The determination and indetermination of service times in manufacturing systems

Wu, Kan; Hui, Keung

2008

Wu, K., & Hui, K. (2008). The Determination and Indetermination of Service Times In Manufacturing Systems. IEEE Transactions on Semiconductor Manufacturing, 21(1), 72-82.

https://hdl.handle.net/10356/79917

https://doi.org/10.1109/TSM.2007.914334

The Determination and Indetermination of Service Times in Manufacturing Systems

Kan Wu[a] and Keung Hui[b]

[a]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA USA (e-mail: kanwu@gatech.edu).
[b]18/F 2–5 Moreton Terrace, Hong Kong (e-mail k.hui@graduate.hku.hk).

Abstract

The notion of service times is of such fundamental importance in the analysis of queues that it has long been taken for granted. Intuitively, it is used to represent the time interval that a server is capable of completing a dispatched job. However, actual measurements of service times under simple queues in production lines have encountered practical difficulties, in spite of its seemingly deterministic nature.

Previous studies have introduced concepts of effective process times to quantify service times. Besides notions of theoretical processing times, raw process times and queueing times, among others, are commonly used in various applications. Their existence causes confusion in the determination of service times and clarification of such terminologies is needed.

A simple model is examined to quantify the various concepts and establish their interrelationships. This paper brings out new properties of effective process times with a dynamic dependence on utilization. Discrete event simulations are conducted to verify these properties and explain the phenomenon of indetermination of service times. Both theoretical prediction and simulation results show that unless the system is fully loaded, service time and effective process time are not equivalent and it cannot be measured directly from observations of effective process times.

Index Terms—Cycle time, effective process time, queueing time, service time, takt time, theoretical processing time, utilization.

I. INTRODUCTION

IN theoretical analyses of queues, the notion of service times is intuitive and it is used to describe the period a server takes to complete a dispatched job. It plays so fundamental a role in the development of queues that it is taken for granted. After all, with ordinary mechanical systems in manufacturing factories, motions of the moving parts can be precisely clocked to determine the "service times."

Against this intuition, actual determinations of service times defined in most queues have encountered practical difficulties in manufacturing systems. In this paper, we will

show that if down time is apparent, then systematic errors will be induced due to the interferences of work-in-process (WIP)-independent events.

Among other reasons, existence of different states of the servers (or manufacturing lines) complicates assignments of service times. For single-server systems, queueing lots exist when the arrival interval is shorter than its "service time." A machine is classified as idle when no WIP exists in front of this machine and classified as busy when WIP exists, whether it is broken down or actually running [1]. Conventional wisdom has it that impacts by loading conditions are constant in the determination of service times.

Previous studies have tried to give operational definitions to service times in order to manage manufacturing processes using queueing theory. To apply M/M/1 (or G/G/1) queues, Hopp and Spearman [1] defined service time by the notion of effective process time (EPT), which includes theoretical processing time, setup, breakdown, operator availability, and all operational times due to variability effects. Sattler [2] defined an average lot service time as "*all cycle time except time waiting for another lot*" but acknowledged that it "*seems to be a measurable variable tends to be difficult to measure accurately*." Jacobs *et al.* [3] adopted the concept of capacity claiming and were the first to propose algorithms for the practical measurements of EPT, defined as "*the total amount of time a lot claims capacity of a machine, even if it is not yet being processed*."

On the other hand, Hopp and Spearman defined utilization as

$$\text{utilization} = \frac{t_E}{t_A m} \qquad (1)$$

where m denotes the number of identical machines, and t_A and t_E are the mean interarrival time and mean EPT [1, p. 268], which can then be calculated from (1) if utilization is known.

Apart from descriptive arguments, it is yet to be proven that the above definitions of EPT are consistent with each other or equivalent to service time under any circumstances.

Service time is often assumed independent of arrival rate. In communication networks this is expressed as the "Kleinrock In-dependence Assumption" which treats queues at each server as independent of the flows between servers [4]. On the other hand, there are also studies on queues for which the service times are dependent on the arrival processes [5]. In either case, service time is assumed constant and measurable.

However, when EPT is measured, there exists a systematic gap from the mean service time as expected from theoretical values. This discrepancy brings about questions on the sources of indetermination of service times. Preliminary investigation readily points out one of the major sources as in the notion of service time itself, defined as statistical aggregate only, rather than for individual jobs. Another is the presence of disturbances, inherent in any server system but rarely investigated.

These observations call for clearer definitions to different kinds of process times (such as service time, EPT, and raw process time, etc.) as employed in manufacturing lines. Next, it demands elucidation of the differences between EPT and service times in the

presence of disturbances. This discrepancy depends on utilization levels and cannot be explained away by any known results of queueing theory.

With the introduction of a fundamental concept classifying all disturbances into two categories, this paper first restates definitions of the various terms in Section II and, assisted by a graphical illustration of the events, mathematical relationships are established to quantify these terms with respect to each other. For the first time in literature, these relationships thoroughly clarify the differences among notions of "service times" on a consolidated basis. An analytical estimate of EPT is formulated in Section III to elucidate its properties. For readers interested in conceptual clarifications only, they may skip the theoretical predictions of EPT in Section III. Simulation results are shown in Section IV to verify the predictions. Concluding remarks are summarized in Section V.

II. DEFINITIONS

Many kinds of operational process times are used to describe various behaviors of machines, such as theoretical processing time, raw process time, effective process time, and service time, etc. Definitions of these terms lead to different values when actual determinations are made [1]–[3] and they have yet to establish standard interpretations to unify their usage. The following proposes an interpretation of these terms and the subsequent establishment of interrelationships among themselves. Whereas the definitions adopted here may not necessarily claim originality nor pretend to be the most fundamental, the attempt to clarify the current confusion is genuine. It is hoped to facilitate subsequent determinations of these terms in practice.

WIP are semi-products in production lines. Impacts of stochastic events on service times can be classified into two categories: Type-I for WIP-dependent and Type-II for WIP independent events. For example, machine failure appears to be time dependent, whether there is WIP (as idling tools break down on their own all the time), and should be classified as a Type-II event. On the other hand, natural fluctuations due to differences in operators, machines, and materials, constitute Type-I events, as their impacts are only apparent when WIP exists. Further examples are listed as follows:

1) Type-I product mix, natural fluctuations, setups;
2) Type-II machine failures, unavailability of operators, engineering time, preventive maintenance.

For subsequent definitions of various terms, the time instants of lot arrival, start to processing, failure, recovery and unloading, of a typical single-server system are illustrated in Fig. 1. These time instants can be exactly tracked and are thus deterministic. This general model unifies analysis and not all the time segments necessarily occur for a particular lot in practice.

Fig. 1 is of fundamental importance to an understanding of the following development. It is a combination of two related scenarios. The upper part describes the lifecycle of a lot

which comes when the machine is idle. The lower part describes the lifecycle of a lot which comes when the machine is busy.

