The Business of Scaling

Rakesh Kumar
TCX Inc., Technology Connexions
San Diego, CA
rakesh@tcxinc.com

In addition to technical challenges, managing the economics of scaling and increasing demand have been key factors in driving the semiconductor industry to nearly $250B over the last 40+ years. The functionality per chip has increased 2x every two years[1,2]. Although the cost of wafer fabs and manufacturing has increased significantly over the years, the semiconductor industry has maintained a reduction of about 29%/year in the cost per function (CPF)[3]. This translates to a halving of the CPF every two years[1]. In this paper we will provide an overview of salient business aspects and economics of scaling.

1. Introduction

Since the introduction of the first commercial integrated circuit in 1961 and the introduction of the first microprocessor in 1971, the semiconductor industry has experienced a healthy growth of approximately 15% CAGR[4]. In the mean time semiconductor sales have grown more rapidly than the worldwide electronics sales and the worldwide GDP and are now roughly 20% of worldwide electronics sales and about 2% of the worldwide GDP[4]. Fueling the growth has been increasing demand for components for personal computers, automotive, mobile wireless and consumer products. Although the growth rate is predicted to slow down, the industry has demonstrated much resilience in combating technical and business challenges.

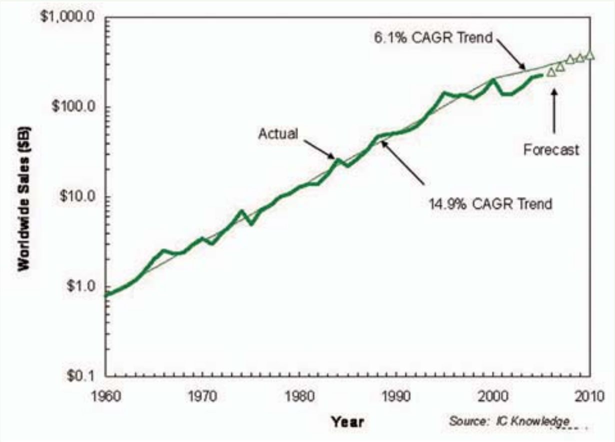

Figure 1 Worldwide semiconductor sales

Taking advantage of scaling, the industry has increased the number of components per chip steadily, as shown in Figure 2. This figure shows the historical increase in the number of transistors per chip (39% per year average) in industry leading microprocessors[4]. This trend shows a doubling of the transistors per chip every two years. This trend was predicted by Gordon Moore and has become known as "Moore's Law"[1,2]. The figure also shows the reduction of minimum feature size at an average rate of 12% per year. The number of transistors per chip has increased 6 orders of magnitude while the minimum feature size has been scaled down over two orders of magnitude during the last 35 years.

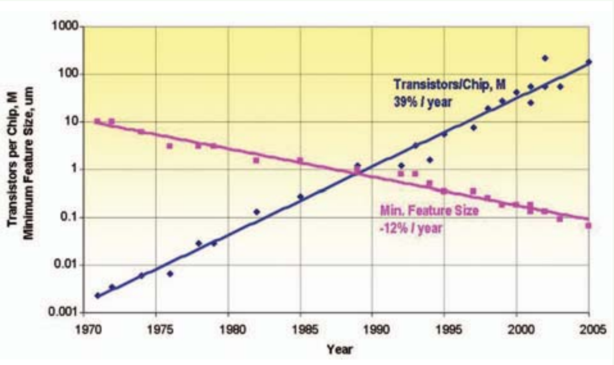

Figure 2 Historical trends of transistors per chip and minimum feature size

2. The Basic Cost Equations

The basic equation for predicting the cost of an integrated circuit die (or "chip") is:

Die Cost = Wafer Cost / Net Die per Wafer

where wafer cost is determined by factors such as facilities and equipment depreciation, materials, labor and processing cost, and

Net Die per Wafer ("NDPW") = Yield* Gross Die per Wafer ("GDPW")

Gross Die per Wafer ("GDPW") = Total usable Area on the Wafer / Die Area

Yield is a function of defectivity (or defect density) and critical area. Contributors to defectivity are usually categorized as systematic (or gross) and random defects[3]. Many different yield models have been used in the industry. Simple models, such as the Poisson and the Murphy models using the die area as the critical area were prevalent in the early days. The Bose-Einstein model using die area but identifying a defectivity per critical layer has been used extensively in recent years[8,9]. Custom models exist at captive suppliers. More recently, sophisticated calculations of critical area based on information embedded in the design database are being used to estimate yield. Examples are the number of vias and contacts in a design, the number of metal layer cross-overs, and the like. A detailed discussion of these is beyond the scope of this paper.

