Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

MEAN CYCLE TIME APPROXIMATIONS FOR G/G/M QUEUEING NETWORKS USING DECOMPOSITION WITHOUT AGGREGATION WITH APPLICATION TO FAB DATASETS

Jinho Shin
James R. Morrison

Department of Industrial & Systems Engineering
Korea Advanced Institute of Science and Technology
291, Daehak-ro, Yuseong-gu
Daejeon, KS015, REPUBLIC of SOUTH KOREA

Dean Grosbard
Adar Kalir

Fab/Sort Manufacturing Division
Intel Corporation
2 HaZoran St.
Qiriat-Gat 82109, ISRAEL

ABSTRACT

The modern semiconductor fabricator needs both accurate and fast cycle time (CT) forecasts. Due to complexity of development and computational intractability, simulation may be supplemented by queueing network methods. In this paper, we develop extensions to approximation methods for queueing networks that are suited for fab modeling using decomposition without aggregation. We conduct simulation experiments based on a semiconductor industry-inspired dataset. For sensitivity analysis, we mainly focus on the interarrival distribution, service time distributions, and bottleneck toolset loading. The results show that the approximations predict the total CT fairly well in various cases.

1 INTRODUCTION

Queueing networks can serve as models for semiconductor wafer fabricator (fabs). However, exact analytic expressions for key performance measures such as the mean CT rarely exist. One alternative to exact analysis is to use approximation methods. For networks of G/G/m queues, Whitt's QNA (1983) used an approach termed decomposition with aggregation (DWA) to determine the variability parameters for the internal flows. It is most applicable for systems with probabilistic routing. Reiman (1990) extended the approach to include customer priorities. Approximations for fabs using these ideas were developed in Connors et al. (1996). Unfortunately, fabs feature mostly deterministic routing and, as Bitran and Tirupati (1989) explain, the splitting operator in DWA is a cause of errors in the approximations. Decomposition without aggregation (DWOA) is considered more appropriate for networks based on fabs; see, Kim (2005). Since fabs often feature both deterministic routing and probabilistic routing (due to metrology for example), Grosbard et al. (2013) extended DWOA to accommodate both deterministic and probabilistic routing. For further information, Whitt (1983) provides details on queueing network systems and Kim (2005) reviews the existing DWOA methods.

In this paper, we modify and extend the DWOA method suggested in Grosbard et al. (2013) to improve the mean CT approximations. To test the quality of the approximations, we conduct numerous simulation studies on two models inspired by actual fab settings. Sensitivity analysis on bottleneck toolset loading, service time distribution and interarrival time distribution are considered.

The paper is organized as follows. In Section 2, we provide the modified traffic variability equations and mean total CT approximations for a network of G/G/m queues with both deterministic and probabilistic routing. In Section 3, we provide a description of the datasets and simulation setup. The

overall performance of the proposed approach is reviewed in Section 4. The results of sensitivity studies are reviewed in Section 5. Concluding remarks are provided in Section 6.

2 MODEL FORMULATION

In this section, we define our basic notation and describe the calculation of traffic rates, traffic variability, and CT approximations. In this section, we focus on the mathematical steps leading to our modified DWOA method. Due to space limitations, illustrations which may be of help are omitted. We refer the interested reader to Whitt (1983) for intuitive and helpful figures.

2.1 Basic Notation

2.1.1 Notation for parameter sets

$T = \{1, \cdots, W\}$	:	Set of G/G/m queues (toolsets) in the network
$O = \{1, \cdots, \nu\}$	:	Set of operations in the network
$P = \{\nu + 1, \cdots, \nu + \theta\}$	:	Set of PM type operations in the network
$T_{k,O} \epsilon O$	:	Set of operations served at queue k
$T_{k,PM} \epsilon P$	:	Set of PM type operations conducted at queue k
T_k	:	$T_{k,O} \cup T_{k,PM}$

In this paper, we consider the toolsets in semiconductor manufacturing facility as queues in G/G/m network. Each operation represents a buffer at a queue in which customers (lots) await service. Numerous operations are allowed at each queue. Operations are for the products. Preventive maintenance (PM) type operations are used to model tool failures. Customers may arrive to any operation (except the PM type operations). PM's are modeled as high priority customers that may arrive to any PM type operation. Customer are low priority customers. Multiple PM type operations may be present at each queue to model PMs with different character (e.g., monthly, quarterly and annual PMs).

After service at operation i, a lot next proceeds to operation j with probability q_{ij}. Operations i and j may be processing, metrology or rework operations. With probability $1\text{-}q_{i1}\text{-}q_{i2}\text{-}\ldots\text{-}q_{iv}$ a lot finishing operation i is complete and exits the network.

2.1.2 Notation for queueing network variables

Throughout, we use CV to denote coefficient of variation.

