A SOCKET INTERFACE FOR GALS USING LOCALLY DYNAMIC VOLTAGE SCALING FOR RATE-ADAPTIVE ENERGY SAVING

Tormod Njølstad[1], Odd Tjore[1], Kjetil Svarstad[2,1], Lars Lundheim[2,1], Tor Ø. Vedal[2,3], Jukka Typpö[1], Tor Ramstad[1], Lars Wanhammar[1,4], Einar J. Aas[1], Helge Danielsen[2,3]

[1]Norwegian University of Science and Technology (NTNU), Norway. (tormod@fysel.ntnu.no)
[2]SINTEF Telecom and Informatics, Trondheim, Norway.
[3]Presently with Nordic VLSI ASA, Trondheim, Norway. [4]Linköping University, Sweden.

ABSTRACT

A low-power amendment to the globally-asynchronous, locally-synchronous paradigm (GALS) is presented. A proposed socket interface provides locally dynamic voltage scaling (LDVS) adapted to the actual processing rate requirements for each module. Local clock generation by ring oscillators tracks the local device speed at the local supply voltage. Asynchronous I/O handshakes provide the elasticity required to permit the substantial intra-die delay variations present in low voltage, deep sub-micron digital designs to give high yield. By the proposed socket interface, synchronous design methodology can be continued at module level including use of industry standard HDL-based synthesis tools, while module reuse and system design can be substantially simplified.

I. INTRODUCTION

Synchronous design and test methodology is mandatory in current industrial quality ASICs and SoCs. State-of-the-art HDL-based verification, synthesis, layout and test tools efficiently support the synchronous design flow, while new methodologies and standards emerge for designing reusable cores to increase the design productivity. Synchronous design methodology has several proven qualities and has come to stay. However, as the semiconductor technology scales down and complexity increases, fully synchronous designs will eventually not be feasible on the chip level. For increased design productivity, simplified system design, higher yield, lower energy consumption and higher performance, it is imperative to depart from the *fully* synchronous chip design paradigm, and replace it with a globally-asynchronous locally-synchronous paradigm (GALS). As will be discussed below, the GALS principle and the voltage scaling principle, may join forces to provide the required performance at minimum energy for each module, locally adapted to actual process parameters and temperature. Low-voltage digital deep sub-micron designs suffer from intra-die parameter variations, i.e. substantial delay variations over the chip [6]. The asynchronous handshake mechanisms between the modules in GALS represent the elasticity and the encapsulation required to design and optimize each module separately, in contrast to the fully synchronous design paradigm which ultimately requires chip level timing optimization.

The paper is organized as follows: In the first part, we review the Semiconductor Industry Roadmaps from 1994, 1997 and 1999 regarding their predictions about design methodology, to identify a sudden breakthrough for the GALS concepts in the 1999 Edition. Then, the GALS concept from its origin in the 1960s to more recent contributions is briefly reviewed. Furthermore, various aspects of the application of the voltage scaling principle are discussed, including validation of a performance model at gate and module level, and a discussion concerning intra-die delay variations at low voltage supplies.

In the second part, we present key elements of the Socket Interface concept, proposed and established in the WIRAC research project [7]. The Socket Interface supports GALS and module level local dynamic voltage scaling (LVDS) to provide rate-adaptive energy saving, while continuing to adhere to the established synchronous HDL-based synthesis design flow at module level. Finally, we identify obstacles and future challenges towards the implemention of the proposed Socket Interface.

II. MOTIVATION FOR GALS

A. The 1999 Roadmap urges development of GALS methodology.

Four years ahead, in year 2005, the expected on-chip global clock rates are approaching 3 GHz in ASICs and SoCs. The switching time of the fastest buffered global signal will be more than 100 times higher than the gate delay, thus fully synchronous operation on-chip will be very challenging and will imply multi-cycle latencies in the interconnect, according to 1999 ITRS Roadmap [9].