3. Overall Cost Reduction

A key factor in managing the business feasibility of scaling is the semiconductor industry's ability to maintain an overall CPF reduction of 29%/year[3] to 35%/year[4]. Within any given process technology node the die cost and CPF are reduced due to the manufacturing and defectivity learning curves. This is shown graphically in a conceptual chart, Figure 3. As the volume of wafer and product shipments ramps up in each technology node, there is a reduction in die cost (and therefore CPF) due to a reduction in wafer cost; this decrease is due to process optimization and the manufacturing learning curve. Also, die cost is reduced as yield enhancement efforts are implemented, defectivity is reduced, yield increases and therefore NDPW increases. A compilation of defect density trends indicates an average reduction of 19% per year over the last 35 years[4]. The technology "cross-over" occurs when the CPF in the newer technology is below the CPF in the older technology.

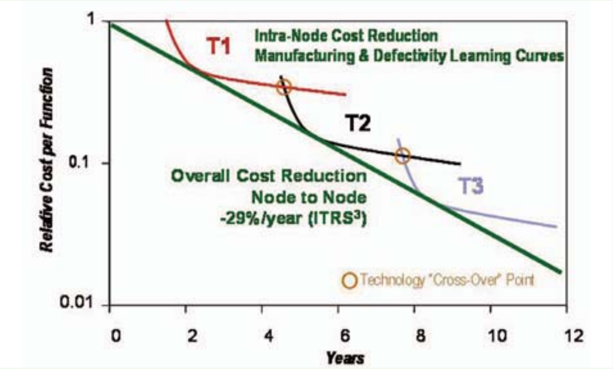

Figure 3 Cost per function and technology "cross-over" points

4. Cost Reduction from Technology Scaling

An industry target has been to reduce minimum feature size by around 30% at every process technology transition. Table 1 shows the various process technology generations or "technology nodes" used since the mid 1980's.

Table 1 Scaling ratio for various technology nodes since the mid 1980's

Technology Node, um	Scale Factor, k
1.5	
1	0.67
0.8	0.80
0.6	0.75
0.5	0.83
0.35	0.70
0.25	0.71
0.18	0.72
0.13	0.72
0.09	0.69
0.065	0.72
0.045	0.69
0.032	0.71

Such technology scaling was achieved typically in the following manner:

a. Drive new photo lithography equipment and processes that allowed printing and patterning of dimensions 30% smaller than in the previous generation.

b. Make improvements to other parts of the process, e.g., gate oxidation, ion implantation, diffusion, etching, interconnect metallurgy etc.

c. Engineer and optimize the transistor device structure and various aspects of the process to meet performance and cost goals, and be manufacturable and reliable.

d. Execute a "**Linear Shrink**" of an existing product reducing the die size by a scaling factor such as 0.7. Due to various intricacies of the process, the design rules and device characteristics at shrinking geometries, such scaling became increasingly difficult. In the mid-1980's such an approach, which was referred to by some people as a "dumb shrink" became known as an "intelligent laborious shrink" at some companies.

e. A new set of design rules - both physical and electrical - were usually used to design new products that took full advantage of the new technology capability. While the shrink approach was able to get an initial product out in the new technology node, the "**Re-Design**" approach was necessary to maximize performance and minimize cost of products in the new node.

f. In addition, the new technology usually had some new features aimed at increasing the packing efficiency, design productivity and device performance. Some examples are: increasing the number of metal interconnect layers, self-aligned polysilicon gate structure, oxide and trench isolation, standard cells, EDA tools and re-usable IP blocks.

