INTEGRATION LIMITS

Livio Baldi
Numonyx Italy S.r.l.

Keywords: Microelectronics, CMOS, semiconductors, Moore's Law, physical limits, economic limits.

Contents

Summary

Since its beginning in the late 1950's, microelectronics has been characterized by an impressive growth in performances, unmatched by any other technology. This exponential evolution has been formalized in the empiric Moore's Law that states that the complexity of integrated circuits doubles every 18-24 months. The end of the growth has been announced several times, on different grounds, but all forecasts have

been proved to be wrong. However limits exist, if nothing else because of physical principles, and as the size of elementary devices is nearing the one of molecules, they appear to loom not too far ahead. In this article we will shortly describe the reasons underlying the seemingly unstoppable progression of Microelectronics, based on CMOS technology, into the nanometer range, and then move to consider the limits that are appearing to the continuation of the present approach, not only on physical and technological grounds, but also considering the economic implications. Then we will briefly discuss the possible emerging alternatives and their outlook and bottlenecks.

1. Introduction

The official start of Solid-State Electronics, the first step towards Microelectronics, is officially set in 1947, when John Bardeen, Walter Brattain and William Schockley realized the first contact transistor at the AT&T Bell Labs, an achievement for which they received the Nobel Prize in 1956. But it was only ten years later in 1958 that Jack Kilby and Robert Noyce, working independently at Texas Instruments and Fairchild developed the concept of the Integrated Circuit, demonstrating the possibility of integrating several components on the same piece of semiconductor material to produce a functional circuit element. At that point Microelectronics was really born: functional devices requiring the assembly of several discrete components on a board with a size of several centimeters could be replaces by a tiny piece of silicon, a few millimeters square. From that moment on the progression of Microelectronics has been incredibly fast, outperforming all other technologies.

2. The Growth of Microelectronics

2.1. The Moore's Law

The exponential growth of Microelectronics found its first formal definition in 1965, when Gordon Moore, to become one of the co-founders of Intel, published a paper on "Electronics", in which he made the remark that, from its beginning, less than 10 years before, the transistor density of minimum cost integrated circuits had been increasing at an exponential rate, doubling every year. This observation was confirmed and better defined in a later publication in 1975, in which the rate of growth, based now on 17 years of history was revised to doubling every two years. This observation became soon popular as the "Moore's Law", or the "First Moore's Law". Even if it can be hardly considered a "law" in the physical science, this empirical relationship between density of integrated circuits and time has proven to be valid over a period of more than 50 years.

The average growth rate has been different for different types of integrated circuits (with memories growing faster than microprocessors), but the average time to doubling of device density has kept between 18 months (for memories) and two years (for microprocessors). In spite of recurring prophecies about its end, Moore's Law has been widely accepted as a basic characteristic of Microelectronic industry, and has become a basic assumption for all technology planning.

There have been several formulations of the Moore's Law: the first one referred to the number of devices per integrated circuit *at minimum cost*, but it was soon simplified, for reasons that we will see in the following chapter, to the maximum transistor count per integrated circuit.

Additional exponential laws have grown around the Moore's Law, most notably the ones referring to maximum transistor operating frequency, and to the clock rate of logic circuits.

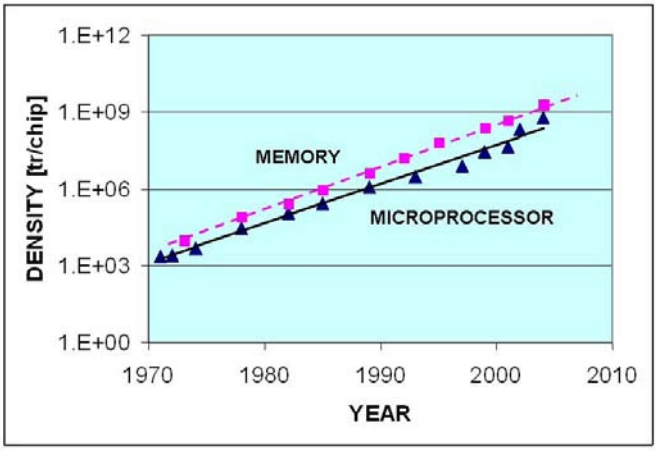

Figure 1. evolution of Moore's Law

2.2. What is behind Moore's Law

In his presentation of 1975 Moore identified three major contributions to this growth:

- increase in chip area, made possible by the progress in technology and material quality;
- the use of finer scale structures, made possible by more advanced lithographic technology, and
- device and circuit cleverness, which covered not only the circuit design but also the architecture of the basic switching device that, in that time, was moving from bipolar to MOS.

