Title: **Minimization of Adjustment Frequency to Achieve Low-Cost Fabrication in Product-Mix VLSI Manufacturing**

Authors: **Kazuyuki Saito and Mitsuru Nagashima**

Affiliation: The University of Aizu,
Tsuruga, Ikkimachi Aizu-Wakamatsu, Fukushima 965-8580 Japan

Tel: 81-242-37-2658, Fax: 81-242-37-2597, e-mail: k-saito@u-aizu.ac.jp

Abstract:
In recent years, there has been strong demand for a product-mix in semiconductor manufacturing to meet various customer' demands. This paper focuses on how to allocate work-in-process (WIP) to a machine while maintaining the required response time. A product-mix requires an adjustment (or setup) of the machine, and this leads to increased production costs due to the additional required human resources and reduced throughput. First, a periodical WIP allocation algorithm is introduced. It estimates the operator request rate and machine utilization rate when two types of WIP with different arrival rates are processed in a single machine. Next, this paper discusses how to make a set of WIP that minimizes the adjustment frequency when we have more than two types of WIP in a processing station. Traffic intensities of each WIP are key parameters for making a WIP set. The adjustment frequency can be minimized when the sum of the traffic intensities is maximized. A resource estimation system considering the adjustment of the machine was evaluated in a real facility, and the system was confirmed to be applicable in weekly resource planning.

Key Words:
VLSI manufacturing, resource planning, adjustment frequency, setup, product-mix, machine utilization, operator request rate.

Minimization of Adjustment Frequency to Achieve Low-Cost Fabrication in Product-Mix VLSI Manufacturing

Kazuyuki Saito and Mitsuru Nagashima
The University of Aizu
Tsugura, Ikkimachi, Aizu-Wakamatsu, Fukushima 965-8580 Japan

1. Introduction

VLSI manufacturers must meet various customer' demands, and being able to make a mixture of products is as important in factory management as the micro-fabrication technologies for producing highly integrated LSIs [1-2]. With a product-mix, it is necessary to adjust (or setup) the machine when the product changes. This adjustment involves two major problems. One is that it requires additional resources, which are usually human resources because the adjustment is too complicated to be done by a robot; these resources increase the production cost. The other is that the machine cannot be used during the adjustment period, reduces the throughput of products, and also increases the production cost. We are studying resource planning and factory management for product-mixes. Recently, we have proposed an algorithm for periodical allocation of machine operators when two kinds of work-in-process or work-in-progress (WIP) share a single machine [3-4]. But usually, we have various kinds of WIP to be processed at the same time by multiple machines. For example, in wire bonding in VLSI assembly, more than 30 kinds of WIP are processed in nearly 100 machines. We must determine sets of WIP to be processed in the same machine. The purpose of this research is to find an algorithm for making sets of WIP that minimizes the adjustment frequency of product-mix adjustment.

2. Periodical allocation of two kinds of WIP

2.1. Input inventory control method to maintain a required response time

Now, we consider multiple types of WIP being processed at a processing station, which consists of several machines. A machine shares multiple types of WIP and achieves high utilization. When the arrival rate of each type of WIP is very low, it is very natural to store the WIP in an input inventory buffer and process each type when some amount of it has accumulated. The average amount of each WIP in the buffer is related to the time in the buffer by Little's equation [5]:

$$\overline{N} = \lambda \overline{T} . \qquad (1)$$

The average number of WIPs in the buffer $\overline{N}$ is given by the product of the arrival rate of WIP λ and the average time in the buffer $\overline{T}$. Although we could control the average number of WIPs by monitoring the average time in the buffer, it is very difficult to know the average number of WIPs because it changes dynamically. We analyze the dynamics of WIP in the buffer as shown in Fig. 1.

