Queues! Kingman: The Equation of Lean

Why are Queues relevant?

- Lead time (over 90% of lead time is in queues in manufacturing, and up to 90% in service)
- Design and New Products
- Software and 'Agile'
- Inventory and space
- Cost
- Defect detection

Kingman's equation (written in a shorthand form) is

Ave Queue length is approximately = $(V_a + V_p)/2$ x U x T_p

Where

V_a is coefficient of arrival variation, made up of value demand variation and unnecessary demand variation. (This is 'Mura' type 1)

- Value demand variation is the 'natural' variation of first time 'true' demand.
- Unnecessary demand variation is demand variation caused internally (for instance order batching, 'end of month hockey stick', or discount promotions), and/or externally (for instance by the 'bullwhip' effect, or 'lumpy' demands.)

V_p is coefficient of process variation, made up of natural process variation and unnecessary process variation. (This is 'Mura' type 2)

- Natural process variation is the 'natural' variation of process activity time. There may be subgroups here, when there are several operators with different competency.
- Unnecessary process variation is process variation caused internally (for instance by breakdowns, stoppages, lack of 5S), and/or externally (for instance by shortages of materials or wrong information.) Note: Similar cautions for SPC.

U is utilization of the process = load / capacity

Load is the work that must be done and = value demand + unnecessary demand.

- Value demand is first time, real demand (also sometimes called 'fresh demand')
- Unnecessary demand is generated externally by, for example, failure demand, supply chain 'bullwhip' ordering), and/or internally (e.g. rework, overproduction, overprocessing, preparing unread reports and KPI's)

Capacity is the capacity of the system to do the work and = base capacity – waste.

- Base capacity is the capacity of the system if everything works perfectly (e.g. MTTR=0, zero changeover time, no shortages, no interruption, design speed, perfect info.)
- Waste is anything that detracts from this capacity. (e.g. This could include many of the 'classic' wastes of Ohno – and any non perfect as above). (Waste is 'Muda'). Waste may include an operator not being fully trained, or fully competent.

U is = u/(1-u), with u being utilization expressed as a decimal. This would mean that if utilization is 100% or 1, then U would be infinite. In practice this means that the average queue is VERY SENSITIVE above about 85%. (This represents overload and can be called 'Muri'). Here the curve is HIGHLY NON-LINEAR so that reducing unnecessary demand, or removing waste has a DRAMATIC effect on reducing queue length. **Muri begins at less than 100% utilization!** (Hence the queue should be limited by controlling input (DBR or CONWIP)). AND queue uncertainty (or the range between min and max queue length) also increases sharply with utilization! A queue is better represented as an expanding cone rather than a line.

Tp is the average process time. This of course may also contain waste. OEE and the losses may be relevant. There is only ONE way to fully eliminate a queue, and that is by making Tp =0 (i.e. eliminating the process). (Eliminating all variation is unlikely!)

Note: The equation can be applied to each of a series of processes, in service, manufacturing, or development. In this case arrival variation is determined by the process immediately upstream, and by the move quantity (batch size), any rework, and by replenishment signaling such as kanban.