Approaching nanoscale integration

Dieter Draxelmayr
Infineon Technologies Microelectronic Design Centers Austria GesmbH,
Siemensstr. 2, 9500 Villach, Austria

ABSTRACT

Technological progress is inevitably linked with decreasing feature size. During the past we have learned that shrinking brings many benefits: Higher speed, lower power consumption (CU^2), and higher levels of integration. This manifests itself in giga-speed for processors, highly complex SoCs, and this even for battery operated products like hand-held phones. However, dark clouds are rising on the sky: Processor developers are talking about a power crisis, meaning that they don't know how to cool their chips. Experts are stating that analog scaling has come to an end. Development and processing cost start to become overwhelming.
Why does this happen and how will it continue?

Keywords: Moore's law, shrinking, ITRS roadmap, leakage currents, power limitation, low power design

1. INTRODUCTION

In 1965 Gordon Moore stated his famous observation, that transistor numbers double every year /1/. Although this number has been slightly revised, the core observation is still true, predicting exponential progress in the semiconductor industry /2/. In the meantime we have learned that this is not for free, but we have to pay for it with exponential cost in process development and plant erection. Nevertheless we continue our efforts because we have seen that the benefits are convincing: Better performance in terms of speed, power consumption and circuit complexity at a significant decrease of cost per function. As for any exponential growth law we can expect that also this system will be limited somehow. Experts assume that this may happen perhaps in 15 years, when quantum and single-atom effects start prohibiting the application of our very successful set of semiconductor formulas. However, it is the purpose of this paper to show that there are some major roadblocks on the way to this target, even in considering the plain conventional semiconductor behavior.

2. WHAT IS THE BENEFIT OF SHRINKING?

At first sight the answer to this question is quite obvious: Higher speed and complexity at lower power levels and cost / function. We can demonstrate this with the development in the memory business.
In 1984 we sold 64k DRAM with an access time of approx. 200ns for 1.60$. In 2001 we sold 64M DRAM with an access time of 70ns for 3.20$. This is a 500x decrease in prize/bit and a 3x increase in speed despite the much higher complexity. Fig. 1 shows the development of memory complexity versus time, whereas fig. 2 shows the achieved prize - normalized to a single bit - versus time.
We got more speed, less power and less cost at the expense of a higher initial invest in process technology. Although this is becoming increasingly unpleasant – we cross the 100 million dollar border for process development and the 1 billion dollar border for plant erection – the benefits still seem to be convincing.

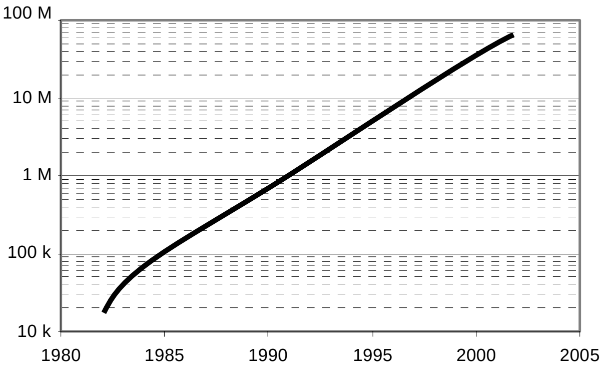

Fig. 1: DRAM: Complexity increase over time

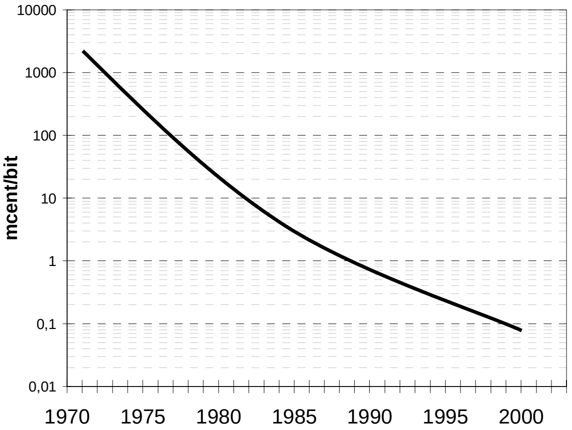

Fig. 2: DRAM: Price [milli-cent per bit] decline over time

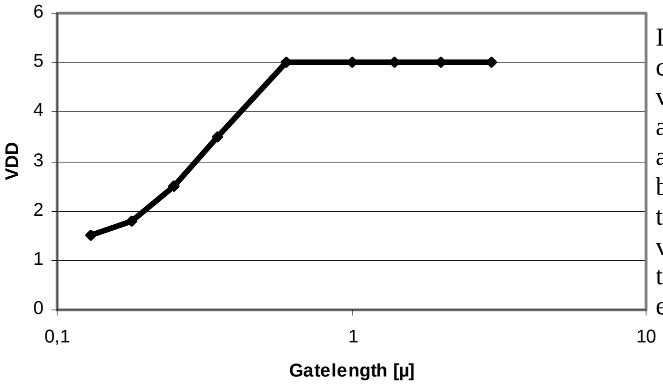

Fig. 3: Supply voltage vs. gate length

In the older days we had a supply voltage of 5V for logic chips, independent of feature size. We call this constant voltage scaling. This was valid until gate lengths of approx. 0.6µm. With further reduced feature sizes we had also to reduce the supply voltage in order to avoid breakthrough. So the basic idea of shrinking is to keep the electrical fields constant at some maximum, uncritical value. We call this constant field scaling. In first order this means that the gate length in µm multiplied by 10 is equal to the supply voltage VDD.

