Approximating the Performance of a Batch Service Queue Using the M/M^k/1 Model

Kan Wu, Leon McGinnis, and Bert Zwart

Georgia Institute and Technology, Atlanta, GA 30332 USA

Batching plays an important role in semiconductor fabs, and it can lead to inefficiency if not treated with care. The performance of parallel batch processes is often approximated by the G/G/1 based approximate models. By carefully examining the existing models, the dependence between queueing time and wait-to-batch time has been identified. A new improved model for parallel batch systems is proposed to exploit this dependence. The computation of the new model is still simple and fast, but it gives better approximation by reducing the systematic error in earlier models which ignored the dependence between queueing time and wait-to-batch time.

***Index Terms*—Parallel batch, Queueing model, Performance evaluation.**

Note to Practitioners:

Practical manufacturing systems are usually very complex. In order to analyze system behavior, assumptions are made to enable the analysis. A decomposition approach is commonly adopted based on an independence assumption. However, any two components in a complex manufacturing system are seldom completely independent. When there are real dependencies, we should take special care with analyses based on decomposition. This paper examines parallel batch processing and through detailed analysis, identifies a particular dependence and proposes an improved model that is validated through experimentation.

I. Introduction

A parallel batch is defined as processing a pre-determined group of jobs simultaneously without being interrupted by other product groups [1]. The job group has no specific composition and may consist of a single product or multiple products, as long as they use the same recipe. The number of jobs, i.e. batch size, is constrained by the process maximum batch size (if any).

Furnaces and ovens are typical examples of parallel batch machines. For example, the common physical capacity of a furnace in a 300mm semiconductor fab is four lots. However, a furnace parallel batch capacity is sometimes reduced to three lots due to process quality concerns.

Modeling parallel batch processing accurately is critical in understanding the performance of semiconductor fabs. Batch processing has captured researchers' attention for a long time and been rigorously studied. The first paper on this topic may be traced back to Bailey [2], who modeled a simple queueing process in which customers arrive at a single queue at random, and are served in a batch with a fixed maximum batch size.

Chaudhry and Templeton [3] summarized the state of the art up to that time in their book, *A First Course in Bulk Queues*. Two types of batch processing were addressed: bulk-arrival queues and bulk-service queues: bulk arrivals correspond to transfer batches and bulk services correspond to parallel process batches. Because they focused on queueing models which can be solved exactly, they did not address G/G^k/1 approximate models. Furthermore, models of serial process batches were not discussed.

In the late 1980s, the G/G/1 based approximate models for G/G^k/1 queues were proposed by Bitran and Tirupati [4] and Segal and Whitt [6]. Their approximations are based on decomposing the cycle time into three parts: wait-to-batch time, queueing time and service time, where the queueing time is obtained by G/G/1 approximations. This approximate model is generally applied to understand the behavior of parallel batches in practical manufacturing systems. In the 1990s, Hopp and Spearman [1] summarized previous work and introduced models for parallel process batches, serial process batches and transfer batches.

The approximate parallel batch model proposed by Bitran and Tirupati [4] offers us a flexible and powerful tool for describing the behavior of parallel batching machines. However, a potential issue of this approach is the assumption of independence between wait-to-batch time and queueing time, which is not satisfied in general. By carefully examining this issue, we develop an improved model to approximate the performance of parallel batching machines using the analytical solution from the M/M^k/1 model.

Our objective is to identify and correct errors in a commonly employed approximate cycle time formula for single server batch queues. Although we still adopt the decomposition approach, we make our decomposed model be the same as the M/M^k/1 model when the arrival process is Poisson and the service time is exponential. By assuming the variability term and the utilization term in Kingman's G/G/1 approximation [5] are approximately independent, the results are then further generalized to the G/G^k/1 queues. The new decomposition approximation yields the exact solution for the M/M^k/1 case and, based on the simulation results, gives smaller errors for the G/G^k/1 cases.

This paper is structured as follows: Section II provides the analysis of issues caused by the standard decomposition approach. In section III the new approximate model is

Manuscript received June 20, 2009 (date on which paper was submitted for review). Corresponding author: Kan Wu (e-mail: kanwu@gatech.edu).

Digital Object Identifier inserted by IEEE

proposed. Simulation results are given in section IV. Extensions to multiple server queues are given in section V and conclusion is given in VI.

II. The Analysis

The G/G/1 based approximate models for parallel process batches decompose the cycle time into three parts: wait-to-batch time (WTBT), queueing time (QT) and service time (ST). The structure of the model is illustrated in Fig. 1. The intention behind this decomposition is to approximate the duration of the three time segments. To guarantee the success of this decomposition, we need to make sure that each segment performs independently of the others.

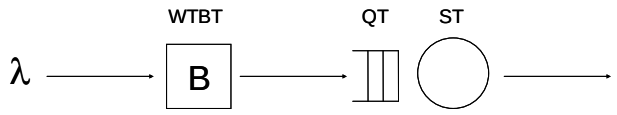


Fig. 1. The structure of parallel process batches.

Based on this scheme, if the job inter-arrival times are independent and identically distributed, Hopp and Spearman [1] propose the following model to approximate the average cycle time of parallel process batches,

$$CT \cong \frac{k-1}{2\lambda} + (\frac{c_a^2/k + c_b^2}{2})(\frac{\rho}{1-\rho})\frac{1}{\mu} + \frac{1}{\mu}, \quad (1)$$

where $\rho = \lambda / k\mu$, and k is the fixed parallel batch size, λ is the arrival rate of jobs (jobs/hour), μ is the service rate of the batching machine (batches/hour), c_a is the coefficient of variation (CV) of inter-arrival times, and c_b is the CV of batch service time.

