Noise Reduction in Nanometre CMOS

Mart Coenen[#1], Arthur van Roermund[*2]

[#1] *EMCMCC bv, Eindhoven, the Netherlands – mart.coenen@emcmcc.nl*
[*2] *Eindhoven University of Technology, the Netherlands – a.h.m.v.roermund@tue.nl*

***Abstract* — With nanometre scaling, the amount of transistors per 100 square millimetre will increase following Moore's Law. The maximum power will, without additional cooling, be limited to a few watt whereas the on- and off-chip clock and data speeds will increase further. To accommodate this, the core supply voltages are reduced further down to below 1 volt as where the peripheral supply voltages will have to follow international agreed voltages levels to enable interfacing. While lowering the core supply voltages, the on-chip noise margin will drop accordingly and tight on- and off-chip decoupling measures are necessary. However by application, RF switching noise from nanometre CMOS designs are forced out of their packages through the supply and ground pins when applying conventional off-chip decoupling is applied.**

In this paper, the state-of-the-art, as well as a new noise reduction technique, which is possible with today's nanometre CMOS processes, will be discussed together with guidance to accompanying complementary off-chip measures.

I. Introduction

With most sub-micron CMOS designs, noise reduction techniques have been introduced to reduce both on- as well as off-chip noise. The rationale for these micron and sub-micron measures are based on long-lasting approaches which have been used for decades. These are mostly a combination of using "standard" off-chip decoupling capacitors at all supply pins, when possible with multi-value capacitors in parallel on the PCB, even in or on-top of IC-packages.

With the introduction of nanometre technology CMOS: 90 nm and beyond, triple-well options were introduced, necessary to minimize leakage current [4, 12-14]. New on-chip circuit decoupling topologies are available which demand for complementary measures at the PCB application level.

An alternative approach was chosen. RF noise generated by the CMOS circuitry has to be minimized towards the PCB and itself. On-chip measures are derived of which the technological implementation constraints have to be evaluated.

At first, known on- and off-chip noise reduction measures will be discussed, followed by application constraints. Then, the new concept for further on-chip improvement is presented. The latter needs complementary off-chip measures.

II. State of the art RF noise reduction measures

Kelvin contacts for off-chip decoupling, figure 3, has been developed and found suited for most of the existing CMOS designs [1].

Functional identical CMOS test chip designs in 90 nm can vary more than 40 dB in RF emission, with nearly any design and layout complexity penalty [2, 3].

Various measures have been introduced to lower the amount of noise generated by CMOS ICs to ensure that the on-chip noise remains below the circuitry's noise margins and that off-chip RF emission is reduced too.

As core supply voltages for CMOS ICs have lowered from 5 volt down to less than 1 volt, on-chip noise margin has decreased proportionally. With most µCs, DSPs and even standard logic designs, the I/O supply voltage is higher than the core or functional cell voltages to ensure off-chip signal integrity. Internationally agreed standards for the various signal interfaces have been defined with accompanying levels for either asymmetric, differential or multi-level signalling.

Variable voltage and variable clock (domain) frequency have been introduced for digital cores, to lower power consumption when and where possible [4, 5, 12, 13, 14] and to achieve lower RF emission e.g. by using spread spectrum clocking techniques, see figure 1.

Embedded supply regulator concepts, based on (low drop-out/voltage regulator module) LDO/VRM regulators or even switching concepts [12], have been introduced to isolate the digital core´s internal supply from the external voltage. The main disadvantage of these series regulators is the additional power loss as all core supply current has to pass through. Additionally, on either side of the regulator, sufficient capacitance (to provide instantaneous charge) shall be present to buffer instantaneous current variations [12-16].

