The Economic Limit to Moore's Law

Karl Rupp, *Student Member, IEEE,* and Siegfried Selberherr, *Fellow, IEEE*

***Abstract*—There have been numerous papers and discussions about the lives and deaths of Moore's Law, all of them dealing with several technological questions. In this paper, we consider economic limitations to the exponential growth of the number of components per chip. As the presented growth model shows, economics constitute indeed a potential slow-down mechanism.**

***Index Terms*—Components per chip, fabrication costs, growth modeling, Moore's Law.**

I. Introduction

Back in 1965, Moore observed in his famous paper [1] that "The complexity for minimum component costs has increased at a rate of roughly a factor of two per year (...)." In 1975, Moore refined his component count estimation to a doubling every two years, and thus a reduced exponential growth compared to his initial estimation. Indeed, looking at the history of integrated circuits from 1975 to 2008, a doubling of transistor counts every two years was a good estimation. This prediction known as *Moore's Law* has become a business-dictum for the whole semiconductor industry.

The higher component density has led to a decrease in end-consumer prices. However, the costs for producers follow a converse trend: research and development, manufacture, and tests become more and more expensive with each new generation. This observation is known as *Rock's Law* and sometimes also referred to as *Moore's Second Law* [2]; fabrication facility (fab) costs also follow an exponential growth. Despite this exponential growth of facility costs, the cost per shipped unit decreases at an exponential rate.

Moore himself already observed in 1995 that the semiconductor industry cannot continue its fast exponential growth indefinitely, since it would exceed the gross world product (GWP) at some time. In contrast to all previous publications dealing with technological limitations to Moore's law, e.g., [3], we investigate economic limitations to the semiconductor business. A summary of our results has already been published in [4], but no details on the derivations have been given. Thus, the model derivation is presented here and each step is discussed.

Manuscript received March 20, 2010; revised August 26, 2010; accepted October 15, 2010. Date of publication October 28, 2010; date of current version February 4, 2011.

The authors are with the Institute for Microelectronics, Technische Universität Wien, A-1040 Wien, Austria (e-mail: rupp@iue.tuwien.ac.at; selberherr@iue.tuwien.ac.at).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSM.2010.2089811

II. Growth Limited by Economy

Moore's Law states that the number of components—taken to be transistors in the following—per chip doubles every year, which is mathematically expressed as follows:

$$\frac{\mathrm{d}n_{\mathrm{comp}}(t)}{\mathrm{d}t} = \alpha n_{\mathrm{comp}}(t) \tag{1}$$

where $n_{\mathrm{comp}}(t)$ is the number of transistors per chip, t is the time measured in years from the beginning of 2010 (to ease the formulation of initial conditions), and $\alpha \approx \ln(2)/2 = 0.34$ is the growth parameter in accordance with Moore's Law. This differential equation can also be read as "The change of the number of transistors per chip that is proportional to the number of transistors available at a given time," or "The current generation of technology is the basis of the next one." The solution of (1) can be written as follows:

$$n_{\mathrm{comp}}(t) = n_{\mathrm{comp},2010} \exp(\alpha t) \tag{2}$$

where $n_{\mathrm{comp},2010} \approx 10^9$ is the number of transistors per chip in the beginning of 2010 (which corresponds to $t = 0$).

A more general formulation of (1) is as follows:

$$\frac{\mathrm{d}n_{\mathrm{comp}}(t)}{\mathrm{d}t} = r(t) \tag{3}$$

where the progress variable $r(t)$ denotes a collection of available research output, computational power, and fab capacity at time t. Clearly, in case $r(t) = \alpha n_{\mathrm{comp}}(t)$, we fall back to (1), which means that no other factors apart from the current available technology limit research.

According to Moore's Second Law, the exponential increase in fab costs over time, the costs $c_{\mathrm{fab}}(t)$ fulfill a differential equation similar to (1) as follows:

$$\frac{\mathrm{d}c_{\mathrm{fab}}(t)}{\mathrm{d}t} = \beta c_{\mathrm{fab}}(t).$$

The growth parameter β is determined from fab costs in the past decades (Fig. 1) and indicates a growth of a factor 100 within 30 years, thus $\beta \approx \ln(100)/30 \approx 0.13$, which corresponds to an average growth of fab costs of 14% per annum [5]. A closed expression for $c(t)$ is as follows:

$$c_{\mathrm{fab}}(t) = c_{\mathrm{fab},2010} \exp(\beta t) \tag{4}$$

where c_{2010} is approximately five billion dollars.

