

## LONG-TERM PRODUCTIVITY MECHANISMS OF THE SEMICONDUCTOR INDUSTRY

Randal Goodall, Denis Fandel, Alan Allan<sup>‡</sup>, Paul Landler<sup>†</sup>, and Howard R. Huff

International SEMATECH  
2706 Montopolis Drive, Austin, TX 78741

A small set of business and technology relationships can characterize the continuous upward revenue growth of the semiconductor industry. An essential success metric has been functional cost or “productivity.” Falling functional cost is assumed to be responsible for the continuously growing demand for electronics and semiconductor products. In addition to the well-known productivity measure, \$/transistor, other functional cost metrics are described that play a role in the semiconductor industry’s continued success. In the future, additional productivity enhancements will be required to keep functional costs dropping at historical rates (and the industry continuing to grow), and some existing mechanisms may become ineffective. Key productivity mechanisms are assessed for their effectivity. The International SEMATECH Industry Economic Model used throughout this analysis is described.

### INTRODUCTION

The semiconductor industry is the most productive the world has ever seen. Semiconductors have become so pervasive (even in less affluent cultures) that they may be the most impacting technology (in rate and effect) in human history.<sup>1</sup> The productivity growth underpinning the semiconductor industry can be envisioned in various startling examples: (1) the exponential revenue growth of the semiconductor industry (averaging ~16% per year) has continued unabated for 40 years; (2) there are more bits of memory on a single 300 mm wafer produced today than were produced by the *entire* industry in 1984; (3) there are more transistors produced per year (~1 quintillion) than grains of rice, and each rice grain can buy 100’s of transistors.<sup>2</sup>

Surprisingly, a small set of business and technology relationships can characterize this remarkable result. The continuing contribution of these business relationships to industry productivity is the topic of this paper. Cost per transistor, device switching speed, and design efficiency have been key productivity drivers. The exponential drop in the cost of a memory bit (with the business benefits to all other chip types coming “naturally”) has for many years been the most suitable tracking parameter for cost per transistor. In the first section below, productivity is defined with the primary focus on cost per transistor. In the second section, productivity metrics and monitoring are described. A popular visualization is noted, but set aside for an alternative and more truly representative but,

---

<sup>‡</sup> Intel, 5000 W. Chandler Blvd., M/S CH6-312, Chandler, AZ 85268.

<sup>†</sup> Formerly with IBM and ISMT. Now: 165 Thorpe Cove Road, Charlotte, VT 05445.

unfortunately, more complex set of diagrams. The subsequent section assesses whether cost per bit is sufficiently encompassing to track all industry productivity issues. Using the *International Technology Roadmap for Semiconductors* (ITRS)<sup>3</sup> developed through global consensus and the Industry Economic Model (IEM)<sup>4</sup> developed at International SEMATECH (ISMT), costs and benefits are assessed for both “traditional” and new productivity enhancers anticipated over the next decade, including non-planar devices, 3-D integration, and scaling for increased density, rapidly shifting lithography generations, possible alternatives to a 450 mm wafer generation, and non-traditional circuitry (nano-sized, quantum mechanical, and biological devices).

## PRODUCTIVITY

Figure 1 shows how revenue, capital investment, and output (measured in transistors) are related to the traditional metric of functional cost or “productivity” (\$/transistor in the figure). The falling functional cost is assumed to be responsible for the continuously growing demand for electronics and semiconductor products. Note that while the units of the various curves are arbitrarily scaled on the vertical axis, the growth rates (slopes) are relatively correct.

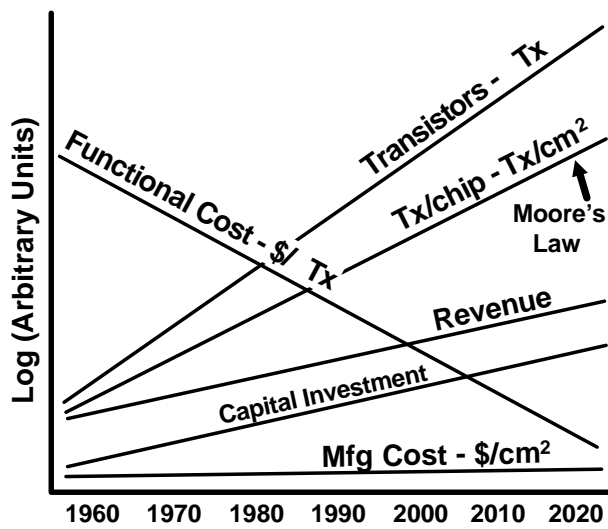


Fig. 1. Functional form of key semiconductor industry business trends (Tx=transistor).

It is proposed here (but not proven) that the productivity trend of the semiconductor industry represents a kind of demand saturation for the information processing that is essential to a post-industrial economy. It is, in effect, a sociological result. Certainly,

technology innovation and business investment rates strongly influence the continuing introduction of new semiconductor parts, but because of the *ubiquity* of electronics and semiconductor-based information processing specifically, society has, through market forces, agreed to “tax” itself at a certain rate (and about 16% more each year) to provide “thinking assistance machines” to everyone.

### Metrics

To understand and study productivity in the semiconductor industry, metrics of productivity must be developed. From the analysis presented here, cost per function (e.g., density, speed, power) is found to be the most useful metric for semiconductors. Over the last few years, International SEMATECH and others have refined the analysis of cost per function to demonstrate its effect in history and to assess its impact in the future. The rate of technology innovation required in the future, as documented in the *International Technology Roadmap for Semiconductors* (ITRS), is based upon these assessments.

An essential characteristic of a useful metric is its embodiment of the complexities of a domain while maintaining a relative simple conceptual exposition and functional form.

When Gordon Moore first proposed what is now called Moore's Law, he was commenting on his observation about the number of transistors on a chip. At that time (prior to very large scale integration), more transistors meant more functions directly. Logic elements (AND, OR, NOR, etc.) and other similar devices with a few transistors per function increase in literal functionality as the number of transistors rises.

In fact, however, the historical trend data in cost per function (\$/Fn) came to be based on memory, and in particular dynamic random access memory (DRAM). By the late 1980's, DRAM manufacturers had fallen into a concerted routine for design and manufacturing technology advancement that continuously delivered four times as many bits every 3 years with average chip size and cost increasing only 41% over 3 years (about 29% per year decline in \$/bit). Even with abnormal market conditions,<sup>5</sup> the cost per bit has maintained its nominal long-term trajectory (see Figure 2). Note, however, that to keep the cost of *chips* about constant, the time of introduction of 4X bits (a new DRAM generation) has increased to four years.

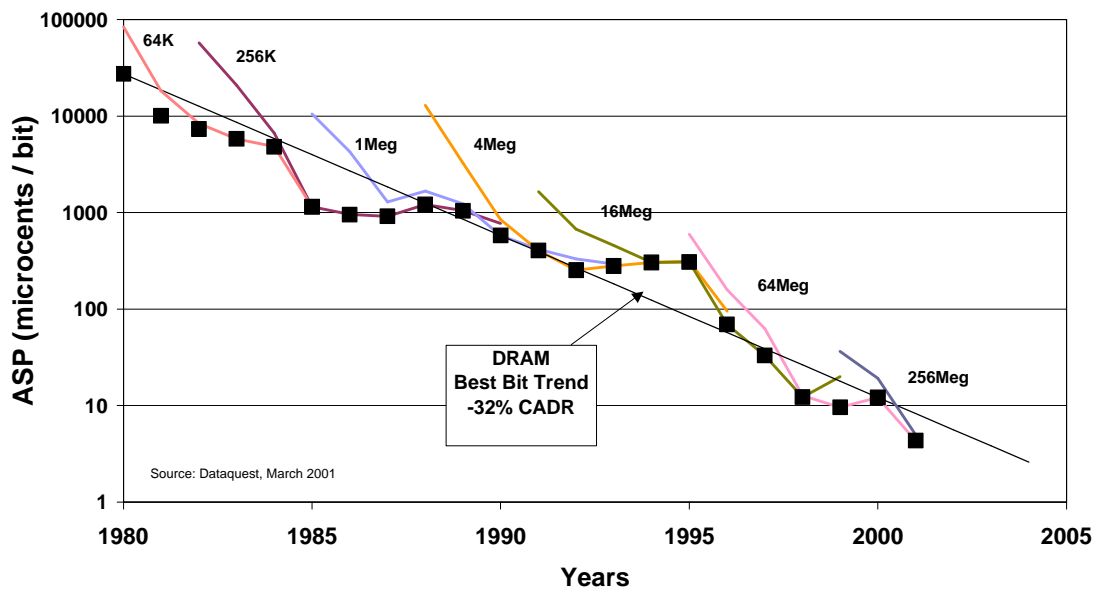


Fig. 2. Actual average price per bit for DRAM memory. The black boxes represent the "best bit price" over time as DRAM chip generations evolved (solid lines).

