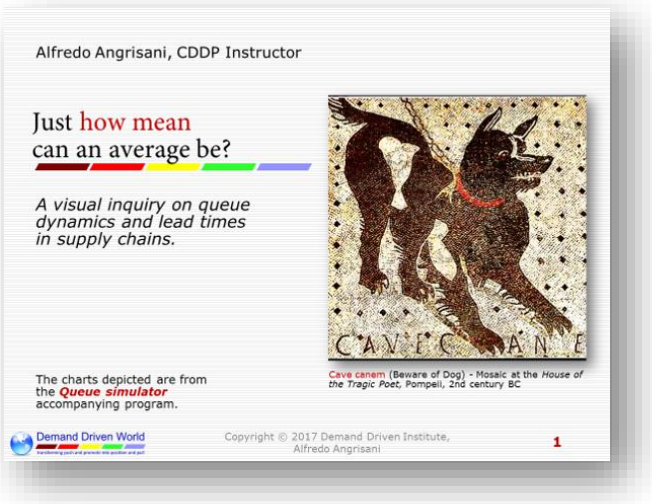
A Visual Enquiry on Lead Times

Just how mean can an average be?



As we all know, industry and services usually exhibit a **huge disproportion** between mere **touch times** and **cycle times**. In the mechanical industry, typically, this ratio is **1:10**, if one is clever.

So, how do orders, materials and all transactions **spend most of their time?**

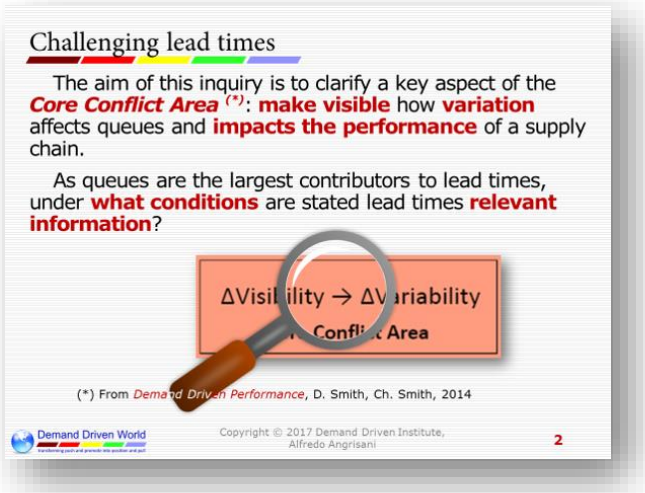
Simple enough: **they sit in queues!**

In fact, queues determine the **actual response time** of entire supply chain segments and, eventually, the supply chain's **overall ability to respond** to demand and generate cash flow.

Indeed, a stated lead time, in some conditions, might turn out to be a **very mean affair**, and can seriously impair a **crucial piece of information** at the **very heart** of the **core conflict area**, where the correctness (relevance) of information has its **maximum leverage**

Challenging lead times

For all practical purposes, these service times (queue plus processing) are expressed by a single number, the **ubiquitous lead time** but, as they depend on the timing of both demand and service, which are subject to (often huge) variation, encapsulating the behavior of the queues in **any single number** -i.e. stating a lead time- **is inherently insufficient** and, as we'll soon see, potentially **very misleading**.



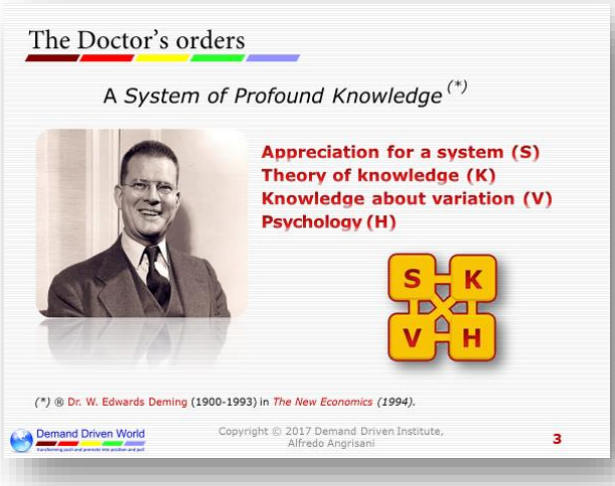
In the **Core Conflict Area diagram**[1], the arrow pointing from Visibility to Variability means that **the more we know about the system** through its relevant messages**, the better we can address variability**, enemy to process capability. But **how do we deal with variation,** in the first place?

