

Smooth Operator: The Value of Demand Aggregation

Working Paper, February 27, 2011

Joe Weinman¹

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

Abstract

In industries such as cloud computing, lodging, and car rental services, demand from multiple customers is aggregated and served out of a common pool of resources managed by an operator. This approach can drive economies of scale and learning curve effects, but such benefits are offset by providers' needs to recover SG&A and achieve a return on invested capital. Does aggregation create value or are customers' costs just swept under a provider's rug and then charged back?

Under many circumstances, service providers—which one might call “smooth” operators—can take advantage of statistical effects that reduce variability in aggregate demand, creating true value vs. fixed, partitioned resources serving that demand.

If each customer's demand is flat or varies identically except for a scaling factor, aggregating the demand and serving it out of a shared pool of resources has no value. In fact, no matter what else happens off-peak, if a set of customers that have a build-to-peak capacity strategy all have a simultaneous peak, there is no benefit to aggregation.

When n customers' demands are independent but have the same mean and standard deviation, combining them has substantial value: the coefficient of variation—a measure of degree of variability, is reduced by a factor of \sqrt{n} , i.e., the aggregate is smoother than its components.

Aggregation increases the likelihood of meeting an availability Service Level Agreement with the same number of resources, or, equivalently, lets such an SLA be met with fewer resources.

If n demands are independent but normally distributed, then with the same quantity of resources, the penalty cost associated with the excess capacity and/or unserved demand is reduced to $1/\sqrt{n}$ of the unaggregated penalty cost.

Customers need to build capacity near peak, service providers can build capacity near average.

Sade, in her hit song *Smooth Operator*, sings of “higher heights,” and how “minimum waste” is tied to “maximum joy.” Operators who smooth variability via demand aggregation can enable individual customers to achieve higher peaks, while minimizing wasted resources. Joy may be in the eye of the beholder, but minimizing waste surely helps maximize economic value.

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

Cloud computing is a rapidly emerging paradigm in IT, in which computing resources and applications, rather than being owned by the end-user or business, are offered as a service over a network. Although new to IT, the *business model* of an operator that services multiple geographically dispersed customers out of a common, possibly geographically dispersed set of resources on a pay-per-use, on-demand basis has been around for decades.

Hotel and rental car chains have service nodes in many locations, renting rooms or cars on an on-demand basis (as well as via reservations). Large retailers and coffee shop chains also offer services in many locations, offering goods on a pay-per-use basis. I've called such business models "CLOUDs," i.e., Common, Location-independent, Online, Utility, on-Demand services, since they aggregate demand into common, or shared resources, offer broad reach to users regardless of location, are Online, or used via a network, provide Utility, or usage-sensitive pricing, and have resources that may be accessed "nearly" instantaneously.

In 2008, I coined the term 'Clouconomics' and proposed the *10 Laws of Clouconomics*² to characterize the underlying behavior of such "clouds."

Elsewhere³, I've explored in depth my 1st Law of Clouconomics: *Utility Services Cost Less Even Though They Cost More*, which explores scenarios where total cost can be reduced by utilizing pay-per-use services even if they have a higher unit cost than dedicated capacity. For example, the unit cost of an "owned" server for a given period of time may be lower than the unit cost of a "rented" one. However, rented resources may still have a lower total cost than owned ones under certain types of variable demand, because unlike owned, dedicated resources, they *aren't* paid for when *not* used.

I have also explored⁴ in some depth my 2nd Law of Clouconomics: *On-Demand Trumps Forecasting*, showing that the benefit of on-demand resource allocation is zero for flat or predictable demand, but can be sublinear, linear, or even exponential, for other types of demand.

This paper will be focused on the 3rd and 4th Laws, namely benefits achieved by aggregating demand and fulfilling it from pooled resources, which is what a service provider, or "smooth operator" does.

² Joe Weinman, "The 10 Laws of Clouconomics," at <http://gigaom.com/2008/09/07/the-10-laws-of-clouconomics/>, and also http://www.businessweek.com/technology/content/sep2008/tc2008095_942690.htm

³ Joe Weinman, "Mathematical Proof of the Inevitability of Cloud Computing," <http://www.joeweinman.com/papers.htm>

⁴ Joe Weinman, "Time is Money: The Value of On-Demand," <http://www.joeweinman.com/papers.htm>

Smooth Operator: The Value of Demand Aggregation

Any business is likely to have variation in demand, due to factors ranging from instantaneous fluctuations in user demand to multi-year business cycles and macroeconomic drivers. An individual customer who attempts to use dedicated capacity to meet this variation is unlikely to have the correct amount of capacity. As shown on the left, below, using a build-to-average approach means either that there are excess resources during periods of low demand, or insufficient resources and thus unserved demand during periods of high demand. On the other hand, as shown on the right, below, a build-to-peak strategy eliminates unserved demand, but increases costs associated with excess resources. Either way, “do-it-yourself” can be costly.

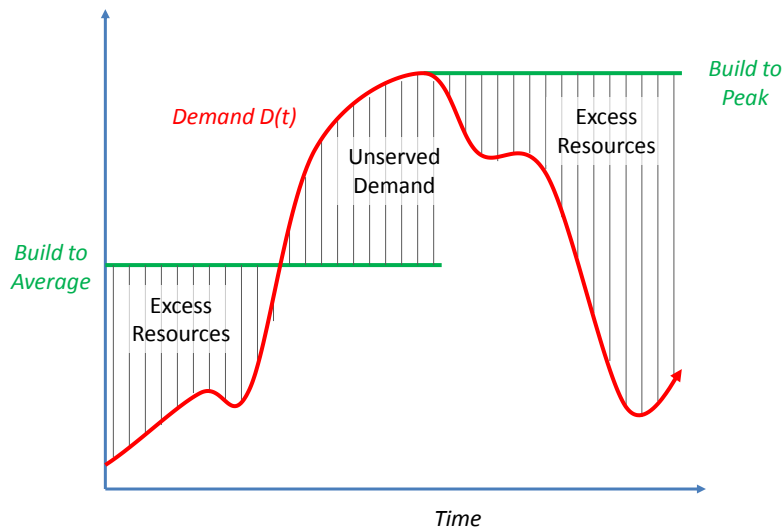


FIGURE 1: The Challenge of Rightsizing Capacity

A better plan may be to leverage a service provider. However, if all the service provider did was to implement private, walled-off fixed capacity to meet the same varying demand, the service provider faces the same challenges of either too little or too much capacity. The economics of the situation then become a tradeoff between the need of the operator to turn a fair profit vs. economies of scale, competencies, or unique intellectual property or tooling, say in resource management.

However, if the service provider, instead of implementing partitioned, fixed capacity, aggregates, or combines, demand from multiple users, real economic value can be generated. If one customer's peaks correspond to another customer's troughs, less total capacity is required. If the gains from such sharing outweigh the margin, overhead, transaction costs, and multi-tenancy management tooling and labor costs that service providers must add, it can become a win-win solution: the enterprises save money, while the service providers make it. If scale economies, learning curve effects, unique IP and the like are also added to the mix, a very

Smooth Operator: The Value of Demand Aggregation

compelling value proposition can arise with quantitative justification in addition to any strategic benefits, such as agility and core vs. context focus.

We will explore the situations in which such win-wins can arise, as well as characterize the nature of the gains. We also will formally prove my 3rd Law of Cloudonomics: *The Peak of the Sum is Never Greater than the Sum of the Peaks*, as well as my 4th Law of Cloudonomics: *Aggregate Demand is Smoother than Individual*. This also implies that operator utilization is higher and thus economics are advantaged.

Finally, we will explore what might be termed the “ \sqrt{n} effect,” where under a number of different circumstances and objectives, there appears to be a reduction in important cost metrics by \sqrt{n} whenever demand from n customers is aggregated. The root cause appears to be the fact that the variance σ_+^2 of a sum of n independent variables is the sum of the variances, i.e., $\sigma_+^2 = n \times \sigma^2$, and therefore the standard deviation of such a sum, σ_+ , follows $\sigma_+ = \sqrt{n} \times \sigma$.

2. PRELIMINARIES

Let there be a set of n customers, $1, 2, \dots, n$. For each customer i , let the demand for that customer over time be represented by $D_i(t)$. We can then define the aggregate demand,

$$D^+(t) = \sum_{i=1}^n D_i(t)$$

We can also characterize some of the statistics of $D_i(t)$ in the usual way: E for expected value, μ for mean, σ^2 for variance, and σ for standard deviation. We will sometimes use A for average, as in $A_i = \mu(D_i(t))$, and P for peak, as in

$$P_i = \max_{-\infty < t < \infty} D_i(t)$$

P^+ will denote the maximum of $D^+(t)$, i.e.,

$$P^+ = \max_{-\infty < t < \infty} D^+(t)$$

We will primarily be interested in understanding the relationship between statistics of the individual demand functions D_1, D_2, D_3 , etc., vs. the aggregate demand D^+ , especially as these impact capacity requirements under different resourcing strategies, and, in particular, the costs of those strategies. For example, we can look at P^+ vs. $\sum_{i=1}^n P_i$, and the implications for capacity utilization, which is the ratio between A and P . The peak-to-average ratio tells us the “spikiness” of demand, and its inverse, the average-to-peak ratio, is equivalent to utilization when capacity is built to peak. In other words, if $R(t)$ is a function describing resources over time, a build-to-peak strategy sets $R(t) = P$, where P is a constant.

