http://sajie.journals.ac.za

ISSN 1012-277X S.Afr.j.ind.eng.

S A *Tydskrif vir Bedryfsingenieurswese*, Vol 7, No 1, Junie 1993, pp.29-42

THE EFFECT OF VARIABILITY ON WORK IN PROCESS (WIP) AND ON CAPACITY

Keith Sandrock, School of Mechanical Engineering,
University of the Witwatersrand,
PO BOX 249, WITS, 2050, RSA.

ABSTRACT:

Capacity concepts are usually illustrated by Wight's funnel analogy in which the queue is the volume of water above the throat of the funnel and capacity is the volume flowing out. This simple illustration implies that input and output rates remain constant over time.

Since this is usually not the case in practice, a better model was sought and industry latched on to simple single-server queueing models (which are incidently also admirably represented by the funnel analogy). Whereas Wight's funnel implies constant I/O rates, the simple queueing model implies statistical equilibrium between I/O rates - and this is a methodological improvement. However the approach still does not represent real-world situations well enough. Industry needs even better simple-to-use models. This paper attempts to resolve this issue through the use of two alternative approaches.

OPSOMMING:

Kapasiteit word gewoonlik verduidelik deur gebruik te maak van Wight se tregter analogie waarin die water bokant die nek van die tregter as die werktou beskou kan word en die water wat uitloop as die kapasiteit van die stelsel. Hierdie benadering neem aan dat die insette en uitsette relatief konstant bly. Omdat dit gewoonlik nie in die werklikheid die geval is nie, m.a.w. dat variasie wel plaasvind, het bedrywe as alternatief gebruik gemaak van eenvoudige toustaanstelsels (wat terloops ook deur Wight se prentjie verduidelik kan word). Hierdie modelle is gegrond op die aanname dat alhoewel die insette en uitsette mag varieer, statistiese stabiliteit tog sal geld. Hulle mag dus 'n verbetering op die tregter analogie wees maar kan nog nie die werklikheid ordentlik weerspiel nie.
Ingenieurs moet dus beter modelle toepas. Twee sulke modele word hierin bespreek.

INTRODUCTION

In industry we have production cycles which are variable in their duration and in their output.
Because of this variability there are times when the demands made on a workcentre are temporarily greater than its capacity.
A bottleneck situation naturally develops, and a queue forms as items start to build up behind the workcentre.
However, these 'peak' periods are randomly alternated with 'valley' periods during which the input is below the capacity of the workcentre so that excess work in process (WIP) is cleared, and the queue diminishes in size or vanishes altogether, only to start growing all over again.
The concertina-like queueing phenomenon results in extended manufacturing lead times, and also in capacity problems, hence it must be eliminated if we are to make optimum use of our resources.
The first step in this direction is to analyse the system so as to develop a picture of what is happening. Queueing theory is a useful tool for initiating a study, and for providing analytical results that can be used in subsequent decision making.

MEASURES OF VARIABILITY

There are two commonly used measures of variability - the Mean Absolute Deviation (MAD), and the Standard Deviation (SD). These two measures are related by the useful approximation:

SD = 1,25 MAD

A third measure is the Coefficient of Variation (CV) which is given by:

CV = SD/Average

The coefficient of variation is extremely useful because it provides an estimate of the ABSOLUTE variability of the system. For example if someone says: "This process has a standard deviation of two minutes" he supplies no meaningful information because, for argument's sake, two minutes on a job with an average processing time of ten minutes is far more variable than two minutes on a job that takes one hundred minutes. The respective coefficients of variation are 0,2 and 0,02 respectively, which provides a clear picture of the situation.
Another way of viewing variation is to study the queue which forms behind a workcentre (or some other server system). The queue builds up and dies down in accordance with fluctuations in the input rate and fluctuations in the processing time of the system. This probabilistic behaviour can be characterised and summarised using queueing theory. Hence it is not surprising that elementary queueing models have been used to study shopfloor problems.
However, simple text-book models are really inadequate for the purpose. Although these models do provide answers, the gap between their underlying assumptions and what actually exists in the real world is often too large to be ignored.