As it happens, **we don't like the very idea of variability**: it suggests a vague territory, a state of things that we cannot seize with sufficient confidence for our purposes. Marshy grounds, indeed. Therefore, we are keen and **circumvent -hide- variation**, even in our lexicon. Lead times are one of these **carpets** under which we try and hide a fact of life that anyway **won't go away just so easily!**

[1] Debra Smith, Chad Smith: *Demand Driven Performance* (2014) p. 94.

How can we go about that?

When addressing complex issues, it's a safe practice to select and possibly follow a robust guideline. This time we don't have to look far; a great man -a giant- provides such guide: **Dr. Deming's *System of Profound Knowledge*[2]** is possibly the best backbone we can hope for, so we'll consider this subject from four angles: **System, Knowledge, Variation, Human;** and try not to miss too much from any of them. The letters of the **SPK symbol** will highlight these **specific angles** as we proceed.



A simple, everyday case

To investigate under which conditions the time spent in a queue is (or is not) predictable and, by that, make a consistent basis for a lead time, let's take the case of your **post office** where a (perfect) operator serves each client in **exactly 60 seconds**.

Clients' arrivals are distributed randomly during the **8-hour opening**.

Intuitively, our clerk should be able to satisfy a flow of up to 60 clients/hour, 480 in a day, **without a problem.**

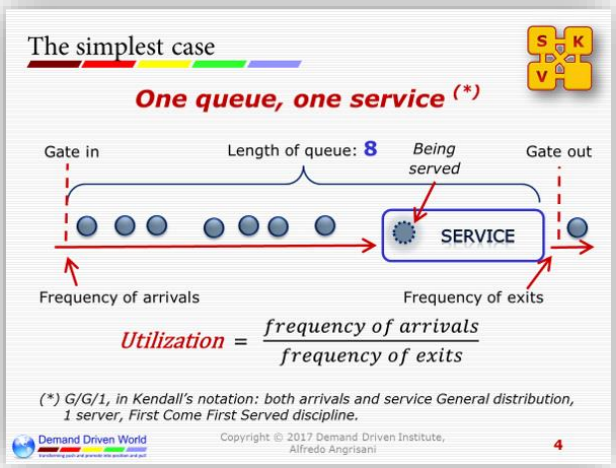
Or can he?

To check this hasty answer, we'll use an Excel program to model the behavior of the queue under different conditions, and… we might be in for **some surprises!**

Some (easy) theory first

Unbeknown to most laymen, queues are the object of a specific science: the **queue theory**, in fact. The problem with this theory is that, beyond the simplest cases, it quickly becomes extremely complex; its workings can be understood only by very smart statisticians (which I am not) who, in their quest for solutions, even in networks of medium complexity, usually need huge computational power.

Yet, for our purposes, some very simple conventions and mathematics will



[2] Dr. W. Edwards Deming, *The New Economics* - MIT Press (1994)

be sufficient to provide enough understanding and **give us an idea** about the process' fundamental workings.

Slide 4 depicts our very simple case of **one service fed by a single queue.**

This configuration is indicated in the **Kendall notation**, the standard adopted by the Theory, as **G/G/1**, meaning that arrivals (first letter) and service times (second letter) both exhibit a *general* distribution of probability.

In addition to the statistical *quality* of arrivals and service (which we'll discuss soon), the key factor that determines the behavior of a queue is **utilization** (U):

U = Rate of service / Rate of arrivals.

Notice that if utilization is -say- 90%, this means that the remaining 10% of the time the service is idle, i.e. ***the queue is empty:*** *but…* ***there is no telling when that is going to happen!***

Experiments

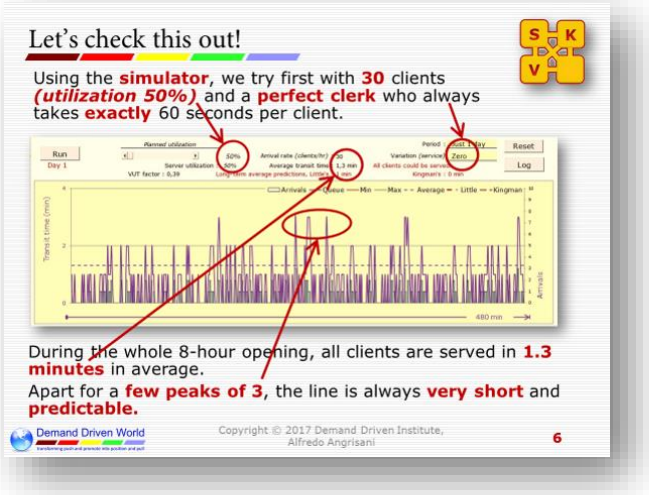
Simulations give us a **visual sense** of what happens with our queue during **a whole day** (480 mins) under different load conditions.

