Semiconductor Technology & Manufacturing Status, Challenges, and Solutions – A New Paradigm in the Making

C. R. Helms

International SEMATECH
Austin, Texas
Present Address: University of Texas at Dallas
P.O. Box 830688, EC 32
Dallas, Texas 75083-0688

Abstract. The phenomenal growth of the semiconductor industry has been made possible by our ability to deliver more functionality at a lower and lower cost, with a reduction in cost per function of approximately 25% per year. Technology advancements have also enabled higher performance, lower operating voltages, and associated lower power consumption. Dimensional shrinks based on 248 nm lithography led the revolution in recent years. However, with the fundamental limit for optical lithography at about 1/3rd the wavelength, the limits for 193 nm and 157 nm lithography are about 65 nm and 50 nm respectively. With EUV and most other NGL technologies 7 years in the future the limit for the remainder of the decade is about 50 nm. Technology challenges, whether they are lithographic in nature or based on needed new materials with advanced properties, are summarized in this paper along with suggestions for potential solutions for the remainder of the decade. Even if we meet the technology challenges, continued revenue growth and profitability will become more and more challenging. The need for larger and larger technology R&D budgets may make it difficult for the current number of semiconductor manufactures and equipment and materials suppliers to remain profitable. The solution to this conundrum is now clear – partnerships and collaboration. Meeting this challenge of creative cooperation with existing and new partnerships is the new paradigm that is discussed here in some detail.

INTRODUCTION

Growth in revenue and continued and improved profitability is a benchmark for any industry, leading to high valuations with significant multiples. We have seen enormous revenue growth in the semiconductor industry with a compounded annual growth rate (CAGR) of 17% over 30 years in the 60’s, 70’s, and 80’s. This growth was comprised of two components, growth in unit shipments and growth in average selling price (ASP). Unit shipments increased at a CAGR greater than 10%, with ASP’s increasing at a rate greater than 5% during this period. At the end of the 80’s the industry had become a major force in the global economy with revenues of $50B on shipments of 120B units.

The 90’s saw some major shifts in the industry. DRAM’s, which had been the critical driver for technology advancement and revenue growth, were replaced by the microprocessor. Various pricing pressures also led to stagnation in the growth of ASP’s that peaked in late ’95. On the technology side improvements in manufacturing productivity also began to saturate leading to a wafer size change from 150nm to 200mm and an increase in the shrink rate from a 30% reduction every 3 years to a 30% reduction every 2 years. Significant improvements in lithography enabled this revolution, with the move to the 248 nm wavelength and a move from steppers to scanners late in the decade.

Given the technical and business challenges, exacerbated by the poor global economy in 2001 and 2002, some argue that the heyday of semiconductors is over and that the industry needs to focus on a situation where minimal unit growth will continue to be compounded by falling ASP’s, leading to a relatively flat revenue picture. There is little question that the 17% CAGR seen over 3 decades cannot not be sustained indefinitely. Indeed, the semiconductor industry is beginning to show signs that it is entering into a more mature phase.
How well each company or institution responds to these business and technical challenges will determine success or failure. Speed of technology development and introduction at the lowest possible cost will be a critical differentiator. Both of these factors are driving business and technical leaders to enter into partnerships and alliances that would have been unthinkable even a few years ago. Those who embrace this new paradigm will be successful and be able to ride the wave of revenue growth I believe we will continue to see through the end of the decade.

The first topic of discussion here will be the revenue picture of the semiconductor industry and what history can tell us about the future. Industry roadmaps will be our next topic, but with a slightly different focus than we normally think of from the International Technology Roadmap for Semiconductors (ITRS) [1]. The discussion of roadmaps naturally leads to the discussion of technology challenges, the R&D cost to overcome these challenges, and finally the move to R&D partnerships.