Better models exist but require statistical knowledge for their use. Since this commodity is scarce in industry, user-friendly yet adequate models are what is ultimately desired. Such a model is discussed later in this paper.

THE SIMPLE M/M/1 QUEUEING MODEL (AND A DRAWBACK)

If it is assumed that the input/output of the system remains in a state of statistical equilibrium then the behaviour of the queue can be described by a Markov process (Cooper[1]) in which:

$\alpha P_j = \mu P_{j+1}$ for all $j = 0, \ldots\ldots n$.

Where:

α = average input rate in items per unit time
μ = average throughput rate in items per unit time
P_j = proportion of time there are j items in the queue.

From this elementary relationship the following well known results are obtained (Cooper[1]; Ravindran[2]):

Average Waiting-line Length $E(Q') = a^2/(1-a)$
Average Waiting Time to Service $E(W') = a/\mu(1-a)$

Where $a = \alpha/\mu$. The ratio 'a' must be less than 1 otherwise the queue will grow to infinity.

From the above expressions for queue length and waiting time it is seen immediately that:

$E(Q') = \alpha \times E(W')$

an expression known as Little's formula. It is this expression which is used by Williams[3] (in a paper to the 1990 APICS Conference) for studying shopfloor capacity.
For a simple proof of Little's unifying result see Jewell[4].

Now consider the condition stipulated above, namely that $a = \alpha/\mu$ must be less than 1 for queues to be of finite length. Many industrialists feel (intuitively) that if this condition pertains then there should never be a queue at all, and that when a queue does develop it does so because something has gone wrong, i.e. it has occurred by default. This assumption would be absolutely correct if there was no variation in either the input rate or the throughput rate.
In reality, this shopfloor situation is not deterministic but stochastic in its behaviour, and hence queueing is a predictable occurrence, not a random one. The simple M/M/1 model takes stochastic behaviour into account even though only the average input and throughput rates appear in the above expressions. Implicit in the M/M/1 model is the assumption that we are dealing with negative exponential input/output inter-arrival- time distributions which have a coefficient of variation equal to one.

The question is: "How realistic is this assumption?"
To answer this question, consider the following times (in minutes) for a certain operation on a drill press:

9 11 10 12 10 8 9 8 10 12 11 12 8

These times - all for the identical operation - are about as variable one would be prepared to accept under normal (average) conditions from a drill press operator - perhaps they are even too variable.
For these data the CV = SD/average = 1,572/10 = 0,1572 which falls far short of unity.
Hence, in the light of this little illustration, should one really believe that in industry it is possible to regularly find distributions with variation of such a magnitude that the CV = 1? This is most unlikely. Which means that the use of elementary queueing theory for examining shopfloor problems must be suspect - must be almost guaranteed to provide wrong answers.
In their excellent treatment of queues Cox and Smith[5] state that: " ... there are numerous complications arising in practical applications which can make these assumptions seriously untrue." These authors go on to discuss some specialized models which although elegant, are not practical enough for shop floor use.

AN ALTERNATIVE MODEL

As an alternative, it is a good idea to use a queueing model which takes the variance of the input/output distributions explicitly into account.
This means more work of course since it is now necessary to determine the average inter-arrival time of items coming into a workcentre, the average time they are worked on (service time), as well as the standard deviations of these times.
Yet this in itself is an excellent exercise because it leads immediately to the determination of the CV and hence to a very good estimate of how variable the process really is.
Having obtained these initial estimates of the parameters of the input and output distributions, the application of the alternative model is made extremely simple via the use of a specially constructed set of tables provided in the appendix. These tables are based on the Erlang Distribution. In this paper the aim is not to provide a detailed discussion of Erlang's distribution - but simply to show how it can be used to study real world queueing systems quickly and effectively.
A good way of doing so is through examples.

Example 1

At a certain workcentre jobs arrive at an average rate of
α = approximately 10/hour.
Actual observations taken over an eight hour shift are shown below:

HOUR NUMBER	1	2	3	4	5	6	7	8
No. of JOBS	8	12	6	15	9	7	10	15

The average INTERARRIVAL TIME in minutes may be estimated as 60/8 for the first hour, 60/12 for the next hour, and so on, which works out over the entire shift at an average of 6,47 minutes with SD = 2,16 minutes.
Hence the CV = 2,16/6,47 = 0,3335. Call this CV_a.

