An Examination of Some Factory Physics Principles

Hong Chen
Shanghai Advanced Institute of Finance (SAIF)
Shanghai Jiaotong University, Shanghai, China
E-mail: hchen@saif.sjtu.edu.cn

and

Heng-Qing Ye
Faculty of Business, Hong Kong Polytechnic University
Hong Kong, China
E-mail: lgtyehq@polyu.edu.hk

May 2015; revision: July 2015

Abstract

In this paper, we examine some principles in managing manufacturing systems. These principles are concerned with the variability, the utilization, the rework, the lead time, and the CONWIP efficiency. While these principles are developed through analyzing some simpler disconnected flow line manufacturing systems, we examine whether they can have broad applications. For some of these principles, we provide sufficient conditions, while for others, we provide counterexamples. Our analysis suggests that we should be very cautious about these laws when applied to non-Markov and non-tandem systems.

Key words: factory physics, manufacturing system, variability, utilization, rework, CONWIP, lead time.

1 Introduction

> "For the scientist, there is only 'being', but no wishing, no good, no evil - in short, no goal." (Albert Einsten 1951)

Though there has been a more than century long history of scientific management and study of manufacturing system, most intensive research on manufacturing system has happened in the 1980s and the early 1990s. (For a history account, see, for example, Chapter 1 of Hopp and Spearman (2011).) The more recent research is built upon the success of operations research during the World War II and its rapid development in the 1960s with the emerging computer technology, and is stimulated by the challenge from Japanese manufacturers in the 1980s. There have been hundreds, if not thousands, of research papers on studying various manufacturing models from their performance analysis, optimal design to optimal control. For some references in this vast literature, readers are referred to references in books by Kelly (1979) and Buzacott and Shanthikumar (1992) and Yao (1994).

Manufacturing systems vary with products and manufacturing processes. Based on the material flows in manufacturing systems, manufacturing processes have been classified into job shops, flow lines and continuous flow processes. While most research results in manufacturing systems are about specific systems and summarized as Theorems or Propositions, it seems natural to search for results that have broad applicability and that can be called principles or laws. In physics, there are laws such as Newton's 2nd Law, and in probability theory, there are laws such as Strong Law-of-Large-Numbers and Law-of-Iterated-Logarithm. In manufacturing systems, the most widely applicable result, Little's Formula, has also been known as Little's Law, which states that the average number of jobs in a system equals the average rate of jobs entering the system multiplied by the average time jobs spent in the system.

In the award-winning book by Hopp and Spearman (2011, first edition published in 1996), the authors proposed some factory physics principles through analyzing a class of what are called disconnected flow line manufacturing systems. These principles include the laws such as Little's Law, the law of variability, the law of utilization, the law of rework, the law of lead time, the law of CONWIP efficiency (where CONWIP is short for Constant Work-In-Process), among others.

All of these laws are simple and easy to communicate. According to *Oxford Dictionary*, a scientific law is a "factual statement of what always happens in certain circumstances". In physics, the validation of a law is usually done through numerous experiments. In mathematics, the validation of a law is usually done through a proof (under some more fundamental axioms), though experiments have also been used by mathematicians. For example, a number of mathematicians experimented with flipping coins to test the Strong Law-of-Large-Numbers. Computer simulation is also a convenient tool of use for experiments. The experiments or the proofs can be used not only to validate a law, but also to identify the circumstances under which it is valid. Little's Law was introduced by Little (1961) and proved by many authors under different circumstances; interested readers are referred to Whitt (1991) and El-Taha and Stidham (1999) for references. However, all of the other factory physics laws mentioned above have not been as well-studied in terms of validation and their general applicability. Our preliminary analysis seems to suggest that except for the well-studied Little's Law, all of the other factory physics laws mentioned above have limited applicability when applied to non-Markov and non-tandem systems. Thus, it may be questionable whether they should be called laws, unless further research can establish a broader class of circumstances under which they hold.

In this paper, we examine the above mentioned factory physics laws. Instead of identifying the exact circumstances or conditions that are necessary and sufficient for a law to be true, we have a modest objective: for some laws, we identify or cite some sufficient conditions for them to hold, while for some other laws, we find some counterexamples. In our analysis, we do not restrictive ourselves to the class of manufacturing systems studied by Hopp and Spearman (2011), which are essentially single class tandem networks. We put these laws into test among a much larger class of manufacturing systems, since modern manufacturing processes have more variety and are more complex. The most notable example is the wafer fabrication process, where the raw wafer may repeatedly visit the same set of stations for more than one hundred operations. Our motivation is to understand whether the proposed laws have broader applicability and what their limitations are. Therefore, the counterexamples should not be taken as criticism towards the laws that were proposed from studying a more restrictive class of systems. (Even the celebrated

Newton's 2nd Law has limitations that give the way to Special Relativity, Quantum Mechanics and General Relativity.) Indeed, in proposing the laws of factory physics, Hopp and Spearman (2011) noted that "some of these 'laws' are *always* true (e.g., Law 5 [Capacity]), while others hold *most of the time*."