**SEMICONDUCTOR REVENUE HISTORY & EXTRAPOLATION**

This topic has been discussed in numerous forums over the years. A critical strategic question for business leaders is how fast will the semiconductor industry mature and thus, what level of revenue growth is likely. Given that such phases for other industries such as steel, autos, electrical power generation and distribution, telephone communication, etc. can take many decades, I would argue that although we may be in the early stages of this maturation, we have a long way to go at very healthy growth rates. To see the evidence for this statement, let's first look at the long-term picture for revenue growth shown in Fig. 1. At first glance the revenue curve appears linear on this graph. Indeed for the 70’s well into the 80’s there is a linear CAGR of 17%. However, if that line were extrapolated we would find it crossing $500B by 2005. Even the wildest optimist would not predict $500B in 2005! Furthermore, it is not difficult to see how some believe that the revenue picture has been flat and may continue in the same vein. If we look at the revenue picture since 1995, discounting 2000, one can easily draw that conclusion.

**FIGURE 1.** Total semiconductor industry revenue from 1970 to 2002 is plotted as the solid diamonds. A model based on the maturation of industries such as autos is shown as the line, with a cyclical component as the gray diamonds.
However, that analysis is clearly somewhat artificial and it is difficult to understand how the CAGR of a whole industry would change from around 15% up through the mid 90’s to zero all in one year. The maturation cycle for most industries is many decades.

We are beginning to see a maturation for the semiconductor industry, which is evident if we inspect Fig. 1 carefully. The line drawn through the data is a model based on saturation in growth not unlike what other industries experienced in the last century (i.e. autos). The “model” indicates that the growth rate was indeed 17% on average in the 70’s, 16% in the 80’s, 14% in the 90’s and projects to be 10% on average for this first decade of the new millennium. The cyclical nature of this growth industry is also apparent, with a periodicity of 5 years. For example local peaks (or inflections) occur in 1974, 1980, 1984, 1990, 1995, and 2000, approximately every 5 years. If we extrapolate to 2005, near-term growth rates should reach or even exceed 20% per year. A word of caution is appropriate in that these models are based solely on historical analysis and extrapolations and could be impacted by major economic forces external to the semiconductor industry itself. I note, however, that there are a number of predictions in a similar range [2].

From the above discussion I draw three conclusions. First, we are clearly seeing the beginnings of maturation of growth of the semiconductor industry. Second, this maturation will take place relatively slowly leading to expected average revenue growth rates around 10% per year through the end of the decade as a likely scenario. Third, the cyclical nature of the industry should lead to revenue growth averaged over the next 3 years of around 20% per year as we move from a local bottom in late-2001/early-2002 to another top in 2005. We can only hope that the next downturn predicted for the latter half of the decade will not be quite so rapid (and quite so deep) as the one we are currently experiencing.

Revenue growth has been driven historically by deeper penetration of markets by continuing price reductions on the one hand and by the development of whole new classes of products enabled by lower costs, better performance, high memory capacities, and lower power consumption. In the next section I will discuss projections of these success factors based primarily on the ITRS, with additional reference to projected product requirements.

REALIZING THE RIGHT ROADMAP

The semiconductor industry is probably the most advanced in history in establishing technology requirements and schedules through a global, formalized technology roadmap, the International Technology Roadmap for Semiconductors (ITRS). It is useful, however, to expand the discussion all the way up the value chain to semiconductor products and end equipment. A roadmap hierarchy, including these factors is shown in Fig. 2. From the perspective of the integrated device maker (IDM) or fabless semiconductor house the end equipment markets they choose to serve and the products that they will sell into these end markets are, by far, the most important considerations. The hierarchy continues down through modules that will be needed for each chip, the high level specifications associated with the modules and overall chip, the architecture to be used, the materials and structures needed to support the architecture, and finally the processes and tools needed to deliver the materials and structures. Working our way down the hierarchy drives the requirements level-by-level, leading to solutions as we move back up the hierarchy. Out of the 7 levels shown in the Figure, the ITRS deals primarily with the bottom four. It is important to realize that an IDM that can meet or beat the requirements associated with the top 4 levels of hierarchy will win independent of the details of the technology solutions provided. This is where much of the competitive advantage for IDMs lies today.