3. The Third Law of Clouconomics

As we will see later, to determine the value V that can be ascribed to aggregating, consolidating, or statistical multiplexing of demand into pooled resources, we need to value the cost of unserved demand due to insufficient capacity, which we can model as a unit cost c_d , weighted by its probability-adjusted expected value, as well as the cost of excess capacity: a unit cost c_r , also weighted accordingly. For example, in a given context, there might be a 1/3 chance of being undercapacity, a 1/3 chance of having the right capacity, and a 1/3 chance of being overcapacity. And, in the first case, the expected degree of undercapacity might be 5 units, each costing \$100; in the last case, the expectation might be 2 units, each at \$30.

Later, we will examine the cost implications of unserved demand, and for now just look at the reduced resource quantity or capacity requirement from a strategy of building to the peak of the aggregate demand P^+ versus a requirement of aggregating individual capacities each built to peak, of total size $\sum_{i=1}^n P_i$, times the cost of resources c_r , in other words:

$$V = c_r \times \left\{ \left(\sum_{i=1}^n P_i \right) - P^+ \right\}$$

Although c_r may have varying costs due to volume discounting, nonlinearities, shifting marketing conditions, depreciation, and the like, we will assume that it is a constant. For most of the analysis, then, we will focus on the behavior of the sum of the individual demands vs. the behavior of the aggregate.

We begin by proving my 3rd Law of Clouconomics: “*The Peak of the Sum is Never Greater than the Sum of the Peaks.*” The proof is perhaps overly rigorous, but formalizes an important notion:

Proposition 1: If the demand for n customers $1, 2, \dots, n$ is defined by $D_i(t)$, with $P_i = \max_{-\infty < t < \infty} D_i(t)$, then $P^+ \leq \sum_{i=1}^n P_i$.

Proof: By counterexample. Suppose $P^+ > \sum_{i=1}^n P_i$. Then there is at least one time \hat{t} where $D^+(\hat{t}) = P^+$, and thus

$$D^+(\hat{t}) = P^+ > \sum_{i=1}^n P_i$$

But since

$$D^+(t) = \sum_{i=1}^n D_i(t)$$

Smooth Operator: The Value of Demand Aggregation

We also know that

$$D^+(\hat{t}) = \sum_{i=1}^n D_i(\hat{t})$$

However, by the definition of P_i ,

$$P_i = \max_{-\infty < t < \infty} D_i(t)$$

And the basic property of the maximum of a function, which is that

$$\forall x, f(x) \leq \max f(x)$$

We know that

$$\forall t, D_i(t) \leq \max D_i(t) = P_i$$

So it must be the case that for all i ,

$$D_i(\hat{t}) \leq \max D_i(\hat{t}) = P_i$$

But

$$D^+(\hat{t}) = \sum_{i=1}^n D_i(\hat{t}) \leq \sum_{i=1}^n \max D_i(\hat{t}) = \sum_{i=1}^n P_i$$

Clearly it cannot be the case that both

$$D^+(\hat{t}) = P^+ > \sum_{i=1}^n P_i$$

And that

$$D^+(\hat{t}) = P^+ \leq \sum_{i=1}^n P_i$$

So there cannot be such a \hat{t} , therefore it cannot be the case that $P^+ > \sum_{i=1}^n P_i$. ■

This proof says that if the peak of the aggregated demand were higher than the sum of the individual peaks, there would have to at least one point in time where such a peak was reached. However, all of the summands at that point in time *can't* be large enough to allow the sum to reach such a peak.

What this means in practice is that the total capacity required for serving aggregated demand via common (i.e., shared, or pooled resources) is no greater than the total capacity required by building capacity to peak for each of the individual demands, and may be less, depending on the individual $D_i(t)$ demands, i.e., thus the value of aggregating demand $V \geq 0$.

This agrees with intuition. After all, we could consolidate capacity and create n virtual partitions, each sized to match P_i , and be no worse off than before.

4. A Service Provider's Utilization is No Worse Than Its Average Customer's

We now show that if the peak of the sums is not greater than the sum of the peaks, the flip side is that a service provider's utilization is not less than the average weighted utilization of its customers.

Let the utilization Z_i of a customer with demand $D_i(t)$ and fixed capacity built to peak P_i be $\mu(D_i(t))/P_i$, or more simply, $Z_i = A_i/P_i$. A customer with flat demand, where $A_i = \mu(D_i(t)) = P_i$ can then get extremely high utilization, namely 100%, whereas one with spiky demand will have lower utilization, as its utilization is the ratio of average to peak, and the larger the peak is relative to the average, the smaller the utilization gets.

We want to measure the total loss associated with overcapacity, or to put it another way, understand how utilization improves. We could *try* to define average utilization something like this:

$$Z^? = \frac{Z_1 + Z_2 + Z_3 + \dots + Z_n}{n}$$

But this won't really work. We can't just take the average of each customer's utilization, because then a customer fully utilizing, say, one server, would get averaged with say, a customer with a trillion servers not using any of them for an apparent average utilization of .5, which would overstate things *slightly* given that only one server in a trillion and one was ever used. We thus need to weight the utilization of each customer by how large the customer is, which would lead us to something like:

$$\bar{Z} = \frac{P_1 Z_1 + P_2 Z_2 + P_3 Z_3 + \dots + P_n Z_n}{P_1 + P_2 + P_3 + \dots + P_n}$$

But there is a simpler way to view this, by substituting $Z_i = A_i/P_i$, we simply have that

$$\bar{Z} = \frac{A_1 + A_2 + A_3 + \dots + A_n}{P_1 + P_2 + P_3 + \dots + P_n}$$

Using a summation symbol, instead, we can just rewrite this as:

$$\bar{Z} = \frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n P_i}$$

We also can denote the utilization of the service provider as $Z^+ = A^+/P^+$.

Proposition 2: $Z^+ \geq \bar{Z}$.

Proof: From Proposition 1, we know that $P^+ \leq \sum_{i=1}^n P_i$. Since the mean of the sum is always equal to the sum of the means, we know that

$$A^+ = \sum_{i=1}^n A_i$$

Given that, for $a, x, y > 0$, if $x \geq y$ then $\frac{a}{x} \leq \frac{a}{y}$, we see that

$$\bar{Z} = \frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n P_i} = \frac{A^+}{\sum_{i=1}^n P_i} \leq \frac{A^+}{P^+} = Z^+ \blacksquare$$

What all this means is very important. The cost of resources either to “customers” attempting to do it themselves, or a service provider, is partly a function of things like volume discounts, overhead, etc. However, the *effective* cost of a resource is a function of how much excess capacity needs to be maintained. If the utilization is better at a service provider—potentially much, much better—then the effective cost of a resource is closer to the acquisition and ongoing management and maintenance costs of those resources. In other words, a service provider can potentially enjoy a lower cost structure by aggregating demand from multiple customers.

5. The Fourth Law of Cludonomics

We now prove the 4th Law of Cludonomics: “Aggregate Demand is Smoother than Individual,” and show that service providers can get higher utilization.

A well known statistic is the standard deviation. However, knowing that the standard deviation is, say, 50, doesn’t tell us much. Is that a deviation of 50 in a distribution that may have a range of 0 to 100, or a deviation of 50 in one that has a range of 0 to 1 billion? The first is very wide, the other is unnoticeable. It’s something like the difference between a pauper and a billionaire losing a fifty dollar bill.

To address this, a somewhat useful metric that is sometimes used to describe smoothness is the *coefficient of variation*. This is defined as the standard deviation σ divided by the mean μ , as long as the mean is non-zero. Like most statistics, it is imperfect, characterizing only one dimension of a random variable with perhaps complex behavior. However, it is useful because it is a proxy for relative variability that corresponds to utilization—which standard deviation or variance can’t tell us by itself.

Smooth Operator: The Value of Demand Aggregation

Proposition 3: If customers 1, 2, ... n have demand $D_1(t)$, $D_2(t)$, ... $D_n(t)$ respectively, and each $D_i(t)$ is independent but identically distributed with non-zero mean μ and standard deviation σ , i.e.,

$$\mu(D_1(t)) = \mu(D_2(t)) = \dots = \mu(D_n(t)) = \mu \neq 0$$

and

$$\sigma(D_1(t)) = \sigma(D_2(t)) = \dots = \sigma(D_n(t)) = \sigma$$

then each coefficient of variation $c_v(D_i(t)) = \frac{\sigma}{\mu}$ but the coefficient of variation of the aggregate demand $c_v(D^+(t)) = \frac{\sigma}{\mu\sqrt{n}}$.