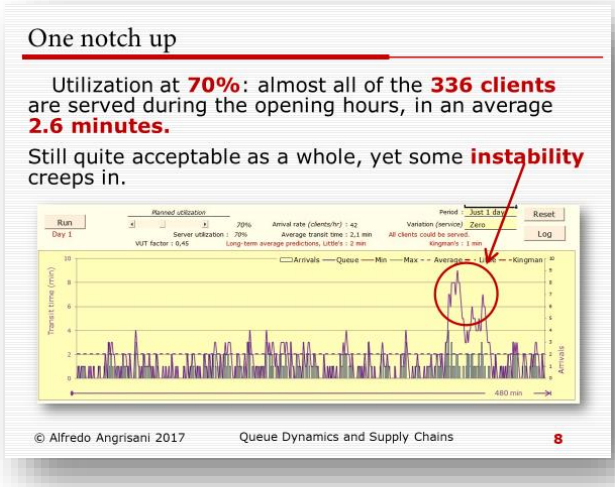
The next four slides depict simulation runs where we progressively increase utilization **from 50% up to 95%.**

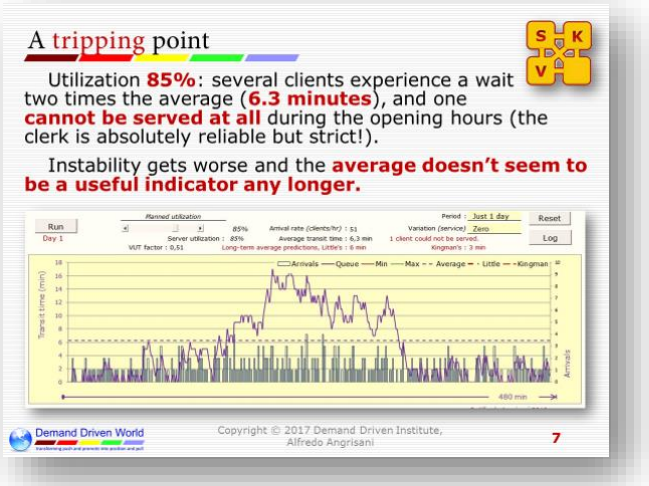
At **moderate** utilization (below 70%), waiting time is very low and **predictable**, but as we increase the load, the average wait increases, and some peaks appear **occasionally**.

The amplitude of **peaks and valleys increases sharply** as utilization is pushed above **85-90%.**

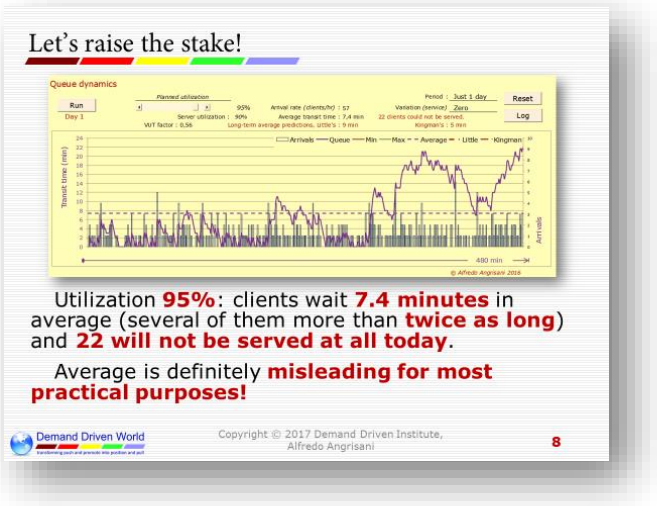
Consequently, as utilization grows **above this point,** the average wait quickly **loses relevance** as a signal of **predictable behavior.**





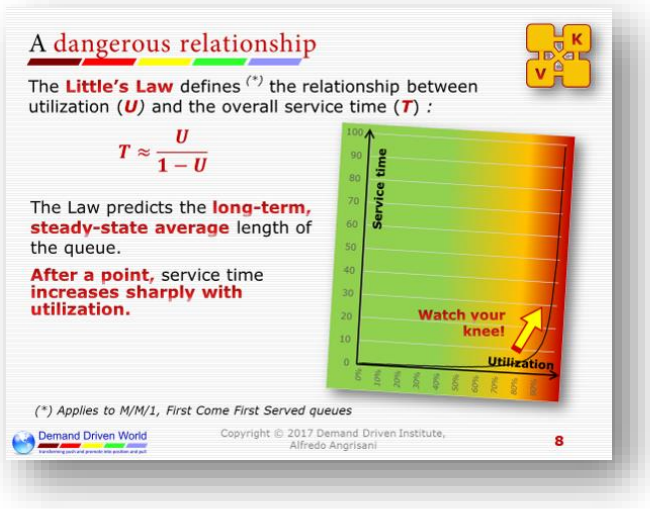


In addition, common variation is **effectively amplified** and the queue **initiates a 'noise'** that **damages the quality of all calculations** and **fosters the bullwhip effect.**



The average question

Turning to the queue theory, thanks to the simplicity of our configuration, we can make good use of one of its fundamental theorems, the **Little's Law,** to estimate the **long-term average length** of a queue.