In the context of an organization like International SEMATECH or the equipment and materials suppliers, “realizing the roadmap” normally applies to the lower two categories of the hierarchy. A focus on the bottom of the value chain in this manner can, however, lead to significant risk. If a solution is found higher up the value chain, the there may be no need for a lower level requirement, let alone solution. The best example of this occurring in recent years is for low-K interconnect dielectrics. The 1997 ITRS indicated a REQUIREMENT for interconnect structures this year (2003) with effective dielectric constants in the range of from 1.5 to 2.0. Indeed the use of such effective dielectric constants in mainstream technologies is years away (possibly not even in this decade). If this were a true requirement, we would not be shipping 130nm product today let alone 90/100nm class products this year. So, when we look at lower level roadmaps we need to realize that there is a possibility that a breakthrough at a higher level may mitigate the need for a specific requirement or solution at the lower level.

Although the top levels of the roadmap hierarchy shown in Fig. 2 are beyond the scope of this paper a few comments are appropriate as paradigm shifts at the higher levels do have a major impact. The most significant change that we see is the move to more integration, either in package or on-chip.
The integration of processor cores with on-chip SRAM (and flash) has been a common practice for a number of years. Logic with embedded DRAM, even given all the process complexities is also appearing in mainstream products. Stacked die are a serious contender for applications that need powerful processing on the one hand and significant capacity for fast DRAM. More and more analog functionality is also appearing in hand-held and wireless chips and some believe that a true system-on-a-chip (SoC) will arrive soon (single chip phone, radio, etc.). The move to SOI, Si-Ge, and alternate memory technologies such as embedded MRAM, FeRAM, and other alternate memories may accelerate this move toward SoC’s.

With this in mind, we will however limit the discussion here to the bottom four categories of the roadmap hierarchy, noting that cost - both manufacturing and R&D cost - is a natural part of this discussion. A critical part of the roadmaps and, indeed, a critical driver for many of the improvements delineated in any roadmap is related to cost reductions that have historically been necessary to open new markets and drive further penetration in existing markets. If we combine the potential for revenue growth and the cost associated with this revenue growth we have almost the totality of the financial model for a chip maker. Thus, in many ways, the cost roadmap may be the most critical of all. Manufacturing cost (per chip, per function, and per unit area) and R&D cost are the two dominant factors that must be considered. The former must be integrated over the life of the facilities and equipment used and also must comprehend fluctuations in factory loading that will lead to an average utilization well below 100%. The $3+B cost of a new 300mm factory has now limited the number of companies that can afford to be independent device manufacturers down to the 5-10 range (>5B in semiconductor revenues). We will discuss technology R&D cost later. However, this may be even a more serious limitation for firms that seek to be tier-1 leading edge producers [3].

The model developed at International SEMATECH to assess manufacturing cost breaks the problem into two fundamental components functionality per unit area and cost per unit area [4]. Furthermore if we assume that functionality is related to the number of transistors, then it is easy to analyze the cost per function by comprehending the cost per unit area and transistors per unit area and dividing one by the other. Again caution is necessary here. If more functionality can be delivered with the same number of transistors, this simple analysis will be flawed. Thus any reduction in the number of
components for a particular application enabled by SoC or system-in-package (SIP) concepts will generate significantly improved functionality per chip or per unit area without the need for either increased transistor density or reduction in manufacturing cost per wafer or per unit area.

Another example is the move from processors with embedded SRAM to chips with processors plus embedded DRAM. Typical SRAM cells contain 6 transistors (6T cell); if that number can be reduced without a significant cost sacrifice, the cost per function will obviously drop. One approach is to replace much of the SRAM with embedded DRAM; unfortunately the process integration challenges here have not provided the ability to use this technique for many mainstream applications, since lower yields and longer development cycle-times mitigate the cost saving.

Until about 1995 a good part of the improvement in cost per function was driven by improvements in fab and process productivity through yield improvements and overall improvement in equipment productivity. Two other important factors were wafer size changes that occurred about every 10 years and shrinks that provided a 2x improvement in transistor density every 3 years. The improvement obtained by a combination of these factors was between 20 and 30% per year in cost per transistor. In the mid-90's, however, the low hanging fruit for global yield improvements and equipment and overall fab productivity had been harvested. Furthermore, the move to 200mm wafers was well underway, but improvements possible from the move to an even larger wafer size were a number of years off. Thus, the year-by-year rate of cost per function decrease had fallen and was in the low 20% per year range.