In describing the above laws, a manufacturing/production system can be viewed as a queueing system. Materials that move through the manufacturing system are jobs in the queueing system. Two of the key performance measures in the manufacturing system, the WIP inventory and the cycle time, are corresponding to the number of jobs (or the queue length) and the sojourn time in the queueing system, respectively. Therefore, we examine the above principles in the context of queueing models. For the law of variability, we present two counterexamples: a GI/G/1 queue and a multiclass queueing network. In the latter case, we show that it is possible to make a stable system into an unstable system by reducing the variability of the interarrival times and service times. For the law of utilization, we point out some positive results for Jackson networks, and provide a counterexample of a multiclass queueing network where by reducing the utilization a stable system can be made unstable. Counterexamples are also presented for the law of rework where reduced rework can make a stable system into an unstable system, and for the law of lead time where under the same lead time, a system with larger mean and standard deviation of the cycle time may offer a better service level. We show that for a Jackson network, the law of CONWIP is true if the service rates at all stations are increasing concave, and the converse is true if the service rates are increasing convex. Some results and examples presented here are either known before (e.g., the law of variability does not apply for many GI/G/1 queues (Whitt 1984)) or are variations of existing ones in literature, and wherever possible we will highlight the sources of the original ones below. The result in Theorem 6.1, concerning the CONWIP for the Jackson network, is a new addition to the literature.

The paper is organized by the laws mentioned above with each corresponding to a section, and we conclude in section 7.

2 Law of Variability

> *Law of Variability:* Increasing variability always degrades the performance of a production system.

The most appropriate measure for the variability of a random variable is probably its coefficient of variation, defined by the ratio of its standard deviation to its mean; or equivalently the squared coefficient of variation. Two most common performance measures of a production system are the average work-in-process (WIP) inventory and the average cycle time. If the system is modeled by a queue, the WIP is the queue length or the number of jobs in the system, and the average cycle time is the average sojourn time or the average waiting time (including service time). In general, it is clear that it is desirable for a system to have a lower WIP and a smaller cycle time. With the throughput rate (or equivalently the same average exogenous arrival rate) fixed, by Little's Law, the WIP increases if and only if the cycle time increases.

With the coefficient of variation as the measure of the variability and the WIP as the performance measure of the system, we provide two counterexamples to the law of variability

First, we compare two $GI/G/1$ queueing systems with the same interarrival times, the same mean service times but different coefficients of variation for service times. Specifically, we assume that there are no initial jobs in both queues, and both queues have deterministic interarrival times of $u = 2$. The service time distribution of the first queue is assumed to be a uniform distribution $U[1-\theta, 1+\theta]$, where $\theta \in (0,1)$ is a constant. Whereas, the service time distribution of the second queue takes value $1-\epsilon$ with probability p and a with $1-p$, where $p = (a-1)/(a-1+\epsilon)$, and $a > u = 2$ and $\epsilon \in (0,1)$ are constants. It is clear that the mean service times of both queues are equal to 1. Hence, their squared coefficients of variation equal their variances, which are given by

$$\begin{aligned} \sigma_1^2 &= \frac{\theta^2}{3}, \quad \text{and} \\ \sigma_2^2 &= \epsilon^2 \frac{a-1}{a-1+\epsilon} + (a-1)^2 \frac{\epsilon}{a-1+\epsilon}, \end{aligned}$$

respectively. Note that $\sigma_2^2 \to 0$ as $\epsilon \to 0$; hence, when ϵ is sufficiently small, $\sigma_1^2 > \sigma_2^2$. In words, when ϵ small enough, the second queue has a smaller variability than the first queue. On the other hand, jobs in the first queue never need to wait since their interarrival times are exactly 2 units of time and their service times never exceeds $1 + \theta < 2$ units of time. Jobs in the second queue may have to wait since their service times (which can take a value of $a > 2$ with probability $1-p$) may exceed the interarrival times ($u = 2$ units of time). Note that both of the queues have the same mean service time; hence, the first queue has a smaller average sojourn time (or cycle time) than the second queue. In summary, the second queue with a smaller variability performs worse than the first queue with a larger variability. This contradicts the law of variability.

The violation of the law of variability can be more extreme. Consider a multiclass queueing network, which is a variation of the network studied by Bramson (1999), as shown by Figure 1. We shall show that under certain parameters, this network is stable (implying a finite expected cycle time); but with reduced variability, it can be unstable, implying that the cycle time approaches infinity. This network consists of four single-server stations indexed by $j = 1, \ldots, 4$. At each station, there are two classes of jobs indexed as class a and b respectively. Each job is processed at stations 1 and 2 sequentially as a class b job first, and then is routed to station 3 and station 4 alternately. At station 3 or 4, the job is processed as a class a and then as a class b job sequentially, and then is turned back and processed by stations 2 and 1 sequentially as a class a job. The interarrival times and service times of jobs are assumed to be i.i.d. sequences with uniform distributions; the specific parameters are given in the following table, where $\theta \in [0, 1]$ is a constant.

Interarrival time			$U[1-\theta, 1+\theta]$
Service time	Station 1	a	$U[0.6(1-\theta), 0.6(1+\theta)]$
		b	$U[0.1(1-\theta), 0.1(1+\theta)]$
	Station 2	a	$U[0.1(1-\theta), 0.1(1+\theta)]$
		b	$U[0.6(1-\theta), 0.6(1+\theta)]$
	Stations 3 and 4	a	$U[1.2(1-\theta), 1.2(1+\theta)]$
		b	$U[0.05(1-\theta), 0.05(1+\theta)]$

We also assume that all the interarrival times and service times are mutually independent. Jobs are processed following work-conserving preemptive priority disciplines with class a having a higher priority at each station. In other words, jobs of class b cannot be processed at a station unless there are no jobs of class a at the same station.