The Law directly **connects service time (T_S) to utilization.**

$$T_S \approx \frac{U}{1-U}$$

The curve that depicts this relationship is **asymptotical** as utilization approaches saturation (U = 100%), and **exhibits a 'knee'** in the vicinity (again!) of 85%. Consequently, this **'tripping point'** turns out to be an area where **both the average service time and the spikes take off.**

The *general* case

The Little's Law applies only to *Marcovian* processes, also informally called 'memoryless', in the sense that one can make predictions about the future of the process based solely on its present state just as well as one could knowing the process's full history, hence independently from such history[3], which is **hardly the case** in real organizations where multiple causes affect the queue dynamics.
As we move away from the 'memoryless' system and into a ***general*** distribution, we need to account for the specific properties of arrivals and service. The **Kingman's theorem** does just that: it connects, for heavily loaded resources, **Variation, Utilization and Time** (hence the **VUT** name for the equation):

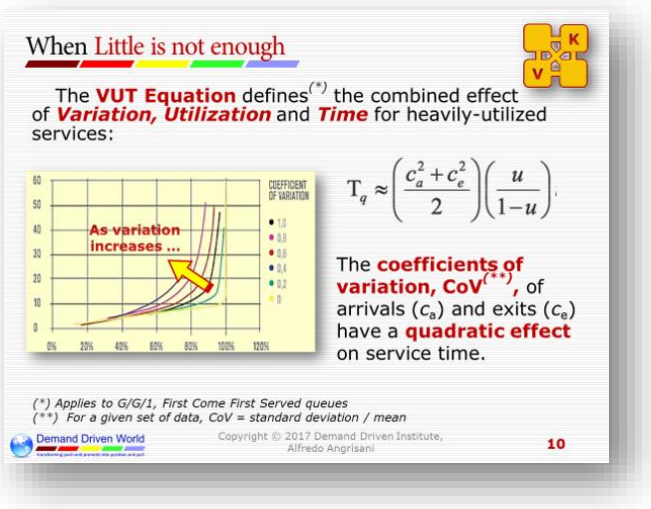
$$T_s \approx \frac{c_a^2 + c_e^2}{2} \cdot \frac{U}{1-U}$$

The VUT equation has the same 'core' as Little's Law, plus a 'correcting factor' that accounts for the statistical properties of arrivals and exits through the [squares of] ***CoV's,*** the coefficients of variation (c), ratio of standard deviations to the means.

[3] Quote from *en.wikipedia.org/wiki/Markov chain*

The effect of either increase in variation is to **flex** the 'Little's curve' **backwards and upwards**, thus **lowering the tripping point** and further **increasing the average service time**.

Only in the (very) special cases, when the CoV's sum of squares equals 2, then the two equations yield the same **average long-term equilibrium value.**



The shape of the demand

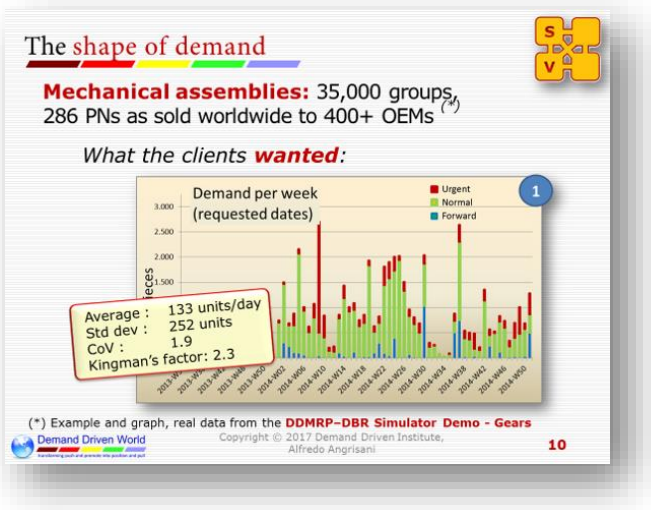
The VUT equation immediately raises our next question: **what CoV's can we expect** in real life?

Of course, the right answer is 'it depends'! But perhaps we can do better than this: the next slide portrays a graph of the actual demand for **35,000 mechanical groups** (**286 PN's**), over a year, to be assembled on the same line (the service).