Even though fab costs have increased in the past exponentially, this cannot continue indefinitely. A natural barrier is in principle given by the total revenue of the semiconductor industry. Since this revenue has experienced a tremendous growth in the past century itself, we better formulate a bound

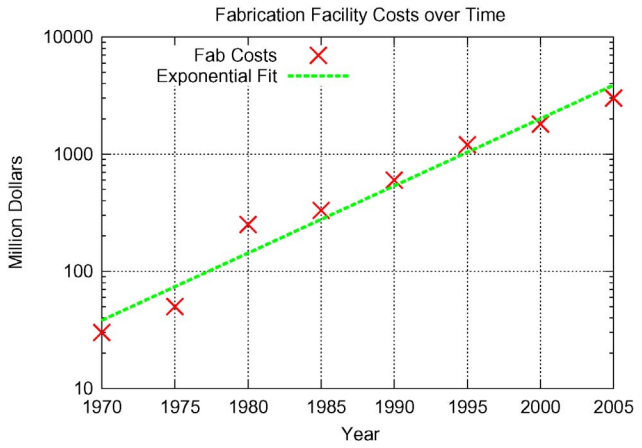

Fig. 1. Fabrication facility costs over time.

for fab costs in terms of a fixed share ε of the GWP $g(t)$ as follows:

$$c_{\text{fab}}(t) = \min\left\{\varepsilon g(t), c_{\text{fab},2010}\exp(\beta t)\right\}$$

where $\varepsilon \approx 0.02\%$ appears to be plausible considering that the semiconductor business revenue has been about half a percent of the GWP over the past ten years and several fabs are built each year. How reasonable is the choice of $\varepsilon = 0.02\%$? Moore himself wrote [6]:

> *I do not know how much of the GWP we can be, but much over one percent would certainly surprise me.*

If we assume that the demand-driven semiconductor industry revenue becomes at most 1% of the GWP, $\varepsilon = 0.02\%$ for one fab means 2% of the semiconductor revenues. If we further suppose that five such fabs are built each year, we already arrive at 10% of the semiconductor industry's revenue used for fabs only. At the present GWP share of 0.5%, $\varepsilon = 0.02\%$ corresponds to 4% of the total semiconductor industry's revenue per fab.

The GWP growth is much more stable than the growth of a single economic branch. It grows also exponentially, but slower than the semiconductor business in the past

$$g(t) = g_{2010}\exp(pt).$$

We assume an average GWP growth of $p = 3\%$ per year, which is estimated by looking at growth rates from the past decades, compare Table I. With this we can write the fab cost estimation as follows:

$$c_{\text{fab}}(t) = \min\left\{\varepsilon g_{2010}\exp(pt) c_{\text{fab},2010}\exp(\beta t)\right\}. \quad (5)$$

Essentially, fab costs c_{fab} are given by the averaged price per chip c_{chip} times the number of chips n_{chips} leaving the factory

$$c_{\text{fab}}(t) = c_{\text{chip}}(t) \times n_{\text{chips}}(t). \quad (6)$$

On the contrary, *economy of scale*, as already addressed in Moore's paper [1], states that the chip price at time t decreases as the number of produced chips $n_{\text{chips}}(t)$ increases. Furthermore, the better the production process, quantified by the progress variable $r(t)$, is developed, the lower the costs per chip are. On the contrary, the price per chip saturates due to

TABLE I
COMPARISON OF GWP (MARKET EXCHANGE RATES), SEMICONDUCTOR INDUSTRY REVENUE (S-REV.), AND GWP GROWTH (PURCHASING POWER PARITY EXCHANGE RATES) FROM 1998 TO 2008 [7]