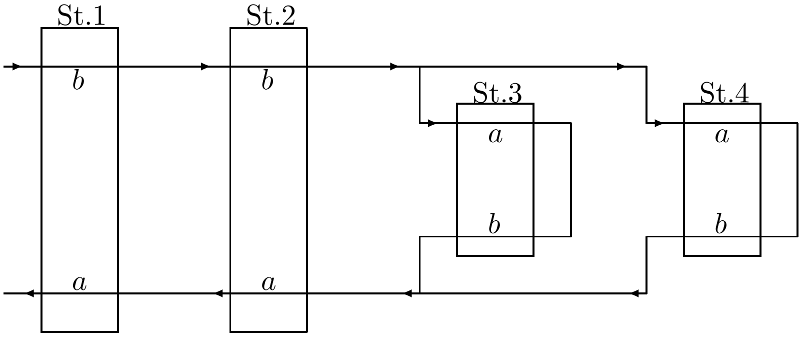

Figure 1: A network that shows the Law of Variability may fail.

It is clear that as θ increases, the variability of interarrival times and service times increases. We perform three simulation runs of the network corresponding to the network with high variability ($\theta = 1$), low variability ($\theta = 0.001$) and no variability ($\theta = 0$). In all the simulations, initially the number of jobs for each class is set at zero, except that there are 2 jobs of class b at station 2. (The non-zero initial queue length condition is essential for the conclusion to hold in the deterministic case ($\theta = 0$), but not for the other two cases.) Figures 2-4 plot the total number of jobs in the network for each of the three simulation runs. From these figures, the network with high variability ($\theta = 1$) is stable while the network with low variability ($\theta = 0.001$ or $\theta = 0$) is not stable. In fact, by a tedious pathwise construction, we can prove that for the deterministic case ($\theta = 0$) the total number of jobs in the network goes to infinity with time. For a similar but more complex network (adding many more stations like stations 3 and 4 to the network shown in Figure 1), Bramson (1999) proves that the network is unstable under the Poisson arrival and exponential service times, but its corresponding deterministic fluid model is unstable. (The network with deterministic interarrival times and deterministic service times provides a very close approximation to an unstable solution of the fluid model.) In summary, this example provides another counterexample to the law of variability.

In this last network example, if the network operates under a different service discipline, then we may reach a different conclusion. In particular, if the network operates "optimally" under

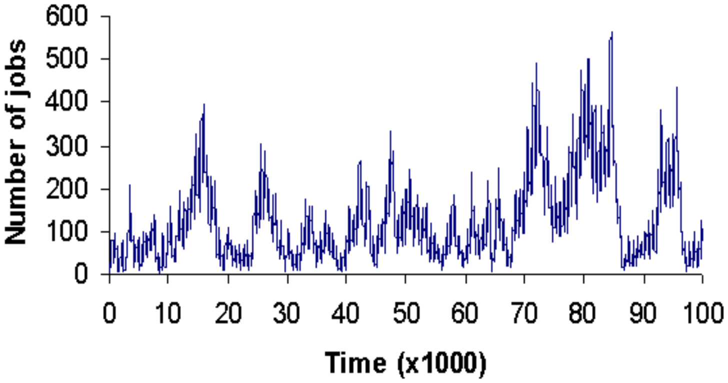

Figure 2: The network as shown by Figure 1 is stable with uniformly distributed service times of high variability ($\theta = 1$).

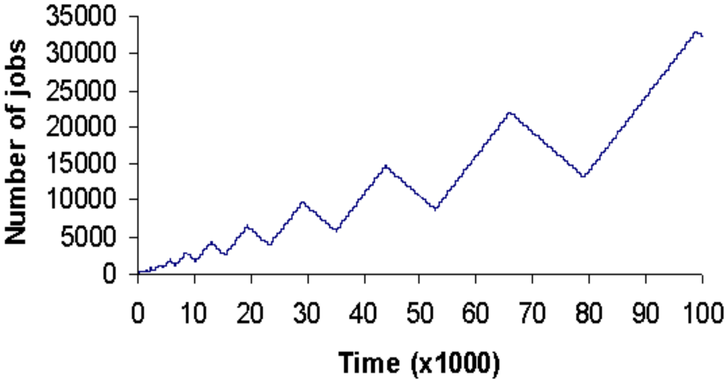

Figure 3: The network as shown by Figure 1 is unstable with uniformly distributed service times of low variability ($\theta = 0.001$).

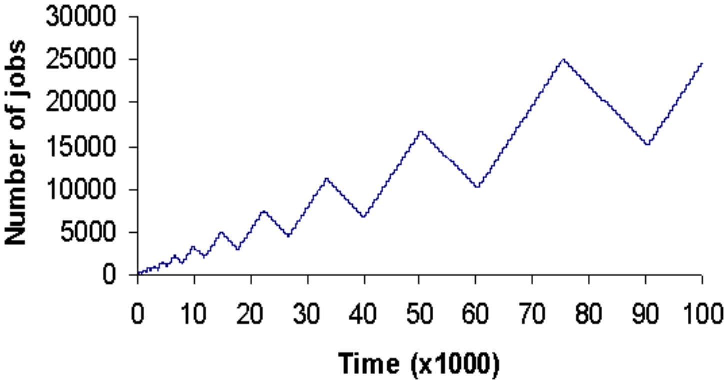

Figure 4: The network as shown by Figure 1 is unstable with deterministic service times ($\theta = 0$).

each of the given parameters, then it could well be that increasing variability would degrade the performance of the network. However, the optimality depends on the specific objective; even with a given objective, it is in general an open problem in terms of identifying an optimal service discipline and its corresponding performance measure.