λ_i^{EX}	:	Exogenous mean arrival rate of (product) customers to operation i
λ^{EX}	:	Column vector of exogenous mean arrival rates of (product) customers to each operation
$\lambda_{i,PM}$	:	Exogenous mean arrival rate of type i PM customers
λ_{PM}	:	Column vector of mean arrival rates for the PM customers
$C_{a_i^{EX}}$	:	CV of the interarrival times of exogenous arrivals to operation i
C_{a_i}	:	CV of the interarrival times for all arrivals (both exogenous and endogenous) to operation i
C_{a,PM_i}	:	CV of the interarrival times for PM type i customers
$q_{i,j}$	:	Probability that a customer departing from operation i is routed next to operation j
Q	:	The routing matrix of $q_{i,j}$ values
λ_i	:	Mean total arrival rate of customers to operation i (both exogenous and endogenous)
λ	:	Column vector of mean total arrival rates of customers to each operation (both exogenous and endogenous)

S_i : Mean service time of operation i (which is served at queue k with $i \epsilon T_{k,O}$)
C_{S_i} : CV of the service time for operation i
D_i : Mean downtime of PM type i customers
C_{D_i} : CV of type i PM downtime
m_k : Number of servers dedicated to serving customers at queue k
σ(i) : Queue at which operation i performed
Γ_k : Mean total arrival rate of all customers to queue k

We consider a network of G/G/m queues. Each queue k consists of m_k dedicated servers. The servers are not prone to failure – we model the downtime as service for high priority customers (PMs). Each queue caters to distinct operations. Customers may arrive to an operation from outside the network (an exogenous arrival process) with arrival rate λ_i^{EX}. The interarrival times are IID random variables for each process. Customers departing operation i are routed to operation j with probability $q_{i,j}$. If a customer is not routed to some operation, it departs the network. The service duration for a customer in operation i is a random variable with mean value S_i. All interarrival times and service time are independent of each other. We treat the servers as non-idling. When a server completes the current service on a customer, it immediately finds another customer to serve. Customers are selected from those waiting for service at operations catered to by queue k in a FIFO manner within each priority class. High priority customers are given non-preemptive priority over the product customers.

Note that because we model the PMs as high priority customers, they do not behave exactly as would a PM in the real world. For example, in reality an incoming PM event is associated with a specific tool in the set of servers. However, in the G/G/m model, high priority customers can be served by any available tool serving the queue.

To calculate the mean total CT of customer in the network, we define the variables for expected number of visits to each operation.

$n_{i,j}$: Expected number of visits to operation j by a customer from the exogenous arrival process to operation i
N_i : Column vector of $n_{i,j}$ values, j = 1, …, ν, for customers from the exogenous arrival process to operation i

2.2 Traffic Rates

Given exogenous mean arrival rate of customers to specific operation, we can derive the mean total arrival rate of customer by solving the following set of linear equations.

$$\forall i \in O, \ \lambda_i = \lambda_i^{EX} + \sum_{j \in O} q_{j,i} \lambda_j \quad (1)$$

As we mentioned, PM type customers only visit designated PM type operation once, and then leave the network. For that, equation (1) is simplified to $\forall i \in P, \ \lambda_i = \lambda_{i,PM}$.

Equation (1) in matrix notation:

$$\lambda = (I - Q^T)^{-1} \lambda^{EX} \quad (2)$$

Each row of Q sums to a value in [0,1], with sum less than 1 indicating a positive probability for the customers to depart the network. This allows both deterministic and probabilistic paths. Since the routing matrix has no sense of history, if one wishes to model an "operation" that should be revisited (say 3 times), a separate operation for each such visit (thus 3 operations) with the same service statistics should

be created. As each operation is unique and only visited once by customers, routing matrix Q can be organized as block diagonal structure. For example, in figure 1, there are totally 70 operations and 30 PM type operations in Q. Operation 1 and 40 have exogenous customers arrivals, and PM type operation 71 through 100 have their own PM type customers arrivals. This means that the customers arriving to operation 1 only visit operation 1 through 39 in its manufacturing route. Similarly, the customers arriving to operation 40 go around operation 40 through 70. As the PM type operations have no interaction with other operations, their $q_{j,i}$ values are all equal to 0.

$$Q = \begin{bmatrix} q_{1,1} & \cdots & q_{1,39} & & & & & & \\ \vdots & \ddots & \vdots & & 0 & & & 0 & \\ q_{39,1} & \cdots & q_{39,39} & & & & & & \\ & & & q_{40,40} & \cdots & q_{40,70} & & & \\ & 0 & & \vdots & \ddots & \vdots & & 0 & \\ & & & q_{70,40} & \cdots & q_{70,70} & & & \\ & & & & & & 0 & \cdots & 0 \\ & 0 & & & 0 & & \vdots & \ddots & \vdots \\ & & & & & & 0 & \cdots & 0 \end{bmatrix}$$

Figure 1: Block diagonal matrix Q.

If every customers can eventually leave, the network is called open and $(I-Q^T)^{-1}$ exists. This structure can readily be used to model separate customers classes each with its own dedicated operations (and as mentioned "operations" visited multiple times in a reentrant process flow should be separated into identical but separately labelled operations in the model).

Given λ^{EX}, the solution to the traffic equations $\lambda = (I-Q^T)^{-1} \lambda^{EX}$ exists. Based on the solution of traffic equation, we can get mean total arrival rate of all customers (both non-preemptive high priority customers and non-preemptive low priority customers) to queue k.

$$\forall k \in T,\ \ \Gamma_k = \sum_{\forall i \in T_{k,O}} \lambda_i + \sum_{\forall j \in T_{k,PM}} \lambda_{j,PM}$$

The vector N_i can be obtained by a process similar to solving the traffic equations. That is, $N_i = (I-Q^T)^{-1} e_i$, where e_i is a ν x 1 vector of zeros with a single 1 in the i^{th} row. (Note that this only has meaning for operations i that host an external arrival process.)