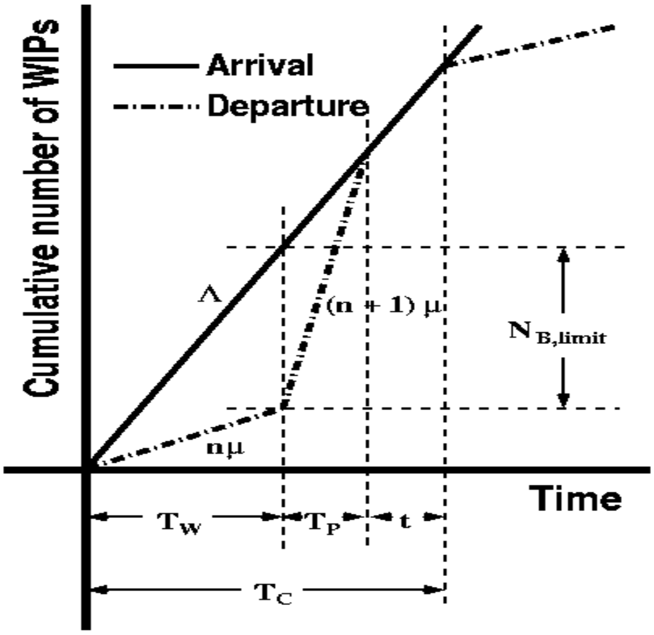

Figure 1. Arrival and departure at a processing station. During T_W, the number of machines in the processing station is *n*, and the number of WIPs in the buffer accumulates.

The WIP arrival rate Λ is assumed to be $n\mu < \Lambda < (n+1)\mu$, where μ is the average service rate of the machines. WIP in the buffer increases when the number of machines is *n*. When this number reaches $N_{B,lim}$, a new machine is allocated for the type of WIP and the number of WIPs in the buffer decreases. Instead of monitoring the average number of WIPs, we can control the average time in the buffer easily by monitoring the maximum number of WIPs. The average time in the buffer can be given by:

$$\overline{W_B} = \frac{N_{B,lim}}{\lambda}. \qquad (2)$$

To keep the average time spent in the buffer below $\overline{W_B}$, the number of WIPs in the buffer must be less than $N_{B,\lim}$.

2.2. Periodical assignment of WIP

Let us consider two types of WIP sharing a single machine. Figure 2 shows the periodical assignment of two types of WIP to a single machine.

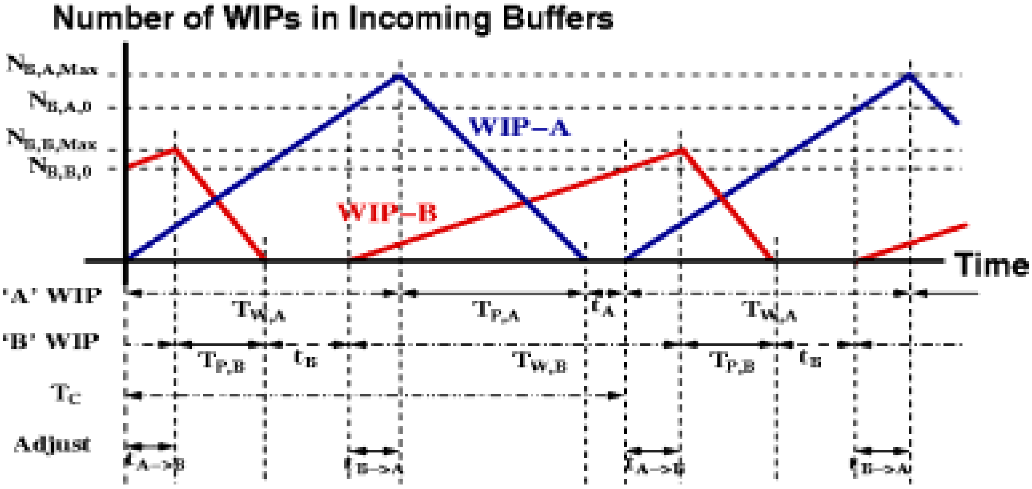

Figure 2. Periodical assignment of WIP. The required response time at the processing station is kept when $N_{B,\max}$ is less than $N_{B,\lim}$.