For lot-k to depart from the unloading port of a machine at time instant $D(k)$, two possibilities exist: either the lot arrives at $A_1(k)$ after, or at $A_2(k)$ before, the previous lot-$(k-1)$ departs from the unloading port at $D(k-1)$. For $A_1(k)$, there is a period of idling time as there is no WIP in the system. For $A_2(k)$, there is always WIP in the system and no idling time presents.

$A_1(k)$	Arrival instant of lot-k for the case of empty WIP.
$A_2(k)$	Arrival instant of lot-k for the case of existing WIP.
$D(k-1)$	Departure time of lot-$(k-1)$ from the load port.
$D(k)$	Departure time of lot-k from the load port.
$F(k)$	Instant of failure after processing of lot-k started.
$R(k)$	Instant of recovery after $F(k)$.
$S(k)$	Instant of lot-k processing starts.
V_1, V_2	Fictitious instant collectively partitioning Type-I and -II events after $A_1(k)$ or $A_2(k)$, respectively.
a	Duration of processing resumed upon recovery.
b	Duration between tool failure and recovery (including all Type-II events after processing of lot starts).
c	Duration of processing taken before failure.
d_1, d_2	Durations of all Type-II events for $A_1(k)$ and $A_2(k)$, respectively, before processing of lot starts.
e_1, e_2	Durations of all Type-I events for $A_1(k)$ and $A_2(k)$ respectively before processing of lot starts [collectively denoted by $e(k)$].
f	Idling duration for the case of empty WIP.
g	Waiting duration of lot-k between its arrival and the departure of lot-$(k-1)$.
h_1	$e_1 + d_1 =$ duration of lot-k between claiming capacity and actual start of processing after $A_1(k)$.
h_2	$e_2 + d_2 =$ duration of lot-k between claiming capacity and actual start of processing after $A_2(k)$.
h	h_1 or $h_2 =$ collective representation of h_1 and h_2.
π_1	$d_1 + b =$ total duration of Type-II events for $A_1(k)$.
π_2	$d_2 + b =$ total duration of Type-II events for $A_2(k)$.
π	π_1 or $\pi_2 =$ collective representation of π_1 and π_2.
Remarks:	Arguments of all durations are omitted for simplicity, i.e., $a = a(k)$, etc.

Without loss of generality, it assumes that all the Type-I events (before processing starts) are grouped together and their completion collectively marks a fictitious time

$A_1(k)$

$D(k-1)$ $A_1(k)$

instant (V_1 or V_2) after the lot claims capacity of the tool. Henceforth, all the Type-II events (before processing starts) collectively follow this fictitious time instant immediately. Once processing of the lot starts, all the Type-II events are assumed to occur between the time instants of machine failure and recovery. It is assumed that processing is resumed upon recovery if failure ever interrupts. Afterwards, all the other Type-I events are diffused somewhere throughout the two segments of processing durations.

The analysis that follows caters to simple nontandem systems of single-servers processing single-lots, thus excluding those multichamber toolsets, multilot furnaces, and continuous flow wet benches. Multitank wet benches are networks of servers in tandem formation and cannot be taken as single-server systems. For this paper, any action parallel to the critical task (such as concurrent loading–unloading) is taken as time instant of zero duration. Segment b is measurable for disruptive events. For nondisruptive interferences and/or breakdowns in the running of the machine which one may not know from the outside, b may not be measurable and is left for further study.

According to SEMI E79 [7], setup time can be considered as part of the processing duration if it occurs in every cycle. Various definitions of processing times shall be interpreted in terms of the time instants and durations as noted in Fig. 1.

A. Service Time, Takt Time, and Theoretical Processing Time

State changes of WIP for a single-server queueing system are results of the combined effects of arrival and service times [6]. Service time in this paper is defined as:

"the time interval that a machine is capable to complete a dispatched job taking into account all the stochastic effects such as machine failure, setups and unavailability of operators".

This value reflects the true capability of a machine and for single-server systems processing single-lots, its capacity can be occupied by one and only one lot at a time.

According to SEMI E79 [7], theoretical unit throughput (TUT) is defined as "*the number of unit per period of time that could be processed by the equipment under ideal conditions of continuous productive time operation in which no equipment effectiveness losses are present*", where the *productive time* is "*a period of time when the equipment is performing its intended function*." During this observation period, WIP always exists and there are no idling times, no failures, and there is no time lost due to either Type-I or Type-II events.

Following the same concept from E79, theoretical processing time (TPT), being the reciprocal of TUT, is defined as "*the time interval that a machine is capable of completing a dispatched job under ideal conditions of continuous productive time operation*."

Takt time is a notion employed in continuous productive operations. It is originally defined as "*the time that a complete product is finished*" [8], or "*the desired time between*

units of production output, synchronized to customer demand" [9]. Bulletins and application of takt charts were reported [10], [11].

From Fig. 1, takt time without Type-II events for lot- in this paper is defined as "*the time interval between the departure times of two consecutive lots, excluding all the idling times* (*if any*) *and all the Type-II events* (*if any*) *occurred in between*."

Symbolically, takt time $T(k)$ is given by (see Fig. 1)

$$T(k) = \begin{cases} D(k) - D(k-1) - f(k) - \pi_1(k), & \text{for } A_1(k) \\ D(k) - D(k-1) - \pi_2(k), & \text{for } A_2(k). \end{cases} \tag{2a}$$

which collectively simplifies to

$$T(k) = a(k) + c(k) + e(k). \tag{2b}$$

Let there be a total of n lots observed over some time period, a metric of TPT for lots is selected as

$$T_{\min}(n) = \min\{a(k) + c(k)\ k = 1 \ldots n\}. \tag{2c}$$

Equation (2b) still encounters practical difficulties as impacts of other Type-I events diffused throughout the processing segments, $a + c$, are hard to gauge. For practical purposes, TPT for the simple nontandem server is approximated by TPT for the lots when the observation is made over a sufficiently long period of time

$$\begin{aligned} \text{TPT} &\cong \lim_{n\to\infty} T_{\min}(n) = T_{\min}(\infty) \\ &= \min\{a(k) + c(k)\ k = 1 \ldots \infty\}. \end{aligned} \tag{2d}$$

In this way, TPT for the server may be expected to be free from impacts of both Type-I and Type-II events.

For the n lots, impacts on $T_{min}(\infty)$ from Type-I events can be assessed through the average deviation of $T(k)$ from $T_{min}(n)$. Define an intensity factor for these impacts as

$$I_{\text{SE}} \triangleq \frac{\text{mean}T(k)}{T_{\min}(\infty)} = \frac{\overline{T(k)}}{T_{\min}(\infty)} \triangleq \frac{\frac{1}{n}\Sigma_{k=1}^{n} T(k)}{T_{\min}(n)} \geq 1. \tag{3}$$

I_{SE} is for Type-I but free from Type-II events. It is the reciprocal of rate efficiency as defined in E79. When product mixes, setups, or natural fluctuations exist, I_{SE} exceeds unity.