The first term, (k-1)/(2λ), is the expected wait-to-batch time experienced by each single job. The second term is the expected queueing time from Kingman's G/G/1 approximation. Because each batch contains k jobs, and the job inter-arrival times are independent and identically distributed, the squared CV of batch inter-arrival times is k times smaller than the squared CV of job inter-arrival times. The third term is the expected batch service time.

To examine the effectiveness of this approximate model, we first compare the results of (1) with an M/M^k/1 queue, since an M/M^k/1 can be solved exactly [7]. In an M/M^k/1 queue, job arrivals occur as a Poisson process, and service times are exponentially distributed with a fixed parallel batch size k. The machine will process a batch if the batch size is exactly k. If the number of jobs in queue is less than k, they wait until k have accumulated. Batches are served on first-come-first-serve (FCFS) basis, and there is no limit on the queue length.

Under the assumption of an M/M^k/1 queue, (1) simplifies to

$$CT \cong \frac{k-1}{2\lambda} + (\frac{1/k+1}{2})BQT(M/M/1) + \frac{1}{\mu}, \quad (2)$$

where $BQT(M/M/1) = (\frac{\rho}{1-\rho})\frac{1}{\mu}$.

The queueing time in (2) is calculated based on Kingman's approximation, which is composed of a variability term, (1/k+1)/2, and an M/M/1 queueing time. We call the variability term a *G/G/1 transformer*, since it transforms an M/M/1 queueing time to a G/G/1 approximate queueing time. We call an M/M/1 queueing time a base queueing time (BQT). Thus, if the batch size is three, (2) can be illustrated by Fig. 2.

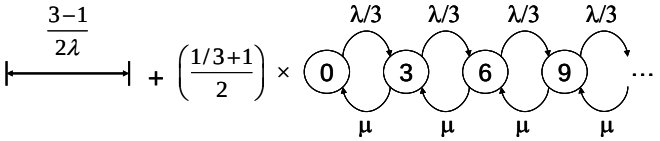


Fig. 2. Graphical illustration of Equation (2)

On the other hand, an M/M^k/1 queue is analyzed using a continuous time Markov chain model, which has a state transition rate diagram. When the batch size is three, the diagram is depicted in Fig. 3.

The state diagram of Fig. 3 is approximated by a flow equivalent birth and death process in Fig. 2. The reasons for the differences between Fig. 2 and Fig. 3 are apparent: one is only an approximation, but the other one is the exact analysis. Investigating their differences may bring us valuable insight for further improvement on the current approximate model.

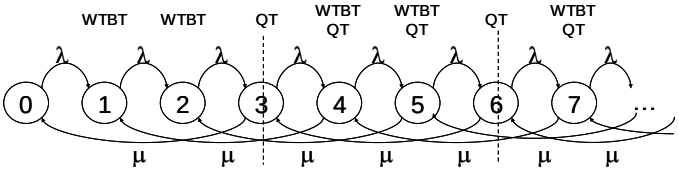


Fig. 3. The state transition rate diagram of an M/M^k/1 queue

In Fig. 3, since the batch size is three, the durations of states 1 and 2 are clearly the wait-to-batch times. The durations in state 3 and 6 are queueing times, since a complete batch is formed at those states which are the multiple of 3. Although the classification is clear at the above states, it is not so clear for the rest. For example, in states 4, 5 and 7, there are both wait-to-batch time and queueing time, since a complete batch is not formed yet, but there are some formed batches in queue.

Wait-to-batch times and queueing times are not independent, at least not in states 4, 5, and 7! This observation suggests rethinking the suitability of (1), which ignores the overlap between wait-to-batch time and queueing time.

Comparing Fig. 2 and Fig. 3, the transition from state 4 to state 1 in Fig. 3 has been ignored in Fig. 2 (similar situation for state 5 to state 2, etc.). The state changes can only occur at states 3, 6 and 9 in Fig. 2. This implies the queueing time in (1) has been *overestimated*, since it takes longer than it should to return to a lower state. Another source of errors comes from Kingman's approximation. As pointed by Shanthikumar and Buzacott [8], Kingman's approximation overestimates the true value when the service time coefficient of variation is smaller

than 1. However, this error becomes small when utilization is high. When service time variability is smaller than 1, both approximation errors overestimate the exact cycle time of the system. This tendency also can be observed in the simulation results of Fowler, et al. [9], where they studied the multiproduct G/G/c model with batch processing.

To gain better understanding of the structured errors caused by (1), we have compared the approximate results from (2) and the exact results from an $M/M^k/1$ queue by a numerical example. Based on Gross and Harris [7], the procedure to analyze an $M/M^k/1$ model is as follows:

(1) Solve for x in the characteristic equation (where $0 < x < 1$):

$$\mu x^{k+1} - (\lambda + \mu)x + \lambda = 0. \tag{3}$$

(2) Calculate the limiting probabilities p_0 and p_n,

$$p_0 = \frac{(1-x)}{k},$$
$$p_n = \begin{cases} \dfrac{p_0(1-x^{n+1})}{1-x} & (1 \le n < k), \\ \dfrac{p_0 \lambda x^{n-k}}{\mu} & (n \ge k). \end{cases}$$

(3) Calculate WIP, cycle time (CT), wait to batch time (WTBT), and queueing time (QT),

$$\begin{aligned} WIP &= \sum_{n=1}^{\infty} n p_n, \\ CT &= WIP / \lambda, \\ WTBT &= \frac{k-1}{2\lambda}, \\ QT &= CT - WTBT - ST. \end{aligned} \tag{4}$$

The definitions of parameters are the same as the parameters in (1). Rather than calculate p_0, and p_n, we may also get cycle time directly as follows,

$$CT = \frac{1}{\lambda k}\left(\begin{aligned} &\frac{k(k-1)}{2} - \frac{x^2(1-x^{k-1})}{(1-x)^2} \\ &+ \frac{(k-1)x^{k+1}}{1-x} + \frac{\lambda k}{\mu} + \frac{\lambda x}{\mu(1-x)} \end{aligned} \right). \tag{5}$$

One disadvantage of the above procedure is that it can only be solved numerically instead of explicitly, since we need solve (3) first. An alternative is to approximate x by a two term Taylor series expansion as follows,

$$x \cong 1 - \frac{2}{k+1}(1-\rho) - \frac{4}{3}\frac{(k-1)}{(k+1)^2}(1-\rho)^2. \tag{6}$$

Eq. (6) can greatly reduce the calculation effort, and gives accurate results when utilization is high. In the cases which we have examined, the errors of average cycle time are less than 3% as long as utilization is higher than 30%. The derivation of (6) is given in the Appendix.