The dynamics of WIP-A is explained. When the number of WIP-As reaches $N_{B,A,0}$, we start to adjust the machine for WIP-A. Then WIP-B starts to accumulate in the input buffer. When the adjustment has finished, the number of WIP-As reaches $N_{B,A,\max}$ and starts to decrease. If we can keep $N_{B,A,\max}$ less than $N_{B,\lim}$, which was defined in the previous section, we can keep the average time in the buffer bellow its maximum allowable value. Figure 2 gives us the following equations to be solved.

$$\frac{1-\rho_A}{\rho_A} T_{P,A} = T_{P,B} + T_0 + t_B \,, \qquad (3)$$

$$\frac{1-\rho_B}{\rho_B} T_{P,B} = T_{P,A} + T_0 + t_A \,, \qquad (4)$$

$$T_{P,A} \le \frac{N_{B,A,\lim}}{\mu_A(1-\rho_A)}, \tag{5}$$

$$T_{P,B} \le \frac{N_{B,B,\lim}}{\mu_B(1-\rho_B)}, \tag{6}$$

and

$$T_0 = t_{A->B} + t_{B->A}. \tag{7}$$

Here, $T_{P,A}$ and $T_{P,B}$ are the processing times of the two kinds of WIPs; ρ_A and ρ_B are their traffic intensities ($\rho = \lambda / \mu$); μ is the service rate; T_0 is the sum of the adjustment times; and t_A and t_B are idle times after processing the WIPs. The number of WIPs must be kept less than $N_{B.A,\lim}$ or $N_{B,B,\lim}$ for each WIP.

We introduced new parameters δ and γ.

$$\delta = t_A - t_B. \tag{8}$$

$$\delta + \gamma = t_A + t_B. \tag{9}$$

As shown in Fig. 3, $T_{P,A}$ and $T_{P,B}$ can be solved. If we define a parameter $K(\delta,\gamma)$ as

$$K(\delta,\gamma) = T_{P,A} + T_{P,B}, \tag{10}$$

then the cycle time of processing for the WIP is given by

$$T_C(\delta,\gamma) = K(\delta,\gamma) + T_0 + \delta + \gamma. \tag{11}$$

Two adjustments are required during T_C, so the adjustment frequency is given by

$$R = \frac{2}{T_C(\delta,\gamma)}. \tag{12}$$

And the machine utilization rate is given by

$$U = \frac{K(\delta,\gamma)}{T_C(\delta,\gamma)}. \tag{13}$$

The adjustment frequency and the machine utilization can be optimized using δ and γ.

$T_{W,A}$ in Fig. 2 is given by

$$T_{W,A} = \frac{1-\rho_A}{\rho_A} T_{P,A}, \tag{14}$$

$$N_{B,A,0} = \lambda_A (T_{W,A} - t_{B->A}). \tag{15}$$

When the number of WIPs in the buffer reaches this value, we have to start an adjustment of the machine.

More generally, when we consider *k* types of WIP, we can formulate the allocation scheme as follows:

In place of (3) and (4),

$$\left(\frac{1-\rho_j}{\rho_j}\right) T_{P,j} \geq \sum_{i=1}^{k} T_{P,i} + T_0. \qquad (j \neq i) \tag{16}$$

In place of (5) and (6),

$$T_{P,i} \leq \frac{N_{B.i.\lim}}{\mu_i (1-\rho_i)}. \tag{17}$$

And in place of (7),

$$T_0 = \sum_{i=1}^{k-1} t_{i->i+1} + t_{k->1}. \tag{18}$$

Equations (16) and (17) are *k*-dimensional inequalities and generally cannot be solved.

3. Combination of WIP to achieve high machine utilization and low adjustment frequency

3.1. Traffic intensities of two types of WIP and machine utilization

When we consider multiple types of WIP in the processing station, we have to determine a set of WIPs to be processed in a particular machine. First, we discuss how to choose two types of WIP. Figure 3 shows the simulated machine utilization rate when the traffic intensities for the WIP are ρ_A and ρ_B. Here, we assumed that $\mu_A = \mu_B = 1$, $\overline{W_{B,A,\lim}} = \overline{W_{B,B,\lim}} = 120$, $\delta = 0$, and $\gamma = 0$. If the traffic intensity of WIP-B is more than $\rho_{\max}$ (= 0.83 for $T_0 = 20$ in this example), then only a single type of WIP (namely WIP-B) can be assigned to the machine.