2.3 Traffic Variability

Following the nature of semiconductor fabrication systems, in which the manufacturing processes are mostly deterministic, the DWOA method is adopted to calculate the traffic variability. We use DWOA similarly to Kim (2005), but extend it to allow both deterministic and probabilistic routing.

To calculate the CV of interarrival time of lots, additional variables should be calculated first. Let

$$\forall i \in T_k,\ \ \varphi_{k,i} = \lambda_i / \Gamma_k \tag{3}$$

$$\forall i \in O \cup P,\ \ \rho_{\sigma(i),i} = \lambda_i \cdot \frac{S_i}{m_{\sigma(i)}} \quad \left(\lambda_{i,PM} \cdot \frac{D_i}{m_k}, \text{in case of PM type operation}\right) \tag{4}$$

$$\forall \sigma(i) \in T,\ \ \rho_{\sigma(i)} = \textstyle\sum_{i \in T_k} \rho_{\sigma(i),i} \tag{5}$$

In (3), for each operation i served at queue k, the fraction of lots arriving to queue k that require operation i is obtained. The loading due to operation i at queue k is obtained in (4). By summing the loading brought by all operations served at queue k, we obtain the total loading of queue k (5).

We define the set of operations that route to an operation j as follows.

$\forall i \in O,\ G_i = \{\forall j \in O | \mathrm{q_{j,i}} > 0\}$: Set of operations j which satisfy $\mathrm{q_{j,i}} > 0$ for operation i

The core of DWOA is the departure variability approximation from Whitt (1994). The SCV of the departure process from each operation is approximated as a convex combination of two limiting cases: a heavy traffic case (with $\rho \to 1$) and a light traffic case (with $\rho \to 0$). Whitt (1994) provides an approximation for the departure process SCV from operation j in the case of a single server queue is

$$\mathrm{Cd_j^2} \approx \rho_{\sigma(\mathrm{j})}^2 \left\{ \left(\frac{\rho_{\sigma(\mathrm{j}),\mathrm{j}}}{\rho_{\sigma(\mathrm{j})}} \right)^2 \mathrm{Cs_j^2} + \left(1 - \frac{\rho_{\sigma(\mathrm{j}),\mathrm{j}}}{\rho_{\sigma(\mathrm{j})}} \right)^2 \mathrm{Ca}_j^2 + \sum_{\substack{\mathrm{k} \in \mathrm{T_j} \\ \mathrm{k} \neq \mathrm{j}}} \left(\frac{\rho_{\sigma(\mathrm{j}),\mathrm{k}}}{\rho_{\sigma(\mathrm{j})}} \right)^2 \frac{\lambda_\mathrm{j}}{\lambda_\mathrm{k}} \left(\mathrm{Cs_k^2} + \mathrm{Ca_k^2} \right) \right\} + (1 - \rho_{\sigma(\mathrm{j})}^2) \mathrm{Ca_j^2} \tag{6}$$

To approximate the SCV in a multi-server queue, we first replace the variables $\mathrm{Cs_j^2}$ and $\mathrm{Cs_k^2}$ with

$$1 + \frac{\mathrm{Cs_j^2} - 1}{\mathrm{m}_{\sigma(\mathrm{i})}^{0.5}} \tag{7}$$

and

$$\frac{\mathrm{Cs_k^2}}{\mathrm{m}_{\sigma(\mathrm{j})}^{0.5}} \tag{8}$$

respectively. We thus obtain the SVC of the departure process from operation j with a multi-server queue

$$\mathrm{Cd_j^2} \approx \rho_{\sigma(\mathrm{j})}^2 \left\{ \left(\frac{\rho_{\sigma(\mathrm{j}),\mathrm{j}}}{\rho_{\sigma(\mathrm{j})}} \right)^2 \left(1 + \frac{\mathrm{Cs_j^2} - 1}{\mathrm{m}_{\sigma(\mathrm{j})}^{0.5}} \right) + \left(1 - \frac{\rho_{\sigma(\mathrm{j}),\mathrm{j}}}{\rho_{\sigma(\mathrm{j})}} \right)^2 \mathrm{Ca_j^2} + \sum_{\substack{\mathrm{k} \in \mathrm{T_j} \\ \mathrm{k} \neq \mathrm{j}}} \left(\frac{\rho_{\sigma(\mathrm{j}),\mathrm{k}}}{\rho_{\sigma(\mathrm{j})}} \right)^2 \frac{\lambda_\mathrm{j}}{\lambda_\mathrm{k}} \left(\frac{\mathrm{Cs_k^2}}{\mathrm{m}_{\sigma(\mathrm{j})}^{0.5}} + \mathrm{Ca_k^2} \right) \right\} + \left(1 - \rho_{\sigma(\mathrm{j})}^2 \right) \mathrm{Ca_j^2} \tag{9}$$

Using the approximation from Kuehn (1979), the SCV for the portion of departure process from operation j that is routed to a particular next operation i is approximated by

$$\mathrm{Cd}_{j,\mathrm{i}}^2 \approx \left(1 - \mathrm{q_{j,i}} \right) + \mathrm{q_{j,i}} \mathrm{Cd_j^2} \tag{10}$$

