Determining the maximum resource utilisation for optimal system profit

by Olufemi Adetunji and Prof Sarma Yadavalli

A major shift in production and supply chain management is that from a predominantly push to a predominantly pull paradigm. However, not many organisations are based entirely on one or the other. The extent to which each paradigm is implemented is usually captured in the decoupling point of the organisation.

In a typical push environment, planning drives the production, and more inventory is allowed to support the production system. A pull environment is more critical of inventory, and management of flow is very important. This is the main thrust of lean principles, theory of constraints (TOC) and the more recent constant work-in-process (CONWIP) systems. Wallace Hopp, a specialist in supply chain science, defined a pull system as a system in which work is released based on the status of the system, thereby placing an inherent limit on the work-in-process (WIP) inventory. This is in contrast to a push system in which work is released based on plan, irrespective of the state of the production system. The magic of pull, he said, is the cap on the WIP in the system. This, in his opinion, is why many pull systems are more profitable than push systems.

However, this WIP cap is implemented differently in the various pull systems, and job scheduling is always implicitly linked to inventory control. Lean uses the Kanban, TOC uses the drum buffer and rope (DBR) and the CONWIP monitors the exit of jobs from the system. One implicit assumption in all such systems, however, is that the demand environment – and by extension, the production system – could be steadied somehow. It is apparent from all these approaches that a key component of every pull technique is the conscious management of the job flow rate through the system and the implicit containment of the level of the WIP in the system.

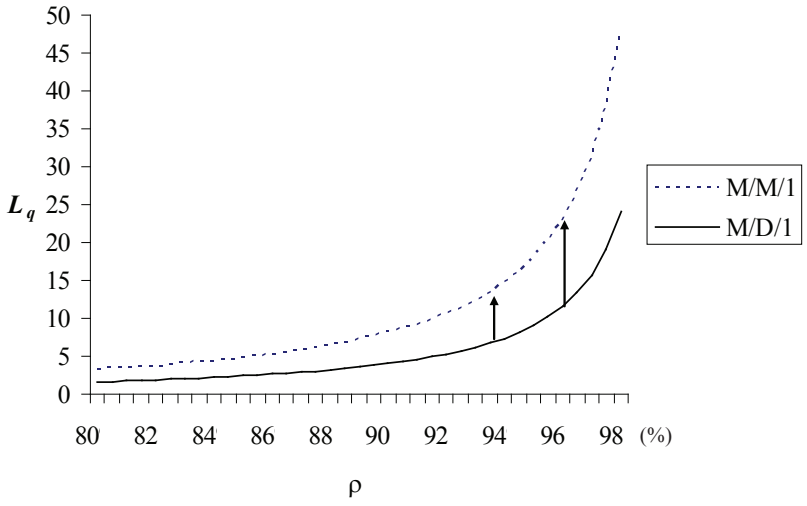
Therefore, an understanding of why WIP grows significantly in every production system (goods or services) would enhance the management of such systems. The interesting thing is that the WIP level is closely related to two main issues: the level of variability in the system and the level of utilisation of the system. The utilisation effect is more easily represented through a simple mathematical relationship called Little's law. Simply put, it states that in any system with variabilities, WIP level = throughput rate x cycle time.

This is similar to most other conservation laws in engineering. It is one of the most fundamental equations of queuing systems. Queues are pervasive models that have been (and are still being) widely researched.

The problem of variability is more intricate than that of utilisation. It is directly dependent on the nature of arrival to the system and that of processing at the resource. It can be captured in simple terms by the coefficient of variation, or in more explicit terms, by the distribution of these variables. Because of the numerous possible combinations of such systems, notations have been developed to manage these systems.

The seminal work in this regard was done by Kendal, who proposed a four-field notation: $A/B/C/D$, where A indicates the nature of variability of arrival pattern, B the nature of variability of processing time, C the number of processing resources available to the input, and D the size of the calling population.

