EL IMPACTO DE LA DEMANDA ENDÓGENA EN LOS SISTEMAS DE DE PRODUCCIÓN PUSH-PULL

Esta investigación construye y analiza un modelo de cadena de suministro de semiconductores en la cual la demanda del cliente responda a la disponibilidad del producto, basado en un largo y profundo estudio del campo de la cadena de suministro de Intel, el modelo captura los flujos de material de producción y de la respuesta de los clientes al porcentaje de disponibilidad o nivel de servicio del fabricante.

Ana Maria Zuluaga Betancur
Práctica Investigativa III

1. INTRODUCCIÓN

Desafiando la revolución en las cadenas de suministro de la década pasada, las compañías en diversas industrias como computadores, electrónicas, autos, juguetes, semillas, y farmacéuticas, continúan luchando con los retrasos en la producción y el envió. Un efecto importante de esos retrasos es generado por la *inestabilidad y los problemas de la cadena de suministros*; y por *la reducción del valor de los accionistas*. Hendricks y Singhal en el 2003 mostró que un decremento anormal del 10% en el valor del las acciones es causado por falta de partes, cambios en las ordenes de los clientes y los problemas de altas y bajas de producción entre otros. Adicionalmente ha sido reconocido a lo largo que el efecto látigo tiende a amplificar la inestabilidad en las órdenes, cuando se mueve rio arriba en la cadena de suministros, potencializando el efecto que también se mueve rió arriba, por ejemplo fabricantes de semiconductores son más propensos a transferir fallas imprevistas en sus cadenas de suministros, para la muestra, a causa de falta de partes para el Boeing que paro la producción de su avión 747 (por cerca de un mes) y retrasando el ensamble final del avión 737 y conduciendo a "entregas más retardadas, a costos más altos, clientes molestos y beneficios reducidos" (Holmes 1997). La corporación Intel ha luchado constantemente con la escasez de partes, alta variabilidad de la demanda y cambios en las órdenes y cancelaciones de clientes. En noviembre de 1999 enfrentando la escasez de los procesadores Pentium III, Intel planeó introducir una nueva planta para el siguiente año. En el 2000, culpaba las cancelaciones de las órdenes de los clientes grandes y la caída de la economía, Intel advirtió que sus ingresos quedarían cortos de los proyectados y las ventas serian planas para el trimestre (Gaither 2001).

El desafió de la variabilidad de la demanda, inestabilidad y amplificación de órdenes es complicado para la producción de periodos largos. En semiconductores, los largos tiempos de throughput (aproximadamente 13 semanas) afecta la habilidad de los manufactureros para mantener adecuados niveles de inventarios en el enfrentamiento de la variabilidad de la demanda. Cuando la demanda de los clientes varía, los administradores de la fábrica deben ajustar la utilización de la capacidad para mantener adecuados niveles de servicio mientras que evitan el exceso de inventario. Sin embargo la

combinación de la variabilidad de la demanda y largos periodos de producción conduce a menudo a *periodos que se alternan entre escasez y exceso en el suministro*. La variabilidad resultante en el suministró puede retroalimentar la demanda y la rentabilidad de los clientes como una inhabilidad de la compañía para satisfacer la demanda conduce a ventas perdidas, erosionado la reputación y decremento del goodwill (buen nombre, confiabilidad). La interacción de la inestabilidad de la cadena de suministros y las respuestas de los clientes da paso a preguntas interesantes:

- ¿Cuál es el impacto de la demanda endógena en la variabilidad de la cadena de suministros?
- ¿Cuál es el impacto de la variabilidad de la cadena de suministro en la respuesta del cliente?
- ¿Qué políticas puede implementar Intel u otras compañías para estabilizar sus cadenas de suministro?

Para responder estas preguntas, esta investigación construye y analiza un modelo de cadena de suministro de semiconductores en la cual la demanda de cliente responda a la disponibilidad del producto, basado en un largo y profundo estudio del campo de la cadena de suministro de Intel, el modelo captura los flujos de material de producción y de la respuesta de los clientes al porcentaje de disponibilidad o nivel de servicio del fabricante. En particular, incorpora dos efectos de la respuesta al cliente. Primero el *efecto de las ventas* captura la realimentación negativa que hace que los clientes busquen fuentes alternas de suministro, reduciendo la demanda y facilitando (o volviendo menos grave) la escasez. Esto es, un cambio en la demanda retroalimenta para mitigar el impacto de disturbio inicial.

En segundo lugar, *efecto de la producción* captura el efecto retrasado de cambios en la demanda en las decisiones de los productores a través de un ciclo de realimentación positivo. Si la demanda cae, reducen sus pronósticos de demanda y la utilización de capacidad para evitar excesos de inventario. Después de un retraso de producción, la producción baja y deja el inventario y la disponibilidad baja, causando una futura pérdida de demanda de clientes. La demora en el *efecto de la producción* genera una reacción que refuerza el impacto del disturbio original.

Veremos que la respuesta endógena del cliente a la disponibilidad deja más inestabilidad comparado con modelos en los que la demanda de cliente se trata como exógena. Sin embargo los modelos de inestabilidad de la cadena de suministros que asumen demanda exógena pueden subestimar la amplificación en la demanda y el valor del búfer de inventario. Más aun, tratando la demanda endógenamente conduce a diferentes políticas de inventario y utilización de los actuales inventarios en uso por la firma. En particular, el suministro debe mantener más altos niveles de inventario de seguridad del producto en proceso (WIP) y producto terminado (FGI); y reducir la sensibilidad de la utilización a cambios en la demanda del cliente causados por niveles de servicio inadecuados. Basado en los costos asociados con ventas perdidas y el mantener inventario en ensamble y bienes terminados, se derivará una recomendación para la óptima localización y cantidad de inventarios de seguridad. La política heurística provee una reducción aguda en la inestabilidad de la cadena de suministro y minimiza el impacto de las ventas perdidas. El modelo analizado en este artículo brindará una comprensión en los costos que apoya las estrategias de inventario y políticas de utilización en respuestas al contexto de sistemas de producción híbridos y demanda endógena.

2. REVISIÓN LITERARIA

La inestabilidad de la cadena de suministro y la influencia del nivel de inventario en la demanda han atraído la atención de investigadores y expertos en diferentes campos como economía, sistemas dinámicos, y administradores de operaciones. En economía, expertos en cadenas de suministro se conocen desde Thomas Mitchell (1924) y su descripción del mecanismo a través del cual los minoristas conjugaron sus pedidos incrementando las ordenes a los distribuidores. En sistemas dinámicos, el estudio de la inestabilidad de la cadena de suministros ayudo a diseñar los fundamentos necesarios para crear el campo (Forrester 1958,1961) investigaciones subsecuentes exploraron aplicaciones en diversos áreas incluyendo interacciones entre la cadena de suministros la fuerza trabajadora (Mass 1975); el comportamiento de la planeación de requerimientos de materiales (MRP) (Morecroft 1980). Los estudios de laboratorio de la habilidad de las personas para manejar sistemas complejos como la cadena de suministros en el juego de la cerveza (Sterman 1989[a], 1989b); La toma

de decisiones bajo diferentes niveles de complejidad de realimentación (Dilehl and Sterman 1995); y el impacto de ciclos de negocios en capital para equipos de la cadena de suministros (Anderson and fine 1999). En adición, modelos en dinámicas de sistemas también incorporaron la realimentación de la disponibilidad de inventario en la demanda de los clientes como el modelo "market growth" de Forrester (1968) y el modelo de Gram (1977) que investigó el impacto de adicionar un lazo de menor importancia a los sistemas oscilatorios. Mientras este artículo enfatiza la combinación de la respuesta de clientes endógenos con la inestabilidad de la cadena de suministros, la mayor contribución para la literatura de la dinámica de sistemas es la investigación del impacto de los dos en los sistemas de producción híbridos.

En la administración de operaciones, los expertos en investigaciones de la inestabilidad de la cadena de suministro asumen típicamente demanda exógena, y estudios de exploración de la influencia del nivel de inventarios en la demanda de los clientes sin consideran múltiples etapas en la cadenas de suministros. Ejemplos de lo anterior incluye los modelos de procesamiento de señales de demanda, racionamiento, loteo de órdenes y variaciones de precios de Lee *et al* (1997a, 1997b). El modelo jerárquico de Baganha y Cohen (1998); el sistema de inventario de un sólo artículo con demanda no estacionaria de Gaves (1999), el modelo con técnica de pronóstico de demanda y retraso de órdenes. Ejemplos del modelo después incluido de Dana y Petruzzi del vendedor de periódicos donde el cliente escoge entre la compañía y suministrador externo; el modelo dinámico de Gans del comportamiento individual del cliente, donde los clientes actualizan sus principales creencias acerca de la compañía, después de cada contacto; y el modelo de Hall y Porteus (2000) donde el nivel de sevicio esperado es una función de la capacidad y la competencia de la firma de invertir en capacidad para servir a los clientes.

Nuestra investigación llena un vació en la literatura de la administración de operaciones explorando ambos efectos de las respuestas de clientes endógenos y la inestabilidad de la cadena de suministros. Nuestros resultados extienden los resultados del caso de Dana y Petruzzi (2001) el resultado del caso de múltiples etapas una cadena de suministro con de demoras en la producción, mostrando que cuando una compañía contabiliza los efecto de la disponibilidad

del inventario en la demanda es óptimo para mantener mas inventario de seguridad.

3. SITIO DE INVESTIGACIÓN

La investigación es el resultado de un año de un profundo análisis de la cadena de suministro de Intel. Intel es líder en tecnología de manufactura de microprocesadores. Entre muchas introducciones originales, Intel fue el primero en producir tecnología de 0.13 micrones, permitiendo así doblar el tamaño de memoria del procesador reduciendo su tamaño en un 30% al mismo tiempo. Tales innovaciones resultaron en microprocesadores más rápidos y en un incremento del número de chips manufacturados por wafer. La compañía era también la primera en la transición de wafers de 200mm a 300mm llevando a mayor eficiencia en la producción de los chips. Para administrar la variabilidad en la línea de producción, la producción, y demanda, Intel empleó cerca de 1500 planeadores que dirigen decisiones de producción a corto y largo plazo, usando sistemas sofisticados y pautas detalladas para dirigir sus decisiones. El modelo de desarrollo exigió entrevistas con planeadores de diversos alcances de decisión y responsabilidades para entender el proceso de toma de decisiones en el sistema de producción de Intel. En adición, el equipo de investigación entrevistó directores en diversas áreas de la corporación, como operaciones, administración de la cadena de suministros, tecnología de información, pronósticos de demanda, marketing y ventas. En total conducimos casi 100 entrevistas semiestructuradas entre entrevistas en el sitio y conferencias telefónicas semanales. La investigación también requirió repasar los registros de Intel que detallaban pautas para la toman de decisiones y recolección de datos cuantitativos y cualitativos relacionados. Lo anterior incluyendo datos de series de tiempo de capacidad trimestral, utilización, comienzos de wafer, envíos, pronósticos, niveles de servicios y mercados de partes. Las decisiones heurísticas posteriormente incluidas de los administradores, lineamientos e incentivos de las compañías, y las dependencias de información entre áreas del negocio. Estos datos ayudan a establecer las suposiciones usadas en el modelo que captura la manufactura semiconductora del Intel.