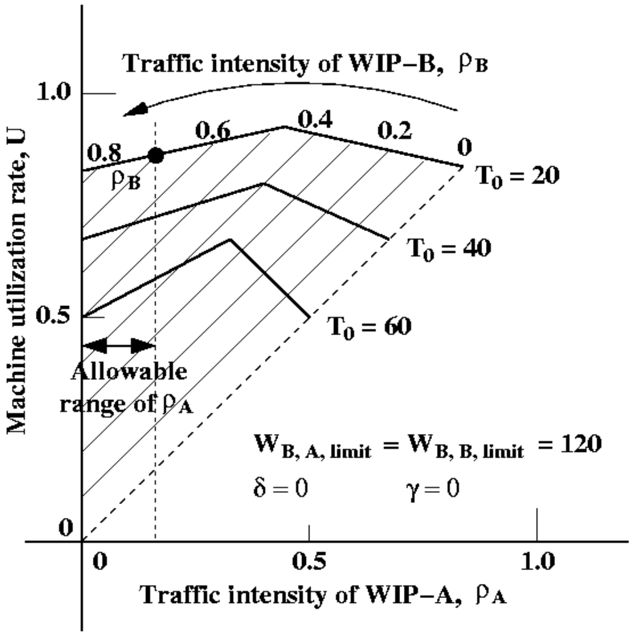

Figure 3. Machine utilization rate for a set of WIPs with different traffic intensities. If the traffic intensity of WIP-B is given as 0.7 for $T_0 = 20$ (black circle), then WIP-A may have traffic intensity up to 0.16.

If the traffic intensity of WIP-B is given as 0.7 (black circle in Fig. 3), then WIP-A may have traffic intensity up to 0.16. It should be noted that the sum of traffic intensities is less than $\rho_{\max}$, i.e., $\rho_A + \rho_B \le \rho_{\max}$, and $\rho_{\max}$ is less than one. When $\rho_A = \rho_B$, the machine utilization can be maximized. The machine utilization rate decreases as the adjustment time T_0 increases.

3.2. Adjustment frequency and machine utilization rate

From equations (11), (12), and (13), we can get the relationship between the adjustment frequency and machine utilization rate U:

$$U = 1 - \frac{T_0 + \delta + \gamma}{2} R . \tag{19}$$

The machine utilization rate is linearly dependent on the adjustment frequency. When the adjustment frequency is minimized, the machine utilization rate can be maximized. From equations (3) and (4), we can get:

$$T_C = \frac{T_0 + t_A(1 - \rho_A) + t_B(1 - \rho_B)}{1 - (\rho_A + \rho_B)} . \tag{20}$$

Here, we recall that $\rho_A + \rho_B \leq 1$, as discussed in 3.1. When $\rho_A + \rho_B$ is large, T_C becomes large, so the adjustment frequency becomes small.

3.3. WIP combination rule

When we have more than two types of WIP at a processing station, we must determine combinations of WIP to be processed in the same machine. We consider 9 types of WIP, as an example. Their traffic intensities are listed in Table 1. We examined several rules for combining two types of WIP, and Table 2 shows typical results of applying them.

Table 1. Traffic intensities ρ of WIP

WIP	A	B	C	D	E	F	G	H	I
ρ	0.605	0.480	0.448	0.433	0.370	0.214	0.214	0.201	0.105

Table 2. Utilization rate and adjustment frequency for different combination rules

Rule	Combination	Average utilization rate	Sum of adjust. frequency
#1	A-F, C-E, B-G, D-H, I	0.569	0.0114
#2	A-F, B-G, C-H, D-I, E	0.579	0.0134
#3	A-I, B-H, C-G, D-F, E	0.576	0.0135
#4	B-I, C-H, D-G, E-F, A	0.573	0.0157