We will now discuss migration of designs from one node to the next using either the "**Linear Shrink**" or the "**Re-Design**" approach. To illustrate the "**Linear Shrink**", consider Figure 4(a), which depicts a square die with dimension y and having N transistors, in technology node T1. A simple shrink of the die into technology node T2 would reduce the die size by the scale factor k, where 0<k<1. It should be noted that this scaling factor corresponds to the factor 1/ used by Dennard in his papers[5]. Table 2(a) is a summary of the resulting scaling parameters as well as typical values for such a scaling. Although the cost to process the wafer in the new technology node increases by a factor C (typically a 20% premium), the die cost and the CPF reduces to Ck^2 or 60% of the cost in the technology node T1, for k=0.7. This initial analysis assumes the new technology is processed using the same wafer size, and that the yield is the same in both technologies.

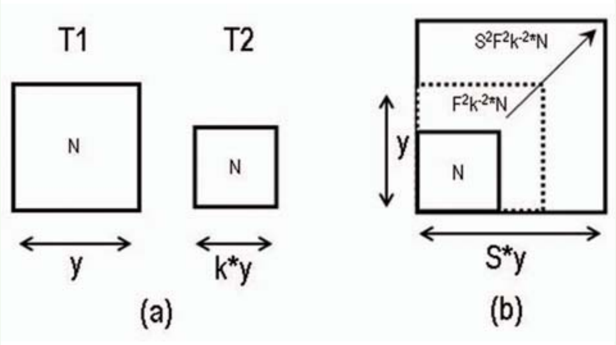

Figure 4 (a) "Linear Shrink" from technology T1 to T2 and (b) "Re-design"

(a) LINEAR SHRINK			
Constant Wafer Size			
Constant Yield			
	T1	T2	Typical
Technology Scale Factor	1	k	0.7
Die Size	1	k	0.7
Wafer Cost	1	C	1.2
GDPW	1	k^{-2}	2
Die Cost	1	Ck^2	0.6
# Functions	1	1	1
CPF	1	Ck^2	0.6

Table 2 (a) Summary of scale factors for a "Linear Shrink"

The "Re-Design" approach is illustrated via Figure 4(b) which depicts increased packing density achieved by taking advantage of more aggressive technology features and design rules and a "Cleverness Factor", F. The number of transistors packed in the same size die increases by a factor F^2k^{-2}. Further increases in packing density resulted from the use of larger die sizes. Manufacturing enhancements of the process, the equipment and the clean room environment resulted in lower defect densities. This allowed the fabrication of larger dice with acceptable yields in the new technology node in spite of the tighter geometries. The increase in the maximum allowed die size is represented by the factor S. For simplicity, we assume a square die and "die size" represents one linear edge of the die. Table 2(b) summarizes the scale factors and typical values. These typical values show a 29% annual reduction in CPF, a 4x increase in functions over a 3 year period, which is consistent with Moore's Law[1, 2] and the ITRS 2005[3].

(b) RE-DESIGN including Increased Die Size and New Technology Cleverness			
Constant Wafer Size			
Constant Yield			
Increase Die Size to Increase Packing Density			
	T1	T2	Typical
Technology Scale Factor	1	k	0.7
Die Size	1	S	1.1
Wafer Cost	1	C	1.2
GDPW	1	S^{-2}	0.9
Die Cost	1	CS^2	1.4
Cleverness Factor	1	F	1.3
# Functions	1	$S^2F^2k^{-2}$	4
CPF	1	$CF^{-2}k^2$	0.36
CPF reduction/year, 3yr cycle			29%

Table 2 (b) Summary of scale factors for "Re-Design"

Such a scaling methodology has been reported by Intel for their 80x86 microprocessors. Figure 5 shows the migration of the 8086, 80286 and the 80486 processors with increasing transistors per chip[6]. For example, in 1989 the 8086 and 80286 microprocessors fit into an area that was a fraction of the area in previous technology generations. Then the 80486 was introduced in the new node with a larger die size and 4x the number of transistors of the previous processor in the previous node.

