The Impact of Endogenous Demand on Push-Pull Production Systems: Supplement

0. Overview

This supplement provides a detailed description and supporting documentation for a system dynamics model used to analyze the impact of endogenous demand on a semiconductor manufacturer hybrid push-pull production system. The model is based on the experience with Intel Corporation, a major semiconductor manufacturer located in Santa Clara California. This supplement should be used in conjunction with Gonçalves, Hines and Sterman (2005). The paper provides motivation for the study, background information about the research, a few main formulas, model results, analyses and policy recommendations. This supplement provides a detailed description of the formulas used in the model as well as the rationale for their use. Additional materials available on the web site include the working model, command scripts, optimization controls and graph definitions used to generate the results in the paper.

The model structure consists of two major flows: the flow of materials through the semiconductor company's supply chain and the flow of information and managers' decision rules, governing material flows. Materials flow thought the company's supply chain from wafer starts through assembly according to the manufacturing process. Information flows control the flow of materials (e.g. wafer starts, assembly starts, assembly completion rate, and shipments) through the company's supply chain.

The model is run in continuous time and implemented in Vensim® and formulated as a system of nonlinear ordinary differential equations. The first section provides an overview of semiconductor manufacturing. Section 2 provides a description of the production and inventory control and section 3 describes distribution and logistics.

1. Manufacturing Process

Semiconductor manufacturing is commonly divided into two major phases: fabrication and assembly. The first phase (fabrication) takes place in a wafer fabrication facility, or Fab. It takes 200 mm/300 mm polished disk-shaped silicon substrates (wafers) as inputs and through a complicated sequence of steps transforms them into fabricated wafers, composed of hundreds of

-inch square integrated circuits (ICs or dies).[1] A vertical cross-section of an integrated circuit reveals a number of layers formed during the fabrication process. Lower layers, produced at the "front-end" of the fabrication process, include the critical electrical components (e.g., transistors, capacitors). Upper layers, produced at the "back-end" of the fabrication process, connect the electrical components to form circuits. In addition, fabrication is characterized by a re-entrant flow process, i.e., the same equipment perform multiple steps at different stages of fabrication (e.g., photolithography, etching, thin films, diffusion, ion implantation).

In the second phase of manufacturing (assembly) wafers are cut into dies and stored in Assembly Die Inventory (ADI) warehouses, collocated with Assembly/Test plants. The dies receive a protective package protecting the integrated circuit from the environment and allowing the attachment of metal connectors. The completed microprocessors (or chips) are then tested to ensure operability. Upon passing the tests, the chips can be stored in finished goods warehouses. The model proposed represents the manufacturing process by a three stage supply chain consisting of fabrication, assembly, and distribution (Figure 1).

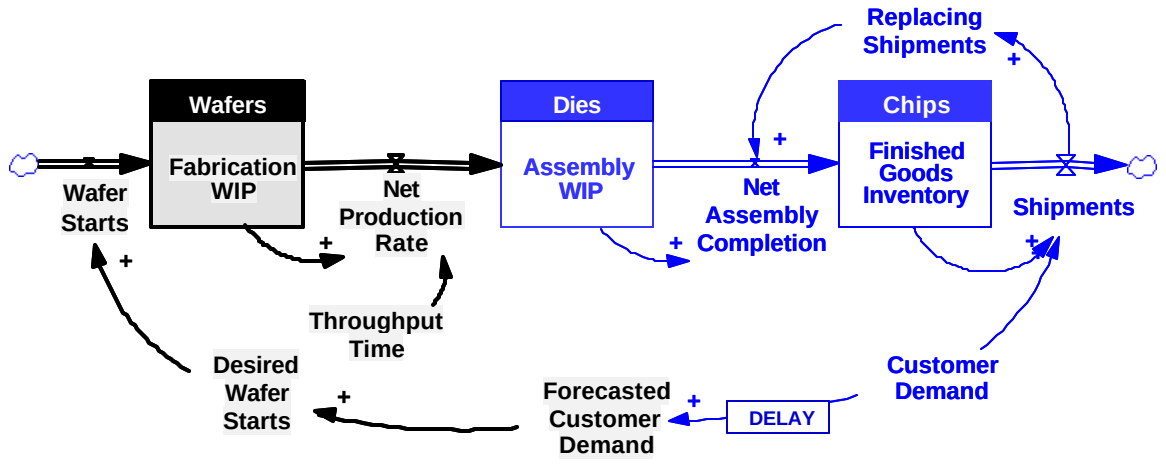

Figure 1 –Hybrid push/pull production system for semiconductors.