4. SUPOSICIONES DEL MODELO.

La manufactura semiconductora es comúnmente dividida en dos fases principales: fabricación y ensamble. La primera fase (fabricación) toma lugar en la instalación de fabricación de Wafer, or Fab. Se toma de 200 mm – 300 mm en forma de disco pulido (Wafers) de substrato de silicio en la entrada y a través en una secuencia complicada de pasos los trasforma en Wafer fabricado, compuesto de cientos de pulgadas cuadradas de circuitos integrados (ICs o dados). Una sección cruzada vertical de un circuito integrado revela un número de capas formadas durante el proceso de fabricación. Las capas más bajas producidas en el "front-end" (frente final) del proceso de fabricación incluye los componentes eléctricos de cristal (transistores y capacitares). Las capas de más arriba producidas en el "back-end" o (trasero final) del proceso de fabricación, conectan los componentes eléctricos para formar el circuito. En adición, la fabricación se caracteriza por un flujo reentrante al proceso, es decir, el mismo equipo desarrolla múltiples pasos en diferentes etapas de la fabricación (tal como: litografía, grabado, películas delgadas y difusión).

En la segunda fase de manufactura (ensamble) los wafers se cortan en dados y se almacenan en inventario de dados ensamblados (ADI) en la bodega, ubicada con la planta de ensambles/prueba. Los dados se reciben en un paquete que protege los circuitos integrados del ambiente y permite la conexión de los conectores metálicos. Los microprocesadores completos (Chips) son entonces probados para asegurar su operatividad. Una vez pasan la prueba, los chips pueden almacenarse en la bodega de bienes terminados. El modelo propuesto representa el proceso de manufactura por una cadena de suministros de 3 etapas consistentes en fabricación; ensamble y distribución (Fig. 1).

Además, la producción de microprocesadores toma lugar en un sistema de producción hibrido push-pull, combinando un sistema push en las etapas río arriba y un sistema pull en las etapas río abajo. Por esto la fabricación es caracterizada por un sistema de producción push: pronósticos de demanda a largo plazo, actualizaciones semanalmente, y ajustes de fabricación y ensamble de WIP que sirven como base para la tasa de producción deseada de wafers o de iniciar los wafers. En contraste, las pruebas de ensamble y la distribución operan como sistemas pull, es decir, con envíos basados

en las órdenes actuales de los clientes.

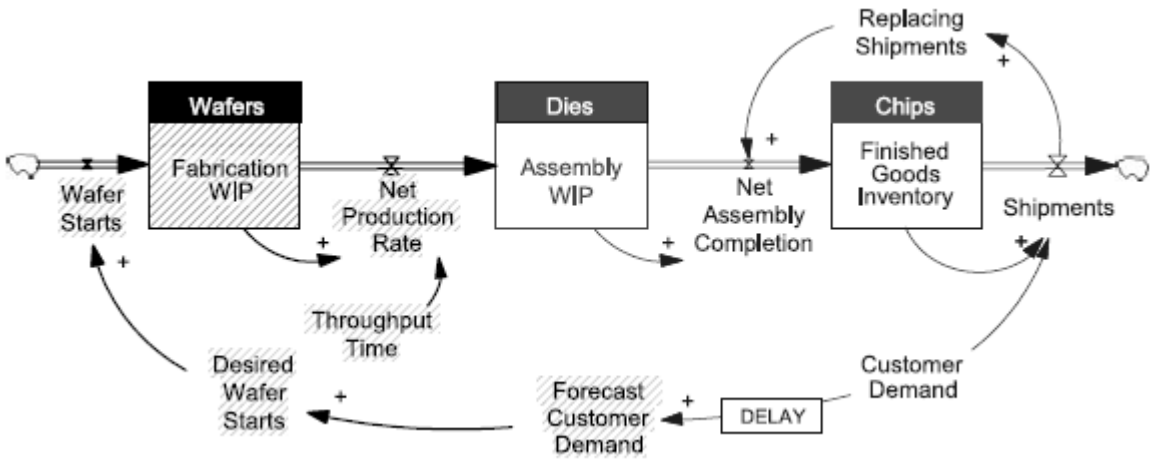

Figura 1. Sistema híbrido de producción push-pull para semiconductores.

Los cuatro principales supuestos están basados en el manejo del trabajo de campo del comportamiento del modelo. Las primeras tres suposiciones conducen decisiones de administradores relacionadas con:

- Producción push y pull
- Utilización de capacidad.
- Pronósticos de demanda.

Estas suposiciones reflejan el racionamiento heurístico de los administradores y gerentes de Intel para controlar sus sistemas. Mientras no son óptimos, ellos reflejan el uso de administradores heurísticos para tomar sus decisiones del día a día y se desarrollan por que están locamente adaptados para las condiciones en la compañía y sus fábricas. La cuarta suposición captura la respuesta del cliente a la disponibilidad de inventario.

a. Decisiones de producción push y pull.

Las decisiones de producción últimamente dependen de la demanda del cliente. Actualmente la demanda conduce a envíos y ensambles completados; los pronósticos de demanda a largo plazo influyen en el inicio de la producción. Todas las órdenes entrantes son registradas por el sistema de información de Intel y seguidas hasta ser entregadas o

canceladas por los clientes. Si los microprocesadores están disponibles en el inventario de bienes terminados (FGI) las órdenes se completan inmediatamente. Por consiguiente, las órdenes entrantes de los clientes halan los microprocesadores disponibles del FGI . A su vez, reemplazar el FGI entregado al cliente "hala" microprocesadores de ensamble. Si el microprocesador no está disponible en FGI , las órdenes retrasadas halan partes directamente desde el ADI (inventario de ensamble de dados). Como las partes deben ser ensambladas, llenar las órdenes desde el ADI incrementa la espera de la entrega para el cliente y reduce el flujo de embarques, por debajo de las órdenes del cliente como el inventario en ADI y el ensamble limitan la capacidad de envíos.

- FGI *y ensamble pull*: Para modelar las características pull de ensamble y de bienes terminados, debemos capturar su dependencia de la demanda actual. Las envíos actuales desde el $FGI(s)$ son dadas por el mínimo del deseado (pull) y la posible (push) tasa de envío. Por diseño, las entregas operan en modo pull con entregas determinadas por la tasa deseada; sin embargo si no hay suficiente FGI disponible el sistema entrega sólo lo que es posible.

$$S(t) = \min(S^*(t), S_{\max}(t)) \tag{1}$$

$S^*(t)$: las entregas deseadas.

Las entregas deseadas dependen de la proporción del retraso $B(t)$ y la demora de entrega deseada DD^* . Las entregas posibles dependen del inventario de $FGI(t)$ y el mínimo tiempo de procesamiento de una orden τ_{OP} ; un proceso de primer orden se asume por simplicidad.

$$S^*(t) = \frac{B(t)}{DD^*} \tag{2}$$

$$S_{\max}(t) = \frac{FGI(t)}{\tau_{OP}} \tag{3}$$

Mientras las entregas $S(t)$ agoten el FGI , la tasa neta de completación de ensambles A_G las llenara. El producto de la tasa gruesa de terminación $A_G(t)$ y el rendimiento unitario Y_u es decir la fracción de chips buenos por ensamble terminado, definen la tasa neta de completar los ensambles A_G . La tasa de completar el

ensamble bruto $A_G(t)$ está dado por la mínima deseado (señal de pull) o el posible (señal de push) de la tasa de completar el ensamble bruto.

$$FGI(t) = Y_u A_G(t) - S(t) \tag{4}$$

$$A_G(t) = \min(push.A_G(t), pull.A_G(t)) \tag{5}$$

Por diseño, el ensamble opera en modo pull, con una salida bruta de ensamble al ser determinada por la tasa bruta deseada. Sin embargo, si no hay suficiente $WIP(t)$ disponible el sistema puede completarlo solo cuando sea posible. La tasa de ensamble bruta posibles es determinada por la disponibilidad de $WIP(t)$ de ensamble $AWIP(t)$ y el tiempo de ensamble τ_A; por simplicidad se usa un retraso de primer orden. La tasa bruta de ensamble deseada $A_G^* = pull.A_G$ está determinada por la tasa neta de ensamble deseada $A_N^*(t)$ ajustada por el rendimiento unitario Y_u.

$$pull.A_G(t) = \frac{A_N^*(t)}{Y_u} \tag{6}$$

$$push.A_G(t) = \frac{AWIP(t)}{\tau_A} \tag{7}$$

La determinación de la tasa de ensamble neta deseada $A_N^*(t)$ por los planeadores de la división, comienza con las entregas recientes (ES), determinada para la demanda actual de manera más estable y confiable que las órdenes. Los chips netos deseados salen entonces ajustados o según las entregas recientes para cerrar cualquier diferencia entre la meta y el actual FGI y para eliminar el exceso de backlog.

$$A_N^*(t) = \max(0, ES(t) + \frac{FGI^*(T) - FGI(T)}{\tau_{FGI}} + \frac{B(t) - B^*(t)}{\tau_B} \tag{8}$$

Donde FGI^* y FGI son la meta y el inventario actual de bienes terminados, B^* y B que son la meta y el backlog actual, y τ_{FGI} y τ_B son los tiempos ajustados para la eliminación de las diferencias entre ellos.

- *La fabricación pull.* La producción wafers en el proceso de fabricación son empujadas al inventario del dado del ensamble ADI

donde ellos son almacenados hasta ordenarlos para halar los productos específicos desde ADI dentro del ensamble y distribución. Mientras que la tasa bruta del ensamble de terminación (A_G) agota $AWIP(t)$, la tasa de terminación del dado D_I se reabastece.

$$AWIP(t) = D_I(t) - A_G(t) \tag{9}$$

La tasa de terminación del dado D_I medida en dados/mes, es dado por la tasa bruta de fabricación F_G, medida de wafer/mes, ajustada por el numero de dados por wafer DPW el rendimiento del dado Y_D es decir la fracción de dados buenos por wafer y el rendimiento de la línea Y_L es decir la fracción de wafers bien fabricados. La tasa bruta de fabricación F_G se determina por la disponibilidad de fabricación $WIP(FWIP)$ y el tiempo de fabricación τ_F, por simplicidad se toma retraso de primer orden y un tiempo constante de fabricación es usado.

$$D_I(t) = F_G(t) \cdot DPW \cdot Y_D \cdot Y_L \tag{10}$$

$$F_G(t) = \frac{FWIP(t)}{\tau_F} \tag{11}$$

Mientras la tasa de fabricación bruta F_G agota la fabricación de $WIP(FWIP)$, los wafers iniciales WS deben ser remplazados. La decisión en la tasa de producción actual, los wafers iniciales WS, son directamente basados en los wafers iniciales deseados WS^*.

$$FWIP(t) = WS(t) - F_G(t) \tag{12}$$

Los planeadores de la fábrica determinan los wafers iniciales deseados considerando el flujo de entrada deseado de dados D_I^* requeridos por las plantas de ensamble/test y ajustado para el work-in-process en fabricación. Lo anterior está basado en heurísticos de los administradores de mantener la fabricación $WIP(FWIP)$ en el nivel deseado $FWIP^*$. La siguiente ecuación muestra el heurístico de los planeadores de fabricación para manejar los Waters iniciales.

$$WS^*(t) = \max(0, \frac{D_I^*(t)}{DPW \cdot Y_D \cdot Y_L} + \frac{FWIP^*(t) - FWIP(t)}{\tau_{FWIP}} \tag{13}$$

donde τ_{FWIP} es el tiempo de corrección del WIP de fabricación y la

constante de no negatividad previene metas negativas de fabricación.