In the example, we assume the batch size (k) is 10, and μ is 300 min. Both service times and inter-arrival times are exponentially distributed. The results are shown in Table I.

TABLE I
COMPARISON BETWEEN TWO MODELS WHEN K = 10 (UNIT: MIN)

Utiliza-	Arrival	Hopp and Spearman			$M/M^k/1$				HCT
tion	Interval	WTBT	HQT	HCT	WTBT	QT	CT	ST	Error %
10%	300.0	1350.0	18.3	1668.3	1350.0	0.3	1650.3	300	1.09%
20%	150.0	675.0	41.3	1016.3	675.0	5.6	980.6	300	3.63%
30%	100.0	450.0	70.7	820.7	450.0	21.4	771.4	300	6.39%
40%	75.0	337.5	110.0	747.5	337.5	50.5	688.0	300	8.66%
50%	60.0	270.0	165.0	735.0	270.0	97.6	667.6	300	10.10%
60%	50.0	225.0	247.5	772.5	225.0	174.8	699.8	300	10.39%
70%	42.9	192.9	385.0	877.9	192.9	306.3	799.1	300	9.85%
80%	37.5	168.8	660.0	1128.8	168.8	577.0	1045.8	300	7.93%
90%	33.3	150.0	1485.0	1935.0	150.0	1418.5	1868.5	300	3.56%
95%	31.6	142.1	3135.0	3577.1	142.1	3049.3	3491.4	300	2.45%

In Table I, HQT and HCT are the queueing time and cycle time calculated based on (1). We first find that the error percentages of HCT, (HCT-CT)/CT, are all positive, which is consistent with our previous observation that (1) tends to overestimate the cycle times. Furthermore, the errors are smaller when the utilization becomes high or low. This means (1) may give us good approximations when the utilization is very high or very low. Understanding this regular pattern of errors will provide insight for a better approximate model.

The errors in (1) mainly come from two sources: Kingman's heavy traffic approximation; and the missing transitions in Fig. 3, which are represented as dashed lines in Fig. 4. The reader may refer to (2) and Fig. 2 for a better understanding of the two sources of errors.

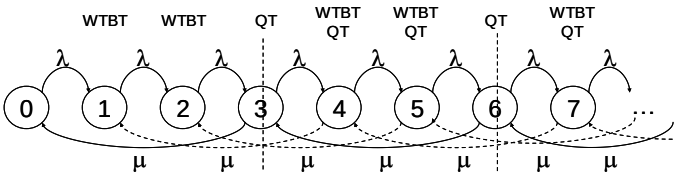


Fig. 4. The state transition rate diagram of an $M/M^k/1$ queue

If a job arrives when the batch processing machine is idle, it will cause no errors in (1), since the machine is in states 0, 1 or 2, (not in 4, 5 and 7, etc.). This implies lower utilization would lead to less error, since machine has longer idle time at lower utilization. On the other hand, Kingman's approximation exhibits larger errors in queueing time at lower utilization. However, the error percentages (i.e. queueing time errors / cycle time) from Kingman's approximation are relatively small compared with WTBT and service times, since queueing time itself is short. Therefore, these two sources of errors both tend to be small at low utilization.

If a job arrives when the machine is busy, the probability that it finds the machine in state 3t or 3t+1 is substantially greater than the probability it finds the machine in state 3t+2, where t is a natural number. If it arrives when the machine is

in state 3t+2, it causes no errors, since all the transitions in 3t+3 are considered in (2). However, if it arrives when the machine is in state 3t or 3t+1, the state will become 3t+1 or 3t+2, respectively, which may cause errors, since compared with Fig. 2, some of the transitions (presented by the dashed lines) are missing in those states. States 3t or 3t+1 are therefore called *incomplete states*.

Since the probability for a state to increase from t to t+1 is $\lambda/(\lambda+\mu)$ for t equal or greater than 3, the state will have higher probability to increase instead of decreasing when λ is larger. This transition will cause no errors, since it has been considered in Fig. 2 (by considering the effect of λ's in Fig. 4 into the $\lambda/3$ in Fig. 2). However, on the other hand, if a job arrives at an incomplete state and its state decreases (with probability $\mu/(\lambda+\mu)$), it will cause errors, since this transition is missing in Fig. 2. This observation suggests that higher utilization leads to smaller errors. At the same time, Kingman's approximation has smaller errors for higher utilizations. Therefore, the two sources of errors both tend to be small at high utilization.

From the above analysis, we know that there are two opposite forces which affect the error percentages. To get smaller errors, one force prefers higher utilization, and the other prefers lower utilization. This explains why the error percentages become smaller at both high and low utilizations in Table I.

When parallel batch size becomes larger, an incoming job may see an idle machine with higher probability, especially at lower utilization, which leads to smaller errors. Likewise, the errors from Kingman's approximation also become smaller due to the relatively longer WTBT. On the other hand, an incoming job may also drop into the incomplete states with higher probability, which leads to larger errors. However, when utilization is high, the incoming state tends to increase (towards right), which would cause no errors, even if the job arrives at an incomplete state. At the same time, Kingman's approximation also gives smaller errors at high utilization.