In the times of constant voltage scaling the power reduction came by the fact that due to shrinking lateral dimensions all capacitances got smaller and smaller. Since in CMOS logic power dissipation is mainly given by the capacitive charge and discharge currents, shrinking automatically also saves power. In the times of constant field scaling, scaling of capacitor values degraded due to the increasing role of fringing fields. Nevertheless, at the same time we started to reduce the supply voltage, which also reduces power by VDD^2. So a reduction in supply from 5V to 3.3V at constant capacitive load gives a saving factor of 2.3.
On the other hand we know that the transit frequency of a MOS transistor scales with $1/L^2$. More precisely,

ωt is related to $gm/C = 2*\mu*(Ugs-Uth)/L^2$ /3/.

We see that the previous statement is a little bit misleading, since it assumes constant gate drive and constant mobility. Nevertheless, the quadratic appearance of L helps a lot.

All these considerations lead to the conclusion, that shrinking automatically also leads to better circuit performance. We get more speed, less power and less cost at the expense of a higher initial invest in process technology and manufacturing capability. Although this is becoming increasingly unpleasant and forces us to form alliances to allow for the expenses the benefits still seem to be convincing.
In the ITRS roadmap 2001 we see the following shrink path (near term): In 2001 we have the 130nm node whereas in 2007 we expect to have the 65nm node. This corresponds to physical gate lengths for µP devices of 65nm and 25nm respectively. Researchers have already demonstrated a functional 10nm transistor /4/. So – what can we expect from these devices in terms of circuit performance?

3. RECENT ACHIEVEMENTS

If we have a look to recent conferences, we can confirm this view. We find circuits in CMOS technology, which have been ought to be impossible only a few years ago. In 2001 researchers reported a 50GHz VCO in a standard CMOS technology /5/. Of course there is a long way from making oscillators towards the creation of complete SoCs. Oscillators belong to the fastest circuits we can build in a given technology.

On ISSCC 2003 circuits have been reported with some logic function running at 40Gbit/s /6/, /7/ implemented in a standard CMOS technology. These are still no real products but tend to go into that direction. Fig. 4 shows a 40Gbit/s MUX (from /7/). We can notice that the real circuitry is more or less invisible. This is due to the modern process technology incorporating cheesing, filling and CMP (chemical mechanical polishing).

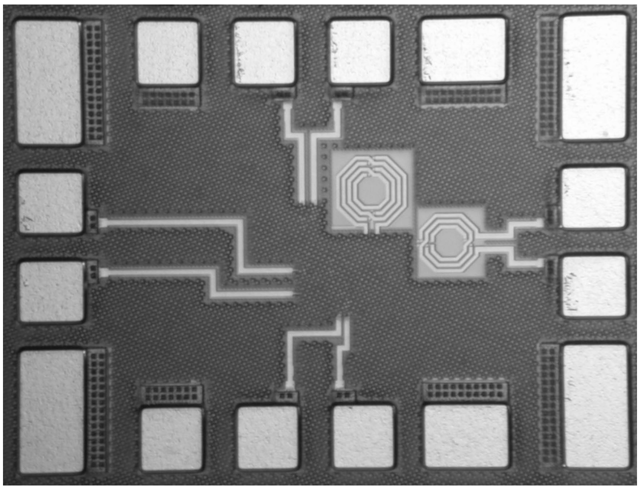

Fig. 4: 40Gbit/s MUX in 0.12µm CMOS

Fig. 5: 10Gbit/s transceiver product

Fig. 5 shows a real product: It is a 10Gb/s ethernet transceiver with a logic complexity of approx. 750k gates /8/. Due to the high speed requirements some circuit blocks are designed in a CML-type logic, which takes some extra power. Naturally, main parts of the chip are implemented in normal CMOS logic for avoiding the static current flow associated with CML logic. So for normal applications we want to dissipate the minimum necessary amount of power, which is given by driving the capacitive load on the signal lines in order to generate the necessary logic swing.

4. POWER CONSUMPTION – REVISITED

So far we have stated that the power dissipation in standard CMOS logic mainly comes from capacitive charge and discharge currents. There is another term coming from the direct current through p- and n-device during switching. Usually this term is essentially smaller than capacitive currents. In case we lower the supply to be less than Vth,n + Vth,p it is almost negligible. However, this consideration implies that there is no static leakage path. Static leakage (fig. 6) can go from drain to source, drain to bulk or gate to drain or channel. For quite a while (using high threshold voltages) drain to bulk was the main leakage source. For reasonable temperatures this was negligible. However, our scaling strategy also brought smaller and smaller threshold values. Fig. 7 shows the development of threshold voltages over technology node. It shall be noted that in older technologies the threshold voltages were pretty much similar all over the world. However, in modern technologies this has changed. Even for a specific process we usually have several threshold voltage options available. So fig. 7 shows a specific example, where threshold voltage scaling saturates. This is not a fundamental necessity, and it will be discussed later in more detail. Decreasing threshold voltage means, that subthreshold leakage increases. Taking a look in the ITRS roadmap again, the table for “high performance logic” also tells us that the supply voltage will go down from 1.2V in 2001 to 0.7V in 2007. At the same time the subthreshold leakage level will go up from 0.01µA/µm to 1µA/µm at room temperature. This is an increase of two decades within 6 years. Fig. 8 shows this dependency. There seems to be an exponential increase in the near future but it is clear that this is not an option at the long term.