3 Law of Utilization

> *Law of Utilization:* If a station increases utilization without making any other changes, average WIP and cycle time will increase in a highly nonlinear fashion.

We first provide some positive results for Jackson networks and then a counterexample. Note that the utilization of a station is commonly defined as the long-run average fraction of time this station is busy.

First consider a birth-death queue with a constant arrival rate λ and a state-dependent service rate $\mu(n)$, where $n \geq 0$ is the number of jobs in the system. Then its queue length process has a stationary distribution if and only if

$$\sum_{n=1}^{\infty} \frac{\lambda^n}{M(n)} < \infty, \qquad \text{where} \qquad M(n) = \prod_{k=1}^{n} \mu(k).$$

When the stationary distribution exists, the utilization of the server is

$$1-\left[1+\sum_{n=1}^{\infty}\frac{\lambda^n}{M(n)}\right]^{-1}.$$

This utilization increases as either the arrival rate λ increases or the service rate $\mu(n)$ decreases for any given n. (Note that by changing the time scale, increasing λ can also be viewed as decreasing $\mu(n)$ for all $n \geq 1$.) It is also known that as the arrival rate λ increases or the service rate μ decreases, the number of jobs in the system (equivalent to WIP) increases. (In fact, the latter increase has been proved in terms of likelihood ratio ordering by Shanthikumar and Yao (1986a,b); also see Example 1.15 in Chen and Yao (2001).) By Little's Law, the same monotonicity also holds for the cycle time. Therefore, the law of utilization holds for this birth-death queue. This conclusion immediately extends to an open Jackson network, because of the product form stationary distribution of its queue length process.

Next, consider a variation of the network first introduced by Lu and Kumar (1991). The network is as shown by Figure 5 with three single server stations. Jobs enter the network visiting or revisiting stations 1, 2, 3, 3 and 1 sequentially. At station 1, jobs of class 5 have a preemptive higher priority over jobs of class 1, and at station 2, jobs of class 3 have a preemptive higher priority over jobs of class 4. Assume Poisson arrivals with rate 1 and exponential

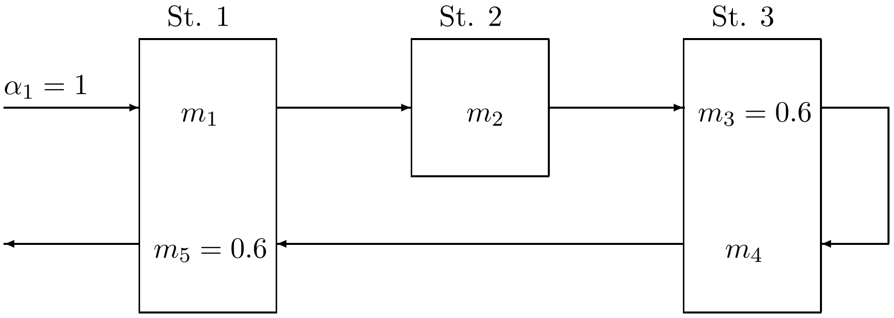

Figure 5: A network that shows the Law of Utilization may fail.

service times. Consider two sets of parameters for the mean service time $m = (m_1, \ldots, m_5)$: $m^1 = (0.2, 0.9, 0.6, 0.2, 0.6)$ and $m^2 = (0.1, 0.1, 0.6, 0.1, 0.6)$. It seems clear that under m^1, the utilizations are 0.8, 0.9 and 0.8 at stations 1, 2 and 3, respectively, and that under m^2, the utilizations are 0.7, 0.1 and 0.7 at stations 1, 2 and 3, respectively. In particular, we note that the utilization under m^1 are greater than the utilization under m^2 at each of the corresponding stations. Under m^1, we can show that this network is positive recurrent (using, for example, the sufficient condition in Chen and Zhang 2000). In particular, the (long-run) average number of jobs at each station is finite. Figure 6 provides a simulation result for this case. On the other hand, under m^2, we can show that this network is unstable, meaning that the total number of jobs in the network approaches infinity (using, for example, the conditions in Dai (1996) or Meyn (1995)). Figure 7 provides a simulation result. Therefore, we find a counterexample that increasing the utilization at every station of a network can decrease average WIP dramatically (from an infinite average to a finite average).

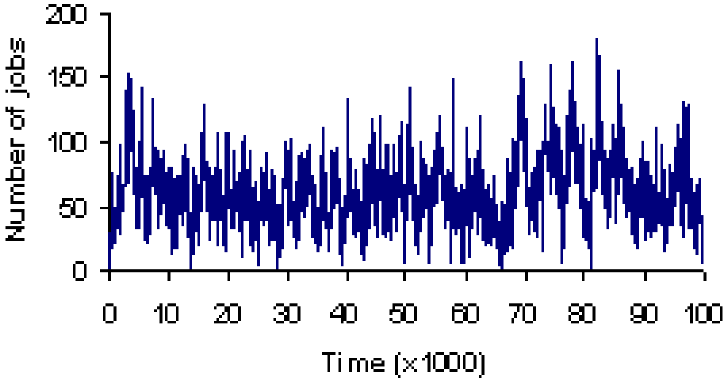

Figure 6: The network as shown by Figure 5 is stable when $m = (0.2, 0.9, 0.6, 0.2, 0.6)$ and the (nominal) utilizations are 0.8 at station 1, 0.9 at station 2 and 0.8 at station 3.