When first developed, the data regarding cost per bit was clearly a DRAM phenomenon and well understood as such. As the industry attempted to understand itself better, the idea of the "bit" was generalized to the more conceptually relevant, but less specifically applicable idea of "function." The switch or transistor (1 of which is assumed to be in a bit of DRAM memory) became the essential unit of function ("remembering something" is an easily recognized information processing capability). The trend lines continued to make sense, but primarily because the number of non-DRAM transistors created is a perturbation to the total (<15%). Additional productivity metrics will be described below that broaden the coverage of semiconductor functional benefits.

## MONITORING PRODUCTIVITY

### Initial Approach – "Staying on the Productivity Curve"

The chart in Figure 3 was developed and popularized by SEMATECH in the mid-1990s (prior to the startup of 300 mm wafer size development). The chart purports to show

components of the historical \$/Fn reduction experienced by the semiconductor industry. The purpose of this widely presented chart was to illustrate perceived impending changes in cost contributors, particularly the need (at that time) to improve the performance of

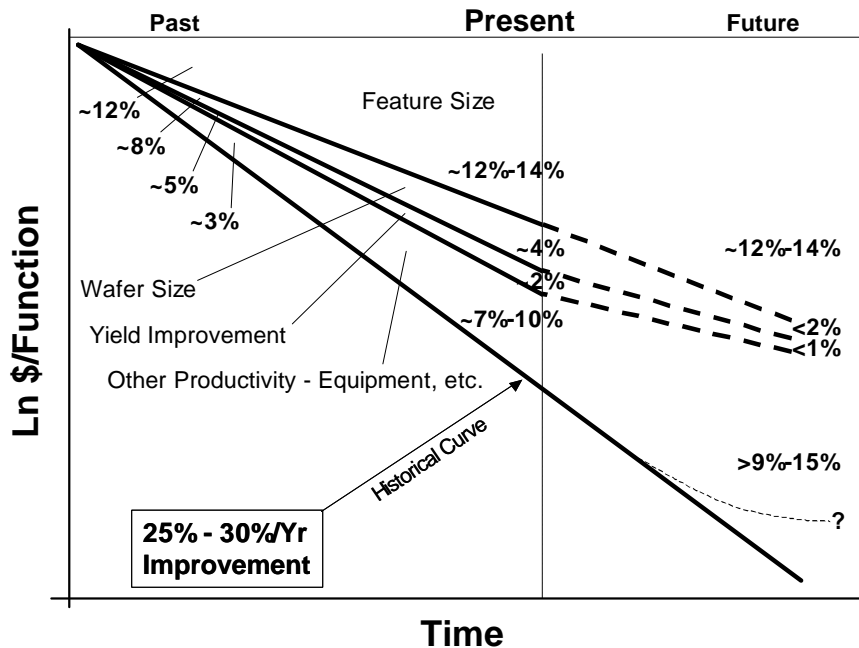


Fig. 3. Early representation of industry productivity drivers. Difficulties with this chart are described in the text.

process equipment. The chart notes the importance of maintaining the productivity trend, with the implied downside of falling off the historical productivity trend being a concomitant falling off the revenue and growth trends.

The following are the key points illustrated in this chart (with corresponding conceptual difficulties in *italics*):

- The main driver in productivity improvement is the continual shrinking of the transistor size. *This feature size reduction is shown to account for a 12-14% reduction in cost per transistor each year. Although the feature size itself (i.e., CD) historically drops at 12% year, its contribution to productivity goes as the square of the feature size (i.e., a decrease in the “area” of a feature by ~50% every node is a ~25% per year raw productivity driver). Therefore, feature size reduction accounts for most of functional cost improvement (other costs remaining about equal). While this was forecast in the chart to remain steady, it actually accelerated since that time to two years per node.*
- One of the components of productivity improvement shown is wafer size increase. An individual wafer size change is expected to give ~40% cost decrease, but roughly ten years elapse between wafer size changes.<sup>6</sup> This lowers the annual benefit to ~4% per year. *In contrast to what is shown on this chart, 300 mm wafer manufacturing will be >40% more productive, so wafer size increases do not lose their productivity benefit.*
- Yield continually improved over the 30-year life of integrated circuit manufacturing and now rises to nearly 100% fairly quickly in a new or upgraded fab. Since yield is capped at 100%, it is not a factor in future productivity enhancement.

- The last category, sometimes shown as “OTHER,” focuses on improvements in the effectiveness of the equipment, usually as measured by Overall Equipment Efficiency (OEE). The implication of the chart is that future productivity was strongly dependent on the increasing of this factor. *Since OEE is capped at a (practically unreachable) value of 100%, however, the illustration of this factor contributing increasingly more to overall productivity is unrealistic.* Of course, there continues to be a concerted effort to improve the reliability and throughput of tools and gains have been made.

Several additional difficulties with this chart limit its usefulness for future productivity assessment:

- The vertical axis (“Ln \$ / Function”) is not properly defined for the sub-factors.
- Lithography aside, Yield, Wafer Size, and Other (OEE) are issues of *wafer-level* manufacturing. They are not associated with individual transistors, per se.
- The relative contributions of sub-factors to \$/Fn are severely (even absurdly) misrepresented, and their time evolution as indicated is improper. This comes about because the spaces *between* the lines, which are shown to be growing larger as time progresses, are interpreted as exponential contributions of the sub-factors. The various lines might be conceptualized as productivity trends without contribution from the sub-factors, but *only the slopes* have meaning, and that is only relative. The vertical translation of the curves is arbitrary, and any ascribed meaning to the gaps between them has no analytical basis.

A new representation for monitoring productivity is needed. An improved view for understanding, tracking, and ultimately ensuring industry productivity is presented below. It is being considered at ISMT for managing development initiatives and productivity programs.

### ***Improved Approach – Manufacturing Effectiveness***

A few key factors are sufficient to drive the reduction in functional cost. The foremost effect is the ability to decrease the size of the smallest feature so that more transistors can be built per area of silicon (Tx/cm<sup>2</sup>). The rise in density is almost completely attributed to advances in photolithography, with necessary tracking improvements in the technology of deposition, etch, and measurement. A second and necessary, adjunct effect is the manufacturing capability to maintain a relatively stable cost to process silicon wafers (\$/cm<sup>2</sup>). There is an additional component of density improvement that obtains from design, whereby inventive configuration and layout of device elements has allowed the space required per function to drop.<sup>7</sup>

Figure 4 illustrates this division between technology and manufacturing productivity enablers. \$/Tx (functional cost) on the left of the diagram can be thought of as being \$/cm<sup>2</sup> ÷ Tx/cm<sup>2</sup>. At the bottom of the chart are density issues. These are solved by technology driven development activities. This is where the “miracle” side of our industry (and to some extent only our industry) occurs. The less widely understood component of manufacturing effectiveness is shown across the upper portion of the chart. It represents the contribution (in the form of nominally constant \$/cm<sup>2</sup> for silicon processing) of the manufacturing sciences. \$/cm<sup>2</sup> can thought of as \$/wafer ÷ cm<sup>2</sup>/wafer.