TABLE II
COMPARISON BETWEEN TWO MODELS WHEN K = 5 (UNIT: MIN)

Utiliza-	Arrival	Hopp and Spearman			M/M^k/1				HCT
tion	Interval	WTBT	HQT	HCT	WTBT	QT	CT	ST	Error %
10%	600.0	1200.0	20.0	1520.0	1200.0	1.3	1501.3	300	1.25%
20%	300.0	600.0	45.0	945.0	600.0	10.6	910.6	300	3.78%
30%	200.0	400.0	77.1	777.1	400.0	31.2	731.2	300	6.28%
40%	150.0	300.0	120.0	720.0	300.0	65.5	665.5	300	8.19%
50%	120.0	240.0	180.0	720.0	240.0	118.9	658.9	300	9.28%
60%	100.0	200.0	270.0	770.0	200.0	203.5	703.5	300	9.45%
70%	85.7	171.4	420.0	891.4	171.4	349.4	820.8	300	8.61%
80%	75.0	150.0	720.0	1170.0	150.0	645.5	1095.5	300	6.80%
90%	66.7	133.3	1620.0	2053.3	133.3	1543.5	1976.8	300	3.87%
95%	63.2	126.3	3420.0	3846.3	126.3	3332.9	3759.2	300	2.32%

Therefore, when batch size increases, the errors at the middle utilization range will increase due to the increase of the incomplete states. The impact on low and high utilization will not be so significant. This phenomenon can be seen in Table II, where the batch size is 5. The errors in the mid-range loading regime are smaller than the errors in Table I (reducing from 10.39% to 9.45% at 60% utilization) as expected.

III. THE NEW APPROXIMATE MODEL

The previous G/G/1 based approximate model is convenient. However, systematic errors caused by overlap between waiting time and batching time exist in the model even for the M/M^k/1 case, where exact solutions are available. A new approximate model is proposed based on this new piece of information.

When the arrival process is Poisson and the service time is exponential, instead of getting the base queueing times from an M/M/1 model, we get the base queueing times (*BQT*) from the M/M^k/1 model by using (2) and (4) as follows,

$$CT(M/M^k/1) = \frac{k-1}{2\lambda} + (\frac{1/k+1}{2})BQT + \frac{1}{\mu}. \quad (7)$$

In this equation, we treat BQT as the unknown variable. CT(M/M^k/1) can be obtained from (5), where x can be determined by solving either (3) or (6). Therefore,

$$BQT = \frac{2}{1/k+1}\left(CT(M/M^k/1) - \frac{1}{\mu} - \frac{k-1}{2\lambda}\right). \quad (8)$$

By assuming the variability term and the base queueing time in Kingman's equation are independent, queueing times (QT) and cycle times (CT) can be obtained as follows,

$$QT \cong (\frac{c_a^2/k + c_b^2}{2})BQT, \quad (9)$$

$$CT(G/G^k/1) \cong WTBT + QT + ST = \frac{k-1}{2\lambda} + QT + \frac{1}{\mu}. \quad (10)$$

The definitions of parameters are the same as in (1). In (9), queueing times are the product of the base queueing times and the G/G/1 transformer.

Although the new approximation does not completely avoid the dependence between wait-to-batch time and queueing time, the error (caused by the dependence) will be zero in the M/M^k/1 case. Thus, one can expect that the errors in the G/G^k/1 cases can be reduced relative to previous models. The performance of this approximation will be tested by simulations in the next section.

IV. SIMULATION EXPERIMENTS

The improvement of the new approximate models is demonstrated in three cases. In all three cases, the mean batch service times are 300 min, and the parallel batch size is 10. To demonstrate the true improvement from the new model itself, and avoid errors from other factors, the M/M^k/1 cycle times are calculated based on the results from the standard procedure (i.e. (3)) instead of the Taylor series expansion (i.e. (6)). However, when using Taylor series expansion, except for very low utilizations, only small errors are introduced.

The first case examined has Poisson arrivals and constant service times. The results are shown in Table III. As in Table I, HQT and HCT are the queueing times and cycle times

calculated based on (1). SWTBT, SQT and SCT are the wait-to-batch times, queueing times and cycle times from simulation.

TABLE III
SIMULATION RESULTS FOR POISSON ARRIVAL AND CONSTANT SERVICE TIMES (CASE 1)

Utiliza-	Arrival	Simulation					Hopp and Spearman			New Approximation		
tion	Interval	SWTBT	95% CI	SQT	95% CI	SCT	WTBT	HQT	HCT	WTBT	QT	CT
10%	300.0	1350.2	±0.05%	0.0	±198.4%	1650.2	1350.0	1.7	1651.7	1350.0	0.0	1650.0
20%	150.0	674.8	±0.06%	0.0	±27.26%	974.8	675.0	3.8	978.8	675.0	0.5	975.5
30%	100.0	450.1	±0.05%	0.0	±5.81%	750.1	450.0	6.4	756.4	450.0	1.9	751.9
40%	75.0	337.4	±0.06%	0.3	±2.17%	637.8	337.5	10.0	647.5	337.5	4.6	642.1
50%	60.0	270.0	±0.05%	1.4	±1.14%	571.5	270.0	15.0	585.0	270.0	8.9	578.9
60%	50.0	224.9	±0.05%	4.6	±0.68%	529.5	225.0	22.5	547.5	225.0	15.9	540.9
70%	42.9	192.9	±0.05%	12.5	±0.51%	505.4	192.9	35.0	527.9	192.9	27.8	520.7
80%	37.5	168.8	±0.05%	32.9	±0.59%	501.7	168.8	60.0	528.8	168.8	52.5	521.2
90%	33.3	150.1	±0.06%	103.3	±0.88%	553.3	150.0	135.0	585.0	150.0	129.0	579.0
95%	31.6	142.1	±0.05%	250.8	±1.53%	692.9	142.1	285.0	727.1	142.1	277.2	719.3

At each specific utilization level, each reported WTBT, SQT and SCT is the mean of 100 replications. In each replication, we collected data for 200,000 jobs after a 50 year warm-up period (thus, the number of jobs discarded in the warmup can be approximated as "50 years * 365 days * 1440 min / arrival-interval"). The above parameters are chosen to reduce the confidence intervals, but with a tolerable simulation run time. The half-width of 95% confidence intervals (CI) are listed right after the corresponding simulation values. Service times are omitted in the table, since they are all about 300 min, and the 95% CI are all smaller than 0.1%. The units of wait-to-batch times, queueing times, and cycle times are minutes. Since the service times are constant (c_b is 0), the queueing times tend to be small compared with the mean service times, even at high utilization.