The week in which we have allocated the demand, is the one actually **requested** by the ***400+ clients worldwide***.

For such industrial -no frills- mainstream products, one would hope for a reasonable CoV (especially if one needs to forecast!), but the result of this case (c_a= 1.9) is **not uncommon.**

If you consider the **true demand** (the **voice of your customer**, not the **voice of your sales** nor that of **your planners**), this is oftentimes what the **real 'engine' of market variation** looks like.



So far, we have assumed a single, independent and reliable service, but **what happens in the real world,** when the ability to perform can be seriously hampered by **all kinds of impediments?**

Shortages, delays, poor synchronization with other processes, unavailability of key resources, missing or wrong instructions, quality issues, delayed approvals… What is the **cumulative impact** of these disruptions on the predictability of a queue?

What CoV can **possibly gauge** that?

Shaking consequences

If we run the program over a longer period (a week or a month), we get a visual confirmation of two facts:

1. The ***average*** service time is quite **in agreement with the VUT prediction,**
2. As expected, at some moment, **the queue is empty**.

But **what does that imply?**

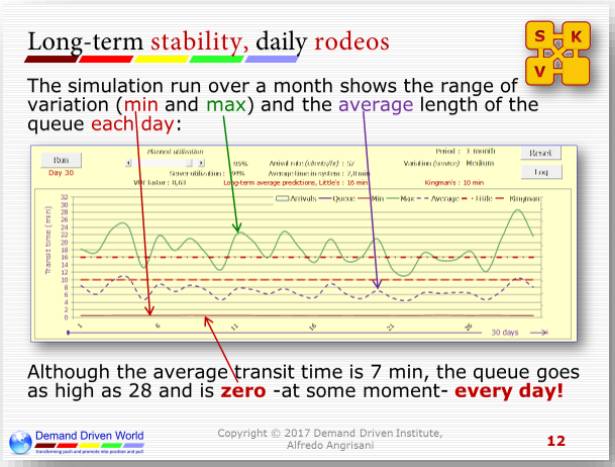
Take a look and **compare the daily average to the peaks.**

! *As utilization reaches the 'tripping point', the mean increases sharply and, as a consequence,* ***as there are always some ground points, the crests must skyrocket upwards****.*

In other words, it's not just a matter of how long but, chiefly, of **how unpredictable** the wait will be!

What are the consequences of such instability on your **daily management** of operations?

Unfortunately, the price is steep:



Highs:

- ➔ **Erratic priorities**,
- ➔ Bad **multitasking**,
- ➔ **Expediting**, overtime costs
- ➔ Loss of **service performance**.

Lows:

- ➔ Potential work disruptions,
- ➔ Potential **loss of market opportunities,**
- ➔ Misjudgments about the **actual process capability.**

Riding the rodeo of such waves can **quickly become impossible.**

Or, at least, **very expensive!**

Are you in for surfing or for tsunaming?

If utilization is acceptable, say 70%, then variation is limited and you can expect with some confidence that:

- if you are in a 'low', some load can be safely anticipated,
- if you are in 'high', you can expect (possibly with a 30% probability) that the pressure will relieve soon.

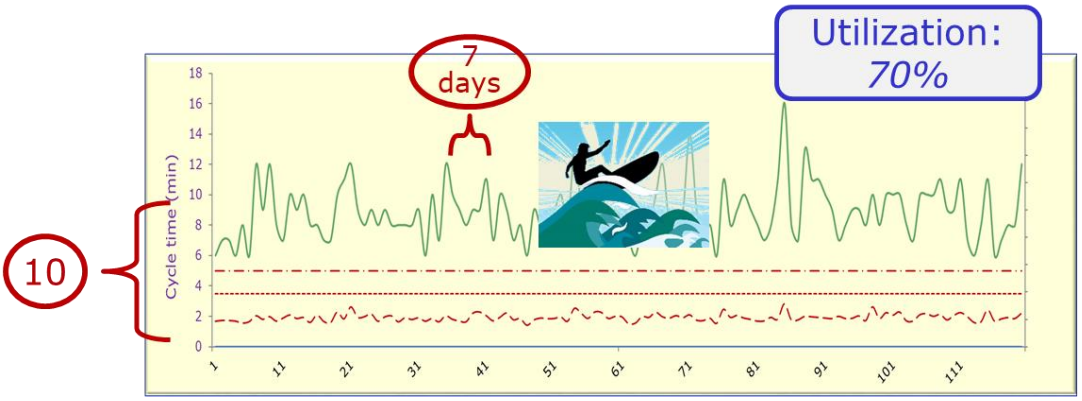
In other words, **you can manage**.