The processing times at this workcentre are also variable and from a similar exercise to the one above, the average run time was found to be 5,26 minutes with SD = 3,73 minutes.
Let CV_b = 3,73/5,26 = 0,709.
[Note that the throughput μ = 60/5,26 = 11,4 jobs/hour].

This is all the information needed to make a realistic assesment of the workcentre's performance.

Step 1: Calculate the utilization of the workcentre given by the ratio of service time to input time as follows:

Utilization U = 5,26/6,47 = 0,812 or 81,2%

Step 2: Enter the table (Appendix) for U= 0,8 with the two values of the coefficients of variation CV_a and CV_b to obtain the value of the waiting time factor.
Some interpolation is required here since CV_a = 0,3335 lies between two values in the table. CV_b is so close to the value of 0,707 (a table column heading) that we can ignore the difference. The correct procedure is to use log-interpolation in these tables, but linear interpolation provides a quick rough estimate.
The factor we arrive at is 1,0064.
Similarly, a factor of 2,5234 is obtained from the table for which U = 0,9. Linear interpolation between these factor values yields a final value of 1,1884 for the waiting time relevant to U = 0,812.
[Note that log interpolation should be used WITHIN tables and linear interpolation BETWEEN tables].

Step 3: Multiply this factor by the average processing time to obtain the average waiting time of jobs prior to service.

$E(W') = 1,1884 \times 5,26 = 6,25$ minutes.

Step 4: Use Little's formula (which is extremely robust) to obtain an approximation of the expected queue length.

$E(Q') = \alpha\ E(W')$

$= 9,27 \times 6,25/60$

$= 1$ item

Use of the simple single-server M/M/1 model for this problem provides the following results:

$$E(Q') = a^2/(1-a) \text{ where } a = \alpha/\mu = 0,812$$
$$= 3,5 \text{ items.}$$
$$E(W') = 22,6 \text{ minutes.}$$

This answer is grossly in error. It is wrong because the distributions in question do not have CV's equal to 1.

It is easy to see from the tables that as CV_a and/or CV_b increases the waiting time factor also increases for a given utilization. This simply means that the manufacturing lead time increases and the WIP increases EVEN THOUGH THE AVERAGE INPUT AND OUTPUT RATES REMAIN UNCHANGED.

Example 2

Refering to EXAMPLE 1 assume that a decision is made to train the operator. Assume further that as a result of training his SD drops to 1,75 minutes but that his AVERAGE OUTPUT RATE REMAINS THE SAME. In other words he is working more consistently but not more quickly. Assume that all other factors such as input variation etc. remain unchanged.

Then CV_b becomes 0,333 and the waiting time factor (from tables) becomes 0,374 so that the new value of E(W') is 1,97 minutes - a reduction of about 63%.

Furthermore, E(Q') is insignificant, and there is no longer a waiting-line of items. In the light of these new findings it is obvious that there exists a potential for increasing the utilization of the workcentre (and its capacity) by increasing the input rate.

Note once again that the improvement was obtained by reducing variability only, and not through any reduction in the average job processing time.
The benefits of variability reduction are obvious from the tables in the appendix. For example, in the case of a workcentre with utilization 0,9 the waiting time factor is found to be 9,0 for the M/M/1 case, but it reduces to 0,74 if the CV's of the distributions are around 0,3. THIS REPRESENTS A 92% REDUCTION IN WAITING TIME.

THE KANBAN (J-I-T) OPTION AND BALKING QUEUES

The use of a PULL philosophy instead of the conventional PUSH system is another method for reducing lead times and WIP. However, the introduction of JIT involves many organization-wide changes (philosophical and otherwise). Lead time reduction via JIT is, therefore, not simply an exercise in reducing CV's. It involves a cultural revolution in the company, and takes a long time to implement.
Before introducing KANBAN systems on the shopfloor, it is often

advisable to get some idea regarding the extent to which WIP reduction and capacity increase are likely to be achieved.
This is easily done and can be illustrated using the M/M/1 model for a balking system as shown below.