We note that under m^2, the long-run average total number of jobs at each of stations 1 and

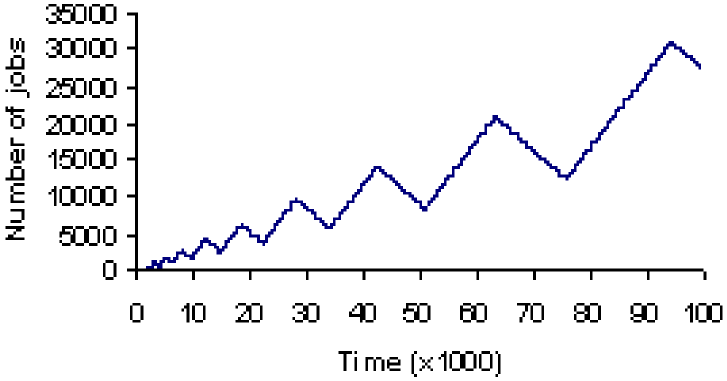

Figure 7: The network as shown by Figure 5 is unstable when $m = (0.1, 0.1, 0.6, 0.1, 0.6)$ and the (nominal) utilizations are 0.7 at station 1, 0.1 at station 2 and 0.7 at station 3.

3 does not exist, i.e., the limits of

$$\frac{1}{T}\int_0^T [Q_1(t) + Q_5(t)]dt \qquad \text{and} \qquad \frac{1}{T}\int_0^T [Q_3(t) + Q_4(t)]dt$$

as $T \to \infty$, do not exist. However, their limit superiors are infinity. We also note that if the utilization is defined as the long-run average fraction of time a station is busy, then under m^2 the utilization does not exist for any of the three stations. This is because none of the following limits as $T \to \infty$,

$$\frac{1}{T}\int_0^T 1_{\{Q_1(t)+Q_5(t)>0\}}dt, \qquad \frac{1}{T}\int_0^T 1_{\{Q_2(t)>0\}}dt,$$
$$\frac{1}{T}\int_0^T 1_{\{Q_3(t)+Q_4(t)>0\}}dt,$$

exist, where 1_A is an indicate function for event A. (However, we note that their limit inferiors are less than the utilizations under m^1 at the corresponding stations.) This suggests the network under m^2 do not have a utilization under our usual definition of the utilization (i.e., the utilization is the ratio of (actual) arrival rate to the effective service rate). The utilizations we calculated for the network under m^2 above (namely, 0.7, 0.1 and 0.7 respectively for stations 1, 2 and 3) should be called nominal utilizations, rather than (actual) utilizations.

In fact, the counter-intuitive phenomenon of the above network example would occur for other realistic scheduling rules as well. Specifically, it would not be difficult to construct similar counter-examples where shortest-expected-remaining-processing-time-first (SERPT) rule, shortest-expected-service-time-first (SEPT) rule, or first-in-first-out (FIFO) rule is applied. These can be done by suitably adding a regulator (such as the station 2 in Figure 5) to some existing network examples, for example, the six class version of Kumar-Seidman network in Chen and Yao (2001, 339-241). Finally, it should be noted that these network examples are simplified models of more complex manufacturing systems in particular in the semi-conductor industry; see for example Kumar (1993). A better understanding of the impact of (nominal) utilization on the network performance would be helpful in managing more complex systems.

4 Law of Rework

> *Law of Rework:* For a given throughput level, rework increases both the mean and standard deviation of the cycle time of a process.

We describe a counterexample for this law, by the network as depicted in Figure 8, which is modified from the network in Figure 5. The modification is that with probability p_r, each job after processed at station 2 (regardless its history) will require a rework. The mean service times are as indicated in the figure. Since we have assumed that all service times are exponentially distributed, the queueing length process of this network has the same distributional behavior as the network in Figure 5 with $m_2 = 0.1/(1 - p_r)$. Therefore, it follows from the discussion for the network in Figure 5 that the network shown by Figure 8 is unstable when there is no rework, i.e., $p_r = 0$, and it can be shown that the network is stable when the rework probability is increased to $p_r = 8/9$. This clearly contradicts the Law of Rework.

In the context of complex manufacturing systems, the Law of Rework is valid for some important classes of networks, for example, the Jackson network and BCMP network; see for example Jackson (1957, 1963), Baskett, *et al.* (1975) and Kelly (1979). For these networks, the so-called product-form solution exists, which enables one to analyze each station in the network

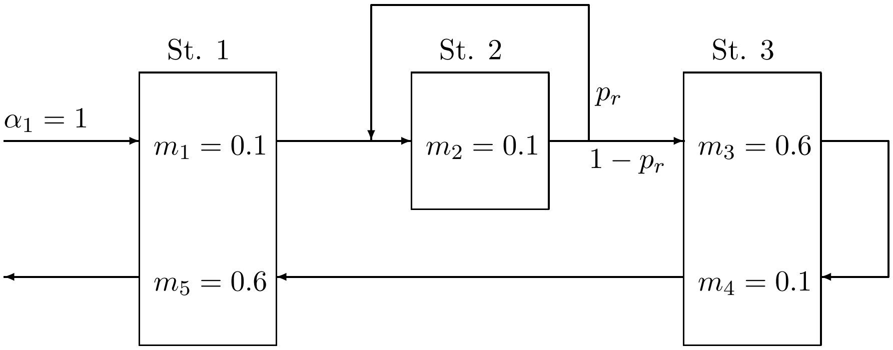

Figure 8: A network that shows the Law of Rework may fail.

independently. In such cases, the rework leads to an increase in the traffic intensity and thus the WIP of some stations.