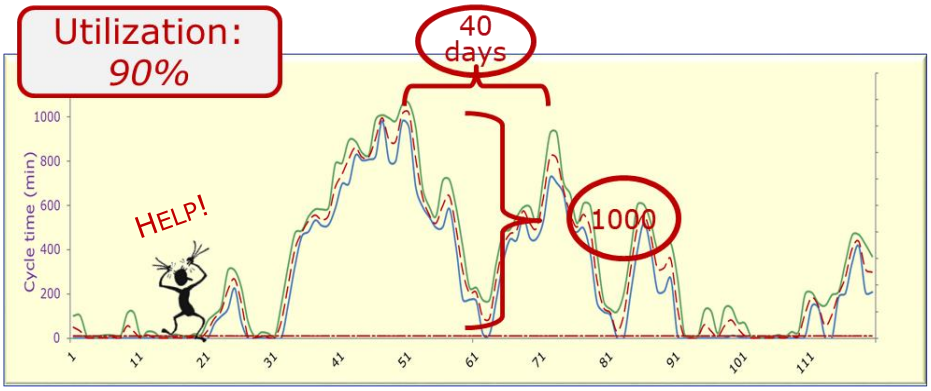
By the way, note that if the probability distribution of arrivals were Markovian, the past could not be used at all to forecast the future; if a condition has already lasted for some time, this would not be an indicator that it's going to end any time sooner![4]

As utilization and variation increase, not only the amplitudes of variation increase dramatically, but visibly they also tend to shape 'mountains' and 'valleys' of extended ranges, so there goes your visibility and chance to adjust execution based on a reliable advance information or even feeble leading indicators.

In these conditions, work looks more like a **perpetual crisis management.**

The figures show these vast differences in both traits as utilization is increased from 70% to 90%



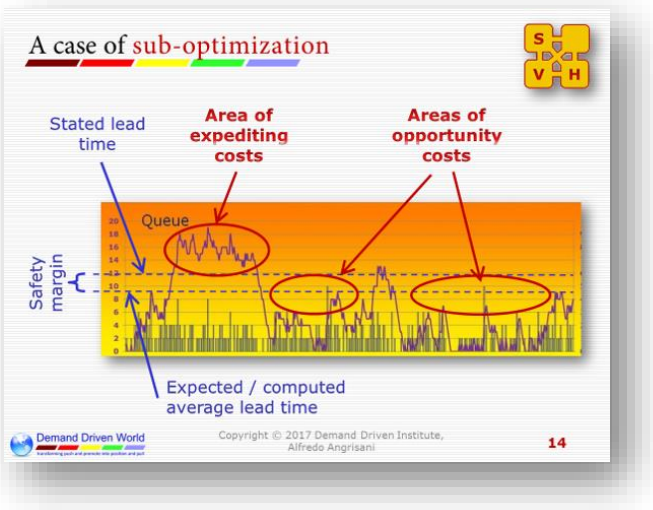


[4] A common example of 'memoryless' distributions are phone calls: if in a particular area and period, the telcom company determines that the average length of calls is -say- 3 minutes, and we check on a number of actual calls that *have already lasted for 2 minutes*, then the average *further duration* of these calls *would still be 3 minutes!* This is not the case with most industrial orders situations, but could be for a web-based business.

A *very* compelling reference

Once lead times are used in contracts, data bases (like MRP's) or algorithms, they become **compelling references** that don't just drive calculations, but **chiefly human behavior.**

Stating lead times is essential, of course, but what are the **effects** and **costs** of **unreliable** lead times as they play the role of universally **accepted standards** and **operational targets?**



To mitigate the problems that stem from incorrectness of all sorts, the lead times usually incorporate some **'safety margins'**, but how **do these slacks help?**

Not much, as they simply set **'new zeros'** and increase the overall system **inertia**.

For instance, if you are **late,** you will:

- ➔ *Do your best to recover the delay and thus incur into **expediting costs,***
- ➔ *Accept a **loss in service** or in **product quality,***
- ➔ *If delays are frequent, you may even **increase the slack** without being aware of the consequences on the **flow of cash** and **materials** and on the **response to market.***

If, on the contrary, you are **early,** you may:

- ➔ *Delay the work and, having used up the safety capacity, risk **to be late** if something goes wrong,*
- ➔ *If these 'lows' are interpreted as 'waste' or inefficient use of a resource, possibly believe that the lead time in use is **too long** and **reduce it!***
- ➔ *The Parkinson's Law[5] points out that* "work expands so as to fill the time available for its completion" *so, most of the time, you will hardly get any gains at all but will incur in **opportunity costs** instead.*

In other words, the lead times might become **self-fulfilling prophecies** and cause the system to **fluctuate**, without much control, in a swamp of **sub-optimization.**

Non-measurable costs, or: how to get yourself a great constraint!