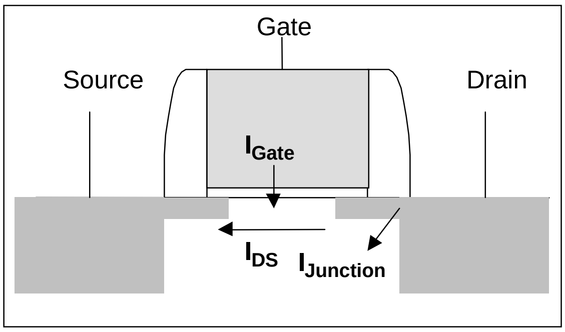

Fig. 6: Transistor leakage currents

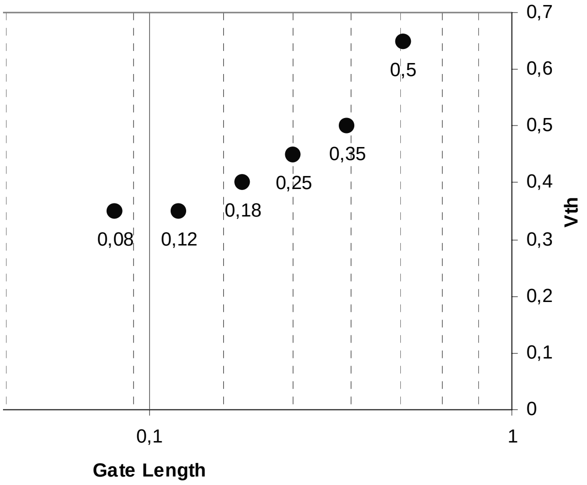

Fig. 7: Threshold voltage vs. gate length

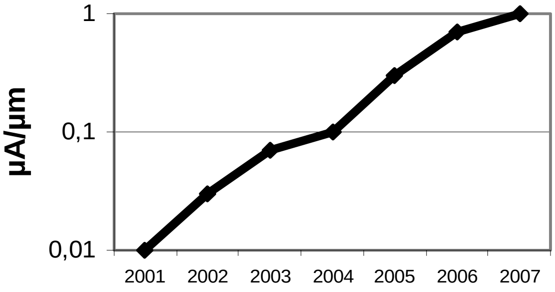

Fig. 8: Subthreshold leakage vs. year (from /2/)

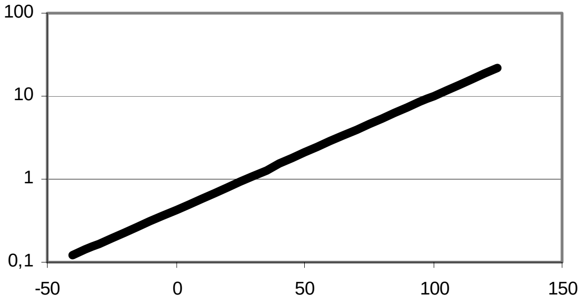

Fig. 9: Subthreshold leakage vs. temp

What does it mean for chip design? We are just entering the 100-million-transistor complexity area and are expecting a 1 billion transistor chip being available within this decade. So let's assume that we have (several) 10 million gates on a chip and each gate has a leakage current of 1µA. This leads to a total leakage of (several) 10A. Depending on temperature, this value may increase between one and two orders of magnitude (fig. 9). It is clear that this is a very severe limitation and unacceptable in most cases.

The most straightforward solution is to go to the process development people and tell them that they have to come up with devices with lower leakage levels. This has been done and it leads to a somewhat different device philosophy. The most important countermeasure is to increase the threshold voltage. Having again a look into the ITRS roadmap we find a section about "low operating power logic". We find that the MOS subthreshold current is 100pA/µm in 2001 and rises to 700pA/µm in 2007. At the same time the current drive capability is somewhat smaller than that of the high-performance device at somewhat longer gates and somewhat higher supply voltages. This implies that we have to pay for lower leakage levels with lower device speed. In fact, there is even a third roadmap available - "low standby power". Consequently, leakage levels are further reduced at the expense of even lower device speed at even higher supply voltage. It shall be noted, that not all dates out of the ITRS roadmap are completely clear today – some of them are indicated in yellow or red colours, meaning that manufacturable solutions do not yet exist or even worse, manufacturable solutions are yet unknown.

Another interesting observation is the fact, that as we try to decrease the power consumption we increase the supply voltage. This is in direct contrast to the classical paradigm that low voltage is a synonym for low power. And it does not mean that by supplying a higher voltage to a given circuit we will reduce the power consumption. It means that for certain constraints, a dedicated process technology operating at a higher supply voltage can do a certain operation at a lower power level than a process technology optimized for high speed and operating at a lower supply. To gain a better understanding we first should have a look to the speed performance of future devices.