La decisión de la tasa de los dados entrantes (D_I^*) depende de los pronósticos de demanda a largo plazo (ED) y un ajuste del WIP de ensamble. Los componentes de el ajuste del WIP de ensamble reflejan el objetivo de los planeadores de ensamble de reemplazar el WIP de ensamble cuando el nivel actual esta por debajo de la meta para corregir la discrepancia de sobre el tiempo τ_{AWIP}. La ecuación 14 muestra el heurístico de los planeadores de la división para manejar el (D_I^*) incorporando información de pronóstico de demanda (ED). Los planeadores de división proveen información del flujo deseado de dados entrantes a los planeadores de fabricación permitiéndoles programar los inicios de producción

$$D_I^*(t) = \max\left(0, \frac{ED(t)}{Y_U} + \frac{AWIP^*(t) - AWIP(t)}{\tau_{AWIP}}\right) \tag{14}$$

donde τ_{AWIP} es el tiempo de corrección del WIP de ensamble; y la constante de no negatividad previene tasas negativas de la entrada de dados.

b. *Utilización de capacidad*

Para fijar la utilización de la capacidad (CU) de sus fábricas, los administradores consideraron la tasa deseada de producción y la capacidad disponible. La utilización de capacidad es una función no lineal del cociente de los inicios deseados de wafers (WS^*) y la capacidad disponible (K) operando a un nivel de utilización de capacidad normal (CU_N)

$$CU(t) = f_U\left(\frac{WS^*(t)}{K \cdot CU_N}\right) \tag{15}$$

Donde la producción deseada (WS^*) iguala la capacidad normal utilizada, la utilización de la capacidad es fijada al punto de operación normal (90%), permitiendo alcanzar toda la producción deseada. El restante 10% de capacidad es a menudo usado para propósitos de ingeniería (corridas de mejoramiento de procesos y desarrollos) como también para acomodar la inestabilidad de la

manufactura. Cuando la producción deseada es más grande que la capacidad normalmente utilizada, los directores de la fábrica incrementan la utilización, reduciendo la capacidad que esta disponible para ingeniería. Lo contrario toma lugar cuando la producción deseada cae por debajo de la capacidad normalmente utilizada. Si la función cae de la línea de referencia de 45°, la utilización será muy suficiente para asegurar que los inicios de wafers igualarán siempre lo deseado (sujeto a las limitaciones de capacidad). El estudio de campo muestra sin embargo, que la función de utilización caracterizando las decisiones actuales de los wafers iniciales se muestra sobre la línea de referencia de 45° y se aplana en el punto de operación normal. Los directores de fábrica prefieren evitar parar y mantenerla operando incluso cuando los inicios deseados se bajan de lo normal, prefiriendo aumentar el inventario; similarmente, ellos incrementan las mínimas salidas para que sean suficientes para conocer completamente las decisiones iniciales cuando lo deseado inicialmente excede las salidas normales esto para mantener cierto espacio para propósitos de ingeniería y evitar problemas con el rendimiento. Una función cóncava donde $f_U \geq 0$, $f'_U > 0$, $f''_U < 0$, $f_{U1}(0) = 0$, $f_U(1) = CU_{Norm}$, $f_U(2) = CU_{Max}$ captura la respuesta de los directores de la fábrica a variaciones en los wafers iniciales deseados relativos a la capacidad.

Mientras la forma general de la función (f_U) es plausible, la pendiente de la función alrededor del punto de operación normal y la fracción de utilización normal juega un papel importante en el comportamiento del modelo. Datos para estimar tales parámetros son especificaciones propias y específicas de la fábrica. Por tanto, se proporciona un análisis de sensibilidad sobre un amplio rango de parámetros plausibles para la función de utilización de la capacidad e investigar el impacto de esos parámetros en el comportamiento del modelo.

c. Pronostico de demanda

La organización de marketing es responsable de los pronósticos de demanda en Intel. Como en muchas firmas, el marketing genera un pronóstico inicial para los microprocesadores basándose en estimaciones de la demanda del cliente, de diferentes regiones geográficas y tipos de cliente (para otro ejemplo en la industria de semiconductores ver Sterman 2000 pp 449-462).

Un proceso conocido como "Judged Demand" o demanda estimada es usado para ir del pronóstico inicial al final. El proceso de "Judged Demand" recibe su nombre debido a la estimación y ajustes subjetivos envueltos en la elaboración del pronóstico. Primero indicadores de macroeconomía se incorporan para adoptar las estimaciones iníciales basado en el mercado total disponible para computadores personales y de negocios. Segundo un proceso de orden ejecutivo "ejecutive order" frecuentemente ajusta el pronóstico agregado hacia arriba para reflejar los objetivos optimistas y aspiraciones de los ejecutivos de la compañía. Finalmente el grupo de marketing "filtra" (suaviza) las estimaciones de demanda de diferentes regiones para contar los incentivos locales. Con respecto a la información regional, de acuerdo con la plataforma de administración en marketing: "los números de clientes se recogen, son agregados y juzgados con un grupo de suposiciones que pueden ser o no correctas; y la comprensión de niveles de clientes, cuando se provee, se recibe aguas abajo". En particular, cuando la demanda de un producto especifico es alta, los directores de bodegas regionales tienden a incrementar sus órdenes para asegurar que podrán cumplir la demanda, generando las familiares "órdenes fantasmas" (phantom ordering) cuando diferentes clientes compiten por porciones más grandes de lo que ellos perciben será un pastel reducido (Forrester 1961, Sterman 2000 pp. 743-755; Gonçalves 2003). En contraste, cuando la demanda es baja, los directores regionales tienen la tendencia a decrecer las órdenes para aseguran no estar atrapados en acumular inventarios no deseados. Por eso, marketing filtra el pronóstico para surgir el pronóstico final. Análisis de los datos del pronóstico confirman esto: Cuando se comparan pronósticos regionales, son mas variables estos que los de marketing. En Intel, el pronóstico de demanda incorpora un componente de tendencia para contar el crecimiento exponencial en las ventas de semiconductores.

Debido a que nos enfocamos en la interrelación entre la respuesta del cliente y la inestabilidad de la cadena de suministro, exploramos la señal sin tendencia de la demanda. Por esto, modelamos el pronóstico de demanda (ED) como suavización exponencial de primer orden de las ordenes actuales (D) – en la práctica obtenida de las agregaciones de órdenes regionales- puestas en el periodo de un mes (τ_{DAdj}), la frecuencia con la que marketing actualiza su pronóstico.

$$ED(t) = \frac{D(t) - ED(t)}{\tau_{DAdj}} \quad (16)$$

Por simplicidad, no tomamos en consideración la factores macroeconómicos aleatorios que influencian los pronósticos de demanda y proceso ejecutivo de adicionar, haciendo la suposición de que marketing puede filtrar el ruido causado en el proceso.

d. Respuesta del cliente

En este modelo capturamos la respuesta del cliente a la disponibilidad de la entrega, medida por la fracción de órdenes cumplidas (FoF). Los clientes responden a una fracción de bajo cumplimiento mirando otras fuentes de abastecimiento; cuando esto sucede sus órdenes caen. El atractivo de Intel para los distribuidores (A_I) es una función no lineal de la percepción de los clientes a la confiabilidad de entrega del suministro ($PFoF$). En respuesta, a la percepción del cliente ajustada a la confiabilidad de la entrega del suministro (FoF) del nivel actual de confiabilidad de entrega – Fracción de cumplimiento de las ordenes (FoF) - con un retrasado de tercer orden Erlang (λ) con un tiempo promedio constante de 6 meses. La distribución Erlang de tercer orden captura la distribución aparentemente buena de respuestas del OEM_S. En el instante de un decrecimiento de nivel de servicio, todos las OEM_S seguirán percibiendo al suministrador como confiables y no habrá búsquedas de recursos alternativos del suministro. Por esto, la respuesta inmediata del retraso del distribuidor debe ser cero. Si el nivel de servicio sigue bajo, algunos clientes cambiaran su percepción y buscaran otros proveedores confiables. La distribución de OEM_S reacciona eventualmente en un pico y luego decrece, alcanzando cero después que un tiempo suficiente ha pasado. La demora captura el tiempo requerido por OEM_S para percibir cambios en la disponibilidad, para determinar que el cambio no es temporal y garantiza búsqueda de otros proveedores y cerrar tratos con ellos.

Por simplicidad asumiremos que los competidores permanecen constantes en su habilidad de entrega (por ejemplo, un atractivo constante A_C sobre el tiempo) esta suposición permite medir cambios en el comportamiento del sistema debido a reacciones de clientes

solo por cambios en condiciones del proveedor: esta relajación es prominente dirección de trabajos futuros.

La función no lineal (F_O) caracterizando el atractivo de Intel (A_I) es una curva logística (la figura 9 provee un ejemplo) el atractivo varia en una escala de 0 a 1. $(0 \le A_{LMin} < A_{LMax} \le 1)$. Una curva logística captura respuestas suaves del cliente a cambios pequeños en la disponibilidad de entrega, y una respuesta más importante a cambios grandes.

$$A_I(t) = f_A(PFoF(t)) \tag{17}$$

Mientras que la forma logística de la función es creíble, el comportamiento del modelo depende fuertemente de la pendiente de la función y el valor mínimo. Al mismo tiempo, los datos para estimar los parámetros no son fácilmente confiables y fáciles de conseguir. También se brindará un análisis de sensibilidad sobre un rango de parámetros plausibles de la función que gobierna la respuesta del cliente e investiga el impacto de esos parámetros en el comportamiento del modelo.

e. *Estructura de realimentación.*

La heurística de los directores de Intel (producción pull-push, utilización de capacidad, pronóstico de demanda, y las decisiones de la respuesta al cliente) cierra el ciclo de realimentación mostrado en Figura 2.

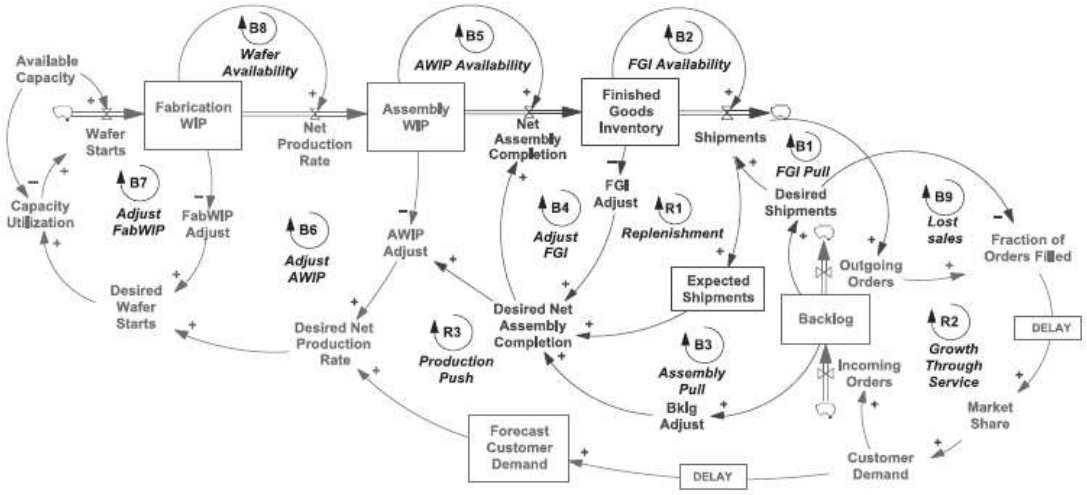

Figura 2. Proceso de realimentación del suministro de demanda para Intel en un sistema de producción híbrido.