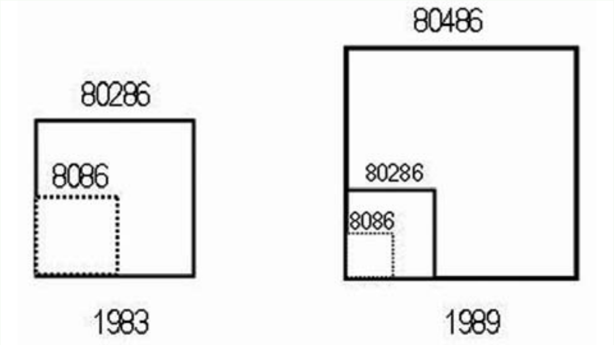

Figure 5 Technology scaling methodology reported by Intel

5. Die Cost Reduction by Increasing Wafer Size

The industry has successfully increased wafer size[7] from 50mm (2") to 300mm (12") as shown in Figure 6. The wafer diameter steps result in either a 1.33x or a 1.5x diameter ratio versus the previous size. An increased number of gross die per wafer results from the use of larger diameter wafers, as shown in Figure 7. The available silicon area is either 1.78x or 2.25x for the two different diameter ratios. The actual ratio of GDPW is generally higher and is a function of the die size, as shown in Figure 8. This is due to improved optimization of die-stepping algorithms to maximize the number of full die. Larger diameter wafers also allow a reduction of the number of partial die around the perimeter of the wafer; this effect is more dominant for larger die sizes. Manufacturing on larger diameter wafers offers an improved economy of scale.

The use of larger diameter wafers does increase wafer cost. However, we will show that there is a reduction in the die cost. Early on in the introduction of a new wafer size, a 70% increase in wafer cost is reasonable[4]. In mature production the cost to process a larger diameter wafer could increase 30%.

Relative Die Cost on larger diameter wafers = W/g,

where W is the relative wafer cost for the larger wafer and g is the relative GDPW

As mentioned earlier, the range of values for W are 1.3-1.7 and for g are 1.8-2.5. Therefore, the range of relative die cost is 0.5-0.9, a 10-50% die cost reduction when using larger diameter wafers.

A couple of examples for a mature and a relatively new technology are shown here:

a. For a 10mm die in 0.8um technology processed on 150mm and 200mm wafers, W=1.35, g=1.95. Therefore, die cost on 200mm wafers = 69% of die cost on 150mm wafers.

b. For a 10mm die in 130nm technology processed on 200mm and 300mm wafers, W=1.75, g=2.45. Therefore, die cost on 300mm wafers = 71% of die cost on 200mm wafers.

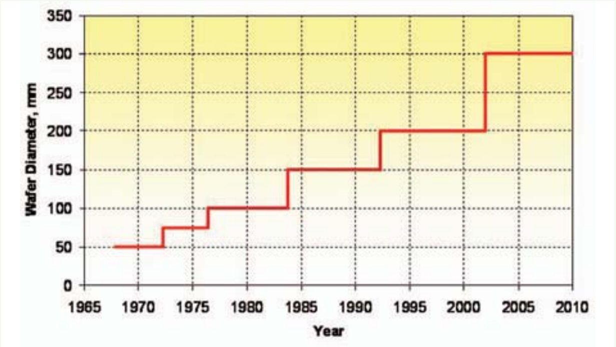

Figure 6 Silicon wafer diameter increase over time

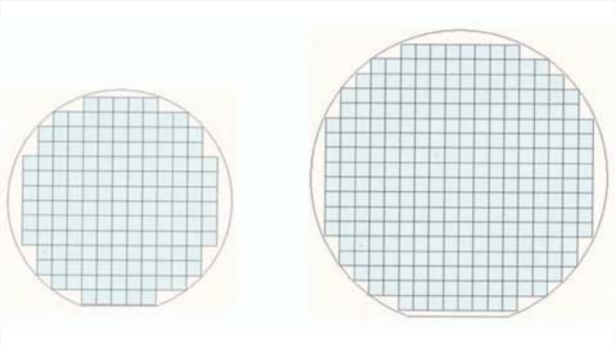

Figure 7 Increased gross die from a wafer diameter increase in the same technology