A practical definition of utilization is "*ratio of the total number of lots arrived and the number of lots capable of being processed by the server over a long period of time*."

With assumptions of process-resumption upon recovery and all Type-II events being grouped as machine failures, utilization and service time of a single-server can be expressed as in

$$\text{utilization} = \frac{\text{Total number of lots arrived}}{\text{Number of lots capable of being processed}}. \tag{4a}$$

For a server to become productive, it requires that the machine is up and there is available labor to man the tool (if necessary). If tool and operator availability are independent, total availability A of the server is defined as

$$A = A_{\text{tool}} \times A_{\text{operator}} \tag{4b}$$

where tool availability A_{tool} = MTBF/(MTBF + MTTR) [1]. MTBF is mean time between failures and MTTR the mean time to repair for the server. Operator availability (as well as other availabilities of Type II events) $A_{operator} = 1$ is often assumed.

From definition of TPT, the expected number of lots capable of being processed N_{cap} over any observation period T_{obs} is

$$N_{\text{cap}} = \frac{T_{\text{obs}} \times A}{\text{mean}T(k)} = T_{\text{obs}} \times \text{Service Rate}. \tag{4c}$$

Based on the assumptions of M/M/1 queue, utilization is

$$\begin{aligned} u = \text{utilization} &= \frac{T_{\text{obs}} \times \text{Arrival Rate}}{T_{\text{obs}} \times \text{Service Rate}} \\ &= \frac{\text{Arrival Rate}}{\text{Service Rate}}. \end{aligned} \tag{4d}$$

The expected service time $\bar{S}_T$ is therefore

$$\bar{S}_T = \text{Mean Service Time} = \frac{1}{\text{Service Rate}}. \tag{4e}$$

In terms of $T(k)$ from (3) and (4c), it becomes

$$\bar{S}_T = \frac{\overline{T(k)}}{A} = \frac{T_{\min}(\infty) \times I_{\text{SE}}}{A}. \tag{4f}$$

This relates *service time* to *theoretical processing time* on the metric of statistical means. So far, none of the above defines the *individual* service time, $S_T(k)$, for lot k. That is, the distribution profile of service time has not been specified anywhere.

The mean service time in (4f) consists of two components: mean takt time and server availability. The former contains all the Type-I events and the latter all Type-II ones. Type-I events lengthen the expected TPT to reflect the true throughput capability,

whereas Type-II events reduce the amount of available times to serve. In either case, it lengthens $\bar{S}_T$.

Identification of the two components enables investigation of their effects separately. Focusing on impacts of Type-II events, it suffices to study situations where $\overline{T(k)}$ remains constant. In this paper, we will show that impacts of Type-II events lead to some of the practical difficulties in the determination of service times and how to model these impacts.

B. Effective Process Time

Although a method to calculate the service time in principle as a statistical mean is demonstrated above, measurements of the service times for individual lots are not clear. Previous studies have tried to gauge the real service time through EPT [1]–[3]. From Jacobs, *et al.* [3], EPT is "*the total amount of time a lot claims capacity of a machine*" even if it is not yet being processed. It includes waiting for machine failure, waiting for operator, setup time, load, orientation, pumping down, robot transferring, processing, unload, and all other operational times due to variability effects. It is expressed as

$$\text{EPT} = \text{Lot departure time} - \text{The time that the lot claims capacity of the machine} \quad (5a)$$

where lot departure time is the time that a lot departs from the unloading port of the server. With the first-come-first-serve (FCFS) dispatching policy, a lot claims capacity of a machine if:

1) the lot is present in front of this machine;
2) it is after the previously processed lot departs;

whether the lot is actually loaded into this machine or not. Subsequently, EPT of lot k in Fig. 1 is given by

$$E(k) = \begin{cases} D(k) - D(k-1) - f, & \text{for } A_1(k) \\ D(k) - D(k-1), & \text{for } A_2(k). \end{cases} \quad (5b)$$

which collectively simplifies to

$$E(k) = a(k) + c(k) + e(k) + \pi(k). \quad (5c)$$

Consequently, EPT is related to $T(k)$ as

$$E(k) = T(k) + \pi(k) \quad (5d)$$

and the relationship to mean service time is

$$\overline{E(k)} = \overline{T(k)} + \overline{\pi(k)} = A\bar{S}_T + \overline{\pi(k)} \quad (5e)$$

which literally translates into

$$\begin{aligned}&\text{the expected value of EPT}\\&\quad=\ \text{availability}*\text{mean service time}\\&\qquad+\ \text{the expected duration of all Type-II events.}\end{aligned}$$

Explicitly, EPT is subject to the combined impacts of Type-I and Type-II events. For actual measurements over an observation period, if a total of n lots are processed, the mean EPT of the server during this period is given by

$$\begin{aligned}\overline{E(k)}&=\frac{\text{summation of all EPTs}}{\text{total number of lots processed}}\\&=\frac{1}{n}\sum_{k=1}^{n}E(k).\end{aligned}\tag{5f}$$

With a FCFS policy, the summation of EPT of a single server is equivalent to the sum of all periods for which WIP exists.

From (5e), determination of the expected EPT practically boils down to an even attribution of the impacts of all Type-II events on the total number of lots processed over the observation period. Accordingly, it implies that the mean EPT depends on the utilization levels as illustrated next.

C. Raw Process Time and Queueing Time

Raw process time here is defined as "*the total duration that a lot stays in a tool and is engaged in process related activities*." With the capability to resume processing upon recovery, the raw process time (PT) of lot k for both cases of WIP in Fig. 1 is

$$P(k)=D(k)-S(k)-b(k)=a(k)+c(k)\tag{6}$$

which naturally contains impacts of Type-I events diffused somewhere in $a(k)$ and/or $c(k)$. Since $T(k)=P(k)+e(k)$, EPT is explicitly related to PT as

$$E(k)=P(k)+h(k)+b(k).\tag{7}$$

Literally, EPT is the sum of PT, duration after claiming capacity but before actual processing starts, and repair time (possibly with all Type-II events between process-start and departure).