Thick lines and patterned background refer to a push system, indicating that the upstream production process operates as a push. Thin lines and clear background refer to a pull system, indicating that the downstream stages operate as a pull.

[1] The actual number of dies per wafer range from 100 to 1000, depending on the chip size which varies with its architecture – whether the chip is "logic" or "memory" – and its specific design. Each die is composed of individual devices such as transistors and memory cells.

In addition, microprocessor production takes place in a hybrid push-pull production system, combining a push system at upstream stages and a pull system at the downstream stages. Therefore, fabrication is characterized by a push production system: long-term demand forecasts, updated weekly, and adjustments from fabrication and assembly WIP serve as the basis for the desired wafer production rate, or wafer starts. In contrast, assembly/testing and distribution operate as a pull system, with shipments based on current customer orders. Since not all assembled chips may be in tune with customer demand, manufacturers assemble and test only those chips that are adequate for market consumption. Orders for specific products pull die from ADI into assembly/testing. The assembled products can either be shipped directly to customers to meet demand or simply be used to adjust the finished goods inventory to desired levels. Therefore, semiconductor manufacturers operate a hybrid push/pull system, starting production based on long-term forecasts and assembly based on customer demand.

2. Production and Inventory Control

This section describes the hybrid push-pull production process of a semiconductor manufacturer. The description characterizes first the push fabrication process and then it explores the pull assembly process.

2.1. Production Push

The number of wafers pushed into production is determined by the wafer starts rate (*WS*), which is given by the product of capacity utilization (CU) and available capacity (*K*). Hence, when production managers receive requests to increase Fab output, they can boost wafer starts by either increasing capacity or capacity utilization. Since it takes a long time to add new capacity, however, in the short run production managers can only accommodate increases in wafer starts by changing capacity utilization. For the purpose of this model, we assume that available capacity is fixed and it is set at a value just above to the desired production start rate. This assumption captures the manufacturer's policy to run the factory as close as possible to maximum capacity and make the best use of capital investment. In addition, this assumption does not change the dynamic behavior of the model in a significant way. In fact, all it does is to require a stronger exogenous shock to drive the system to the observed behavior.

To set the capacity utilization (*CU*) of their Fabs, managers consider the desired production rate and the available capacity. Capacity utilization is a nonlinear function of the ratio of desired wafer starts (WS^*) and available capacity (*K*) operating at the normal capacity utilization level (CU_N).[2]

$$WS(t) = CU(t) \cdot K \tag{1}$$

$$CU(t) = f_U\left(\frac{WS^*(t)}{K \cdot CU_N}\right) \tag{2}$$

When desired production (WS^*) equals the normal capacity utilized, capacity utilization is set at the normal operating point (90%), allowing all desired production to be met. The remaining 10% slack capacity is often used for engineering purposes (process improvement and development runs) as well as to accommodate manufacturing instability. When desired production is large relative to normal capacity utilized, Fab managers increase utilization, therefore reducing the capacity that is available to engineering. The opposite takes place when desired production falls below normal capacity utilization. If the function lay on the 45^o reference line, utilization would vary enough to ensure that wafer starts always equaled desired starts exactly (subject to the capacity constraint). Field study showed, however, that the utilization function characterizing actual wafer start decisions lies above the 45^o reference line and has a flatter slope at the normal operating point. Fab managers seek to avoid shutdown and prefer to keep their Fab running even when desired starts fall below normal, preferring instead to build inventory; similarly, they increase output less than enough to meet desired starts fully when desired starts exceed normal output so as to maintain some room for engineering purposes and to avoid yield problems. A concave function where $f_U \geq 0$, $f_U' > 0$, $f_U'' < 0$, $f_{U1}(0) = 0$, $f_U(1) = CU_{Norm}$, $f_U(2) = CU_{Max}$, captures the response of Fab managers to variations in desired wafer starts relative to capacity.