TABLE IV
SIMULATION RESULTS FOR POISSON ARRIVAL AND ERLANG-2 SERVICE TIMES (CASE 2)

Utiliza-	Arrival	Simulation					Hopp and Spearman			New Approximation		
tion	Interval	SWTBT	95% CI	SQT	95% CI	SCT	WTBT	HQT	HCT	WTBT	QT	CT
10%	300.0	1350.4	±0.05%	0.0	±18.14%	1650.3	1350.0	10.0	1660.0	1350.0	0.2	1650.2
20%	150.0	675.1	±0.05%	1.1	±2.54%	976.0	675.0	22.5	997.5	675.0	3.1	978.1
30%	100.0	450.1	±0.05%	6.1	±1.19%	756.4	450.0	38.6	788.6	450.0	11.7	761.7
40%	75.0	337.4	±0.05%	17.9	±0.87%	655.6	337.5	60.0	697.5	337.5	27.5	665.0
50%	60.0	269.9	±0.05%	39.7	±0.87%	609.5	270.0	90.0	660.0	270.0	53.2	623.2
60%	50.0	225.1	±0.05%	77.8	±0.67%	602.9	225.0	135.0	660.0	225.0	95.3	620.3
70%	42.9	192.9	±0.05%	147.0	±0.82%	639.8	192.9	210.0	702.9	192.9	167.1	659.9
80%	37.5	168.7	±0.05%	290.6	±0.86%	759.2	168.8	360.0	828.8	168.8	314.7	783.5
90%	33.3	150.0	±0.04%	733.5	±1.48%	1183.5	150.0	810.0	1260.0	150.0	773.7	1223.7
95%	31.6	142.1	±0.04%	1626.6	±2.42%	2068.6	142.1	1710.0	2152.1	142.1	1663.3	2105.4

In the second case, the arrival process is still Poisson, but service times follow an Erlang-2 distribution. The squared coefficient of variation (SCV) of service times is 0.5. Comparing Case 2 with Case 1, since the service times are changed from constant to Erlang-2 distribution, the queueing times become considerably longer compared with the mean service times.

In the third case, the service times still follow an Erlang-2 distribution, but the arrival intervals follow an Erlang-10 distribution. The SCV of arrival intervals is 0.1.

TABLE V
SIMULATION RESULTS FOR ERLANG-10 ARRIVALS AND ERLANG-2 SERVICE TIMES (CASE 3)

Utiliza-	Arrival	Simulation					Hopp and Spearman			New Approximation		
tion	Interval	SWTBT	95% CI	SQT	95% CI	SCT	WTBT	HQT	HCT	WTBT	QT	CT
10%	300.0	1343.3	±0.99%	0.0	±198.4%	1643.5	1350.0	8.5	1658.5	1350.0	0.1	1650.1
20%	150.0	675.1	±0.02%	0.1	±7.81%	975.3	675.0	19.1	994.1	675.0	2.6	977.6
30%	100.0	448.9	±0.49%	2.2	±7.53%	751.2	450.0	32.8	782.8	450.0	9.9	759.9
40%	75.0	337.5	±0.02%	9.2	±1.18%	646.7	337.5	51.0	688.5	337.5	23.4	660.9
50%	60.0	270.0	±0.02%	25.3	±0.84%	595.5	270.0	76.5	646.5	270.0	45.2	615.2
60%	50.0	225.0	±0.01%	55.3	±0.84%	580.3	225.0	114.8	639.8	225.0	81.0	606.0
70%	42.9	192.8	±0.02%	111.9	±0.84%	604.6	192.9	178.5	671.4	192.9	142.0	634.9
80%	37.5	168.6	±0.22%	237.8	±4.04%	706.3	168.8	306.0	774.8	168.8	267.5	736.3
90%	33.3	150.0	±0.02%	598.8	±1.69%	1048.5	150.0	688.5	1138.5	150.0	657.7	1107.7
95%	31.6	142.1	±0.02%	1370.6	±3.00%	1812.7	142.1	1453.5	1895.6	142.1	1413.8	1855.9

The errors of estimated queueing times from the old and new models are shown in Table VI. HQT error is "HQT/SQT - 1", and QT error is "QT/SQT - 1". Improvement is "HQT error/QT error - 1", which gives the improvement of the new model. In all three cases, both old and new models give large errors (HQT error and QT error) at low utilization and relatively small errors at high utilization, since Kingman's formula is a heavy traffic approximation.