Year	GWP (B$)	S-Rev. (B$)	S-Rev./GWP (%)	GWP Growth (%)
1998	29.998	135	0.45	2.30
1999	31.078	169	0.54	3.19
2000	32.037	220	0.69	4.11
2001	31.810	150	0.47	1.51
2002	33.070	156	0.47	1.86
2003	37.207	183	0.49	2.65
2004	41.917	227	0.54	4.18
2005	45.292	237	0.52	3.49
2006	49.021	260	0.53	3.93
2007	55.117	269	0.48	3.77
2008	60.557	258	0.43	3.20

inevitable wafer costs as the number of produced chips grows. Writing these dependencies explicitly, (6) becomes

$$c_{\text{fab}}(t) = c_{\text{chip}}(r(t), n_{\text{chips}}(t)) \times n_{\text{chips}}(t). \quad (7)$$

With increasing progress $r(t)$, process tools have proven to be more cost-effective as the number of chips $n_{\text{chips}}(t)$ per tool rises. This links the number of chips leaving a fab with the available progress by

$$n_{\text{chips}}(t) = C_{\text{n,r}} r(t)^{1/\gamma} \quad (8)$$

where $C_{\text{n,r}}$ is a constant of proportionality and the parameter γ takes different exponential growth rates of $n_{\text{chips}}(t)$ and $r(t)$ into account. As can be seen later, the qualitative results of our model do not depend on $C_{\text{n,r}}$.

For a given number of chips to be produced, higher progress $r(t)$, such as improved yield rates or smaller feature sizes, reduces the costs per chip. However, once chip costs fall below a certain threshold, one of the following scenarios typically occurs.

1) Chip functionality is increased to remain competitive, thus higher costs for additional functionality compensate savings.
2) Chip functionality is reused as an Intellectual Property block in a future, more complex chip at costs similar to the original chip.
3) Chip production becomes unattractive and is shut down.

Consequently, chip costs c_{chip} can be safely assumed to be essentially constant, i.e., there is a negligible relative change over time compared to that of fab costs.

Substitution of (8) into (7) leads to

$$\begin{aligned} c_{\text{fab}}(t) &= \tilde{C}_{\text{n,r}} r(t)^{1/\gamma} \\ \Leftrightarrow \quad r(t) &= c_{\text{fab}}^{\gamma}(t)/\tilde{C}_{\text{r}}^{\gamma} \end{aligned} \quad (9)$$

with $\tilde{C}_{\text{n,r}} = c_{\text{chip}} C_{\text{n,r}}$. Since $\tilde{C}_{\text{n,r}}$ shows at most negligible exponential growth rates, γ is found to be α/β to account for the different growth rates of $r(t)$ and $c(t)$. Note that in case $r(t) = \alpha n(t)$, as it has been in the past, (2) and (4) exactly fulfill (9).

With the economic barrier on fab costs in (5), we find

$$\begin{aligned} r(t) &= \min\left\{\varepsilon g_{2010} e^{pt}, c_{2010}\mathrm{e}^{\beta t}\right\}^{\alpha/\beta}/\tilde{C}_{\text{n,r}} \\ &= \min\left\{C_{\text{GWP}} e^{\mu t}, C_{\text{Moore}}\mathrm{e}^{\alpha t}\right\} \end{aligned}$$

TABLE II
CONSUMER PRICE INDEX WEIGHTS FOR URBAN CUSTOMERS (CPI-U) AND URBAN WAGE EARNERS AND CLERICAL WORKERS (CPI-W) FOR CATEGORIES HIGHLY RELEVANT FOR SEMICONDUCTOR INDUSTRY IN THE U.S. [8]

Category	CPI-U	CPI-W
Personal computers and peripheral equipment	0.214%	0.202%
Televisions	0.135%	0.117%
Audio equipment	0.104%	0.097%
Photographic equipment and supplies	0.072%	0.062s%

where $C_{\mathrm{GWP}} = (\varepsilon g_{2010})^{\alpha/\beta}/\tilde{C}_{\mathrm{n,r}}$, $C_{\mathrm{Moore}} = c_{2010}^{\alpha/\beta}/\tilde{C}_{\mathrm{n,r}}$, and $\mu = p\alpha/\beta$. Substituting this expression into (3), we obtain

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} = \min\left\{C_{\mathrm{GWP}} e^{\mu t}, C_{\mathrm{Moore}} \mathrm{e}^{\alpha t}\right\}. \tag{10}$$

A direct comparison with (1) and (2) shows the economic limit; while Moore's Law predicts a growth constant $\alpha \approx 0.34$, economic limits would result in a reduced asymptotic growth parameter of $\mu = p\alpha/\beta \approx 0.08$, which is by a factor of four smaller than α. Hence, a reduced asymptotic growth of transistor counts by a factor of two every eight years follows. The GWP share parameter ε does not influence the reduced growth rate due to economy, it only influences the time at which we run into economic limitations.