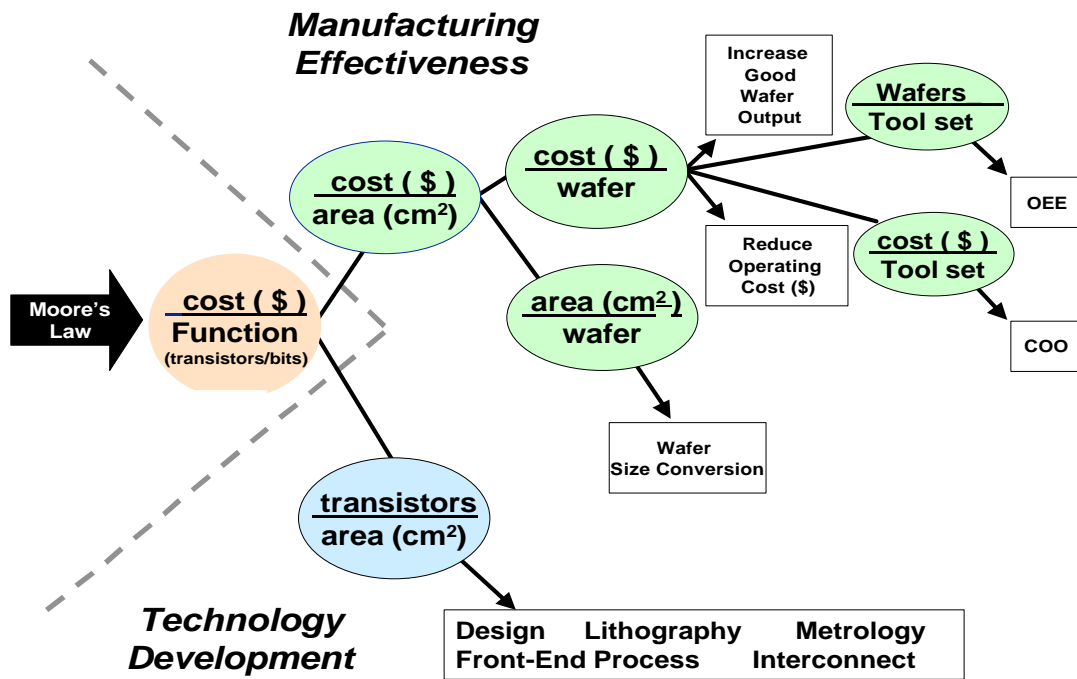


Fig 4. Manufacturing Effectiveness and Technology Development.

The latter factor increases as the regular drumbeat of wafer size changes moves through the industry. Since the wafer is the basic unit of manufacturing for semiconductors, controlling the cost of wafer manufacturing is a key industry issue. The tide against which manufacturing resists is evolving technology. In the lower branch of the chart, technologies are developed to drive density; in the upper branch, their costs are accommodated. Of course, every effort is made to develop cost-effective equipment, but ultimately in this industry, capability wins and costs are accommodated.

Cost per wafer can be viewed in two ways in terms of productivity. First the cost of manufacturing each wafer can be driven down and/or the number of wafers for a given cost can be driven up. The former is executed in two ways: (1) operating costs, per se, are reduced (e.g., the cost and amount of consumables, energy, labor, maintenance) and (2) fixed costs, are managed, like the depreciated cost of equipment and facilities, which are heavily negotiated. The number of wafers finished per dollar can be driven upward by faster equipment, improved equipment reliability, higher yield (although high yields cannot be driven much higher, they can be reached faster), and better factor automation, logistics, planning, and scheduling. As illustrated in this figure, by continuing to “divide out” effects, even common performance metrics such as COO and OEE (cost of ownership and overall equipment efficiency, respectively) can be evidenced in this kind of analysis.

### A Bigger Picture

Figure 5 shows a logical extension of the above concept to the next higher level. This view is important for understanding what is presumed to be meant by “function.” The users of semiconductors, with whom the industry maintains its precarious social contract of increasing functionality for increasing dollars, *never* interact with a semiconductor directly. They use electronic systems of some sort. Within the supply chain (at least today) between silicon chip and system are packaged chips and boards. Of course, this

hierarchy can be expanded in detail by sub-dividing, but the message is the same—at every level below the user, technology and manufacturing work together to drive down functional cost. Technology tends to increase the functions per manufactured unit, and manufacturing controls unit production costs.

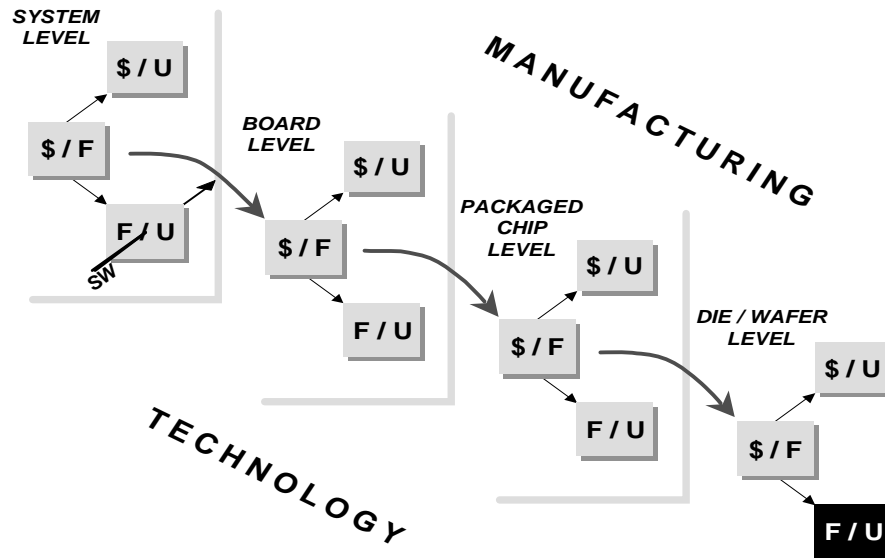


Fig.5. Productivity hierarchy. Functional cost requirements cascade down from the highest level.

*F*=functions and *U*=manufacturing units.

Historically, the greatest leverage in this chain has always been Functions/Unit at the wafer level (and will be for the intermediate-term future). This is because (1) the long-term power of 2-D VLSI scaling is so overwhelming compared to all other productivity measures and (2) the benefits to the end user of semiconductors pass through the supply chain fairly directly. For example, the effect of a faster microprocessor and more memory at the chip level are directly measurable at the system level. Whenever one layer of this hierarchy cannot drive productivity as strongly as desired, the others above have to pick up the difference. Today, the coordination of this need is not integrated vertically (although it is well considered at each layer individually, e.g., through the ITRS in the semiconductor industry and similar efforts and trade association arrangements in other segments).

Of particular note are the crosscutting pervasiveness of software technology (“SW” in the chart) and the requirements on it to support the delivery of productivity generated within hardware.<sup>8</sup> Recent data<sup>9</sup> suggests that while the hardware side of the electronic systems industries has been improving productivity (even accelerating it), the software side of the industry has done the opposite. Two “partners,” one increasingly productive, but commoditized by technology singularity and competition and the other decreasingly productive, but specialized and without competition, will naturally evolve a relationship where the profit margin they symbiotically generate will be systematically distributed inversely to productivity (i.e., proportional to cost). The semiconductor industry must comprehend this systemic reality in the future to maintain sufficient overall margin to cover its increasing development costs.

### **Modeling Productivity**

One of the principles of the ISMT efforts to track productivity is that it must be data-based, and the approach to maximizing data utility is to incorporate the data into models of the industry’s business behaviors. Before looking at model data, however, it is

important to understand how various factors combine to give the on-going productivity enhancements.<sup>10</sup> Figure 6 is a schematic of the contribution of wafer size and technology nodes (and implicitly, yield and process complexity) to productivity improvement. The left vertical axis of the chart is functions per area, and DENSITY (only) is mapped against that axis as the major contributor (as discussed above). The other vertical axis here is more conceptual. It is the cost per area multiplier for technology and wafer size effects (those two curves are normalized arbitrarily to 1.0 for early times). Each contributor is described below.