As we approached the mid-90’s the industry was losing one of its historical growth advantages, the 25+% per year cost per function reduction. To restore the productivity curve to this trend, the industry accelerated the shrink roadmap from a 3-year cycle to a 2-year cycle. Many argue that this has not yielded to benefits promised; however, in my view, the benefits are clear. We can see this by looking at a cost comparison analysis shown in Figure 3.

We start with the premise that a 25% per year reduction in cost per function is the target and seek to understand how the various factors will lead to this number. The advantage of the 2-year roadmap cycle in terms of manufacturing cost is obvious since the rate of increase in functions (transistors) per unit area goes from 25% per year for the 3-year cycle to 40% per year with the 3-year cycle! All of this benefit is not realized however, since there will be differential manufacturing cost increases between the 2-year and 3-year cycles. International SEMATECH studies suggest that the manufacturing cost at constant wafer size has increased at about 4% per year for a 3-year cycle (since about 1990) and increases to 8% per year for the 2-year cycle. In the 70’s and 80’s the manufacturing cost per unit area had actually been decreasing due to significant improvements in yield, yield ramps, and overall fab productivity advances.

<table>
<thead>
<tr>
<th>Productivity Factors</th>
<th>2-Year Cycle</th>
<th>3-Year Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required Annual Manufacturing Cost per Function Decrease</td>
<td>25% per Year</td>
<td>25% per Year</td>
</tr>
<tr>
<td>Annual Transistor per Unit Area Increase</td>
<td>40% per Year</td>
<td>25% per Year</td>
</tr>
<tr>
<td>Nominal Annual Manufacturing Cost per Unit Area Increase - Constant Wafer Size</td>
<td>8% per Year</td>
<td>4% per Year</td>
</tr>
<tr>
<td>Average Annual Reduction for Wafer Size Conversion Every 10 Years</td>
<td>4% per Year</td>
<td>4% per Year</td>
</tr>
<tr>
<td>Predicted Annual Manufacturing Cost per Function Decrease</td>
<td>27% per Year</td>
<td>21% per Year</td>
</tr>
</tbody>
</table>

FIGURE 3. Comparison of cost factors between a 70% every 2 years and 70% shrink every 3 years.
As discussed above, however, the industry has matured to the point where these factors no longer provide the benefits of the early years. Finally, we can put the complete cost picture together by comprehending the long term, average benefits of a wafer size change every 10 years, which are found to be about 4% per year.

This analysis shows that the 3-year cycle can be expected to generate a cost reduction of 21% per year compared to the 27% per year cost per function reduction of the 2-year cycle. Figure 4 shows the benefits of both the move to the 2-year roadmap cycle as well as the wafer size change for cost per transistor for leading edge logic products from 1995 to the present and projected out to 2005. The diamonds are results from analysis of data available and extrapolations into the future for the lowest cost node (and wafer size) for any given year for volume production. The trend-lines show a slope that is about 20% per year from 1995 through 1998 and drops to 27% per year from 1998 to 2005. The breakpoint in 1998 is due almost completely to the move to the 2-year roadmap cycle (the 0.25|Lim node was the first node to be developed in 2 years). The 300 mm wafer size is projected to be the lowest cost manufacturing solution in 2004 for 130 nm technology. This projection, made in 2001, is consistent with recent information from the leading edge chip-makes with 300 mm factories [5]. Thus, the need for both the 2-year roadmap cycle as well as a wafer size change every 10 years is well established if we are to continue to achieve a 25% per year manufacturing cost per function.

Unfortunately manufacturing cost is not the only issue. The increased rate of research and development needed to support the accelerated roadmap cycle and its associated cost can be a mitigating factor. The technical challenges faced by our R&D communities will be discussed next; the R&D cost situation will be discussed later in the paper.

TECHNOLOGY CHALLENGES

The discussion to this point has focused on the need to shrink transistor dimensions at an ever-increasing rate which was 11% per year for most of our history (3-year shrink cycle) and is now 16% per year (2-year shrink cycle).