0-7803-6741-3/901/$10.00 © 2001 IEEE

TCL 1009
Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on June 27,2024 at 18:51:27 UTC from IEEE Xplore. Restrictions apply.
p. 1

The previous roadmaps have *not* been discussing such design issues. While concerning several researchers at that time [17], the increasing gap between interconnect delay and gate delay related to scaling was not an issue in the first 1994 SIA Roadmap, even though it was indeed well-known, e.g. by the book of Bakouglu [16] in 1990. Moreover, the 1997 edition just briefly compared interconnect delay and gate delay, only to address the need for interconnect technology improvements without discussing future design obstacles for the fully synchronous design paradigm.

However, the 1999 ITRS edition imperatively expresses both short-term and long-term needs for design methodology and tools to cover asynchronous, multi-cycle paths as well as locally synchronous, high-performance near-neighbor communication. The ITRS editions are based on concensus within the leading industry experts and the research community world-wide. Therefore ITRS' recognition of the globally-asynchronous locally-synchronous (GALS) design paradigm definitely is a strong incentive to further and enforced research on related GALS design methodology and tools.

B. The renaissance of self-timed locally synchronous regimes with stoppable clocks.

The idea of self-timed locally synchronous modules with stoppable clocks dates back to the 1960s with several design projects conducted by Charles L. Seitz [18]. The GALS concept was the topic of the PhD Thesis of Chapiro in 1984 [10], and several contributions have since then been reported, among them [11][12][13][14][15].

When Seitz back in 1980 [18] discusses local stoppable clocks and stretchable clocks, it is interesting, from a technological and historical point of view, that his motivations are strikingly relevant to the design challenges of *this* decade's complex ASICs and SoCs:

- Global clock distribution with problems related to skew limitations and electromigration. By means of such local clocks, these problems will be significantly reduced.
- Performance spread and yield problems due to variation in process parameters, aging, temperature and supply voltage. Seitz argues for using local ring oscillators as clock sources, since "on the same piece of silicon these variable factors are nearly the same for the clock and for the system".
- Design complexity management. Seitz' self-timed locally synchronous modules with asyncronous communication encapsulate and hide details of the underlying physical implementation, which is not relevant to the system architects. By such design abstraction "free of hidden rules", Seitz introduces a system design domain similar to that of a programmer, called the sequence domain abstraction. This is in many aspects relevant to today's discussion about increasing the design reuse to meet the productivity gap and time-to-market requirements, and the need for new abstraction levels to manage complexity.

C. Current GALS methodology research.

Today, several research groups work on complete design methodology for globally asynchronous on-chip communication networks [12][13][14][15][23], which may connect both locally-synchronous and entirely asynchronous modules. Muttersbach *et al* [14] give a very clear overview of GALS, present problems to be solved, and propose solutions very instructively.

III. MOTIVATION FOR GALS-LDVS

Digital design for low power consumption by using low supply voltages was exploited already in the 1970s in digital wrist-watches [21]. In the 1990s, however, design methodology to reduce power consumption was systematically reviewed and enhanced by Chandrakasan and Brodersen [1], among several others. Several researchers argued for utilizing the voltage scaling principle leading to a quadratic reduction in energy per operation, and to treat actual supply voltage as a design variable [1][2]. Use of complete systems with variable supply voltages has been demonstrated in academia, e.g. [8][19][20]. Recently, voltage scaling is introduced in several consumer products for mobile computing based on e.g. Transmeta's Crusoe microprocessor [26].

A. The performance model for voltage scaling applied at gate level and at module level.

As shown in Figure 1, we have validated by measurements on several test chips [5] that the simple model of delay vs. supply voltage [3] given by $t_{pd} = k \cdot V_{dd}/(V_{dd}-V_t)^{\alpha}$, with $\alpha = 1.55$ for the AMS 0.8 µm process, is quite accurate for simple CMOS inverters. It is also rather accurate for various cells using different circuit techniques, if increased threshold values due to the body effect of stacked devices are used.