The fabrication work-in-process (*FWIP*) is increased by production starts and decreased by the good wafers outflow, the net fabrication outflow (F_N) and rejected fabrication (F_R). Gross fabrication rate (F_G) is obtained by producing the fabrication work-in-process (*FWIP*) over the manufacturing cycle time (t_F). The line yield (Y_L) determines the fraction of gross wafer outflow,

[2] We assume the normal capacity utilization level at Intel to be equal to 90% of maximum capacity.

or gross fabrication (F_G), that is good for assembly. We assume that bad production is rejected without rework.

$$\dot{FWIP}(t) = WS(t) - F_N(t) - F_R(t) \quad (3)$$

$$F_G(t) = FWIP(t)/t_F \quad (4)$$

$$F_N(t) = F_G(t) \cdot Y_L \quad (5)$$

$$F_R(t) = F_G(t) \cdot (1 - Y_L) \quad (6)$$

The desired wafer starts (WS^*) – a metric managers use to determine actual starts – is given by the sum of desired gross wafer starts (desired gross fabrication F_G^*), and a term for fabrication WIP adjustment ($FWIPAdj$), constrained to be non-negative. The fabrication WIP adjustment term reflects the firm's willingness to produce more (less) when fabrication WIP is below (above) the desired level, to correct the discrepancy over time (t_{FWIP}). Managers set the desired level of fabrication WIP ($FWIP^*$) in order to produce the average gross wafer outflow rate over the manufacturing cycle time (t_F).

$$WS^*(t) = MAX(0, F_G^*(t) + FWIPAdj(t)) \quad (7)$$

$$FWIP^*(t) = F_G^*(t) \cdot t_F \quad (8)$$

$$FWIPAdj(t) = \frac{FWIP^*(t) - FWIP(t)}{t_{FWIP}} \quad (9)$$

The desired gross fabrication (F_G^*), i.e., the desired gross wafer starts, is determined by the desired net wafer start rate (WS_N^*) adjusted by losses in the production line, the line yield. In turn, desired net production rate is determined by the desired die inflow (D_I^*) in assembly adjusted by the number of dies per wafer (DPW) and the die yield (Y_D). Where the former variable determines the number of die can be obtained from each wafer and the latter determines the fraction of good die per wafer.

$$F_G^*(t) = WS_N^*(t) / Y_L \quad (10)$$

$$WS_N^*(t) = \frac{D_I^*(t)}{DPW \cdot Y_D} \quad (11)$$

Rewriting equation (7) in terms of the desired die inflow (D_I^*) and the components of the adjustment for fabrication WIP ($FWIPAdj$) we get equation (12).

$$WS^*(t) = MAX\left(0, \frac{D_I^*(t)}{DPW \cdot Y_D \cdot Y_L} + \frac{FWIP^*(t) - FWIP(t)}{t_{FWIP}}\right) \quad (12)$$

Hence, the desired die inflow ultimately drives the desired wafer starts, we note that this system is pushed by production requests from downstream the Fab. In addition, the sum of the long-term expected customer demand (*ED*) and the adjustment from assembly work-in-process (*AWIPAdj*) determine the desired die inflow (D_I^*).

