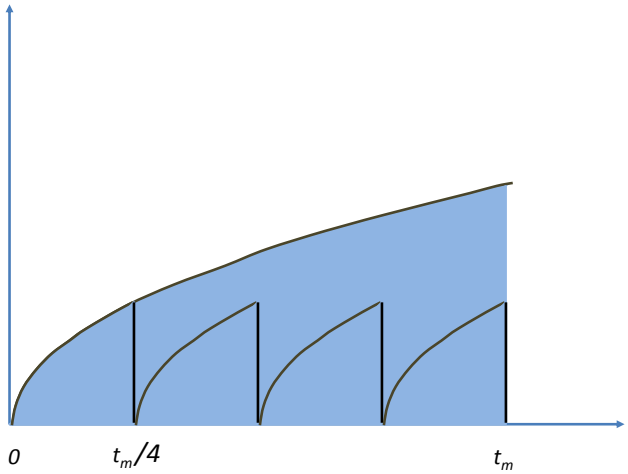
We can now assess the ratio of the two

$$\frac{E(L_{\frac{t_m}{n}})}{E(L_{t_m})} = \frac{n \times \left[\frac{1}{2}\times\frac{2}{3}k\sqrt{(\frac{t_m}{n})^3} \times c_d + \frac{1}{2}\times\frac{2}{3}k\sqrt{(\frac{t_m}{n})^3} \times c_r\right]}{\frac{1}{2}\times\frac{2}{3}k\sqrt{{t_m}^3} \times c_d + \frac{1}{2}\times\frac{2}{3}k\sqrt{{t_m}^3} \times c_r}$$

We can multiply fractions, rearrange and eliminate terms and the like to arrive at:

$$\frac{E(L_{\frac{t_m}{n}})}{E(L_{t_m})} = \frac{n \times \left[\sqrt{(\frac{t_m}{n})^3}\right]}{\sqrt{{t_m}^3}} = \frac{1}{\sqrt{n}} \blacksquare$$

A way to understand the expected loss better is by looking at the relative areas. As can be seen in the example below, compressing the time by $n = 4$ has less area (i.e., loss) than the original approach, showing some benefit to "mid-course corrections."



Clearly, there is a benefit from time compression, but it is sub-linear. To recap some earlier results for context, when demand is flat, the benefit from “on-demand” is zero. When demand is growing or declining linearly, the benefit from “on-demand” is linear. When it is a random walk, the expected change over time is subdued relative to linear growth, so the benefit from time compression is less compelling. To put it another way, because the random steps in random walks tend to cancel each other out, the loss is not as great as if they didn’t, and so the move to on-demand is slightly less compelling than in circumstances with greater variability over time.

12. Behavioral Factors and Cognitive Biases

“Behavioral Economics,” a rich field of research at the intersection of psychology and economics, has empirically shown that human beings don’t always behave according to the predominant assumptions of the last century that humans are rational optimizers. Instead, as I’ve written elsewhere[16], people tend to be lazy, hazy, and crazy: *lazy*, in minimizing cognitive, physical, emotional, and dollar costs; *hazy,* using short cuts and heuristics to solve problems; and *crazy,* using emotional (i.e., irrational) mechanisms for decision-making.

There is a rich body of literature and research studies in the field. Wikipedia has an excellent list[17], and a number of books explore these and related topics in depth: *Predictably Irrational* by Dan Ariely, *Sway* by the Brafmans, *How We Decide* by Jonah Lehrer, *Why We Buy: The Science of Shopping* by Paco Underhill, *Iconoclast* by Greg Berns, and *Your Brain at Work* by David Rock to name a few.

Behavioral economics and neuroeconomic results regarding cognitive biases and heuristics impact the strictly mathematical results above. Hyperbolic discounting and perception of wait times support an on-demand orientation, but the normalcy bias and related biases cause the importance of on-demand to be discounted.

The time value of money, present value, and discounted cash flow calculations tell us that a dollar today is worth more than a dollar a year from now, because that dollar can be invested in such a way that it will gain in the intervening year. The “hyperbolic discounting bias” suggests that people overvalue this difference, especially in the short term, above and beyond any financial benefit. At an interest rate of 5% annually, one dollar would be worth $1.05 in a year (excluding taxes). Therefore, in a day, one dollar would be worth $1.0001337, just over a hundredth of a cent more. However, just ask a five-year-old (or anyone else) whether she’d rather have a one dollar chocolate bar *right this instant* vs. receiving it tomorrow together with,

[16] “Lazy, Hazy, Crazy: The Ten Laws of Behavioral Economics,” http://gigaom.com/2010/06/06/lazy-hazy-crazy-the-10-laws-of-behavioral-cloudonomics/
[17] http://en.wikipedia.org/wiki/List_of_cognitive_biases

say, a quarter, a clear net profit of almost 25 cents, even after discounting, and a financially superior option.

Perception of wait times is surprisingly complex. Briefly, memories of wait times are encoded by a few factors, such as experience at the start, experience at the end, and average quality. Consequently, a longer wait that that culminated in a positive experience may have a more positive remembrance than a shorter wait. Wait times of up to two minutes are accurately perceived, but after that, there is a time expansion effect, where a three minute wait for example, might be perceived as having been four or five minutes. As Albert Einstein quipped, “When you sit with a nice girl for two hours, you think it's only a minute. But when you sit on a hot stove for a minute, you think it's two hours. That's relativity.” David Maister[18], argues that there are a number of rules in the perception of wait times, e.g., “occupied time feels shorter than unoccupied,” “uncertain waits are longer than known, finite waits,” and “solo waits feel longer than group waits.” These same biases are not unlikely to carry over to say, provisioning computing resources.