The GWP share parameter ε is bounded by the demand. In Table II, the consumer price index weights for the main electronic equipment business branches in the U.S. are listed, making up a total of about 0.5%, thus representing the semiconductor industry's revenue quite accurately. A considerable change in these weights cannot be expected due to the maturity of these market segments. Under the assumption that similar figures hold true for other countries, the global demand for semiconductor devices does not allow GWP shares of the whole semiconductor industry larger than about 1%, even if we allow for a generous doubling in the GWP share over the present value. Recalling market shares from Table I, we conclude that it is the saturated demand that prohibits higher market shares.

The reduced growth model (10) can be interpreted in the following way; as long as fab costs increase with the same rate as they did in the past, the number of transistors per chip also increases at the same rate as in the past. However, as soon as fab costs hit an economic barrier given by $\varepsilon g(t)$, fab costs can only increase at the same rate as the GWP does. Consequently, transistor counts will also grow at a reduced rate.

The prediction of the time at which we run into economic limitations is very sensitive with respect to the fab costs parameter ε; choosing $\varepsilon = 0.03\%$, a growth reduction is predicted around 2025, whereas the choice $\varepsilon = 0.01\%$ shows first signs of reduced growth already in 2015. Thus, with joint funding of large fabs, an economic growth capping can be shifted many years into the future so that we might face limitations imposed by physics first.

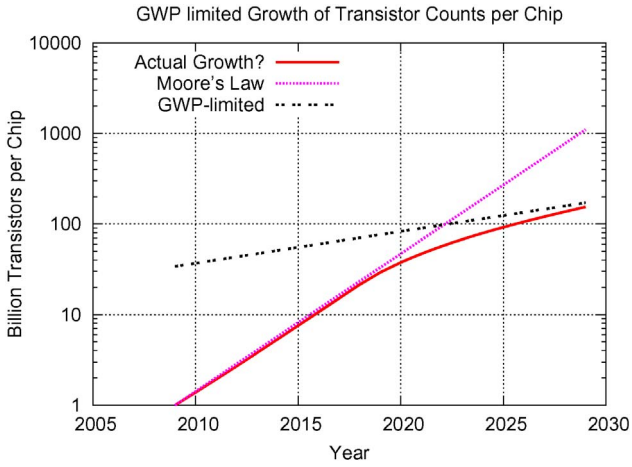

Fig. 2. If costs for a single fab are at most 0.02% of the GWP (i.e., $\varepsilon = 0.0002$), a reduced growth of transistor counts per chip for economic reasons is likely to happen around 2020.

III. FROM FREE MARKET TO OLIGOPOLY

Considering the market structure of the semiconductor business, integrated device manufacturers (IDMs), who both design and manufacture, will only be able to afford fab costs up to $\varepsilon_{\mathrm{IDM}} \times$ GWP, simply because they have to cover both fab expenses and investments into research and development. Actually, $\varepsilon_{\mathrm{IDM}}$ has to be determined for each IDM separately. In contrast, foundry companies can fully focus on device manufacture, potentially being able to afford $\varepsilon_{\mathrm{foundry}} \times$ GWP with $\varepsilon_{\mathrm{foundry}} > \varepsilon_{\mathrm{IDM}}$. As soon as fab costs exceed $\varepsilon_{\mathrm{IDM}} \times$GWP, the business is thus subjected to a restructuring process; IDMs are unable to afford a fab for the latest generation of devices any longer. The only way to stay competitive for IDMs in mass markets is to become a fabless company then, fully focus on chip design, and rely on foundry companies for fabrication. Such a transition, which can already be observed today, is to a certain degree self-energizing; with each IDM becoming a fabless company, another foundry company grows and pushes its individual GWP share parameter $\varepsilon_{\mathrm{foundry}}$ to a larger value.