Since digital systems are collections of such cells, one may expect that the inverse of the delay model can be applied as a model for performance versus supply voltage. We have demonstrated the validity of this performance model at module level [5] as shown in Figure 2, by scaling and overlaying the performance model to the plot of measurements of maximal sample rate vs. V_{dd} from [22] of complete digital filters functions, implemented in the 0.8 μm AMS technology, and where we have used $\alpha = 1.55$ as above. Kai Chen and Chenming Hu have reported $\alpha = 1.3$ to be valid for

TCL 1009
Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on June 27,2024 at 18:51:27 UTC from IEEE Xplore. Restrictions apply.
p. 2

subhalf-micrometer MOSFET's for a wide range of technologies from L_{eff} = 0.4 μm down to L_{eff} = 0.1 μm [4].

B. The justification of using ring oscillators as local clocks.

The performance model is strikingly accurate, and is a clear justification of the principle of using a ring oscillator as clock source, since the cells in the ring oscillator follow the same delay model as the complex module itself. Some special functions, like certain sense amplifier topologies used in RAM, do not follow this performance model, and have to be characterized carefully, or must be completely avoided if the performance model is applied at the module level.

C. Extended voltage scaling for reducing the local speed vs. global speed mismatch.

Liu and Svensson [2] even suggests more aggressive voltage scaling as a means to alleviate the mismatch between the MOS device speed and the global signal speed, since this

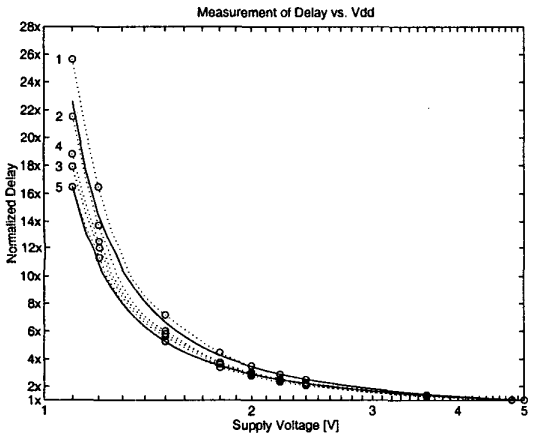


Figure 1: Validation of the delay model at cell level. Dotted line curves: Measured delay vs. supply voltage for different circuit techniques (5 = CMOS inverter chain). Lower solid line curve: delay model with no body effect (inverters). Upper solid line curve: Delay model with increased threshold voltages (due to stacked devices) [5].

mismatch increases dramatically due to the technology scaling.

In this way synchronous design methodology can be utilized further, without need for special high speed interconnect since the clock rate is lower, while parallelism may compensate for degraded performance due to voltage scaling.

D. Low voltage means increased delay sensitivity and intra-die delay variations.

By reducing the supply voltage the delay is increased, while the energy per operation is quadratically reduced. However, an unwanted side effect of voltage scaling is that delay sensitivity to process parameter variations, temperature and supply voltage fluctuations increases substantially. The required design margins are dramatically higher, compared to designing at conventional supply voltages [2][6]. Furthermore, intra-die variations can not be ignored. Highly pipelined designs, and designs with a large number of critical paths and with a small logic depth are most sensitive to uncorrelated gate delay variations in state-of-the-art and future technologies [6]. Supply voltage regulation techniques can be used to compensate for process parameter variations for maintaining given throughput performance requirements, but this implies ultimately *local* supply voltage regulation mechanisms to track intra-die variations.

IV. THE PROPOSED SOCKET INTERFACE

As a part of the Wideband Radio Access project WIRAC [25], supported by the Norwegian Research Council's Basic Telecommunication Research Programme, we have studied implementation issues concerning low-power digital design and digital signal