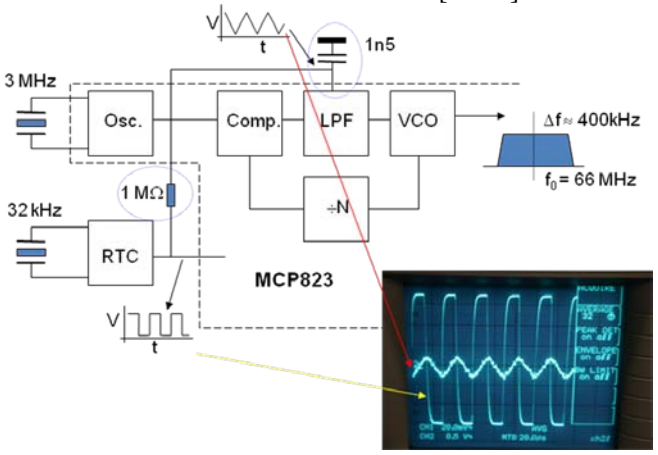

Figure 1 - Power PC application with spread-spectrum clock modulation

Clock enabling has been introduced as well as supply enabling to minimize the mean supply current, thus power. However, by disabling the clock with CMOS, when no leakage is assumed, the supply current will drop instantaneous to zero accordingly. The supply enabling is achieved by using switches in series with the supply of the functional logic. These are already in place by the triple-well topology and these can be controlled easily, figure 5. When switching on and off circuit blocks, care shall be taken that globally distributed clock and control signals are not pulled down by clamping diodes at inputs. Also critical is the instantaneous change of the on-chip currents leading to unacceptable on-chip voltage variations. For both issues, patented gradual supply enabling as well as gradual clock enabling have been developed and implemented as follow-up.

Differential signalling, the two data signals in opposite phase at every moment in time, should lead to continuous supply current. The slightest mismatch however, in switching amplitude (due to loading impedances) or by transistor topology leads to differences in rise and fall times which creates steep current spikes in the supply. These current spikes can be compensated for by on-chip or off-chip decoupling. In the latter case, unintended RF emissions will result.

On-chip decoupling capacitance can provide instantaneous charge while sustaining the supply voltages within tight bounds: typically less than 10 or 5 % drop of the nominal supply voltage is tolerated, see figure 2. This requires large amounts of on-chip decoupling, typically 10 to 20 times the simultaneous switching capacitances. Like with clock synchronous INTEL Pentium® processors, on-chip decoupling capacitance above 200 nF is required to sustain the internal supply voltage over just a few clock periods.

A substantial amount of decoupling is in fact already provided inherently in digital CMOS designs, as the number of transistors/gates which is switching simultaneously is small in ratio, figure 3. All non-active transistors/gates will act as decoupling for the dynamic transistors/gates. E.g. with memory banks, less than 1‰ of the memory cells will be active simultaneously, while all others cells are static and contribute to decoupling. On the opposite, the row and column drivers for read and write are active with every memory action.

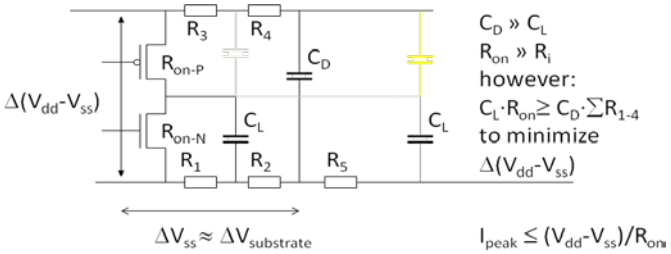

Figure 2 - Various impedances to be considered with on-chip decoupling

III. APPLICATION CONSTRAINTS

On- or off-chip voltage regulators can adapt to the on-chip current changes, but the change shall occur in the µs-region rather than within ns to make this possible. Therefore, on-chip decoupling has to take care of the supply voltage in the ns-region. Off-chip decoupling capacitances have to take over in the ns to µs-region while the series regulator takes over in the µs-range and beyond (closed loop bandwidth over 1 MHz is required).

IP blocks should be decoupled individually. This requires at least 10 to 20 times this simultaneous switching capacitance, being the sum of all internal logic cell nodes and interconnect wiring, see figure 2. Blocks which satisfy this constraint can be combined and supplied from a single supply rail without affecting one another, irrespective of simultaneous or asynchronous switching. Borrowing instantaneous charge from a critical adjacent block e.g. memory is one of the worst approaches when re-use of blocks (IP) is considered with future CMOS designs.