5 Law of Lead Time

> *Law of Lead Time:* The manufacturing lead time for a routing that yields a given service level is an increasing function of both the mean and standard deviation of the cycle time of the routing.

The lead time of a given routing or line is the time allotted for production of a part on that routing or line. Hence, it is a management constant. In contrast, cycle times are generally random. Therefore, in a make-to-order environment (where parts are produced to satisfy orders with specific due dates), an important measure of line performance is the service level, which is defined as

$$\text{Service level} \quad = \quad \mathsf{P}\{\text{cycle time} \leq \text{lead time}\}.$$

Fix the service level at s $(0 < s < 1)$. If the cycle time follows a normal distribution with mean m and standard deviation σ, then the lead time is given by $L_n = m + \sigma\Phi^{-1}(s)$, where Φ^{-1} is the inverse function of the standard normal cumulative distribution function. In this case, it is clear that the lead time is increasing in both the mean (m) and the standard deviation (σ). Next suppose that the cycle time follows a uniform distribution between $(m-a)$ and $(m+a)$ (where $0 \leq a \leq m$). Then the lead time is given by $L_u = m + (2s-1)a$, which is also increasing in both the mean (m) and the standard deviation ($a/\sqrt{3}$), provided that the service level $s > 0.5$. Now suppose that the cycle time follows an exponential distribution with mean m. In this case, the lead time is given by $L_e = -m \log(1-s)$, which is again increasing in both the mean (m) and the standard deviation (m).

As a brief summary, the above examples suggest that the Law of Lead Time hold when the cycle time varies its mean and standard deviation within the same *family* of a distribution, though it remains to identify all such families of distributions. It seems reasonable to assume that the distribution of the cycle time of a production process will approximately remain within the same family as the process involves a marginal improvement (reducing the mean and the standard deviation of its cycle time); hence, the Law of Lead Time would apply.

The next example shows that this law may fail when different families of cycle time distributions are involved (which may correspond to different manufacturing processes). Suppose that the service level $s = 0.5$. The first cycle time follows a uniform distribution between $(80 - a)$ and $(80 + a)$ (where $0 \leq a < 80$), and the second cycle time follows an exponential distribution with mean 100. Clearly the first cycle time has, respectively, smaller mean (80) and smaller standard deviation ($a/\sqrt{3}$) than the mean (100) and the standard deviation (100) of the second cycle time. On the other hand, the lead time for the first case is $L_u = 80$ is greater than that for the second case $L_e = 100 \times \log 2$ ($69.3 < L_e < 69.4$). We obtain a similar counterexample for the service level $s = 0.75$, by assuming the first cycle time has a uniform distribution between 5 and 185, and the second cycle time has an exponential distribution with mean 100.

The Law of Lead Time may still fail if it is rephrased as "the service level of a manufacturing process for a given lead time is an increasing function of both the mean and standard deviation of the cycle time of the process." Consider two production processes with a given common lead time $L_1 = L_2 = 2$. The cycle times are X_1 and X_2, respectively for the first and the second system. Suppose that the distributions for X_1 and X_2 are $\mathsf{P}\{X_1 = 1\} = 0.8$ and $\mathsf{P}\{X_1 = 3\} = 0.2$, and

$\mathsf{P}\{X_2 = 1\} = 0.9$ and $\mathsf{P}\{X_2 = 10\} = 0.1$. Then it is clear that the first system has a smaller mean and standard deviation than the second system (namely, $\mathsf{E}(X_1) = 1.4 < \mathsf{E}(X_2) = 1.9$ and $\sigma(X_1) = 0.8 < \sigma(X_2) = 2.7$). However, the second system provides a higher service level than the first system (namely, $\mathsf{P}\{X_1 \leq L_1\} = 0.8 < \mathsf{P}\{X_2 \leq L_2\} = 0.9$). This contradicts the rephrased Law of Lead Time.

The Law of Lead Time will always hold with the following modification, "the service level of a manufacturing process for a given lead time increases as the cycle time of the process decreases in stochastic ordering." This positive result follows immediately from the definition of stochastic ordering (see, for example, Ross 1996).

Similar to the comment at the end of the last section, we note that similar counter-examples can also be constructed for more realistic scheduling rules such as SERPT, SEPT and FIFO rules.

6 Law of CONWIP Efficiency

> *Law of CONWIP Efficiency:* For a given throughput, a push system will have more WIP on average than an equivalent CONWIP system.

In this section, we first offer a short explanation for the CONWIP system, and a counterexample to the law of CONWIP Efficiency. Then for a Jackson network, we provide a sufficient condition such that the Law of CONWIP Efficiency holds.

A push system is an *open* system, where jobs enter (or are pushed into) the system following some distribution independent of the number of jobs already in the system. A corresponding CONWIP system is obtained from the open system through controlling the external arrivals of jobs: specifically, a new job enters the system when and only when a job leaves the network. Hence, the CONWIP system has a constant number of jobs in the system (which supports the term, "constant work-in-process (CONWIP) inventory"), and is a *closed* system.