$$D_I^*(t) = MAX\left(0, \frac{ED(t)}{Y_U} + AWIPAdj(t)\right) \quad (13)$$

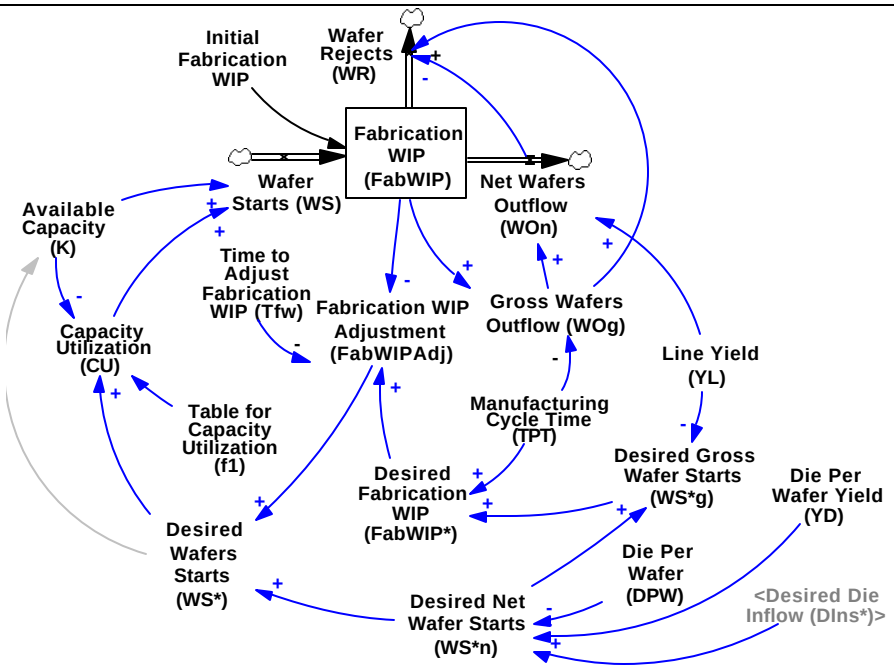

Figure 2 – Push system for die fabrication

Figure 2 shows a system dynamics representation of the push production system for the die fabrication process. Expected demand is simply an exponential smooth of actual orders updated over one year. And the assembly WIP adjustment (*AWIPAdj*) term reflects the firm's goal to replenish (reduce) assembly WIP when the current level is below (above) the target to correct the discrepancy over time (t_{AWIP}). The desired level of assembly WIP ($AWIP^*$) will be explained in the next section.

$$\dot{ED}(t) = \frac{D(t) - ED(t)}{t_{DAdj}} \quad (14)$$

$$AWIPAdj(t) = \frac{AWIP^{*}(t) - AWIP(t)}{t_{AWIP}} \quad (15)$$

We can rewrite the equation for the desired die inflow (D_I^*) by substituting equation (15), which provides further information about the adjustment from assembly work-in-process (*AWIPAdj*), into equation (13).

$$D_I^*(t) = MAX\left(0, \frac{ED(t)}{Y_U} + \frac{AWIP^{*}(t) - AWIP(t)}{t_{AWIP}}\right) \quad (16)$$

2.2. Demand Pull

The wafers out of Fabrication are pushed into the Assembly Die Inventory. In assembly, the wafers are cut into small square dies. Due to the disk-like shape of the wafer and variability of the fabrication process, only a fraction of the die produced are good enough to proceed into final assembly. For instance, dies at the margins of the wafer are commonly scraped. The die per wafer yield (Y_D) indicates the fraction of good die. So, the product of gross fabrication, die per wafer, line yield, and die per wafer yield determines the inflow of dies (D_I) into assembly. While the inflow of dies increase assembly work-in-process (*AWIP*), net assembled chip outflow (A_N) and assembly rejects (A_R) decrease it. The unit to die yield (Y_U) determines the fraction of gross assembled chip outflow (A_G) that are good and continue to finished goods inventory (*FGI*); the remainder, bad assembly, are rejected.

$$D_I(t) = F_G(t) \cdot DPW \cdot Y_D \cdot Y_L \quad (17)$$

$$\dot{AWIP}(t) = D_I(t) - A_N(t) - A_R(t) \quad (18)$$

$$A_N(t) = A_G(t) \cdot Y_U \quad (19)$$

$$A_R(t) = A_G(t) \cdot (1 - Y_U) \quad (20)$$

Gross assembled chip outflow is given by the minimum between the indicated gross assembled chip outflow rate determined by the production push ($PushA_G$) and the desired gross assembled chip outflow originated by the pull from demand signals ($PullA_G$). The former is given by the feasible completion rate, which is the ratio of available assembly WIP and the time to complete assembly (t_A). The latter is determined by the ratio of the desired net assembled chip outflow (A^*_N) and the unit to die yield (Y_U). Hence, when assembly WIP is sufficiently high

assembly is driven by the downstream demand. However, when inventory is low assembled chip outflow takes place at a rate that is feasible from the available assembly WIP.

$$A_G(t) = MIN(PushA_G(t), PullA_G(t)) \quad (21)$$

$$PushA_G(t) = AWIP(t)/t_A \quad (22)$$

$$PullA_G(t) = A_N^*(t)/Y_U \quad (23)$$

Assembled dies increase finished goods inventory and shipments decrease it. The company will ship as many goods to customers as the desired shipment rate (S*) or as many as the finish goods inventory can support, that is, the maximum shipment rate (S_{MAX}). Hence, the minimum of the desired and maximum shipment rate determines actual shipments (S). In addition, the volume of orders in backlog (B) divided by the target delivery delay (DD^*) determines the desired shipments rate. And the maximum shipment rate is given by the ratio of finished goods of inventory (FGI) and order processing time (t_{OP}).