TABLE VI
ERRORS OF THE THREE MODELS BY USING (3)

		Case 1: $c_a^2 = 1$, $c_b^2 = 0$			Case 2: $c_a^2 = 1$, $c_b^2 = 0.5$			Case 3: $c_a^2 = 0.1$, $c_b^2 = 0.5$		
Utiliza- tion	Arrival Interval	HQT Error	QT Error	Improve- ment	HQT Error	QT Error	Improve- ment	HQT Error	QT Error	Improve- ment
10%	300.0	8124942.8%	127787.1%	98.4%	49125.9%	674.8%	98.6%	400452.9%	6204.7%	98.5%
20%	150.0	344446.7%	46975.6%	86.4%	2035.4%	191.8%	90.6%	15569.3%	2040.9%	86.9%
30%	100.0	16519.6%	4928.6%	70.2%	534.2%	91.9%	82.8%	1413.6%	358.0%	74.7%
40%	75.0	3050.1%	1344.8%	55.9%	234.4%	53.4%	77.2%	454.8%	154.5%	66.0%
50%	60.0	961.5%	527.7%	45.1%	126.6%	34.0%	73.1%	202.4%	78.8%	61.1%
60%	50.0	388.0%	244.7%	36.9%	73.5%	22.5%	69.3%	107.7%	46.7%	56.6%
70%	42.9	179.8%	122.6%	31.8%	42.8%	13.6%	68.2%	59.5%	26.9%	54.8%
80%	37.5	82.2%	59.3%	27.9%	23.9%	8.3%	65.2%	28.7%	12.5%	56.4%
90%	33.3	30.7%	24.8%	19.1%	10.4%	5.5%	47.4%	15.0%	9.8%	34.4%
95%	31.6	13.7%	10.5%	22.8%	5.1%	2.3%	56.0%	6.0%	3.1%	47.9%

The improvement percentage of the new model decreases with increasing utilization. In Case 1, the improvement from the new models decreases from 98% (at 10% utilization) to 23% (at 95% utilization). Furthermore, among these three cases, the improvement increases when either c_a or c_b is close to one. This observation is consistent with our assumptions, since we know our approximate model yields exact solutions when the service time is exponential and the arrival process is Poisson. On the other hand, this new approximation may not perform very well when c_a and c_b are much larger than one.

However, in practical manufacturing systems, in order to maintain competitiveness, service time SCV is desired to be small. Therefore, c_b^2 is chosen to be 0 in Case 1 and 0.5 in Case 2 and 3. Among the three cases, due to the Palm-Khintchine theorem, Case 2 may be representative of the situations where the machine is fed by multiple upstream workstations, and each workstation is composed of multiple machines. In this situation, the arrival process can be close to a Poisson process [10]. If the machine is only fed by one or two upstream machines, the c_a^2 can be small. Case 3 may be representative in this situation. As we have seen in Table VI, in both cases, the original errors can be around 10% and the improvement can be around 50% at high utilization.

If we use (6), the Taylor series expansion approximation, the errors will be larger at low utilization, but almost the same at high utilization (see Table VI). In the examined cases, because the value of x is overestimated at low utilization by

the Taylor series expansion, the estimated queueing time indeed becomes negative when utilization is less than 20%. However, it causes no significant impact to the overall system, since the true queueing time in this situation is less than two minutes compared with the WTBT 1350 minutes.

TABLE VII
ERRORS OF THE THREE MODELS BY THE TAYLOR SERIES EXPANSION

		Case 1: $c_a^2 = 1$, $c_b^2 = 0$			Case 2: $c_a^2 = 1$, $c_b^2 = 0.5$			Case 3: $c_a^2 = 0.1$, $c_b^2 = 0.5$		
Utilization	Arrival Interval	HQT Error	QT Error	Improvement	HQT Error	QT Error	Improvement	HQT Error	QT Error	Improvement
10%	300.0	8124942.8%	--	--	49125.9%	--	--	400452.9%	--	--
20%	150.0	344446.7%	--	--	2035.4%	--	--	15569.3%	--	--
30%	100.0	16519.6%	4707.9%	71.5%	534.2%	83.5%	84.4%	1413.6%	337.9%	76.1%
40%	75.0	3050.1%	1704.4%	44.1%	234.4%	91.5%	60.9%	454.8%	217.8%	52.1%
50%	60.0	961.5%	642.2%	33.2%	126.6%	58.5%	53.8%	202.4%	111.5%	44.9%
60%	50.0	388.0%	278.9%	28.1%	73.5%	34.7%	52.8%	107.7%	61.2%	43.1%
70%	42.9	179.8%	134.1%	25.4%	42.8%	19.5%	54.5%	59.5%	33.4%	43.8%
80%	37.5	82.2%	62.5%	24.0%	23.9%	10.5%	56.1%	28.7%	14.7%	48.6%
90%	33.3	30.7%	23.6%	23.1%	10.4%	4.4%	57.5%	15.0%	8.7%	41.6%
95%	31.6	13.7%	10.6%	22.7%	5.1%	2.3%	55.8%	6.0%	3.2%	47.7%

The robustness of the model was tested for service times and inter-arrival times that do not belong to the family of gamma distributions. The model still gives good approximations when the service time is lognormally distributed with c_b^2 = 0.1 and the inter-arrival time is triangular distributed with lower limit a, mode b and upper limit c as shown in Table VIII. The c_a^2 is kept at 0.5 at all utilizations.

TABLE VIII
SIMULATION RESULTS WHEN SERVICE TIME IS LOGNORMAL DISTRIBUTED WITH C_B^2 = 0.5 AND INTER-ARRIVAL TIME IS TRINGULAR DISTRIBUTED WITH C_A^2 = 0.1 (CASE 4)