El ciclo balanceado $FGI.pull(B1)$ describe el sistema de operación pull de la compañía en los niveles de inventarios de bienes terminados FGI. Un incremento en órdenes pendientes (debido a órdenes adicionales) incrementa los envíos deseados, fomentando los envíos y reduciendo las órdenes pendientes – si hay suficiente FGI. Cuando la disponibilidad es limitada de FGI, el ciclo negativo de la disponibilidad de FGI $(B2)$ limita los envíos. Con FGI limitando los envíos, el sistema pull no puede operar en el nivel de FGI. Sin embargo, el sistema puede aun halar (pull) desde el WIP de ensamble. El ciclo de ensamble pull (Assembly pull) $(B3)$ es análogo para el ciclo $FGI.pull$, pero la salida de chips halados del WIP de ensamble, se demoran mas. Por esto, cuando el FGI es limitado, el sistema sigue operando como pull pero con mas retraso en el cumplimiento. La tasa actual de ensambles completados es también ajustada con dos ciclos: un ciclo que corrige con los niveles de inventario de producto terminado (Ajuste FGI - $(B4)$) y un ciclo reforzador (Reabastecimiento $-(R1)$) que reemplaza todos los envíos desde el FGI. El sistema puede halar desde el ensamble tanto como el inventario de WIP es suficientemente grande. Cuando la disponibilidad del WIP decrece el primer orden para controlar el ensamble – ciclo $AWIP$ disponible $(B5)$ – previene que el $AWIP$ se vuelva negativo. Si el ensamble WIP neto limita el sistema no puede seguir halando de $AWIP$ y todo el sistema se vuelve un sistema push.

Producción en la etapa de aguas arriba es basada en los pronósticos de demanda de largo plazo, información de la rata deseada de finalización del ensamble y ajustes debidos para el inventario (Ajuste $AWIP$ $(B6)$) y fabricación (Ajuste $FWIP$ $(B7)$).

La parte de empujar (push) el sistema de producción es determinada por el control de primer orden para fabricación – ciclo *Wafer Availability* $(B8)$. En términos de respuesta del cliente, el ciclo refuerza *Growth Through Service* $(R2)$ describe la habilidad de la compañía de incrementar su parte del mercado mientras es capaz de cumplir la demanda. En contraste, el ciclo balanceador – *Lost Sales* $(B9)$ - describe la dinámica inversa. Mientras la demanda de los clientes crece, la habilidad de la compañía de mantener el nivel de servicio (fracción de órdenes satisfechas) decrece, reduciendo su habilidad de retener los clientes. Si la compañía no puede satisfacer

adecuadamente las órdenes del cliente, perderá mercado frente a los competidores. Finalmente, la realimentación de la cadena de suministros de la compañía para la demanda de los clientes se describe en el ciclo reforzador, *Producción push* $(R3)$, que captura la larga espera asociada con producción y reacciones de clientes. Si la demanda cae, manufactura reduce el pronóstico y la utilización de la capacidad para evitar excesos de inventario. Después de la demora de producción, la producción más baja deja inventario bajo y nivel de servicio, causando caída adicional en la demanda. Este proceso de realimentación es capaz de la generación del comportamiento dinámico observado en la compañía y replicarlo en el modelo.

5. ANÁLISIS DEL MODELO Y RESULTADOS

El modelo constituye un sistema de ecuación diferencial no lineal de noveno orden. Dado que la ecuación es altamente no lineal, no es posible obtener soluciones muy cerca de su forma. Por lo tanto, se realizó una simulación del modelo para obtener intuitivamente una solución. Aunque los parámetros escogidos para el caso base (tabla 1) refleja el sistema de producción Intel, estos se cambian para mantener la confidencialidad.

Parámetro	Definición	Valor	Unidades
D	Demanda del cliente	5	Millón de unidades/mes
MS	Segmento que participa en el mercado inicial	80	%
DPW	Número de dados por wafer	200	Dados/wafer
CUN	Utilización normal de la capacidad	90	%
YL	Rendimiento de la línea: Fracción de dados buenos por el total	90	%
YD	Rendimiento de dados: Fracción de dados buenos por wafer	90	%
YU	Rendimiento de unidad: Fracción de chips buenos por dado bueno	95	%
K	Capacidad disponible	28,9	miles de wafers/mes

Para una demanda de cliente (D) dada, la capacidad de equilibrio (k) que se requiere para cumplir la demanda se puede computar desde la utilización normal de capacidad y rendimientos. La formula es: $k = \frac{D \cdot MS}{CU_N \cdot DPW \cdot Y_D \cdot Y_L \cdot Y_U}$.

La Figura 3 muestra el comportamiento de registros y FGI convergentes para 3 corridas de simulación. El modelo es inicializado en equilibrio con la demanda constante de la industria. Como se menciona antes, mientras la demanda de semiconductores ha crecido exponencialmente por décadas, el trabajo se enfocará en una señal de demanda sin tendencia porque nos interesamos sólo en los factores que respectan con la estabilidad del sistema. En equilibrio el sistema de producción hibrido (pull-push) funciona como lo propuesto: el proveedor consigue su meta de retraso de entrega, llena el 100% de las órdenes entrantes, y mantiene cantidades deseadas de FGI, AWIP y Wafers. Desde el equilibrio introducimos un pulso en la demanda incrementando un 5% y después un 20% respectivamente para un mes al final del primer año simulado. La demanda incrementa los registros (Figura 3a), ellos rápidamente matan la necesidad de incrementar la producción y el número de inicios de Wafers deseados aumentan (Figura 3b). Como la capacidad es fijada en un corto plazo, los administradores deben aumentar la utilización de capacidad (Figura 3c) para incrementar los inicios de Wafer. Los administradores de la fábrica rápidamente ajustan la utilización a su máximo. Una utilización mas alta incrementa el nivel de fabricación (Figura 3d). Después de los retrasos en fabricación y ensamble, eventualmente los productos terminados llegan a estar disponible para cumplir la demanda.

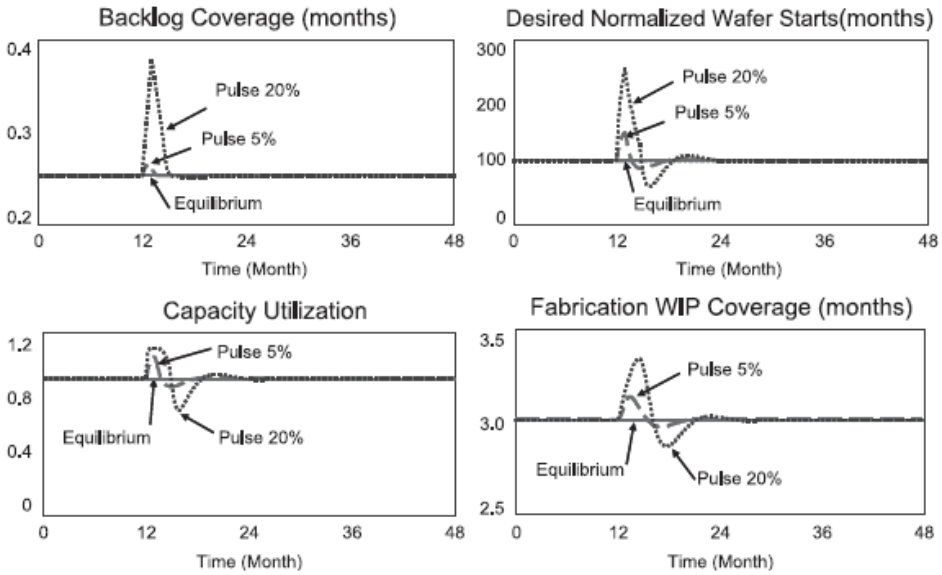

Figura 3. Retrasos cubiertos (3a), Inicialización deseada de wafers (3b), Utilización de la capacidad (3c), Fabricación de WIP cubierta (3d).

En ambos casos el sistema responde inmediatamente a la onda en los registros incrementando las entregas (que no se muestran) y halando más chips desde FGI. En el caso del incremento del 5%, el agotamiento en FGI (Fig 4b) es insuficiente para afectar los registros. Mientras el choque de la demanda crea algo de inestabilidad en la cadena de suministros (Fig 3b y 4a), el stock de seguridad en FGI y AWIP permiten al sistema operar como se desea (como un hibrido push-pull). A pesar del inventario, la compañía es capaz de conseguir su meta de entrega y de cumplir el 100% de sus órdenes entrantes (Figura 4c).

Sin embargo, el impacto del 20% produce algo diferente. El inventario de seguridad no es capaz de mantener el sistema en su modo de operación deseada. En este punto el sistema se comporta como un sistema push puro, reaccionando a cambios en la demanda con demoras de producción más largos. El desabastecimiento en FGI limita las entregas y la fracción de órdenes cumplidas (Figura 4c) entonces el sistema no puede operar con el nivel de FGI. El sistema compensa con lo que hay de FGI, sin embargo, halla el chip del WIP de ensamble incrementando la tasa o velocidad de ensamble. Conforme la disponibilidad de WIP de ensamble decrece, eventualmente limita el ensamble. Ahora, el sistema no puede halar del WIP y se transforma en un sistema de empujar (push).

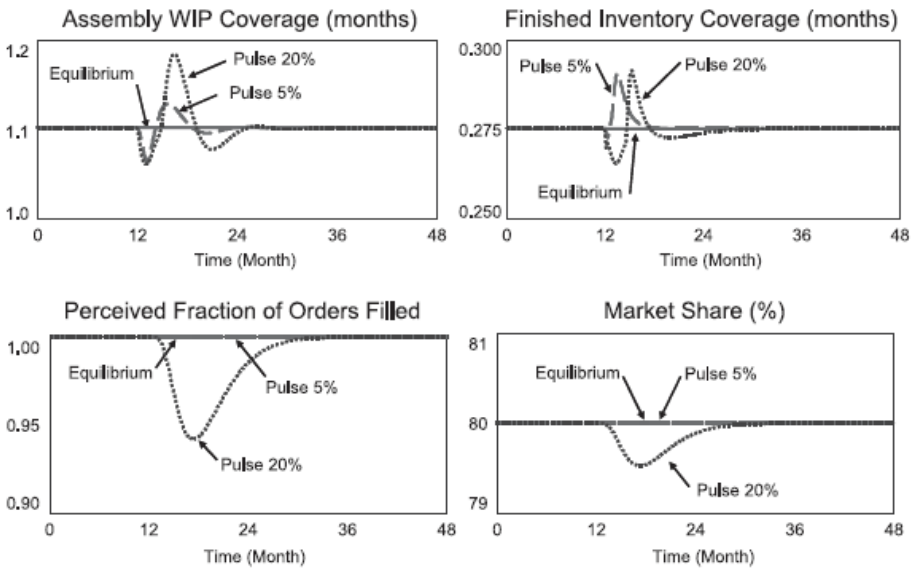

Figura 4. Ensamble de WIP (4a), Inventario terminado cubierto (4b), Percepción de la fracción de órdenes completadas (4c), Participación del mercado (4d).