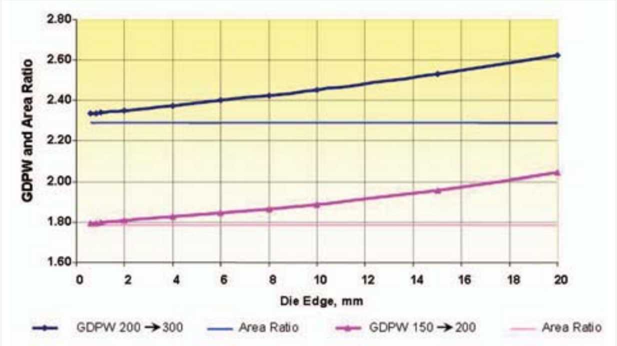

Figure 8 GDPW increase as a function of die size for two different wafer size transitions

6. Optimizing the Die Size and Packing Density per Chip

Selecting the optimum packing density and the die size becomes a challenge in this dynamic industry. We have developed models to predict the optimum die size and functions per chip. In Figures 9 and 10 we show examples of the cost/gate for 90nm and 180nm technologies as a function of die size and millions of gates per chip. The curves have a U-shape. If the die size is too small the cost is dominated by the overhead of the input/output structures, the scribe lane, etc. If the die size gets too big, the cost per gate increases due to the increased complexity. For simplicity, gate count is assumed here to be an equivalent 2-input NAND gate count. Each equivalent gate uses four transistors. The optimum gate density and cost per gate can be converted to transistor density and cost per transistor. The actual transistor count per chip increases rapidly as larger amounts of memory is included on the die. For reference, one of Intel's Pentium processors is reported with 55M transistors (14M equivalent gates) in a 90nm technology[4]. Referring to Figure 10, this data point will be considered reasonably well optimized in our analysis, since it is located near the minimum, just at the cusp of the steep slope and marked by the arrow. The shape of the curve is affected by parameters such as wafer cost, defect density, physical and electrical design rules, design tools' packing efficiency.

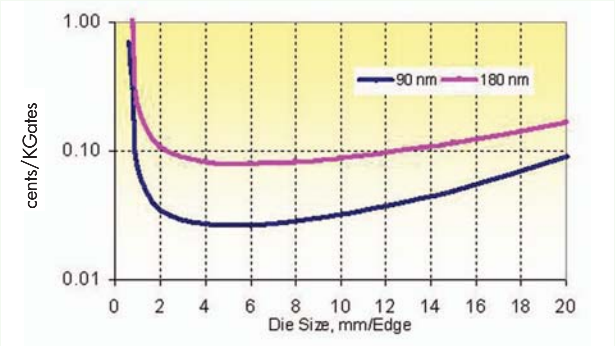

Figure 9 Cost per gate as a function of die size for 90nm and 180nm technologies

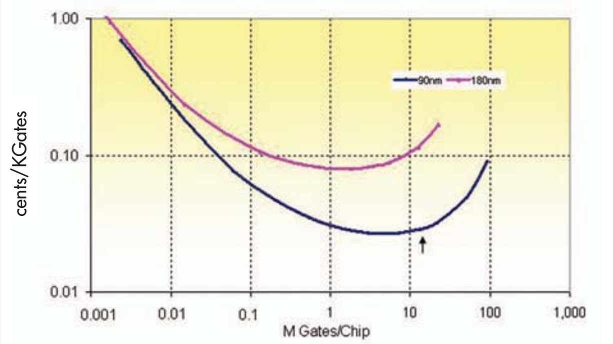

Figure 10 Cost per gate as a function of packing density for 90nm and 180nm technologies

7. Current Trends

This paper has focused on providing a historical perspective of business aspects of scaling. While a detailed discussion of the current status of technical and business challenges is beyond the scope of this paper, we will provide some highlights of current trends in this section.