Tringular (a, b, c)			Utilization	Arrival Interval	Simulation					HQT Error	QT Error	Improvement
a	b	c			SWTBT	95% CI	SQT	95% CI	SCT			
67.62	300.0	532.4	10%	300.0	1350.0	±0.02%	0.0	±23.16%	1649.9	34624.8%	446.6%	98.7%
33.81	150.0	266.2	20%	150.0	674.9	±0.01%	0.9	±5.14%	975.4	2128.7%	204.5%	90.4%
22.54	100.0	177.5	30%	100.0	449.9	±0.01%	4.7	±2.11%	754.4	590.7%	109.0%	81.6%
16.91	75.0	133.1	40%	75.0	337.5	±0.02%	14.1	±1.27%	651.7	260.7%	65.4%	74.9%
13.52	60.0	106.5	50%	60.0	270.0	±0.02%	32.0	±1.25%	602.1	138.8%	41.2%	70.3%
11.27	50.0	88.73	60%	50.0	225.0	±0.02%	62.7	±1.03%	587.9	83.1%	29.3%	64.7%
9.66	42.9	76.05	70%	42.9	192.9	±0.02%	120.7	±1.09%	613.5	47.9%	17.6%	63.2%
8.453	37.5	66.55	80%	37.5	168.7	±0.02%	242.4	±1.33%	711.0	26.3%	10.4%	60.4%
7.513	33.3	59.15	90%	33.3	150.0	±0.02%	613.7	±1.9%	1063.5	12.2%	7.2%	41.2%
7.118	31.6	56.04	95%	31.6	142.1	±0.02%	1362.7	±3.25%	1804.8	6.7%	3.7%	43.8%

Comparing Case 3 with Case 4, since the mean and SCV of the service times and inter-arrival times are the same, the approximate queueing times are the same (not shown in Table VIII), and the simulated queueing times in Case 3 and 4 are close to each other as expected. Therefore, the HQT errors, QT errors and the improvement are similar.

V. EXTENSION TO MULTIPLE SERVER QUEUES

In this section, we extend the previous model to a queueing system having c homogeneous servers, each with maximum parallel batch size k. When one of the servers is available, then: (a) If the number in queue > k, then the first k customers enter service immediately. (b) If 0 < number in queue <= k, then all the waiting customers are taken for service. (c) If number in queue = 0, service stops until a new customer arrives, after which service resumes immediately.

Based on Chaudhry and Templeton [3], the procedure to analyze the above $M/M^k/c$ model is as follows:

(1) Solve for x in the characteristic equation (where $0 < x < 1$):

$$\lambda x^{k+1} - (\lambda + c\mu)x^k + c\mu = 0. \tag{11}$$

(2) Calculate the limiting probabilities,

$$p_{0,0} = \left(\frac{(\lambda/\mu)^c}{(1-1/x)c!} + \sum_{r=0}^{c-1} \frac{(\lambda/\mu)^r}{r!} \right)^{-1},$$

$$p_{n,c} = x^{-n} (\lambda/\mu)^c \frac{p_{0,0}}{c!}, \qquad m \geq 0$$

$$p_{0,l} = p_{0,0} \frac{(\lambda/\mu)^l}{l!}, \qquad 0 < l < c.$$

where n is the number of customers in the queue (i.e. waiting customers) and l is the number of busy servers. n can be greater than zero only when all c servers are busy.

(3) Calculate WIP, cycle time (CT) and queueing time (QT),

$$\begin{aligned} WIP &= \sum_{n=1}^{\infty} (n+c) p_{n,c} + \sum_{l=1}^{c} l p_{0,l}, \\ CT(M/M^k/c) &= WIP/\lambda, \\ WTBT &= \frac{k-1}{2\lambda}, \\ QT &= CT - WTBT - ST. \end{aligned} \tag{12}$$

Using (12) to replace the $CT(M/M^k/1)$ in (7) and (8). We can then obtain $CT(G/G^k/c)$ by (10).

The above approach is developed based on the observation from Sakasegawa [11], who finds the variability term and utilization term in Kingman's equation behave approximately independently in a G/G/c queue.

VI. CONCLUSION

Through a detailed examination of the earlier G/G/1 based approximate models for parallel batch process, the information lost during the decomposition process has been identified. By partially recovering the lost information, a new approximate model has been proposed. The new model shows notable improvement over the previous approaches.

Decomposition is a powerful and convenient technique, especially when we want to analyze large complex systems in a practical environment. However, in complex manufacturing systems, the independence assumption inherent in decomposition is often not satisfied. As a result, the approximation errors from decomposition can be significant if the potential for information loss is not recognized and dealt with, as illustrated in this paper.

Although we have gained notable improvement by the new approach, the new model only considers the case of parallel batches. The approximate model considering a workstation with the combinations of parallel process batches, serial process batches and transfer batches is left for future research.

APPENDIX: DERIVATION OF (6)

From (3),

$$\mu x^{k+1} - (\lambda+\mu)x + \lambda = 0, \;\Rightarrow\; \mu(x - x^{k+1}) = \lambda(1-x),$$

$$\Rightarrow\; \mu x \frac{(1-x^k)}{(1-x)} = \lambda, \;\Rightarrow\; \frac{1}{k}\frac{x(1-x^k)}{(1-x)} = \frac{\lambda}{\mu k} = \rho,$$

$$\Rightarrow\; \frac{1}{k}\sum_{n=1}^{k} x^n = \rho.$$

Let $f(x) = \frac{1}{k}\sum_{n=1}^{k} x^n$ and $g(\cdot) = f^{-1}(\cdot)$.

Therefore, $x = g(\rho)$ and $f(x) = \rho$.

