Stepping up to a New Hybrid AMHS Design

GEORGE W HORN, Middlesex Industries SA

Contributions to variance that result in high WIP content are identified. Augmenting the AMHS with autonomous flow elements provides a solution.

CLEARLY, POOR FAB EFFICIENCIES OF around 50% (true process time vs fab cycle time) cannot be improved without reduction in fab work-in-process (WIP) content. And yet, current automated material handling systems (AMHS) cannot work without extra WIP to the fab. According to the probabilistic operating characteristic (OC) relation of Throughput and Cycle Time of a given factory, and the NP-Hardness [5] of algorithms dealing with its inherent process variability (dispatch), a new way to achieve a more favorable OC is via the revision of its logistics, while its IC and process design remain unchanged.

Considering the variance in the contemporary IC front end manufacturing as being the cause of low fab efficiency, a proof of the AMHS as a contributor is offered. A steady state model of the fabrication flow is adopted, where the inter process gaps are flow elements. It is shown that these inter process flow elements, controlled by the AMHS, increase the variance of arrivals. It is shown that the variance received from upstream tools is multiplied by the inter process AMHS. The evidence shows that such contributions to variance result in high WIP content, which in turn is the cause of inefficiency (cycle time vs. process time). To improve on this requires reduction in WIP content of the fab, achievable only through augmenting the AMHS with autonomous flow elements, described as hybrid infusions.

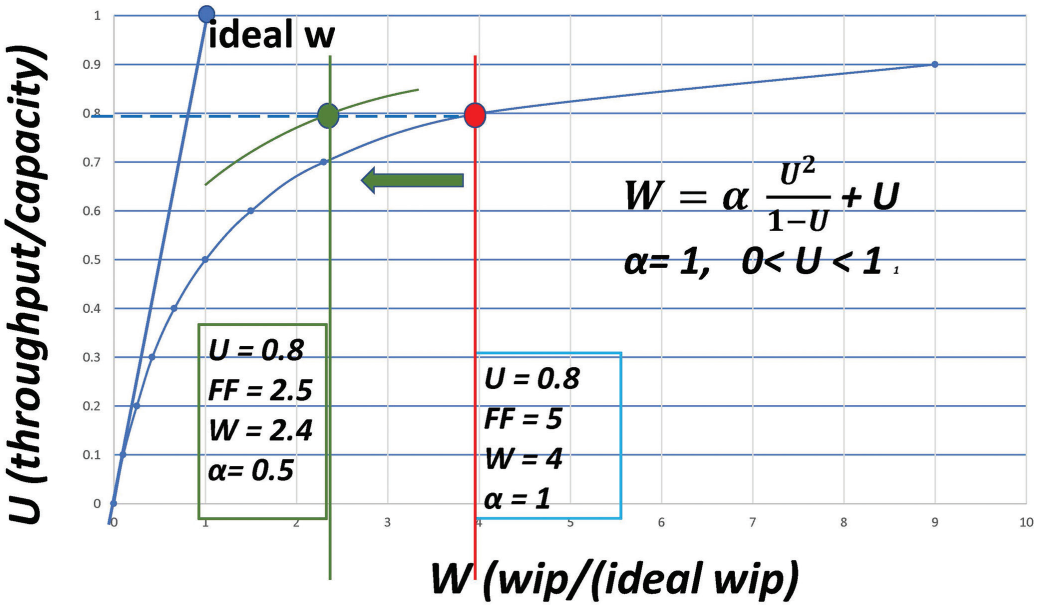

Figure 1. WIP vs. throughput of a fab at 80% capacity, with α=0.5 requires a WIP content 2.4 times the ideal WIP.

Background

There may be between 1000-2000 such flow elements in the front-end manufacturing process potentially contributing to the global fab process variance. Axiomatically, 1) wafer flow is considered steady state on long term, 2) the physical process variance, and the variance introduced by stochastic arrivals to a process manifest themselves in the wafer lot output of that process. Thus, examining the inter process phenomena — as the wafer lot evolution from the upstream process and the arrivals to the downstream process — will yield conclusions about the global fab variance as the detractor of fab efficiency. Such assumptions are justified, as simulation studies show little long-term effect of erratic equipment failures and scheduling upsets.