Example 3

Assume that the average input and output rates for a bottleneck workcentre are 10 and 12,5 jobs per hour respectively. The utilization is 0,8. Using the conventional M/M/1 model we can derive a 'worst case' scenario in which:

$E(Q') = 4$ units and $E(W') = 24$ minutes.

However to convey a more complete picture of the situation use must made of the probability distribution of the queue given by:
$P(n) = a^n P(0)$ $n >= 0$;
where $P(0) = 1 - a = 0,2$ is the proportion of time that the workcentre is idle, and n is the number of units in the queue. From this relationship it can be shown that the queue is composed of more than four items 27% of the time, and is composed of more than 8 items (i.e. twice the average) for 15% of the time. This represents a substantial amount of WIP - not obvious from consideration of the expected value of the queue alone.
The long right hand tail is characteristic of WIP distributions and is the reason why bottleneck stations are often purposefully under supplied in practice in an attempt to come to grips with the WIP problem. But there is a disadvantage in under supplying a bottleneck, and this is simply that it automatically results in under-utilization of resources.

In order to gain better control of WIP and to limit it to a maximum number of items a pull philosophy can be used. To do this in the above example, a bin of size four (say) can be placed at the workstation, and the rule associated with the bin could simply be that when it is full no more items are to be supplied. The resulting behaviour is that of a balking system as shown in the table below:

WIP	INPUT RATE	OUTPUT RATE
0	$\alpha_0 = 10$	$\mu_0 = 0$
1	$\alpha_1 = 10$	$\mu_1 = 12,5$
2	$\alpha_2 = 10$	$\mu_2 = 12,5$
3	$\alpha_3 = 10$	$\mu_3 = 12,5$
4	$\alpha_4 = 0$	$\mu_4 = 12,5$

[Note that the system can have progressive balking in which α_n reduces systematically from 10 to zero instead of having the abrupt cut-off shown in the table. The mathematics remains the same, only the values of α_n changing in what follows. Note also that the WIP need not be limited to $n <= 4$ but can be of any quantity - the only mathematical change will be a logical extension of the series in the calculations.]

The equation describing statistical equilibrium for this system is:

$$\alpha_n P(n) = \mu_{n+1} P(n+1) \text{ for } n = 0, 1, 2, 3, 4$$

Hence $P(0) = [1 + \alpha_0/\mu_1 + \ldots\ldots + (\alpha_0\alpha_1\alpha_2\alpha_3/\mu_1\mu_2\mu_3\mu_4)]^{-1} = 0,3$

$U = 0,7 = 70\%$
$E(Q') = 1,6$ units
$E(W') = 9,6$ minutes
Q' max = 4 units

The points to note are that waiting time is reduced to 9,6 minutes, while the utilization of the centre is now only 70%. This immediately opens up opportunities for increasing the throughput of the centre by increasing the input rate α - in other words by REVERSING the previous practice of under supplying the bottleneck. Assume that the input rate is doubled so that $\alpha_n = 20$ in the above table. The results become:

$P(0) = 0,06$
$U = 94\%$
$E(Q') = 2,7$ units
$E(W') = 8,1$ minutes
Q' max = 4 units

It is seen from these answers that once a balking (KANBAN-like) system is implemented WIP is controlled while manufacturing lead times are automatically reduced, and increases in throughput are possible.
In the above example the M/M/1 model (and its assumptions) have been used to illustrate the benefits of Kanban-like systems. However, further dramatic improvement is possible if the CV of the service time is reduced. In a simulation study of J-I-T systems Muraldihar et al [7] have shown that reducing the CV from 1 to 0,5 increases capacity by approximately 18%, and if the CV is still further reduced to 0,2 capacity increases by yet another 20%. Thus, theoretically, a decrease in the CV from 1 to 0,2 can boost capacity by 38%.

CONCLUSION

While elementary queueing theory has often been used to help solve production bottleneck problems, it is too unrealistic to be of great practical value. The main weakness is its lack of sensitivity to changes in the coefficients of variation of the input/ouput distributions.
It is only sensitive to changes in the AVERAGE input/output rates per se.
This drawback is a consequence of the fact that it assumes that the CV's for both time distributions are always equal to 1.
In the real world of industry this assumption is unrealistic.
Industrialists have known about these shortcomings for some time of course, but have continued to use the elementary models because of the need for quick answers, and because the use of more

sophisticated approaches requires a level of statistical expertise which they often do not posses.
This paper addresses that very problem.
It provides a means whereby virtually anyone (no matter how unskilled in the use of quantitative methods) can calculate expected queue lengths and expected job waiting times so as to be able to understand, and to improve, existing processes.