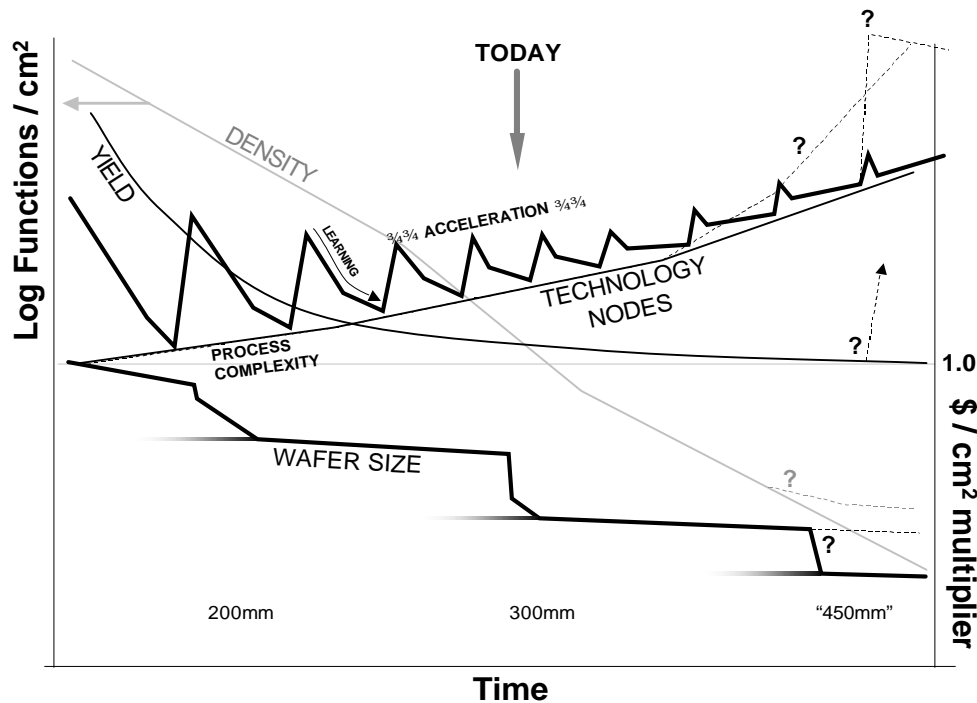


Fig 6. Productivity contributors and their form. Details are explained in the text.

**Wafer Size.** WAFER SIZE spans the 150 mm to “450 mm” eras. Wafer size changes have been regular productivity enhancements over the years. The productivity benefit is trivial to identify: when the wafer area increases by  $>2$  times, but the cost of the new tool set for the same number of wafer starts increases by only 30-40% (which is typical), the cost per area decreases by 30-50% — an annualized improvement of about ~4% when wafer size changes occur about ever 10 years. This means that every wafer generation brings an intrinsic productivity boost. This is necessary (as shown below) because the technology generations alone (which drive up density) cannot keep pace with the historical industry trends. The case for another wafer generation after 300 mm, however, is discussed later.

**Technology Nodes.** Each of the TECHNOLOGY NODES (a generation of DRAM half-pitch), improves productivity by shrinking the size of a device. These nodes, however, and the processes they represent, bring with them an increasing manufacturing cost (10-15% per node). The following features are notable:

- General product yield has risen in the industry to  $>90\%$  for most high volume chips in production mode. Most of this improvement occurred in the 1970’s and does not contribute to productivity enhancement now, so the YIELD curve goes to a cost benefit of ~1.0.

- In each generation, however, the yield loss due to defects, mis-processing, slow or unreliable equipment must be overcome to move to the target productivity point. Over the time period shown, learning has improved dramatically so that the overall starting yield degradation as well as the speed with which high yield is attained have improved. Learning how to upgrade fabs (rather than build new fabs for each node) has been a key element in this improvement. The yield improvements that used to take 2 years to achieve in a new fab now occur within 6 months in an upgraded fab. Further improvements will be incremental, not dramatic.
- A significant part of the slowly rising process cost is due to process complexity. This is most easily seen in additional metal layers (and required additional processes like CMP) and in additional, more costly mask layers for compensating the increasingly less forgiving lithography processes.
- The pace of technology nodes is a critical factor in their effective productivity. From the late 1990's through the early 2000's, technology is moving forward on a two-year per node basis (rather than the previous and eventual 3 years per node). There was, of course, a corresponding acceleration in density. Even with higher capital costs, the earlier availability of denser, faster devices reduces the functional cost. One key result of this is that the 300 mm productivity boost was not required as early as originally forecast. In general, every year the industry can stay on a two-year technology cycle delays the next wafer transition by one year.

**Failure Modes.** Each component in Figure 6 has a failure mode identified (dotted lines and "?"). While yield, per se, is not expected to collapse from a manufacturing point of view (although failure to keep pace with defect metrology could introduce risks), effective yields will fall if a technology node "fails," i.e., cannot execute the desired dimension at any cost. There is also risk that process complexity for late- and post-optical patterning will be so high that a technology node cannot deliver its necessary productivity boost. The failure of the next wafer size to provide a sufficient boost is discussed below, and the industry's growth might be mortally wounded if another wafer generation were introduced which did not lower costs.

## MODELING PRODUCTIVITY

International SEMATECH began the development of a comprehensive industry business/economic model in late 1998. The detailed assumptions, construction, operation, and output capability of the IEM are described elsewhere.<sup>11</sup> The basic approach to productivity modeling is described in Appendix A, but Figure 7 illustrates the point for the specific case of leading edge semiconductor products (dense memory and high performance logic/MPUs). Note the IEM has the capability to compare productivity between widely divergent industry segments by normalizing to "technology equivalent transistors" (referred to here simply as transistors). For this analysis combining logic and memory, 10 bits of DRAM are about equivalent to 1 logic transistor.

Figure 7a and 7b are the density (area per transistor) and manufacturing cost (\$ per area), respectively. As noted earlier, density productivity (\$/Tx) can be understood as  $\$/\text{cm}^2 \div \text{cm}^2/\text{Tx}$  (although not computed so simply in the IEM).