5. SPEED – REVISITED

It has been stated above, that by shrinking speeds are always going up, since device speed is strongly related to $1/L^2$. Fig. 10 shows the maximum speed of CMOS gates (gate delay) in different process technology nodes. So in parallel to the feature size also VDD is changed. In addition, also Vth is changed, as indicated in fig. 7. However, in a power limited world we face the problem that we should not lower threshold voltages arbitrarily – so let's consider a scenario, where the threshold voltage is fixed and does not scale. Taking now the plain formula for device speed $ft \sim \mu*(Ugs – Uth)/L^2$ we can try to apply a shrink scenario. First, we replace Ugs by UDD (supply voltage). Next, we replace L by UDD (constant field scaling). µ and Uth remain fixed. Of course this is a quite crude approximation, because of effects like mobility degradation, velocity saturation, non-proportional scaling of UDD and L, 3-D effects in small devices. We then get $ft \sim \mu*(UDD\text{-}Uth)/UDD^2$. This is plotted in fig. 11. The formula says that speed is zero for UDD = infinity and UDD = Uth. The first case describes a technology with infinite feature size, whereas the second case tells us that there is no speed when there is no current (the formula neglects subthreshold operation). Obviously, somewhere in between must be a maximum. This can be found by taking the first derivative and set it to zero. The result is UDD = 2* Uth. Doing the same calculation with the assumption, that speed is related to 1/(C*Ron) gives the same result. Doing this calculation with the assumption, that speed is related to (saturation current)/(C*swing) gives UDD = 3*Uth. I do not want to judge which one of these calculation methods (if any) is best, but I want to point out that by applying plain scaling at a fixed threshold voltage leads to a point, where technology is not getting faster, but slower again. The ultimate reason for this is that with reduced feature size we have to reduce the supply voltage. When the supply voltage approaches the threshold voltage, speed degrades.

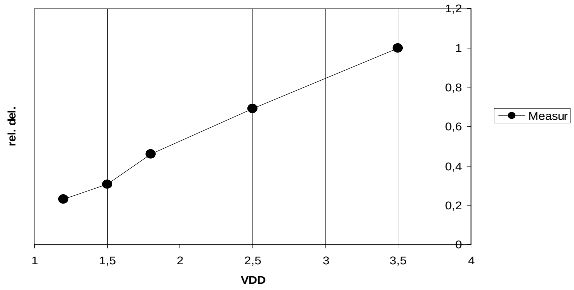

Fig. 10: (Relative) gate delay vs. shrink

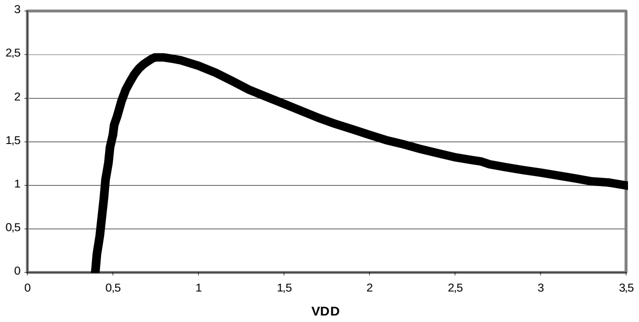

Fig. 11: (Relative) max. operating frequency vs. shrink

6. LOW POWER OPERATION

For calculating the total power we have to take into account static power dissipation as well as dynamic power dissipation. As already pointed out before, it is important to know about the application – intense computing requires a different optimization than a circuit without noticeable activity. In the past there were many optimization strategies based on the assumption, that static power is negligible. This is true for applications which operate at the maximum speed they can get, using a process with relatively high threshold voltage. This is the typical case for number crunching in an old process. Consequently, the optimization strategy is as follows:

Try to operate in parallel. Since the number of required operations is constant, the number of events will not increase very much. But since the circuits now operate at relaxed speed, it is possible to reduce the supply voltage. So, as a grand total: For a little overhead in events we get a big benefit in power saving due to reduced supply.

Now let’s discuss the opposite scenario: We have a circuit waiting for an event to occur once and a while. But when this event occurs, we want to have fast reaction. Assuming the right design style (shutting down clocks …) a circuit waiting for an event dissipates primarily static power. For reducing static power we have to raise the threshold voltages. It shall be noted, that a linear increase of threshold voltage leads to an exponential saving in subthreshold leakage. Currently, a typical value for the subthreshold slope in a bulk CMOS process is 85mV/decade. This means, that a change in threshold voltage of 85 mV brings a factor of 10 in leakage current. For getting a high Ion/Ioff-ratio we would like to reduce the subthreshold slope as much as possible. The theoretical minimum value is at about 60mV/decade. One reason for moving to SOI (Silicon on insulator) is to get closer to this value.

For getting still fast response, we have to keep the supply voltage well above the threshold voltage. Due to exponential saving in subthreshold current this is still beneficial, even if we have to go to a less aggressively scaled process. Of course dynamic power consumption is greatly increased, but due to the assumption that the total power dissipation is dominated by static leakage, raising of threshold voltages and thus also rising the supply is overall beneficial.