FAR MORE IMPORTANT, it enables engineers to design better new systems.
They can preempt future bottleneck situations, and find out how to eliminate them, long before they become a reality.
Regarding the tables supplied in the appendix, there is one underlying assumption, namely, that the real world data follows an Erlang Distribution. This is not an unrealistic assumption, and in practice it is usually quite unnecessary to check it out using statistical hypothesis tests.
Muralidhar et al [7] have shown that capacity increase as a result of reduction in the CV is virtually independent of the nature of the underlying service time distribution.

REFERENCES:

1. COOPER R B, INTRODUCTION TO QUEUEING THEORY. Macmillan (1972).

2. RAVINDRAN A, ET AL OPERATIONS RESEARCH PRINCIPLES AND PRACTICE. Wiley (1987)

3. WILLIAMS B R, Capacity's Hidden Dimension: Variability, APICS CONFERENCE PROCEEDINGS, 1990, pp 212-214.

4. JEWELL S J, A Simple Proof of L = αW, OPERATIONS RESEARCH, Vol(15) (6) 1967, pp1109-1116.

5. COX D R SMITH W L QUEUES, Chapman & Hall, London, 1971.

6. PAGE E, QUEUEING THEORY IN OR, Butterworths, 1972.

7. MURALDIHAR K et al Describing Processing Time When Simulating JIT Environments, INT. J. PROD RES, 30(1) 1992, pp 1-11.

AVERAGE WAITING-TIME FACTOR FOR A SINGLE SERVER.

UTILIZATION = 0,5

VALUES OF CV_a	VALUES OF CV_b									
	1	0,707	0,577	0,500	0,447	0,408	0,378	0,354	0,333	0,316
0,316	0·3246	0·1326	0·0785	0·0545	0·0415	0·0336	0·0282	0·0245	0·0217	0·0195
0,447	0·3962	0·1924	0·1313	0·1029	0·0867	0·0763	0·0691	0·0638	0·0598	0·0566
0,548	0·4692	0·2559	0·1897	0·1581	0·1396	0·1276	0·1192	0·1129	0·1081	0·1043
0,632	0·5432	0·3221	0·2520	0·2150	0·1978	0·1846	0·1753	0·1683	0·1630	0·1587
0,707	0·6180	0·3904	0·3170	0·2810	0·2597	0·2456	0·2356	0·2281	0·2223	0·2177
0,775	0·6935	0·4602	0·3842	0·3467	0·3243	0·3095	0·2990	0·2911	0·2850	0·2801
0,837	0·7696	0·5313	0·4531	0·4142	0·3911	0·3756	0·3647	0·3564	0·3501	0·3450
0,894	0·8460	0·6034	0·5232	0·4833	0·4594	0·4435	0·4322	0·4237	0·4171	0·4118
0,949	0·9229	0·6764	0·5945	0·5537	0·5292	0·5129	0·5012	0·4925	0·4857	0·4803
1,000	1·0000	0·7500	0·6667	0·6250	0·6000	0·5833	0·5714	0·5625	0·5556	0·5500

AVERAGE WAITING-TIME FACTOR FOR A SINGLE SERVER.