Proof: The variance of any $D_i(t)$ is σ^2 , and its mean is μ . Since the sum of the variances is the variance of the sum, and the sum of the means is the mean of the sum, as long as the random variables are independent:

$$\sigma^2(D^+(t)) = \sigma^2\left(\sum_{i=1}^n D_i(t)\right)$$

Therefore, since the variance of the sum is the sum of the variances,

$$= \sum_{i=1}^n \sigma^2(D_i(t))$$

However, since each $D_i(t)$ has the same variance, this is merely

$$= n \times \sigma^2(D_i(t))$$

Therefore, since

$$\sigma^2(D^+(t)) = n \times \sigma^2(D_i(t))$$

We simply take the square root of both sides to get the standard deviation of $D^+(t)$, so

$$\begin{aligned}\sigma(D^+(t)) &= \sqrt{n \times \sigma^2(D_i(t))} \\ &= \sqrt{n} \times \sigma(D_i(t))\end{aligned}$$

Since we also know that the sum of the means is the mean of the sum, we have

$$\mu(D^+(t)) = \mu\left(\sum_{i=1}^n D_i(t)\right)$$

so

Smooth Operator: The Value of Demand Aggregation

$$\begin{aligned}\mu(D^+(t)) &= \sum_{i=1}^n \mu(D_i(t)) \\ &= n \times \mu(D_i(t))\end{aligned}$$

The coefficient of variation of the aggregate demand is then

$$\begin{aligned}c_v(D^+(t)) &= \frac{\sigma(D^+(t))}{\mu(D^+(t))} \\ &= \frac{\sqrt{n} \times \sigma(D_i(t))}{n \times \mu(D_i(t))} \\ &= \frac{\sigma}{\sqrt{n} \times \mu} \blacksquare\end{aligned}$$

In short, as we aggregate independent demand with a high degree of variability, the overall variability relative to the total tends to decrease, meaning lower peak requirements and thus higher average utilization. Since the reduction in this case is proportional to \sqrt{n} , it means that it always pays to get bigger, i.e., become a larger service provider pooling increasing demand, but there are “decreasing returns to scale.” The benefit of aggregating sixteen customers is about as high as also aggregating the next forty-eight (since $\sqrt{16} = \frac{1}{2}\sqrt{64}$). Higher average utilization means lower unit costs per “utilization-weighted” resource.

Letting μ and σ be identical across all the independent demands that are aggregated to access pooled resources does not mean that all the functions have an identical shape. For example, a normal distribution and a uniform distribution can have the same μ and σ , but represent different curves. This is important, because it means that adding together a broad range of different types of demand, e.g., seasonal / cyclical, spiky for Mother’s Day, spiky for the holidays, 9-5 workdays, etc., will still result in this smoothing.

Suppose the means are identical but the standard deviations are *not*? Then all we can know is that the pooling will result in a lower coefficient of variation than the coefficient of variation of the demand with the highest variance, as we show in the next proof.

We will use the terminology $c_v(\{\sigma_1, \sigma_2, \dots, \sigma_n\}, \mu)$ to refer to the coefficient of variation of n random variables with possibly varying standard deviations σ_i , and common mean $\mu \neq 0$.

Proposition 4: If customers 1, 2, ..., n , $n \geq 2$ have demand $D_1(t)$, $D_2(t)$, ..., $D_n(t)$ respectively, and each $D_i(t)$ is independent but with the same non-zero mean μ and without loss of generality, non-decreasing standard deviation σ_i , such that

$$\mu(D_1(t)) = \mu(D_2(t)) = \dots = \mu(D_n(t)) = \mu \neq 0$$

and

$$\sigma(D_1(t)) = \sigma_1 \leq \sigma(D_2(t)) = \sigma_2 \leq \dots \leq \sigma(D_n(t)) = \sigma_n$$

then the coefficient of variation $c_v(D_i(t)) = \frac{\sigma_i}{\mu}$ but the coefficient of variation of the aggregate demand $c_v(D^+(t)) < c_v(D_n(t))$.

Proof: Consider two sets of customers with nearly identical demand, but where the standard deviation of customer i is changed from say, a to b , where $a \leq b$.

We know $c_v(\{\sigma_1, \sigma_2, \dots, \sigma_{i-1}, a, \sigma_{i+1}, \dots, \sigma_n\}, \mu) \leq c_v(\{\sigma_1, \sigma_2, \dots, \sigma_{i-1}, b, \sigma_{i+1}, \dots, \sigma_n\}, \mu)$, since $a \leq b$, by the definition of coefficient of variation, and since standard deviation is the positive square root of variance and thus $a, b \geq 0$, so

$$c_v(\{\sigma_1, \sigma_2, \dots, \sigma_{i-1}, a, \sigma_{i+1}, \dots, \sigma_n\}, \mu) = \frac{\sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_{i-1}^2 + a^2 + \sigma_{i+1}^2 + \dots + \sigma_n^2}}{n \times \mu}$$

whereas

$$c_v(\{\sigma_1, \sigma_2, \dots, \sigma_{i-1}, b, \sigma_{i+1}, \dots, \sigma_n\}, \mu) = \frac{\sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_{i-1}^2 + b^2 + \sigma_{i+1}^2 + \dots + \sigma_n^2}}{n \times \mu}$$

In other words, increasing the size of one term and leaving everything else the same can't help but make the overall expression larger.

By the premise,

$$c_{v,1}(D^+(t)) = c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_{n-3}, \sigma_{n-2}, \sigma_{n-1}, \sigma_n\}, \mu)$$

The σ_i 's are ordered, hence we can replace any σ_i that is less than σ_n , observing in turn that:

$$c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_{n-3}, \sigma_{n-2}, \sigma_{n-1}, \sigma_n\}, \mu) \leq c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_{n-3}, \sigma_{n-2}, \sigma_n, \sigma_n\}, \mu)$$

and then that

$$c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_{n-3}, \sigma_{n-2}, \sigma_n, \sigma_n\}, \mu) \leq c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_{n-3}, \sigma_n, \sigma_n, \sigma_n\}, \mu)$$

and then that

$$c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_{n-3}, \sigma_n, \sigma_n, \sigma_n\}, \mu) \leq c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_n, \sigma_n, \sigma_n, \sigma_n\}, \mu)$$

and then that

$$c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_n, \sigma_n, \sigma_n, \sigma_n\}, \mu) \leq c_v(\{\sigma_1, \dots, \sigma_n, \sigma_n, \sigma_n, \sigma_n, \sigma_n\}, \mu)$$

and so forth. So by induction, we see that

$$c_v(\{\sigma_1, \dots, \sigma_{n-4}, \sigma_{n-3}, \sigma_{n-2}, \sigma_{n-1}, \sigma_n\}, \mu) \leq c_v(\{\sigma_n, \dots, \sigma_n, \sigma_n, \sigma_n, \sigma_n\}, \mu)$$

Smooth Operator: The Value of Demand Aggregation

Or replacing the left side with the coefficient of variation of the aggregate demand,

$$c_v(D^+(t)) \leq c_v(\{\sigma_n, \dots, \sigma_n, \sigma_n, \sigma_n, \sigma_n, \sigma_n\}, \mu)$$

From proposition 3 we know that

$$c_v(\{\sigma_n, \dots, \sigma_n, \sigma_n, \sigma_n, \sigma_n, \sigma_n\}, \mu) = \frac{\sigma_n}{\sqrt{n} \times \mu}$$

And, since $n \geq 2$, we also know that

$$\frac{\sigma_n}{\sqrt{n} \times \mu} < \frac{\sigma_n}{\mu} = c_v(D_n(t))$$

Putting this all together, we see that

$$c_v(D^+(t)) \leq c_v(\{\sigma_n, \dots, \sigma_n, \sigma_n, \sigma_n, \sigma_n, \sigma_n\}, \mu) = \frac{\sigma_n}{\mu\sqrt{n}} < \frac{\sigma_n}{\mu} = c_v(D_n(t)) \blacksquare$$

To put this more colloquially, when customers are about the same size, the aggregate demand is always “smoother” than the spikiest customers.

In the extreme, consider two customers, one with flat demand and one with spiky demand. What this says is that the one with flat demand won't gain anything, but a service provider that serves both will be able to generate a net benefit to the spiky demand customer. Depending on how prices are allocated, a customer with flat demand actually may lose something, in effect, helping to defray the costs associated with the untamed variability of the spiky one.

6. Normally Distributed Demand

The problem with trying to understand the behavior of peaks as we have been lies in the amplitude dimension as well as the time dimension. Consider a normally distributed random variable with a non-zero standard deviation. No matter how small that standard deviation is, there is no positive or negative upper bound, so the peak is actually infinity. The other issue is that since we assume that time is continuous, we have an infinite number of time intervals in any finite or infinite non-zero interval, and within these intervals, at any given instant, any random demand may have any legal value. This means that, while we expect that some customer demands will be high while others will be low, letting the law of large numbers have the demands cancel each other out, there is no reason why all of the customer demands can't be arbitrarily “close” to peak at any given time—even though the probability may be infinitesimally low, it is still non-zero.

Smooth Operator: The Value of Demand Aggregation

Suppose we assume that the demand functions' distributions are Gaussian (*aka*, normal, or bell-shaped), and rather than provide infinite capacity, we merely try to ensure that there is sufficient capacity to handle demand “most of the time.”

Let us assume that each of the demand functions is normally distributed, with mean μ and standard deviation σ .

We know that roughly 68% of values drawn from a normal distribution will be within one standard deviation from the mean. Roughly 97% of values will be within two standard deviations, roughly 99.7% will be within three standard deviations, 99.99% within four standard deviations, roughly 99.9999% within five standard deviations, and almost nine nines, or 99.9999998% will be within six standard deviations (the famous “six sigma”).