For an operation i with multiple predecessors, the asymptotic method described in Whitt (1983) is used to approximate the SCV of the total arrival process (including both exogenous and endogenous arrivals) to that operation as

$$\mathrm{Ca_i^2} \approx \sum_{j \in G_i} \frac{\lambda_\mathrm{j} \mathrm{q_{j,i}}}{\lambda_\mathrm{i}} \mathrm{Cd}_{j,\mathrm{i}}^2 + \frac{\lambda_{\mathrm{0,i}}}{\lambda_\mathrm{i}} \mathrm{Ca_{0,i}^2} \tag{11}$$

Plugging the approximations from (9) and (10) into (11), we obtain the following expression for $\mathrm{Ca_i^2}$. For $\forall i \in O$:

$$\mathrm{Ca}_{\mathrm{i}}^2 = \mathrm{A}_{\mathrm{i}} + \sum_{\substack{\mathrm{j}\in\mathrm{G}_{\mathrm{i}} \\ \mathrm{i}\neq\mathrm{j}}} \left(\mathrm{B}_{\mathrm{i,j}}\right)\mathrm{Ca}_{\mathrm{j}}^2 + \sum_{\substack{j\in G_i \\ \mathrm{i}\neq\mathrm{j}}} \sum_{\substack{\mathrm{h}\in T_j \\ \mathrm{h}\neq\mathrm{j}}} \left(C_{j,h}\right)\mathrm{Ca}_{\mathrm{h}}^2 \tag{12}$$

with A_{i}, $\mathrm{B}_{\mathrm{i,j}}$, $\mathrm{C}_{\mathrm{j,h}}$ are given by:

$$\mathrm{A}_{\mathrm{i}} = \frac{\lambda_{\,0,\mathrm{i}}}{\lambda_{\,\mathrm{i}}}\mathrm{Ca}_{0,\mathrm{i}}^2 + \sum_{\mathrm{j}\in\mathrm{G}_{\mathrm{i}}} \frac{\lambda_{\,\mathrm{j}}\mathrm{q}_{\mathrm{j,i}}}{\lambda_{\,\mathrm{i}}}(1-\mathrm{q}_{j,\mathrm{i}})$$

$$+\sum_{\mathrm{j}\in\mathrm{G}_{\mathrm{i}}} \frac{\lambda_{\,\mathrm{j}}\mathrm{q}_{\mathrm{j,i}}^2}{\lambda_{\,\mathrm{i}}}\left(\rho_{\,\sigma(\mathrm{j})}^{\,2} \sum_{\substack{\mathrm{h}\in\mathrm{T}_{\mathrm{j}} \\ \mathrm{h}\neq\mathrm{j}}} \frac{\left(\varphi_{\sigma(j),\mathrm{j}}\,\rho_{\,\sigma(j),\mathrm{h}}^{\,2}\right)}{\left(\varphi_{\sigma(j),\mathrm{h}}\,\rho_{\,\sigma(j)}^{\,2}\right)} \frac{\mathrm{Cs}_{\mathrm{h}}^2}{\mathrm{m}_{\sigma(\mathrm{h})}^{0.5}}\right) \tag{13}$$

$$+\sum_{\mathrm{j}\in\mathrm{G}_{\mathrm{i}}} \frac{\lambda_{\,\mathrm{j}}\mathrm{q}_{\mathrm{j,i}}^2}{\lambda_{\,\mathrm{i}}}\left(\rho_{\,\sigma(\mathrm{j}),\mathrm{j}}^{\,2}\left(1+\frac{\mathrm{Cs}_{\mathrm{j}}^2-1}{\mathrm{m}_{\sigma(\mathrm{j})}^{0.5}}\right)\right)$$

$$\mathrm{B}_{\mathrm{i,j}} = \frac{\mathrm{q}_{\mathrm{j,i}}^2\,\lambda_{\,\mathrm{j}}}{\lambda_{\,\mathrm{i}}}\left[\rho_{\,\sigma(\mathrm{j})}^{\,2}\left(1-\frac{\rho_{\,\sigma(\mathrm{j}),\mathrm{j}}}{\rho_{\,\sigma(\mathrm{j})}}\right)^2 + (1-\rho_{\,\sigma(\mathrm{j})}^{\,2})\right] \tag{14}$$

$$\mathrm{C}_{j,\mathrm{h}} = \rho_{\,\sigma(\mathrm{j})}^{\,2}\left(\varphi_{\sigma(j),\mathrm{j}}\,\rho_{\,\sigma(j),\mathrm{h}}^{\,2}\right)/\left(\varphi_{\sigma(j),\mathrm{h}}\,\rho_{\,\sigma(j)}^{\,2}\right)\frac{\mathrm{q}_{\mathrm{j,i}}^2\,\lambda_{\,\mathrm{j}}}{\lambda_{\,\mathrm{i}}} \tag{15}$$

Note that the coefficient A_{i} (13) has no interarrival SCV terms; it contains only mean arrival rates and service time statistics. $\mathrm{B}_{\mathrm{i,j}}$ (14) captures the SCV for all operations that are joined together to create the arrival stream for operation i. The coefficient $\mathrm{C}_{j,\mathrm{h}}$ (15) is used to adjust $\mathrm{B}_{\mathrm{i,j}}$ in the case that operation h is serviced at the same queue as an operation j.