External contributions to the overall variance of downstream arrivals are 1) the upstream wafer lot release and 2) the dispatch strategy implemented at downstream process inputs. For example, in dispatch, a FIFO priority at inputs to a process will result in chaotic determinism. At the same time, release rate rules will contribute to setting general queue sizes at the input to each element. These queues, if sufficiently large, will eradicate some variance of lot arrivals to the downstream process. These imposed contributions to variance are under the control of fab operators. They can create more or less variance with their actions. Higher fab throughput will result with lesser created variance. And it can be assumed that release and dispatch policies are adopted to control the fab at a favorable point on its OC curve. But, today's algorithms, developed to control release and dispatch, categorically ignore the physics of moving wafer lots between process steps, which moving, in this study, is considered as an integral part of variance and method of manufacturing. Even though both dispatch and release have tangential controls for queue sizes at process arrivals. And even though these algorithms recognize the throughput gains achieved by minimizing queue sizes at process arrivals. The question remains if these algorithms will adopt queue size reductions to the point where the logistics of inter process moves becomes relevant to fab throughput. Today, wafer lot release and dispatch algorithms avoid operations where the AMHS would shows its relevance to throughput. The practice of maintaining a large enough queue at all process inputs creates intentional AMHS irrelevance, while increasing cycle times.

The Inter Process gap (exponentially distributed probabilities)

If the wafer lot output of a process is a stochastic Poisson process, estimated as an exponential distribution of the time intervals between output events, and 1/λx mean and 1/λx^2 as variance, and the transport of an output lot similarly being an independent Poisson process (discrete vehicles), estimated with exponential distribution of its service time, and 1/λy mean and 1/λy2 as variance. Then, the variance of lot arrivals to the next step will be the combined variance of the two independent variables. One: the variable of wafer lot output times from the first process and the variable of hat lot's service time by the AMHS. The probability distribution of arrivals to the next process is computed via multiplication of the two independent variables.

In general, the outcome of two independent random variables X and Y acting as a system is a probability distribution obtained by the multiplication of the two, XY = Z. [4]
multiplication of the two, $XY = Z$. [4]

$$P\ (X \cap Y) = P\ (X)\ P\ (Y) = P\ (Z) \qquad (1)$$

If X and Y are independent, and continuous, described by probability density functions $f_X(x)$ and $f_Y(y)$, then the probability density function $f_Z(z)$ of XY is [1]

$$f_Z(\mathrm{z}) \ = \int_{-\infty}^{\infty} f_X\,(x)\; f_Y\,(z/x)\,\frac{1}{1x} \qquad (2)$$

And the variance of the product XY in general is [1]:

$$\mathrm{VAR}(XY) = (\sigma_x{}^2 + \mu_x{}^2)(\sigma_y{}^2 + \mu_y{}^2) - \mu_x{}^2\mu_y{}^2$$

Where σ^2 is variance and μ is the mean of the distribution.

With the symbolism of exponential distributions,

$$\mathrm{Var}(XY) = [1/(\lambda_x)^2 + (\lambda_x)^2][1/(\lambda_y)^2 + (\lambda_y)^2] - \\ -(\lambda_x)^2(\lambda_y)^2 \qquad (3)$$

Solving for the ratio:

$$\mathrm{Var}\ (XY)/\mathrm{Var}(X) = 3\mathrm{Var}(Y) = 3/(\lambda_y)^2. \qquad (4)$$

Meaning a multiplication of discharge variance from upstream tools, resulting in arrival variance to downstream tools, as proportional to the AMHS variance.

The resultant of multiplication of the two random and independent variables is a bivariate density function of X & Y, $f_Z(z)$. This bivariate density function, and its variance describe the wafer lot arrivals to a typical process in the fab. Let $f_X(x)$ be the density of time intervals between wafer lots exiting a process and let $f_Y(y)$ be the density of AMHS service times in moving those wafer lots to the next process, with the following characteristics,

$$\{X(a \le x \le b) \cap Y(c \le y \le d)\}$$

where the height of space above the area (a-b) · (c-d) is bounded by the upper surface $f_Z(z)$. In other words, the probability of the area in the x-y plane given by the multiplicand of XY as a resultant variate Z.

If X=Y, then the AMHS service is synchronized with the evolution of wafer lots from a process. In such a case Z is fixed and defined only by the common frequency of X & Y. Wafer lot arrivals to a process are determined by wafer lot evolution from the previous process, i.e. $\lambda_x = \lambda_y = \lambda_z$, and $1/(\lambda_x)^2 = 1/(\lambda_y)^2 = 1/(\lambda_z)^2$. Therefore, the existing variance of a previous process has not been multiplied, or if considering the global fab, inherent variance in the process has held steady. This case would define an idealized AMHS, and close coupling of the processes. Such idealization with the OHT type of AMHS (i.e. discrete vehicle principle) is not possible.

While at the same time, the conveyor (hybrid) type of AMHS would closely approximate it.