Suppose that rather than *absolutely* ensuring sufficient capacity (which would require infinite capacity for any unbounded distribution such as normal or exponential), we merely wanted to ensure that there was sufficient capacity, “most of the time,” and asked what the trade-off was between each individual customer providing such availability assurance vs. a service provider ensuring it for a group of customers whose demand had been aggregated and served by a set of pooled resources? We formalize “most of the time” by introducing $R_q(D(t))$ as the minimum fixed capacity that ensures that $P(x \leq R_q) = q$, $x \in D(t)$. In other words, if $q = .9999$, R_q is the capacity required to ensure that 99.99% of the time we will have sufficient capacity to meet the demand.

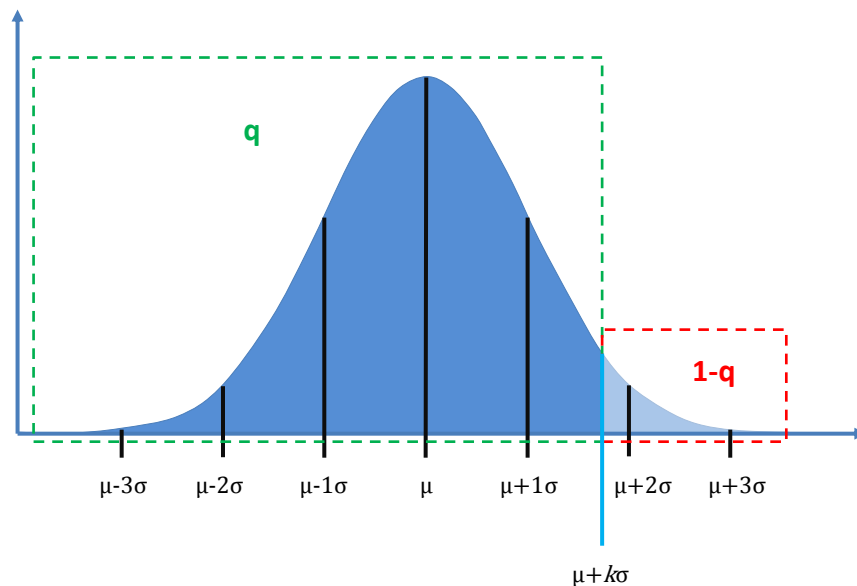


FIGURE 2: Ensuring Availability of Sufficient Capacity “Most of the Time”

Smooth Operator: The Value of Demand Aggregation

For this analysis we will assume that $q > .5$, a reasonable assumption, i.e., that we want to serve our customers at least half the time. As shown in the diagram above, there is a point $R_q = \mu + k\sigma$, such that q is the area of the normal density function where $P(x \leq R_q) = q$, and conversely, where $P(x > R_q) = 1 - q$.

Since $D(t)$ is normally distributed, this can be calculated by recollecting that not only are there calculated results such as that the area within the range $\mu \pm 1\sigma$ is $\sim .682689$, but also that the area in the range $\mu \pm k\sigma$ is $\text{erf}\left(\frac{k}{\sqrt{2}}\right)$, where erf is the “error function,” defined by

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Note that this is the area where $P(\mu - k\sigma \leq x \leq \mu + k\sigma)$, rather than the area of $P(x \leq \mu + k\sigma)$. However, since the normal distribution is symmetric, we can find k easily enough from the above equalities by slicing $1 - q$ off not only the high end but also the low end, and thus finding the k where $\text{erf}\left(\frac{k}{\sqrt{2}}\right) = 1 - (2 \times (1 - q))$.

However, it turns out that we don't even need to be able to exactly calculate k from q , because we can demonstrate the value of demand aggregation much more simply from the information we have, as the next proposition shows.

Proposition 5: Let customers $1, 2, \dots, n$ have independent demand $D_1(t), D_2(t), \dots, D_n(t)$ respectively, where each $D_i(t)$ is normally distributed with the same mean μ and standard deviation σ , i.e.,

$$\mu(D_1(t)) = \mu(D_2(t)) = \dots = \mu(D_n(t)) = \mu$$

and

$$\sigma(D_1(t)) = \sigma(D_2(t)) = \dots = \sigma(D_n(t)) = \sigma$$

As always, let the aggregate demand be

$$D^+(t) = \sum_{i=1}^n D_i(t)$$

If $n > 1$, then

$$\sum_{i=1}^n R_q(D_i(t)) > R_q(D^+(t)) = R_q\left(\sum_{i=1}^n D_i(t)\right)$$

Proof: To begin with, we need to recall that when we sum independent random variables that are normally distributed, the result is also normally distributed. This is referred to as the “reproductive property.” Moreover, the mean of the sum is the sum of the means:

Smooth Operator: The Value of Demand Aggregation

$$\mu(D^+(t)) = \sum_{i=1}^n \mu(D_i(t)) = n \times \mu(D_i(t)) = n \times \mu$$

And, since they are independent, the variance of the sum is the sum of the variances:

$$\sigma^2(D^+(t)) = \sum_{i=1}^n \sigma^2(D_i(t))$$

Since the standard deviation of each $D_i(t)$ is σ , the variance of each $D_i(t)$ is σ^2 . After summing n demands characterized by $D_i(t)$ all with variance σ^2 , the variance of $D^+(t)$ is $n \times \sigma^2$, so the standard deviation of $D^+(t)$, which we'll refer to as σ_+ , is $\sqrt{n} \times \sigma$.

Let k be the value such that, as defined above, $\text{erf}\left(\frac{k}{\sqrt{2}}\right) = 1 - (2 \times (1 - q))$ and thus fixed capacity set to $R_q(D(t)) = \mu + k\sigma$ will be sufficient to meet demand $D(t)$ with probability q .

Since $n > 1$, we know that $\sqrt{n} < n$.

But

$$\sqrt{n} < n$$

implying that

$$\sqrt{n} \times \sigma < n \times \sigma$$

Therefore

$$\sigma_+ < n \times \sigma$$

Multiplying both sides by k

$$k \times \sigma_+ < k \times n \times \sigma$$

As we know, the sum of the means is the mean of the sums, so

$$\mu(D^+(t)) = \sum_{i=1}^n \mu(D_i(t)) = n \times \mu(D_i(t)) = n \times \mu$$

Which if we add to the prior equation leads to

$$\mu(D^+(t)) + k \times \sigma_+ < n \times \mu + k \times n \times \sigma = n \times (\mu + k \times \sigma)$$

These $\mu(f(x)) + k \times \sigma(f(x))$ terms should look familiar, as they are just $R_q(f(x))$, and so substituting, we see that we have shown that

$$R_q(D^+(t)) < \sum_{i=1}^n R_q(D_i(t)) \blacksquare$$

In other words, the sum of the capacities required for each individual customer to be “pretty sure” they have sufficient capacity, is greater than the capacity required for a service provider to be just as “pretty sure” that it has enough shared capacity to meet the aggregate demands of those same customers. And this is true no matter how close q gets to 1.

7. The Costs of Unused Resources and Unserved Demand

In *Time is Money*, I introduced two kinds of costs. The first is the cost of resources, c_r , which we introduced above. The other was the cost of unserved demand c_d . The reader is referred to that paper for additional detail. Briefly however, if capacity is fixed, but demand varies, at any given time the demand may be lower or greater than the amount of available capacity. If demand is lower than capacity, there are costs associated with resources that are idle. If demand is higher than available capacity, there can be costs associated with not serving that demand, e.g., lost revenue, lost productivity, dissatisfied customers, and so forth. Generally, it is safe to assume that $c_d > c_r$, that is, that when a resource is deployed, it is for some objective that has value greater than the cost of deploying the resource. As a rule, this means that, for a fixed capacity strategy, there is a breakeven point F that optimizes the balance of costs between the two, based on the expected value at that resource level of being overprovisioned, and thus incurring the resource cost times the expected amount of overprovisioning, and the expected value of being underprovisioned, and thus incurring the unserved demand cost times the expected amount of unserved demand.

For example, consider a uniformly distributed demand on the interval $[0, P]$. If $c_r = c_d$, then selecting $F = \frac{P}{2}$ minimizes the expected value of the costs of unserved demand and excess capacity. If, on the other hand, $c_r \ll c_d$, to take an extreme case, that $c_r = 0$ and $c_d = \infty$, then a fixed capacity $F = P$ is appropriate. After all, in such a scenario, resources are free, but the cost of even a trivial (but non-zero) amount of unserved demand is infinite, so it would be smart to ensure that there are sufficient resources at all times.

Between these two extremes, F climbs from $\frac{P}{2}$ to P as the ratio between c_d and c_r increases, or to restate Proposition 6 from *Time is Money* regarding selecting the right balance:

Proposition 6: Let $D(t)$ be uniformly distributed on $[0, P]$, with a cost of resources of c_r and a cost of unserved demand of c_d . Then the optimal fixed capacity F for uniformly distributed demand is $(P \times c_d) / (c_r + c_d)$.