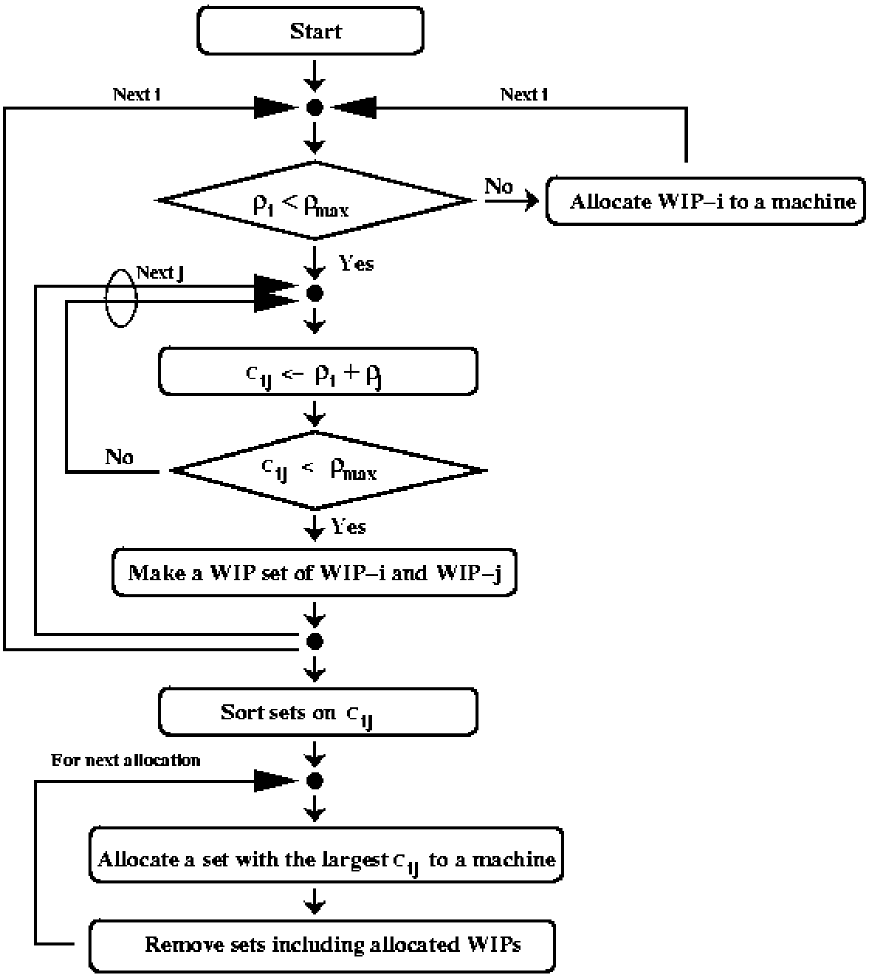
Start
Next i
Next i
$\rho_i < \rho_{max}$
No
Allocate WIP–i to a machine
Yes
Next j
$c_{ij} <- \rho_i + \rho_j$
No
$c_{ij} < \rho_{max}$
Yes
Make a WIP set of WIP–i and WIP–j
Sort sets on c_{ij}
For next allocation
Allocate a set with the largest c_{ij} to a machine
Remove sets including allocated WIPs

Figure 4. A WIP set formation algorithm for rule #1. A sum of the traffic intensities $C_{ij}(=\rho_i+\rho_j)$ of WIP-i and WIP-j is evaluated. A WIP set with the largest C_{ij} is allocated first. If ρ_i is larger than ρ_{max} (=0.83 in an example in Fig. 3), a machine will be used for a single type of WIP.

The sum of the adjustment frequencies at the processing station differs among the combination rules, although the average utilization rate differs by only 2% between the maximum and minimum rates. As discussed in 3.2, the adjustment frequency can be small when $\rho_A + \rho_B$ is large. The best allocation algorithm is given in Figure 4. A traffic intensity of each WIP is evaluated. If ρ_i is larger than ρ_{max}, a machine will be used for a single type of WIP. A sum of the traffic intensities $C_{ij} (= \rho_i + \rho_j)$ of WIP-i and WIP-j is evaluated. A WIP set with the largest C_{ij} is allocated first. The combination of WIP type is done to maximize $\rho_A + \rho_B$. Rule #1 in Table 2 is based on this idea.