En modo de empujar, el distribuidor esta incapacitado para cumplir todas las órdenes de los consumidores. Los clientes (OEMs) perciben la caída en el nivel de servicio (Fig 4c), después de un lapso (contando la toma de dediciones y el tiempo de reporte en los sistemas de información) y busca otras alternativas de proveedores, la inhabilidad de su compañía para cumplir la demanda resulta en reducción del mercado compensando el incremento inicial de la demanda. Como la demanda sigue decreciendo, eventualmente iguala el volumen de entregas permitiendo el cumplimiento de los registros (Fig. 3a) para frenar el incremento y la fracción de órdenes llenadas paren de declinarse. Incluso cuando FGI adicional esta disponible, el mercado sigue disminuyendo hasta la percepción del cliente. La respuesta de los clientes a la disponibilidad de inventario retroalimenta las decisiones de producción, La utilización de la capacidad (Fig 3c) sube mientras el distribuidor reacciona a la declinación de la demanda. El decremento en utilización disminuye el nivel de fabricación de WIP de ensamble, y la cobertura de FGI. Cuando los clientes finalmente perciben la mejora en el desempeño de la compañía, incrementan las órdenes y el sector del mercado. Con tiempo, el incremento en órdenes sobrepasa las entregas, permitiendo un incremento en los registros. Una vez mas las entregas no son suficientes para cumplir la demanda y la fracción de órdenes satisfechas decrece. El aumento

de 20% en la demanda genera una respuesta oscilatoria que deja algunos excesos de demanda perdida y el distribuidor cierra cualquier vacío de demanda con la utilización de la capacidad por encima del nivel normal, a diferencia del pulso del 5%, cuando el sistema esta sujeto a un stock lo suficientemente largo, la interacción de las decisiones racionales de la firma sobre entregas y utilización de capacidad con la repuesta del mercado o disponibilidad de producto resulta en oscilaciones ligeramente amortiguadas, deprimiendo el mercado de la firma.

a. Impacto del sistema pull.

Para obtener información adicional de las causas de la oscilación, primero consideramos como los lazos de balanceo de FGI pull (B1) y ensamble pull (B3) influencian el comportamiento del sistema. El halar desde FGI permite a la compañía cerrar los vacíos entre las entregas deseadas y las entregas actuales ejecutando los registros. Naturalmente este lazo solo puede operar mientras hay suficiente FGI para permitir las entregas. La habilidad del lazo FGI pull para operar tan eficientemente se da por un corto tiempo (1 semana) asociado con la tasa deseada de entregas. Cuando este lazo está apagado – los envíos son función del nivel de FGI- El sistema opera como un sistema push y el comportamiento oscilatorio del sistema incrementa (Figura 5). Similarmente el ciclo de Assembly pull hala inventario desde ensamble para permitir al lazo FGI pull operar como desea. Este lazo solo puede operar cuando hay suficiente AWIP. Para permitir que los ensambles se completen. Esto también tiene una constante de tiempo corta (1 mes) determinada por el tiempo para ajustar los registros. Si no se puede halar del ensamble se reduce la estabilidad del sistema al restringir el lazo pull de FGI para operar eficientemente. La Figura 5 muestra el efecto de apagar el lazo de FGI y el de ensamble pull, comparados con el caso base.

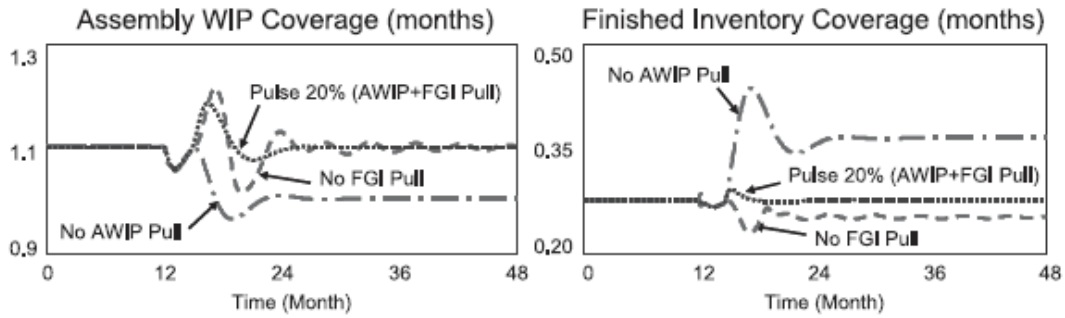

Figura 5. Ciclos pull de FGI y AWIP apagados comparados con el 20% del pulso de la demanda

b. Impacto de la demanda endógena.

El impacto de la demanda endógena en el sistema ofrece más profundización en las causas de oscilación. Para hacer demanda endógena, se debe derivar el ciclo de la respuesta del cliente dejando el tiempo de percepción de la fracción de órdenes llenadas en un numero muy grande ($T_{FOF} = \infty$). Este cambio rompe la realimentación desde la fracción de órdenes satisfechas (*FoF*) a la demanda del cliente (*D*), haciendo la demanda exógena mientras se reduce el sistema a la producción y la respuesta de registros a un cambio de demanda. El lazo de producción establece "inicio de Wafers" basados en la demanda esperada y ajustes de inventario en fabricación, ensamble, y productos terminados son completamente visibles a los directores y ellos usan la misma constante de tiempo para ajuste de inventario. El sistema puede ser reducido a un sistema efectivamente de primer orden. El comportamiento será entonces un incremento suave en producción para alcanzar el pulso adicional de demanda seguido de una disminución también suave.

Le estructura del ciclo de producción sin embargo es diferente. Primero mientras el sistema tiene un comportamiento completamente visible, los productos terminados (FGI) solo son usados para establecer el nivel deseado de inventario de ensambles, en lugar de ser también usados para establecer la tasa deseada de salida de ensamble. Segundo, la demanda esperada, suavizada con la larga constante de tiempo, es usada para establecer la producción, pero las entregas esperadas, suavizadas con una constante de tiempo corta, son usadas para informar los ajustes necesarios para el inventario deseado. Finalmente los ciclos de FGI pull y Assembly pull introducen complejidad adicional al proceso de producción. Mientras las

entregas esperadas ajustan el FGI influencian también demanda esperada, utilización de capacidad. Por lo tanto una caída de entregas debido a la baja cantidad de FGI disponibles envía una señal falsa a la producción ya que las salidas adicionales no son exactamente necesarias cuando se desea lo opuesto. El comportamiento resultante es una oscilación amortiguada de la producción en la cadena de suministro (Figura 6). Incrementando el ajuste de tiempo para correcciones de inventario o suavizando la demanda esperada sobre un largo tiempo constante, podemos amortiguar las oscilaciones.

La interacción de la respuesta del cliente con el resto del sistema amplifica el comportamiento oscilatorio de la producción. Mientras la demanda incrementa la compañía incrementa la producción y es inmediatamente hábil para cumplir cuando la demanda decrece. Con el tiempo, los clientes perciben que los niveles de servicio están decreciendo. Después de un retraso en la manufactura los bienes terminados adicionales permiten a la compañía cumplir con una mayor fracción de demanda que la que cumpliría en otros casos. Mientras mas bienes terminados llegan a estar disponibles, los retrasos de respuesta para los clientes reducen las órdenes. Cuando el productor se encuentra con mayor cantidad en inventario de FGI y demanda reducida, los administradores de la fábrica reducen la utilización de la capacidad, limitando la habilidad de la compañía de cumplir con la demanda futura. El sistema eventualmente converge, sin embargo, la interacción de la respuesta del cliente y producción amplifica la inestabilidad de la cadena de suministro.

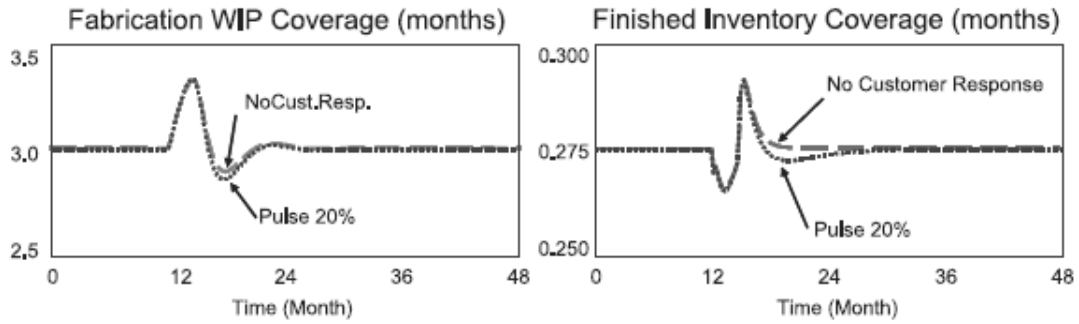

Figura 6. Respuesta de la producción a un cambio en la demanda compara con un 20% de pulso en la demanda.

Mientras la demanda endógena amplifica dicha inestabilidad, mas importantemente afecta las políticas de la compañía que tienen que ver con la utilización de la capacidad y el control de inventarios, desde que la reacción del sistema a un cambio en la demanda es más estable cuando esta se vuelve exógena, una política de inventario firme, con niveles reducidos de inventarios o de seguridad, es también capaz de proveer un nivel de servicio alto, sin incurrir en costos extras de inventario. Por esto si la demanda se asume exógena, una política de inventario estrecha llevará a bajar costos y se preferirá a una política de inventario de seguridad. Sin embargo, un sistema mas inestable, como el caso de la demanda endógena requiere mas buffers de inventario para proveer el mismo nivel de servicios. En este caso los costos asociados con ventas perdidas deben compensar los costos de mantener inventario. Por esto, los buffers de inventario se preferirán cuando la demanda es asumida endógena.

Ahora se considera la política de utilización de capacidad. Cuando la demanda es endógena, la disponibilidad de inventario afecta la demanda. Los agotamientos de inventario que restringen las entregas decrementan el nivel de servicio y la demanda, enviando una señal clara de que no se necesita más salidas. Desde que la disminución de la demanda es causada por un agotamiento de inventario, las salidas adicionales son altamente deseadas. Una política de utilización de capacidad que no responde porqué no disminuye el nivel de producción cuando la demanda decrece, proveerá un nivel de servicio más alto cuando la demanda es endógena. En contraste, cuando la demanda es exógena, la disponibilidad de inventario no afecta la demanda. Una política de capacidad que responde permite a la compañía prevenir acumulaciones de excesos de inventario durante periodos de demanda baja. Entonces, cuando la demanda se asume exógena se recomienda políticas de utilización de capacidad de respuesta.

