Fluctuation Smoothing Production Control at IBM's 200mm Wafer Fabricator: Extensions, Application and the Multi-Flow Production Index (MFPx)

James R. Morrison
Engineering and Technology Dept.
Central Michigan University
Mount Pleasant, MI, USA
morri1j@cmich.edu

Elizabeth Dews
Production Control
IBM Corporation
Essex Junction, VT, USA
edews@us.ibm.com

John LaFreniere
Production Control
IBM Corporation
Essex Junction, VT, USA
jlafreni@us.ibm.com

Abstract
To increase the flexibility of existing production control algorithms and reduce the variation and mean of fabricator cycle times, a Fluctuation Smoothing for the Variation of Cycle Time (FSVCT) policy was implemented at IBM's 200mm semiconductor wafer fabrication facility. Extensions allowing for products with different cycle times and enabling the change of cycle time targets during production were developed. The policy was named the Multi-Flow Production Index (MFPx), reflective of its capabilities. Increased production agility and a controlled variation of cycle time resulted from the implementation.

Keywords
Fluctuation smoothing, production control, WIP management.

I. INTRODUCTION

As construction costs for state of the art 300mm semiconductor wafer fabrication facilities rise (reaching US$4 billion or more) and competition between existing facilities increases, efficient manufacturing operation is essential. One facet of efficiency is a competitive cycle time, which can lead to reduced operating costs, faster time to market, shorter yield learning cycles, improved customer satisfaction and increased market share and profit. Despite complications arising from the reentrant structure of semiconductor wafer fabrication, simulation studies ([1, 2, 3]), analytic performance evaluation ([4]) and implementation results ([5,6]) have demonstrated that a careful choice of which lot to next process (production control or work in process (WIP) management) can have dramatic implications for the mean and variance of fabricator cycle time.

Many control policies have been developed (and some deployed) ranging from relatively simple policies (e.g., critical ratio) to more complex ones with the potential for superior performance (e.g., the mathematical programming based approach of [7]). At IBM's 200mm semiconductor fabrication facility, a multi-objective production control policy combining elements of critical ratio and continuous flow methodologies had been employed for many years. One difficulty experienced in the application of such a policy was in treating the ever changing mix and volume of products released for production. As a consequence, lots whose production was deemed imperative were prioritized (sometimes overly so) at the expense of the nominal control decisions and the remaining lots.

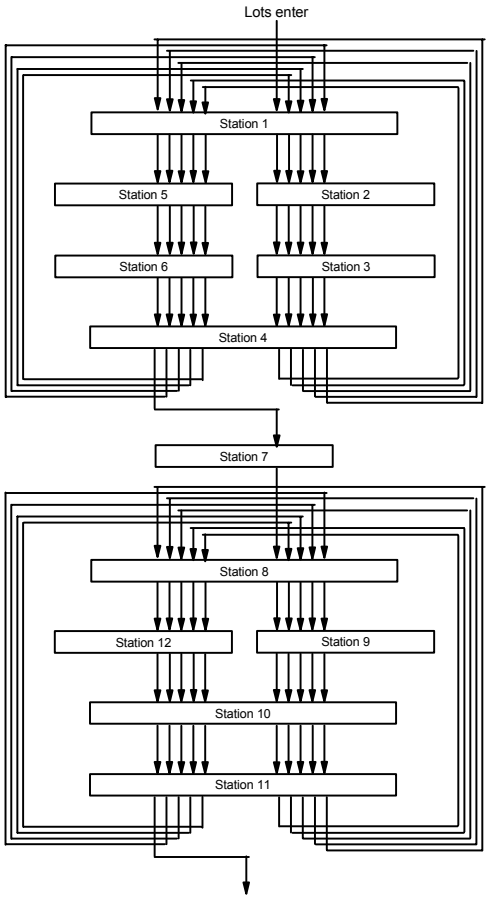

Figure 1. A simple reentrant process flow for a single product.

IBM's 200mm semiconductor fabrication facility features a highly reentrant process flow (lots of wafers return to various toolsets time and again) with as many as thirty or more visits to certain tool groups. The menu of products offered contains approximately thirty different semiconductor technologies each containing a dozen distinct products or more. To create a single semiconductor wafer, as many as four hundred stages of production may be required from on the order of one hundred distinct tool groups. Cycle times vary

by technology, product and business priority and generally range from forty to sixty days with an underlying production time of around twenty days. Figure 1 depicts a simple example reentrant process flow for a single product.