However, the number of semiconductor foundries can only be small, if the fab costs keep increasing at a rate higher than the GWP; the large, cost-effective fabs provide best and most transistors per dollar. Higher unit costs of a process tool in combination with a process engineer at simultaneously increased output makes it harder for smaller companies to run production effectively. They cannot reach the cost efficiency of larger companies due to lower capacity utilization and are consequently pushed out of the market. Hence, the free market of (leading edge) chip manufacturing may turn into an oligopoly. The consequences of such an oligopoly can be manifold; research and development could be focused and accelerated, while syndicates may lead to a considerable slow-down of innovations, because the remaining companies will not be interested in a technology race. Since a successful market entry for new companies is almost impossible, existing fabs could be sufficient to skim the market within such an oligopoly.

In the long run another limitation stems from (8), because the increase in the number of chips produced per fab is

potentially higher than the increase in demand. Thus, assuming that fundamental limits of physics are not reached beforehand, a point where a single fab is sufficient to fulfill world demand might eventually be reached, if we continue to follow Moore's Law. After that point, fab costs per chip would start to rise due to the impossibility of reaching economic scale.

IV. Conclusion

There have been numerous papers and discussions about the lives and deaths of Moore's Law, all of them dealing with several technological questions. As the presented growth model showed, economics must not be left unconsidered and constitute a slow-down mechanism. With a reduced growth of transistor counts, it will take some additional years until we finally hit a fundamental barrier imposed by physics.

Acknowledgment

The authors wish to thank the anonymous reviewers for their helpful comments. In particular, Reviewer 4 provided very valuable suggestions.

References

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.
[2] P. E. Ross, "5 commandments," *IEEE Spectrum*, vol. 40, no. 12, pp. 30–35, Dec. 2003.
[3] J. R. Powell, "The quantum limit to Moore's law," *Proc. IEEE*, vol. 96, no. 8, pp. 1247–1248, Aug. 2008.
[4] K. Rupp and S. Selberherr, "The economic limit to Moore's law," *Proc. IEEE*, vol. 98, no. 3, pp. 351–353, Mar. 2010.
[5] Y. Nishi and R. Doering, Eds., *Handbook of Semiconductor Manufacturing Technology*, 2nd ed. Boca Raton, FL: CRC Press, Jul. 2007.
[6] G. E. Moore, "Lithography and the future of Moore's law," *Proc. SPIE*, vol. 2437, no. 8, pp. 2–17, 1995.
[7] World Bank. *World Development Indicators Database* [Online]. Available: http://devdata.worldbank.org/data-query
[8] United States Department of Labor, Bureau of Labor Statistics. (2010). *Consumer Price Index* [Online]. Available: http://www.bls.gov/CPI

Karl Rupp (S'10) was born in Austria in 1984. He received the B.E. degree in electrical engineering from the Technische Universität Wien, Wien, Austria, in 2006, the M.S. degree in computational mathematics from Brunel University, London, U.K., in 2007, and the Diplomingenieur degree in microelectronics and technical mathematics from the Technische Universität Wien in 2009. Since 2009, he has been pursuing the Doctoral degree from the Institute for Microelectronics, Technische Universität Wien.

His current research interests include generic and generative programming of discretization schemes such as the finite element method for the use in multiphysics problems as well as deterministic numerical solution approaches to the Boltzmann equation.

Siegfried Selberherr (M'79–SM'84–F'93) was born in Austria in 1955. He received the Diplomingenieur degree in electrical engineering and the Doctoral degree in technical sciences, both from the Technische Universität Wien, Wien, Austria, in 1978 and 1981, respectively.

He has been holding the "venia docendi" on computer-aided design since 1984. Since 1988, he has been the Chair Professor with the Institute for Microelectronics, Technische Universität Wien. From 1998 to 2005, he served as the Dean of the Fakultät für Elektrotechnik und Informationstechnik. His current research interests include modeling and simulation of problems for microelectronics engineering.