The advantage of the many standard models derived from this classification is that their steady-state solutions are readily available. These steady-state solutions also determine how the system behaves, based on its level of utilisation.



→ *1. Curse of utilisation and variance.*

One of the simplest cases is the Markovian System of ***M/M/1/∞***, where the first ***M*** is Poisson and the second ***M*** is exponential by nature of their distributions. The pattern of WIP builds up as a function of the utilisation and is captured in a term referred to as the curse of utilisation.

This is shown in Figure 1 for the ***M/M/1*** and ***M/D/1*** , where ***ρ*** is the system level of utilisation and L_q is the WIP level (expected queue length). ***D*** means that the processing time is uniform.

From Figure 1, it can be seen that while the WIP level is dependent on the system level of utilisation, its expected value balloons as the utilisation level approaches full resource utilisation. This makes it imperative for a production manager to watch this trade-off as he or she tries to push more and more products through the pipeline to meet more customer demands. There comes a time when it is better to allow customers' demands to go unmet than to increase system throughput, even when it is not yet at full utilisation.

A simple flow model could be built around this characteristic by defining a profit function around the holding cost of the expected WIP that results from the level of utilisation of the system and the profit earned from sale or throughputs from such system.

This is defined as:

$$NP = TH - OE \qquad 1$$

Using the steady-state solution of the ***M/M/1/∞*** queue that the arrival and processing time is assumed to follow, and optimising the utilisation factor relative to this profit function, one can derive an optimal job flow rate to be

$$\rho^{\cdot} = 1 - \sqrt{\frac{C_{OE}}{\mu C_{TH}}} \qquad 2$$

$$NP^{\cdot} = (\sqrt{\mu C_{TH}} - \sqrt{C_{OE}})^2 \qquad 3$$

where ***NP*** is the net profit, ***TH*** is the throughput rate, ***OE*** is the operating expense (incurred during the same time window as the throughput, and assumed here to be made up of only the holding cost), μ is the rate of service at the resource over a stated time interval, C_{TH} is the profit earned from selling a unit of output and C_{OE} is the inventory cost per unit (product-time).

From the model in Equation 2, one can conclude that if the level of utilisation required of the resource by the customer demand exceeds the optimum utilisation level, it is better to allow the customer demand to go unmet. There may, however, be instances when it is important to factor in the cost of not having the product available. Assuming a once-off cost is paid for not having the product available on demand, the equations may be modified as:

$$NP = TH - OE - SH \qquad 4$$

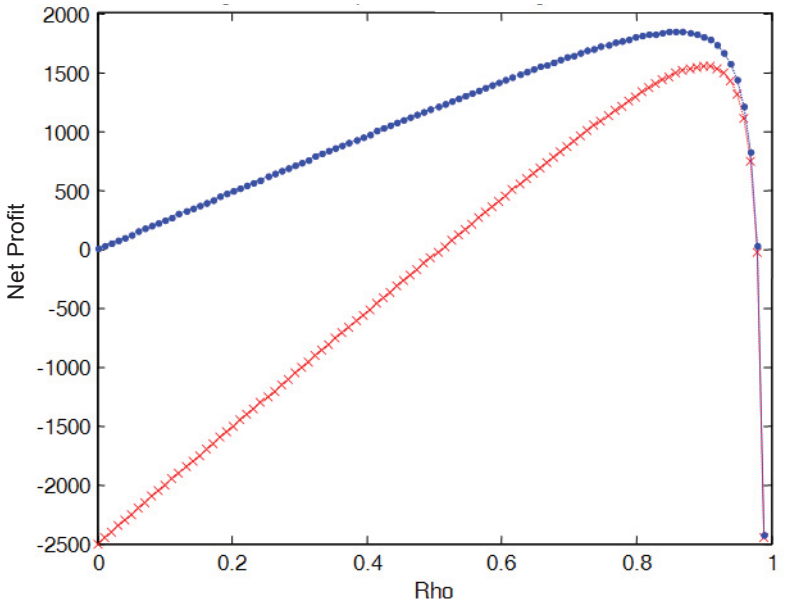
$$\rho^{\cdot} = 1 - \sqrt{\frac{C_{OE}}{\mu(C_{TH} + C_{SH})}} \qquad 5$$

$$NP^{\cdot} = (\sqrt{\mu(C_{TH} + C_{SH})} - \sqrt{C_{OE}})^2 - \mu C_{SH} \qquad 6$$