To account for the highly volatile mix and volume, enable the treatment of lots with diverse cycle time requirements and reduce the variation of the cycle times about their expectations, a new methodology based on the fluctuation smoothing concepts of [1] was developed, see [6]. Existing batching and setup rules were not modified, only the relative importance of each lot was adjusted. The control policy, termed the MFPx for **M**ulti-**F**low **P**roduction Inde**x**, was implemented April 12, 2005. One of the MFPx's primary objectives is to reduce the variation of expected cycle time for collections of lots about their mean. As suggested by the Pollaczek-Khinchin formula [8], one expects that a reduction in the mean cycle time will result.

In this paper, two generic forms for a multiple cycle time FSVCT policy slack are developed. Also, this paper generalizes the approach in [6] to enabling lots to change their cycle time targets during production. The generalization is a key to the ability of the control to adjust for changes in the wafer release levels and allow for lots to change priorities at the dictates of business requirements. Further, the consequences of the control policy to the variation of achieved cycle time and WIP levels following the implementation of the MFPx control are reviewed.

The paper is organized as follows. Section II reviews the fluctuation smoothing for the variation of cycle time (FSVCT) production control policy and develops generic extensions for the presence of multiple cycle times. Section III develops two methods enabling a change in cycle time targets for lots during production. Section IV highlights the consequences of the policy implementation. Concluding remarks are presented in Section V.

II. THE FSVCT POLICY AND EXTENSIONS

The Fluctuation Smoothing for the Variation of Cycle Time (FSVCT) production control policy was first proposed in [1] and assigns to each lot a number termed slack. When a tool becomes available to accept another lot into production, it selects that lot with the least slack from among those lots to which it caters. In general, the FSVCT policy slack values for each lot may serve to guide batching (e.g., furnace and cleaning tools) and setup (e.g., ion implant and photolithography tools) decisions without providing explicit recommendations.

II.1. The Basic FSVCT Policy

Let *Now* be the time at which the control decision is to be made (i.e., the time at which a tool becomes available to accept another lot into production) and let $\alpha(l)$ be the arrival time of lot l to the fabrication facility (equivalently, the release time into production). For a manufacturing facility with a single product, the FSVCT policy defines the slack for a lot l as

$$s(l) := -[\text{Now} - \alpha(l)] - \rho(\sigma(l)).$$

The first term in brackets is the time that the lot has been in the facility. The term $\sigma(l)$ is the stage of production at which the lot resides (e.g., the 235th stage of production along the 400 stage route) and $\rho(\sigma)$ denotes the expected *remaining* cycle time for a lot at stage σ until completion (including production time, queueing time and other overheads). The FSVCT slack value is thus the negative of the expected cycle time for lot l, if it travels at the expected pace from its present stage $\sigma(l)$ until the end of production.

The $\rho(\sigma)$ may be found via simulation, analytic performance evaluation, measurement of actual performance or approximation. The key is that the values used in the slack calculation should be the expected values obtained when the fabricator is operated under the slack policy. The authors in [6] discuss an approximate method to determine $\rho(\sigma)$ in the absence of a fabricator model.

Example 1. *Consider two lots l_1 and l_2 with [Now – $\alpha(l_1)$] = 10 days and [Now – $\alpha(l_2)$] = 21 days. Let the production route consist of 360 stages, each with an expected cycle time of four hours (perhaps consisting of 1.5 hours of actual production and 2.5 hours of queueing and other overheads). The total expected cycle time is thus sixty days (360 stages * 4 hours/stage). Let $\sigma(l_1)$ = 61 (l_1 has completed 60 stages of production) and $\sigma(l_2)$ = 121 (l_2 has completed 120 stages of production).*

Since l_1 and l_2 have 300 and 240 stages of production remaining, respectively, and each stage is expected to require four hours, $\rho(\sigma(l_1))$ = 50 days and $\rho(\sigma(l_2))$ = 40 days. The FSVCT slack values are thus $s(l_1)$ = –10 days – 50 days = –60 days and $s(l_2)$ = –21 days – 40 days = –61 days. Since the slack for l_2 is least (it's expected total cycle time is the greatest), this lot will receive priority over lot l_1.