$$\dot{FGI}(t) = A_N(t) - S(t) \quad (24)$$

$$S(t) = MIN(S^*(t), S_{MAX}(t)) \quad (25)$$

$$S^*(t) = B(t)/DD^* \quad (26)$$

$$S_{MAX}(t) = FGI(t)/t_{OP} \quad (27)$$

The desired level of finish goods inventory (FGI^*) is given by the product of desired weeks of inventory (WOI^*) and the expected shipments (ES). The latter is simply an exponential smooth of actual shipments updated over half a week. Managers set weeks of inventory coverage as the sum of the order processing time (t_{OP}) and the safety stock coverage (t_{SS}). While inventory coverage may change throughout the life-cycle of a product, for simplicity we assume a constant coverage policy.[3] This assumption is consistent with our investigation of the production behavior of mature products. Furthermore, by comparing the desired level of finished goods inventory with the actual level managers can order upstream to adjust any existing gap in FGI.

$$FGI^*(t) = WOI^* \cdot ES(t) \quad (28)$$

[3] For instance, at the early stages of a product life when demand is highly uncertain, inventory managers may adopt a policy of high (e.g. two weeks) inventory coverage. For mature products, with low demand variability, a policy of low (e.g. one week) coverage may suffice.

$$\dot{ES}(t) = \frac{S(t) - ES(t)}{t_{SAdj}} \tag{29}$$

$$WOI^{*} = t_{OP} + t_{SS} \tag{30}$$

$$FGIAdj(t) = \frac{FGI^{*}(t) - FGI(t)}{t_{FGI}} \tag{31}$$

Managers use the information about expected shipments (ES), finished goods inventory adjustment ($FGIAdj$), and backlog adjustment ($BAdj$) to determine the desired net assembled chip outflow (A_N^{*}). In addition, managers ensure that the desired net assembled chips are always non-negative. This request for assembled upstream chips is grossed up into the desired gross assembled chip outflow (A_G^{*}) with the yield for good units (Y_U) in the assembly line.