Smooth Operator: The Value of Demand Aggregation

This formula applies to uniformly distributed demand. For normally distributed demand, we are trying to find the level F where the cost is minimized. Given μ , σ , c_d , and c_r , we analytically or at least numerically can find this value based on minimizing the total cost

$$p(D(t) < F) \times c_r \times r + p(D(t) > F) \times c_d \times d$$

where d is the expected value of the amount of unserved demand, if there is undercapacity, and r is the expected value of the amount of unused resources, if there is overcapacity. As can be seen from *Figure 3* below, $F - r$ is the expected value of the dark blue area, while $F + d$ is the expected value of the light blue one.

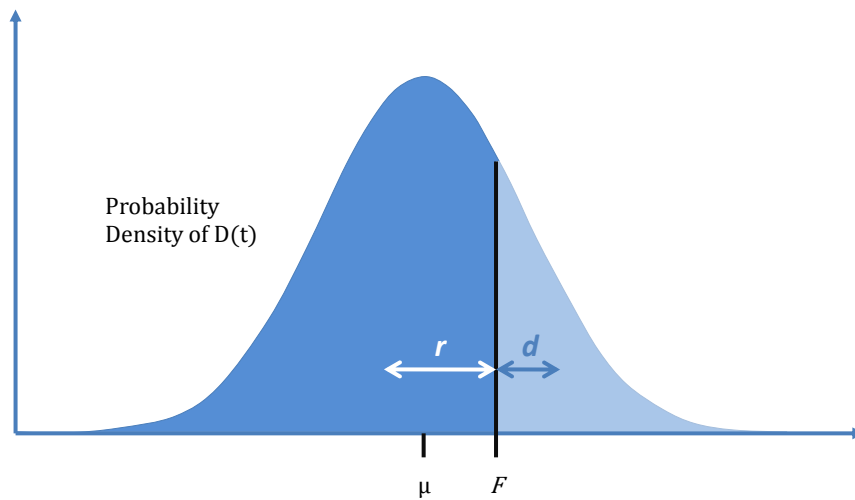


FIGURE 3: The Optimal Fixed Capacity F When $c_d > c_r$

We can observe that $F - r < \mu$, since the expected value of D is μ , so the expected value of D excluding all values higher than F must be lower than that. Although we can expect that $c_r < c_d$, finding the value of F that minimizes cost would appear to be somewhat of a challenge, so next we will examine the simpler case where $c_r = c_d$.

Consider having fixed capacity at a point F exactly at the mean μ of a normally distributed demand $D(t)$. When F is so set, given the nature of the distribution, it is certainly possible that at any given instant, i.e., for any sample drawn from this distribution, the *actual* demand will be less than or greater than this fixed capacity F .

The expected value of this difference is *not* the standard deviation, but the “mean absolute deviation,” also known as the “average absolute deviation,” or the somewhat ambiguous “MAD”

Smooth Operator: The Value of Demand Aggregation

(which also can signify *median* absolute deviation). The mean absolute deviation is the mean of the absolute deviation from the mean, i.e., $E(|X - E(X)|)$.

When $c_r = c_d$, the point F will be exactly at the mean μ . Perhaps interestingly, this is not generally the case for all distributions, but it is known that the point that minimizes the average absolute deviation is the median, and in the case of a normally distributed random variable, the mean is the same as the median.

For any distribution, the mean absolute deviation is always less than or equal to the standard deviation. Specifically, however, for a normally distributed random variable, the relationship between the mean absolute deviation and the standard deviation is that $MAD(X) = \sigma \sqrt{\frac{2}{\pi}}$, in other words, $MAD(X)$ is about 4/5ths of σ .

We now show that a service provider can reach a lower optimal total cost point than the cost of its individual customers.

Proposition 7: Let customers 1, 2, ... n have independent, identically distributed demand $D_1(t)$, $D_2(t)$, ... $D_n(t)$ respectively, where each $D_i(t)$ is normally distributed with the same mean μ and standard deviation σ , i.e.,

$$\mu(D_1(t)) = \mu(D_2(t)) = \dots = \mu(D_n(t)) = \mu$$

and

$$\sigma(D_1(t)) = \sigma(D_2(t)) = \dots = \sigma(D_n(t)) = \sigma$$

As always, let the aggregate demand be

$$D^+(t) = \sum_{i=1}^n D_i(t)$$

Let the cost of resources equal the cost of unserved demand, i.e., $c_r = c_d$.

If $n > 1$, then the lowest penalty cost due to excess resources and unserved demand for the operator is only $\frac{1}{\sqrt{n}}$ of the lowest penalty cost when capacity is not aggregated.

Proof: For each individual demand $D_i(t)$, the minimal cost occurs when the MAD is minimized, which occurs when $F = \mu$, since $c_r = c_d$. If served separately, there is a 50% chance that the actual demand will be lower than the mean, and a 50% chance that it will be higher, since the normal probability density function is symmetric about the mean. The expected value of the difference is the mean absolute deviation from the mean, which is $\sigma \sqrt{\frac{2}{\pi}}$. Therefore the expected value of the lost cost C_i when the capacity to serve $D_i(t)$ is set to F is

Smooth Operator: The Value of Demand Aggregation

$$C_i = \frac{1}{2} c_r \times \sigma \sqrt{\frac{2}{\pi}} + \frac{1}{2} c_d \times \sigma \sqrt{\frac{2}{\pi}}$$

Given that this holds for each individual demand, and since $c_r = c_d$, the total cost of solving each individual demand separately, using the best possible fixed capacity, is:

$$n \times C_i = n \times c_r \times \sigma \sqrt{\frac{2}{\pi}}$$

If instead we were to aggregate the demand via a service provider, we set the aggregate to be

$$D^+(t) = \sum_{i=1}^n D_i(t)$$

Since the sum of the variances is the variance of the sum, and since $D^+(t)$ is normally distributed, we know that $\sigma^2(D^+(t)) = n \times \sigma^2(D_i(t))$, thus, using the notation σ^+ to stand for $\sigma(D^+(t))$, we know that $\sigma^+ = \sqrt{n} \times \sigma$. But similar relationships hold as before: the lowest cost fixed capacity is at $F^+ = \mu^+ = n \times \mu$, and there, the expected value of the difference of the actual demand from the fixed capacity F^+ is $\sigma^+ \sqrt{\frac{2}{\pi}}$. Given this, in the aggregated case we know that the penalty cost associated with excess resources or demand is

$$C^+ = \frac{1}{2} c_r \times \sigma^+ \sqrt{\frac{2}{\pi}} + \frac{1}{2} c_d \times \sigma^+ \sqrt{\frac{2}{\pi}} = c_r \times \sigma^+ \sqrt{\frac{2}{\pi}} = c_r \times \sqrt{n} \times \sigma \sqrt{\frac{2}{\pi}}$$

The relative lowest cost of the aggregated demand vs. the total cost the sum of the best individual costs for the unaggregated demand is then:

$$\frac{c_r \times \sqrt{n} \times \sigma \sqrt{\frac{2}{\pi}}}{n \times c_r \times \sigma \sqrt{\frac{2}{\pi}}}$$

Which is of course just $\frac{1}{\sqrt{n}}$. ■

It is important to put this number in context. We are not saying that the total cost of infrastructure is $\frac{1}{\sqrt{n}}$, or else we would have proved something obviously incorrect, namely that to serve an infinite number of customers with non-zero demand, the infrastructure costs zero. What we are saying is that if we size infrastructure for each demand based on the mean demand at $F = \mu$, the base cost for n customers will be $n \times \mu$, which is no different than if we sized infrastructure for the aggregate demand at the level of $F^+ = n \times \mu$. However, the “penalty

costs” associated with over or under capacity will be reduced as we increase the number of customers. Even if we were to ignore the cost of unused resources and consider the “right half” of the pdf where $D(t) > \mu$, we would still have this $\frac{1}{\sqrt{n}}$ reduction effect.

8. Constant Demand

Just because demand aggregation is not a bad idea doesn't necessarily mean that there is a benefit associated with it; it may be neutral:

Proposition 8: If all customer demand is constant, then $P^+ = \sum_{i=1}^n P_i$.

Proof: If for each customer, demand is constant, then let $D_i(t) = k_i$. Then,

$$\begin{aligned} P^+ &= \max_{-\infty < t < \infty} D^+(t) \\ &= \max_{-\infty < t < \infty} \sum_{i=1}^n D_i(t) \\ &= \max \sum_{i=1}^n k_i \\ &= \sum_{i=1}^n k_i \end{aligned}$$

Since $D_i(t) = k_i$, we know that $P_i = k_i$, so we can substitute, concluding that

$$P^+ = \sum_{i=1}^n P_i \blacksquare$$

By the way, the reverse is not true. The easiest counterexample demonstrating this is where $n - 1$ customers have constant demand, and one customer, say, customer i , has variable demand, i.e., $\exists t, A_i < D_i(t) \leq P_i$.

9. Correlated Demand

Constant demand is a specific case of the more general case of correlated demand. While perfectly correlated demand may not exactly occur in the real world, it can come very close. For example, consider U.S. physical and online retailers. They have different patterns of usage (see below), but all have very busy periods around the interval from Thanksgiving to Christmas. Moreover, they all have their busiest online day on “Cyber Monday.” Whether they built their own data centers or combined their IT into one single data center wouldn’t make a big difference in terms of statistical multiplexing benefits (although it might, say, in overhead costs). However, if any of them were to try to combine forces with say, a flower delivery company and a tax preparation firm, then the various peaks on or near Valentines Day (February), tax filing deadlines (April), Mothers Day (May), and Cyber Monday (end of November or early December), would not occur at the same time, consequently there would be benefits from statistical multiplexing.