![Figure 4](image.png)

**Figure 4.** Cost per transistor is shown on a relative scale for leading edge logic products. The black diamonds are the results of the model calculation, validated by industrial average data. The large diamonds indicate a cross-over year where the latest technology has matured to the point of being lowest cost (i.e. 0.25 micron technology became a lower cost technology than 0.35 micron technology in 1999). The gray diamonds are an extrapolation. Note that the year when 300 mm manufacturing is predicted to be lowest cost is 2004. A break point in 1998 due to the move to 2-year shrink cycles is indicated.
This has been accomplished over the last decade, even though the critical level exposure wavelength has remained at 248nm. This >3x improvement in resolution has been enabled by a 50% increase in the numerical aperture of the exposure tools, process control advancements, and numerous resolution enhancement techniques (RET's), including aggressive phase shift mask technologies, with a reduction in the K1 factor by nearly a factor of 2. Thus what was once thought as an insurmountable barrier – critical dimensions less than the wavelength – was broken in 1995. The year 2002 saw the ½ wavelength barrier also broken. The history and projections are shown in Figure 5. However, given that the industry has a history of breaking barriers once thought to be impenetrable, the ultimate limit as a fraction of the lithography wavelength is not clear. The ½ and 1/3rd wavelength regions are, however, indicated in the Figure.

The reduction in critical dimension as a fraction of wavelength, especially given the difficulty driving 193 nm technology into the mainstream until mid-last year, and potential delays for 157 nm, are providing pressure for more aggressive RET’s and mask technology. Thus, masks - availability, cost, and cycle time – are considered the #1 technical challenge faced by the industry today and in the near future [1].

Mask and the other top technology challenges for the next few years are listed in order of perceived importance and difficulty:

1. Mask availability, cost, and cycle time.
2. Post-193 nm exposure tools and infrastructure.
3. Compatible materials and processes to enable continued improvements in transistor performance and power consumption.
4. Design methodologies and interconnect materials and processes to enable continued improvements in circuit complexity, wiring density, and performance.
5. New materials, processes, and structures to enable the move to more on-chip integration at a low cost. (i.e. low cost embedded memory).

As just discussed the advances in mask technology needed are significant indeed, exacerbated by the delay in mainstream application of 193 nm technology and the risk of additional delays in the availability of post-193 nm exposure systems and infrastructure.

---

**FIGURE 5.** 2-year node dimensions are plotted as a function of time as diamonds; the trend-line is also shown. The dark line is an indication of the availability of lithography wavelengths in the past, with projections for the future. Regions where the critical dimension falls below ½ and 1/3rd of the wavelength are indicated.
This is driving not only mask challenges but also changes in design methodology and architecture that may be necessary to achieve the needed increases in performance and functions per unit area. Given the relatively constant field size and magnification factor, the increased feature density, improvement in resolution of features on the mask, and improved tolerances are providing a significant technical challenge. Coupled with the desire to have cycle times that are measured in hours, the task is daunting indeed. The use of significant optical proximity correction (OPC) and more recently attenuated and alternating phase shift has also increased the complexity. There are significant issues across the entirety of mask-making infrastructure including pattern generation (mask writers), mask inspection, and mask repair. For non-classical technologies (i.e. EPL and EUV) fundamental materials issues are also on the critical path. The introduction of phase shifters in mainstream technology has led to the requirement to project features that may be \(< \frac{1}{4}\) of the wavelength (even if they are not resolved directly).

All of these issues have led to poor yield, difficulty in achieving first-pass success, long cycle times, and associated high costs. 130 nm-class technology mask costs have ranged up to over $1M and are on the order of $500K today. There are projections that the costs could continue to escalate, doubling every generation (per 30% reduction in linear dimension). Thus a production mask set for 90nm-class technology could reach $2M this year dropping to $1M in 2005; a set for 65nm-class technology would be $4M in 2005, dropping to $2M in 2007, and so on.

The mask cost is, of course, important for all technologies. However, for large volume products that may require thousands or tens of thousands of wafers the overall impact is significantly reduced. For a $1M mask set and a 10,000 wafer run, the cost associated with the mask set is $100 per wafer. This may not be desirable, but it is not a total showstopper. Thus for large volume microprocessors, DRAMs, and chips for consumer products, the mask cost can be accommodated. For product runs that consume fewer wafers, the mask cost alone may drive to other lower cost approaches. Some of the tradeoffs are illustrated in Figure 6. The fabrication cost per unit area including one full mask set is plotted vs. the number of wafers in the run (see the Fig. caption for assumptions). The cost cross-over for the assumptions used is about 500 wafers between the advanced technology solution (solid black line) and the solution based on the previous node technology (and thus lower mask cost). Note this comprehends only the manufacturing cost for one mask set; any additional masks for prototypes, etc will move the cross-over to larger run sizes.