2.4 CT Approximation

2.4.1 G/G/m Waiting Time Approximation

Let Wq_k denote the mean queueing delay time at queue k for (product) customers. (Not for the high priority PM customers, though they delay our low priority customers.)

$$Wq_k \approx \frac{\left(\sum_{i\in T_k}\rho_{\sigma(i),i}\right)^{\sqrt{m_k-1}}}{m_k^2} \frac{\sum_{i\in T_{k,O}}\left(Ca_i^2+Cs_i^2\right)\left(\lambda_i {S_i}^2\right)+\sum_{i\in T_{k,PM}}\left(\mathrm{C}_{a,PM_i}^2+CD_i^2\right)\left(\lambda_{i,PM}{D_i}^2\right)}{2\left(1-\sum_{i\in T_k}\rho_{\sigma(i),i}\right)\left(1-\sum_{i\in T_{k,PM}}\rho_{\sigma(i),i}\right)} \tag{16}$$

In this study, we deal with the G/G/m queueing network with non-preemptive high priority (NPPR) customers. There for CT approximation for single toolset with G/G/1/NPPR queueing system in Wu (2014) and G/G/m/NPPR approximation in Connors (1996) are considered and adjusted by using the values of $\mathrm{Ca}_{\mathrm{i}}^2$, $\forall i \in O$ which are unobtainable under the DWA method. The proposed G/G/m/NPPR approximation is shown in (16).

2.4.2 Total CT Approximation

$$TC_i = \sum_{j\in O}\{n_{i,j}S_j\} + \sum_{k\in T}\{n_{i,k}Wq_k\}, \quad \text{for } \{i: \lambda_i^{EX} > 0\} \tag{17}$$

To see the prediction accuracy of our approximation model, we calculate the mean total CT of customers using (17). In this paper, the mean total CT is generated via aggregating all the mean service times and mean queueing delay times in customer route. For probabilistic routing, we apply the mean number of visit to each operation and toolset. As a reminder, TC_i is the mean total CT of customers which start the manufacturing processes by first arriving to operation i.

3 DATASET DESCRIPTION AND SIMULATION SETUP

For two systems, we test our total CT approximation. The systems used are listed below.

- Industry inspired MIMAC dataset 7: We explore the performance of our total CT approximation in this deterministic routing fab model.
- Industry Inspired FAB Dataset: We explore the performance of our total CT approximation in both deterministic and probabilistic routing fab model.

In this section, we describe the datasets used and provide an overview of our simulation study approach and how they are converted into a network of G/G/m queueing networks.

3.1 Industry Inspired MIMAC Dataset 7

To test the conformity of our model to the simulation result in deterministic route system, we use one of the well-known MIMAC datasets which are based on industrial data. We adjust the data to apply to our G/G/m network model as follows.

We use set7 from the MIMAC datasets. It contains 24 toolsets (all of them are single lot processing toolsets), 1 product, and 21 PM types. We use the mean service durations given in the original dataset for the product customers; they are per-lot processing times. Since queues 16, 18, and 23 do not have PMs specified, we give them default PM plans with 360 hours mean interarrival time and 33 hours of mean PM activity duration.

For the basic analysis, the service durations are uniformly distributed in the range from 10% below to 10% above the mean value. The interarrival times for the single customer product are set to 55.029 hours (this gives 90% loading on the bottleneck queue); they are exponentially distributed. All PM setup durations have mean value ½ of the default PM activity duration. The PM service times are Erlang distributed (consisting of two exponential sub-stages). The assumptions on service times, PM service times, and interarrival times distributions are followed by Morrison et al. (2014).

We conduct sensitivity analysis using dataset 7 on system loading, the service time distributions and the interarrival time distributions. There are 10, 5 and 6 different cases considered for each, respectively.

- Sensitivity to bottleneck queue loading: 10 (loading = 90, 91, 92, 93, 94, 95, 96, 97, 98, 99%)
- Sensitivity to service time distributions: 5 (SCV = 0.003, 0.030, 0.083, 0.163, 0.270)
- Sensitivity to interarrival time distributions: 6 (SCV = 0.0625, 0.125, 0.25, 0.5, 1, 2)

The numbers in parenthesis above are the specific values of % loading of the bottleneck toolset, squared CV (SCV) of service times and SCV of the interarrival times. There are 22 experiments implemented with this system.

3.2 Industry Inspired FAB Dataset

To check the conformity of our model to the simulation result in both deterministic and probabilistic route systems, we use industry inspired fab dataset. Unlike MIMAC datasets, this is not publically available simulation model. In here, we only describe the adjusted version of this dataset.

It contains 14 toolsets (all of them are single lot processing toolsets), 1 product, and 14 PM types. There are totally 39 servers, 31 normal operations, and 10 probabilistic routing processes in this network.

Followed by previous work, the service durations are uniformly distributed with 10% around the mean value for the basic analysis,. The interarrival times for the single customer product are set to give 90% loading on the bottleneck queue. They are exponentially distributed. The PM service times are Erlang distributed (consisting of two exponential sub-stages).

We also conduct sensitivity analysis using this dataset on similar scope of MIMAC dataset analysis.

- Sensitivity to bottleneck queue loading: 6 (loading = 90, 91, 92, 93, 94, 95%)
- Sensitivity to service time distributions: 5 (SCV = 0.003, 0.030, 0.083, 0.163, 0.270)
- Sensitivity to interarrival time distributions: 6 (SCV = 0.0625, 0.125, 0.25, 0.5, 1, 2)

The numbers in parenthesis above are the specific values of % loading of the bottleneck toolset, squared CV (SCV) of service times and SCV of the interarrival times. There are 17 experiments implemented with this system.