An operational definition of queueing time is

$$Q_{\text{op}}=\text{Lot start time}-\text{Lot arrival time.}\tag{8a}$$

Simplification leads to

$$\begin{aligned}&Q_{\text{op}}(k)\\&\quad=\begin{cases}S(k)-A_1(k)=h_1(k), & \text{for no WIP.}\\ S(k)-A_2(k)=h_2(k)+g(k), & \text{for WIP.}\end{cases}\end{aligned}\tag{8b}$$

An effective queueing time is defined when the WIP level is greater than the server counts. For a workstation of m identical machines, the effective queueing time Q_E is defined as

$$\begin{aligned}&[1]\quad \text{WIP} <= m \quad Q_E = 0.\\&[2]\quad \text{WIP} > m \quad Q_E = \text{The time}\\&\text{that the lot claims capacity of the machine}\\&\quad - \text{Lot arrival time}\end{aligned} \tag{9a}$$

giving

$$Q_E(k) = \begin{cases} 0, & \text{for } A_1(k).\\ g(k), & \text{for } A_2(k).\end{cases} \tag{9b}$$

Henceforth

$$Q_{\text{op}}(k) = h(k) + Q_E(k). \tag{9c}$$

With cycle time defined as

$$\text{Cycle Time} = \text{Lot departure time} - \text{Lot arrival time} \tag{10a}$$

which simplifies to

$$\begin{aligned}C(k) &= P(k) + Q_E(k) + h(k) + b(k)\\&= P(k) + Q_{\text{op}}(k) + b(k).\end{aligned} \tag{10b}$$

Notions of queuing time by Q_{op} or Q_E are not equivalent to the queueing time as constructed in queueing theory. Statistically the latter is expressed as

$$\bar{C} \triangleq \bar{S}_T + \bar{Q}_T. \tag{10c}$$

As with service time $S_T(k)$, queueing time for individual lots $Q_T(k)$ is not specifically defined and, therefore, not measurable.

To assign a distribution profile to $Q_T(k)$, from 10(b)

$$Q_T(k) \triangleq P(k) + Q_{\text{op}}(k) + b(k) - S_T(k) \tag{10d}$$

if the service time for individual lots $S_T(k)$ is ascertained. Equation (10d) is intended for a simple nontandem system only. For tools outside scope of this paper, there may be other terms to consider such as load and unload time.

D. Normalized Cycle Time of Single Server System

An approximation of queueing time $\overline{Q_T}$ by the P-K formula for a single-server queueing system of the *G/G*/1 model is [13]

$$\bar{Q}_T(G/G/1) = \alpha \frac{u}{1-u} \bar{S}_T \qquad (11)$$

with utilization u and variability index α denoting

$$\alpha = \frac{1}{2}\left(C_a^2 + C_s^2\right) \qquad (12)$$

where C_a^2 is the squared coefficient of variation (SCV = σ^2/μ^2) of the arrival interval, and C_s^2 is that of the service time, with variance σ^2 and mean μ.

The normalized cycle-time, defined as "*ratio of cycle time divided by raw process time*," has its expected mean given by

$$\begin{aligned} \overline{C_N(k)} &= \frac{\overline{S_T(k)}}{\overline{P(k)}} + \frac{\overline{Q_T(k)}}{\overline{P(k)}} \\ &= \left(1 + \alpha \frac{u}{1-u}\right) \frac{\overline{S_T(k)}}{\overline{P(k)}}. \end{aligned} \qquad (13)$$

Two confusions often arise in the literature. One is in (11), queueing time is approximated by service times, not by EPT [12]. The second is raw process time $\overline{P(k)}$ was incorrectly used in (13) in place of the mean service time $\overline{S_T(k)}$ [14]–[17].

III. PREDICTION OF EFFECTIVE PROCESS TIME

Definitions in Section II provide the practical means of actually measuring the various terms. This section formulates statistical models in predicting the dynamic behaviors of EPT subject to changing conditions. The result is in contrast to the static analysis as tendered in [1]. Simulation studies are carried out to verify these dynamic predictions.

A logical choice of observation period adopted is the basic cycle consisting of one duration of the mean time between failure f_T and one duration of mean time to repair r_T.

A. Dynamic Analysis of Statistical Model

Let n be the total number of lots arrived during this period, $(f_T + r_T)$. Server availability is (assuming $A_{opereator} = 1$)

$$A = \frac{f_T}{f_T + r_T}. \qquad (14a)$$

With arrival interval t_A, the total number of lots arrived is

$$n = \frac{f_T + r_T}{t_A}. \qquad (14b)$$

The expected number of lots capable of being processed is

$$m = \frac{f_T + r_T}{\bar{T}/A} = \frac{f_T}{\bar{T}}. \tag{14c}$$

Subsequently, system utilization is at a level of

$$u = \frac{n}{m} = \frac{\bar{T}/A}{t_A} \tag{14d}$$

and the total number of lots arrived is

$$n = \frac{f_T + r_T}{\bar{T}} \cdot Au = \frac{f_T u}{\bar{T}}. \tag{14e}$$

The total effective process time for the n lots is

$$\sum_{k=1}^{n} E(k) = \sum_{k=1}^{n} T(k) + \sum_{k=1}^{n} \pi(k). \tag{15a}$$

If interest is in the study of consequences due to the single event of machine failure and repair, then all Type-II events sum to

$$\sum_{k=1}^{n} \pi(k) = \sum_{k=1}^{n} [d(k) + b(k)] = d(n) + b(n) = \pi(n). \tag{15b}$$

From Fig. 1, for lot-n, machine failure could have interrupted:

1) after lot claims capacity of tool, with or without queuing WIP, that is, after $A_1(k)$ or $D(k-1)$ for $A_2(k)$;
2) before claiming capacity and in the absence of queuing WIP, that is, during idling time $f(n)$.

Let w be the probability of existence of queueing WIP when lot-$(n-1)$ departs and p be the probability that failure interrupts after lot arrival in the absence of queueing WIP. Then, there are a total of five possible cases to consider as summarized in Fig. 2.

$w = w(n)$	Probability that WIP exists when lot-$(n-1)$ departs $= P[g(n) \geqslant 0]$ from Fig. 1.
$p = p(n)$	Probability that failure interrupts after lot-n claims capacity in the absence of queueing WIP.
$q = q(n)$	$P[r_T \leqslant f(n)] =$ probability that repair time is less than idling time in the absence of WIP for lot-n.
$S = s(n)$	$= (1 - r_T/f) =$ probability that failure interrupts within the interval $(f - r_T)$ after lot-$n-1$) departs for$q = 1$.
n_1, n_2, n_{21}, n_{22}	Probable substates for lot-n in the absence of WIP leading to cases of possible outcomes.
A, B, C, D, E	Probable cases of outcomes with different distributions of Type-II event $\pi(n)$.