UTILIZATION = 0,6

VALUES OF CV_a	VALUES OF CV_b									
	1	0,707	0,577	0,500	0,447	0,408	0,378	0,354	0,333	0,316
0,316	0·5786	0·2600	0·1643	0·1200	0·0951	0·0793	0·0684	0·0606	0·0547	0·0501
0,447	0·6786	0·3488	0·2463	0·1973	0·1689	0·1504	0·1375	0·1280	0·1207	0·1149
0,548	0·7796	0·4407	0·3330	0·2808	0·2500	0·2299	0·2156	0·2050	0·1968	0·1903
0,632	0·8813	0·5348	0·4232	0·3684	0·3360	0·3146	0·2994	0·2880	0·2793	0·2723
0,707	0·9835	0·6306	0·5157	0·4590	0·4253	0·4029	0·3870	0·3751	0·3659	0·3586
0,775	1·0862	0·7278	0·6102	0·5519	0·5171	0·4940	0·4775	0·4652	0·4556	0·4480
0,837	1·1892	0·8259	0·7061	0·6465	0·6108	0·5871	0·5702	0·5576	0·5477	0·5399
0,894	1·2926	0·9249	0·8031	0·7424	0·7061	0·6819	0·6646	0·6516	0·6416	0·6335
0,949	1·3962	1·0247	0·9012	0·8395	0·8025	0·7778	0·7603	0·7471	0·7369	0·7287
1,000	1·5000	1·1250	1·0000	0·9375	0·9000	0·8750	0·8571	0.8438	0·8333	0·8250

AVERAGE WAITING-TIME FACTOR FOR A SINGLE SERVER.

UTILIZATION = 0,7

VALUES OF CV_a	VALUES OF CV_b									
	1	0,707	0,577	0,500	0,447	0,408	0,378	0,354	0,333	0,316
0,316	1·0198	0·4908	0·3252	0·2462	0·2006	0·1711	0·1506	0·1355	0·1240	0·1150
0,447	1·1642	0·6248	0·4526	0·3689	0·3198	0·2876	0·2649	0·2480	0·2350	0·2247
0,548	1·3093	0·7612	0·5840	0·4970	0·4456	0·4116	0·3875	0·3695	0·3556	0·3445
0,632	1·4548	0·8994	0·7183	0·6288	0·5756	0·5404	0·5153	0·4966	0·4821	0·4705
0,707	1·6006	1·0391	0·8547	0·7633	0·7088	0·6726	0·6468	0·6275	0·6125	0·6005
0,775	1·7468	1·1798	0·9927	0·8997	0·8442	0·8072	0·7808	0·7611	0·7458	0·7335
0,837	1·8932	1·3214	1·1321	1·0378	0·9813	0·9437	0·9169	0·8968	0·8812	0·8687
0,894	2·0397	1·4637	1·2724	1·1770	1·1198	1·0817	1·0545	1·0342	1·0183	1·0056
0,949	2·1865	1·6066	1·4136	1·3172	1·2595	1·2209	1·1934	1·1728	1·1568	1·1439
1,000	2·3333	1·7500	1·5556	1·4583	1·4000	1·3611	1·3333	1·3125	1·2963	1·2833

AVERAGE WAITING-TIME FACTOR FOR A SINGLE SERVER.

UTILIZATION = 0,8

VALUES OF CV_a	VALUES OF CV_b 1	0,707	0,577	0,500	0,447	0,408	0,378	0,354	0,333	0,316
0,316	1·9222	0·9744	0·6692	0·5205	0·3916	0·3758	0·3355	0·3056	0·2826	0·2644
0,447	2·1523	1·1947	0·8831	0·7299	0·6391	0·5791	0·5366	0·5049	0·4803	0·4608
0,548	2·3826	1·4168	1·1005	0·9440	0·8509	0·7891	0·7452	0·7124	0·6869	0·6667
0,632	2·6132	1·6405	1·3203	1·1614	1·0665	1·0035	0·9586	0·9250	0·8989	0·8781
0,707	2·8440	1·8653	1·5419	1·3811	1·2849	1·2209	1·1753	1·1411	1·1146	1·0934
0,775	3·0750	2·0910	1·7650	1·6026	1·5053	1·4406	1·3944	1·3598	1·3329	1·3114
0,837	3·3061	2·3174	1·9892	1·8254	1·7273	1·6620	1·6153	1·5804	1·5532	1·5315
0,894	3·3573	2·5444	2·2143	2·0494	1·9506	1·8847	1·8377	1·8025	1·7750	1·7531
0,949	3·7686	2·7720	2·4402	2·2743	2·1749	2·1086	2·0612	2·0257	1·9981	1·9760
1,000	4·0000	3·0000	2·6667	2·5000	2·4000	2·3333	2·2857	2·2500	2·2222	2·2000