For now, we look at what happens when multiple demand curves are identical except for a scaling factor.

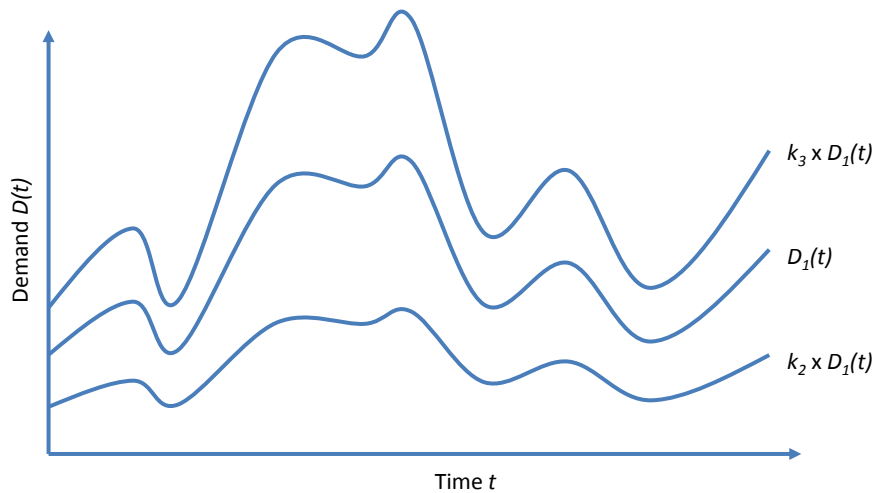


FIGURE 4: Demands $D_1(t)$, $\frac{1}{2} D_1(t)$, and $1.5 \times D_1(t)$,

Smooth Operator: The Value of Demand Aggregation

The illustration above shows three demands related by a scaling factor. Their individual statistics μ and σ are very different, but it turns out their ratio is a constant.

Proposition 9: Given n customers, let $D_1(t)$ represent the demand of the first customer, and for each $D_i(t)$, $2 \leq i \leq n$ let $D_i(t) = k_i \times D_1(t)$. Then $c_v(D^+(t)) = c_v(D_1(t)) = c_v(D_i(t))$, $2 \leq i \leq n$

Proof: Let $K = \sum_{i=1}^n k_i$. Then $D^+(t) = K \times D_1(t)$. Note that the variance of a constant C times a random variable X conveniently follows the rule $\sigma^2(CX) = C^2 \times \sigma^2(X)$, from which we can deduce that $\sigma(CX) = C \times \sigma(X)$. Also, the mean of a constant C times a random variable X is simply $\mu(CX) = C \times \mu(X)$. Using the definition of coefficient of variation, which for non-zero μ is $c_v(X) = \frac{\sigma(X)}{\mu(X)}$, we see that

$$\begin{aligned} c_v(D^+(t)) &= \frac{\sigma(D^+(t))}{\mu(D^+(t))} \\ &= \frac{\sigma(K \times D_1(t))}{\mu(K \times D_1(t))} \\ &= \frac{K \times \sigma(D_1(t))}{K \times \mu(D_1(t))} \\ &= \frac{\sigma(D_1(t))}{\mu(D_1(t))} \\ &= c_v(D_1(t)) \end{aligned}$$

Note that we can substitute any k_i for K in the above equalities, leading to the conclusion that $c_v(D^+(t)) = c_v(D_1(t)) = c_v(D_i(t))$, $2 \leq i \leq n$. ■

Simply put, aggregating any number of demand curves which are the same except for a scaling factor does not change the relative smoothness or spikiness, nor therefore, utilization.

The implication is that there are no utilization gains to be had by aggregating demand from similar industries. Having, say, all the tax preparation companies—who have peaks on April 15th—or all the retailers—who have peaks on Cyber Monday—aggregate their demand and serve it out of pooled resources will not lead to any net benefit in utilization (although it might in say, purchasing power, spreading overhead expenses across a wider base, or other economies of scale).

10. Peak Alignment

In the analysis above, we showed that if all the demands were identical except for a scaling factor, there was no benefit to aggregation. In fact, we can weaken the conditions even further. If all demand curves have at least one peak at the same time, even if they have nothing else in common, there won't be any utilization benefit to aggregation.

Proposition 10: Let $D_i(t)$, $1 \leq i \leq n$ represent demands as usual, where $P_i = \max(D_i(t))$ and $A_i = \mu(D_i(t))$. There exists at least one time \hat{t} where $\forall i, 1 \leq i \leq n, D_i(\hat{t}) = P_i$ **if and only if** $Z^+ = \bar{Z}$, i.e., the utilization Z^+ of resources needed to serve the aggregate demand D^+ equals the utilization \bar{Z} of the unaggregated demand.

Proof: We know that there is at least one point in time \hat{t} , where

$$D^+(\hat{t}) = D_1(\hat{t}) + D_2(\hat{t}) + \dots + D_n(\hat{t}) = P_1 + P_2 + \dots + P_n$$

It is clear that, in any scenario where there is an ordered set S , if $s \in S$, then $\max(S) \geq s$. For example, if we run into someone six feet tall at a party, we know that the tallest person at that party is at least six feet tall (and perhaps is named "Max").

So, if $S = \{D^+(t), -\infty \leq t \leq \infty\}$, and $s = D^+(\hat{t})$, we know that $s \in S$ and therefore that $P^+ = \max(D^+(t)) \geq D^+(\hat{t})$. Therefore we know that

$$P^+ \geq \sum_{i=1}^n P_i$$

But from Proposition 1 we know that

$$P^+ \leq \sum_{i=1}^n P_i$$

The only way this can happen is if

$$P^+ = \sum_{i=1}^n P_i$$

We have already seen that, using our oft-repeated rule that the mean of the sum is the sum of the means,

$$A^+ = \sum_{i=1}^n A_i$$

Smooth Operator: The Value of Demand Aggregation

Recall that average utilization for an environment with capacity built to peak is:

$$\bar{Z} = \frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n P_i}$$

Then the equivalence is clear, because replacing terms we have

$$\bar{Z} = \frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n P_i} = \frac{A^+}{P^+} = Z^+$$

The reverse is also true. If $\bar{Z} = Z^+$, since $A^+ = \sum_{i=1}^n A_i$, it must be the case that $P^+ = \sum_{i=1}^n P_i$. Proposition 1 tells us that $P^+ \leq \sum_{i=1}^n P_i$, therefore, per the logic in the counterexample of the proof of Proposition 1 and above, there must be a \hat{t} as we've defined it. ■

This insight is *key*. No matter what else happens off-peak, if the peaks align there is no benefit to aggregating demand. For example, consider the online “page view” data in *Figure 4* below from Alexa.com for three (r)etailers. This shows the total number of web pages viewed at each web site.

The first retailer is known for online sales, and has healthy business throughout the year, with roughly a 1.5:1 peak-to-average ratio. The second is a large bricks and mortar as well as online retailer, with roughly a 3:1 peak-to-average ratio. And the last is an even larger (r)etailer, with roughly a 5:1 peak-to-average ratio.

Means, standard deviations, and peak-to-average ratios differ, but from the perspective of determining whether there is value in aggregating this demand, the only factor to look at is that their peaks all occur on the same day, Cyber Monday, and possibly even in the same hour.