Para explorar el impacto de las suposiciones de demanda en inventario y utilización consideramos un estructura simple de costo, donde el costo total (Tc) es la suma del costo de mantener inventario en ensamble WIP (*Hc Awip*) y bienes terminados (*Hc FGI*) y todos los costos de ventas perdidas (LSc). (Por simplicidad, no contamos con los costos de fabricación) Los costos de mantenimiento en cada etapa están dados por el producto del volumen de inventario en cada etapa (*wip* y *FGI*) y el costo unitario respectivo de mantener inventario

($\beta\phi$ y $\delta\phi$, eso es, una fracción del costo unitario de bienes terminados ϕ). El costo de las ventas perdidas es el producto de un factor (ϕ) del costo unitario de bienes terminados y la suma de ventas perdidas, dada por la diferencia entre el nivel de mercado inicial (*MSo*) y el actual (*MSt*)

$$H_{CAWIP} = Awip \times \beta \times \phi \tag{18}$$

$$H_{CFGI} = FGI \times \delta \times \phi \tag{19}$$

$$LS_C = (MS_O - MS_T) \times \alpha \times \phi \tag{20}$$

El criterio para evaluar la mejor política es la comparación del valor presente neto de los costos descontados acumulativos (CDC) como una tasa de descuento (r).

$$CDC = \int_0^\infty e^{-rT} \left\{ \left[(MS_O - MS_T) \times \alpha + Awip \times \beta + FGI \times \delta \right] \times \phi \right\} dT \tag{21}$$

La Figura 7 muestra que la prescripción política de adoptar inventario ajustado y las políticas de utilización de respuesta están en efecto reservadas cuando la demanda es endógena. Figura 7 (a,b) muestra el valor presente neto de los costos asociados con la demanda endógena y exógena para cada política de inventario. Cuando la demanda es endógena una política de inventario de seguridad deja costos más bajos que el inventario ajustado, sugiriendo que los costos de ventas perdidas contrapesan los costos de mantenimiento. Desde que mantener inventario previene pérdidas de ventas entonces tasas de descuentos mas altas cargan los costos de mantenimiento, de ese modo reduciendo los beneficios asociados con bajar las ventas perdidas. La figura 7(c,d) muestra el valor presente neto del costo asociado con demanda endógena y exógena bajo las políticas de utilización de capacidad. Una política de utilización sin respuesta permite a la compañía construir inventario durante periodos de baja demanda, causada por disponibilidad pobre de inventario, cuando la demanda es endógena. Ya que los beneficios de una política que no responde bien a su construcción de inventario, si la compañía no adopta una política de inventario de seguridad los beneficios de una política de respuesta se podrán reducir. Así pues, buffers de inventario y políticas de utilización de no respuesta rinde menos costos con demanda endógena.

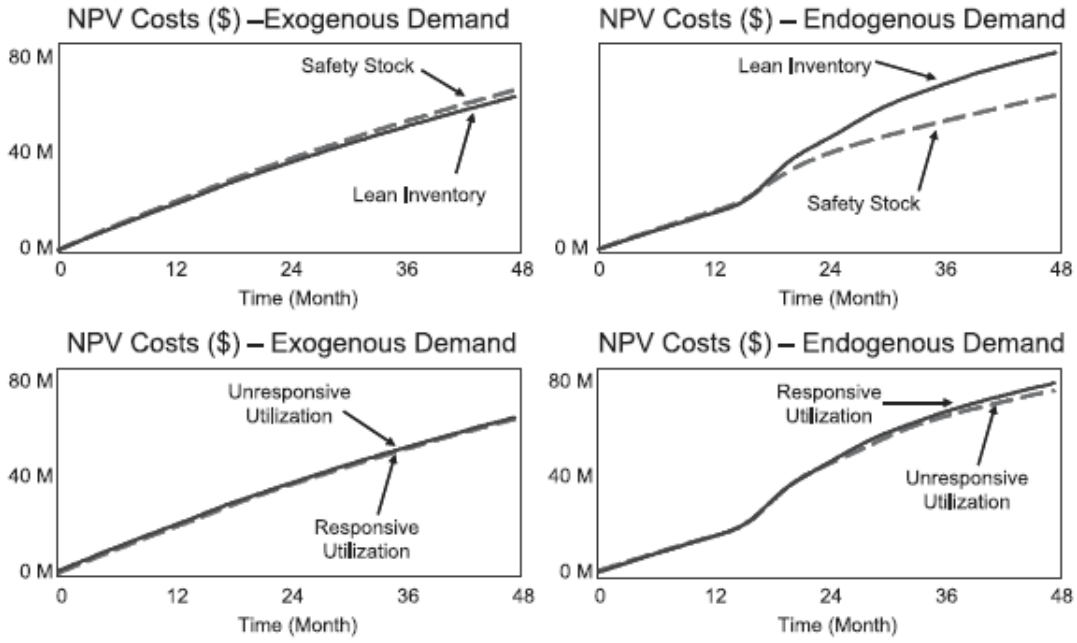

Figura 7. Impacto de la demanda endógena y exógena en el inventario y en la utilización de políticas.

La conclusión anterior refleja el comportamiento del sistema del conjunto original de costos. Después, se cambia la relación entre costos de mantenimiento y costos de ventas perdidas para explorar el impacto que debe tener en nuestras conclusiones. El modelo se simula 2,500 veces con parámetros independientes seleccionados aleatoriamente de una distribución uniforme con rangos especificados en la tabla 2.

Parámetro	Símbolo	Unidades	Mínimo	Base	Máximo
Costo de participación en la fracción unitaria de FGI	δ	1/mes	0,005	0,01	0,2
Razón de una fracción unitaria de AWIP al costo de FGI	β/δ	dmnl	0,125	0,25	0,75
Fracción unitaria del costo de las ventas perdidas	a	dmnl	0,5	1	5

Tabla 2. Rango de valores para parámetros de costos diferentes. Nota: la simulación se corre con un costo unitario de bien terminado = 50 $/unid y una rata de descuento r=0.01/mes; dmnl, significa adimensional.

La tabla 3 presenta la media, mediana, desviación estándar e inventario de confianza (50%, 90%, 95%) para el valor presente neto de los costos de descuento acumulados para políticas de utilización e inventario cuando los clientes responden o no a la disponibilidad de inventario. Los estadísticos son evaluados al final de la simulación (en un tiempo t=48 meses). (Ver tabla 3)

La conclusión anterior se mantiene para un rango de costos de mantenimiento y costos de ventas perdidas. Adaptando inventario ajustado y las políticas de utilización de respuesta permite minimizar costos cuando la demanda se asume exógena. Sin embargo, cuando ese no es el caso, los buffers de inventario de seguridad y las políticas de utilización de la capacidad que no responde logran bajar los costos. El mejor potencial de ahorro toma lugar cuando la demanda se asume endógena. Cuando las políticas se prueban independientemente ahorros de 26% vienen de adoptar políticas de inventario de seguridad y 5% viene de la política de utilización sin respuesta. Cuando se explora la combinación efectiva de políticas de inventario y utilización sin embargo, se verifica que las políticas de respuesta y las que no responden llegan a probar resultados similares. Esto es que hubo un balance entre los costos adicionales de inventario obtenidos con políticas que no responden y el beneficio de la reducción en los costos de la pérdida de las ventas. Por lo tanto, cuando se asume la demanda endógena se recomienda una política de seguridad debido a su alto potencial de ahorro.

Costo NPV (millones $): demanda exógena

Políticas	Media	Mediana	SD	50% CI	90% CI	95% CI	100% CI	Ahorros promedios
Inventario de seguridad (SS)	533	477	374	224,783	29,1241	21,1387	14.2,1548	
Inventario Lean (LI)	511	457	358	215,75	28,1189	20,1323	13.6,1483	4.2%
Ulilización de respuestas (RU)	511	457	358	214,75	28,1187	20,1321	13.7,1481	0.2%
Utilización de no respuestas (UU)	512	458	359	215,752	28,1192	20,1326	13.7,1487	

Costo NPV (millones $): demanda endógena

Políticas	Media	Mediana	SD	50% CI	90% CI	95% CI	100% CI	Ahorros promedios
Inventario de seguridad (SS)	602	549	374	294,851	97,1317	68,1429	27,167	26.3%
Inventario Lean (LI)	790	745	383	499,1043	234,1499	170,1617	65,1987	
Ulilización de respuestas (RU)	769	722	378	480,1019	224,1469	161,1589	61,1947	
Utilización de no respuestas (UU)	731	683	374	437,979	200,1431	143,1549	54,1884	5.4%

Tabla 3. Políticas de utilización e inventario en las salidas para diferentes parámetros del costo.

c. Análisis de Sensibilidad

El comportamiento del modelo es altamente sensible a las suposiciones incluidas en las funciones de utilización de la capacidad (f_U) y respuesta del cliente (f_A). En particular, el modelo es sensible a (1) las pendientes de las funciones no lineales f_U y f_A, (2) la utilización máxima de la capacidad, y (3) el mínimo de la respuesta de la demanda del cliente. Él análisis de sensibilidad sigue un procedimiento común para obtener estos resultados. Se representa cada función no lineal (utilización de la capacidad y respuesta al cliente) como una combinación lineal de dos casos polares, capturando suposiciones extremas. Variando el peso en la combinación lineal es posible obtener un rango de comportamientos en el modelo.

Sensibilidad a la utilización de la capacidad: Considere dos tipos extremos de gerentes de Fab que reaccionan a la producción deseada: un gerente que responde (obediente) y uno que no responde (insensible). Ambos responden diferentemente para un volumen por debajo de la producción deseada. Cuando la producción deseada es baja, el gerente que no responde, caracterizado por la función (f_{U1}), prefiere guardar el funcionamiento de Fab y acumular los niveles de inventario por debajo de la cadena en vez de retardar la línea o cerrarla. En el límite, por supuesto, la utilización debe bajar a cero mientras que la producción deseada baja a cero. La política que no responde se muestra en la figura 8; la utilización tiene una pendiente plana cerca de la región de funcionamiento normal. En contraste, un gerente que responde, caracterizado por la función (f_{U2}), responde agresivamente a las disminuciones de la producción deseada cortando la utilización en proporción a la declinación en la producción deseada, evitando el aumento de inventario y haciendo la capacidad innecesaria disponible para la mejora del proceso, funcionamientos de prueba o mantenimiento preventivo (figura 8). Los casos intermedios se obtienen de la combinación lineal de los dos extremos; en el caso que sigue $w_1=0.5$.

$$CU = w_1 \cdot f_{U1} + (1 - w_1) \cdot f_{U2}; w_1 \in [0,1] \quad (22)$$

La figura 8 muestra parte del mercado para diferentes especificaciones de la función de utilización de la capacidad. La variabilidad del sistema se incrementa con las respuestas de los