It is easily seen that $\pi(n) = r_T$ for cases {A,B} and $\pi(n) = 0$ for case *C*. For case *D*, if failure interrupts within the time interval r_T before lot-n arrives to claim tool capacity, then the remaining idling time cannot fully recover the tool from breakdown and a portion of the repair time shall be counted as a Type-II event before lot n starts processing. The expected duration is

$$\pi(n) = \mathrm{E}[d_1(n)] = \int_0^1 r_T \cdot x dx = \frac{r_T}{2}. \qquad (15c)$$

For case *E* where the repair duration r_T is longer than all the available idling time $f(n)$, then there exists a distribution of Type-II events as $d_1(n)$, ranging from the minimum duration $[r_T - f(n)]$ to the maximum of r_T, as

$$P_{d_1(n)}(x) = r_T - (1-x)f. \quad x \in [0,1]. \qquad (15d)$$

With $b(n) = 0$, the overall expected Type-II disturbance is

$$\begin{aligned} \pi(n) = \mathrm{E}[d_1(n)] &= \frac{1}{\int_0^1 dx} \cdot \int_0^1 [r_T - (1-x)f] dx \\ &= r_T - \frac{1}{2} f(n). \end{aligned} \qquad (15e)$$

Therefore, the expected impact of Type-II event is

$$\mathrm{E}[\pi(n)] = \begin{cases} w \cdot r_T + (1-w) \cdot p \cdot r_T \\ +(1-w) \cdot (1-p) \cdot q \cdot [s \cdot 0 + (1-s) \cdot \frac{1}{2} r_T] \\ +(1-w) \cdot (1-p) \cdot (1-q) \cdot (r_T - \frac{1}{2} f). \end{cases} \qquad (15f)$$

Let $\lambda = r_T/f$, and from (15a) the expected EPT

$$\overline{E(k)} = \frac{1}{n}\sum_{k=1}^{n} E(k) = \overline{T(k)} + \frac{1}{n}\cdot \mathrm{E}[\pi(n)] \qquad (16a)$$

may be normalized as

$$\frac{\overline{E(k)}}{\overline{T(k)}} = \begin{cases} 1 + \frac{[1-(1-w)\cdot(1-p)]}{u} \cdot \frac{r_T}{f_T} \\ + \frac{(1-w)\cdot(1-p)\cdot[q\cdot\frac{1}{2}\cdot\lambda^2 + (1-q)\cdot(\lambda-\frac{1}{2})]}{u} \cdot \frac{f}{f_T}. \end{cases} \qquad (16b)$$

This is the general expression for the predicted EPT, depending on the level of utilization and probabilities of queueing WIP, idling times, and ratio of repair to idling times. The determination of effective process time now translates into the determination of probabilities $\{w, p, q\}$.

B. Upper Bound of Effective Process Time at High Utilization

For a fully loaded system with utilization $u = 1$ and infinite WIP $(w = 1)$, (16b) yields

$$\frac{\overline{E(k)}}{\overline{T(k)}}[u = 1, w = 1] = 1 + \frac{r_T}{f_T} = \frac{1}{A} \qquad (17a)$$

which identically regenerates the expected EPT as predicted in static analysis [1, p. 266], recalling that in the absence of Type-I events, $\bar{T} = \bar{P}$ as from (7).

In Fig. 3 is shown the curve for the case of

$$\frac{\overline{E(k)}}{\overline{T(k)}}[w = 1] = 1 + \frac{1}{u}\cdot\frac{r_T}{f_T} \qquad (17b)$$

as compared to results of a simulation study. Details are covered in Section IV. The match is excellent for high utilization when the probability of existing WIP is high $(w \to 1)$. As utilization drops, so does the deviation of from unity and (17b) gives an upper bound for EPT at high utilization.

C. Estimates of Effective Process Times

For utilization levels below 0.5, various estimates of EPT have been obtained and shown in Fig. 4 (See Appendix.) Different probabilities $\{w, p, q\}$ lead to different estimates of EPT. Each has its own range of validity. Prediction of mean service time is indicated by the straight line marked "static $u = 1$."

D. Concluding Remarks

Several important conclusions can now be established. First, the notion of service time is a statistical metric defined in terms of some expected means of static nature, independent of utilization levels [see (4f)]. On the other hand, EPT is an explicit function

of utilization and henceforth is a dynamic quantity depending on the operational behaviors of the single-lot non-tandem server. These two different metrics are therefore not equivalent, much to the contrast of conventional wisdom. They describe the same operational behavior *if and only if* the server is fully loaded. This follows naturally from (17a) and (4f) which link up mean service time and EPT explicitly as

$$\overline{E(k)}(u=1,w=1)=\frac{\overline{T(k)}}{A}=\bar{S}_T. \qquad (18)$$

Subsequently, based on the above observations for single-lot nontandem servers, which is understandably the simplest, it is *argued* that *all* servers, be they tool groups composed of single-lot nontandem servers in series or in parallel formation, or be they servers operating in tandem, the same characteristics of static mean service time and dynamic EPT are universally inherited and they become equivalent if and only if under full load conditions. This summary is restated as follows.

Conclusion I: The notions of mean service time and effective process time are equivalent if and only if the server is under full load condition with the existence of queueing WIP.

Secondly, for each lot-k, its EPT as defined in (5b) can be precisely clocked; therefore, each $E(k)$ is measurable and its variance and subsequently, its variability, can all be measured. On the other hand, individual service time is not measured (or even measurable as explained in Section II-A).

However, as clearly seen from the analyses, prediction of EPT is not warranted without the assumption of certain distributions about the probabilities of queueing WIP and failure interruptions. This necessarily implies that for any practical measurements over finite length of observation periods, the EPT thus obtained naturally forms a probabilistic distribution on its own, rather than a deterministic figure of static merit worthy of representing service times.

Conclusion II: Service time of individual lots cannot be measured directly or represented by effective process time under dynamic conditions.

IV. SIMULATION RESULTS

Simulation studies were conducted using AutoSched AP[1] from Brooks Automation, Inc. A single nontandem server with a constant TPT is assumed. Intervals of lot release, failures, and repairs are all exponentially distributed. Each of the 20 experiments consists of ten runs of 30 000 lots. The various parameters and results are summarized in Table I.

Arrival interval increases from 1.25 to 70 time units (for $\searrow = 1, t_A = 13, u_3 = 0.096$; for $n = 1, t_A = 60, u_2 = 0.021$). The EPT at $u \equiv 1$ gives the mean service time from static analysis (17a). E_{sim} denotes the EPT measured from the simulation runs.

[1] Registered trademark.

The gap between mean service time and E_{sim} is listed. $E(p = Au)$ predicts EPT using (A4b) or (A4c). The error between E_{sim} and $E(p = Au)$is also computed. The curves are shown in Fig. 4.

When arrival interval increases, so that utilization level drops off from the full load condition, the difference between mean service time and E_{sim} increases, thus providing the concrete evidence that the notions of mean service time and EPT are not equivalent. They are the same only under the conditions of (18).

From Fig. 3, the approximation of (A4b) for $u_3 \leq u \leq 1$ gives sufficiently good estimates of the simulation results to within 1% error when utilization is above 50%.

Service times for the 20 instances in Table I are 1.25. EPT ranges from 1.255 to 3.220. The gap increases from 0.4% to 61.2% as utilization drops to 1.8%.

V. CONCLUSION

This paper establishes the fact that notions of service time and effective process time are not equivalent and the former cannot be measured from the latter. It brings the investigation back to how to define service time for individual lots, apart from its statistical means. The importance of this lies in a practical need to gauge performance of production lines using the notion of variability (12), which is defined by the variance of service times for individual lots.

In this paper, we have clarified definitions of various terms employed in the management of production lines, namely, service time, theoretical processing time, effective process time, raw process time, queueing times, and normalized cycle time. Takt time is borrowed from lean manufacturing to thread through their explanations to enable practical measurements of these terms on the shop floor.