By using Taylor Series Expansion,

$$x = g(1) + (\rho-1)g'(1) + \frac{1}{2}(\rho-1)^2 g''(1) + O((1-\rho)^3), \quad (A.1)$$

where

$$\because\; f(g(x)) = x,\; f(g(1)) = 1 \text{ and } f(1) = 1, \;\Rightarrow g(1) = 1,$$

$$\because\; f'(g(x))g'(x) = 1, \;\therefore g'(1) = \frac{1}{f'(1)}$$

$$\because\; f'(x) = \frac{1}{k}\sum_{n=1}^{k} n x^n,$$

$$\therefore\; f'(1) = \frac{1}{k}\sum_{n=1}^{k} n = \frac{1}{k}\frac{k(k+1)}{2} = \frac{k+1}{2}, \Rightarrow g'(1) = \frac{1}{k+1}.$$

$$\because\; f''(x) = \frac{1}{k}\sum_{n=1}^{k} n(n-1)x^n,$$

$$\therefore\; f'(1) = \frac{1}{k}\sum_{n=1}^{k} n(n-1) = \frac{1}{k}\frac{(k-1)k(k+1)}{3} = \frac{(k-1)(k+1)}{3},$$

$$\because\; f''(g(x))(g'(x))^2 +' f(g(x))g''(x) = 0,$$

$$\therefore\; f''(g(1))(g'(1))^2 + f'(g(1))g''(1) = 0,$$

$$\Rightarrow g''(1) = -\frac{f''(1)(g'(1))^2}{f'(1)} = -\frac{\frac{1}{3}(k-1)(k+1)(\frac{2}{k+1})^2}{(k+1)/2}$$

$$= -\frac{8}{3}\frac{(k-1)}{(k+1)^2}.$$

Therefore, (A.1) becomes

$$x = 1 - \frac{2}{k+1}(1-\rho) - \frac{4}{3}\frac{(k-1)}{(k+1)^2}(1-\rho)^2 + O((1-\rho)^3),$$

$$\cong 1 - \frac{2}{k+1}(1-\rho) - \frac{4}{3}\frac{(k-1)}{(k+1)^2}(1-\rho)^2.$$

ACKNOWLEDGMENT

The authors would like to acknowledge the Keck Foundation and Gwaltney Chair for Manufacturing Systems for their financial support to conduct the research.

REFERENCES

[1] W. J. Hopp and M. L. Spearman, *Factory Physics*. Chicago, IL: IRWIN, 1996.

[2] N. T. Bailey, "On Queueing Processes with Bulk Service", *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 16, no. 1, pp. 80–87, 1954.

[3] M. L. Chaudhry and J. G. C. Templeton, *A First Course in Bulk Queues*, New York: Wiley, 1983.

[4] G. R. Bitran and D. Tirupati, "Approximations for Product Departures from a Single-Server Station with Batch Processing in Multi-Product Queues", *Management Science*, vol. 35, no. 7, pp. 851–878, 1989.

[5] Kingman, J. F. C., "Some Inequalities for the Queue GI/G/1", *Biometrika*, vol. 49, pp. 315–324, 1962.

[6] M. Segal and W. Whitt, "A Queueing Network Analyzer for Manufacturing", *Teletraffic Science for New Cost-effective Systems*, pp. 1146–1152, 1989.

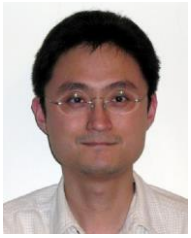
[7] D. Gross and C. M. Harris, *Queueing Theory*, New York: Wiley, 1998.

[8] J. G. Shanthikumar and J. A. Buzacott. "On the Approximations to the Single Server Queue", *International Journal of Production Research*, vol. 18, no. 6, pp. 761–773. 1980.

[9] J. W. Fowler, N. Phojanamongkolkij, J. K. Cochran and D. C. Montgomery, "Optimal Batching in a Wafer Fabrication Facility Using a Multiproduct G/G/c Model with Batch Processing", *International Journal of Production Research*, vol. 40, no. 2, pp. 275–292, 2002.

[10] Inoue, T., Ishii, Y., Igarashi, K., Muneta, T., and Imaoka, K., "Study of cycle time caused by lot arrival distribution in a semiconductor manufacturing line", *IEEE International Symposium on ISSM 2005*, pp. 115–118, 2005.

[11] H. Sakasegawa. "An approximation Formula $L_q = \alpha\rho^{\beta}/(1-\rho)$", *Annual of the Institute for Statistical Mathematics*, vol. 29, pp. 67–75, 1977.



Kan Wu received the B.S. degree in nuclear engineering from National Tsinghua University, Hsinchu, Taiwan, the M.S. degree in Industrial Engineering and Operations Research, the M.E. degree in nuclear engineering from the University of California, Berkeley, and the Ph.D degree in Industrial and Systems Engineering from Georgia Institute of Technology, Atlanta.

He has been a senior engineer with Tefen, Ltd., and Taiwan Semiconductor Manufacturing Company, and an IE manager at Inotera Memories Inc. Currently, he is the CTO and funding team member of Sensor Analytics, Inc. His current research interests focus on quantifying and optimizing the performance of manufacturing systems.

Leon McGinnis is the Gwaltney Professor of Manufacturing Systems at Georgia Tech. Professor McGinnis is internationally known for his leadership in the material handling research community and his research in the area of discrete event logistics systems. He has received several awards for his innovative research, including the David F. Baker Award from IIE, the Reed-Apple Award from the Material Handling Education Foundation, and the Material Handling Innovation Pioneer award from Material Handling Management Magazine. He is author or editor of seven books and more than 110 technical publications. At Georgia Tech, Professor McGinnis has held leadership positions in a number of industry-focused centers and programs, including the Material Handling Research Center, the Computer Integrated Manufacturing Systems Program, the Manufacturing Research Center, and the newly-formed Product/Systems Lifecycle Management Center. His current research explores the application of PLM technologies to the design and management of highly capitalized factories.

Bert Zwart is senior researcher at the Center for Mathematics and Computer Science in Amsterdam and is a Professor at VU University Amsterdam, while being on leave from his tenure at Georgia Tech. Bert holds an M.A. in Econometrics from VU University, Amsterdam, and a Ph.D in Applied Mathematics from Eindhoven University of Technology. He serves on several conference program committees in applied probability and performance analysis, and serves on the editorial boards of six international journals. His research is concerned with modeling, analysis and simulation of stochastic systems arising in actuarial and financial mathematics, computer and communication systems, manufacturing systems, and customer contact centers. His research is partly supported by NSF grants 0727400 and 0805979, an IBM faculty award, and a VIDI grant from NWO.