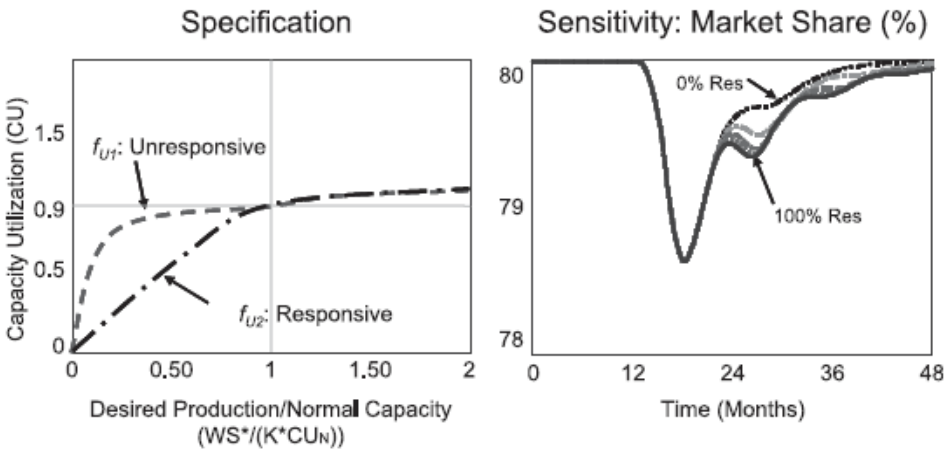

Figura 8. Parte del mercado que responde a las especificaciones de la utilización de la capacidad.

gerentes a los cambios en producción deseada. Parece intuitivo que más respuestas deban reforzar la estabilidad previniendo la acumulación de exceso del inventario durante los períodos de demanda baja. El otro lado se observa, sin embargo, porque la demanda es endógena: corta la producción agresivamente cuando se percibe una baja demanda, la empresa asegura que ese inventario incluso estará menos disponible, manejando partes de mercados con bajas tendencias. La política que no responde empuja el producto dentro de la cadena de suministros, mejorando la disponibilidad y trayendo a los clientes de regreso a la empresa más rápido. Además, retrasos más cortos en la respuesta al cliente y pronósticos acentuados del Impacto de las respuestas en una parte del mercado. Por consiguiente, los canales de distribución con clientes que responden (Ej., ventas online) son particularmente vulnerables a la política de utilización de capacidad que responde.

Sensibilidad a la respuesta al cliente: Ahora considere los 2 casos extremos de respuesta al cliente. Un cliente insensible como base, caracterizado por la función (f_{A1}), que no responde a los cambios en el nivel de servicio percibido. La función de respuesta del cliente insensible opera alrededor del punto (1,1) donde es plana (pendiente es cero). Este caso extremo corta la realimentación desde el nivel de servicio de los suministradores hasta la demanda del cliente. En contraste, un cliente sensible como base, caracterizado por la función (f_{A2}), responde agresivamente a los cambios en el servicio

percibido. La pendiente de la función de respuesta al cliente sensible alrededor del punto de operación es alta. Los niveles de servicio percibidos lo suficientemente bajos pueden reducir lo atractivo de los productos al nivel mínimo posible. Una función de respuesta del cliente se obtiene desde la combinación lineal de 2 casos polares (f_{A1} y f_{A2}); en el caso base $w_2 = 0.5$.

$$CR = w_2 \cdot f_{A1} + (1 - w_2) \cdot f_{A2}; w_2 \in [0,1] \quad (23)$$

La figura 9 muestra el resultado para diferentes grados de respuesta al cliente. La variabilidad del sistema se incrementa en tanto los clientes sean más sensibles a la disponibilidad del producto. Este resultado es el esperado. Cuando la demanda es exógena (por fuera, alrededor), la curva de la producción opera independientemente. Agregando la respuesta al cliente que equilibra la curva, notamos que las oscilaciones en producción aumentan. También es sensato esperar que una base del cliente más sensible, reaccione con un retraso de la percepción de la disponibilidad del inventario, esto introducirá más variabilidad en la demanda y por consiguiente en la producción. La sensibilidad de la respuesta al cliente muestra que la inestabilidad de la cadena de suministros con demanda exógena es mas pequeña que la inestabilidad con demanda endógena. Capturando la realimentación de la disponibilidad del producto con la demanda del cliente se amplifica la inestabilidad de la cadena de suministros. Estos resultados sugieren que los modelos que adoptan una demanda exógena subestiman la inestabilidad en la cadena de suministros, la cual profundiza en las reglas asociadas a la política.

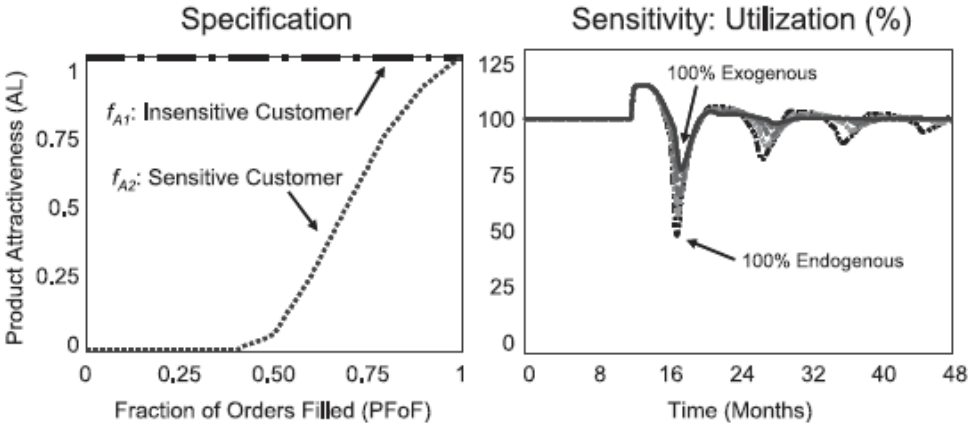

Figura 9. Utilización (% de CU_N) de la sensibilidad a clientes que responden a especificaciones

Sobre "el impacto del sistema pull", se aprende que los ciclos del ensamblaje y FGI son capaces de estabilizar el sistema, pero sólo toman lugar cuando hay suficiente inventario disponible. Por consiguiente, una política importante para mejorar la estabilidad del sistema es mantener stocks de seguridad en ambos (Ensamble y FGI).

d. Análisis de la ubicación del óptimo del stock de seguridad

El nivel deseado de stock de seguridad en ensamble y producto terminado se basa en un control heurístico que optimiza el trade - off entre el costo del inventarío en espera y el costo de ventas perdidas. Usando la misma estructura del costo definida anteriormente ("Impacto de la demanda endógena"), el criterio para evaluar el óptimo nivel de inventario de seguridad es la minimización del valor presente neto de los costos descontados acumulados (CDC por sus siglas en inglés), con una tasa de descuento (r). La fracción de volumen de inventario destinada como inventario de seguridad en cada etapa es dada por un porcentaje (p_{AWIP},p_{FGI}) del volumen inicial en el etapa (AWIP, FGI). El nivel óptimo de inventario de seguridad es el valor del porcentaje del inventario de seguridad de cada etapa que minimiza el valor presente neto de CDC sobre el periodo simulado. La demanda se especifica como la suma del nivel actual y un ruido de auto correlación (Pink) con una desviación estándar (σ) del 5% (representativo para la variación de la demanda en Intel).

Investigamos el volumen óptimo para el inventario de seguridad para 4 diferentes niveles de costos de ventas perdidas (α) y 4 tasas por unidad de costos de inventario en espera en ensamble y productos terminados (β/δ). La figura 10 muestra el resultado de la optimización para (a) el porcentaje del volumen total en ensamble (p_{AWIP}) y (b) el porcentaje del volumen total en producto terminado (p_{FGI}). Como se esperaba, el nivel óptimo de inventario de seguridad en ambas etapas se incrementa con el costo de las ventas perdidas (α). Así mismo, la asignación de "inventario de seguridad entre ensamble y producto terminado es alta dependiendo de la tasa de los costos en espera en las 2 etapas. Las tasas bajas en los costos de espera benefician la asignación del inventario de seguridad en ensamble. Además, para grandes tasas de costo de espera, el fabricante espera más inventario de seguridad en FGI que el aumento en los costos de las ventas perdidas. Este resultado tiene sentido en cuanto a los mas altos stocks de seguridad en FGI que permiten a la compañía hallar la demanda con un tiempo de respuesta mas corto.

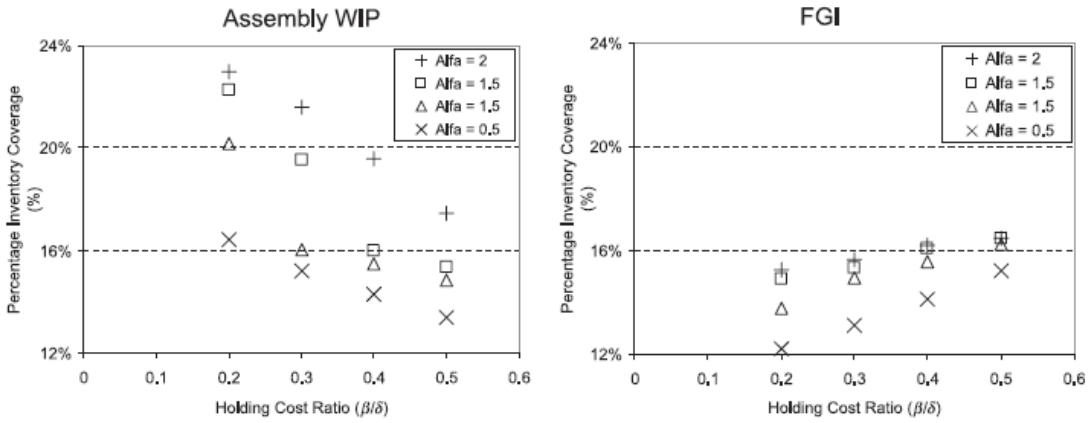

Figura 10. Porcentaje óptimo del inventario de seguridad (a) Ensamble (b) FGI.

Además como el tiempo del throughput en ensamble (τ_A) es mucho mas largo que en producto terminado (τ_{OP}), el mismo porcentaje de volumen de stock de seguridad en ensamble y PT será trasladado a un fondo de inventario de seguridad (respaldo) mas alto en ensamble. Por ejemplo, cuando los tiempos del throughput son de 4 semanas en ensamble (τ_A) comparados con 1 semana en PT (τ_{OP}), un 15% del stock de seguridad en ensamble y en PT se traslada al fondo de stock de seguridad con 0.6 semanas en ensamble y 0.15 semanas en PT. Aquí, la misma inversión del dólar en stock de seguridad produce un alto respaldo de inventario para el ensamble. Este resultado intuitivo, tiene un impacto directo en la parte del mercado que la compañía puede retener y la inestabilidad resultante de la cadena del suministro. Comparando el impacto de la misma cantidad de dólares en el stock de seguridad, el más alto de los fondos de inventario en ensamble tiene un efecto estabilizador en la variabilidad de la cadena de abastecimientos, el cual produce retrasos más cortos, menos órdenes expedidas y más clientes satisfechos. Mientras que contando con la intuición de que el stock de seguridad debe ser ubicado en FGI, debido al incremento de la demanda sensible, el acto se mejora guardando el stock de seguridad en el ensamble. Además, el heurístico de mantener stock de seguridad adicional en el ensamble puede ayudar a estabilizar la operación del sistema y mejorar el nivel de servicio. La figura 11 compara la ejecución del 5% de la política del stock de seguridad adicional en ensamble WIP y FGI a partir de que empieza a correr el caso.