For the very first time, we explicitly quantify the inherent relationships among these terms based on the classification of two types of disturbance events: one is WIP-dependent and the other WIP-independent. Explicit expressions among these quantities explain the underlying natures of determination and indetermination of service times, originating from three sources. First is the structural lack of definition of service time for individual lots, second is the nonequivalence between the notions of service time and effective process time, and the third is the probabilistic dependence of effective process time on the distribution profiles of queueing WIP and failure interruptions under dynamic loading conditions.

Indetermination of service times leads to the same of its SCV. Subsequently, variability of production lines or factories cannot be measured directly as conventionally anticipated. Structural changes are needed to resolve the lack of specific definitions. In addition, the implicit assumption (that breakdowns only occur when machine is active) still needs to be resolved.

We give predictions of effective process times, incorporating probabilities of existence of queueing WIP and failure interruptions, compounded with the distributions of the lengths of repair times, all as a function of utilization levels. From both theoretical predictions and simulation results, we demonstrate that mean service time is static and effective process time dynamic. They are equivalent if and only if under the full load condition with existence of queueing WIP.

Lastly, although the apparent ambiguity of service time is now explained when down time exists, what has not been done is to point out an effective remedy. We are currently exploring various approximations for queues with breakdowns [18], [19] that are expected to help to resolve this issue.

APPENDIX

This Appendix summarizes estimates of effective process times derivable from the general expression (16b).

A. Estimates of Effective Process Time

From (16b), EPT in absence of queueing WIP ($w \to 0$) is

$$\overline{\frac{E(k)}{T(k)}}(0,p,q) = \begin{cases} 1 + \frac{p}{u} \cdot \frac{r_T}{f_T} \\ + \frac{(1-p)\cdot[q\cdot\frac{1}{2}\cdot\lambda^2 + (1-q)\cdot(\lambda-\frac{1}{2})]}{u} \cdot \frac{f}{f_T} \end{cases} \cdot \tag{A1a}$$

Depending on the exclusive combinations of whether repair time is less than idling time, two cases exist. For utilization levels ($r_t \geq f: q \to 0 \; and \; \lambda \geq 1$)

$$\overline{\frac{E(k)}{T(k)}}(0,p,0) = 1 + \frac{p}{u} \cdot \frac{r_T}{f_T} + \frac{(1-p)\cdot(\lambda-\frac{1}{2})}{u} \cdot \frac{f}{f_T}. \tag{A1b}$$

For utilization levels ($r_t < f: q \to 1 \; and \; \lambda < 1$)

$$\overline{\frac{E(k)}{T(k)}}(0,p,1) = 1 + \frac{p}{u} \cdot \frac{r_T}{f_T} + \frac{(1-p)\cdot\lambda^2}{2u} \cdot \frac{f}{f_T}. \tag{A1c}$$

Determination of EPT now depends on the probability of machine failure interrupts after lot-n claims tool capacity.

From Fig. 1, when utilization is low enough that there exists no queueing WIP but idling time for each lot, then

$$f_T + r_T = \sum_{k=1}^{n} [f(k) + E(k)]$$
$$= n[\overline{f(k)} + \overline{E(k)}] \quad \text{(A2a)}$$

From (14e), the expected idling time for each lot is

$$\overline{f(k)} = \frac{f_T + r_T}{n} - \overline{E(k)} = \frac{\overline{T(k)}}{Au} - \overline{E(k)}$$
$$= \overline{T(k)} \cdot \left(\frac{1}{Au} - \nu\right) \quad \text{(A2b)}$$

where $\nu = \overline{E(k)}/\overline{T(k)}$. The expected probability that failure interrupts after claiming capacity is

$$p = \overline{p(k)} = \frac{\overline{E(k)}}{\overline{f(k)} + \overline{E(k)}} = \frac{\overline{E(k)}}{\overline{T(k)}} \cdot Au$$
$$= \nu \cdot Au \quad \text{(A2c)}$$

and the idling time is related to this probability as

$$f = \overline{f(k)} = \overline{E(k)} \cdot \left(\frac{1-p}{p}\right)$$
$$= \nu \cdot \overline{T(k)} \cdot \left(\frac{1-p}{p}\right). \quad \text{(A2d)}$$

B. Exact Solution

For utilization levels within the range

$$1 \geq u \geq u(\lambda = 1) = u_0$$
$$= \frac{2A \cdot \overline{T(k)}}{2A \cdot \overline{T(k)} + (1 - A^2) \cdot f_T} \quad \text{(A3a)}$$

it is easily shown that EPT from (A1b) is given by

$$\frac{\overline{E(k)}}{\overline{T(k)}}(u) = \frac{1}{Au} - \frac{f_T}{A \cdot \overline{T(k)}} \cdot$$
$$\times \left[1 - \sqrt{1 - \frac{2A \cdot \overline{T(k)}}{f_T}\left(\frac{1}{u} - 1\right)}\right]. \quad \text{(A3b)}$$

For utilization levels within the range

$$u_0 \geq u \geq u(n = 1) = u_2 = \frac{\overline{T(k)}}{f_T} \quad \text{(A3c)}$$

EPT from (A1c) is given by

$$\frac{\overline{E(k)}}{\overline{T(k)}}(0,p,1\ \lambda \le 1) = \frac{1}{A} + \frac{1}{2} \cdot \frac{r_T}{\overline{T(k)}} \cdot \frac{r_T}{f_T}. \qquad \text{(A3d)}$$

C. Approximate Solution

If $v = 1$ is approximated in (A2c) and (A2d), then for

$$1 \ge u \ge u_3 = u(\lambda = 1) = \frac{1}{A} \cdot \frac{\overline{T(k)}}{r_T + \overline{T(k)}} \qquad \text{(A4a)}$$

EPT from (A1b) is readily shown to be

$$\frac{\overline{E(k)}}{\overline{T(k)}}(u) = 1 + \frac{1}{\overline{T(k)}} \cdot \frac{2 \cdot r_T \cdot t_A - (t_A - \overline{T(k)})^2}{2 \cdot (f_T + r_T)}. \qquad \text{(A4b)}$$

For $\lambda < 1$ and $u_2 < u < u_3$, EPT from (A1c) becomes

$$\frac{\overline{E(k)}}{\overline{T(k)}}(0,p,1|\lambda \le 1) = 1 + A \cdot \frac{r_T}{f_T} + \frac{A}{2} \cdot \frac{r_T}{\overline{T(k)}} \cdot \frac{r_T}{f_T}. \qquad \text{(A4c)}$$

ACKNOWLEDGMENT

The first author wishes to thank Y.-H. Chiu for writing the simulation program to verify the derived results and W. Welker and Dr. K. Horninger of Infineon Technologies for their encouragement. The authors would also like to thank Prof. L. McGinnis and B. Zwart as well as the reviewers for their invaluable comments.