3.3 Simulation Setup

For simulation, we use AutoSched AP software with 20 years of warm-up and 50 years of data acquisition. Such a long period is required as the duration of time between PM event arrivals can be very long (e.g., 3 months). We want the data to contain at least 500 such events for each PM type. We use 30 replications.

4 NUMERICAL STUDY: TOTAL CT

Here we focus on comparing the mean total CT obtained via approximation and simulation.

4.1 Numerical Study in Deterministic Routing Dataset

Table 1 summarizes the results of the simulation study for MIMAC dataset 7. There the mean total CT for product customers is shown for both the approximation and simulation. The error is -5.77%.

Table 1: Summary of the numerical study with MIMAC dataset 7.

Title	Mean total CT comparison		
	Approximation (h)	Simulation (h)	Difference (%)
MIMAC dataset 7	1513.55	1606.27	-5.77

4.2 Numerical Study in Deterministic and Probabilistic Routing Dataset

Table 2 shows the results of the simulation study using industry inspired fab dataset. The error is 8.84%.

Table 2: Summary of the numerical study with industry inspired fab dataset.

Title	Mean total CT comparison		
	Approximation (h)	Simulation (h)	Difference (%)
Industry inspired fab dataset	1506.96	1384.58	8.84

5 NUMERICAL STUDY: SENSITIVITY ANALYSIS

We conduct sensitivity studies on bottleneck utilization, interarrival time distribution and service time distribution. Table 3 provides the detailed results.

5.1 Sensitivity to Bottleneck Queue Loading

We consider 10 cases in the MIMAC dataset in which bottleneck loading varies from 90% to 99% in 1% increments. We consider 6 cases in the industry inspired fab dataset in which bottleneck loading varies from 90% to 95% in steps of 1%. This is accomplished by changing the lot arrival rate. For each system, the mean total CT obtained via approximation and simulation are shown in Figures 2. For both cases, as the bottleneck queue loading increased, simulated mean total CT also gradually increased. Our approximation model predicts the simulation results fairly well across a broad range of parameter input values. In MIMAC dataset, the minimum and maximum differences of the mean total CT are -0.63% and -8.92%. For industry inspired fab dataset, those differences are 8.84% and 11.91%.

5.2 Sensitivity to Service Time SCV

To check the tendency when the service time distribution changes, we modify the SCV value of service time from almost 0 (0.003) to above 0.25. These values are obtained for uniformly distributed service time with the ranges of [0.9·MS, 1.1·MS], [0.7·MS, 1.3·MS], [0.5·MS, 1.5·MS], [0.3·MS, 1.7·MS], and [0.1·MS, 1.9·MS], where MS is the mean service time for that operation.

Table 3: Summary of the sensitivity analysis.

Category		Mean total CT comparison (Difference, %)	
		MIMAC dataset 7	Industry inspired fab dataset
Sensitivity 1. Bottleneck queue loading	90.0%	-5.77	8.84
	91.0%	-5.86	9.06
	92.0%	-6.64	10.10
	93.0%	-6.57	10.63
	94.0%	-8.92	11.04
	95.0%	-8.74	11.91
	96.0%	-8.11	-
	97.0%	-7.85	-
	98.0%	-4.70	-
	99.0%	-0.63	-
Sensitivity 2. Service time distribution	Uni10%	-5.77	8.84
	Uni30%	-5.69	6.93
	Uni50%	-6.03	4.50
	Uni70%	-5.88	3.97
	Uni90%	-5.92	3.50
Sensitivity 3. Interarrival time distribution	Gam,16	-9.78	7.03
	Gam,8	-7.35	8.10
	Gam,4	-5.27	8.69
	Gam,2	-6.88	7.89
	Gam,1	-5.77	8.84
	Gam,0.5	-5.22	5.73

Figures 3 shows the results of the simulated mean total CT with its approximated value for two systems. For both cases, the mean total CT increases when the SCV value of service times increased. This is well matched with the expected performance.