At each tool group, the FSVCT policy recommends production of the lot with the greatest expected total cycle time (when continuing at the expected cycle time pace). As a consequence, lots proceeding faster than the expected cycle time will slow while lots proceeding more slowly will accelerate. Thus, the FSVCT policy strives to drive all lots to the same total cycle time, so that one expects to reduce the variation of total cycle time. As suggested by the Pollaczek–Khinchin formula (see, for example, [8]), one also expects a reduction in the expected total cycle time. Simulation studies in [1, 2, 3] and implementation results in [6] have demonstrated that the FSVCT policy can improve system performance over baseline policies such as first-in first-out (FIFO), critical ratio and other common dispatching heuristics.

II.2. Incorporating Multiple Expected Cycle Times

For a manufacturing facility with a myriad of products, labeled p_1, p_2, …, p_M, each with potentially different cycle time expectations, the basic FSVCT policy can be extended. Two extensions are developed here. First it is helpful to generalize the notation of Section II.1 for each product p_i. Let $\sigma_i(l)$ denote the current stage of production for a lot l of product p_i. Let $\rho_i(\sigma)$ denote the expected remaining cycle time for a lot l of product p_i at stage σ. Let $c_i(\sigma)$ denote a positive constant associated with stage σ of product p_i lots.

The first generic form for the multiple cycle time FSVCT (MCT–FSVCT) slack is the total cycle time form, which is defined as

$$s(l) := -\left[\left(\text{Now} - \alpha(l)\right) - \rho_i(\sigma_i(l))\right] \cdot c_i\left(\sigma_i(l)\right),$$

for a lot l of product p_i. This is the form employed at IBM's 200mm semiconductor fabricator with $c_i(\sigma) = CT_{NOR} / CT_i$, for all σ, for an arbitrary normalization constant CT_{NOR} (e.g., 40 days) with CT_i denoting the expected *total* cycle time for a lot of product p_i (just entering the facility and proceeding on pace). Due to the capability of the MCT-FSVCT policy to provide variation control properties for multiple products with different expected total cycle times, the control policy (with the choice $c_i(\sigma) = CT_{NOR} / CT_i$) was named the **M**ulti-**F**low **P**roduction Inde**x** or MFPx at IBM's 200mm fabricator. For products with multiple target cycle times corresponding to different customers to which different cycle times have been committed, one may create a separate label p_i for each.

Example 2. *Consider two lots, l_1 of product p_1 and l_2 of product p_2, with [Now – $\alpha(l_1)$] = 11 days and [Now – $\alpha(l_2)$] = 21 days. Suppose the expected remaining cycle time until the lots exit the fabricator are given as $\rho(\sigma(l_1))$ = 20 days and $\rho(\sigma(l_2))$ = 40 days. Further consider that the scaling constants for lots of product p_i are $c_i(\sigma) = CT_{NOR} / CT_i$, where CT_{NOR} = 40 days, CT_1 = 30 days and CT_2 = 60 days.*

The expected total cycle time for lot l_1 is 31 days if it continues at the expected pace and CT_1 = 30 days (lot l_1 will be one day late). The expected total cycle time for lot l_2 is 61 days if it continues at the expected pace and CT_2 = 60 days (lot l_2 will be one day late).

*The MCT–FSVCT slack values are $s(l_1)$ = –[31 days] * CT_{NOR} / CT_1 = –[31 days]*[40 days / 30 days] = –41.33 days and $s(l_2)$ = –[61 days]*[40 days / 60 days] = –40.67 days. Since the slack for l_1 is least, it will receive priority over lot l_2 if they are at the same tool group. Note that both lots are expected to be one day late, but that since lot l_1 has a shorter nominal expected total cycle time (CT_1), one day is of greater import (resulting in less slack).*

The second generic form for the MCT–FSVCT slack is the lateness form, which is defined as

$$s(l) := -\left[\left(\text{Now} - \alpha(l)\right) - \rho_i(\sigma_i(l)) - CT_i\right] \cdot c_i\left(\sigma_i(l)\right),$$

for a lot l of product p_i. Here, CT_i is the expected total cycle time for a lot of product p_i (just entering the facility and proceeding on pace). Note that the term in the square brackets is the expected lateness of lot l if it proceeds at the expected pace from its current stage of production. This form of the MCT-FSVCT policy is not employed at IBM.