REFERENCES

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

[2] L. Sattler, "Using queuing curve approximations in a fab to determine productivity improvements," in *Proc. 1996 IEEE/SEMI ASMC Conf.*, Nov. 1996, pp. 140–145.

[3] J. H. Jacobs, L. F. P. Etman, E. J. J. van Campen, and J. E. Rooda, "Characterization of operational time variability using effective process time," *IEEE Trans. Semiconduct. Manuf.* , vol. 16, no. 3, pp. 511–520, Aug. 2003.

[4] L. Kleinrock, "Communication nets," in *Stochastic Message Flow and Delay*. New York: McGraw-Hill, 1964.

[5] L. Kleinrock, "On queueing problems in random-access communications," *IEEE Trans. Inform. Theory*, vol. 31, no. 2, pp. 166–175, Mar. 1985.

[6] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[7] SEMI E79, Equipment Automation/Hardware Volume, Book of SEMI Standards 1999, Mountain View, CA: SEMI.

[8] [Online]. Available: http://www.isixsigma.com/dictionary/takt_time455.htm

[9] [Online]. Available: http://www.strategosinc.com/takt_time.htm

[10] "Manufacturing Management Technology Institute," *Takt Times* Aug. 1999, Tech. Bull. v.1.

[11] L. Labanowski, "Improving overall fabricator performance using the continuous improvement methodology," in *Proc. 1997 IEEE/SEMI ASMC Conf.*, 1997, pp. 405–409.

[12] K. Wu, "An examination of variability and its basic properties for a factory," *IEEE Trans. Semiconduct. Manuf.*, vol. 18, no. 1, pp. 214–221, Feb. 2005.

[13] L. Kleinrock, *Queuing Systems Volume I: Theory*. New York: Wiley, 1975.

[14] D. P. Martin, "Key factors in designing a manufacturing line to maximize tool utilization and minimize turnaround time," in *Proc. IEEE/ SEMI Int. Semiconductor Manufacturing Science Symp.*, Jul. 1993, pp. 48–53.

[15] D. P. Martin, "How the law of unanticipated consequences can nullify the theory of constraints," in *Proc. IEEE/SEMI ASMC Conf.*, 1997, pp. 380–385.

[16] M. Kishimoto, K. Ozawa, K. Watanabe, and D. Martin, "Optimized operations by extended X-factor theory including unit hours concept," *IEEE Trans. Semiconduct. Manuf.* , vol. 14, no. 3, pp. 187–195, Aug. 2001.

[17] D. R. Delp, “A new X-factor contribution measure for identifying machine level capacity constraints and variability,” in *Proc. IEEE/SEMI ASMC Conf.*, 2004, pp. 334–338.

[18] P. P. Wang, “Queue length distribution of an unreliable machine,” *Opsearch*, vol. 37, no. 2, pp. 99–123, Jun. 2000.

[19] I. Adan and J. Resing, Queueing theory 2001, Lecture Notes. Feb. 14.

Kan Wu received the M.S. degree in industrial engineering and operations research and the M.E. degree in nuclear engineering from the University of California, Berkeley, in 1996. Currently, he is working toward the Ph.D. degree at Georgia Institute of Technology, Atlanta.

He was an Engineer with Tefen, Ltd., and Taiwan Semiconductor Manufacturing Company. From 2003 to 2005, he was an IE Manager at Inotera Memories, Inc. His research interests include production planning, scheduling, and dispatching in the semiconductor industry.

Kheng Hui received the Ph.D. degree in control engineering from the University of Hong Kong in 1995.

Prior to that, he was a Building Services Engineer, installing and designing large-scale HVAC and E&M systems. He earned his professional qualifications as a Chartered Engineer of the Engineering Council, U.K., and was a corporate member of the Chartered Institution of Building Services Engineers, the Institution of Mechanical Engineers, and the Hong Kong Institution of Engineers. He joined the semiconductor industry as a Consultant Engineer on advanced process control and engineering data analyses.

Dr. Hui is a Reviewer for the IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING and other control journals.

List of Tables

List of Figures

	t_A	f_T	r_T	$\overline{T}$	u	EPT		gap	EPT	error
						u=1	E_{sim}		p=Au	
1	1.25	48	12	1	1.000	1.25	1.255	0.4%	1.249	-0.5%
2	1.30	48	12	1	0.962	1.25	1.261	0.9%	1.259	-0.2%
3	1.40	48	12	1	0.893	1.25	1.277	2.1%	1.279	0.2%
4	1.50	48	12	1	0.833	1.25	1.293	3.3%	1.298	0.4%
5	1.70	48	12	1	0.735	1.25	1.324	5.6%	1.336	0.9%
6	2.00	48	12	1	0.625	1.25	1.383	9.6%	1.392	0.6%
7	3.00	48	12	1	0.417	1.25	1.533	18.5%	1.567	2.2%
8	4.00	48	12	1	0.313	1.25	1.653	24.4%	1.725	4.4%
9	4.50	48	12	1	0.278	1.25	1.709	26.8%	1.798	5.2%
10	5.00	48	12	1	0.250	1.25	1.764	29.1%	1.867	5.8%
11	6.00	48	12	1	0.208	1.25	1.876	33.4%	1.992	6.2%
12	8.00	48	12	1	0.156	1.25	2.046	38.9%	2.192	7.1%
13	10.0	48	12	1	0.125	1.25	2.199	43.1%	2.325	5.8%
14	11.5	48	12	1	0.109	1.25	2.285	45.3%	2.381	4.2%
15	13.0	48	12	1	0.096	1.25	2.358	47.0%	2.400	1.8%
16	15.0	48	12	1	0.083	1.25	2.455	49.1%	2.400	-2.2%
17	20.0	48	12	1	0.063	1.25	2.641	52.7%	2.400	-9.1%
18	40.0	48	12	1	0.031	1.25	3.038	58.9%	2.400	-21.0%
19	60.0	48	12	1	0.021	1.25	3.216	61.1%	2.400	-25.4%
20	70.0	48	12	1	0.018	1.25	3.220	61.2%		