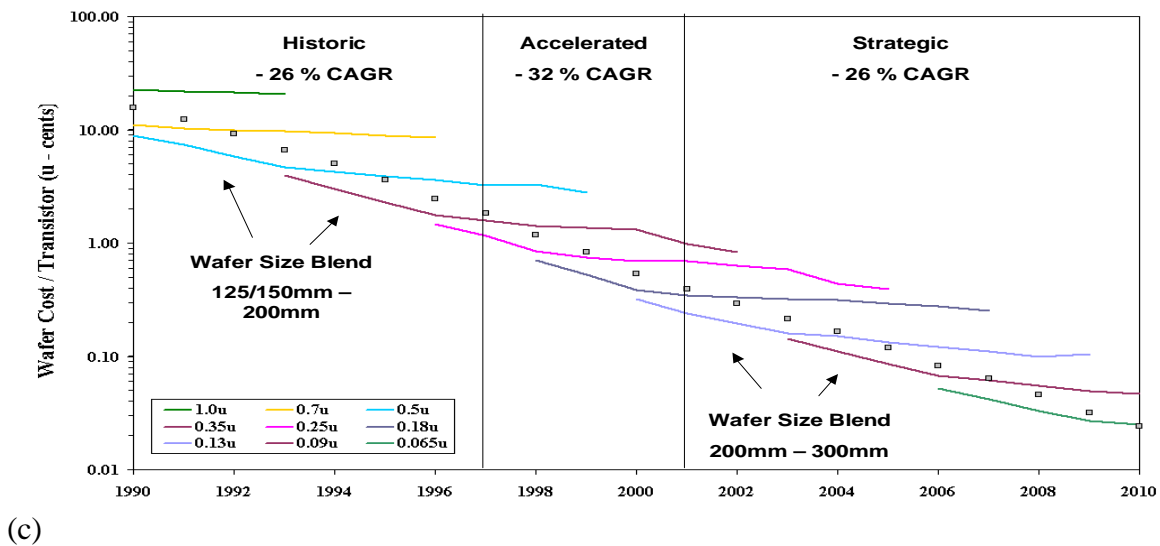
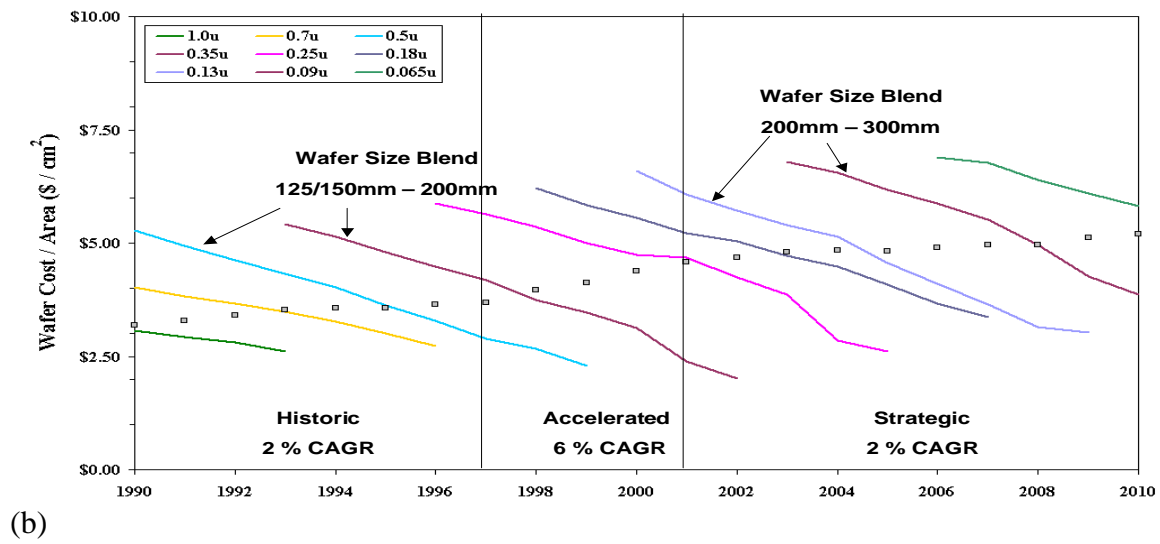
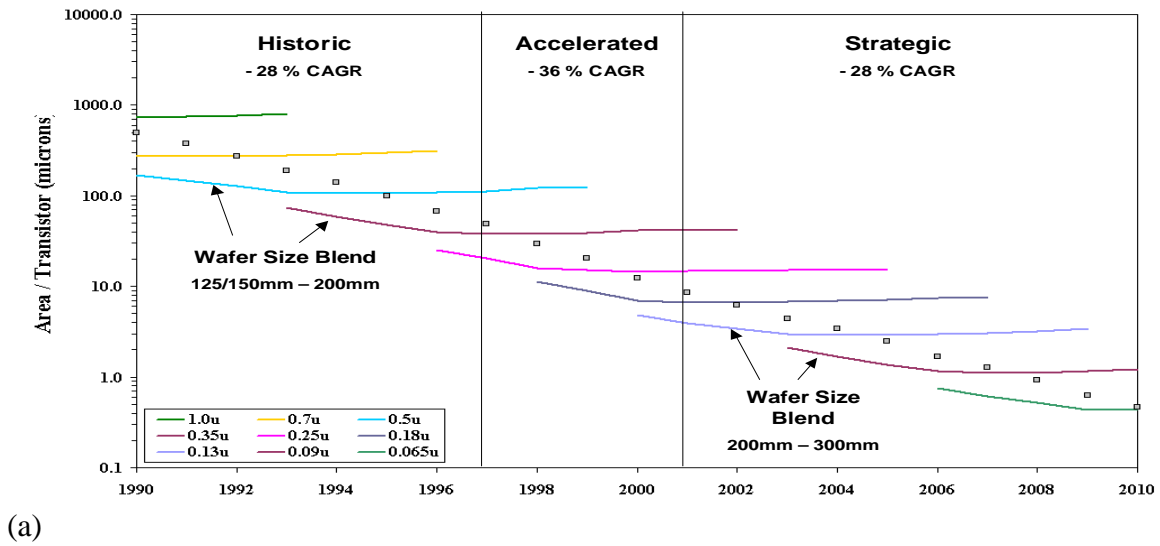


Fig. 7. Trends for Leading Edge Products (e.g. DRAM, MPU). (a) Area per Transistor (Density). (b) Manufacturing Cost per Area. (c) Manufacturing Cost per Transistor (Productivity).

Figure 7c shows the modeled cost per transistor (with 10 bits of DRAM = 1 transistor) dropping about 30% per year. Figures 6 and 7 can be compared to verify that the conceptual representations in Figure 6 correspond to the data-based features in Figure 7. Note that the saw-tooth technology curve in Figure 6 represents the overall leading-edge segment of the industry optimally “jumping” from curve to curve in Figure 7b. Any one fab, of course, may traverse a longer portion of a single cost curve for business reasons. Comparison of Figures 2 and 7c (and accounting for 10 bits per transistor in Figure 7c) shows favorable trend agreement in cost per transistor.

## NEW PRODUCTIVITY MEASURES

For more than a decade, cost per bit has been the key productivity measure for the semiconductor industry. Moore’s (long-standing) Law that the number of transistors per chip doubles every 18-24 months, even if accurate, does not capture the business dynamics that \$/Tx represents. That is, density functional cost does not deal with all the industry’s issues. Other metrics may be needed.

### *High Speed*

As noted above, in addition to \$/Tx, speed of operation of transistors is also an important productivity parameter, and the ITRS is paying increasing attention to establishing speed

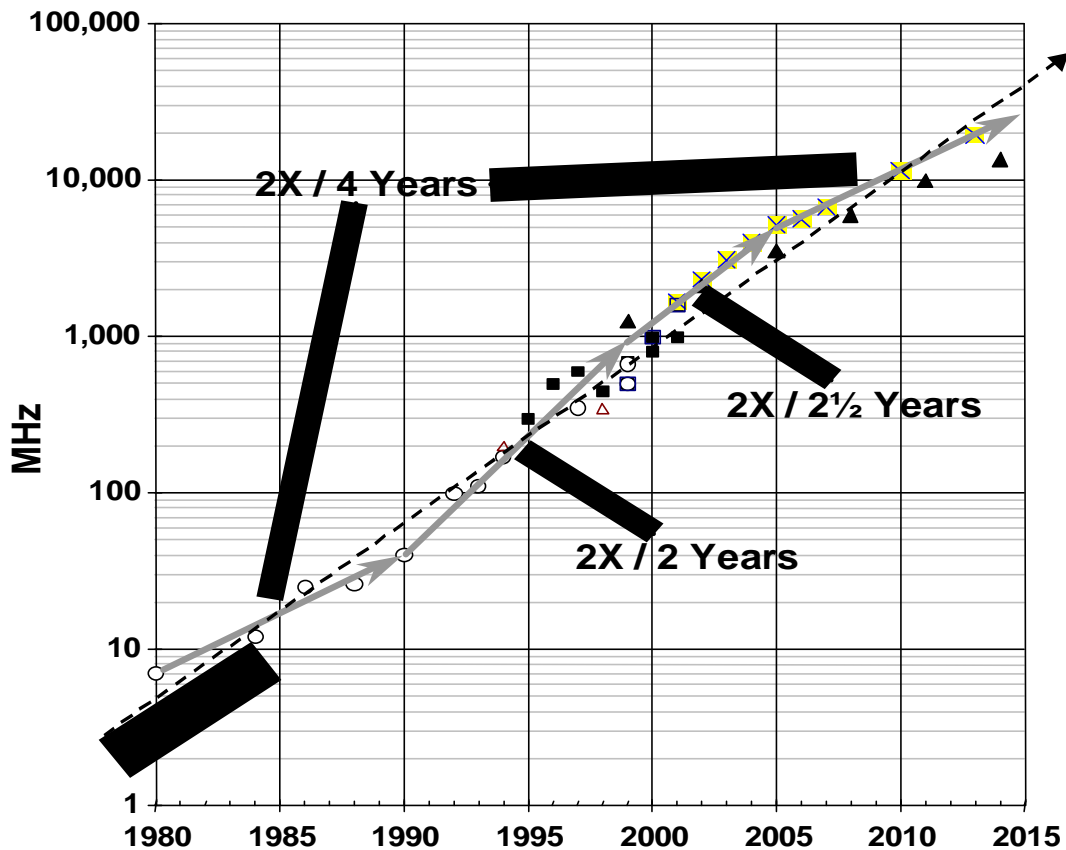


Fig. 8. Speed history of microprocessors. Symbols represent various data sets considered in the source analysis.

requirements and potential solutions. In fact, smaller transistors (specifically those with shorter gate lengths) do go faster because of lower travel distances and improved electrodynamics. Size reduction alone, however, is not the only determinant of speed improvements. Driving the Industry to shrink the sizes of transistors has sufficiently synergized the key need of microprocessor makers, i.e., speed, that the ITRS and the industry has had little need to focus on anything other than \$/Tx.