Far beyond the issue of **measurable costs**, the least desirable effect of lead time instability is the **saturation of the most valuable 'service'** of all: the **mangers' attention and time** and, as a result, the creation of **an extremely effective** (although invisible) **constraint for the whole organization!**

Contrary to common sense, your managers' **ability to 'put out fires'**, to **negotiate hard and unsatisfying settlements** of **unnecessary conflicts**, to **micromanage** and work **long overtime hours**, are generally **valued** by businesses and perhaps **a matter of pride** for the manager himself!

Organizations are thus deprived of their best, most crucial (and most expensive) resources, and **managers waste chances** to improve their professional and personal lives.

[5] Cyril Northcote Parkinson: *Parkinson's Law and Other Studies in Administration* (1957).

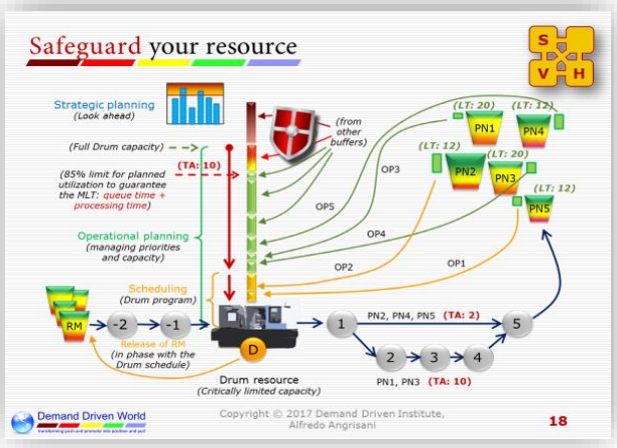
In conclusion, a set of sound recipes

Variation is a fact of life often **ignored** and **swept under the carpet** of **single values**.

Excessive utilization triggers and amplifies variation and has **disruptive effects** on lead times causing them to **lose relevance,** and **jeopardizes** the entire supply chain performance**.**

If you need to **bring back** queues (and flow) into **manageable territory**, here is a **short list** of actions to consider:

1. **Identify and reduce the number** of critically constrained resources in your system.
   - **Optimize** -offload when possible- their work to increase capacity.
   - Consider **Lean initiatives (SMED).**
2. **Feed such resources from controlled sources.**
   - For this and most next points, follow the prescriptions of **DDMRP**[6]
   - Place buffers in front of the resource and release the work orderly at the maximum rate allowed by that resource.
   - If the source is **demand**, consider adopting a mixed-mode MTS/MTO model.
3. **Don't plan for utilization above 85% of resource capacity.**
   - Don't go after misleading efficiency KPIs!
   - Look ahead for potential overloads
   - Perform the **DDS&OP process** regularly and set the appropriate ground
4. **Shorten lead times by system design.**
   - Use buffers to 'slice' lead times into shorter, independent segments.
5. **Stifle the overall process variation.**
   - Again, **by system design**, using strategic buffers.
   - Consider **6 Sigma** initiatives.
6. **Reduce dependence** on lead times.
   - **Assess** supply **less on their compliance** with stated lead times and **more on their ability to coordinate work with your priorities**.
7. **Reduce process time**
   - Consider **Lean** initiatives.
   - Consider **re-engineering** the product
8. **Stifle variation in demand.**
   - Adopt ***Sales Process Engineering*** **(SPE)**[7]**.**
   - Mitigate peaks through appropriate sales/supply chain actions like ***Vendor Managed Inventory*** **(VMI).**
9. **Adopt flow-based *Smart Metrics***[8] to control the system effectively
   - Don't go after (impossible) optimizations, but use these metrics to **drive the PDCA improvement cycle.**



[6] Carol Ptak, Chad Smith: *DDMRP - Demand Driven Material Requirements Planning* (2016).
[7] Justin Roff-Marsh: *The Machine* (2015), ballistix.com
[8] Debra Smith, Chad Smith: *Demand Driven Performance Using Smart Metrics* (2014).

A few technical notes

As a famous quote goes: "*All models are wrong but some are useful*"[9], so in designing the Excel® simulator that goes with this paper, I have selected a limited number of scenarios; the bare number (hopefully) sufficient to give the user a gist of what fundamentally happens within a queue.

To me, writing this program was an almost physical reaction to the **inadequate clarity, of all analyses based on 'averages',** especially in accounting for real-life experimentals. In other words, the **lack of coherence** between those analysis and our direct perception of the behavior of the queues of all sorts.