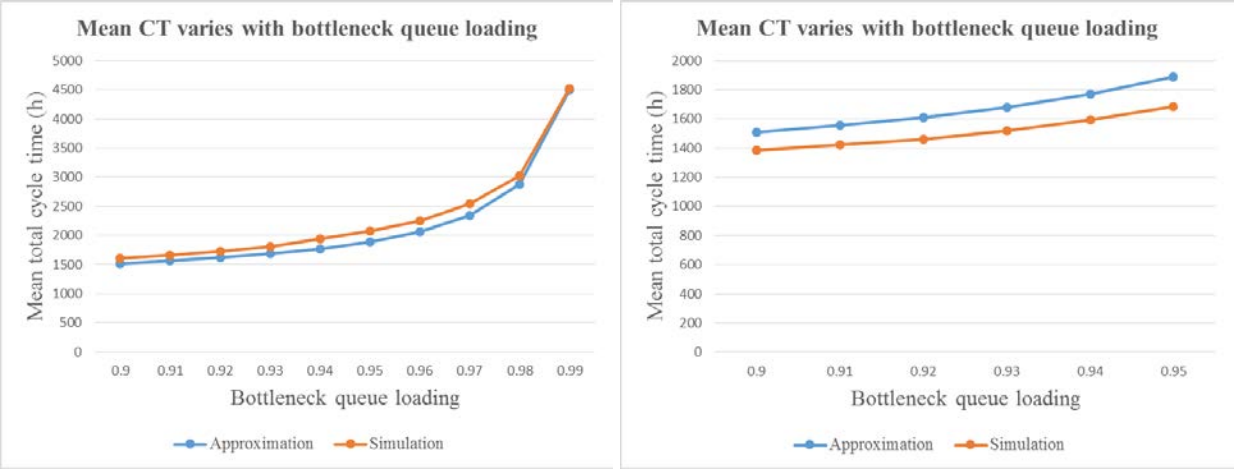

Figure 2: Simulated mean total CT as the product customer arrival rates drive the bottleneck loading from 90-99% (Left: MIMAC dataset, deterministic routing system / Right: Industry inspired fab dataset, deterministic and probabilistic routing system).

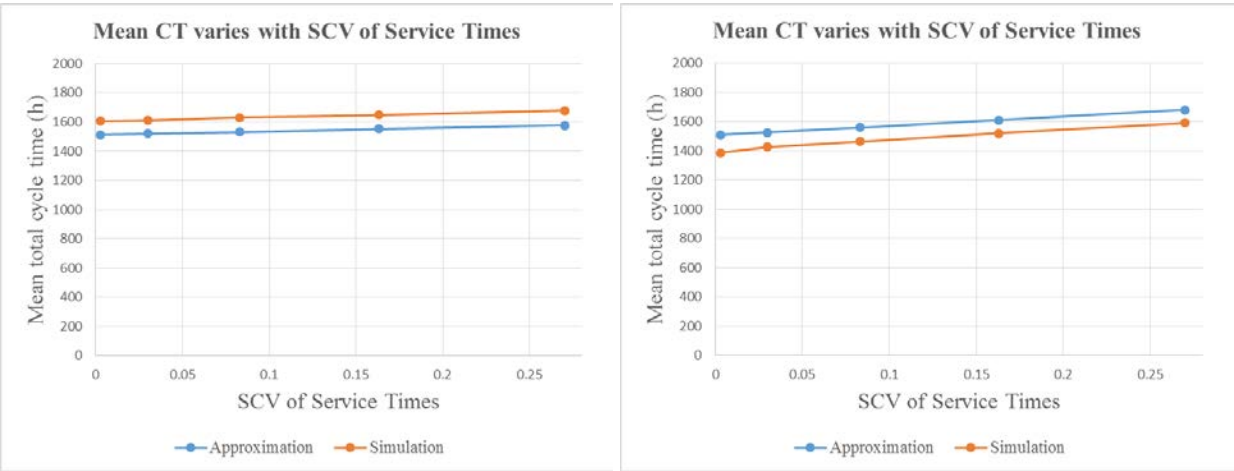

Figure 3: Mean total CT increases with the SCV of service times (Left: MIMAC dataset, deterministic routing system / Right: Industry inspired fab dataset, deterministic and probabilistic routing system).

5.3 Sensitivity to Interarrival Time SCV

We focus on the effect of the SCV of interarrival times. In basic analysis, we consider the interarrival times to be exponentially distributed for both product customers and PM type customers. For this study, we hold the mean values and vary the number of Erlang sub-stages in the range 0.5, 1, 2, 4, 8, and 16. The SCV values are 2, 1, 0.5, 0.25, 0.125 and 0.0625, respectively.

For both systems, Figures 4 shows how the mean total CT for product customers increases when the SCV for interarrival times increases. The observed increasing tendency of the mean total CT is as expected. Our approximation predicts the simulation results evenly in this sensitivity analysis also.

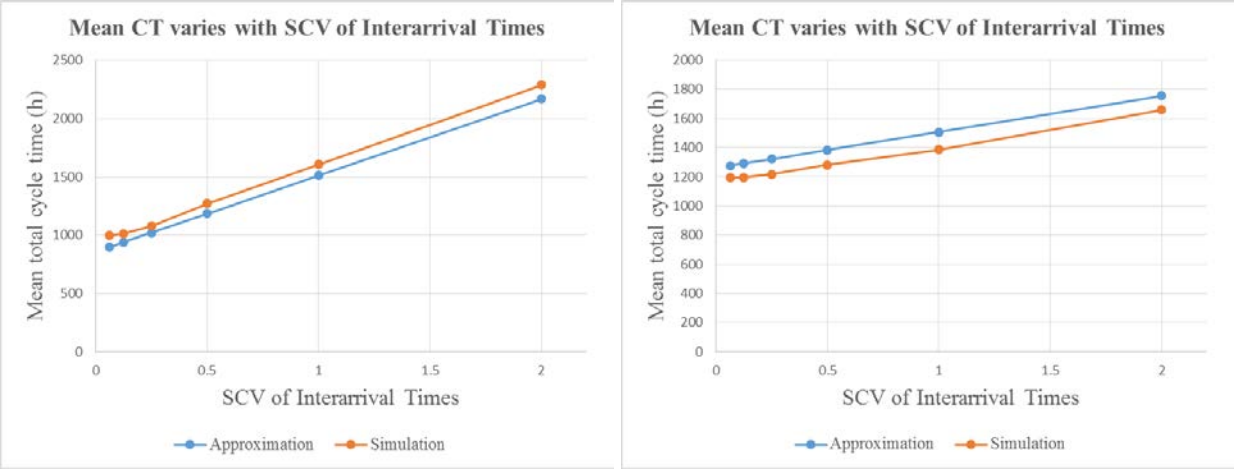

Figure 4: Mean total CT increases with the SCV of interarrival times (Left: MIMAC dataset, deterministic routing system / Right: Industry inspired fab dataset, deterministic and probabilistic routing system).