$$A_N^{*}(t) = MAX(0, ES(t) + FGIAdj(t) - BAdj(t)) \tag{32}$$

$$A_G^{*}(t) = A_N^{*}(t) / Y_U \tag{33}$$

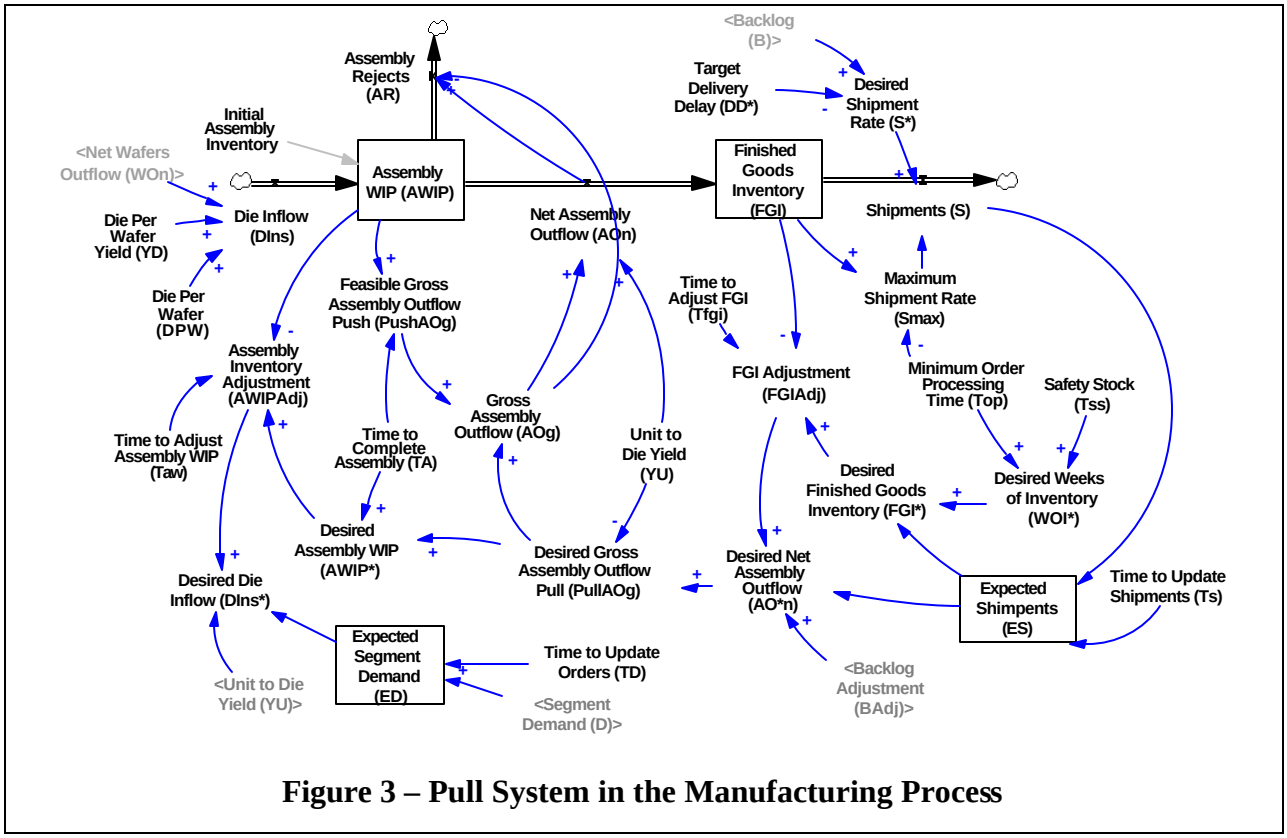

Figure 3 – Pull System in the Manufacturing Process

Figure 3 shows the demand-pull system for the assembly/testing process. The desired level of assembly WIP ($AWIP^{*}$) is set to produce the average gross assembled outflow rate over

the assembly time (t_A). The backlog adjustment (*BAdj*) term reflects the firm's goal to replenish (reduce) finish goods inventory when the current backlog is above (below) the target level, to correct the discrepancy over time. The desired level of backlog (B^*) is set at a level that allows the company to meet customer demand within the target delivery delay.

$$AWIP^*(t) = A_G^* \cdot t_A \tag{34}$$

$$FGIAdj(t) = \frac{FGI^*(t) - FGI(t)}{t_{FGI}} \tag{35}$$

$$BAdj(t) = \frac{B^*(t) - B(t)}{t_B} \tag{36}$$

$$B^*(t) = D(t) \cdot DD^* \tag{37}$$

We can rewrite the equation for the desired net assembled chip outflow (A_N^*) by substituting equation (35) and equation (36), characterizing the finished goods inventory adjustment (*FGIAdj*) and backlog adjustment (*BAdj*), respectively, into equation (32).

$$A_N^*(t) = MAX\left(0, ES(t) + \frac{FGI^*(t) - FGI(t)}{t_{FGI}} + \frac{B(t) - B^*(t)}{t_B}\right) \tag{38}$$

3. Distribution and Logistics

The manufacturer receives orders from OEMs and other customers. Since orders cannot be filled immediately, the company keeps a backlog of unfilled orders (*B*). The backlog accumulates the discrepancy between customer orders received by the company (*CD*) and actual shipments (*S*). If the manufacturer has the finished goods products available in inventory, it can ship them to customer at the desired shipment rate (S^*), otherwise will ship them as fast as it can (S_{MAX}). Overall, the manufacturer's ability to fill orders, that is, the fraction of orders filled (*FoF*) depends on the ratio between actual (*S*) and desired shipments (S^*). When actual shipments equal the desired shipment rate, the company is capable of shipping the full fraction of orders demanded by customers. When actual shipments are lower than the desired, the company fills only a fraction of its orders.