The idea was to have a peek under the famous '**carpet of single numbers**', and it seems to me that the surprising dynamics unveiled visually by the program prove that the effort was worthwhile.

So, with the ***2017 Demand Driven World Conference*** in Lyon, France, **I gladly offer this work to the DDI community** in the hope that this model can belong to the second set in the quote also for someone else than myself!

Scenarios

The simulation program allows for **4 time ranges** and **3 variation levels** both for arrivals and exits.

- Time ranges: 1 day, 1 week, 1 month (30 days), 4 months (120 days)
- Variation levels of arrivals: low, medium, high.
- Variation levels of service: zero, medium, high.

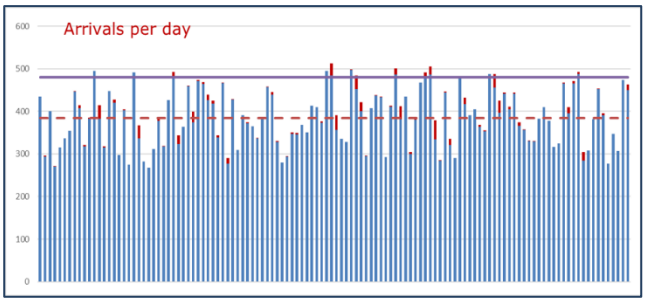
Arrivals

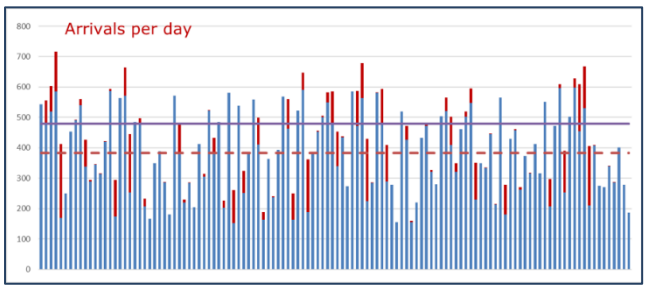
Arrivals are always **evenly spread over the day**, i.e. a second of arrival is assigned randomly between 1 and 28,800 (8 hours x 60 minutes x 60 seconds) by the Excel VBA ***rnd* function** to each planned transaction, whose total **(PDA)** is determined as:

Planned Daily Arrivals (PDA) = Utilization x 480 minutes.

If the time frame is **longer than 1 day**, then the variation levels of arrivals are set according to the following rules:

- **Low**: all days in the period receive **exactly the same PDA** number of clients
- **Medium**: the total number of arrivals (**PDA x Days**) is spread across the period with a +/- variation day-to-day of **20%**





- **High**: the total number of arrivals (**PDA x Days**) is spread across the period with a +/- variation of **50%**

Please note how even such 'high' variation still **looks 'mild'** compared to the actual pattern of demand for the mechanical groups mentioned earlier in this paper.

[9] George E. P. Box in the *Journal of the American Statistical Association,* 1976

Capacity, utilization and carry overs

The **blue horizontal** line in the graphs marks the full standard **average capacity** of the server (480 clients per day) and the **red dotted horizontal** line the overall **average load.**

The **red vertical** segments show the **'carry overs':** arrivals that could not be served on the preceding day and that show up again at the following opening.

Overloading

If the system is overloaded, the **Little and VUT equations are no longer applicable**, as extended saturated periods (planned utilization > 100%) disrupt their underlying assumptions. When this is the case, both equations are still computed, but the diagram in the home page **will not show them**.

Service

Except for the case of Zero variation, server times are again a random value (VBA *rnd* function), whose overall average is (almost exactly) 60 seconds across the entire simulation period.

The 3 levels of variation are as follows:

- **Zero**: **all** service times are **exactly 60 seconds**
- **Medium:** each service time is **60 +/- 20 seconds**
- **High:** each service time is **60 +/- 40 seconds**

Detailed logs

The **detailed log** for the **single-day** scenarios and the intermediate computation results are readily available in the **log page**.

The **longer periods** are summarized by day, but for the durations of **7 and 30** days **detailed logs** (for each transaction and for every minute) are also available as an option from the home page (***Transactions' details*** checkbox).

As this option requires an extended use of the computer resources, it can be quite time-consuming depending on your processor (queues, again!), thus altogether impractical and **definitely useless** for the 120-day period.

Contacts

If you have any comments about this paper or the program, or if you find (very possible) inaccuracies or plain errors, or if you just want to share your thoughts or suggest improvements for the next revision, don't hesitate to contact me:

posta@alfredoangrisani.it

Skype: alfredo.angrisani

Tel +39 340 869 7635

And, if you happen to visit Bologna, also don't forget to drop me a queue… oops, a line, I mean!