It can be seen, that depending on speed, activity and latency requirements there are different results in optimizing for low power operation. /9/ shows a systematic approach to explore this topic, although only a few cases focused on subthreshold operation are shown in detail. The subthreshold region is generally attractive for power efficiency due to its high gm/I ratio. However, we have to pay for this with comparatively low speed.

7. OTHER CONTRIBUTORS

Among lots of effects, two more shall be discussed: Gate leakage and wiring. As gate oxides approach the few nm-dimension we see increasing gate leakages due to direct tunnelling, fig.12 (e.g. /10/).

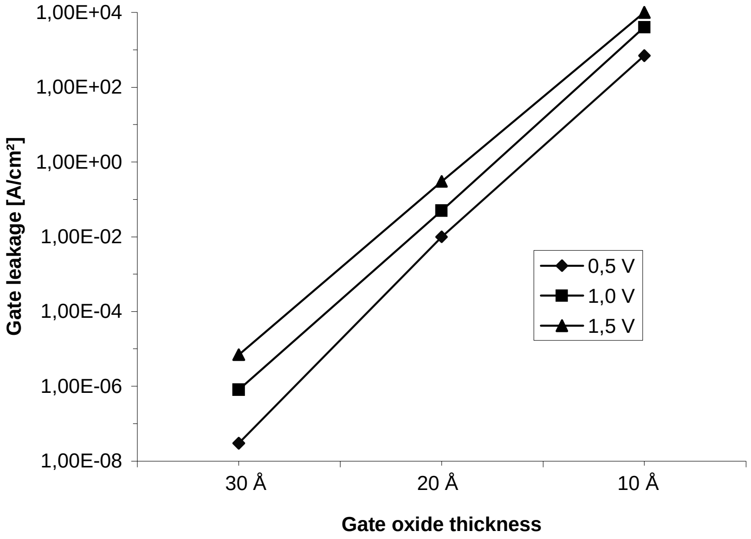

Fig.12: Gate oxide leakage vs. gate thickness

We can identify two major leakage mechanisms: In a turned-on transistor the gate leaks into the channel underneath. In a turned-off transistor the gate leaks to the drain at the gate-drain overlap region. Currently there is much research on high-k materials that will allow to reduce gate currents. Some prospective materials have been identified, but there are still some problems in process integration. In further reducing dimensions and therefore increasing doping levels due to the constant-field scaling model, we see some further effects. High doping levels lead to drain to bulk tunneling due to very thin depletion layers. Ultimately, very short channel lengths even allow direct drain to source tunneling. Today, however, subthreshold leakage is most dominant followed by gate leakage.

Another major contributor to speed and power is the signal routing. Coming from single or double alu layers in the past currently we have approached approximately 10 layers of copper metallization.

Let’s now follow the behavior of interconnects with shrinking dimensions. For old processes a linear shrink lead to a quadratic decrease in capacitance. This is simply caused by the fact that capacitance was more or less related to area. Coming to the 1µm dimension fringing fields became more important, capacitance in first order was related to the perimeter. So a linear shrink led to a linear capacitance decrease. Nowadays the height of the metal tracks is much larger than their spacing – lateral coupling between tracks dominates. Linear shrink of dimensions does not change the capacitance, because a linear decrease in length is compensated by a linear increase due to spacing. We could try to compensate for that with lower metal height – at a higher risk of electro migration and an increased resistance for that wire. Effectively the intrinsic delay of the wire would rise. This leads to constant or even increasing wire delays. In parallel we design for higher clock speeds. As a consequence the so-called isochronous zones on a chip get smaller and smaller. At the same time delays get less predictable: Mismatch starts playing an important role. Intrinsic device mismatch gets higher due to reduced device sizes and the relative error in Vgs-Vth gets larger due to the smaller absolute value. It is also increasingly difficult to define symmetric switching behavior. At reasonable high supply voltages one could consider that the switching speed of p- versus n-transistors is basically given by the ratio of mobility and W/L. At low supply voltage it is increasingly important, that both transistor types see the same gate overdrive, because otherwise the delay ratio also depends on supply voltage. However, this is a nontrivial task. We do not only face global and individual spreads of the Vth-value, we also find that the value of Vth is strongly dependent on geometry. For a minimum length device the difference between threshold voltages of a narrow versus a wide transistor can be more than 100mV.

On top of that delay is also influenced by cross-coupling of neighboring wires; two neighbor wires switching in the same direction speed up themselves, whereas the same wires switching in opposite direction slow down /11/. This problem has been relaxed somehow with the introduction of low-k dielectrics. Several materials have been found to reduce the dielectric constant of the inter-metal dielectric by about a factor of two. This is a limited resource, however. We cannot expect that future research will bring another factor of two, since 1 is a quite fundamental limit to the dielectric constant. In introducing these new materials we also had to solve some problems with bad mechanical properties.