III. ADJUSTING CYCLE TIME TARGETS DURING PRODUCTION

To facilitate business agility, allow for mutable customer demands and respond to random yield fluctuations, IBM's production control organization required the new MFPx policy to enable the implementation of changes in the expected cycle times of lots during production. The desired behavior was for lots to proceed from the moment of change at the pace dictated by the new expected cycle times provided. For lots already in production, the cycle time changes would apply to future stages of production and not to those stages that were *supposed* to have been completed prior to the expected cycle time updates. For all lots released into the fabrication facility following a change in the expected cycle time values, merely calculating the MFPx slack value with updated values for $\rho_i(\sigma)$ and CT_i is sufficient. For lots presently in the fabrication facility, to employ the updated values for $\rho_i(\sigma)$ and CT_i in the slack calculation, one must also adjust the arrival time used.

To define the updated arrival time, additional notation is required. As before, let $\rho_i(\sigma)$ and CT_i be the expected remaining and total cycle times, respectively, for lots of product p_i prior to any adjustment of the cycle time targets. For a lot l of product p_i such that Now – $\alpha(l) < CT_i$ (the lot is not yet expected to have completed production), let $\eta(l)$ denote the stage of production at which a lot l would be located if it had proceeded at the expected cycle time. That is, if Now – $\alpha(l) < CT_i$, let $\eta(l)$ be the stage of production σ such that

$$Now - \alpha(l) \geq CT_i - \rho(\sigma),$$

and

$$Now - \alpha(l) < CT_i - \rho(\sigma+1),$$

where $\rho_i(\sigma+1) = 0$ if σ is the last stage of production for product p_i.

Denote the adjusted (newly updated) expected remaining and total cycle times as $\rho'_i(\sigma)$ and CT'_i, respectively. For a lot of product p_i with Now – $\alpha(l) < CT_i$, define the adjusted arrival time as

$$\alpha'(l) := Now - \left[CT'_i - \rho'_i(\eta(l))\right] - \Delta'_i(\eta(l)) \frac{\left[\left(\text{Now} - \alpha(l)\right) - \left(CT_i - \rho_i(\eta(l))\right)\right]}{\Delta_i(\eta(l))},$$

where $\Delta_i(\sigma) = \rho_i(\sigma) - \rho_i(\sigma+1)$ and $\Delta'_i(\sigma) = \rho'_i(\sigma) - \rho'_i(\sigma+1)$. As Figure 2 depicts, the updated arrival time $\alpha'(l)$ is Now minus the time (based on the new expected remaining cycle times $\rho'_i(\sigma)$) the lot should have been in manufacturing given that it is expected to be at stage $\eta(l)$. The third term in the definition of $\alpha'(l)$ accounts for the proportion of time that the lot would have penetrated into stage $\eta(l)$.

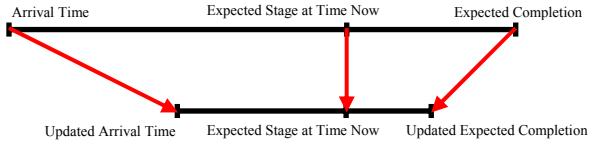

Figure 2. For a lot expected to be at a certain stage of production, the adjusted time since arrival is reduced when the expected cycle times are reduced.

In the case that the lot *should have* already completed production, that is Now – $\alpha(l)$ is greater or equal to CT_i, define

$$\alpha'(l) := Now - CT'_i\left[\frac{Now - \alpha(l)}{CT_i}\right].$$

Example 3. *Consider a product p_1 requiring 100 stages of production with an expected cycle time of six hours per stage. Suppose that a lot l_1 of product p_1 has 50 stages of production remaining and has been present in the fabricator for 330 hours (i.e., Now – $\alpha(l_1)$ = 330 hours). The expected cycle time values are $\rho_1(j) = 6*(101-j)$ hours, for j = 1 to 100, and $CT_1 = 6*100 = 600$ hours.*

Since, Now – $\alpha(l_1) < 600$ hours, one may deduce that lot l_1 should be at the $\eta(l_1) = 56^{th}$ stage of production. This is clear because the expected total cycle time for lots starting production is 600 hours, so that 270 hours remain until lot l_1 is expected to exit the fabricator (if it had proceeded at the nominal cycle times throughout). Since 270 hours corresponds to 45 full stages of production remaining, lot l_1 should have just entered the 56^{th} stage of production. This can be more systematically deduced by applying the definition of $\eta(l_1)$ given above.