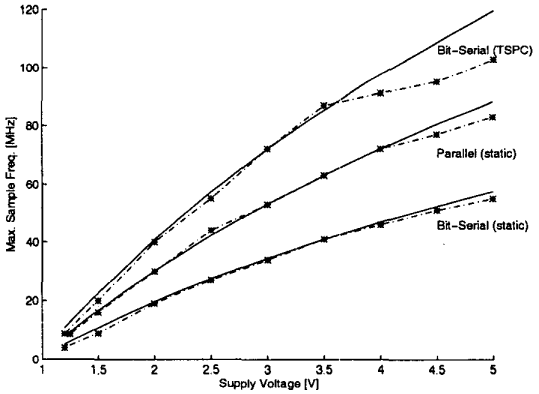


Figure 2: Validation of the performance model at module level. Dotted line curves: Test chip measurements of maximal sample rate of three different wave digital filter architectures [22]. Solid line curves: The performance model, expressed as the inverse of the simple delay model with α = 1.55, (scaled to the measured values at 3 V), see [5].

processing in wireless terminals [27]. Before presenting the Wirac's Socket Interface as a general vehicle for

TCL 1009
Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on June 27,2024 at 18:51:27 UTC from IEEE Xplore. Restrictions apply.
p. 3

implementing rate-adaptive energy-minimal systems by means of state-of-the-art HDL-based synthesis design flow, we will discuss key elements in multiple supply voltage systems and present some results and conclusions from test chip measurements.

A. A model of an integrated radio access system

An integrated digital tranceiver may be partitioned as in Figure 3, where all the modules with their individual performance and latency requirements are clustered into separate supply voltage regions, thus giving potentially the lowest energy per operation while satisfying the performance requirements. The different regions are communicating through logic level shifters, as presented below.

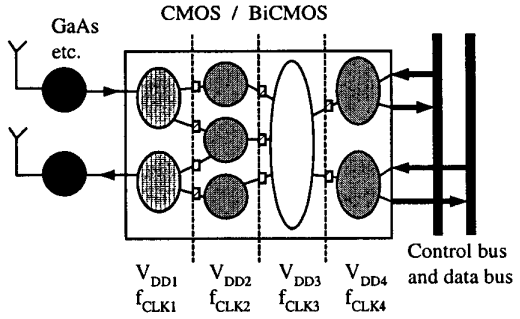


Figure 3: A model of a wideband radio access system using multiple supply voltages and logic level shifters.

B. Characterization of the level shifter cell.

We have evaluated different logic level shifter topologies. The well-known topology reported by Vittoz in 1977 [21], and utilized in [1], based on cross-coupled PMOS devices, was chosen due to its circuit and conversion symmetry. The cell was characterized by test chip fabrication and measurements in 0.6 μm AMS technology, showing good agreement with simulation results. The topology was also simulated in 0.5 μm and 0.35 μm processes from Alcatel Mietec. Such level shifter translates a logic "1" between two regions of different supply voltages V_{dd1} and V_{dd2}.

Table 1: Delay measurements of an inverter (INV) and a level shifter (LS).

V_{dd1} [V]	0.9	1.0	1.1	1.3	1.5	1.7	2.2	2.7	3.3
t_{LS} [ns]	83	22	9.6	6.8	6.1	5.9	5.5	4.7	4.8
t_{INV} [ns]	252	37	12	3.4	1.8	1.2	0.7	0.5	0.4
t_{LS}/t_{INV}	0.3	0.6	0.8	2	3.4	4.9	8.4	10	12.9

Table 1 shows the propagation delay t_{LS} of a level shifter (LS) which translates an input swing of V_{dd1} to output swing V_{dd2} = 3.3 V, compared to the delay of an inverter t_{INV} at V_{dd1}. The delay of the LS is 0.3 – 2.0 times the delay of an inverter for low voltages V_{dd1} = 0.9-1.3 V.
Table 2 shows that the energy W_{LS} per level shifting operation (from V_{dd1}=1.1V to V_{dd2}) is about 2 times that of an inverter W_{INV} supplied by the highest supply voltage V_{dd2}. Thus, level shifting from a low voltage V_{dd1} to a high voltage V_{dd2} represent a delay cost (calculated at V_{dd1}) and translation energy cost (calculated at V_{dd2}) corresponding to about two inverters in series. In contrast, level shifts from high to low voltage do not require level shifters, but can always be performed directly by an inverter or buffer supplied by the low supply voltage.

Table 2: Measurements of energy per operation of an inverter and a level shifter (LS).