In case X is fixed to be constant (i.e. $f_X(x)$ is invariate) the wafer lot evolution from a process is fixed and steady state, the arrivals to the next process will be determined by the AMHS only. This will be exponential in Y, with a variance of $1/(\lambda_y)^2$ modifying arrivals around λ_x mean. The normal industry procedure to reduce or eliminate this introduced variance into the process by the OHT type of AMHS is by creating a queue at the destination process. Of course, this will then increase fab cycle time.

In case Y is fixed to be constant (i.e. $f_Y(y)$ is invariate), the AMHS does not increase the exit variance of a process, and so that variance gets transferred unchanged to the destination process. Variance for the fab is then created embedded in a process. This scenario is generally assumed by fab management and measures implemented to deal with, as tools in a toolbox to increase competitiveness of the fab. These tools are the release and dispatch algorithms adopted by the fab. The current methodology to achieve this state of the fab is by making the AMHS irrelevant in manufacturing. Such irrelevance is achieved by creating high WIP content, which in turn assures high content of queues between process steps. To state this simply, fab cycle time is increased to quiet the instability that would be created by inter process moves through the use of variance multiplying AMHS types.

The transition from a fab where all the tools are fully buffered to a fab where no tools have buffered WIP arrivals is a mixed case of some of each. The thesis is that such fab, in between the two extremes is the most efficient one in practical terms. To this end a moderate number of process steps should be closely coupled (a degree of synchronization) via AMHS systems without variance, i.e. direct and continuous flow transports (hybrid systems – conveyors + local OHT). This process is similar to using cluster tools, except on a macro scale of integration.

A common belief has been that of scale. Meaning that the individual process variances are far larger than the AMHS variance in the process gap, therefore the latter is of little concern. However, the above calculations show that the significance of the AMHS introduced variance is in its multiplying the process variance. As a nonlinear multiplier of process variance, the AMHS should be of significant concern.

Implications of the OC curves

Controlling Q contents by regulating flow rates in the process gap (via the availability of parallel tools or via holding wip idle at upstream tools) implies a degree of underutilization of process tools. This directly contradicts common sense and presents us with a paradox. The problem is highlighted by the OC curve of the fab, which relates throughput to cycle time.

The Pollaczek-Khinchin result from queuing theory for Markovian arrivals to a single server M/G/1 queue is the basis for generating OC curves.

$$\frac{1}{\left(\frac{1}{\mu}\right)} = \frac{\rho(1+C^2)}{2(1-\rho)}$$

T- average time in queue and in service
μ - average service rate (throughput)
ρ - server utilization (1/μ)
C^2- $\sigma^2/(1/\mu)^2$ coefficient of service variance

Combined with Little's Law (by the IBM Consulting Group),

W (WIP) = (service time)·(arrival rate)

And further normalizing Cycle Time with Pure Process Time, as well as Throughput (Service rate) with design Capacity, the current OC relationship is used for factories:

$$CT = \alpha \frac{U}{1-U} + 1 \qquad (5)$$

CT is (cycle time)/(pure process time), also known as FF (flow factor)
α is Coefficient of variance $(C_1^2 + C_2^2)/2$,
U is throughput/capacity,
C_1 is coefficient of variance for inter process arrivals,
C_2 is coefficient of variance for the value-add processes,

Note, however, that the coefficient of variance for inter process arrivals is a result of multiplying the discharge variable of the upstream process with the independent AMHS variable that brings the WIP to the arrival queue of the next process. Therefore the variance of the AMHS becomes a factor in the variance of the arrivals.

It is also important to recognize that the above derivation with queuing theory and little's law does not consider the physics of inter process moves. It merely looks at a system from the outside, may that have any types of physical components. Thus the Pollaczek-Khinchin Queueing model assumes Markovian arrivals with a General service rate distribution. Meaning a service rate with variance unspecified (M/G/1). Consequently, the derivative formula for OC curves is also devoid of any specific probability distribution of service (i.e. AMHS type). Yet, specifying AMHS will alter arrival rates and hence α, the coefficient of the OC. It can be estimated that in practice the discrete vehicle AMHS will move carriers to destination via a stocker stop in 60% of cases (two inter process moves per wafer lot). And in these moves the arrival variance will be purely the AMHS variance (the variance of the second move). At the same time, the 40% direct inter-process moves will have the AMHS multiplying the variance of the upstream process discharge.

According to the probabilistic OC relation of Throughput and Cycle Time of a given factory, and due to the NP-Hardness to create algorithms used to improve its inherent process variabilities (dispatch), the tool remaining to achieve a more favorable OC is via the revision of its logistics.