The throughput of a system is the rate at which jobs leave the system. In an open system, the throughput also equals the rate at which jobs enter the system, provided that the open system is (weakly) stable (i.e., all jobs that enter the system will leave the system in a finite amount of time). In the corresponding CONWIP system, the events of leaving and entering coincide.

First we present a counterexample to the Law of CONWIP Efficiency. Consider an open tandem system and its corresponding closed system, as shown in Figure 9 and Figure 10 respectively. Both systems have two stations and deterministic processing times with mean 1 at station 1 and mean $\epsilon(< 1)$ at station 2. When the closed system has a WIP of 2, its throughput is 1, which equals the processing rate at station 1 (the bottleneck station). On the other hand, consider the open system with a deterministic interarrival time of 1. Then its average WIP is $1+\epsilon$, which is less than the WIP of the close system (2). In summary, with the same throughput 1, this open system has a lower WIP than the corresponding closed system.

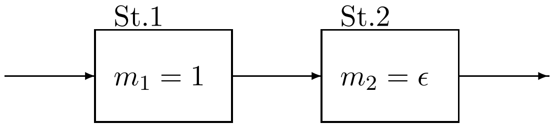

Figure 9: An open network that shows the Law of CONWIP Efficiency may fail.

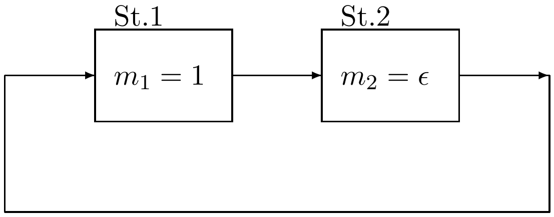

Figure 10: A closed network corresponding to the open network as shown by Figure 9 that shows the Law of CONWIP Efficiency may fail.

Finally, we obtain a sufficient condition for the Law of CONWIP Efficient to hold in a Jackson network. Consider an open Jackson network with J service stations, each with one or several servers. Exogenous arrivals of jobs to the network follow a Poisson process with rate α_0. Each arrived job is independently routed to station j with probability $p_{0j} \geq 0$, $j = 1, \dots, J$, and $\sum_{j=1}^{J} p_{0j} = 1$. The service times of jobs at each station are i.i.d., following an exponential distribution with unit mean. The service rate, i.e., the rate at which work is depleted at each station j can be both station-dependent and state-dependent. Specifically, whenever there are x_j jobs at station j, the processing rate is $\mu_j(x_j)$, where $\mu_j(0) = 0$ and $\mu_j(x) > 0$ for all $x > 0$. Upon service completion at station i, a job may go to another station j with probability p_{ij}, $j = 1, \dots, J$, or leave the network with probability $p_{i0} := 1 - \sum_{j=1}^{J} p_{ij}$, $i = 1, \dots, J$. We assume that the transition matrix $P = (p_{ij})_{i,j=1}^{J}$ is a substochastic matrix with a spectral radius less than unity.

Let $\lambda = (\lambda_j)_{j=1}^{J}$ be the unique solution to the following traffic equation:

$$\lambda_j = \alpha_0 p_{0j} + \sum_{i=1}^{J} \lambda_i p_{ij}, \qquad j = 1, \dots, J.$$

Assume that the condition,

$$\sum_{n=1}^{\infty} \lambda_j^n \Big/ [\prod_{\ell=1}^{n} \mu_j(\ell)] < \infty, \qquad j = 1, \dots, J, \tag{1}$$

holds. In this case, λ_j represents the total arrival rate to station j.

It is clear that the throughput rate of this open network is α_0. However, we would like to provide an alternative representation of the throughput rate. Assume that the network is in equilibrium. Let $\mu(n)$ be the (instant) throughput rate when there are n jobs in the network, and N be the total number of jobs in the network. Then, the throughput of the network is the expected value of $\mu(N)$, $\mathsf{E}\mu(N)$. Let $h(n) = \mathsf{E}[\mu(N)|N = n]$ denote the conditional throughput of the network. Then the throughput can be expressed as

$$\mathsf{E}\mu(N) = \mathsf{E}h(N). \tag{2}$$

The corresponding CONWIP system for the above open network is a closed network, which has the same service stations and processing rates. The only modification is that the transition matrix of this closed network is a probability matrix given by

$$\tilde{p}_{ij} = p_{ij} + p_{i0} p_{0j}, \qquad i, j = 1, \dots, J.$$

Assume that there are n jobs in this closed network and the network is in equilibrium. Then the throughput of this closed network is given by $h(n)$; that is, it is the same as the throughput of the open network conditioned on that there are n jobs in it. (This can be proved following a modification from Section 2.4 of Chen and Yao (2001).)

Note that the Law of CONWIP Efficiency can also be put as follows: a push system has a lower throughput than an equivalent CONWIP system with the same average WIP. If $\mathsf{E}N$ (i.e., the expected number of jobs in the open network) is an integer, then $h(\mathsf{E}N)$ is the throughput rate of the equivalent closed network (CONWIP system) with the same average WIP. In view of (2), the Law of CONWIP Efficiency holds if and only if

$$\mathsf{E}h(N) \leq h(\mathsf{E}N).$$

The above holds if $h(\cdot)$ is a concave function, and the reversed inequality of the above holds if $h(\cdot)$ is a convex function. It follows from Theorem 3.22 in Chen and Yao (2001) that the throughput $h(n)$ in the closed network is increasing and concave (respectively increasing convex) in the job population n if the service rates at all stations are increasing and concave (respectively increasing convex). Hence, we have the following.