V_{dd2} [V]	1.3	1.5	1.7	2.2	2.5	2.8	3.1	3.3	4.0
W_{LS} [pJ], energy per LS op.	0.22	0.35	0.45	0.73	1.1	1.4	1.8	2.2	3.7
W_{INV} [pJ], energy per INV op.				0.37				0.87	
W_{LS}/W_{INV}				2				2.2	

C. Means to generate different local supply voltages

From above, we have concluded that the cost of level shifters is not an obstacle for introducing a multiple-supply voltage methodology. However, regulating and distributing different supply voltages to the different regions in Figure 3 are regarded as more challenging. Different scenarios have been considered:

1. Almost all modules of the chip run at a fixed low supply voltage to minimize power consumption. Parallel processing is utilized to fulfill the performance requirements, as described by [1]. Some very few modules related to the analog front-end, minimum latency DSP functions and fast i/o interfaces require higher supply voltages. The higher supply voltage level may be generated by voltage doublers or triplers. Voltage doublers were already used in commercial wrist-watches in the 1970s [21]. Favrat *et al* [7] have described efficient voltage doublers with external or fully integrated capacitors. We have studied their proposed topologies to specify and design an on-chip dynamically programmable generator to provide local supply voltage of 1x, 2x or 3x of the low input supply. No test design chip has yet been fabricated.
2. Two or more different supply voltages are available

TCL 1009
Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on June 27,2024 at 18:51:27 UTC from IEEE Xplore. Restrictions apply.
p. 4

and distributed on-chip, while the actual supply voltage for each module can be selected by means of DC switches, i.e. large transistors with low $R_{DS(ON)}$. This method is considered to be a viable alternative to the dynamically programmable generator described above, but performance degradation due to IR drop, and other practical aspects of this method, have to be studied further.

3. DC/DC buck converters like in [1][8][20][24] can provide finer regulation of the supply voltage than the two methods above, but require large external components for the low-pass filter [24], so they are not suited for several locally controlled on-chip voltage supply generators.

The candidate methods for local supply voltage generation should be explored further including their suitability for future technologies, because it is only by local supply voltage selection rate-optimal energy saving can be attained.

D. Data-flow oriented processing

Typically in many signal processing system, e.g. as in Figure 3, many tasks are performed at a fixed rate (or at a temporally fixed rate). Each processing element or complex module, see Figure 4, receives an input data set and produces an output data set at this fixed rate. The input and output data sets may be passed along with handshake signals which implement some given protocol, such that e.g. the processing does not start before the module has input data ready. The fixed rate requirement or the fixed average rate requirement can dynamically be configured, by the special control signal *RATE-REQUIREMENT*$_i$. From a system point of view, faster operation of the processing element than given by the rate requirement does not give any better system performance, while slower operation than the requirement can of course not be tolerated. We have

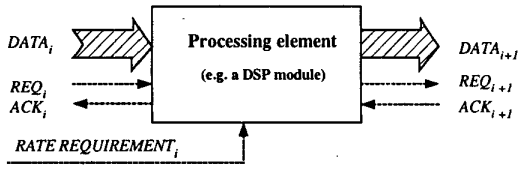


Figure 4: The prosessing element (e.g. a DSP module) executes its calculations fast enough to satisfy the actual performance requirement.

encapsulated all implementation details in the model in Figure 4. In Figure 5 some of the hidden details are revealed. The control signal *RATE-REQUIREMENT*$_i$ can be used to scale down the local supply voltage such that the locally generated clock rate is just high enough to perform the actual processing task and to produce output data at the given rate. The energy per operation will therefore be regulated to the lowest possible value for the actual rate-requirement, the given technology parameters and temperature, since the local clock is generated by a ring oscillator which tracks these local conditions.

Furthermore, the number of clock periods required per processing operation can be predefined, and the local clock can be stopped after the operation is performed. By stopping the clock when no data processing is performed, the power consumption will be further reduced. The clock stopping control circuit will, however, include e.g. counters which represent some contribution to the power consumption.