Table 1

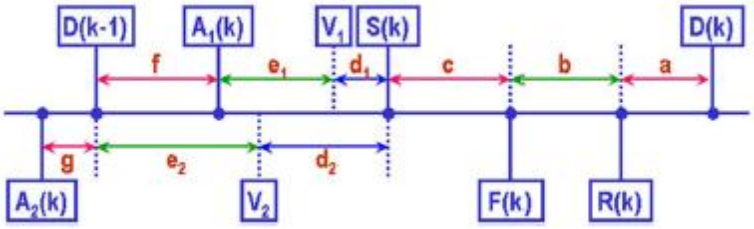

Fig. 1

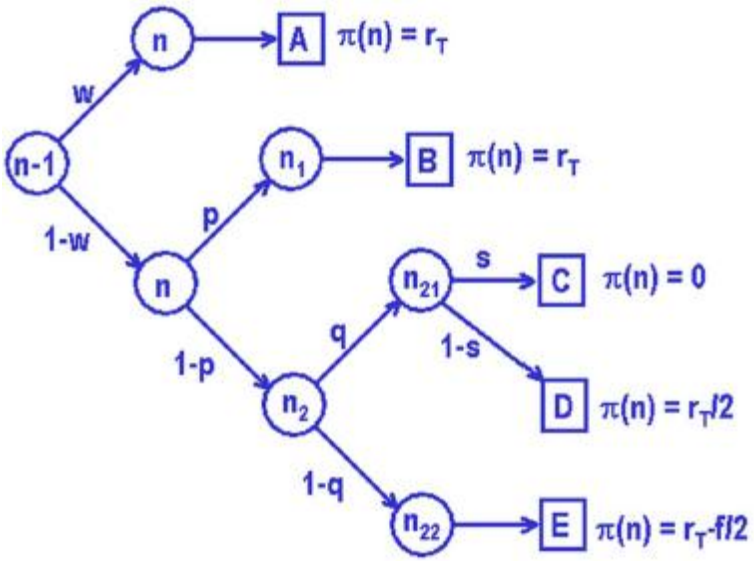
n
A
π(n) = r_T
w
n-1
n_1
B
π(n) = r_T
p
1-w
n
n_21
s
C
π(n) = 0
q
1-s
1-p
n_2
D
π(n) = r_T/2
1-q
n_22
E
π(n) = r_T-f/2

Fig. 2

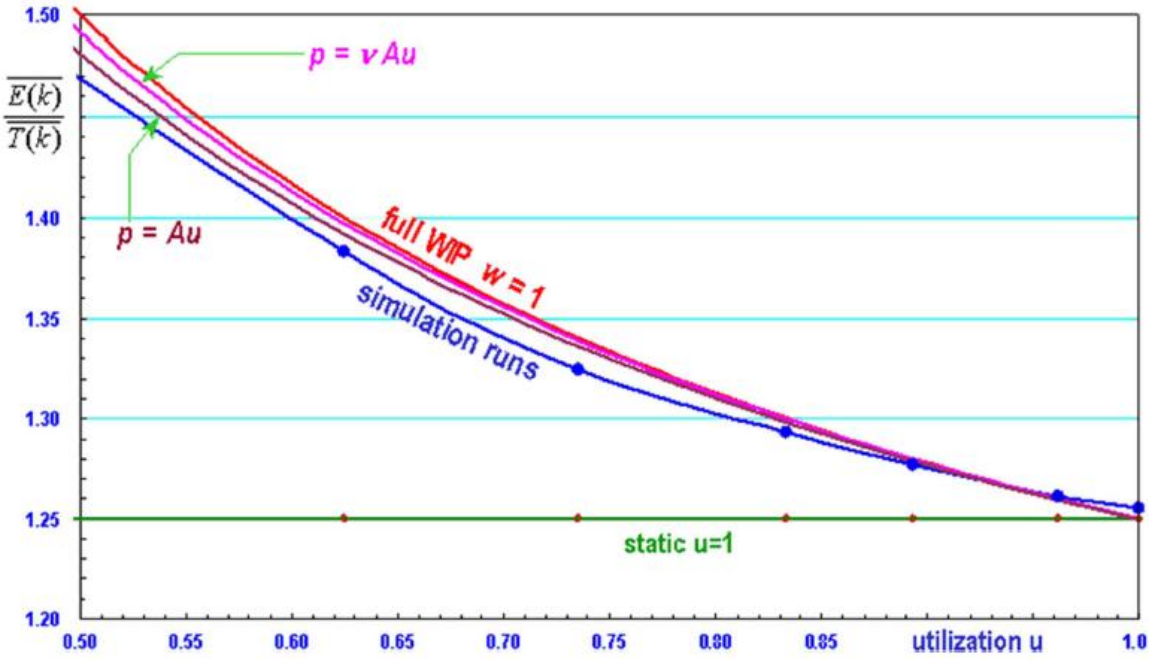
1.50
1.40
1.35
1.30
1.25
1.20
$\frac{\overline{E(k)}}{\overline{T(k)}}$
p = v Au
p = Au
full WIP w = 1
simulation runs
static u=1
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
utilization u
1.0

Fig. 3

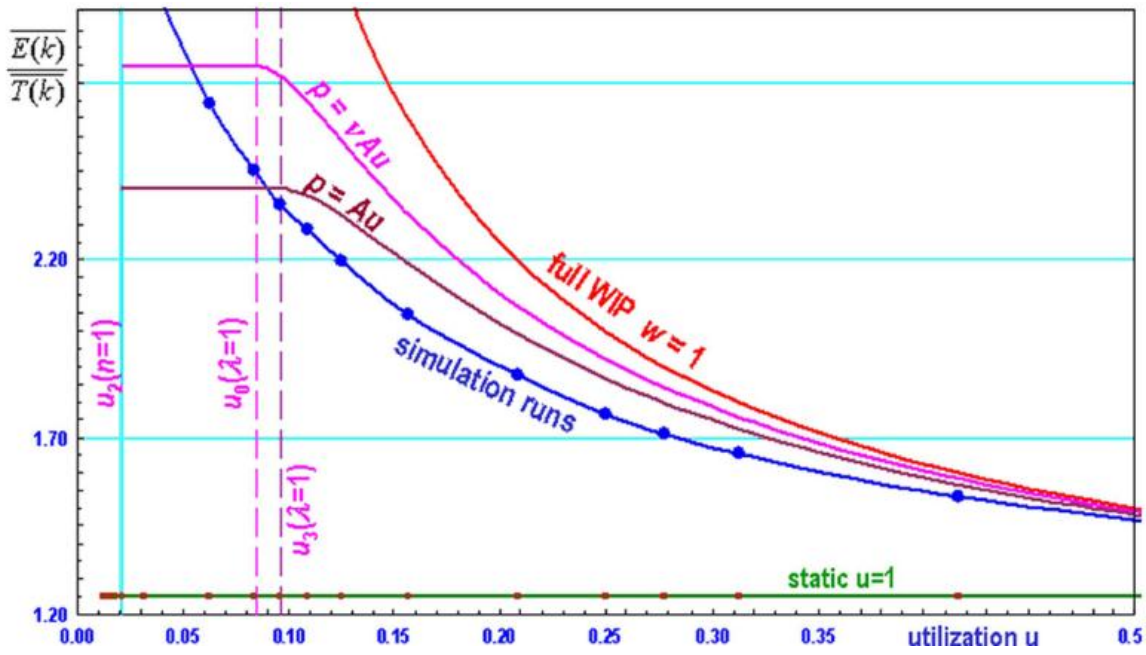
E(k)
T(k)
2.20
1.70
1.20
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
utilization u
0.5
P = vAu
P = Au
full WIP w = 1
simulation runs
static u=1
u2(n=1)
u0(λ=1)
u3(λ=1)

Fig. 4