Theorem 6.1 *Assume that the condition (1) holds for the open Jackson network. Let $\mathsf{E}N$ be the total expected number of jobs in the open Jackson network in equilibrium. If $\mathsf{E}N$ is an integer, then the throughput of the equivalent CONWIP system (i.e., the closed network) with a job population of $\mathsf{E}N$ has a larger (respectively a smaller) throughput if the service rates at all stations are increasing and concave (respectively increasing convex).*

If $\mathsf{E}N$ is not an integer, the above theorem still holds with some minor modification: the job population in the equivalent CONWIP system is the smallest integer no less than $\mathsf{E}N$ (respectively the largest integer no greater than $\mathsf{E}N$) when the service rates are increasing concave (respectively increasing convex).

Note that when a service station has any given number of identical servers, the service rate is increasing concave. When a service station has an infinite identical servers, the service rate is linear. In most applications, the service rates are increasing concave. Hence, the Law of CONWIP Efficiency is a relative safe law to use, when the network is a Jackson network.

7 Conclusions

In this paper, we have examined some factory physics laws proposed in Hopp and Spearman (2011). These laws are simple and intuitive. Most of these laws hold when the system is described by an M/M/1 queue, or a tandem network of M/M/1 queues, or sometimes a GI/G/1 queue. It is probably through studying these simple systems that these laws have been developed. Modern manufacturing systems can be more complex. This motivates us to ask the questions whether and when these laws hold for more complex systems.

Our analysis suggests that we should be very cautious about the applicability of these laws, particularly when the systems are not Markovian or not in tandem. We hope our primitive analysis and counterexamples can stimulate more research interest in the discovery of more general factory physics laws.

Acknowledgements

This research was partially supported by an NSERC Grant (Canada), the National Natural Science Foundation of China (Grant No. 71271131), the Academic Research Fund and the E-Business Center of National University of Singapore, and the Central Research Grant (G-YN82) of Hong Kong Polytechnic University.

References

[1] Baskett, F., K.M. Chandy, R.R. Muntz, and R.G. Palacios. (1975). Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM*, **22**, 248-260.

[2] Bramson, M. (1999). A stable queueing network with unstable fluid model. *Annals of Applied Probability*, **9**, 818-853.

[3] Buzacott, J.A. and J.G. Shanthikumar. (1992). *Stochastic Models of Manufacturing Systems*, Prentice Hall.

[4] Chen, H. and D.D. Yao. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*, Springer-Verlag, New York.

[5] Chen, H. and H. Zhang. (2000). Stability of multiclass queueing networks under priority service disciplines. *Operations Research*, **48**, 26-37.

[6] Dai, J.G. (1996). A fluid-limit model criterion for instability of multiclass queueing networks. *Annals of Applied Probability*, **6**, 751-757.

[7] Einstein, A. (1951). “Forward” in the book, “Relativity - A Richer Truth”, P. Frank, Alden Press, Oxford, London.

[8] El-Taha, M. and S. Stidham. (1999). *Sample-Path Analysis of Queueing Systems*, Kluwer Academic Publishers.

[9] Hopp, W.J. and M.L. Spearman. (2011). *Factory Physics*, 3rd edition, Waveland Press, Long Grove, Illinois, USA.

[10] Jackson, J.R. (1957). Networks of waiting lines. *Operations Research*, **5**, No.4.

[11] Jackson, J.R. (1963). Jobshop-like queueing systems. *Management Science*, **10**, 518-521.

[12] Kelly, F.P. (1979). *Reversibility and Stochastic Networks*, Wiley, New York.

[13] Kumar, P.R. (1993). Re-entrant Lines. *Queueing Systems: Theory and Applications, Special Issue on Queueing Networks*, **13**, No.1-3, 87-110.

[14] Little, J.D.C. (1961). A proof of the queueing formula $L = \lambda W$. *Operations Research*, **9**, No.3, 383-387.

[15] Lu, S.H. and P.R. Kumar. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control*, **36**, 1406-1416.

[16] Meyn, S. (1995). Transience of multiclass queueing networks via fluid limit models. *Annals of Applied Probability*, **5**, 946-957.

[17] Ross, S. (1996). *Stochastic Processes*, Wiley.

[18] Shanthikumar, J.G. and D.D. Yao. (1986a). Preservation of likelihood ratio ordering under convolution. *Stochastic Processes and Their Applications*, **23**, 259-267.

[19] Shanthikumar, J.G. and D.D. Yao. (1986b). The effect of increasing service rates in closed queueing networks. *Journal of Applied Probability*, **23**, 474-483.

[20] Yao, D.D. (1994). *Stochastic Modeling and Analysis of Manufacturing Systems*, Springer-Verlag.

[21] Whitt, W. (1984). Minimizing Delays in the GI/G/1 Queue. *Operations Research*, **32**, No.1, 41-51.

[22] Whitt, W. (1991). A Review of $L = \lambda W$ and Extensions. *Queueing Systems, Theory and Applications*, **9**, No.3, 235-268. (Correction Note on $L = \lambda W$. *Queueing Systems,* **12**, No. 4, 1992, 431-432.)