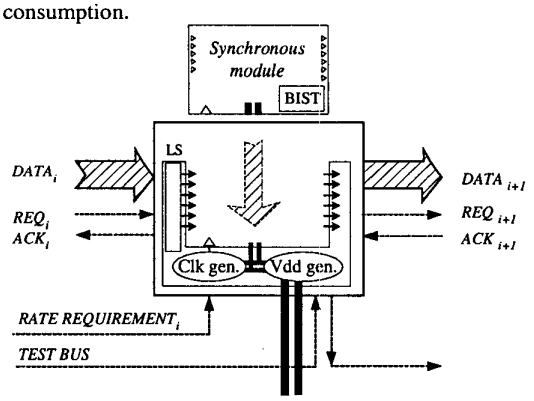


Figure 5: A synchronous version of the actual prosessing element (DSP module) is inserted into the Socket Interface.

The synchronous module in Figure 5 can be generated by state-of-the-art HDL-based synthesis tools including built-in self-test support, and can be "inserted" into the Socket Interface. The socket interface provides level shifting of all data inputs to convert from the low voltage data signals outside the socket, to the module's actual supply voltage level.

A clock signal with the correct frequency given by the logical depth of the synchronous module and by the actual supply voltage level, is active for a predefined number of clock periods for the actual task. The synchronization tasks related to the asynchronous interfaces can be performed as proposed by [14].

V. CONCLUSION

We have described GALS-LDVS, a low-power amendment to the GALS paradigm where locally rate-adapted dynamic voltage scaling can be provided by means of the proposed Socket Interface. Local clock generation by using local ring oscillator tracks the local device speed. The module's speed is proportional to the

TCL 1009
Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on June 27,2024 at 18:51:27 UTC from IEEE Xplore. Restrictions apply.
p. 5

device speed with the same dependence of local power supply level, process parameters and temperature variations. Asynchronous I/O handshakes support the elasticity required to permit low voltage intra-die delay variations, thereby increasing the performance yield.
By exploiting the proposed methodology, synchronous design and HDL-based synthesis flow support can still be advocated, and design reuse and system design can be substantially simplified.

Further research will focus on

- fail safe synchronization methods [14] when voltage scaling is applied,
- generation and regulation of local supply voltages,
- utilization and enhancements of synthesis tools and layout generator tools to support the proposed Socket Interface for the GALS-LDVS methodology, and
- development of GALS-LDVS applications.

Acknowledgement

The work described in this paper was supported in part by the Norwegian Research Council's Basic Telecommunication Research Programme through research project 119393/431 – WIRAC (Wideband Radio Access).

References

[1] A. Chandrakasan and R.W. Brodersen. Minimizing power consumption in digital CMOS circuits. *Proceedings of the IEEE*, 83(4):498-523, 1995

[2] D. Liu and C. Svensson. Trading speed for low power by choice of supply and threshold voltages. *IEEE J. Solid-State Circuits*, 28(1):10-17, 1993.

[3] C. Hu. Low-voltage CMOS device scaling, *Proc. IEEE International Solid-state Circuits Conference*, ISSCC'94, 1994.

[4] Kai Chen and Chenming Hu. Performance and V_{dd} Scaling in Deep Submicrometer CMOS. *IEEE J. Solid-state Circuits*, 33(10): 1586-1589, October 1998.

[5] T. Njølstad and E.J. Aas. Validation of an accurate and simple delay model, and its application to voltage scaling. *Proc. of IEEE 1998 International Symposium on Circuits and Systems (ISCAS'98)*, Vol. II, pages 101-104, Monterey, USA, 1998.

[6] M. Eisele et al. The Impact of Intra-Die Device Parameter Variations on Path Delays and on the Design for Yield of Low Voltage Digital Circuits. *IEEE Trans. on VLSI*, 5(4): 360-368, December 1997.

[7] P. Favrat et al. A High-Efficiency CMOS Voltage Doubler. *IEEE J. Solid-state Circuits*, 33(3): 410-416, March 1998.