For the first factor, Moore indicated a limit around 0.3 sq. inches (around 2 cm^2) which corresponded to doubling the maximum size available at that time. This prevision has been verified, and current maximum size of integrated circuit does not usually exceed this limit, for technical reasons, related to lithography, and for economical reasons, related to defect density.

There is no doubt that the second factor has been the one giving the largest contribution. Already Moore considered possible to reduce typical device size to the micron or submicron region. What happened was that typical minimum geometry size has been reduced from about 5 micron in 1975 to less than 50 nanometers in 2010, with a reduction in the area of the basic device of four orders of magnitude. If we consider that the maximum density shown by Moore in 1975 was around 50K transistors, and the top-

of-the-line Intel microprocessor in 2005 is around 500Mtransistors, we can recognize that scaling of device structures has been the main driving factor.

The last contribution, according to Moore was already almost exhausted. It still plays a role, as shown from the difference in integration density between memories and logic circuits (around one order of magnitude), but has not increased significantly. On the contrary, often design approaches privilege design speed over circuit density.

If we try to understand what is behind this exponential growth, which surprisingly spans more than 50 years, we have to distinguish between *physical* feasibility, *technological* viability and *economical* motivations that justify the effort.

All three factors have been important for the evolution of Microelectronics till now, and must be considered in defining its limits.

2.3. The Physical Feasibility

The evolution of Microelectronics has been based on the use of the MOS transistor as basic switching element. The main advantages of MOS over bipolar transistors are:

- the low intrinsic complexity of the transistor architecture;
- the possibility to realize logic circuits based on complementary MOS architectures (CMOS) that do not dissipate power in stand-by, but only during operation (a sort of pay-for-use approach);
- the high input impedance that simplify design and makes large fan-outs possible;
- the *intrinsic scalability* of the basic transistor, which allows to reduce the size of the device, when the technology is available, without impacting the physical mechanism.

The last factor has been surely the most important, since it has allowed scaling the size of critical devices along an evolutionary path, with a sequence of progressive improvements to basic technology. In simple terms scalability means that two transistors of different size behave in the same way if their physical and geometrical parameters are scaled by the same factor. Besides insuring the continuity of the physical mechanism, scaling laws, first described by Dennard in 1974, allow to derive in a simple way the main physical parameters that the scaled transistor must have to work properly.

2.4. The Technical Factors

Optical lithography has been the main technological enabler of transistor scaling, allowing defining the geometrical elements of the transistors and of the integrated circuits with always increasing resolution. The basic principle, that is in its essence the use of “an optical microscope running backwards”, was presented by Feynman in 1959 in its famous talk at the annual meeting of the American Physical Society: an image, 4-5 times the size of the final structure to be realized, is generated on a glass plate (mask) by e-beam writing, and afterwards reproduced on silicon wafers through a reduction optical system. This approach combines the high resolution of the electron beam systems, with the low cost and high throughput of optical printing. Optical lithography has started to

show its limitation when the features to be imaged have approached the wavelength of the light used in the system. Shorter wavelengths, in the UV range have been introduced, down to 193nm, and by using water immersion to decrease the effective wavelength, coupled with image correction techniques it has been possible to extend the limits of optical lithography to the 45nm range. Lack of transparent materials for lenses at lower wavelengths has stopped further progress in this direction.

Several technological breakthroughs have been needed to overcome the practical problems related to the down-scaling of transistor size, but they have not implied any fundamental change to the basic device architecture, which has allowed a smooth transition from one generation to the other. It is not the purpose of this article to analyze in detail the CMOS technology, and we will quote only the major innovations that have allowed it to move into the deep submicron region:

- Massive use of ion implantation for precisely controlled silicon doping
- Introduction of CVD and, more recently, ALCVD as layer deposition technology;
- Use of refractory metal silicides as low resistivity interconnections;
- Introduction of CMP for dielectric planarization, which has made possible to increase drastically the number of metal interconnections;
- Introduction of Copper for low-resistivity metal interconnection layers.

In parallel, the quality of the materials, of the equipment and of the manufacturing plants has gone through a continuous improvement process to reduce defect size and defect density to allow achieving high production yields for devices including hundreds of millions, if not billions, of deep submicron components.

All this progress in manufacturing technology has implied huge investments in R&D and equipment development that have been compensated by the increase in productivity.