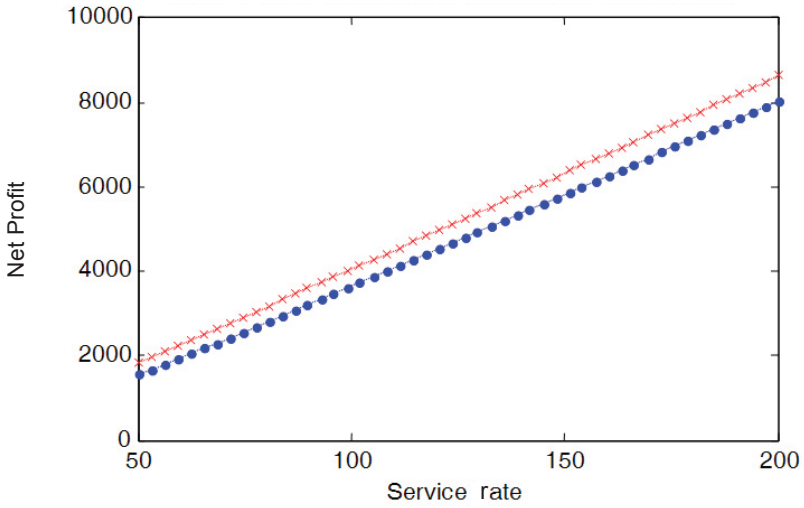
where ***SH*** is the cost of not having the product when demanded (also incurred during the same time window as the throughput), and C_{SH} is a once-off shortage cost charged per unit product for not having the product available for the customer.

To understand why these models are important to the production system, a simple example can be used by initialising all the variables to some values, and the qualitative behaviours of each of the system parameters can be explored. If each of the variables is initialised to 50 units, for instance, and the behaviour of other variables and functions is explored as one of the variables is varied, the following can be observed, starting with the net profit relative to the level of system utilisation:

In Figure 2, the blue-coloured graph is the case where the unit shortage cost is zero, and represents Equation 2. The red-coloured graph represents Equation 5. It can be seen that, in both cases, the net profit declines very rapidly after the optimal



→ *2. Plot of optimal net profit against utilisation.*



→ *3. Plot of optimal net profit against service rate.*

utilisation level. This effect is directly traceable to the rapid non-linear increase in the WIP level as the system gets close to full utilisation, as shown in Figure 1. This shows that it might not be profitable in any way to meet customer demands beyond the optimal utilisation level.

Investigation of the qualitative behaviour of the various input parameters with respect to μ shows, even from the equations, that the optimal utilisation level increases with an increase in the maximum processing capacity, μ, the unit profit rate, C_{TH}, and the unit shortage cost rate, C_{SH}, while it decreases with an increase in the unit holding cost rate, C_{OH}. However, it appears from the diagrammatic plots that the holding cost has a greater effect on the models.

The qualitative behaviour of the net profit pattern is shown in Figure 3. The net profit function for models with shortage cost (Equation 6) is below that of the one without shortage cost (Equation 3). Changes in other input parameters like shortage costs, unit profit and unit holding cost can be plotted in a similar manner to Figure 3, as well, but will be omitted here.

Prof Sarma Yadavalli *is a professor and head of the Department of Industrial and Systems Engineering at the University of Pretoria.*

Olufemi Adetunji *is a lecturer in the Department of Industrial and Systems Engineering at the University of Pretoria.*