[8] V. Gutnik and A. Chandrakasan. Embedded Power Supply for Low-Power DSP, *IEEE Trans. on VLSI*, 5(4): 425-435, December 1997.

[9] International Technology Roadmap for Semiconductors. 1999 Edition + 2000 Update, http://public.itrs.net/

[10] D. Chapiro. Globally-Asynchronous Locally-Synchronous Systems. *PhD Thesis, Department of Computer Science, Stanford University*, October 1984.

[11] I.E. Sutherland. Micropipelines. *Communications of the ACM*, 32(6): 720-738, June 1989.

[12] K.Y. Yun and A. E. Dooply. Pausible Clocking-Based Heterogeneous Systems. *IEEE Trans. on VLSI*, 7(4): 482-488, December 1999.

[13] D.S. Bormann and P.Y.K. Cheung. Asynchronous Wrapper for Heteregeneous Systems. *Proc. International Conf. Computer Design (ICCD)*, pp. 307-314, Oct. 1997.

[14] J. Muttersbach, T. Villiger, W. Fichtner. Practical Design of Globally-Asynchronous Locally-Synchronous Systems. *Proc. 6th Int. Symp. on Advanced Research in Asynchronous Circuits and Systems (ASYNC 2000)*, pp. 52-59, Eilat, Israel, April 2000.

[15] A. Hemani *et al.* Lowering Power Consumption in Clock by using Asynchronous, Locally Synchronous Design Style. *Proc. ACM/IEEE 36th Design Automation Conference*, 873-878, 1999.

[16] H.B. Bakouglu. Circuits, Interconnections, and Packaging for VLSI, *Addison-Wesley*, ISBN 0-201-06008-6, 1990.

[17] T. Njølstad. SIA Roadmap – trends towards 2010. Invited lecture for the 11th DAK forum, Trondheim, Norway, October 1996.

[18] C. Mead and L. Conway. Introduction to VLSI Systems. *Addison-Wesley*, ISBN 0-201-04358-0, 1980, Ch. 7: System Timing, by C.L. Seitz.

[19] L.S. Nielsen, C. Niessen, J. Sparsø and C. H. van Berkel. Low-Power Operation Using Self-Timed Circuits. IEEE Trans. on VLSI,2(4):391-397, 1994.

[20] T.D. Burd and R.W. Brodersen. Design Issues for Dynamic Voltage Scaling. *Proc. IEEE Int. Symp. on Low Power Electronics and Design (ISLPED'00)*, pp. 9-14, Italy, July 2000.

[21] E. Vittoz. LSI in Watches. 2nd European Solid State Circuits Conference (ESSCIRC'1976), Toulouse, France, September 21-24, 1976. Invited paper. Reprinted in *Solid State Circuits, 1976, Editions du Journal de Physique*, France.

[22] M. Vesterbacka. On Implementation of Maximally Fast Wave Digital Filters. *Ph.D. Thesis, Linköping University. ISBN 91-7871-947-X*, Linköping,

TCL 1009
Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on June 27,2024 at 18:51:27 UTC from IEEE Xplore. Restrictions apply.

Sweden, 1997.

[23] A.E. Sjogren and C.J. Myers. Interfacing Synchronous and Asynchronous Modules Within a High-Speed Pipeline. *IEEE Trans. on VLSI*, 8(5):573-583, Oct. 2000.

[24] T. Kuroda et al. Variable Supply-Voltage Scheme for Low-Power High-Speed CMOS Digital Design. *IEEE J. Solid-State Circuits*, 33(3):454-462, March 1998.

[25] http://www.tele.ntnu.no/wirac/

[26] http://www.transmeta.com/crusoe

[27] L. Lundheim, T. Njølstad, O. Tjore. On the Efficiency of Dynamic Voltage Scaling in a Digital Front End for a Burst-Based Radio Receiver, *Accepted for publication in Proc. of 1st Mobile Communications Summit,* Barcelona, Spain, Sept. 12-15 2001.

TCL 1009
Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on June 27,2024 at 18:51:27 UTC from IEEE Xplore. Restrictions apply.