It is proposed here that two quantities are conceptually necessary to convey the functional importance of semiconductors in information processing, i.e., the number of bits and the rate at which they can be accessed or changed. The productivity of the industry relative to its customers depends on both a falling cost/bit and a falling cost/bit-rate.

The number and cost of memory bits is fairly well documented, but the bit-rate data is not nearly so well understood or measured. Figure 8 shows the nominal speed in MHz of microprocessors.<sup>12</sup> To obtain \$/bit-rate (in a way analogous to \$/bit), we need a measure of the aggregate yearly cost of all “bit changers” — MPUs and DSPs are used here for this purpose — and the total number of bit changes they theoretically represent. Of course in neither the bit nor bit-rate case do we (ever) assume that all the functionality is simultaneously (or even ever!) in operation or use. This is a discussion of potentials.

Figure 9 shows the bit-changes per second history, obtained by multiplying the data in Figure 8 by the nominal historical bit width of microprocessors (noted at the top of the figure<sup>13</sup>). Bit widths for future times are assumed to provide continuous exponential

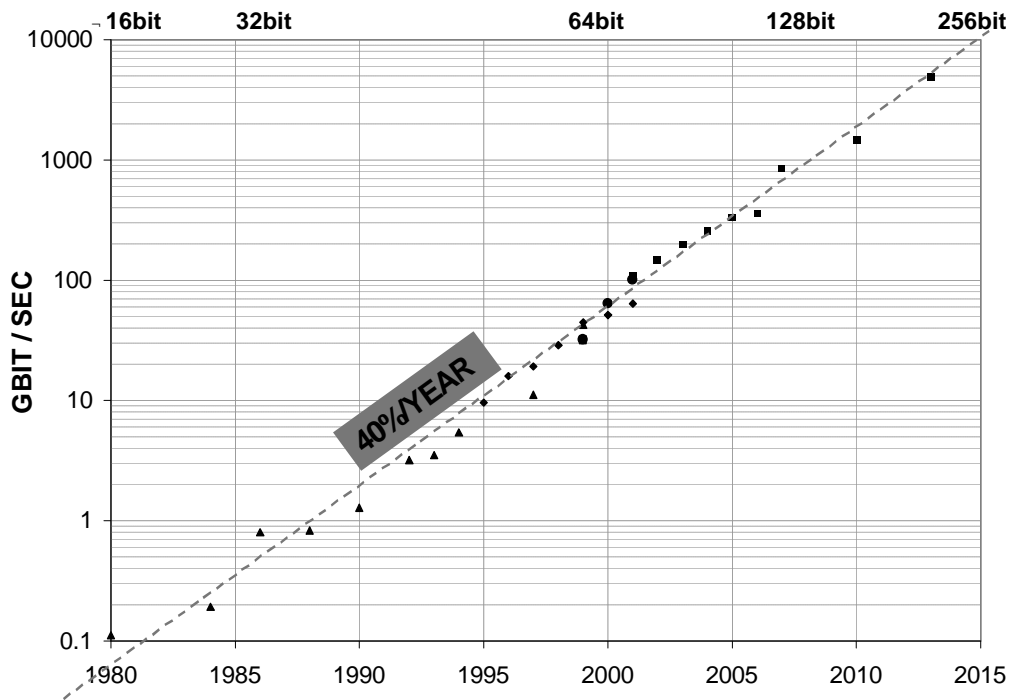
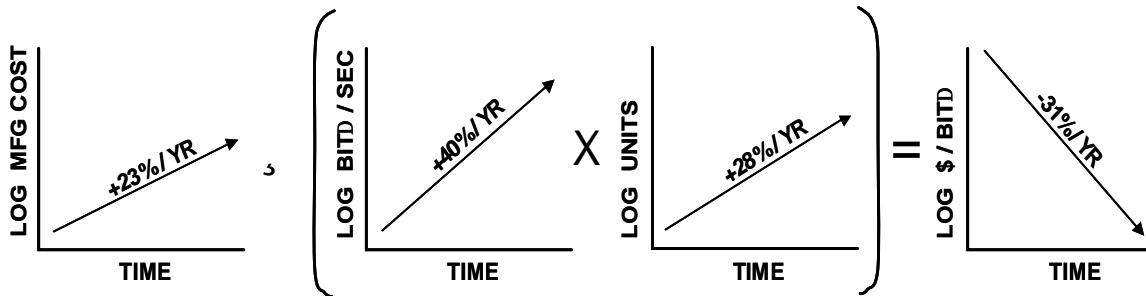
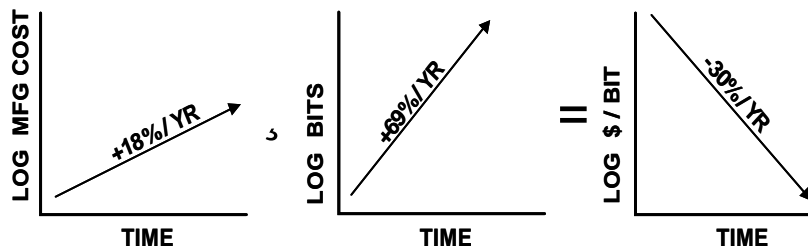


Fig. 9. Bit changes per second trend.

growth of bit-rate. It must be noted that the end user’s functional expectations are not with the microprocessor, per se, but with his electronic product, so any mechanism (e.g., multiple processors) that obtains these nominal bit-rates is “OK” for our purposes. On the other hand, the nominal bit widths assumed are conservative given video controller architectures now in development.



(a)



(b)

Fig 10. Schematic representation of relative productivity calculation. (a) Cost per Bit-rate for MPUs and DSPs. (b) Cost per Bit for DRAMs.

The absolute scaling of these exponential functions is a secondary (and complex) issue.<sup>11</sup> Figure 10 illustrates the conceptual closure of the bit-rate metric (bit-rate is abbreviated “BIT  $\Delta$  / SEC” in the figure). The nominal industry growth rates for manufacturing cost and units for MPUs (including DSPs) during the 1990’s are shown, along with the analogous, well-known DRAM result. The final productivity ( $\$/Fn$ ) growth is computed in the now familiar way: Productivity = Total Cost  $\div$  (Functions/Unit  $\times$  Units). Note that implicitly multiplying the last curve on the right by the number of seconds in a year ( $\sim 3.15 \times 10^7$ ) to convert “BIT  $\Delta$  / SEC” to “BIT  $\Delta$ ” does not change the growth rate. So, finally, in addition to the long-time productivity metric, “30% decline in  $\$/Bit$ ,” a new metric may be additionally considered, “31% decline in  $\$/Bit$ -rate.”

### Low Power

Low power is an essential attribute of ICs that is and will continue to enable their *ubiquity*. The world is moving toward the era when semiconductors will be: integrated into clothing, food, and building materials; networked and embedded sufficiently that local, wireless access to any data and appropriate processing power are as common as electricity and water today; merged with micro-electro-mechanical systems (MEMS) to create customized intelligent physical agents; and, in general, have enabled machines to exceed us in “thinking power” as much as they do in physical power.

Power (energy-rate) is not quite as intrinsic a function as Bits or Bit-rate. Power is not a potentiality—it is an actuality, meaning that, while society may tax itself for the potential to turn on increasingly powerful computers at any time, anywhere, they will *not* pay in advance for all the energy they *might* consume. The productivity metric for Power is still  $\$/Fn$ , but is more abstract, since the functionality is realized in the semiconductor, while the operational cost for the function is realized, rather, in the energy source (batteries, wall plug, solar panel, etc.) of the system. This metric remains under development.