2.5. The Economic Drive

The real driving factor for the Moore's Law, made possible by the scalability of the basic device architecture and by the timely development of the critical technology steps, has been the economical push. If we look at the growth of the market of Microelectronics it shows a constant average growth of around 10-12%, in spite of wide periodical oscillations.

These oscillations are less evident if we look at the number of sold units, rather than at the economical value, because they were largely caused by price variations. The exponential growth of any product is normally related to the presence of an unsaturated market: if all production is sold and the profit is reinvested to make more products, the consequence will be an exponential growth until the available market is saturated. What is surprising is that the Microelectronics market has kept the characteristic of an unsaturated market over such a long period of time.

The reason goes back to the Moore's Law:

- production cost of integrated circuits is largely proportional to the area and independent from technology generation, because the increase in complexity, going

to smaller geometry size, has been traditionally compensated by improvements in manufacturing efficiency;

- therefore by making devices smaller more product are obtained at the same costs, and price can be lowered;
- for the same component, reducing the area, the price is reduced and the potential market is increased; and
- for the same area, that is the same price, more complex functions can be offered which opens the way to new applications and creates new markets.

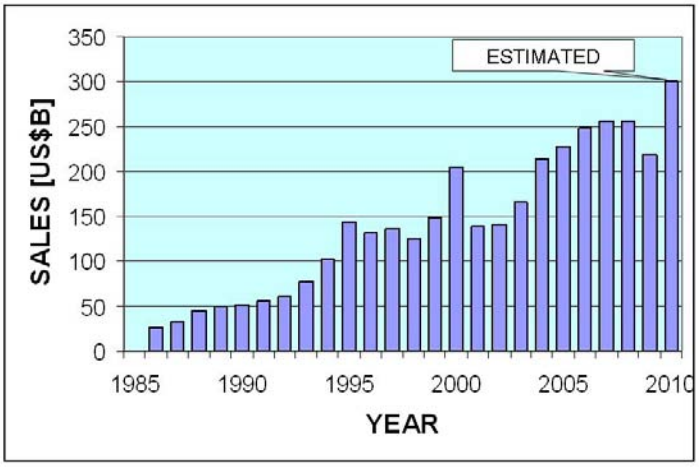

Figure 2. Market of Microelectronics (source: WSTS, IC Insights)

Of course there are additional costs that do not scale as the area, like the ones for testing and packaging, and effective area scaling does not really go as the square of the geometrical linewidth scaling, because of peripheral contact structures for chip wiring to package that do not scale, but in general chip costs scale almost as the square of geometrical scaling factor.

An additional push towards the continuation of Moore's Law has been given by the additional benefits that can be obtained by scaling the geometrical and physical parameters of transistors, according to the principles proposed by Dennard. As shown in the table below, by applying the same scaling factor k to geometrical, physical parameters and to supply voltage, the typical delay time is also reduced by the same factor, and the power*delay figure is reduced as the third power of the scaling factor, while total power dissipation is reduced and power density is kept constant.

The advantages in terms of reduction of delay time can be even larger (up to the square of the scaling factor) if the supply voltage is scaled less than the geometrical parameters, which is what happened for high performance CMOS technology where going from the 6 μm technology of 1974 to the 65 nm technology of 2006 (a factor of almost 100 of geometrical scaling), the supply voltage has been scaled from 5 V to 1.2 V (for a scaling factor of less than 5).

Device/Circuit Parameter	Scaling Factor
Device dimension *tox, L, W*	$1/k$
Doping concentration *Na*	k
Voltage *V*	$1/k$
Current *I*	$1/k$
Capacitance $\varepsilon A/tox$	$1/k$
Delay time/circuit *VC/I*	$1/k$
Power dissipation/circuit *VI*	$1/k^2$
Power density *VI/A*	1

Table 1. Constant voltage scaling rules and impact on device performances.

As a result the operating speed of all logic devices has increased exponentially, following a trend that parallels closely the Moore's Law for device density, as shown by the evolution of the clock frequency of microprocessors, given in Figure 3.

The continuous increase in performances joins the economical push from the reduction in specific production costs towards the extension of Moore's Law, in order to insure a continuous growth of the market. A demonstration of the success of this strategy is shown not only by the persistence of the Moore's Law over the past 50 years, but by the presence of applications based on Microelectronics in all aspects of everyday life: from cellular phones to personal computers, from Smart Cards to Digital Television, from Internet based services to sophisticated safety features in cars.