If the expected cycle time for lots of product p_1 is adjusted to three hours per stage of production, the adjusted cycle time values are $\rho'_1(j) = 3*(101-j)$ hours, for j = 1 to 100, and $CT'_1 = 3*100 = 300$ hours. Application of the definition of $\alpha'(l)$ yields that

$$Now - \alpha'(l) = \left[300 - \rho'_i(56)\right] + 3\frac{\left[(330) - \left(CT_i - \rho_i(\eta(56))\right)\right]}{6}$$

$$= \left[165\right] + 3\left[\frac{330 - 300 + 270}{6}\right] = 165 \text{ hours}.$$

To determine the slack for lots in production following an adjustment to expected cycle times, employ the updated expected cycle time and arrival time values. Below, the formula for the MFPx employed at IBM's 200mm fabricator is provided for a lot of product p_i:

$$s'(l) := -\left[\left(\text{Now} - \alpha'(l)\right) - \rho'_i(\sigma_i(l))\right] \cdot \frac{CT_{NOR}}{CT'_i}.$$

Note that the stage of production is not changed as only the cycle times and arrival time have changed (the lot is still at stage $\sigma_i(l)$).

Example 4. *Let CT_{NOR} = 450 hours. The MFPx (MCT-FSVCT slack value) of lot l_1 in Example 3, before the adjustment to cycle times, may be calculated as $s(l_1)$ = -(330-50*6 hours)*(450 hours)/(600 hours) = -(630 hours)*(450 hours)/(600 hours) = -472.5 hours.*

*The MFPx of lot l_1 in Example 3, after the adjustment to cycle times and arrival time, may be calculated as $s'(l_1)$ = -(165-50*3 hours)*(450 hours)/(300 hours) = -(315 hours)*(450 hours)/(300 hours) = -472.5 hours.*

Note that the slack value did not change after the expected cycle time adjustment.

As suggested by Example 4, if the changes in the cycle time values for lots of product p_i are linear, i.e., $\rho'_i(\sigma) = K_i \rho_i(\sigma)$ and $CT'_i = K_i CT_i$, for fixed constant $K_i > 0$, then the slack value $s'(l) = s(l)$. This fact leads to an alternate and simplified methodology for the determination of the adjusted arrival time, obtained by setting the old slack equal to the new and solving for $\alpha'(l)$:

$$\alpha'(l) := Now - K_i\left[Now - \alpha(l)\right].$$

For linear changes in expected cycle time values, the resulting $\alpha'(l)$ calculated in this manner yields the same result as the more general case. Also in this case, the slack values are unchanged following a change to expected cycle time values.

In general, the two methodologies are different; however the results are often similar. At IBM's 200mm fabrication facility, the second approach is employed for simplicity of implementation even though expected cycle time changes may be nonlinear (as when loading changes cause modifications to the expected cycle time values).

IV. IMPLEMENTATION RESULTS

The ability of MFPx to control lots from different populations to achieve different cycle times while simultaneously reducing the variation of the cycle times about the means was considered a substantial increase in functionality beyond previous WIP management methodologies. During implementation, the fabricator experienced a reduction in wafer releases (and consequently WIP), thus obscuring the impact of our control. However, a substantial reduction in the variation of cycle time behavior does appear correlated with the implementation of MFPx (though we must still

extrapolate from system behavior before implementation to reach this conclusion). As wafer release levels were changing, it is difficult to distinguish the magnitude of variation improvement attributable to the MFPx control.