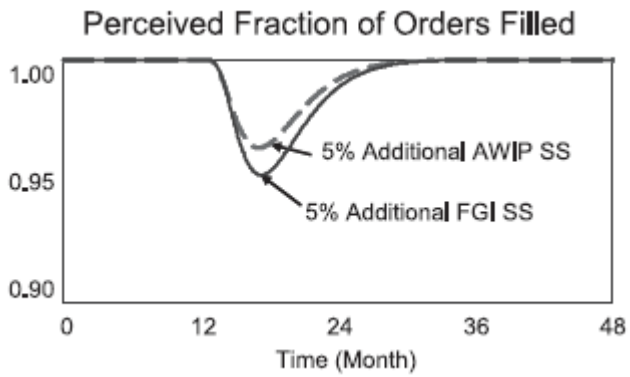

Figura 11. Comparación con el mismo rendimiento de existencias de seguridad en inversiones de AWIP y FGI.

Los costos de inventario en espera se moderan fácilmente, son muy visibles y precisos. En contraste, el costo de las ventas perdidas son de difícil acceso, y la variabilidad de la demanda es fácilmente explicada a lo lejos como el resultado de los factores fuera de la empresa. De esta manera, es probable que los gerentes subestimen el costo de la ejecución, de una entrega pobre (por ejemplo: ventas perdidas, pérdida de la reputación, nuevos clientes, etc.). Si el fabricante subestima el costo de las ventas perdidas, también subvalorara el stock de seguridad óptimo que se debe mantener en ensamble y en productos terminados. Además, no solo se esperaran pequeños inventarios a costos más altos de lo necesario, alterando la dinámica del sistema al reducir la estabilidad de la cadena de abastecimiento. Esta inestabilidad probablemente tiene muchos otros costos que no se tienen en cuenta. Otros costos operacionales incluyen longitudes de recorrido más corto y lotes pequeños, set-ups más frecuentes y cambios de referencia (changeovers), altos errores y tasas de reproceso y más tiempo ocioso entre lotes y entre set-ups.

6. DISCUSIÓN Y DIRECCIONES PARA INVESTIGACIONES FUTURAS

Este paper esta dirigido al comportamiento oscilatorio en la utilización de capacidad de un fabricante semiconductor y el rol de la demanda endógena del cliente influyendo en la producción de la compañía y en el nivel de servicio. El esfuerzo modelado se dibujó en un campo de trabajo extenso, incluyendo observación directa, colecciones de datos y archivos, y entrevistas estructuradas y semi-estructuradas con gerentes de Intel. El paper contribuye a la comprensión del rol que tiene la respuesta del cliente en la amplificación del incremento de la demanda sobre la cadena de

abastecimiento explorando los mecanismos a través de los cuales la demanda endógena del cliente interactúa con los heurísticos de producción de los gerentes. La interacción entre las ventas y efectos de producción amplifica grandemente el comportamiento oscilatorio de la producción. Este resultado sugiere que los modelos que adoptan una demanda exógena subestiman la inestabilidad en las cadenas de abastecimiento, las cuales quebrantan las reglas asociadas a las políticas. Las investigaciones muestran que capturando la realimentación de la disponibilidad del producto con la demanda del cliente se amplifica la inestabilidad de la cadena de abastecimientos, finalmente se invierten las reglas de la política tradicional acerca del inventario y las políticas de utilización de la capacidad. La simulación sugiere que las reglas típicas de la política del inventario pobre y las políticas de utilización sensibles solo se mantienen cuando la demanda se asume exógena. Con demanda endógena, el buffer del inventario y una política de utilización de la capacidad insensible producen reducción de costos.

El análisis del ciclo de knockout sugiere que equilibrando los ciclos de FGI Pull (B1) y Ensamble Pull (B3) se estabiliza el sistema. Sin embargo, ellos solo pueden operar eficientemente cuando hay suficiente inventario disponible, el fabricante debe mantener buffers de inventario más grandes en ensamble y en producto terminado. Mientras el heurístico de la conservación de buffer de inventario en ensamble y en PT para mejorar la potencia del sistema no es nuevo, esta investigación recalca su importancia en la operación de los sistemas híbridos Push-Pull. Además, el análisis de la política soporta una cantidad de stocks de seguridad que se incrementan con el costo de las ventas perdidas y una asignación dependiente de la tasa de costos de espera entre el ensamble y el PT. Por eso, el fabricante puede reducir la inestabilidad de la cadena de abastecimiento efectivamente y reducir el impacto en las ventas perdidas, tanto como reduzca el costo de las ventas perdidas internamente.

En general, fabricantes de semiconductores, así como las empresas en otras industrias, tienden guardar bajos niveles de inventario y ejecutar cadenas e suministros delgadas, permitiéndoles reducir costos de inventario. Esta práctica presenta fabricantes con estrategia para evitar costos asociados a la obsolescencia del inventario en industrias con ciclos de vida del producto cortos. El modelo mental que motiva inventarios delgados asume que la

variabilidad de la demanda es exógena: en un mundo con cambios en la demanda impredecibles, costoso FGI y rápida obsolescencia tecnológica, guardando inventarios delgados se minimiza el riesgo que puede tener la empresa con exceso de inventario si la demanda cae inesperadamente. Sin embargo, la demanda no es exógena— la disponibilidad del producto afecta la demanda, la cual entonces retroalimenta la disponibilidad. Bajos inventarios de PT y PP incrementa la oportunidad de salida de inventarios (stockouts) en diferentes etapas en la cadena de abastecimiento, influyendo en la posibilidad de que el sistema opere de manera indeseada (por ejemplo un sistema push). Considerando la administración heurística del típico inventario adoptada por las compañías, como el ajuste constante del nivel de inventario deseado que refleja las señales de la demanda actual, y el incremento potencial en la variabilidad de la demanda introducida por la respuesta del cliente, se nota que las compañías pueden subestimar los verdaderos costos asociados con salidas de inventario (stockouts).

Por otra parte, los heurísticos de los gerentes, del ajuste de la utilización de la capacidad para responder a la variabilidad de la demanda— causada por la inhabilidad de los proveedores para satisfacer al cliente— puede amplificar la variabilidad de la demanda. Como la demanda es endógena, cortando la producción agresivamente cuando se percibe que la demanda va a ser baja, la empresa garantiza que el inventario estará incluso menos disponible, manejando partes del mercado con tendencia bajas. Aquí, la fuerza de los proveedores para encontrar la demanda del cliente en un corto tiempo puede herir el servicio al cliente a la larga. En contraste, la política insensible empuja el producto dentro de la cadena de abastecimiento, mejorando la disponibilidad y causando el regreso del cliente a la empresa más rápido. Sin embargo, si la compañía adopta una política de stock de seguridad, las políticas de utilización sensible e insensible producen resultados similares. Esto quiere decir que el inventario adicional resultado de una política insensible reduce los costos de ventas perdidas en una cantidad similar a la del aumento de los costos de espera.

Hay numerosas oportunidades para las investigaciones futuras motivadas por este estudio. Primero, hay muchas otras industrias (ej., automóviles, electrónicos) donde los efectos reportados aquí pueden jugar también un papel importante. Segundo, el resumen del estudio

actualmente está lejos de introducir nuevos productos en el tiempo y los modelos característicos de la demanda durante el ciclo de vida del producto donde a menudo hay escasez al inicio durante el aumento de la producción, seguido por un declive temporal en la demanda final de la vida del producto. Unos niveles de stock de seguridad óptimo y heurísticos de producción pueden cambiar en el transcurso del ciclo de vida. Además, el modelo incorpora sólo la respuesta de los clientes, debido al actual nivel de servicio (por ejemplo, la fiabilidad del suministro). Sin embargo, si los clientes constantemente tienen una experiencia de entrega con fiabilidad pobre pueden optar por hacer de otras empresas sus principales proveedores, reduciendo las ventas de otros productos y para productos futuros, y también el aumento de la variabilidad de la demanda (Risch et al. 1995). Las órdenes de cancelaciones también pueden ser añadidas al modelo. En caso de que las cancelaciones se produzcan como resultado de una disminución de nivel de servicio, es probable que ampliar el efecto causado por pérdida de ventas fortalecería los resultados presentados aquí. Además, el modelo actual no incorpora la posibilidad de la inflación de pedidos por los clientes, creando fantasmas o burbujas de la demanda, cuando múltiples OEM's cobertura contra la escasez de la oferta (Gonçalves, 2003, Sterman 2000, Cap. 18,3). La demanda Phantom es importante, ya que es probable que equilibre efectos de la pérdida de ventas y contrarreste los efectos observados en la investigación.

Notas

a. El número real de dados por wafer está en un rango de 100 a 1000, según el tamaño en el chip, que varía con la arquitectura-si es el chip "Lógica" o "memoria" y su diseño específico. Cada uno se compone de los dispositivos individuales, tales como transistores y células de memoria.

b. Una descripción completa del modelo, las formulaciones, y las hipótesis pueden ser encuentran en Gonçalves (2003).

c. El proceso de fabricación, representado como una sola población, los agregados de un complejo, multi-paso conjunto de actividades dentro de la Fábrica. Si bien el mayor nivel de agregación es coherente con el propósito de este trabajo, muchas interesantes prácticas de investigación y las cuestiones

relacionadas con la optimización del flujo de los diferentes SKU a través de un Fab requeriría mayor desglose.

d. Mientras que la heurística de la tasa de flujo deseado morir no tiene en consideración el ajuste de FGI explícitamente, la información es de FGI utiliza para configurar el montaje WIP (AWIP*).

e. Se asume que el nivel normal de utilización de la capacidad en Intel es igual al 90% de capacidad máxima.

f. Es sencillo agregar ruido aleatorio al pronóstico de captar el impacto de estas fuentes de error y ajuste.

g. En la práctica, la integración de los límites (de 0 a ∞) Van de 0 a un límite (pero grande) para terminar el tiempo de simulación. El tiempo final se ha seleccionado para garantizar que el descuento ha reducido la contribución al resto de NPV de manera significativa.

h. Las políticas de inventarios seguridad se ponen a prueba con el 0% y el 5% de la seguridad stock en AWIP y FGI, respectivamente.

i. La comparación de las políticas de la utilización de la capacidad se realizan con 0% stocks de seguridad en AWIP y FGI para que el impacto de la política de utilización sea más destacados.

j. Usamos auto-correlación aleatoria para la demanda para estudiar la eficacia del los stocks de seguridad bajo una señal de la demanda mas realista. El mismo número aleatorio se utiliza para todas las simulaciones, permitiendo que cada optimización corra usando exactamente las mismas realizaciones de los proceso aleatorios.

k. Asumimos también que los parámetros de los costos unitarios de los productos terminados (Φ) = 50 $/unidad, y una tasa de descuento (r) = 0.01/mes.

l. Para obtener una cantidad comparable a la misma cantidad de stocks de seguridad de FGI y AWIP, se establece la relación entre costes de espera (β/δ) a 0,25, que compensa exactamente por la tasa de cobertura de inventario entre el dos etapas.