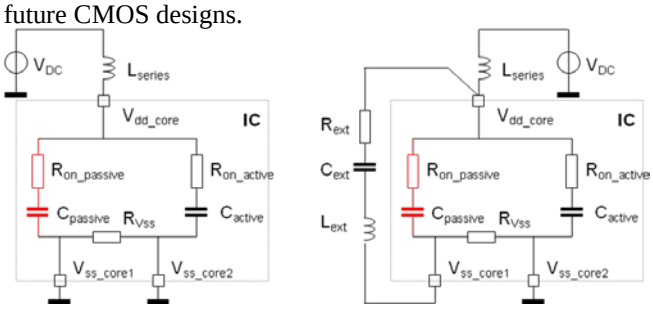

Figure 3 - Off-chip Kelvin contact decoupling

An example: when 1 million flip-flops (FFs) would switch simultaneously with 25 fF of summed load/FF (= 25 nF total) and would get charged to 1 volt within 100 ps, the total peak current would be 250 amps, while the average current at a clock rate of 1 GHz will be 25 amps. To sustain the on-chip supply voltage of 1 volt -5 %, a supply dip of only 50 mV is allowed, requiring total on-chip decoupling capacitance of 500 nF to sustain only a single event. Even when trench capacitors are used with a capacitance of 250 nF/mm^2, 2 mm^2 of non-functional nanometre silicon area will be required [16].

Cu-metal layers, with a square resistance of only 10 mΩ, can lower the interconnect resistance level. Even if various metal layers are taken in parallel, and if the above mentioned load is equally distributed over the die area, the local supply voltage will still collapse due to the on-chip IR-drop in the order of 160 mV. See figure 4 as a real example. Additionally there is a need that the RC-time for decoupling is equal or shorter than the switching event itself (which in the above given example is the case), see figure 2.

Any on- or off-chip inductance in series with the supply e.g. 1 nH, will exceed the voltage limitations definitely; 1 nH would yield a voltage drop far beyond the nominal supply voltage [7 - 10].

Spreading the switching events uniformly over the clock period will lower the peak current down to the average current yielding the lowest possible IR-drop. This approach is in use with pipe-lined processors where the clock ripples along with the data from input to output. This becomes increasingly

effective for enhanced clock skews down to 10 ps as these would yield a 10 times higher peak current but not the average supply current, and thus the average losses, which remain the same.

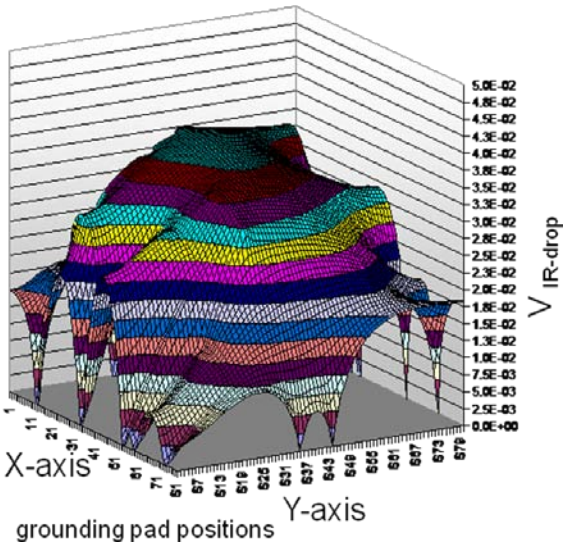

Figure 4 - Example of an Excel based IR-drop pre-estimator for real supply/ground pin allocations showing steep gradients

From the example it will be clear that vast measures are necessary to fulfil the on-chip supply constraints, even when assuming that the peak currents occur homogeneous over the chip area towards the supply pads at the die circumference.

Lowering the supply voltage will lower the peak current proportionally. Lowering the voltage swing will also lower the absolute noise margin proportionally (unless differential signalling is used).

Lowering the clock frequency will lower the average current but will not lower the instantaneous peak current, when all FFs would still switch simultaneously within a small window of time.

Variable supply voltage in combination with variable clock frequency are being used to lower peak and average current. The variable frequency in itself will not affect the peak current, but a variable supply voltage will lower the voltage swing and due to that the rise and fall time will lower too (higher R_{on}, larger depletion capacitance), thus decreasing the peak current.

IV. NANOMETRE PROCESS MEASURES

With the introduction of deep sub-micron or nanometre processes, the phenomenon of current leakage came along which made it necessary to switch-off the fast functional transistors with low V_{Th} from the global supply by slow ones with high V_{Th} when the circuitry is not in use. To lower the leakage currents in active mode, the fast "inner" transistors needed to be back-biased towards their substrates, figure 5 [4].

The high V_{Th} transistors (with low leakage) at bottom and top are typically turned on and off simultaneously, controlled by an on-chip power management sub-processor. The typical supply turn-on time needs to be large (in the µs-region), to enable the (external) supply system regulator to adapt to the changing supply current of the functional block(s) without causing a global supply dip.