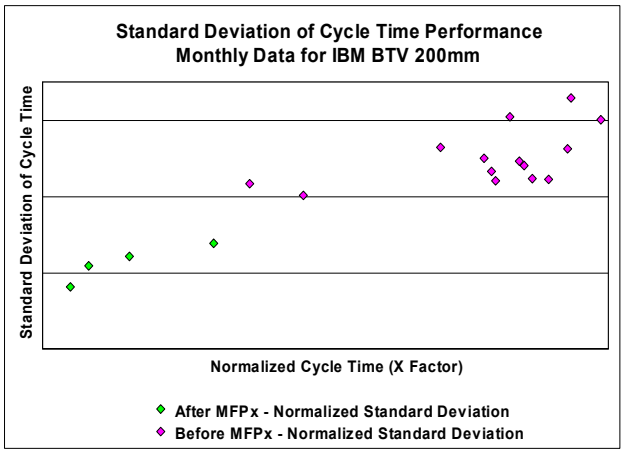

Figure 3. Standard deviation of cycle time (averaged monthly) is reduced and itself has less variation.

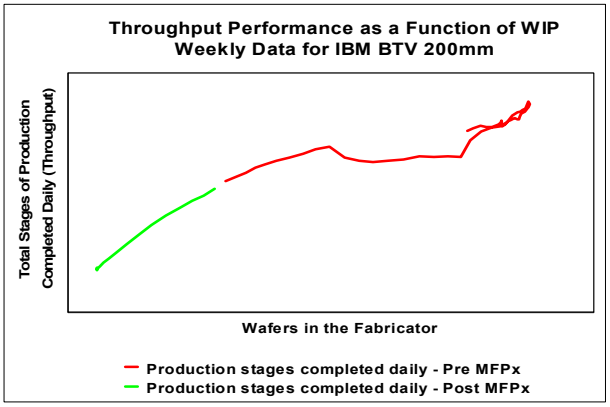

Figure 4. Average throughput trend does not significantly appear to change (in low loading situation).

Figure 3 demonstrates the reduction and tightening of the monthly variation of cycle time data as a function of mean cycle time. Figure 4 provides the aggregate rate at which stages of production that were completed within the fabricator (summed throughput for all tools) as a function of the total number of wafers in the fabricator. In the relatively low loading regime in which the fabricator was operating (roughly 30% of the cycle time was due to queueing), there does not appear to be a difference in the general trend of the performance (though this is not unexpected in a low loading regime).

V. CONCLUDING REMARKS

A multiple cycle time FSVCT production control policy termed MFPx was implemented at IBM's 200mm wafer fabrication facility in April, 2005. Numerous extensions to the basic FSVCT policy were developed to ensure successful implementation and that management needs were fulfilled. Increased production control agility and a reduction in the variation of cycle time were achieved.

REFERENCES

[1] S. C. H. Lu, D. Ramaswamy and P. R. Kumar, "Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 7, No. 3, pp. 374-388, August 1994.

[2] M. Janakiram and J. R. Morrison, "Capacity planning and study of scheduling policies using simulation at Motorola's ACT fab," Proceedings of the 1999 SMOMS Conference, San Jose, CA, 1999.

[3] J. R. Morrison, M. Janakiram and P. R. Kumar, "A comparative study of scheduling policies at Motorola," Proceedings of the International Conference on Semiconductor Manufacturing Operational Modeling and Simulation (SMOMS), San Francisco, CA, pp. 51-56, January 1999.

[4] S. Kumar and P. R. Kumar, "Performance bounds for queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, Vol. 38, No. 7, pp. 1600-1611, August 1994.

[5] R. M. Dabbas and J. W. Fowler, "A new scheduling approach using combined dispatching criteria in wafer fabs," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 16, No. 3, pp. 501-510, 2003.

[6] J. R. Morrison, B. Campbell, E. Dews and J. LaFreniere, "Implementation of a fluctuation smoothing production control policy in IBM's 200mm wafer fab," *Proceedings of the Joint 44th IEEE Conference on Decision and Control (CDC) and European Control Conference (ECC)*, Seville, Spain, 2005.

[7] M. Chen, R. Dubrawski and S. P. Meyn, "Management of demand-driven production systems," *IEEE Transactions on Automatic Control*, Vol. 49, No. 5, pp. 686-698, May 2004.

[8] L. Kleinrock, Queueing Theory, Volume 1: Theory, John Wiley - Interscience, New York, N.Y., 1975.

BIOGRAPHY

James R. Morrison holds a Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign. He works as an Assistant Professor of Electrical Engineering at Central Michigan University.

Elizabeth Dews served in the US Air Force and works as a Staff member in Production Control at the IBM Corporation.

John LaFreniere manages the Industrial Engineering team in the 200mm semiconductor fabrication facility at the IBM Corporation.