The cost for another solution, based on a gate array is also shown. This analysis assumes that a gate array with appropriate functionality can be designed using the advanced technology with only a 30% area penalty (and no important performance penalty), with only 10% of the cost of the mask set associated with the particular custom product. In this case the cross-over is at about 1500 wafers and it always provides a lower cost than the full custom solution using the previous node technology. Recent development and product announcements seem to indicate that this strategy is gaining momentum for smaller wafer runs.

The rapid technological advancement that is needed in pattern generation, mask inspection, mask repair, and handling of large databases is also leading to economic challenges for those firms that choose to service this segment. The limited revenues available to these equipment and software suppliers is beginning to fall short of what is required to pay for the associated R&D cost to generate an appropriate return.

Is there a solution? We believe that the answer is yes. Mask cost (and cycle time) can be mitigated by yield improvement. And the financial issues associated with R&D at the mask maker as well as R&D at their suppliers can also be mitigated by strategic partnerships and consortial activity. As I will discuss below, however, firms must be willing to trade what would once be considered valuable intellectual property for reduced R&D cost and accelerated development cycles.

The challenges for masks relate directly to the overall lithography problem. Looking back at Fig. 5, I am not projecting large scale use of 157nm until 2006 or EUV until 2010. These projections are more pessimistic than other published reports but are consistent with the development cycle for 193nm where I do not put large scale use until sometime last year (2002). Mask challenges are clearly a major part of the barriers to adoption of these technologies. However, there are other significant challenges as well related to exposure tool materials and other infrastructure issues.

The final technology challenge from the list above is the need for new materials with improved properties to respond, in turn, to the need for improved performance, power consumption, and smaller dimensions. This need for new materials is increasing every year. We have seen a move to tungsten for first level metal and local interconnect, titanium and cobalt silicide for contacts, copper interconnects, lower dielectric constant materials for interconnect and barriers, and nitrogen-containing dielectrics for higher dielectric constant applications. SOI substrates and
strained silicon and Si-Ge will be in the mainstream in the near future.

One of the most significant challenges is the need for a MOS dielectric stack with sufficiently small effective oxide thickness and parasitics on the one hand, but sufficiently low leakage currents on the other. Nitroxides will suffice for the 90nm node, but a replacement is highly desirable for the 65nm node, with production in 2005 [6].

In summary, we see the technical challenges of the 90nm node being solved with 193nm lithography, significant use of phase shift mask technology, nitrooxide gate insulators with cobalt silicide contacts, and Cu interconnect with nominal low K dielectrics. The 65nm node will be a significant challenge, especially if we expect to see significant production volumes in 2005. 157nm lithography will likely not be at the maturity level needed and the new materials solutions for both front end and back end may also be lagging. The technical challenges for the remainder of the decade will be no less daunting, especially for lithography. Breakthrough solutions such as electron projection lithography, immersion lithography or perhaps nano-imprint technology show some promise, but the time window is closing. A significant level of R&D resources will also be needed to deliver these technologies. That is the topic of discussion for the next section.

CHIP-MAKER R&D COST: THE RISE OF R&D PARTNERSHIPS

Even with improved productivity and co-located technology R&D and leading edge manufacturing facilities, R&D costs are increasing at a significant rate. The factors are numerous, but the two dominant factors in the recent past are the move to the 2-year roadmap cycle and the acceleration in the introduction of new materials and lithography wavelengths. The move of R&D to the 300mm wafer size is also adding a significant cost as facilities are converted or built for this purpose.