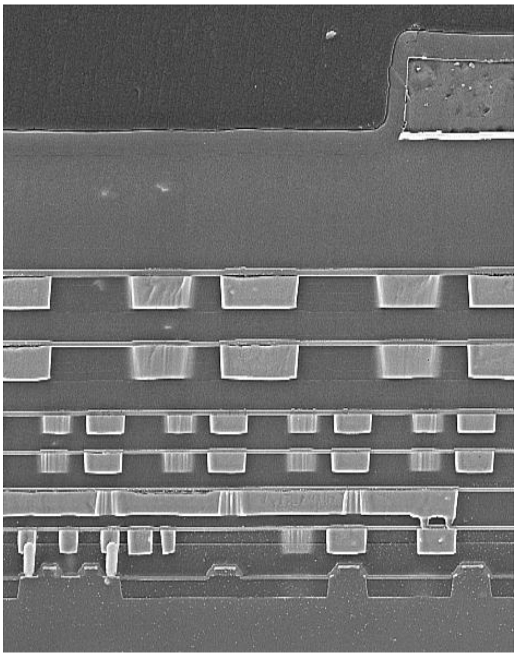

Fig. 13: Metallization cross section

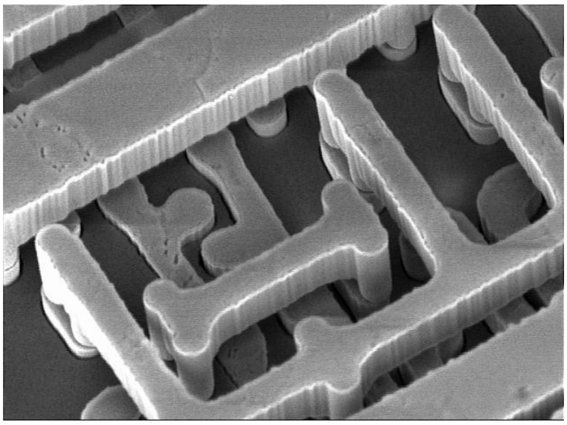

Fig. 14: Metallization photomicrograph

Another problem associated with metallization is linked to our design style: Since the very beginning of circuit design we are used to see interconnect as something immaterial. We define circuit “nodes” by names which are used to define the connectivity of a network. However, from a physical viewpoint metal is something very real and in taking a view on fig. 13 and 14 it is hard to just ignore that. Indeed the handling of metallization has played a more and more severe role in our design systems. Starting from very simple backannotation-tools extracting a simple capacitance to ground we have sharpened our view towards RC-meshes including vertical and lateral coupling between different layers. It is increasingly difficult to make accurate simulations with gigascale integration at nanoscale dimensions. It will be interesting to see who will win the race – the EDA industry in creating always more powerful tools, or the process and electronic engineers in defining highly complex design tasks. Currently it seems that complexity is somewhat ahead capability – that's also why we are talking about "productivity crisis".

8. CIRCUIT OPTIONS

The fundamental question to be answered is: How to get high computing power at low leakage level? There is no clear answer to this. As we have seen, one option is to choose the right supply voltage and the right threshold voltage for a given requirement profile. However, it is not trivial to shift voltages and thresholds. To some extent we can consider active back-biasing for adjusting the thresholds. But this is only of limited use. So for a large circuit with different requirement areas on chip we have to consider different technology options on one chip. This of course adds significant cost. Another idea is to use gate stacking for slower circuit parts. This is based on the observation, that a series connection of two cut off gates has significantly reduced leakage. Several authors have proposed multi-threshold logic (MTL). In principle, they try to combine the high-speed properties of low-Vth and the low-leakage properties of high-Vth devices into a beneficial new gate architecture. Undoubtedly the most effective method for saving power is to turn off unused circuit blocks. Turning off means, in principle, to cut the supply with the aid of a dedicated switch. This switch has to be large enough to let pass the required active power, whereas it has to be able to cut off deeply to suppress leakage. In addition, sending a circuit block to sleep and activating it again also consumes noticeable power and takes some time. Therefore it is not an option to do this clockwise. Instead this is only applicable with a burst-mode type of operation. This means that we have to adjust the architecture of our circuits towards burst-mode operation. Some systems operate in a discontinuous way by nature, like GSM-systems. There power-up and down features are implemented already today.

9. SHRINKING IN THE ANALOG WORLD

So far we have discussed the implications of shrinking in the digital world. How does all this apply for analog? Generally speaking – analog is different. As indicated, digital experts for sure see some problems but they nevertheless expect more or less unlimited shrinking until the intrinsic transistor function starts becoming obsolete. Quite in contrast, according to analog experts, we need a supply voltage of 1.8-2.5V for several analog functions. This has been stated in /12/, /13/ and can also be seen clearly out of the ITRS roadmap - mixed signal device requirements. Two reasons are most often cited as a proof: One is, that shrinking in analog leads to a larger power and area consumption. The other is, that due to low signal swing we need impractical large capacitors for getting the desired SNR (signal to noise ratio). Both reasons are related to each other. In order to achieve lower noise levels, we have to build larger devices. This is true for kT/C noise as well as for matching based devices. Both scale with a square law: Reducing the useful swing by a factor of two leads to an increase in capacitance / area by a factor of 4. For area this is not completely true: In scaled technologies we usually have a larger specific capacitance and a better matching constant. In order to drive a larger capacitance value at the same speed we also need more current. Some more systematic considerations can be found in /14/, /15/, /16/. It is interesting to note, that /14/ and /15/ have been published in the frame of the same conference and come to somewhat different conclusions about the future development. This shows that sometimes it is not completely clear which assumptions to take.