6 CONCLUDING REMARKS

In this paper, we propose extensions to approximation methods for G/G/m queueing networks that are suited for fab modeling using decomposition without aggregation. To study the approximation performance, we first compare the simulated mean total CT to approximated mean total CT using two industry inspired datasets. The model based prediction on total CT shows satisfactory accuracy. By varying some parameters in datasets, sensitivity studies were conducted. The model has errors less than 12% in all cases and less than 5% in most cases. One observation from our numerical study is that the prediction errors are always negative in the deterministic routing system and always positive in the system with probabilistic routing. If this difference is the result of systematic differences between the systems, we may be able to use this observation to improve our approximation model. We will investigate this in more detail in the future.

In the future, we plan to modify and apply the state of art batching approximation to our model. It would be of interest to utilize the proposed approximations in a real fab to promptly identify extreme changes in the cycle time behavior across a segment of operations or a segment of tools and address those as they evolve into issues.

REFERENCES

Bitran, G. R., and D. Tirupati. 1989. "Multiproduct Queuing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference." *Management Science* 34: 75-100.

Connors, D. P., G. E. Feigin, and D. D. Yao. 1996. "A Queueing Network Model for Semiconductor Manufacturing." *IEEE Transactions on Semiconductor Manufacturing* 9: 412-427.

Grosbard, D., A. A. Kalir, I. Tirkel, and G. Rabinowitz. 2013. "A Queuing Network Model for Wafer Fabrication Using Decomposition without Aggregation." Automation Science and Engineering (CASE), *IEEE International Conference* 717 – 722.

Fowler, J. W., and J. Robinson. 1995. "Measurement and Improvement of Manufacturing Capacity (MIMAC): Final Report." *SEMATECH, Technology Transfer.* #95062861A-TR.

Global Semiconductor Forum. 2011. "Elpida Memory 300mm Wafer Fab, Hiroshima, Japan." http://www.semiconductor-technology.com/projects/elpida. [Accessed: August 1, 2013].

Kim, S. 2005. "Approximation of Multiclass Queuing Networks with Highly Variable Arrivals Under Deterministic Routing." *Naval Research Logistics* 52: 400-408.

Kuehn, P. J. 1979. "Approximate Analysis of General Queuing Networks by Decomposition." *IEEE Transactions on Communications* 27: 113-126.

Lapedus, M. 2010. "EUV Tool Costs Hit $120 Million." *EE Times*. http://www.eetimes.com/document.asp?doc_id=1257963.

Morrison, J. R., H Kim, and A. A. Kalir. 2014. "Mean Cycle Time Optimization in Semiconductor Tool Sets via PM Planning with Different Cycles: A G/G/m Queueing and Nonlinear Programming Approach." *In Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 2466-2477. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Reiman, M. I., and B. Simon. 1990. "A Network of Priority Queues in Heavy Traffic: One Bottleneck Station." *Queueing Systems* 6: 33-57.

Whitt, W. 1983. "The Queueing Network Analyzer." *Bell System Technical Journal* 62: 2779–2815.

Whitt, W. 1994. "Towards Better Multi-Class Parametric-Decomposition Approximations for Open Queueing Networks." *Annals of Operations Research* 48: 221-248.

Wu, K. 2014. "Classification of Queueing Models for a Workstation with Interruptions: A Review." *International Journal of Production Research* 52: 902-917.

AUTHOR BIOGRAPHIES

JINHO SHIN is an PhD student in the Department of Industrial and Systems Engineering, KAIST, South Korea. He holds B.S. degrees in Industrial and Systems Engineering, Business and Technology Management, and M.S. degrees in Industrial and Systems Engineering from KAIST, South Korea. His email address is tlswlsgh3@kaist.ac.kr.

DEAN GROSBARD holds a B.S in Mathematics and a B.S and M.S in Industrial Engineering from Ben-Gurion University of The Negev, Beer-Sheva, Israel. He is currently a PhD student at the University of California at Berkeley. His email address is dean.grosbard11@gmail.com.

JAMES R. MORRISON is an Associate Professor in the Department of Industrial and Systems Engineering, KAIST, South Korea. He holds B.S. degrees from the University of Maryland at College Park, USA and M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign, USA. He is a co-Chair of the IEEE RAS Technical Committee on Semiconductor Manufacturing Automation. His email address is james.morrison@kaist.edu.

ADAR A. KALIR is a Senior Principal Engineer with the Fab/Sort Manufacturing Division of Intel Corporation, Qiriat-Gat, Israel. He holds a B.S. and M.S. in Industrial Engineering from Tel-Aviv University, Israel, and a Ph.D. in Industrial Engineering and Operations Research from Virginia Tech, USA. He also serves as an Adjunct Professor at Ben-Gurion University, Israel. His e-mail address is adar.kalir@intel.com.