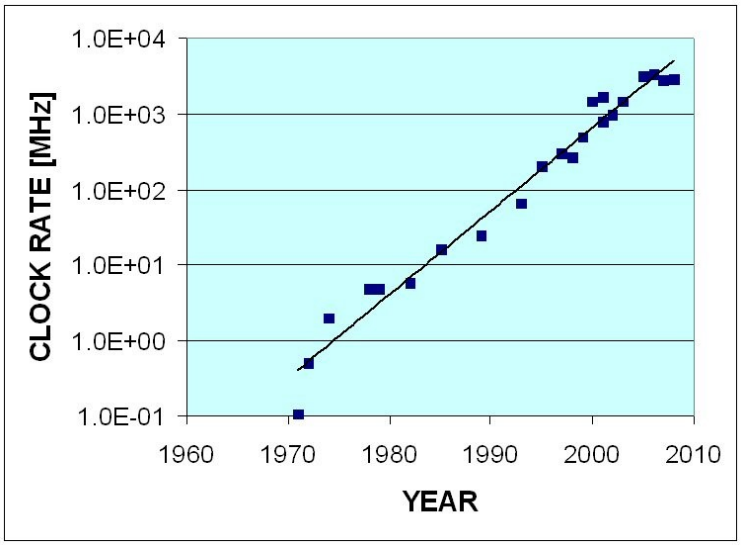

Figure 3. clock frequency evolution in Intel processors (source: Intel)

The strong economical impact of Moore's Law has pushed all major actors in the development of Microelectronic, both industry and research centers, to cooperate to plan the future evolution of the technology, in order to make sure that all conditions are met to continue with this virtuous circle as long as possible. To this purpose all major players in the field worldwide have associated in the ITRS (International Technology

Roadmap for Semiconductors) organization, which prepares and releases on a two year basis a forecast of the evolution of Microelectronics over the next 15 years, evidencing potential roadblocks and indicating priority for research in this sector. For each major application, critical value of all technology parameters are defined, essentially through an extrapolation of scaling laws, and an indication of the possible ways to achieve them is given, with an assessment of the difficulty of the task.

3. Which are the Limits?

Gordon Moore himself is quoted saying "'The important thing is that Moore's Law is exponential, and no exponential is forever… But we can delay forever". Up to now this statement has hold true: the end of Moore's Law on the ground of technology limitations has been forecasted several times, but technological solutions have always been found around them. But there are firm physical limits to size reduction and therefore two seemingly contradictory questions arise:

- Where is the limit? and
- What comes after?

We have identified three main factors driving Moore's Law. All of them must be considered.

-

-

Bibliography

Moore G.. (1965). Cramming more components onto integrated circuits. *Electronics*, Vol.38, 8, April 1965 [First presentation of the exponential evolution of the density of integrated circuits by Gordon Moore]

Moore G.(1975). *Progress in digital integrated electronics*. Proceedings of IEDM, 1975 [Second paper by Moore in which the Moore's Law was confirmed after 10 years and defined in more details]

Dennard R.H. (1974) Design of Ion Implanted MOSFET's with very small dimensions, *IEEE-JSSC*,.SC-9, 256 [First presentation of the principle of MOS scaling]

Feynman R.P. (1960). There is plenty of space at the bottom, *Engineering and Science*, Febr. 1960 [Text of the original presentation by Feynman at the meeting of the American Physic Society which is considered at the basis of the nanotechnology].

ITRS Report 2009, http://www.itrs.net

Doris B. et al. (2002). *Extreme Scaling with Ultra-Thin Si Channel MOSFETs*, Proceedings of IEDM 2002 [First presentation by IBM of the realization of a working 6 nm MOS device, the smallest until now]

Moore G. (2003). *No Exponential Is Forever: But "Forever" Can Be Delayed!*, Digest of Technical Papers of 2003 International Solid-State Circuits Conference [An update of the limits of scaling from the author of Moore's Law]

Frank D.J. (2002). Power constrained CMOS scaling limits, *IBM Journ. of Research and Development*, Vol.46, March 2002 [An analysis of the main mechanisms of power consumption in CMOS, and the effects of scaling]

Asenov, A., Cheng B.,Dideban D., Kovac U., Moezi N., Millar C., Roy G.., Brown A.R., Roy S. (2010) *Modeling and simulation of transistor and circuit variability and reliability*, Proceedings of Custom Integrated Circuits Conference, Sept. 2010 [A recent paper analyzing all the main causes of the statistical spread of transistor parameters]