As an example for this we can consider the scaling behavior of a noise limited SC-amplifier. Usually the current consumption of such a block is dictated by the capacitive load it has to drive. For SC-circuits it is well known that a linear decrease in voltage leads to a quadratic increase in capacitance value. This is due to the so-called kT/C-noise which says, that the noise energy is indirect proportional to the capacitance value. A linear decrease in voltage leads to a quadratic decrease in signal power and therefore to a quadratic increase in capacitance value. If we now consider the slewing condition – how much current do we need to be able to supply the capacitor – we find a linear increase in current. This comes from the quadratic increase in capacitance at a linearly reduced swing. So as a grand total the dissipated power remains the same – linear decrease in supply voltage versus linear increase in supply current.

If we now consider the settling condition – we need a certain number of time constants to get a certain accuracy level – we find different results. In a typical OTA-C structure the settling time constant is given by C/gm. To keep this constant a linear decrease in voltage leads to a quadratic increase in gm, since we already have seen that C has to be increased quadratically. For a given operating point this means that we also need a quadratically increased current. So as a grand total we remain with a linear increase in power due to linear decrease in voltage and quadratic increase in current. For a real SC-circuit we

usually operate neither totally in slewing nor totally in settling condition but in a combination of both. To make things even more complicated, we also have the option of dynamically biasing an amplifier.
Regardless which calculation method we assume, the above assumptions still tend to be somewhat optimistic. This is caused by the fact that signal swings usually have to scale larger than the supply voltage. We never can provide signal swings until we reach the supply rails, we need at least some multiples of UT (= kT/q, the thermal voltage) in order to insure reliable circuit operation by keeping the transistors in saturation. This means that a linear decrease in supply voltage leads to a super-linear decrease in signal swing, which is unfavorable for the power budget, as we already have seen.
It is also quite often disregarded, that for different technology properties we should expect different architectures or circuit implementations to be the optimum solution. This means, that plain shrinking rules are quite often not applicable. A very simple example for this can be seen in fig. 15 (from /17/).

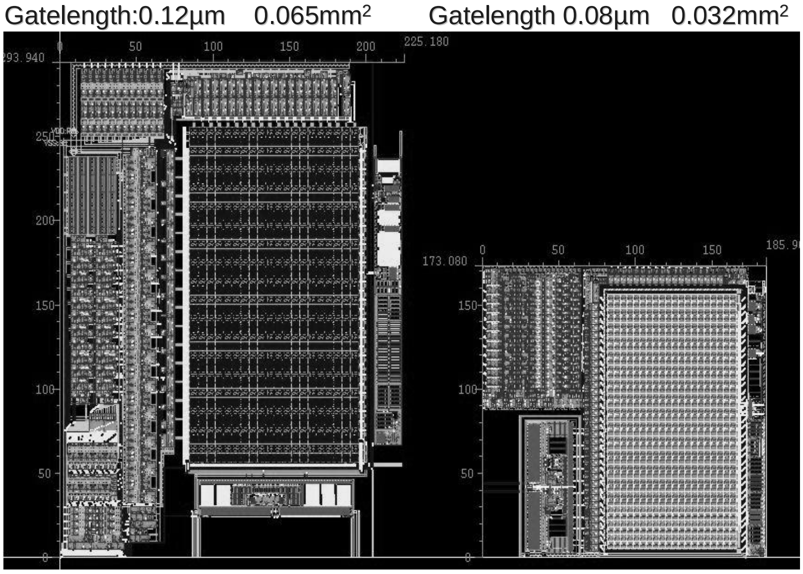

Fig. 15: A/D converter in two different process technologies

Against the straightforward prediction we can see that the 80nm design is much smaller that the 120nm design. One key for that is better area usage: By exploiting multi-layer metallization it is possible to achieve high specific capacitance and in addition we can place active circuit area under capacitance area. This is not possible in old technologies with 2 or 3 layer of metallization, but of course we should use the options available by newer technology generations. This obviously leads to some non-linear shrinking behaviour. Another benefit in newer technologies is that parasitic poles go to higher frequencies for free and effectively at lower power levels that in older technologies. This is caused by reduced capacitances due to reduced dimensions. A third major paradigm change comes by the ability to implement thousands of gates more or less for free. This enables digital algorithms to take over accuracy requirements from analog. A historical example is the advent of sigma-delta conversion plus digital filtering which has proven to be beneficial over conventional high accuracy A/D conversion. In more recent time we see examples like offset-, gain-, and timing correction of A/D converter systems or digital correction for distortion in communication systems.