It is fairly easy to estimate the impact of the move to the 2-year roadmap cycle. Assuming that the R&D effort per node is comparable, the number of nodes that are in active development increases by simple mathematics by 50%; thus the differential R&D cost will be 50% higher! For a large volume producer the

FIGURE 6. Example cost per wafer is shown vs. the number of wafers per mask. The black solid line refers to a technology with a $1700 cost per wafer w/o the mask set and a cost for the mask set of $1M. The black line with the large dashes refers to the same product run at the previous node with a $1400 cost per wafer and $500K cost per mask set (twice as many wafers are needed; the plot is shown in equivalent advanced technology wafers to comprehend the additional factor of 2). The gray dashed line refers to a product run with the advanced technology using a "gate array" approach assuming comparable functionality with a 30% area penalty and only 10% of the cost of the mask set custom for a particular product.
improved manufacturing cost may well offset the increased R&D expenditures, but there will clearly be a chip-maker size (depending on product mix) below which the additional R&D expense cannot be justified.

The actual technology R&D cost for a leading edge chip-maker is a complex function of what segments of the market it serves. However, estimates place the baseline for silicon technology R&D at $150M per year in 1997, $250M in 1999, and $350M in 2001. This increase has been over 20% per year. A part of this increase is due to the move to a 2-year roadmap cycle. This has clearly added 5% to 10% per year to these increases. However, the need to move R&D to 300mm over the next few years will apply continuing pressure, so the 20% increase per year may project through a good part of the decade. Thus, targets for 2003 and 2005 of $500M and $750M per year are not out of the question.

Unfortunately, for an integrated device manufacturer, technology R&D is not even the dominant R&D cost, and may account for only 1/3rd of the total. This places the revenues needed to support leading edge silicon technology R&D in the $5B to $10B range, similar to that needed to support the move to 300mm manufacturing. This, in turn, limits the number of firms capable of a completely independent R&D operation from 5 to 10. R&D partnerships and strategic alliances are one way to reduce this cost burden.

The concept of strategic alliances and partnerships to mitigate cost and provide a more coherent voice to drive the industry agenda is a concept that has been practiced for 20 years or so. The most well known early programs were the formation of the Semiconductor Research Corporation (SRC) and SEMATECH in the mid-80's. One of the best known programs centered at an individual company was the memory partnership at IBM, including Infineon (then Siemens) and Toshiba.

However, the high cost of both manufacturing and R&D has made the participation in such partnerships an economic necessity for many companies that are not in the $5B to $10B revenue club. International SEMATECH in the US, Selecte in Japan, and IMEC and CEA Leti in Europe all have multi-company programs housed in a central facility. There are numerous academic programs and both federal and state governments are supporting R&D as well.

The biggest paradigm shift we see is the willingness of competitors to trade what had been the potential for valuable intellectual property for the potential of lower cost R&D and also possibly accelerated development cycle times. In the semiconductor industry three large programs have been announced recently: one centered at STMicroelectronics in Crolles, one centered at IBM in New York, and one managed by Dupont Photomask in Dresden. The first involves Motorola, Philips, and TSMC, and includes participation by CEA Leti; the second involves Infineon and AMD, with possible participation of SUNY Albany; and the mask program involves AMD and Infineon. Thus we are seeing a major strategy shift to R&D partnerships and alliances to reduce cost and accelerate development cycle times. As 300mm manufacturing matures it is conceivable these R&D partnerships will also develop into partnerships for volume manufacturing.

**SUMMARY**

In this paper I have tried to provide a snapshot of the silicon manufacturing segment of the semiconductor industry from both a technical as well as business prospective. Indeed, both are intimately tied together. Probably the most significant claim I make here is that with the exception of the top five to ten integrated semiconductor companies and the largest foundries, no one company will likely find a business model to support leading edge silicon technology R&D or the large 300mm factories needed for appropriate manufacturing economies of scale. The two choices from a business perspective are mergers and acquisitions to increase market share and revenues or inclusive partnerships and alliances. There are numerous examples of both over the last year or so.

The technical challenges that are driving the large investment in R&D relate to both the drive for smaller feature lithography as well as new materials to deliver the performance, lower power consumption, and higher densities. The top near term challenges are masks and mask cost, maturation of 157nm lithography, and low effective oxide thickness gate stacks, employing high K gate insulators.

**REFERENCES**

2. J. Feldham, as reported in the EE Times, 3/10/03 (United Business Media)
5. D. Wang as reported in the EE Times, 3/21/03 United Business Media).
6. J. Hutchby, CMOS Devices and Beyond: a Process Integration Perspective, this volume; H. R. Huff, Challenges and Opportunities for Front-End Processing in the Twenty-First Century, this volume.