So, finally, the social demand saturation (if it indeed is) for remembering information (density), modifying information (speed or bit-rate), and doing this anywhere (ubiquity or energy-rate) can be seen as a 3-parameter functional over time (shown conceptually in Figure 11). Many solutions exist, but they all share a common underlying set of productivity requirements that the semiconductor industry must continue to satisfy at historical rates if it is to continue to receive historical benefits.

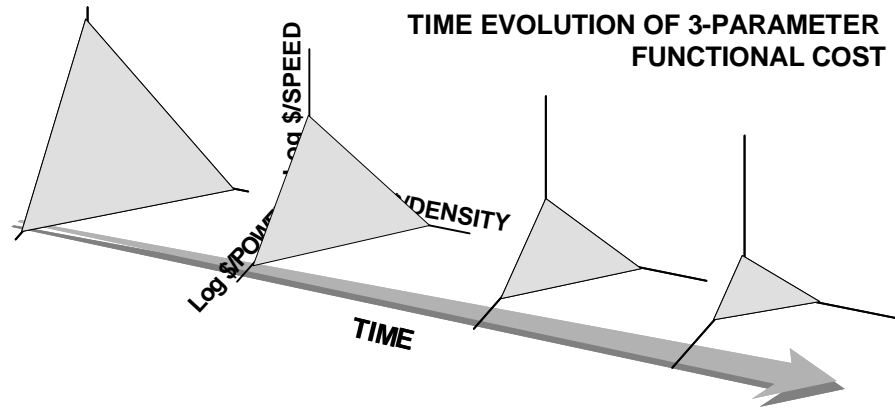


Fig. 11. Time evolution of Density, Speed, and Power functional costs.

## FUTURE PRODUCTIVITY ENHANCEMENTS

In the future, additional productivity enhancements will be required to keep functional costs dropping at historical rates for density, speed, and power. Some of the existing mechanisms may not be effective in the future; some will evolve in new directions. Problems that the industry has elided, evaded, or even unknowingly avoided for decades may rise up with exponential fury. The list below is not exhaustive, but meant to illustrate how the business challenges of the semiconductor industry are reflected in productivity improvement mechanisms.

### *Potential of Existing and Future Productivity Mechanisms*

**Yield.** As noted above, in the early years of the industry (through the 1960's and 1970's), yield increases were the key mechanism by which productivity improved. By the mid-70's, yields in volume production had reached diminishing returns near 100%, and today yield is a baseline requirement for productivity, not a contributor to productivity improvement, per se.

**Wafer Size.** The transition to the next larger wafer size is a formidable, multi-billion dollar, multi-year exercise. Whether this will continue to make sense for the industry is discussed below.

**Design.** Current estimates are that lithographic reduction and design innovation must equally share the burden on continuing MPU speed enhancements.<sup>3</sup> Chip design tools, however, are in no way keeping up with design complexity and may constrain the rate of density innovation. In addition, design tool weaknesses in other areas such as process and equipment may reduce the development rate and reliability of manufacturing improvements. As noted previously, software productivity is not advancing as fast as semiconductor productivity. Since the complexity of semiconductor chip, process, and equipment designs must be modeled using software first (because commitment to manufacturing is so expensive), a slow down in the productivity of software may

ultimately constrain the semiconductor industry. One mitigating opportunity is in the area of system-on-a-chip (SOC), which improve productivity at the system level. These products are driving new design approaches, such as proprietary functional core libraries. If hierarchical complexity isolation can be standardized through defined modules and interfaces, the industry may avoid the capability collapse of design tools.

***OEE and OFE.*** Overall equipment and factory efficiencies are both capped at a value of 1.0, so neither contributes continuously exponential productivity benefit. Like yield, they ultimately become part of “the cost of doing business.” Unlike yield, however, neither is particularly good today (40-60% or lower is not uncommon), so improvements here can contribute to industry-level productivity. As noted above, the key dynamic of the semiconductor industry is exponentially increasing functions per area with slowly rising manufacturing cost per area compensated by regular but infrequent wafer size increases. Even significant improvements in OEE and OFE could not eliminate an entire technology node, but they can slow the rise in manufacturing cost so that the next wafer size increase is delayed. This is a significant result and worth the industry efforts underway to push OEE and OFE to their limits.

While OEE has been defined (and standardized) for some time and is well understood (if not always well executed), OFE<sup>14</sup> is a newer concept and is associated with the more arcane science of manufacturing operations. No matter how well the chips, processes, and equipment are designed, a poorly executed manufacturing layout, equipment scheduler, lot dispatcher, or automation system can drive up manufacturing costs dramatically. ISMT data shows that fabs of similar age, technology, and product mix can have dramatically different fab performance indexes.<sup>15</sup> Until recently, the industry has not collectively addressed this issue. Today the ITRS has a significant section describing requirements in this area, and ISMT has programs specifically targeted to improve manufacturing methods and productivity. These range from driving operations research into commercialization to distributing best-known methods to our member companies for use in current fabs. It should be noted that the 300 mm wafer generation has been conceptualized to support several key paradigms that improve fab efficiency: (1) total hardware and software automation so that the factory control system moves wafers, downloads recipes, starts and stops equipment, and uploads data; (2) mini-environments in equipment and wafer carriers so that environmental defects are reduced and the cleanliness requirement in the fab is relaxed; (3) e-Manufacturing equipment capabilities for using factory networks and the internet to enhance fab and equipment performance, including e-Diagnostics, which enables suppliers to track equipment performance from their offices, predicting failures, optimizing maintenance, and lowering service and repair costs.

***2-D Scaling → 3-D Scaling.*** The limits of 2-D scaling have been “right around the corner” for more than a decade. Sophisticated modeling today more meaningfully establishes the nominal limits of 2-D devices, but within the same formulation must begin the introduction of non-planar constructs to compensate for density constraints. Although there are formidable power and electrical issues to resolve in designing 3-D devices, if manufacturing costs can be controlled, productivity can be maintained. Unfortunately, the industry’s ability to model future manufacturing costs is essentially nonexistent, thus introducing significant risk of an unanticipated productivity collapse.

***Non-Optical Patterning → Self-Arrangement and -Replication.*** If the issue of forecasting manufacturing effectiveness is overcome, the industry can narrow the scope of research and development on post-optical patterning, greatly increasing the likelihood of a successful business solution. While molecular physics and thermodynamic considerations must be accommodated, they are intrinsically modelable (using the faster computers of the future!), so that an era of atom-by-atom designer circuitry is imaginable. Alternatively, non-solid-state (or combination) devices are as easily imagined. In any case, the density side of the industry's long productivity trend has no foreseeable risk, and optical and eventually quantum devices claim to maintain speed trends. Constraining manufacturing costs is a far more uncertain endeavor.

### ***Wafer Size – Is 450mm Wafer Manufacturing Necessary?***

The introduction of new wafer sizes is a first order productivity enhancement approximately equal to one technology node. As described above, when the effective wafer area more than doubles, but the cost of the new tool set (with the same overall wafer throughput) increases by only 30-40%, the cost per area decreases by 30-50%. A wafer size change also has a second order effect, noted above under OFE. Because all equipment and factories must be redesigned to some extent, the opportunity to introduce new factory operations, equipment capability, automation, software systems, and facilities exists at a wafer size boundary. This is a non-trivial factor that is revolutionizing the 300 mm generation of fabs. All this new engineering, however, is not without cost. Not including the costs to ultimately build and outfit fabs, the resources to execute the transition to 300 mm (the most expensive in history) can be estimated at ~\$30B (consider ~500 equipment systems requiring \$50M each in non-recurring engineering and ~10 IC companies spending \$½B each to drive early manufacturing infrastructure). While consortia and standardization were critical to minimizing this cost (and are therefore a kind of productivity enhancement themselves), this industry (any industry) cannot go through such an event very often, yet the semiconductor industry has done it about every 10 years.