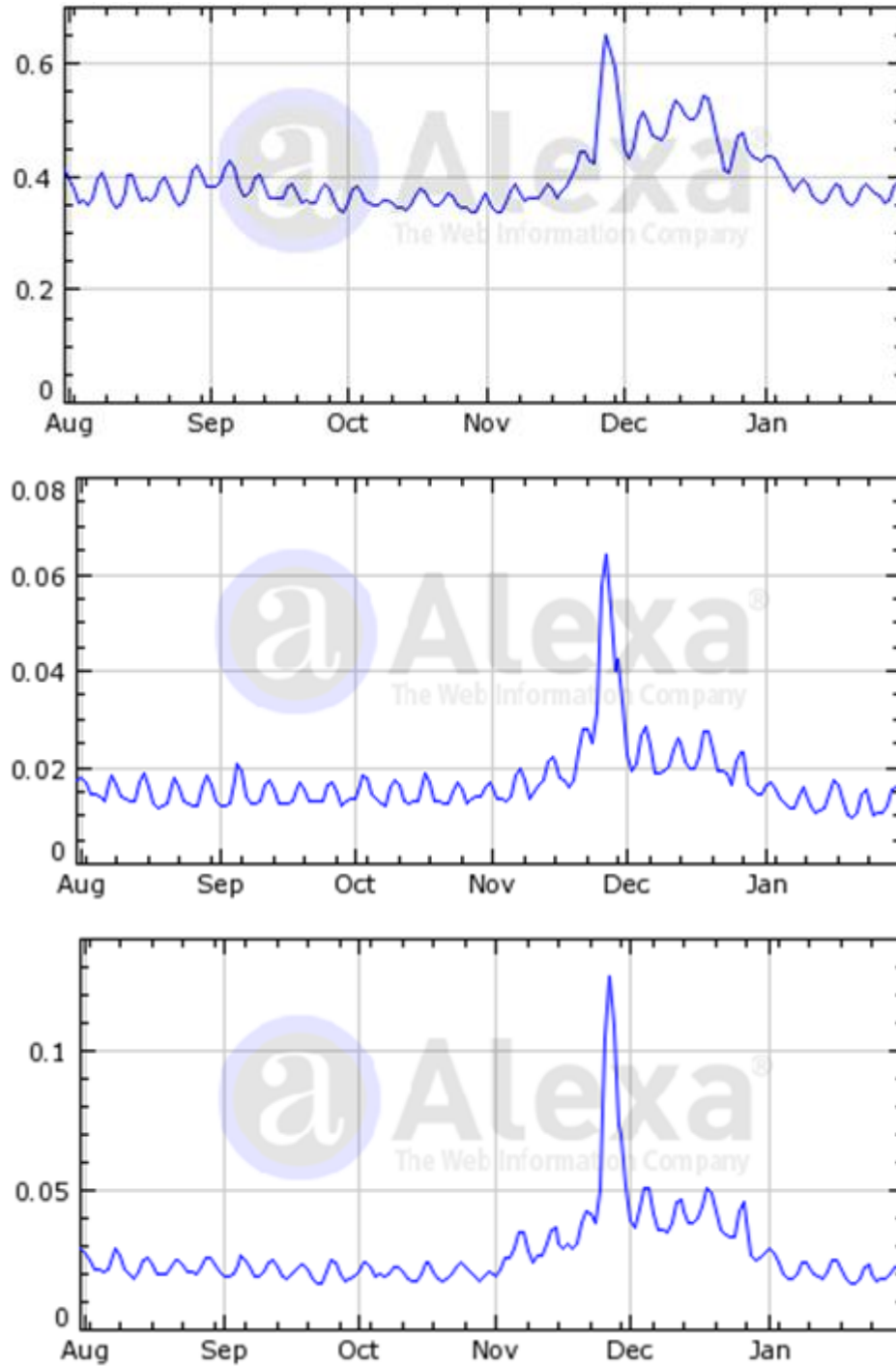


FIGURE 5: Different (R)etailers with Different Demands but Synchronized Peaks

11. In the Limit

In the limit, as the number of customers grows to infinity and the number of samples from a distribution also does, what happens? Briefly, for most distributions, the more samples there are, the more likely it is that the peak of the sample population will be arbitrarily close to the peak of the distribution. And, the more customers there are, the more likely it is that the sum of those demands is close to the sum of the means.

Proposition 11: Let customers $1, 2, \dots, n$ have demand $D_1(t), D_2(t), \dots, D_n(t)$ respectively, where each of the $D_i(t)$ are independent but identically distributed with non-zero mean μ and standard deviation σ . As always, let $D^+(t) = \sum_{i=1}^n D_i(t)$. Then, $\frac{D^+(t)}{n}$ converges⁵ to μ as $n \rightarrow \infty$.

Proof: We rely on a form of the *Law of Large Numbers* based on Chebyshev's Inequality. We know that:

$$P \left[\left| \frac{D^+(t)}{n} - \mu \right| \geq \epsilon \right] \leq \frac{\sigma^2}{n\epsilon^2}$$

In other words, the probability that the sum of the $D_i(t)$'s divided by n —which since they are identically distributed and independent may be considered as average result of n experiment trials—differs by more than an amount of ϵ from the mean, will be less than $\frac{\sigma^2}{n\epsilon^2}$. However, as $n \rightarrow \infty$, it is clear that $\frac{\sigma^2}{n\epsilon^2} \rightarrow 0$, i.e., no matter how small we pick ϵ , the probability that $\frac{D^+(t)}{n}$ differs from μ by more than that grows vanishingly small. ■

A more detailed proof of the following proposition has been described by Thayer Watkins of San Jose' State University,⁶ but this shorter proof should do for our purposes here:

Proposition 12: Let X be a random variable with maximum $\max(X)$, and S_n be a sample taken from X of size n . Then, as $n \rightarrow \infty$, $E(\max(S_n)) \rightarrow \max(X)$.

Proof: Let $\min(X) \leq v < \max(X)$, and let the cumulative distribution function of X be $P(X)$, i.e., $P(x) = p(X \leq x)$.

Let $S_n = \{s_1, s_2, \dots, s_n\}$. For the maximum of S_n to be less than or equal to v , all the s_i 's must be less than or equal to v . Since these are independent samples, we know

$$p(\max(S_n) \leq v) = p(s_1 \leq v) \times p(s_2 \leq v) \times \dots \times p(s_n \leq v)$$

⁵ "Converges in probability" and "converges in distribution"

⁶ Watkins, Thayer, "The Expected Value of Sample Maximums as a Function of Sample Size," <http://www.sjsu.edu/faculty/watkins/samplemax.htm>

Smooth Operator: The Value of Demand Aggregation

$$= P(v) \times P(v) \times \dots \times P(v) \text{ } n \text{ times}$$

Therefore

$$p(\max(S_n) \leq v) = P(v)^n$$

Since $v < \max(X)$, $P(v) < 1$. Therefore,

$$\lim_{n \rightarrow \infty} (P(v))^n = 0$$

Or, more precisely, that, regardless of how large v is (as long as $v < \max(X)$),

$$\lim_{n \rightarrow \infty} p(\max(S_n) > v) = 1 \blacksquare$$

Let us substitute “ $D_i(t)$ ” for “ X ”. Then, the practical implication of this is that if a single customer builds out their own resources with the objective of having sufficient fixed capacity to meet the needs of their own varying demand, it needs to plan for peak, since sooner or later that peak will arrive. As global news clearly demonstrates, events such as “100 year” floods aren’t a moving target, always 100 years away, but outliers that may be happening as you read this sentence.

Figure 6 shows a Monte Carlo simulation run across 200 trials.

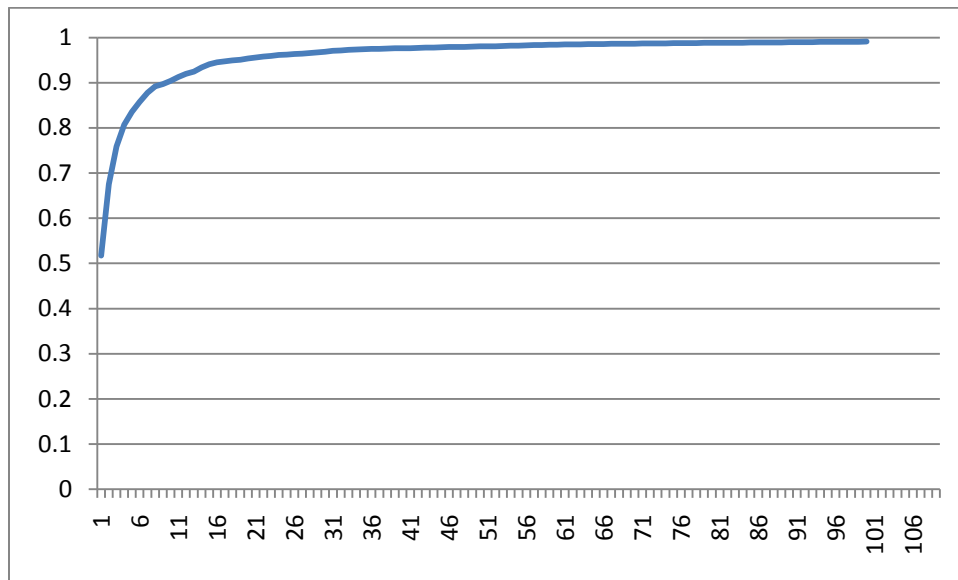


FIGURE 6: Monte Carlo Simulation of $E(\max(S_n))$, $n = 1$ to 100

It shows the average of $\max(S_n)$ when X is uniformly distributed—a proxy for the theoretical expected value as n grows from 1 to 100. This simulation can be conducted easily in a spreadsheet program by setting the upper left cell to “=RAND()”, the cell below it to “=MAX(A1,RAND())”, the cell below it to “=MAX(A2,RAND())”, and so forth by cutting and pasting until row n , and then replicating that column by cutting and pasting as many times as the desired number of trials. Then, using a formula such as =AVERAGE(A1:GR1) to generate the average across each trial (column), the chart can be then generated.

Smooth Operator: The Value of Demand Aggregation

Combining these two propositions, let us consider what happens as both the number of customers increases and as the number of sampling periods increases. As the number of customers increases, the service provider can just build relatively closer to the mean demand. Consequently, a service provider, aggregating demand from more and more customers, can get higher and higher utilization. On the other hand, an individual customer can't build to mean, it has to build to peak if it wants to be *assured* of meeting any demand "event" from the underlying distribution. Therefore, the improvement from consolidating demands at a service provider is ultimately related to the peak to average ratio. If the demand is constant, this improvement is nil. If the demand is uniformly distributed, the improvement is 2 to 1. For demands resembling exponential distributions, the improvement can be even greater. I say "resembling" because exponential, normal, and similar distributions have infinite peaks, which don't *exactly* match conditions in the real world: we experience only finite peaks. An exponential distribution ($\lambda e^{-\lambda x}$) has a mean of λ^{-1} , but no bound on how large x can be (although as x gets larger, its probability drops to zero).