10. FUTURE CHALLENGES

We have seen that many of the aforementioned problems have been addressed by researchers. For reducing delays we have seen the adoption of low-k inter-metal dielectric and SOI processes. We have introduced salicide for reduction of resistance and are considering the introduction of metal gates for further resistance reduction and avoidance of depletion effects. We have replaced aluminum by copper – with exception of the top layer, where aluminum is still needed for bonding. We have lot of ideas about new devices – more advanced elementary transistor constructions as well as new material layers for building memories. We have seen research results that demonstrate digital circuits operating in subthreshold region and analog circuit functions far beyond the predictions from ITRS roadmap. However – how are we putting all this together? Currently a modern CMOS process needs more that 30 masking steps with quite a lot of masks at critical dimensions. This does not include all the goodies we would like to have for a "really useful" process, like several different threshold voltages for optimizing logic, different gate stacks for logic, analog and IO, options for RAM and NVM devices, options for RF design. It is a big challenge to build a process integration concept that allows flexibility at low production cost. In parallel this problem as well as rising demand by rising circuit performance have also lead to a push in package development. We see several approaches for 2-D multi-chip modules as well as for 3-D wafer stacks. This rich choice of options makes it quite difficult to define something like an optimum system solution. And all these options have to be supported by CAD tooling. This leads to a scenario, where we have to design more complex chips in a more complex process and CAD environment within a preferably shorter timeframe. And any decision we take will have great impact, since due to the cost structure the benefits of modern technologies rely on heavy mass production. Out of this scenario it is understandable, that people are trying to leave the traditional ASIC design-style and want to come to something like platform-based design (eg./18/).

11. CONCLUSION

Examining Moore's law we have seen that we have been very successful in creating smaller geometries, and there are many indications that we will continue to be successful. There had been one major paradigm change in the past – which was the transition from constant voltage scaling to constant field scaling. But it seems that several paradigm changes are necessary today since several trends are running into impractical regions. The two most obvious are gate oxide leakage caused by direct tunneling and power dissipation density. Whereas the first quite likely will be cured by process technology means (i.e. introducing new materials), the second is a real mess. It turns out that we are not able to generate smaller, faster and less power consuming devices at the same time just straightforward by a constant-field shrink strategy. This leads to today's picture of diverging processes for different applications. On the other hand we face more complex designs under more complex restrictions in a more complex CAD environment at higher invest. Customer expectations are cheaper chips in shorter time. This leads to the conclusion that we are quite well prepared to continue Moore's law in their original meaning. But since this does not automatically solve all the problems there will be an increasing demand in clever ideas how to exploit this base in terms of product definition, circuit design and cost efficiency.

ACKNOWLEDGEMENTS

The author wishes to thank K. Schrüfer and G. Hirt for their support as well as A. Mikula for the preparation of the paper.

REFERENCES

1. G. Moore: "Cramming more components onto integrated circuits"; *Electronics*, Volume 38, Number 8, 1965.

2. ITRS: http://public.itrs.net

3. Y. P. Tsividis: "Operation and Modeling of the MOS Transistor"; McGraw-Hill

4. B. Yu et al: “Finfetscaling to 10nm gatelength”, *IEDM 2002*

5. H. Wang: “A 50GHz VCO in 0.25µm CMOS”; *ISSCC 2001*

6. J. Lee, B. Razavi: “A 40Gb/s Clock and Data Recovery Circuit in 0.18µm CMOS Technology”, *ISSCC 2003*

7. D. Kehrer et al: “40Gb/s 2:1 Multiplexer and 1:2 Demultiplexer in 120nm CMOS”, *ISSCC 2003*

8. Infineon Technologies: Data Sheet UPF 01012

9. A. Wang, A. P. Chandrakasan, S. V. Kosonocky: “Optimal Supply and Threshold Scaling for Subthreshold CMOS Circuits”; *Annual symposium on VLSI 2002*, Pittsburgh

10. D. Frank: “Power-constrained CMOS scaling limits”; *IBM Journal Res. & Dev,*. **Vol 46, no. 2/3**, March/May 2000

11. S. van Dijk, D. Hely: “Reduction of Interconnect Delay by Exploiting Cross-talk” *ESSCIRC 2001*

12. Panel discussion: “Does Moore’s Law Apply to Analog? Past, Present, and Future Implications of Technology Progress and Higher Levels of Integration for Mixed-Signal Circuits”; *ISSCC 2002*

13. Panel discussion: “Low-Voltage Design or the End of MOSFET Scaling?”; *ISSCC 2002*

14. K. Bult: “Analog Design in Deep Sub-Micron CMOS”; *ESSCIRC 2000*, Stockholm

15. K. Uyttenhove, M. Steyaert: “Speed-Accuracy-Power Trade-off in High-Speed CMOS ADCs: Design Example”; *Workshop on Embedded Data Converters*; September 22, Stockholm

16. A. Marques, V. Peluso, M. Steyaert, W. Sansen: “Analysis of the Trade-off between Bandwidth, Resolution, and Power in SD Analog to Digital Converters”; *International Conference on Electronics, Circuits and Systems*, 1998, Lisboa

17. F. Kuttner: “A 1.2V 10b 20Msample/s Non-Binary Successive Approximation ADC in 0.13µm CMOS”, *ISSCC 2002*, Visuals supplement

18. L. P. Carloni, F. De Bernardinis, A. L. Sangiovanni-Vincentelli, M. Sgroi: “The Art and Science of Integrated Systems Design”; *ESSCIRC 2002*