The 300 mm transition was the first one for which there was a concerted, relatively visionary forecasting and planning of the new generation. How well it went, history will tell. However, the paradigms of manufacturing (those that drive the basis cost/chip) from 200 mm are still mostly intact at 300 mm: round silicon wafers, optical patterning, layering, and etching. In the 300 mm era, the industry will move to wafer-level from batch processing; fully automated, mini-environment fabs; and intensely software-driven factory control. The next productivity boost is forecast in the ITRS to be 450 mm wafers beginning in the early-mid 2010's. The engineering and standardization for that generation will take at least as long as 300 mm (it will be harder, but the industry will be smarter)—about 7-8 years. If production is to begin in 2014, then by 2006 the industry must begin engineering. This means there are 4 years that the industry has never had before—the precious time to decide if there is a better way. During this time, the thinking and earliest experimentation for revolutionary paradigms of manufacturing must be driven. In contrast to the density, speed, and power research (already in place), this new research must address issues such as: continuous flow processing (ribbons or rods or fluids/gases); non-crystalline and/or reusable substrates; high-speed single chip processing; the rise and impact on manufacturing of the micro- and nano-machine eras; mass customization of chips and macro- vs. micro-sized fabs; the issues of technology enablers in other segments of the electronic system supplier chain (packaging, software, communications infrastructure); and the global business issues associated with the magnitude and rate of investment to enable it all.

## SUMMARY

Nothing fundamental impedes the semiconductor industry's maintaining productivity growth near historical levels for the next 10 years. It will, however, become increasingly critical to consider business and manufacturing issues as we assess long-term productivity mechanisms. Other related industries and other segments of the supplier chain for electronic systems must come under the same discipline of technology and manufacturing self-evaluation and roadmapping. Opportunities exist now to alter the trajectory of future manufacturing costs, enabling future technology solutions heretofore unimaginable.

Perhaps the upper bound on the delivery of "thinking productivity" to the world occurs when a significant portion of the gross world product pays for the full human bandwidth delivery of information to and from every person anywhere on earth. If so, we are only setting the groundwork for attaining it, and our children will tell us how we did. Even our imaginations end there.

## ACKNOWLEDGMENTS

The authors would like to acknowledge their ISMT colleagues on the Industry Economic Modeling team, especially Walt Trybula for insights into lithography and Bob Ruliffson for tracking down any and all the required industry data. Discussions within various ITRS working groups have provided significant stimuli the ideas presented here.

## APPENDIX A

### *International SEMATECH Industry Economic Model*

To begin the process of modeling industry productivity trends (historical and future), the major contributing factors (price, function) requires characterization and segmentation. DRAM was selected as the prototype for this modeling effort since its product sales and process evolution were well researched, documented and generally accessible to the public. Consequently transistors (=10 DRAM bits) became the nominal metric for measuring function and FEOL wafer processing cost which historically accounted for over 50% of the average selling price became the nominal metric for price. Calculating DRAM productivity was relatively straight forward since its sales were recorded in density (# of bits) groupings and until recently products and technology processes were tightly coupled allowing for the calculation the total "cost of goods sold" based on modeling the wafer processing cost with the SEMATECH cost resource model (CRM). Finally, time sensitive cost parameters were added for probe yield and equipment throughput learning based on consultation with SEMATECH member companies and primary research of public documents.

Expanding this methodology to the rest of the industry required some significant modifications since the Logic, Analog and Discrete products contributing factors (price, function) were not easily grouped or identified. Since all semiconductors emanate from silicon, area demand was selected as the common denominator for creating an industry productivity model. Semico Research provided historical (1994-2000) and forecasted (2001-005) technology area demand for each SIA category. Products were than aggregated into major groups around common attributes, leading edge (memory, processors and custom logic) and trailing edge to minimize the inconsequential

aberrations of small samples. Stylized technology distributions were developed from actual observation and correlated to the technology roadmap to provide a vehicle to protract manufacturing requirements throughout time. The next step, determining the fab population by wafer size, required the development of an allocation simulator that not only calculated the number of fabs required (after retires due to obsolescent or age) but how they would be constructed (build new or an upgrade of an existing structure). Simulation for each leading edge group is executed in parallel applying throughput learning curves based on fab, wafer size, technology aging and the type of construction. Fab population for trailing edge products are then developed based on any wafer falling of existing facilities / equipment from the leading edge groups or the construction of new entities. Next, to facilitate ease of use and common tools, a simplified manufacturing cost model was developed for each product group by technology, wafer size and construction type to generate the total equipment markets and overall processed wafer cost for the industry. Area per transistor are then assigned to each product group based on primary research of press releases and technical papers, and trended consistent with the technology roadmaps. To complete the equation, total transistors supply is then calculated based on the area demand and overall product yield based on Semico Research original estimates with learning curves comparable to throughput.

## REFERENCES

<sup>1</sup> Michael Cox, "The New Paradigm," *1999 Annual Report*, Federal Reserve Bank of Dallas.

<sup>2</sup> Calculation outline: 500M tons of rice per year = 450B kg/yr. Informal measurements yielded ~60,000 grains/kg (dry). The resulting annual number of grains is ~27 quadrillion. The number of DRAM bits alone manufactured in 2002 will be ~1 quintillion (=1,000 quadrillion).

<sup>3</sup> All references to the *International Technology Roadmap for Semiconductors* (ITRS) are from the 2001 version. See <http://public.itrs.net>

<sup>4</sup> All references to and results from the ISMT Industry Economic Model (IEM) are version 3.0. For more information, see <http://www.sematech.org/public/resources/econmod1>

<sup>5</sup> For example, during the early 1990's DRAM prices were not falling at historical rates due to PC-driven increase in demand and a temporary, regional, near-monopoly in volume production. As capacity came on line around the world, prices subsequently collapsed. Even across this dramatic set of business dynamics, bit demand held fairly steady and price per bit eventually returned to historical trends.

<sup>6</sup> Alan Allan developed the modeling concept of the "133 mm" wafer size between 100 mm and 200 mm. This pseudo-generation accommodates the 125 mm and 150 mm generations and reveals the underlying ~10 year wafer generation timing.

<sup>7</sup> This is measured with some rigor in DRAM as the so-called *A* factor, which relatively assesses DRAM chip designs for density improvements *not* due to merely shrinking the minimum feature size (referred to as *f*). It also is loosely considered as the number of unit-features (of the minimum feature area,  $f^2$ ) that constitute one bit.

<sup>8</sup> Shekhar Wadekar, private communication. In addition to system software, each smart semiconductor chip (e.g., MPU) incorporates embedded software ("microcode") that

---

flexibly controls its hardware sequencing. This critical, but invisible, software is presently “given away” with the semiconductor device, so that the potential for revenue sharing with system software is nullified.

<sup>9</sup> Dale W. Jorgenson and Kevin J. Stiroh, “Raising the Speed Limit: U.S. Economic Growth in the Information Age,” *Brookings Papers on Economic Activity*, Issue 1, 2000.

<sup>10</sup> D. Rose, “The Future of 200 mm Wafers,” *SEMI/JEIDA Joint Technical Symposium: What’s Ahead for 200 mm Wafers?*, SEMICON/West 1993, pp. 11-28. This was an early attempt to illustrate these ideas.

<sup>11</sup> Denis Fandel and Robert Wright, To Be Published: “Modeling the Overall Semiconductor Industry Supply Chain,” *Proceedings of the International Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM 2002)*.

<sup>12</sup> Alan Allan, developed for the International Roadmap Committee of the ITRS. It is an analysis of the historical data from the International SEMATECH Competitive Analysis Group database.

<sup>13</sup> From the Intel web site: <http://www.intel.com/intel/intelis/museum>

<sup>14</sup> A SEMI standard for Overall Factory Efficiency (OFE) is presently under development within the SEMI Standards Metrics Committee. For more information, see <http://www.semi.org>

<sup>15</sup> Mike Schwartz, ISMT Manufacturing Methods Council Project Manager, private communication. The confidential fab performance data collected by this council is analyzed for their exclusive use, including the computation of a fab performance index.