Digh Hisamoto, Member, IEEE, Wen-Chin Lee, Jakub Kedzierski, Hideki Takeuchi, Kazuya Asano, Kuo C., Anderson E., King T.-J., Bokor J., Hu C. (2000). FinFET—A Self-Aligned Double-Gate MOSFET Scalable to 20 nm, *IEEE Trans. on Electr. Dev*., Vol. 47, No. 12, Dec. 2000 [Presentation of the Fin-FET architecture]

K. Rim K., Koester S., Hargrove M., Chu J., Mooney P. M., Ott J., Kanarsky T., Ronsheim P., Ieong M., Grill A., Wong H.-S. P. (2001). *Strained Si NMOSFETs for High Performance CMOS Technology,* Proceedings of Symposium on VLSI Technology, Kyoto, 2001 [First presentation by IBM of the use of strained Silicon to increase electron mobility in MOS devices]

Awano Y. (2009). *Graphene for VLSI: FET and Interconnect Applications*, Proceedings of IEDM, Baltimore 2009 [A review of the possible application of high mobility graphene layers for integrated circuits]

T. Ebihara et al. (2003*). Beyond k1=0.25 lithography: 70nm L/S patterning using KrF scanners*, Proceedings of SPIE, Vol. 5256, 23rd Annual BACUS Symp. on Photomask Technology, 2 [A description of the potential of double patterning technology for extending limits of optical lithography]

Hsiao Y.-H., Lue H.-T., Hsu T.-H., Hsieh K.-Y., Lu C.-Y. (2010). *A Critical Examination of 3D Stackable NAND Flash Memory Architectures by Simulation Study of the Scaling Capability*, Proceedings of IEEE International Memory Workshop, Seoul 2010. [A critical review of tridimensional architectures for Flash non volatile memory]

Bandyopadhyay S., Cahay M. (2008). *Introduction to Spintronics*, CRC Press [An updated introduction to spintronic technology]

Someya T., Sekitani T., Takamiya M., Sakurai T., Zschieschang U., Klauk H. (2009). *Printed organic transistors: Toward ambient electronics*, Proceedings of IEDM, Dec. 2009 [A recent presentation of status of organic electronics, with special reference to low complexity applications]

Hirvensalo M. (2003). *Quantum Computing*, Springer Verlag [A comprehensive presentation of the principles of quantum computing]

A.Chiabrera, E.Di Zitti, F.Costa and G.M.Bisio, *Physical limits of integration and information processing in molecular systems*, J. Phys. D: appl. Phys. 22, 1571-1579 (1989) [The paper considers several physical limits to the scaling of computing systems, with special consideration to interconnection problems]

V.V.Zhirnov, R.K.Cavin, J.A.Hutchby, G..I.Bourianoff, Limits to Binary Logic Switch Scaling – A Gedanken Model, *Proc. of IEEE*, Vol.91, No.11, Nov. 2003. [A theoretical analysis of the quantum and thermodynamic limits of computation]

G.F.Cerofolini, Realistic limits to computation – I. Physical limits, *Appl. Phys*. A, 86, 23-29 (2007) [An accurate analysis of quantum and thermodynamic limits, with special focus on memories]

P. Benioff, Quantum Mechanical Models of Turing Machines That Dissipate No Energy, *Phys. Rev. Lett*. 48, 1581–1585 (1982) [An interesting and provocative paper advancing the possibility of non dissipative reversible computation]

Biographical Sketch

Ing. Livio Baldi graduated in Electronic Engineering in 1973 at the University of Pavia. In 1974 he joined SGS-ATES (now STMicroelectronics), in the Central R&D of Agrate Brianza. He has been

responsible for the development of CMOS processes for EEPROM memories and multifunction logic. In 1999 he moved to lead the NVM Design Platform Development Group, and afterwards he was put in charge of coordinating the participation of ST Italy in cooperation research projects, representing it in the MEDEA+ Steering Group-Technology and in the Support Group of the European Technology Platform (ENIAC). He has acted as consultant for the Commission in the definition of Framework Programmes and he is member of the Expert Advisory Group for Theme 4 (Nanoscience, Materials and Production Technology) of FP7.

From March 31st, 2008 he has moved to Numonyx, the new ST-Intel joint venture on Flash memories, now Micron Semiconductors Italia, in charge of External Relations and Funding in Central R&D, and representing it in AENEAS and CATRENE.

He holds 33 US patents, 17 European patents, and is author of more than 65 papers and communications to conferences.