The focus on Dispatch to minimize the inherent process variance is justified. However in these the regulation of overall WIP content of the Queues is overlooked. The proposal is to reduce Queue sizes to the point, where the logistics executed by the AMHS becomes relevant. And then, at that point to apply the hybrid AMHS, with is near zero variance multiplier in the gap for single step direct inter process moves.

The size of Inter Process queues can be controlled

Inter process flow is monitored via the AMHS move rates between process steps and processing times of the tools.

Thus, $Q_i = \lambda_i \cdot w_i$ and $w_i = (t_{dep} - t_{arr})_i$ according to Little's law.

Q_i is queue content between process steps

λ_i is the mean discharge rate from the i^{th} upstream process, w_i is the mean residence time between process steps, including time in buffer at the *i+1* tool downstream, (or mean residence time in the j^{th} process gap)

t_{arr} is the time the wafer lot arrived at the process, and t_{dep} is the time the wafer lot departed to the next downstream process.

With current data collection capabilities w_i and λ_i in a fab are known and inter process rate of flow can be made a variable. Thus $f_{i+1}(\lambda_{i+1})$ becomes a function releasing wafer lots to the *j+1* gap as a combination of 2 or more independent Poisson processes, depending on the number of identical downstream tools which are brought online to regulate λ_{i+1}, the flow rate out of the current inter process gap. Thus, $\lambda_i \neq \lambda_{i+1}$. To maintain stability of wip flow across all inter process gaps, the upstream queues are concurrently sampled and

Q_{1-1} ……….Q_{i-n} queues also adjusted. This may require the temporary adjustment of release rates.

The thesis is that such flow rate regulation of WIP out of the j^{th} inter process gap establishes a dispatch criteria which maintains inter process, and global, fab queue contents at levels where the AMHS variance multiplier gets regulated. Such a criteria is then incorporated into accepted dispatch algorithms. While today's dispatch and release algorithms may contain functions (such as look ahead at downstream queues) to regulate buffer contents, and purport to achieve maximum throughput rates for the fab, it is believed that these buffer content control mechanisms are too generous in size, thus eliminating considerations of AMHS and making AMHS irrelevant. Therefore, cycle time vs. throughput balances are achieved on a less efficient OC curve.

Reducing global WIP content of the fab, while maintaining target cycle times, will result in reduced throughput, and so, the only way to reduce WIP content at a fixed cycle time is to exit the Fab OC. In other words, to create a new OC. Given this result, the task is impossible without changing fab design, or fab logistics.

Overall WIP and throughput

Considering the relation of Fab throughput and WIP content the OC curve is converted by bringing Wip into the expression.

$$W = \alpha \frac{U^2}{1-U} + U \qquad (6)$$

The idealized basic WIP content in the fab resides in process tools, i.e. none in the transport gap between processes. Or, summing over each process tool group

$\sum_i (ct)_i (cap)_i$ which may be $\sum_i (ct)_i \, (cap)_{bneck}$

Considering FIGURE 1, the tangential to the initial rise of the WIP curve quantifies the ideal WIP in the system as W= 1. Further on Fig. 1 it can be seen that at 80% capacity the Fab WIP content is W=4. This means that 3 times more WIP is in the process gaps (i.e. idle or in transit between processes) than in the process itself. Clearly, an inefficient manufacturing system.

Fab efficiency and WIP

To improve fab performance the WIP content in waiting needs be reduced. This gets us closer to the distant ideal of Just in time. As we reduce WIP content, we step outside of the current OC curve and generate a new and better OC. On this new OC curve at a WIP content of W= 2.4 we achieve a coefficient of overall variance of 0.5. while theoretically such a reduction of WIP content is possible, it would prove to be impossible with current AMHS. Therefore, we need to create a new kind of AMHS to enable the WIP reduction. This is the rationalization of the hybrid AMHS, where we introduce a transport mechanism with less inherent variance than the current OHT system.

REFERENCES

1. An Introduction to Probability and Statistics, Second Edition, Wiley, Vijay K. Rohatgi, A. K. MD. & Ehsanes Saleh
2. An Integrated Release and Dispatch Policy for Semiconductor Wafer Fabrication, International Journal of Production Research, 2014, You LI & Zhibin Liang.
3. Frederich Böbel, PHD, Stefan Halmel, Siemens Semiconductors, ASMC 1998
4. Statistics for Scientists and Engineers, R. Lowell Wine, Prentice Hall.
5. NP-Hard. Wikipedia
6. Development and Simulation of Semi-conductor Production System Enhancements for Fast Cycle Times. Technical University Dresden, Killian Stubbe