Finally, there are a few biases that appear to make it difficult for people to correctly plan for “black swans[19],” i.e., rare events where, e.g., demand shoots up or dramatically ebbs. The *Hindsight Bias* makes us falsely believe that prior events were more predictable than they actually were. The *Dunning-Kruger* effect suggests that we often falsely overestimate our abilities. The *Forward Bias* makes us extrapolate past trends linearly. The *Normalcy Bias* prevents us from wasting cognitive and emotional energy on planning for unlikely events or those not previously experienced, and *Neglect of Probability* indicates that people have difficulty incorporating probabilistic thinking into their planning.

13. Conclusion

We have seen that not only is there a time value of money, there is a **money value of time**, specifically, increased agility and responsiveness lead to reduced loss, including a reduction in missed opportunities. Time **is** money.

From a business perspective, one has to ask whether the reduction in monitoring or provisioning time that potentially results in reduced loss due to unserved demand or unused resources is worth it. I believe in most cases the answer is yes. The reason is that the costs of implementing such on-demand strategies are largely fixed, are a relatively minor portion of the total cost, or are already incorporated, say, into a cloud provider’s offerings. For example, the cost for an enterprise or cloud provider to acquire and deploy dynamic provisioning software compared to the losses associated with unserved demand or unutilized capacity make it an attractive proposition.

[18] David Maister, “The Psychology of Waiting Lines,” http://davidmaister.com/articles/5/52/
[19] Nassim Nicholas Taleb, “Fooled By Randomness” and “The Black Swan”

The bottom line is that when we attempt to deploy resources to serve demand in a variety of scenarios, we can never do any better than on-demand, but we can often do worse. To recap the insights derived here:

1) For flat demand, on-demand provisioning offers no benefit.

2) If demand can be forecasted accurately further out than it takes time to provision, on-demand provisioning also is unnecessary.

3) If on-demand is not an option, and the cost of unserved demand is greater than the cost of resources, it is better to be “safe than sorry” by building in excess capacity

However, when there is variability coupled with unpredictability, the following rules hold:

4) For linearly growing or declining demand, a reduction in time (monitoring cycle or resource provisioning) offers a proportional reduction in cost.

5) For exponential demand, the loss associated with even fixed interval provisioning grows exponentially, so on-demand provisioning is essential.

6) If the demand in each time interval is drawn from a random distribution, on demand is far superior to the expected loss from the best fixed capacity strategy.

7) If demand varies as in a Random Walk, time compression offers sub-linear benefits, for example, a two-fold reduction in cost requires a four-fold reduction in time.

These rules apply when there is no additional cost for resource provisioning or deprovisioning. For fixed, dedicated, owned resources, this is typically not true. However, for utilities such as cloud computing services, they are. Today’s managed / outsourced computing services have largely automated provisioning and dynamic resource allocation, culminating in pure “clouds,” or as I have defined elsewhere, “CLOUDS,” i.e., “Common, Location-Independent, Online Utility on-Demand Services.”

There are a number of ways of evaluating the extent to which Time is Money. The approach described here should help clarify in which circumstances and for what reasons on-demand is of value by providing a formal analytic foundation for such evaluation.

APPENDIX: Random Walk Javascript Simulation

```
<html>
<head>
<title>Random Walk Monte Carlo Expectation Analysis</title>
<script type="text/javascript" language="javascript">

var n = 3000 /* number of trials - set to any value but sim runs in time O(n * t) */
var t = 10000 /* number of random walk steps */
var unserved = new Array(n)
var excess = new Array(n)
var unservedcount = new Array(n)
var excesscount = new Array(n)
var i, j, total, stepsize

function simulate()
   {
   document.write("STEP, EXCESS, UNSERVED<br />")
   for(i=0; i<t; i++)
      {
      unserved[i] = 0
      excess[i] = 0
      unservedcount[i] = 0
      excesscount[i] = 0
      }
   for(i=0; i<n; i++) /* do n trials */
      {
      total = 0
      for(j=1; j<t; j++) /* do t steps in each trial */
         {
         stepsize= (Math.random() * 2.0) - 1 /* generates random variable between -1 and 1 */
         /* use next statement only for simple random walk (-1, or +1), otherwise remove */
         stepsize = stepsize < 0 ? -1 : 1

         total += stepsize
         if(total > 0)
            {
            unserved[j] += total
            unservedcount[j] += 1
            }
         else
            {
            excess[j] += total
            excesscount[j] += 1
            }
         }
      }
   /* print results, divided by sample count to get expectation */
   for(i=1; i<t; i++)
      {
      excess[i] = excess[i] / excesscount[i]
      unserved[i] = unserved[i] / unservedcount[i]
      document.write(i + "," + excess[i].toFixed(6) + "," + unserved[i].toFixed(6) + "<br />")
      }
   }
</script>
</head>
<body onload="simulate();" style="font-family:Courier,monospace;"> </body>
</html>
```