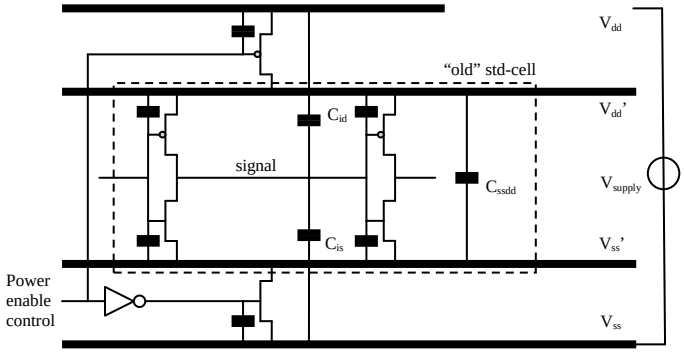

Figure 5 – Supply enabling using triple wells

The opposite problem, a global supply increase due to instantaneous reduction of supply current, occurs when series inductances (bondwires, leadframe) are present in the supply lines.

Another problem is that all the "local" block supply voltages will be less than the global net supply voltage so that level shifters between the various functional blocks, running at different supply voltages, need to be used [4, 6, 9 –16].

If supply enabling, figure 5, is combined with on-chip decoupling, figure 2, the internal circuit between $V_{ss'}$ and $V_{dd'}$ should be decoupled by C_{ssdd}. As C_{ssdd} is considered the decoupling capacitance within that functional block, it needs to be just a little larger than the simultaneous switching capacitance of that functional block. However, this approach is only correct if additional capacitance between V_{ss} and V_{dd} is present and suited (time-constant) for decoupling when the top and bottom transistors are turned on and in saturation.

When both transistors are fully turned on, the global V_{ss} line becomes the signal reference for the asymmetric logic circuit. The back-bias voltage for the fast low-V_{Th} transistors, appearing across the high-V_{Th} transistors, is lost and shall be generated separately by a charge pump. When these switches are turned on, most of the switching currents flow through the global nets rather than the local nets and the model of figure 2 w.r.t the ground and supply net drop can be re-applied.

This problem can be solved by slightly and equally turn on the top and bottom transistors. The voltage drop between V_{ss} and $V_{ss'}$ and V_{dd} and $V_{dd'}$ will be symmetric and can be used to back-bias the fast transistors [4]. The high-V_{Th} transistors than behave as equivalent resistance which together with C_{ssdd} and the capacitance/ impedance of the global V_{ss}/V_{dd}-net will form a π-type low-pass filter. The low-pass corner frequency which is achieved will be in the 100 – 500 MHz region as the on resistance still needs to be low.

The top and bottom transistors can, with a slight change of the control circuit topology, be turned into current sources

(with a small stray capacitance in parallel). By doing so, the corner frequency of the filter will lower drastically into the low MHz region which is typically far less than the functional clock frequencies used. The current source shall be controlled such that a functional and power optimal supply voltage is maintained. As a result, the local switching noise will be filtered off towards the global net. Moreover, different from the on-/off-state of the top and bottom switch application, the "virtual" signal reference now remains symmetric and equal to half the external supply voltage. The digital signal state is determined by the threshold window which is assumed symmetrical across the "virtual" reference voltage. So, in most cases no in-between level shifters will be required as the circuit's output voltage swing will exceed the threshold window at its "virtual" reference.

The RF noise reduction which can be achieved equals the ratio of the stray capacitance over the current source transistors to the capacitance/impedance of the global supply net. Furthermore, as all switching currents become isolated from the global nets, neglectable RF noise will appear between the multiple grounds and substrate coupling will diminish. Typically DC current crowding can remain at the die edge as given in figure 4. However with the above measures, the core's off-chip V_{ss} and V_{dd} pins at the package level will only carry average supply current and ground bounce will be diminished.

The approach given in figure 5 is further enhanced to a topology which partly enables charge recovery too. This extended concept has been patented under WO 2006/051485 A1 under application number PCT/IB2005/053669.

V. Complementary off-chip measures

Off-chip decoupling for the core of a digital IC can be further enhanced by supplying it through series inductance rather than applying (standard) off-chip parallel capacitance to provide additional RF series impedance.