Figure 5 is the allocation results based on rule #1. The human resources for the adjustment can be estimated using a queuing model. This estimation system has been applied in a real VLSI assembly facility, which processes 30 or more types of LSIs and produces a total of more than 5 million LSIs per month. The resource estimations have accuracies of 95 - 102% for machinery resources and 97 – 101% for human resources. The estimation was reduced from 2 - 3 days to less than a few hours for the half-year production resource plan. Thus, this system has a short enough estimation time to be applicable to weekly planning.

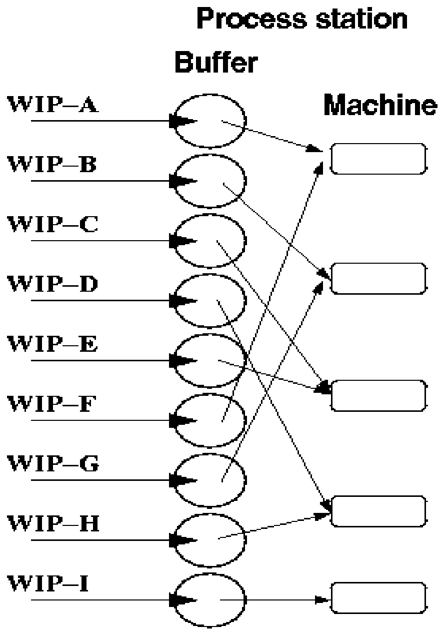

Figure 5. Combination of WIP types when their traffic intensities are given in Table 1. The traffic intensities of the combinations have the following relationship: $0.83 > \rho_A + \rho_F > \rho_C + \rho_E > \rho_B + \rho_G > \rho_D + \rho_H$.

4. Conclusion

We have developed a periodical WIP allocation algorithm for product-mix manufacturing for VLSI assembly facilities. For effective machine utilization, each machine is shared by multiple types of WIP. However when the type of WIP changes, the machine must be adjusted, or setup, for the new WIP. The adjustment reduces the machine utilization and requires additional resources. We discussed how to manage the WIP flow in the product-mix system, and how to reduce the adjustment frequency. When continuous flows of WIP are assumed, the allocation scheme in which two types of WIP are processed at a single machine can be solved under the constraint of the response times. This algorithm is based on the periodical allocation of two types of WIP to be processed in the same machine. This algorithm suggests that WIP flow in the facility can be controlled by measuring the traffic intensities and the number of WIPs in the incoming buffer. The algorithm also gives a rule for combining two types of WIP to be processed in the same machine. When we combine two types of WIP so that the sum of their traffic intensities is maximum, the adjustment frequency can be minimized. On the other hand, the machine utilization rate is not affected by the combination. We tested this estimation system in a real facility, and confirmed that it is accurate enough to be used for weekly planning.

Acknowledgment

Some parts of this research are supported by Fujitsu Tohoku Electronics Inc.

References

[1] K. Ozawa, H. Wada, and T. Yamaguchi, ‘Optimum Tool Planning Using the X-Factor Theory,’ ISSM’99, pp. 49 - 52, 1999.

[2] Y. J. Chen, et al., ‘A Next Queue Algorithm in Real Time Dispatching System of Semiconductor Manufacturing,’ ISSM’2000, pp.67 - 70, 2000.

[3] K. Saito, et al., ‘Application of a Resource Planning System for a VLSI Assembly Facility,’ ISSM’99, pp.345 - 348, 1999.

[4] S. Arima and K. Saito, ‘Operator Allocation Planning for a Product-mix VLSI Assembly Facility,’ IEICE Trans., Electron., Vol. E84-C, pp.832, 2001.

[5] R. Jain, “The art of computer systems performance analysis,” pp. 513, John Wiley & Sons, Inc. 1991.