$$\dot{B}(t) = CD(t) - S(t) \tag{39}$$

$$FoF(t) = S(t) / S^*(t) \tag{40}$$

In this model we capture customers' response to supply availability, measured by the fraction of orders filled (*FoF*). Customers respond to a low fulfillment fraction by seeking alternative sources of supply; as they succeed, their orders drop. Intel's attractiveness to suppliers (A_I) is a nonlinear function of customers' perception of supplier delivery reliability (*PFoF*). In turn, customers' perception of delivery reliability (*PFoF*) adjusts from the actual delivery reliability – Fractional orders Filled (*FoF*) – with a third-order Erlang lag (λ), with an average time constant of six months (t_P). The third-order Erlang distribution captures the plausible distribution of responses by OEMs. At the instant of a decrease in the service level, all OEMs will still perceive the supplier as reliable, and there will be no shifts to alternative sources of supply. Therefore, the immediate response of the distributed lag should be zero. If service level remains low, however, some customers will change their perceptions about supplier reliability and seek other suppliers. The distribution of OEMs' reactions eventually peak, and then decrease, reaching zero after a sufficient time has elapsed. The delay captures the time required for OEMs to perceive changes in availability, to determine that the changes are not temporary and warrant a search for alternative sources, and to close deals with those alternative sources. To capture this delayed perception in the model, we use a third order smooth for the customer perception of fractional orders filled with a six months delay (t_P).

$$\dot{PFoF}_1(t) = \frac{PFoF(t) - FoF_1(t)}{t_P / 3} \tag{41}$$

$$\dot{PFoF}_2(t) = \frac{PFoF_1(t) - PFoF_2(t)}{t_P / 3} \tag{42}$$

$$\dot{PFoF}_3(t) = \frac{PFoF_2(t) - PFoF_3(t)}{t_P / 3} \tag{43}$$

Hence, attractiveness (A_I) is a function of customers' perceived delivery reliability ($PFoF_3$). A nonlinear function (f_A), a logistic curve, characterizes Intel's attractiveness (A_I). Attractiveness varies on a scale going from zero to one ($0 \le A_{LMin} < A_{LMax} \le 1$). A logistic curve captures customers' mild response to small changes in supply availability, and more significant responses to large changes in supply availability. For simplicity, we assume that competitors maintain constant delivery performance (i.e. a constant attractiveness (A_C) over time). This assumption allows us to measure changes in system behavior due to customers' reactions only due to changes in supplier conditions.

$$A_I(t) = f_A(PFoF(t)) \tag{44}$$

The manufacturer's market segment share is given by the ratio of the company's attractiveness divided by total attractiveness, that is, the sum of the company's and competitor's attractiveness. Finally, the product of total demand (*TD*) for chips and the company's market segment share (MSS_I) determines its customer demand (*CD*).

$$MSS_I(t) = \frac{A_I(t)}{A_I(t) + A_C(t)} \tag{44}$$

$$CD(t) = MSS_I(t) \cdot TD(t) \tag{45}$$

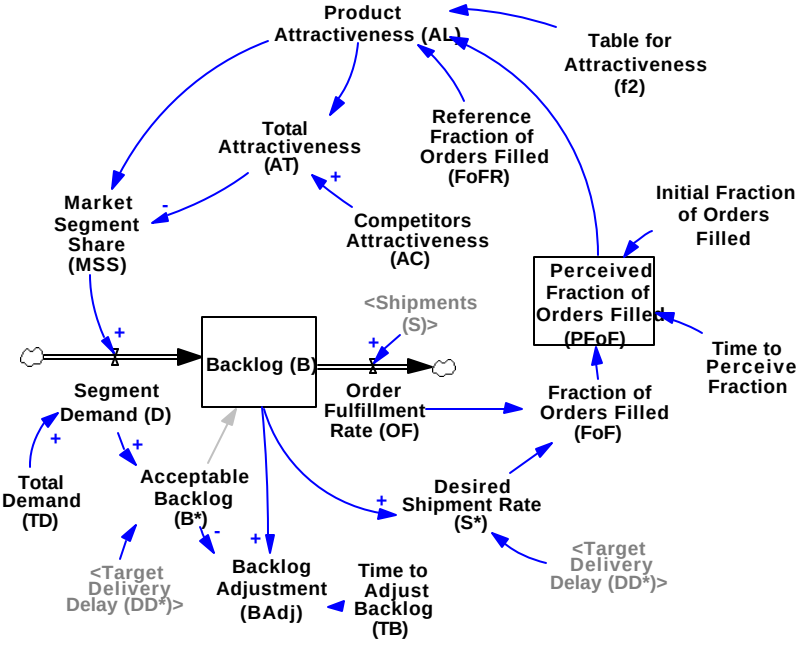

Figure 4 – Distribution and Logistics

Figure 4 shows the model diagrams for the distribution and logistics sector. Now, we rewrite the equations to express the system in the following form:

$$\dot{x} = f(x)$$

where x is the state vector composed by the state variables in our system and $\dot{x}$ is its first derivative with respect to time. The equation suggests that the first derivatives of the state variables can be written in terms of the vector of state variables. After a long and tedious algebraic substitution and disregarding the non-negativity constraints, we obtain a system of nine

non-linear differential equations, given by equations (46) – (54), in which we will base our analysis:

$$\dot{FWIP}(t) = f_U \left\{ \frac{t_A}{DPW \cdot Y_D \cdot Y_L \cdot t_{AWIP} \cdot K \cdot CU_N} \left[\frac{ES(t)}{Y_U} \left(1 + \frac{(t_{OP} + t_{SS})}{t_{FGI}} \right) - \frac{FGI(t)}{t_{FGI} Y_U} \right.\right.$$
$$\left.\left. - \frac{\frac{f_A(PFoF_3(t))}{f_A(PFoF_3(t)) + A_C(t)} \cdot TD(t) \cdot DD^*}{t_B Y_U} - \frac{B(t)}{t_B Y_U} - \frac{AWIP(t)}{t_A} + \frac{ED(t) \cdot t_{AWIP}}{Y_U t_A} \right] \right.$$
$$\left. \cdot \left(1 + \frac{t_F}{t_{FWIP}} \right) - \frac{FWIP(t)}{K \cdot CU_N \cdot t_{FWIP}} \right\} \cdot K - \frac{FWIP(t)}{t_F} \tag{46}$$

$$\dot{AWIP}(t) = \frac{FWIP \cdot DPW \cdot Y_D \cdot Y_L}{t_F} - MIN \left(AWIP(t)/t_A, \frac{ES(t)}{Y_U} + \frac{(t_{OP} + t_{SS}) \cdot ES(t)}{t_{FGI} Y_U} \right.$$
$$\left. - \frac{FGI(t)}{t_{FGI} Y_U} - \frac{f_A(PFoF_3(t))}{f_A(PFoF_3(t)) + A_C(t)} \cdot \frac{TD(t) \cdot DD^*}{t_B Y_U} - \frac{B(t)}{t_B Y_U} \right) \tag{47}$$

$$\dot{FGI}(t) = MIN \left(\frac{AWIP(t) \cdot Y_U}{t_A}, ES(t) + \frac{(t_{OP} + t_{SS}) \cdot ES(t) - FGI(t)}{t_{FGI}} \right.$$
$$\left. - \frac{f_A(PFoF_3(t))}{f_A(PFoF_3(t)) + A_C(t)} \cdot \frac{TD(t) \cdot DD^*}{t_B} - \frac{B(t)}{t_B} \right) - MIN \left(B(t)/DD^*, FGI(t)/t_{OP} \right) \tag{48}$$

$$\dot{B}(t) = \frac{f_A(PFoF_3(t))}{f_A(PFoF_3(t)) + A_C(t)} \cdot TD(t) - MIN \left(B(t)/DD^*, FGI(t)/t_{OP} \right) \tag{49}$$

$$\dot{ED}(t) = \frac{f_A(PFoF_3(t))}{f_A(PFoF_3(t)) + A_C(t)} \cdot \frac{TD(t)}{t_{DAdj}} - \frac{ED(t)}{t_{DAdj}} \tag{50}$$

$$\dot{ES}(t) = \frac{MIN \left(B(t)/DD^*, FGI(t)/t_{OP} \right) - ES(t)}{t_{SAdj}} \tag{51}$$

$$\dot{PFoF}_1(t) = \frac{MIN \left(B(t)/DD^*, FGI(t)/t_{OP} \right)}{(t_P/3)\left(B(t)/DD^* \right)} - \frac{PFoF_1(t)}{t_P/3} \tag{52}$$

$$\dot{PFoF}_2(t) = \frac{PFoF_1(t) - PFoF_2(t)}{t_P/3} \tag{53}$$

$$\dot{PFoF}_3(t) = \frac{PFoF_2(t) - PFoF_3(t)}{t_P/3} \tag{54}$$