The optimal way to supply the peripheral circuitry is determined by the signal interface chosen. On-chip decoupling is very efficient w.r.t. the elimination of sub-nanosecond glitches from differential signalling topologies.

With asymmetric digital signal interfaces, joined Kelvin contacts are recommended for decoupling towards the accompanying ICs, often obtainable from the multiple V_{ssE} pins of the ICs (multiple V_{ddE} pins are not needed, fig. 3).

VI. Conclusions

Compared to the existing CMOS designs where embedded supply regulators LDO/VRM are used to reduce RF noise, further enhancements can be realized by utilizing nanometre on-chip process and circuit topology measures.

Triple-well topologies in nanometre CMOS technologies can be used in various ways: for lowering leakage current, lowering supply voltage thus current and lowering RF noise.

When the local circuit is supplied symmetrically, all signals become "virtually" referenced to half the external supply voltage and no level shifters are needed.

On-chip decoupling measures in nanometre CMOS technologies remain inevitable. The capacitance required will be 10 - 20 times the simultaneous switching capacitance (or 10 - 20 times of the delta capacitance in case of symmetric/ balanced logic).

Individual decoupling shall be provided to each functional block (IP). This is to enable re-use when equal supply voltage variation requirements apply at the global net level.

The RC-time constant for decoupling has to be equal or less than the minimum switching event occurring. Otherwise the charge arrives too late to decouple the switching event.

The requirements for the currents sources, driven as mirror will be determined by functional and power optimisations.

Off-chip decoupling, even through very small series inductances, will have too high time-of-flight and will always be too late to provide instantaneous on-chip decoupling in the sub-ns region.

VII. Acknowledgement

The work carried out is part of an investigation commissioned by CTO NXP Semiconductors, the Netherlands. NXP also holds the WW patent on charge recovery.

VIII. References

[1] M. Coenen, D. de Greef, Optimizing techniques for minimizing IR-drop and supply bounce, EMC Compo, Munich, 2005

[2] L. **v**an Wershoven**,** Characterization of an EMC test-chip, IEEE Symposium on Electromagnetic Compatibility, Washington, 2000

[3] M.J.Coenen, R. Derikx, Design of Experiments on an EMC test chip for the interrogation of SI and EMC measures, IEEE EMC Symposium on EMC, Istanbul, 2003

[4] P.R. van der Meer, Low-Power Deep Nanometer CMOS Logic, Sub-threshold current reduction, PhD thesis TUD, 2003

[5] Sung-Mo Kang and Yusuf Leblebici, CMOS Digital Integrated Circuits, Analysis and Design, McGraw Hill International Editions, 2nd edition, 1999

[6] Harry Veendrick, Deep-Submicron CMOS ICs, From Basics to Asics, Kluwer Academic Publishers, 2000

[7] Danardono Dwi Antono, Modeling of Inductive Interconnect Responses and Coupling Effects in Deep-Submicron VLSI, Master Thesis, Department of Electronic Engineering Graduate School of Engineering The University of Tokyo, January 31, 2003

[8] Brian Young, Digital Signal Integrity: modeling and simulation with interconnects and packages, Prentice Hall

[9] Kaveh Shakeri and James D. Meindl, "Compact Physical IR-drop Models for GSI Power Distribution Networks", Proceedings of IEEE International Interconnect Technology Conference, June 2003, pages 54-56

[10] Sanjay Pant and Eli Chiprout, "Power Grid Physics and implications for CAD", Proceedings of the 43rd Annual Conference on Design Automation (DAC 2006) July 2006, pages 199-204

[11] Stéphane Donnay, Georges Gielen, Substrate Noise Coupling in Mixed-Signal ASICs, Kluwer Academic Publishers, 2003

[12] Volkan Kursan, Eby G. Friedman, Multi-Voltage CMOS Circuit Design, John Wiley & Sons Ltd, 2006.

[13] Mohab Anis, Mohamed Elmasry, Multi-Threshold CMOS Digital Circuits, Managing Leakage Power, Kluwer Academic Publishers, 2003

[14] Christian Piguet, Low-Power Electronics Design, CRC Press, 2004

[15] Madhavan Swaminathan, A. Ege Engin, Power Integrity Modeling and Design for Semiconductors and Systems, Prentice Hall, 2007

[16] P.v.d. Wiel (Philips Semiconductors/NXP), Internal communication on IR-drop calculation methods and application of current spreaders.