a. The cost of wafer fabrication facilities and equipment, masks and chip design have all escalated significantly over the years. Finding solutions to technical challenges at the 32nm node will require ever increasing capital and manpower investments.

b. Manufacturing entities have worked diligently to accelerate the manufacturing and defectivity learning curves.

c. Creative co-design of process and design considerations has been called for by many authors[10] and are being implemented to manage challenges such as increased leakage and standby power.

d. New product introductions on the 65nm technology node have been made at leading edge users in the 2005 time frame; the cross-over point varies but is expected to be in 2007. Lead products on 45nm will likely be announced in 2007 with a cross-over in 2009. These timetables indicate a less than 3 year cycle for the introduction of new technology nodes.

e. As in the past, technical solutions for the next technology (32nm), e.g. the use of double-exposure lithography, will add significantly to capital, process development and therefore wafer cost. The author is confident that the industry will find a new manufacturing and design optimization point that will allow introduction of new products cost-effectively at this node.

f. The increasing cost of wafers, masks and design require users to very carefully assess the selection of the proper technology for their products. The trend is towards the use of leading edge technology nodes only for products with very high volumes, a compelling technical argument and a clear value proposition.

8. Summary

This paper has provided a simplified view of the business aspects of scaling and technology migrations that have been key to sustaining a phenomenal reduction in CPF for integrated circuits. Although trends such as the increasing cost of wafer fabs, masks and the increasing cost of complex designs indicate a possible slow down of the implementation of new technologies, the industry marches onward. The industry has demonstrated resilience in finding solutions to challenges. New technologies are still being introduced at a feverish pace allowing increased packing density, reduced CPF and improvements in performance.

9. Acknowledgements

The author wishes to thank Behrooz Abdi, Bill Bidermann, Mung Chen, James Clifford, Brian Henderson, Professor Dave Hodges, Merrill Hunt, Matt Nowak and Ian Young for reviewing the manuscript and offering their valuable suggestions.

10. References

1. G.E. Moore, “Progress in Digital Electronics”, 1975 IEDM, pp11-13.
2. G.E.Moore, “No Exponential is Forever; but “Forever” can be delayed”, ISSCC 2003, Paper 1.1.
3. International Technology Roadmap for Semiconductors 2005, http://public.itrs.net
4. IC Knowledge, www.icknowledge.com
5. W.Haensch, E.Nowak, R.H.Dennard, et. al., “Silicon CMOS Devices beyond scaling”, IBM J. Res. and Dev., Vol. 50, April/May 2006.
6. P. Gelsinger, P. Gargini, G. Parker, A. Yu, “2001: A Microprocessor Odyssey”, published in “Technology 2001”, MIT Press, pp. 95-113, July 1992.
7. P. Gelsinger, “Moore’s Law – The Genius Lives On”, IEEE SSCS Newsletter, September 2006.
8. R.C.Leachman, “Yield Modeling”, http://www.ieor.berkeley.edu/~ieor130/yield_models.pdf
9. M.Sydow, “Compare Logic-Array To ASIC-Chip Cost per Good Die”, Chip Design Magazine, February/March 2006.
10. T.C. Chen, “Where CMOS is Going: Trendy Hype vs. Real Tecdhnology”, ISSCC 2006, Paper 1.1.

About the Author

Rakesh Kumar is President of TCX, a consulting services company. He is also CEO of ei2, a fabless product integration company. Previously he was VP & GM of the worldwide Silicon Technology business unit at Cadence Design Systems and Tality. During his 32 years of industry experience Rakesh has also been at Unisys and Motorola where he held various technical and management positions with increasing responsibility. He has numerous publications and patents to his credit. Dr. Kumar is on the AdCom of the IEEE Solid State Circuits Society and serves as its Treasurer. He has chaired and served on the Steering committee of the IEEE Custom IC Conference for fourteen years. Rakesh received his Ph.D. and M.S. in Electrical Engineering from the University of Rochester in 1974 and 1971 respectively. He received his B. Tech. in Electrical Engineering from the Indian Institute of Technology, New Delhi in 1969. rakesh@tcxinc.com 858.748.4624