It is important to note that such a strategy may not always be cost-optimal, depending on the relative tradeoffs of the cost of resources vs. the cost of unserved demand, a topic I explored elsewhere,⁷ and we discussed above. However, as the cost of unserved demand increases relative to the cost of resources, the optimal point minimizing the penalty function gets closer to the peak. Therefore, we can combine the prior two propositions to arrive at their business impact.

Proposition 13: Let customers 1, 2, ... n have demand $D_1(t)$, $D_2(t)$, ... $D_n(t)$ respectively, where each of the $D_i(t)$ are independent but identically distributed with non-zero mean μ , standard deviation σ , and finite maximum $P = \max(D_i(t))$. Let $D^+(t) = \sum_{i=1}^n D_i(t)$. As $n \rightarrow \infty$, the ratio of the required capacity for a service provider that consolidates demand versus the required capacity if the demand is unconsolidated converges to $\frac{\mu}{P}$.

Proof (Outline): From *Proposition 12*, we know that the capacity required for *each* unconsolidated demand is arbitrarily close to P , if we want to ensure that we have sufficient capacity. Thus, the total capacity required in such a scenario is $n \times P$. However, from *Proposition 11*, for the aggregated demand $D^+(t) = \sum_{i=1}^n D_i(t)$, the required capacity converges to $\mu^+ = n \times \mu$. The relative required capacity then converges to $\frac{n \times \mu}{n \times P} = \frac{\mu}{P}$. ■

There are all kinds of caveats required for such a proof, regarding the implication of "converges" in a probabilistic sense and that the sum of means is only defined for a finite number of terms. However, what this means is quite significant...in effect, an individual company arguably needs to deploy capacity for its *peak* demand, but a service provider can, in effect, deploy (or perhaps, virtually allocate) capacity for that enterprise's *average* demand. For companies that have 4-to-1, or 10-to-1 or 25-to-1 peak to average ratios, this is quite an advantage, and is part of the driver of the 1st Law of Cloudonomics.

⁷ Joe Weinman, "Time is Money: The Value of 'On-Demand'",
http://www.JoeWeinman.com/Resources/Joe_Weinman_Time_Is_Money.pdf

This also provides an upper bound on improvement. We will never be able to do better than 100% utilization, thus the minimum capacity we need is $n \times \mu$. However, depending on the demand functions $D_i(t)$, we may not need an infinite number of customers to reach that level. For example, if one customer has constant (flat) demand $D_i(t) = k$, its peak equals its average equals k , and a “shared environment” with only itself as a customer will have 100% utilization. If there are two customers, where no matter how variable $D_1(t)$ is, we have $D_2(t) = k - D_1(t)$, that is, the second demand is the mirror image of the first, then it only takes two customers to reach 100% utilization in the shared environment, since $D_1(t) + D_2(t) = k = \mu_1 + \mu_2$.

12. Behavioral Cludonomics

In addition to the mathematical effects described here that create business value, there are also important considerations regarding human cognitive biases that relate to owned, dedicated resources vs. shared, pay-per-use resources. I’ve overviewed a few of them in a short post: “Lazy, Hazy, Crazy: The 10 Laws of Behavioral Cludonomics.”⁸ Here’s a brief recap, with some additional observations.

Need for Control and Autonomy — A basic human need is one of control over the environment. Lack of control can create a feeling of helplessness, and studies show that such a lack of control can actually lead to death, in animals as well as humans. Ownership, e.g., having your own car, or data center, increases the perception of control, whereas using a shared resource may not. Consider the frustration one may experience in waiting for a cab or finding out that a hotel room believed to be confirmed was a victim of overbooking. Before cloud, sharing of Business Continuity / Disaster Recovery sites sometimes led to issues, requiring “zip-code exclusivity” to prevent two victims of a natural disaster such as a flood or hurricane to both claim the same site. Rationally, a service provider may offer better availability than an owned data center does, and also better scalability, but this cognitive bias may impact perceptions and thus decisions.

Neglect of Probability – As we’ve seen for most of this paper, probability plays a key role in assessing the differences between individual and aggregated demand. However, without the formal analysis, most decision-makers can’t intuitively determine the relative trade-offs between approaches.

Illusion of Control – The need for control is so ingrained that rather than *accept* that one has limited or total lack of control over a situation and its outcomes, individuals will imagine that they *do* have control over an environment. For example, in one experiment it was shown that many presumably intelligent college students believe that they can learn to influence the result of flipping a coin. It may be true that an owned resource actually has greater controllability than a

⁸ Joe Weinman, “Lazy, Hazy, Crazy: The 10 Laws of Behavioral Cludonomics,” <http://gigaom.com/2010/06/06/lazy-hazy-crazy-the-10-laws-of-behavioral-cludonomics/>

Smooth Operator: The Value of Demand Aggregation

shared service, but the illusion of control may create the perception of an even greater difference than actually exists.

The Choice-Supportive Bias and the Confirmation Bias – People tend to believe their prior choices are good ones rather than admit otherwise. And, they tend to selectively filter information, paying attention to data and anecdotes that fit their worldview, and ignoring data that does not, to the point of believing, for example, that an objective analysis that disagrees with their belief system is biased. This can work both ways: someone that has always owned a car may not see the value in selling it and only using taxi or limousine services, but a city dweller who has never owned or driven a car may not see the need to ever change, either.

The Endowment Effect and Sunk Cost Fallacy — People value goods that they already own more than they would pay to acquire them. This may partly explain why the asking price for a house doesn't always match the offered price. Moreover, decision-makers rationally should exclude sunk costs and view options merely in terms of their differing “prospective costs” and benefits. Decisions to continue to own a data center (or many) versus shift to a service provider model may be influenced by the Endowment Effect.

Need for Status — People and other social primates have dominance hierarchies and self-images. Status is important, with functional Magnetic Resonance Imaging (fMRI) studies showing areas of the brain dedicated to processing status information, including from playing games. Owning assets and managing organizations associated with such assets may confer status to the leader. On the other hand, being seen as an innovator or change agent may also confer status.

There are numerous other cognitive biases, but this should give a sense that real-world decisions need to account for biases, heuristics, and deep-seated needs, by counterbalancing these with quantitative assessments.

13. Conclusion

We have examined how “smooth operators” can offer advantages relative to “spiky customers.” We've used a few standard distributions to illustrate such advantages, although it must be recognized that such a theoretical viewpoint must be translated judiciously into practical decision-making, e.g., demand that is normally distributed has a non-zero (although small) probability of reaching a negative number with large absolute value, whereas demands are usually non-negative. Beyond unit costs such as the cost of resources or the cost of unserved demand, it is important to understand the fundamental characteristics of demand variation for individual customer demands vs. if aggregated. Consolidating flat demands, or consolidating demands with those from similar customers within the same “community” does not offer much of a benefit, at least as far as resource savings are concerned. For example, if demands are

Smooth Operator: The Value of Demand Aggregation

related by a scaling factor, or even if there is at least one peak that is synchronized among all the customers, there is no difference in total requirements vs. peak requirements.

However, if demands are independent, there is a surprisingly strong effect from using the services provided by a “smooth operator.” If the strategy is “build to peak,” one can look at this from the perspective of total cost, where the relative cost can be viewed as being consonant with the average to peak ratio. Or, if the strategy is “minimize costs due to unserved demand and excess capacity,” the relative penalties under certain assumptions are only $\frac{1}{\sqrt{n}}$ of the penalties associated with n individual demands.

Translating these results to real-world strategy requires an additional level of analysis. Specifically, there are transaction costs associated with use of a service provider that don't exist for relationships within the firm's boundaries, as Ronald Coase⁹ points out in *The Nature of the Firm*. There are additional costs associated with maintaining secure multi-tenancy. Offsetting these costs are the benefits ascribed in this paper to pooling of independent demands.

Also, while I've shown here that there are advantages to aggregating demand and delivering it out of common, pooled capacity, I've also shown¹⁰ that there are occasions where, if the capacity is distributed and not all users can access all capacity, there is an additional problem of computational complexity that needs to be addressed.

Finally, while this paper has looked at each demand as independent and uncorrelated, in practice service providers can specifically target customer segments with negatively correlated demand to achieve even greater benefits. For example, hotels can target business travelers and conventions during the week with corporate discounts and convention packages including blocks of rooms, while targeting consumers via romantic weekend getaway packages that include brunch for two, achieving high utilization. Some resorts target skiers in the winter and then target tennis players and golfers during the warmer months. And, as Nick Carr¹¹ reports in *The Big Switch*, early electric utilities targeted consumers for night time lighting, trolley operators for the morning and evening rush hours, and factories during the day.

⁹ Ronald Coase, “The Nature of the Firm,” *Economica*, Vol 4(16), 1937, pp. 386–405.

¹⁰ Joe Weinman, “Cloud Computing is NP-Complete,” <http://www.joeweinman.com/papers.htm>”

¹¹ Nicholas Carr, *The Big Switch: Rewiring the World, from Edison to Google*, Norton, 2008.